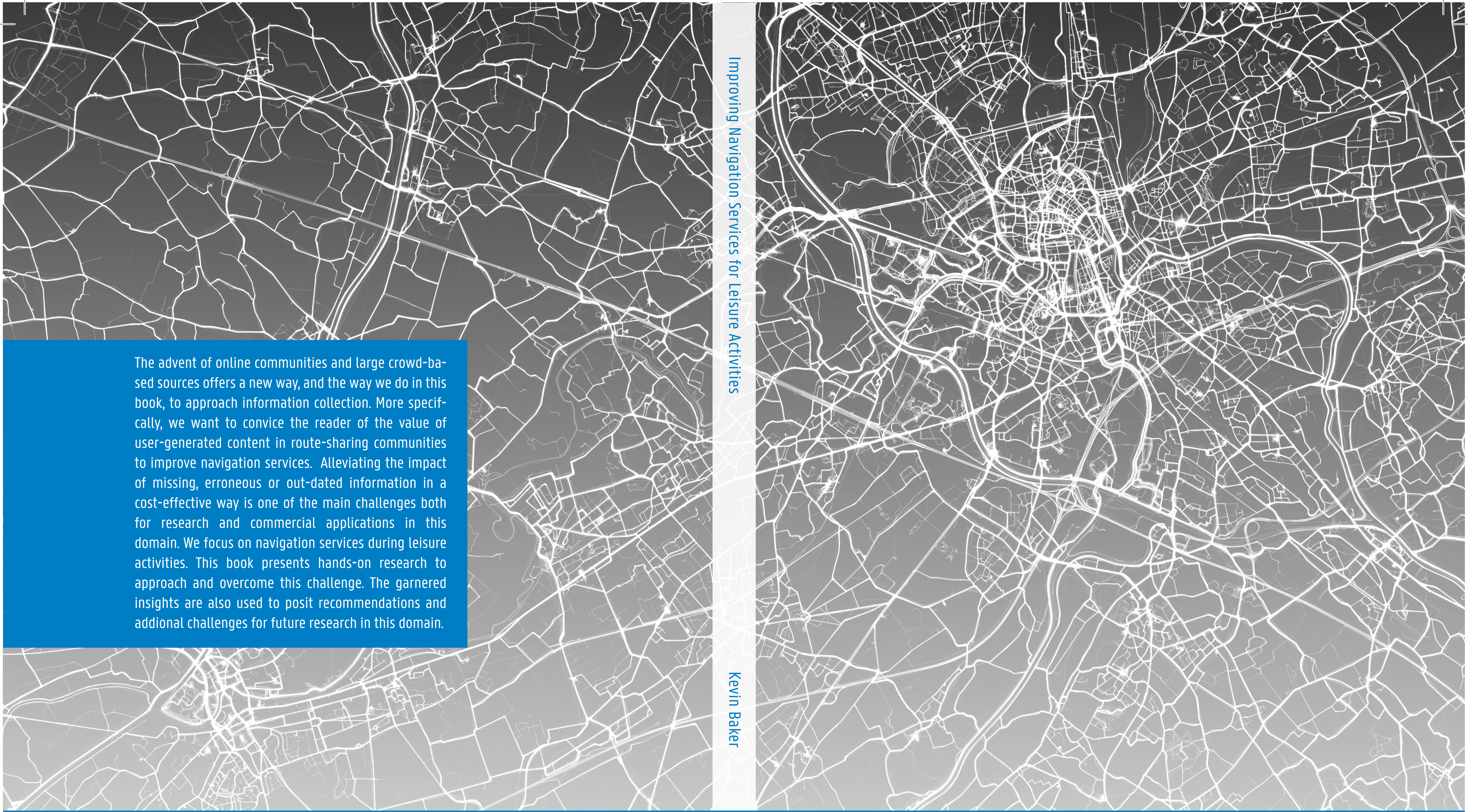


Improving Navigation Services for Leisure Activities

exploiting the full potential of route-sharing communities
and their crowd-based source

Kevin Baker

Improving Navigation Services for Leisure Activities

Kevin Baker

The advent of online communities and large crowd-based sources offers a new way, and the way we do in this book, to approach information collection. More specifically, we want to convice the reader of the value of user-generated content in route-sharing communities to improve navigation services. Alleviating the impact of missing, erroneous or out-dated information in a cost-effective way is one of the main challenges both for research and commercial applications in this domain. We focus on navigation services during leisure activities. This book presents hands-on research to approach and overcome this challenge. The garnered insights are also used to posit recommendations and addional challenges for future research in this domain.

GHENT
UNIVERSITY

route ✱ you

AGENTSCHAP
INNOVEREN &
ONDERNEMEN

Vlaanderen
is ondernemen

Improving navigation services for Leisure Activities

exploiting the full potential of route-sharing communities and their crowd-based sources

Kevin Baker

The research reported in this dissertation was conducted at the Department of Geography, Faculty of Sciences, Ghent University.

Map cover based on routes shared on RouteYou and created with datamaps (https://github.com/ericfischer/datamaps).

UNIVERSITEIT
GENT

Faculty of Sciences
Department of Geography
Academic year 2017-2018

Improving navigation services for Leisure Activities

exploiting the full potential of route-sharing communities and their crowd-based sources

Dissertation submitted in fulfillment of the requirements for the degree of Doctor (Ph.D.) in Science: Geography

Verbeteren van digitale recreatieve navigatiediensten

gebruik maken van het volledige potentieel van online routeplatformen en hun bronnen

Proefschrift aangeboden tot het behalen van de graad van Doctor in de Wetenschappen: Geografie

Kevin Baker

With support of

AGENTSCHAP
INNOVEREN &
ONDERNEMEN

Vlaanderen
is ondernemen

route you
www.routeyou.com

Supervisors

prof. dr. Philippe De Maeyer
Ghent University, *promotor*

prof. dr. ir. Rik Van de Walle
Ghent University, *co-promotor*

dr. Pascal Brackman
RouteYou, *co-promotor*

Members of the examination committee

prof. dr. Veerle Van Eetvelde
Ghent University, *chair*

prof. dr. George Gartner
Vienna University of Technology

prof. dr. ir. Sidharta Gautama
Ghent University

prof. dr. Nico Van de Weghe
Ghent University

prof. dr. Steven Verstockt
Ghent University

dr. Kristien Ooms
Ghent University

Contents

List of Figures

List of Tables

Acknowledgments

If it wasn't for the fun and money, I really don't know why I'd bother.

Terry Pratchett

Over the past four years I had the pleasure of working with and learning from many smart and exceptional people. This book has been strongly influenced by all. Thank you!

In particular, I want to express my gratitude to my supervisors prof. dr. Philippe De Maeyer and prof. dr. ir. Rik Van de Walle for their in-depth discussions, valuable suggestions and help. I am also indebted to RouteYou, and more specifically Pascal Brackman and Philiep De Sutter, for the opportunity to start a PhD and helping me throughout every phase of it. Without their support this book would not have been possible. I am equally grateful to Flanders Innovation & Entrepreneurship (VLAIO) for the funding of my research.

Also a big, big thank you to all my colleagues at RouteYou. Your support, cooperation and evaluative reflections were very important to me. I am equally thankful to all university colleagues. I am most grateful for the many on-topic suggestions and collaborations. Finally, the members of the examination committee are acknowledged for their valuable suggestions to improve the manuscript.

Last but not least, a very big thank you to all my family and personal friends for their support. I want to thank my parents for the chances they gave me. A special thank you to Lies for everything. Thank you also to my son, Sam, keeping me amused and entertained during the final months of my PhD.

Kevin Baker
Gent, December 1, 2017

Abstract

Navigation services are most often used to reduce time spent finding your way during daily travel. They are aimed at improving ones travel experiences. Consider the everyday problem "Which way is the fastest route to get from home to work?". Yet these services also receive a growing attention for activities outside that daily routine, namely during leisure activities. These services provide answers to different question. One example is "Which way will provide me with an adventurous and challenging hiking round trip starting at our vacation house?". This book is about this specific type of navigation services.

The use of navigation services during these leisure activities is conditioned by special needs and trade-offs (e.g. adventurous, challenging). Furthermore, end-users expect a personalized and user-centered experience. Yet the machine-readable maps supporting these services are limited in their ability to answer this growing demand. Missing, erroneous or out-dated information are an important limiting factor in current expert-based models supporting these services. Alleviating this impact in a cost-effective way is one of the main challenges both for research and commercial applications in this domain.

The advent of online communities and large crowd-based sources offers a new way, and the way we do in this book, to approach this challenge. Within this dissertation we gauge the value of online route-sharing platforms. This book presents three cases using route-sharing communities and their specific user-generated content. By applying geographical and contextual analyses, we elicit new ancillary data.

Supported by our findings, we want to convince the reader of the value of this ancillary data to improve navigation services for leisure activities. Both active user interaction and data-driven approaches, or passive information collection, have an observable potential to harvest contextual information from these new data sources. In future valorisation, the synergy between both crowd-sourcing approaches and expert-based models will be imperative. In conclusion, we also want to highlight the value of geographical and contextual analyses in other facets of route-sharing platforms, and by extension other social media, to personalize and improve their user-centered design.

1
General introduction

> *"What do you consider the largest map that would be really useful?" "About six inches to the mile." "Only six inches!" exclaimed Mein Herr. "We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country on the scale of a mile to the mile!" "Have you used it much?" I enquired. "It has never been spread out, yet," said Mein Herr: "the farmers objected: they said it would cover the whole country and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well."*
>
> Lewis Carroll, Sylvie and Bruno Concluded

1.1 Research context

Maps have always had value in solving complex spatial problems. Consider finding your way. Maps have the ability to solve this problem in a clear, quick and comprehensible manner. However, the rigor with which we actually map the world and, hence, the map's problem-solving ability have always had their limits. Information is inevitably left out. While there is a well-known need to generalize reality during the practice of making maps (Koláčný, 1969), the notion of remoteness can also serve to explain why *reality gaps* exist. At one level this may sound trivial. The tangible barriers of latitude, altitude, continentality or insularity have long been discerned as reasons for missing or *under-mapped* places in representations of the world. In an attempt to deal with this remoteness, humans have built tools and technology to bridge the distance to these places and, as a result, weakening its impact on contemporary map making (Blackmore and Harley, 1980; Bocco, 2016).

However, map making is constrained by many other aspects, not the least of which is cost. In that sense, we posit another initiator for remoteness in map making and, hence, an explanatory variable for missing or out-dated information in maps: the combined value to all users is less than the cost to capture this information. Most contemporary mapping companies were established under the presumption that an intricate database describing the geographical domain and its derived cartographic products have a value that justifies the costs to map these remote places. Until the beginning of the 21th century, this was an actual and viable business case. Two things changed this situation.

In 2005 Google Maps was launched, which was to all intents and purposes free for use. While not the sole online map launched in mid 2000s, Google Maps is a clear example of a business case which disruptively changed the way we look at mapping. Google used its maps to target advertising better and, hence, create revenue. While there is no such thing as a free lunch, these maps created an expectation of freely accessible maps for end-users.

Secondly, fueled by the burgeoning popularity of location-based services, the demand for more detailed digital road databases has increased rapidly in the last years. Finding your way, for example, has moved well beyond paper maps and in-car navigation devices. End-users expect personalized services with up-to-the-minute detail meeting their expectations and local knowledge for a gamut of activities. These expectations range from safe, scenic and attractive route suggestions to *fresh* and coherent information about attractions along your sightseeing walk (e.g. its cultural value, historical significance or more practical information such as opening hours). Of particular concern is that navigation services need more than skilfully designed online and mobile tools and technology to succeed. They need to be allied with the right machine-readable maps (Fu et al., 2006) and comprehend a complicated interplay of special needs and trade-offs.

In this book, we focus our attention on navigation services and how they are impacted by the above-described situational change. In particular, we study, as the title of this dissertation implies, navigation services for leisure activities and defining where and how improvements are possible. Both aspects are the cornerstones of the research in this book and shape its context and aim.

1.1.1 Problem statement

Leisure is most often defined as the use of free time, time not spent working or occupied, for enjoyment. While this is a highly interpretable and fuzzy classification by activity, we use this term to identify a niche in navigation services used in combination with leisure activities or active recreational pastime. In the remainder of this book, we often refer to this niche as activity-specific navigation services. Consider, for example, finding your way during a sightseeing walk or a leisurely sports activity such as cycling or hiking.

Navigation services in automotive engineering and industry clearly have a head start on these activity-specific services. Research and development has moved well beyond tools, technology and data to find your way. New potentially disruptive changes such as in-car sensing and processing together with semi- or fully-autonomous vehicles receive much attention to make driving simpler and safer (Fagnant and Kockelman, 2015; Schmidt et al., 2011). These navigation services are highly functional, gauging their tangible value by reducing time to destination and travel costs (e.g. gas money).

Contrastingly, the complexity of activity-specific navigation services during leisurely sports, tourism and other recreational pastime lies in gauging soft factors of the streetscape and meeting aforementioned expectations such as an attractive route suggestion. Herein lies the problem and, hence, the need for improvements. An interplay of several reasons can be put forward why these services are nowadays more prone to being hampered by the introduced remoteness than well-known automotive navigation services. We posit three main drivers.

First, there is a very skewed distribution in the hierarchy of roads if ordered by proportional length in a real-world road network. There are many more local roads, paths and dirt roads than freeways, arterials and collectors put together (an 80/20 ratio as seen in Figure 1.1). Also, a large portion of road networks have traffic restriction, most often inaccessibility for motorized vehicles. Both for reasons of cost and limited time, these places can, as a result, be deemed remote and remain *under mapped* as the expected traffic on these parts of the road network, and the expected combined value or revenue created by up-to-date information, is much lower. This budgetary choice, however, has a large performance impact on the aforementioned navigation services for leisure activities. Local roads, paths and dirt roads are preferred connectors in route suggestions generated by these services.

A second reason can be found in the physical boundaries of contemporary mapping strategies. Remote sensing from mobile, multi-sensor platforms (e.g. mobile mapping vehicles) have significantly lowered the validation and verification cost of procured information. However, the limited access to the roads and paths of interest for these leisure activities hampers capturing road infrastructure and its direct surroundings in a standardized way and boost the time-to-market performance of new and updated map products. Consider Google Street View. Probably one of the most popular and widespread examples of this mobile mapping technology, but still for the majority of its street-level imagery limited to motorized vehicle-accessible roads. As such, on-site mapping teams and human input remain important in the tedious validation and verification process and, hence, increases both cost and time to procure up-to-date and verified information for the services of interest in this research.

Finally, contemporary mapping standards condense information in predefined and constrained attributes. However, as the complexity of queries to these activity-specific services increases, it becomes equally hard to distinguish these properties and condense the necessary information into object attributes. Hence, it is not only difficult to collect the necessary information, but it becomes equally hard to describe these machine-readable maps in actionable and scalable structures.

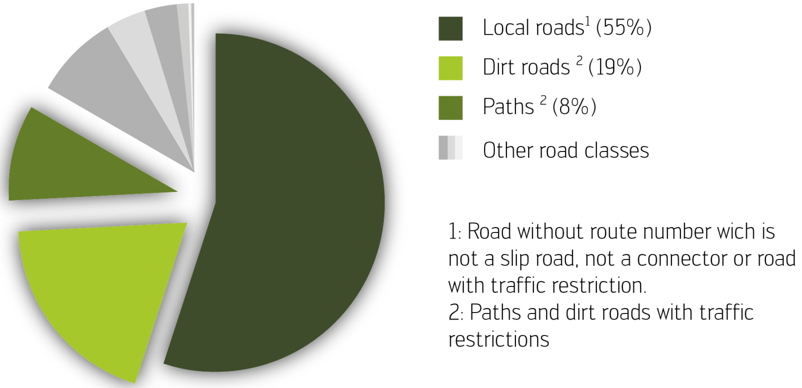


Figure 1.1: Representation of the proportional length distribution of road segments in a real-world network in Belgium (source: NGI - 1 : 10 000 road network).

In this regard, interest in both crowd-based sources and data-driven research has grown continuously, aiming to better understand citizens or consumers and to improve and personalize services in a creative and cost-effective way. We live in a datafied world where many aspects of our daily life are lumped into data streams and shared within diverse online communities. Against the backdrop of a burgeoning amount of devices with communication and sensing capabilities (Srivastava et al., 2012) and a growing digital literacy among its users, our data doppelgängers describe the seemingly *uniform distancelessness world*[1] in an unprecedented pace and detail.

[1] Metaphysical term, a time-space compression as a result of technology (Heidegger, 1971)

Every day, members of the general public (i.e. crowds) are, knowingly or unknowingly, generating georeferenced facts amassing in so-called *data wells* (Goodchild and Li, 2012). This process of creating, collecting and disseminating geographic information on the web has first been coined Volunteered Geographic Information (VGI) by Goodchild (2007). In the last ten years, this term has become a suitcase term for a variety of spatial data and its derived information that is voluntary made available (Elwood et al., 2012) and is, almost always, used interchangeably with spatial user-generated content or crowd-sourced information. For a more in-depth review of specific terms in this research field and their differences and nuances, we refer to the review paper of See et al. (2016).

The potential of this new type of data creation and the resulting *data wells* has been highlighted in both research and commercial applications. However, these crowd-based sources can not be seen as a commodity which can be bought and sold to bridge the described *reality gaps* seen in the practice of making maps. Besides its latent knowledge and great promise, there are challenges and limitations in using these sources that need to be understood.

As such, the overarching theme in this dissertation is the potential of these sources to create ancillary data to improve the problem-solving ability of machine-readable maps in the current myriad of location-based services. In particular, we study how to exploit this potential in route-sharing communities and use their crowd-based sources describing our leisured self to create trustworthy, scalable and actionable knowledge to bridge the aforementioned *reality gaps*. The approach in doing so is focused on two specific types of user-generated content managed in these communities (i.e. routes and points of interest; see Section 1.1.3), aiming to improve activity-specific navigation services.

1.1.2 Valorisation potential

By reframing our main objective to "How can route-sharing platforms exploit the full potential of their community to streamline efforts to improve the fitness of their services?", we also imply there is an optimal solution to this problem. The implementation difficulties of this solution, that already came apparent in the above-described problem statement, can also be grasped in three high-level performance indicators for navigation services in general.

- **Reality** - navigation services and their underlying machine-readable maps should present real-world situation. In other words, the calculated activity-specific fitness for use of an actual path or road represents its real-world suitability for a specific user type. This implies that we should be able to distill the necessary information to maintain these services, track changes and characterize blind spots in the current expert-based models.

- **Time** - End-users expect near real-time updates in services. This is linked with the time to market of updated maps and is strongly connected with the previous performance indicator. As such, performance can only be attained by a continuously growing amount of data to 'learn' from. This flywheel effect is imperative in keeping these approaches sustainable in time.
- **Cost** - Cost should be kept to a minimum to maintain an approach. Resources should be focused on managing the value-creation process instead of actively intervening in this process.

The interplay between these performance indicators forms the logical basis of this research and will be addressed throughout this dissertation. The less obvious yet equally important question to ask within this dissertation is which trade-offs need to be made to attain a sub-optimal but satisfactory solution to the introduced problem for all stakeholders in the value chain (e.g. end-users, data providers and service providers).

This dissertation is a reference work of demand-driven scientific research addressing the unique opportunities, challenges and limitations within the above-described problem statement. Acquiring an intricate road database to facilitate navigation services for leisure, sports and tourism activities is one of the main focuses of the industrial partner in this dissertation, RouteYou. RouteYou manages a route-sharing platform where end-users create, share and use user-generated content in this activity-specific context. Of specific interest to the main objective of this dissertation is the model RouteYou maintains to calculate an *attractiveness* of a road tailored to the needs of specific leisure, sport and tourism activities. Attractiveness can be defined as features or qualities of a road or path which arouses an interest in an activity-specific context. For example, a road could be pigeonholed as *attractive to race cyclists* if it has low traffic counts, lies in a natural setting and has an asphalt surface. Figure 1.2 gives an overview of situations with prevailing characteristics and proxies used in RouteYou's model.

The following chapters discuss certain lacking or irretrievable information for this and similar models and devise an approach to bridge these specific *reality gaps*. The relevance of this research, however, does not limit itself to the niche market of RouteYou or route-sharing platforms. Creative and cost-effective ways to define the suitability, safety or popularity of places during recreational activities have been receiving growing attention in both academic research and popular media about tourism and leisure, health, policy making or spatial planning. Throughout this dissertation we enrich our research context with related work, emphasizing the relevance of this dissertation in a broader frame of reference.

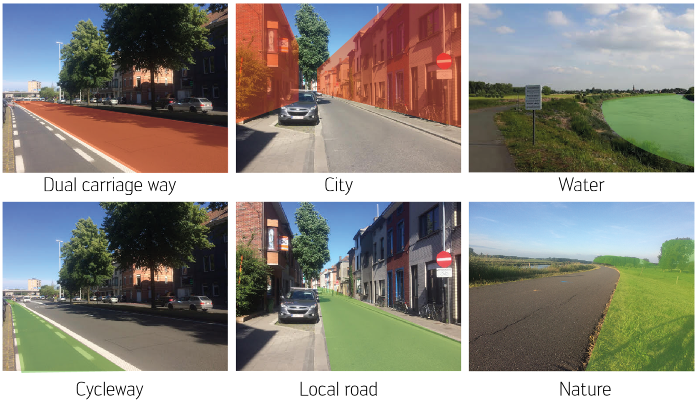
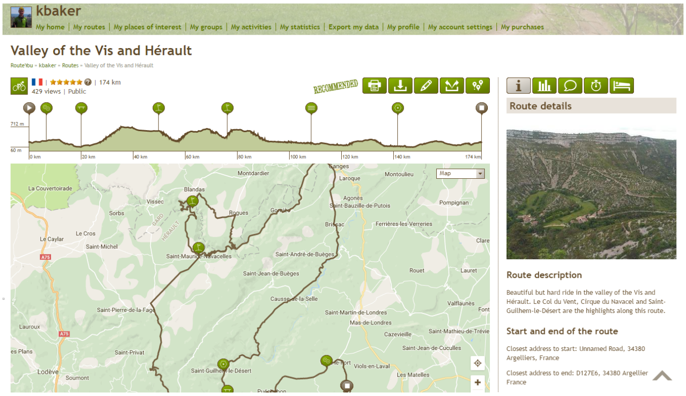

Dual carriage way
City
Water
Cycleway
Local road
Nature

Figure 1.2: Overview of several characteristics determining RouteYou's *attractiveness* model. Green features are positive characteristics; red features are negative characteristics.

In the following sections we give a brief introduction to route-sharing communities and their user-generated content. Next, we present the challenges in using these crowd-based sources to address the presented research objectives.

1.1.3 Route-sharing platforms and their communities

According to the typology of social media by Kaplan and Haenlein (2010), route-sharing communities are, on a first level, content communities with pleasure-seeking and utilitarian values. For example, end-users explore activities in the region of their next vacation or plan their coming Sunday morning ride. In these communities, self-presentation (i.e. the ability to control the impression one makes) is expected to be subservient to the created and shared content in these networks. However, self-presentation becomes increasingly more important both as an individual or within a business-to-business or business-to-customer network. Hence, route-sharing platforms are evolving into social networking sites. These platforms allow end-users to manage their own channel to create, collect and present their location-based information enabling self-presentation and digital storytelling. Within the following chapters in this dissertation we focus on two types of user-generated content within these communities: (i) routes and (ii) points of interest. Following paragraphs introduce both key concepts and highlight their importance for route-sharing communities and, more specifically, RouteYou.


kbaker
My home | My routes | My places of interest | My groups | My activities | My statistics | Export my data | My profile | My account settings | My purchases
RouteYou » kbaker » Routes » Valley of the Vis and Hérault
| ★★★★★ | 174 km
429 views | Public
RECOMMENDED
712 m
60 m
0 km
20 km
40 km
60 km
80 km
100 km
120 km
140 km
174 km
Map
Route details

Valley of the Vis and Hérault

Route description

Beautiful but hard ride in the valley of the Vis and Hérault. Le Col du Vent, Cirque du Navacel and Saint-Guilhem-le-Désert are the highlights along this route.

Start and end of the route

Closest address to start: Unnamed Road, 34380 Argelliers, France

Closest address to end: D127E6, 34380 Argellier France

Figure 1.3: Example of user-generated content within the route-sharing community of RouteYou.

Similar to the reason for travel as proposed by Bovy and Stern (1990), a **route** occurs because "different things exist in different places (p 1)". Within the specific context of this dissertation, a route is a representation of a leisure activity and is used both as guidance aid before or during this activity and as a log of previous activities. Routes on RouteYou originate from (i) RouteYou's route-planning tool, (ii) user uploads in well-known standards such as GPX[2] or FIT[3] or (iii) third-party applications using RouteYou's services and back end. From a technical perspective, a route consists of a set of legs describing a particular way connecting different places. A leg describes a real-world shortest path between two places. In addition to the technical tools to create these route data, RouteYou implements different mechanisms to enable storytelling during this route-creating process. Providing a description and using the tagging system to link a route to a type, theme, group or characteristic allows users to create so-called rich routes. In doing so, every route has its own story to tell, shaped by random agents and complex factors, such as weather, scenery, affordance of the street- and sound-scape, physical difficulty or group dynamics (Bull, 2006; Damant-Sirois et al., 2014; Dill and McNeil, 2013; Downward and Lumsdon, 2001; Pijanowski et al., 2011; Tucker and Gilliland, 2007; Winters et al., 2010). At the same time, this linking

[2] GPX or GPS Exchange Format is an XML standard to describe and exchange route information

[3] Similar to GPX, FIT or Flexible and Interoperable Transfer is a protocol to describe route information

provides valuable information for research, enriching the possibilities of contextual analysis and information retrieval.

Next, a route can have one or more **points of interest** which lie in the vicinity of the proposed way in the route. Points of interest or POI are a well-known concept in popular location-based services. POI can easily, and probably naively, be described as location-aware information such as plain text, images or video's. RouteYou uses contextual (e.g. theme-specific or activity-specific) POIs to enrich the experience before and during an activity and, hence, enabling storytelling through map exploration of a route-specific POI set. Figure 1.3 shows an example of a rich route created with the RouteYou platform. At the time of writing RouteYou holds nearly 4.5 million routes and 2 million points of interest linked to more than 250 000 activated accounts. On average, the website of RouteYou receives 1 million visits per month.

1.1.4 Challenges in using route-sharing platforms

Researchers and academics increasingly turn to crowd-based sources to generate new interdisciplinary research output. In recent years, a broad range of research domains such as social studies and psychometrics (Kosinski et al., 2015; Lazer et al., 2009), healthcare and medicine (Alemdar and Ersoy, 2010), or disaster management (de Albuquerque et al., 2016; Simon et al., 2015) have used these sources to come to new insights. Within a geographic research context, Sui and Goodchild (2011) and Goodchild (2011), among others, recognized the need to find new ways of fusing social media and Geographic Information Systems (GIS) and produce protocols and procedures to link these crowd-based sources to fill gaps in contemporary spatial data infrastructures. While above-described problem statement (see Section 1.1) has introduced the opportunities of this dissertation, several challenges can also be identified. Challenges lie in determining (i) the role end-users are willing to play, (ii) the suitability of harvested data, (iii) the *dirty* nature of harvested data and (iv) the comprehensibility of inferred information. The following paragraphs introduce these challenges and underpin our statement that crowd-based sources can not just be seen as a commodity. Understanding these challenges within our research context and presenting methodological approaches while mitigating their impact is an important goal of this dissertation.

Challenge I - Current research focusing on online communities and their crowd-based sources document two forms of information collection, namely passive and active contributions (She et al., 2015). Both approaches have proven their individual usefulness in creating new information. First, data mining, machine learning and knowledge-discovery techniques (see for example Bishop (2006)) have grown very popular in recent decades as means to passively collect information from shared records (Mayer-Schönberger and Cukier, 2013). Using these data-intensive techniques within a geographical context are one way to exploit the potential in crowd-based sources and address the main objective of this dissertation. Contextual analysis of places occurring in these crowd-based sources can be used to infer new information. This contextual analysis is based on two well-studied geographical principles (Goodchild, 2011; Sui, 2004; Tobler, 1970): (i) the likelihood of finding relations between phenomena occurring at the same location and (ii) the tendency that near phenomena are more related than distant ones. Secondly, crowd-sourcing projects, requiring active interaction with and within IT-mediated crowds to infer new information, have become equally popular (Doan et al., 2011). Consider Google's Local Guides[4] or community-based traffic applications such as Waze[5]. Many efforts are now being made to integrate these crowd-sourcing approaches in commercial applications. Crowdsourcing is used both for reasons of information retrieval and marketing. Van Belleghem (2015) noted that the ultimate goal of these approaches is to make the customer a part of the company and create an emotional customer relationship. Route-sharing platforms thrive on this relationship. We study how knowledge-sharing behavior can be fostered beyond content creation. As such, the first challenge lies in understanding the trade-offs when implementing and combining these methods to address the main objective of this dissertation.

Challenge II - An intrinsic characteristic and, hence, a second challenge of this and similar research is the fact that we use an online community for another purpose than it was conceived for. Similar to other social media, route-sharing platforms are purpose-built applications, most often in a commercial setting. They are user centered and give end-users the tools and freedom to create their content. Furthermore, these tools are developed in a way that is based on practical rather than theoretical considerations and aim at increasing user retention or volume of shared content. Very few, if any, of these platforms are designed as a research instrument and, hence, lack a fully controllable environment to do empirical research. Due to this lack of control, data collection is prone to serendipity. Within leisure-oriented applications, such as route-sharing communities, very personal contributor characteristics such as vacation, spare time or social context can have an impact on contribution patterns. Panciera et al. (2010) noted that research in online communities is often fostered or hampered by what researchers cannot see or measure. Furthermore, we have to be aware that data collection is done in a self-selected population (e.g. RouteYou's end-users) and precedes

[4] https://maps.google.com/localguides

[5] https://www.waze.com

both hypotheses predictions and experiment design. This biased nature of the collected data emphasizes the need to identify the right research question to which these data can be used with reasonable generality (Miller and Goodchild, 2015; Romanillos et al., 2016).

Challenge III - The third challenge is strongly intertwined with the previous one and addresses the *dirty* or *messy* nature of the collected data. A lack of data semantics, structure or level of detail hampers reuse of collected data sources (Li et al., 2016). This lack can result from the second challenge, a difference in creation and reuse purpose of the created content in route-sharing communities. Consider a specific route. This type of content can meet the personal requirements of a *route author* and his goals, but can lack more general quality metrics such as positional accuracy for map making reuse. Furthermore, end-users of a route-sharing platform can lack experience, intrinsic motivators or incentives, knowledge, time or attention to address inaccuracies in both their content or the shared content of other users. As such, omissions, errors or other inaccuracies are inherent and, most often, undocumented in the content-creation procedure and data collection (e.g. Antoniou and Skopeliti, 2015; Flanagin and Metzger, 2008; Goodchild and Li, 2012; Senaratne et al., 2017). By integrating user interaction through comments or content-rating systems (think of 'likes', 'stars' or 'kudos'), many social media try to make the quality control in these communities self-regulated. However, the self-regulation process often fails or is insufficient to meet user expectations or predefined standards. Despite a growing amount of research focusing on trust and quality propagation of user-generated content for research, Antoniou and Skopeliti (2015) noted that more intrinsic quality and trust measures such as data lineage or contributor history are still far from solving the question "How good are these crowd-based data sources?". Li et al. (2016) and Sui and Goodchild (2011) noted that the solution to this challenge could lie in the data itself, that is redundancy within crowd-based sources to alleviate and mitigate quality issues.

Challenge IV - Miller and Goodchild (2015) identify a fourth and final challenge in "how to build data-driven models that are both true and understandable". Mayer-Schönberger and Cukier (2013) argue that as the volume of crowd-based sources increases, it becomes equally hard to infer causality, explanations and, as a result, unifying theories and models. Similar to a Google-like approach of page indexation or Facebook's EdgeRank algorithm to model its news feed, the data deluge is reshaping research, crunching large numbers of records to infer an educated guess without this unifying framework. To extend and improve RouteYou's attractiveness model in its current form, however, it is imperative to know why a road or path is preferred above another. For more in depth discussion of this Big Data paradigm shift and the implications of data-driven or data-intensive research we refer to extensive discussions in literature including, but not limited to, the work of Kelling et al. (2009), Boyd and Crawford (2012), Kitchin (2014) and Miller and Goodchild (2015). Many data-driven companies such as Facebook use the combination of data-intensive approaches, such as their machine-learned news feed ranking algorithm, and human input from *feed quality panels* (i.e. a paid workforce to evaluate algorithms' output), trying to make their data-driven models both true and understandable. As such, this challenge is strongly interwoven with the first challenge introduced in this section.

1.2 Synopsis

This research makes use of an online community and its user-generated content and is conducted for the particular case of route-sharing communities. The goal of this dissertation is twofold. First, we want to present actual use cases of a route-sharing community and its data to address the introduced problem statement. Second, opening the door to new ways of understanding, we study the introduced challenges in using route-sharing communities in research in Section 1.1.4. Documenting methodological approaches that alleviate their impact is imperative to create future valorisation trajectories. This clear-cut goal allows us to posit three overarching research questions which can be addressed within this frame of reference.

1.2.1 Research questions

Following section introduces each research question, gives a brief topical summary and links them back to the introduced challenges and performance indicators.

RQ1: How can large route sets maintained on route-sharing platforms be used to improve activity-specific navigation services?

> This first research question draws attention to routes and strings the four aforementioned challenges together. We focus on set generation, that is the action of semantically and spatially grouping routes. By studying the movements along a road network described in these route sets, we aim at creating ancillary data for navigation services. If successful, this approach has the potential to fill blind spots in current machine-readable maps for niches of leisure activities where navigation services are receiving a growing interest. In doing so, we focus on sourcing route choices consolidated in a route set to increase the *reality* of route suggestion and *time to market* of valuable information while reducing the *costs* to detect

inconsistencies in the expert-based model of navigation services. This research question is used to guide the discussion of opportunities and fallacies in this rationale.

RQ2: Which opportunities do POIs provide to enrich current navigation services beyond well-known map exploration?

Map exploration, the action of searching and consuming information presented on maps, has become very important in web-enabled services focusing on location-based information or POIs. As mentioned before, navigation services use a similar map-exploration approach to show information along your route to enrich the experience. This research question, in contrast, emphasizes the value of POI to guide route suggestion for specific leisure activities and addresses the challenges when reusing POI in our research contexts. As such, we focus on improving the fitness for use of these route suggestion, that is the *reality* in navigation services. While all four challenges have links with this research question, the main focus lies on the third and fourth challenge.

RQ3: Do route-sharing platforms have engaged end-users that are able to help improve the problem-solving ability of maps? If so, what is the value of active user interaction?

The third and final research questions is driven by the first and fourth challenge. Engaged end-users are a potential source of very specific, topical and local information. Moreover, these end-users are, most often, willing to share this information without any financial reward, reducing *costs* of information collection. We seek to discover the added value of this active user interaction in contrast to data-driven approaches and passive information collection. We use this research question to gather insights on both the *reality* and *time* performance indicator. Do end-users provide information which can not be harvested by passive information collection? What can we expect of these end-users (e.g. time to market, spatial coverage).

1.2.2 Outline

This work consists of five chapters. The corpus of this work is structured around three specific use cases[6]. All three use cases fulfill a different need while addressing the main objectives and research questions presented in this

[6] Chapters are based on original research papers. Because these chapters can also be regarded as individual, multi-authored papers, there might exists some overlap. However, all three chapters emphasize their specific frame of reference and link back to the problem statement

introduction. Figure 1.4 gives a schematic overview of the different chapters and their topical focus.

Chapter 2 presents a data-driven approach using routes collected through user uploads and third-party applications using RouteYou's services and back end (RQ1). To enhance existing navigation services for leisure activities, we describe a methodology to integrate a user's perspective in contemporary routing engines designed for these services. Based on movements condensed in a route set a generic popularity model is built approximating a cost to traverse a road in the network for an activity type (i.e. a cyclist's perspective).

Similar to the previous use case, Chapter 3 proposes a method to improve and personalize path suggestion services in a road network for a specific leisure activity based on aggregation of a crowd-based source collected in a route-sharing community. In contrast, however, this work studies the utilization of a set theme-specific POI to infer this fitness for use (RQ2). The goal of this chapter is twofold. First, we focus on harvesting theme-specific information lacking an actionable structure. We present generic building blocks to geographically enrich this information to become actionable within the main objective of this dissertation. Next, we present an approach to aggregate and consolidate this information to be useful in the aforementioned path suggestion services.

In Chapter 4 we direct our attention on active user interaction and its value in this dissertation (RQ3). We analyze an implementation of a web-based feedback tool to query an individual's cognitive map to infer road-network updates. We use this individual's shared routes as a proxy for his cognitive map (RQ1) and infer possible errors, ommissions or inaccuracies in a digital road network where user feedback is required. All routes created by 325 active contributors through RouteYou's route-planning tool and user uploads were used in this case study.

The final chapters link these use cases back to the introduced research framework and present a general discussion and conclusion. Aside from the results garnered in the three use cases, we use the insights collected during the four years of project-based research prior to this dissertation to bring this discussion to address our research goals from different angles and take the long view.

In conclusion to this introductory chapter, we want to provide the frame which fostered this research. Research was funded by Flanders Innovation & Entrepreneurship (VLAIO) and RouteYou under a Baekeland mandate. The purpose of this funding type is to support research that, if successful, has the potential to offer an added value to the company involved in the project. Together with the academic research partners supporting this dissertation,

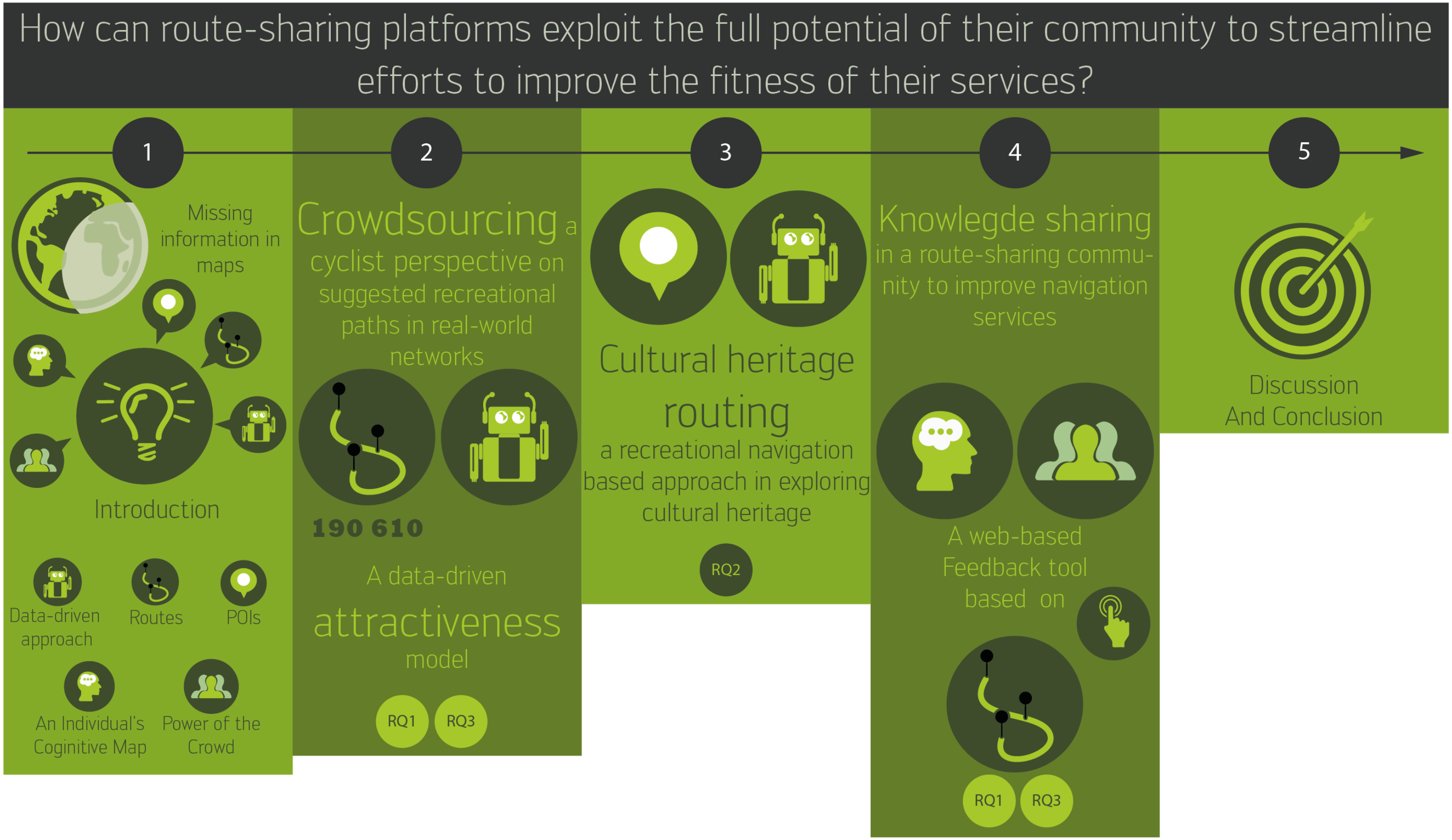

How can route-sharing platforms exploit the full potential of their community to streamline efforts to improve the fitness of their services?
1
Missing information in maps
Introduction
Data-driven approach
Routes
POIs
An Individual's Coginitive Map
Power of the Crowd
2
Crowdsourcing a cyclist perspective on suggested recreational paths in real-world networks
190 610
A data-driven
attractiveness
model
RQ1
RQ3
3
Cultural heritage
routing
a recreational navigation based approach in exploring cultural heritage
RQ2
4
Knowlegde sharing
in a route-sharing community to improve navigation services
A web-based Feedback tool based on
RQ1
RQ3
5
Discussion And Conclusion

Figure 1.4: Outline of this dissertation

this resulted in a multidisciplinary research framework. The growing synergy between geographic research, internet technology and data science has become very clear in recent decades and also resulted in this dissertation. The Department of Geography of Ghent University, one of the research partners in this dissertation, has a long-standing expertise on in- and out-door navigation applications, mapping and Geographic Information Science. In addition, the expertise of the Internet Technology and Data Science Lab (IDLab) of Ghent University, the second research partner, arches over different research areas such as machine learning and data mining, semantic intelligence, distributed intelligence for IoT and multimedia processing. As both research fields become more integrated in daily life, it also becomes imperative to move beyond topical boundaries and domain-specific expertise. The research activities presented in this dissertation combine these fields of study in the approach to the introduced topic. This resulted in a project-based and application-oriented research merging geography with internet technology and data science. In parallel to the presented use cases, efforts have also been made to implement the presented tools, methods and ideas in the thriving community of RouteYou. In addition, valuable insights and tools were also gathered during collaboration on conference or workshop papers and the guidance I provided in four internships and four Master's theses (see Appendix A).

References

Alemdar, H. and Ersoy, C. (2010). Wireless sensor networks for healthcare: A survey. *Computer Networks*, 54(15):2688 – 2710.

Antoniou, V. and Skopeliti, A. (2015). Measures and indicators of vgi quality: an overview. In *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume II-3/W5, pages 345–351, Göttingen, Germany. Copernicus publications.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Blackmore, M. and Harley, J. (1980). *Concepts in the history of cartography: review and perspectives*. Macmillan and Co., London, United Kingdom.

Bocco, G. (2016). Remoteness and remote places. a geographic perspective. *Geoforum*, 77:178 – 181.

Bovy, P. H. and Stern, E. (1990). *Route Choice: Wayfinding in Transport Networks*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Boyd, D. and Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5):662–679.

Bull, C. (2006). Racing cyclists as sports tourists: The experiences and behaviours of a case study group of cyclists in east kent, england. *Journal of Sport & Tourism*, 11(3-4):259–274.

Damant-Sirois, G., Grimsrud, M., and El-Geneidy, A. M. (2014). What's your type: A multidimensional cyclist typology. *Transportation*, 41(6):1153–1169.

de Albuquerque, J. P., Eckle, M., Herfort, B., and Zipf, A. (2016). Crowdsourcing geographic information for disaster management and improving urban resilience: an overview of recent developments and lessons learned. In Capineri, C., Haklay, M., Huang, H., Vyron, A., Kettunen, J., Ostermann, F., and Purves, R., editors, *European Handbook of Crowdsourced Geographic Information*, pages 309–321. Ubiquity Press, London, United Kingdom.

Dill, J. and McNeil, N. (2013). Four types of cyclists? *Transportation Research Record: Journal of the Transportation Research Board*, 2387:129–138.

Doan, A., Ramakrishnan, R., and Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Commun. ACM*, 54(4):86–96.

Downward, P. and Lumsdon, L. (2001). The development of recreational cycle routes: an evaluation of user needs. *Managing Leisure*, 6(1):50–60.

Elwood, S., Goodchild, M. F., and Sui, D. Z. (2012). Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Annals of the Association of American Geographers*, 102(3):571–590.

Fagnant, D. J. and Kockelman, K. (2015). Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, 77:167 – 181.

Flanagin, A. J. and Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*, 72(3):137–148.

Fu, L., Sun, D., and Rilett, L. (2006). Heuristic shortest path algorithms for transportation applications: State of the art. *Computers and Operations Research*, 33(11):3324 – 3343.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.

Goodchild, M. F. (2011). Challenges in geographical information science. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 467(2133):2431–2443.

Goodchild, M. F. and Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1:110 – 120.

Heidegger, M. (1971). *Poetry, Language, Thought. Translations and Introduction by Albert Hofstadter. 1st Perennial Classics edition.* Harper Perennial Classics, New York, USA.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59 – 68.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., and Hooker, G. (2009). Data-intensive science: A new paradigm for biodiversity studies. *BioScience*, 59(7):613.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):1–12.

Koláčný, A. (1969). Cartographic information—a fundamental concept and term in modern cartography. *The Cartographic Journal*, 6(1):47–49.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., and Stillwell, D. (2015). Facebook as a Research Tool for the Social Sciences Opportunities, Challenges, Ethical Considerations, and Practical Guidelines. *American Psychologist*, 70(6):543–556.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Computational social science. *Science*, 323(5915):721–723.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., and Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:119 – 133.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think.* Houghton Mifflin Harcourt Publishing Company, New York, USA.

Miller, H. J. and Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, 80(4):449–461.

Panciera, K., Priedhorsky, R., Erickson, T., and Terveen, L. (2010). Lurking? cyclopaths?: A quantitative lifecycle analysis of user behavior in a geowiki. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1917–1926, New York, NY, USA. ACM.

Pijanowski, B. C., Villanueva-Rivera, L., Dumyahn, S. L., Farina, A., Krause, B. L., Napoletano, B. M., Gage, S. H., and Pieretti, N. (2011). Soundscape ecology: The science of sound in the landscape. *Bioscience*, 61(3):203–216.

Romanillos, G., Austwick, M. Z., Ettema, D., and Kruijf, J. D. (2016). Big data and cycling. *Transport Reviews*, 36(1):114–133.

Schmidt, A., Paradiso, J., and Noble, B. (2011). Automotive pervasive computing. *IEEE Pervasive Computing*, 10(3):12–13.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.-Y., Milčinski, G., Nikšič, M., Painho, M., Pődör, A., Olteanu-Raimond, A.-M., and Rutzinger, M.

(2016). Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information*, 5(5).

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., and Haklay, M. M. (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31(1):139–167.

She, B., Zhu, X., Ye, X., Guo, W., Su, K., and Lee, J. (2015). Weighted network voronoi diagrams for local spatial analysis. *Computers, Environment and Urban Systems*, 52:70 – 80.

Simon, T., Goldberg, A., and Adini, B. (2015). Socializing in emergencies—a review of the use of social media in emergency situations. *International Journal of Information Management*, 35(5):609 – 619.

Srivastava, M., Abdelzaher, T., and Szymanski, B. (2012). Human-centric sensing. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 370(1958):176–97.

Sui, D. and Goodchild, M. (2011). The convergence of gis and social media: challenges for giscience. *International Journal of Geographical Information Science*, 25(11):1737–1748.

Sui, D. Z. (2004). Tobler's first law of geography: A big idea for a small world? *Annals of the Association of American Geographers*, 94(2):269–277.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240.

Tucker, P. and Gilliland, J. (2007). The effect of season and weather on physical activity: A systematic review. *Public Health*, 121(12):909 – 922.

Van Belleghem, S. (2015). *When Digital Becomes Human: the Transformation of Customer Relationships*. Kogan Page Limited, London, United Kingdom.

Winters, M., Davidson, G., Kao, D., and Teschke, K. (2010). Motivators and deterrents of bicycling: comparing influences on decisions to ride. *Transportation*, 38(1):153–168.

2

Crowdsourcing a cyclist perspective on suggested recreational paths in real-world networks

"It can only be attributable to human error"
HAL9000, 2001: A Space Odyssey

In this chapter, we study routes and their capacity to support contextual analysis and information retrieval, creating ancillary data to improve navigation services. As such, we focus on how the potential in routes can be used to distill the necessary information to maintain these services and mitigate the problem of missing information in digital road networks. More specifically, the presented use case analyses an alternative to an expert-based model such as RouteYou's attractiveness model. Instead of using a set of attributes that describe activity-specific features or qualities of a road or path, we describe an adapted workflow and build a data-driven model to create a per-road attractiveness index based on the local flow distribution of recorded movements in a set of 190 610 routes. We use this set of routes to gain insights into the attractiveness of each edge in a real-world network for a specific leisure activity (i.e. road cycling). Next, we compare the devised model with popular navigation services for the activity at hand. Apart from describing the important decisions and set of procedure while processing a route set, we also want to gain insights in the constraints of the proposed data-driven approach.

The work presented in this chapter is based on following publication:
Baker, K., Ooms, K., Verstockt, S., Brackman, P., De Maeyer, P. and Van de Walle, R. (2017). Crowdsourcing a Cyclist Perspective on Suggested Recreational Paths in Real-world Networks. *Cartography and Geographic Information Science*, 44(5):422-435. doi: http://dx.doi.org/10.1080/15230406.2016.1192486

2.1 Introduction

Thanks to the mobile sensing (r)evolution (Srivastava et al., 2012), knowing where you are and where you want to go is becoming easier. As the number of online and mobile navigation services is growing, so has the use of these services for a gamut of leisure activities, such as road cycling or mountain biking. However, the affordance of a certain path for these particular modes lies primarily in subjective preferences and interests of the user group. Pigram and Jenkins (1999) voiced that these outdoor recreational activities lack the orderliness and monotony seen in utilitarian travel (i.e. commuting). Cognition and emotion play an important role (Golledge, 1999). These soft factors of the streetscape make the path suggestion much more complex and difficult. One of the fundamental problems is the insufficient comprehension of the selection criteria and trade-offs in path selection for the aforementioned types of leisure activities. While rigorous user studies are a preferred tool to solve this problem, this approach involves a high cost and needs thorough understanding of the subjects (Rubin and Chisnell, 2008). To face this challenge, Bakillah et al. (2014) stated that the use of sensor data and crowd-sourced information will certainly lead to new insights and improved navigation and routing services.

Recent years have seen a proliferation of motivated and IT-mediated crowds sharing their personal interests, experiences and feelings about their leisure activities through streams of user-generated content. Cyclists, for example, have the online and mobile tools to share and compare their trips and interact with peers worldwide. Concurrently, harnessing new affordable technologies in particular location-aware mobile devices and the mobile web has only amplified this trend and results in large crowdsourced datasets. Engaging these digital communities to elicit, collect and share information about their (leisure) activities is the approach in a wide variety of research (e.g., Huang et al., 2014; Reddy et al., 2010; Srivastava et al., 2012; Swan, 2013; Verstockt et al., 2013) and commercial applications such as RouteYou[1], Ride with GPS[2], MapMyRide[3], Strava[4], Bikemap[5] or Garmin Connect[6].

This use case analyses a route set, more specifically a large set of historical GPS traces, of recreational cyclists to give insights into a cyclist perspective on suggested paths in real-world networks. The goal of this research is

[1] www.routeyou.com

[2] www.ridewithgps.com

[3] www.mapmyride.com

[4] www.strava.com

[5] www.bikemap.net

[6] www.connect.garmin.com

twofold. First, we aim at developing a methodology to improve the adoption of user perspective in weighted graphs and routing engines designed for specific leisure activities, such as road cycling or mountain biking. Second, we want to create a thorough understanding of these leisure activities to facilitate future user studies. These user studies are an important step to achieve a final estimation of the quality of suggested paths. However, the latter is beyond the scope of this use case. Based on the proposed methodology, we gather insights on the first research question posited in the introduction of this dissertation. We focus on three, more specific, research questions:

- RQ1: How can **movements** described in large **route** sets maintained on route-sharing platforms be used to model specific **features or qualities** of a road?

- RQ2: Can this **model** be used to create a route suggestion for specific leisure activities?

- RQ3: How can path suggestions be **evaluated**?

Throughout this chapter, we consecutively address these research questions. First, we model a road segment popularity solely based on movements along edges in a graph condensed in route sets. This model is proposed as an approximation of complex factors and seemingly random agents in the activity at hand, such as scenery, popularity, affordance of the streetscape, physical challenge and group dynamics. This popularity can hence be perceived as an alternative to RouteYou's *attractiveness* of a road. Next, as state-of-the-art routing algorithms in performance demanding applications still consider path suggestion as a shortest-path problem, we also propose a cost function and a shortest-path alternative based on the developed popularity model. Finally, we elaborate on a model evaluation setup. We report on the general road characteristics in the proposed paths and compare our model and popular routing engines with shortest-path alternatives for cycling, i.e. RouteYou, Google Maps[7], Strava and Brouter[8]. This methodology was tested with 190 610 GPS tracking logs of road cyclists in the region of East-Flanders (Belgium) shared on the web platform of RouteYou and combined with an OpenStreetMap road network extract of this region of interest.

[7] www.maps.google.com

[8] www.brouter.de/brouter-web/

2.2 Related work

For many centuries, the resource-intensive task of creating and annotating geographic representations of the world was limited to commercial or governmental mapping agencies (Kimerling et al., 2009; Mooney et al., 2013). However, since the terms citizen science (Haklay, 2013) and crowdsourcing (Howe, 2006) were first coined, several projects have demonstrated the feasibility of crowdsourcing a wide range of IT-enabled tasks to a large community of volunteers. Within a geographical context, this new approach is often exemplified by the community-driven mapping project OpenStreetMap (OSM) (Haklay and Weber, 2008). Notwithstanding the intricate nature of the digital road network in OSM, it still lacks a spatial homogeneity in quality within larger geographic regions (Haklay, 2010; Mooney and Corcoran, 2012). While the widespread use of OSM within slow-traffic projects proves its usability (Kessler, 2011), the lack of homogeneity in specific tagging behavior often hampers calculation of an attractive path for slow traffic. Consider the road class *unclassified* in OSM[9]. This road class is defined as publicly accessible, local roads and, hence, valuable for slow traffic. This class encompasses approximately 10% of OSM road objects[10]. Yet depending on country, region or urban/rural setting, the intrinsic characteristics of this road class changes, for example surface type. Lack of additional object attributes limits their value in rigid expert-based models supporting activity-specific navigation services.

The plethora of user-generated content, sensor data and crowd–sourcing projects receives a burgeoning interest to enrich digital road networks with additional information. For example, Huang et al. (2014) collected people's in-situ affective response to enhance path suggestion in routing engines. Similarly, Quercia et al. (2014) combined crowdsourced perceptions of streetscape imagery in cities to discover happy, quiet or beautiful urban places. In addition, photo-sharing platforms such as flickr[11] or panoramio[12] have also been used to discover popular or attractive places (e.g., Popescu et al., 2009; Rattenbury et al., 2007; Zheng et al., 2009,1). In a similar way, Pippig et al. (2013) combined geo-referenced wikipedia concepts and shared geotagged photos to discover theme-based routes in cities. Furthermore, much research has successfully proven the applicability of raw GPS traces as a proxy for the movement of objects (Bakillah et al., 2014). In Route Choice Modeling (RCM), studies increasingly incorporate GPS traces and revealed

[9] http://wiki.openstreetmap.org/wiki/Tag:highway%3Dunclassified

[10] https://taginfo.openstreetmap.org/keys/highway#values

[11] www.flickr.com

[12] www.panoramio.com

preferences as a cost-effective alternative to active data collection through stated preference, such as questionnaires and travel diaries (e.g. Alivand et al., 2015; Broach et al., 2012; Dhakar and Srinivasan, 2014; Hood et al., 2011; Menghini et al., 2010; Prato, 2009; Snizek et al., 2013). Similar to our approach, the basic idea behind this is that a recorded movement encompasses all the route selection criteria and trade-offs which are important to the end-users. However, Bierlaire and Frejinger (2008) also pointed out the bias and errors which can be introduced while transforming network-free data to network-constrained events. However, Newson and Krumm (2009) argue that high sampling rate (i.e. below 30 seconds) while recording positional data during movement is an important prerequisite to reduce errors resulting from this transformation.

Managing routing engines that adopt a user perspective and compute routes of interest entails a weighted digital road network and a routing algorithm. From a theoretical perspective, the weighting involves the approximation of the mode-specific travel expense of every road segment in a network. In practice, a cost function is often modeled as a linear combination or a decision tree based on specific attributes in a road database. In routing applications, this is often called the routing profile. In a cycling profile, for example, suitability could be approximated by the appropriate weighting of physical and scenic road characteristics and traffic regulations. However, modeling and maintaining these profiles requires considerable time and effort. Furthermore, Mooney and Winstanley (2006) stated this single-cost approach is an oversimplification of a complex problem and generates undesired paths in routing applications. However, state-of-the-art routing algorithms in performance-demanding applications still reduce path suggestion to a standard shortest-path problem, implementing the algorithm of Dijkstra (Luxen and Vetter, 2011).

Despite the many commercial applications which offer path suggestions for leisure activities, academic research and validation remain, to our knowledge, rather sparse. Within RCM research few efforts have been made to discover characteristics or factors determining leisure activities. Alivand et al. (2015) used scenic routes extracted from route-sharing platforms to discover and extract scenic attributes in route choices. However, the capturing of these factors in geographical databases remains a labor-intensive task and limits these approaches. Alternatively, Oksanen et al. (2015) discussed heat maps of GPS traces as a more holistic way of analyzing movements during leisure activities. Our approach implements a similar holistic view, using the intensity of a phenomenon (i.e. identical movements along road network segments) and its explanatory value in research. In essence, a heat map is a graphical representation of matrix where each cell (or image pixel) encompasses an area on the earth surface. A heat map of routes, for example, represents the number of times a route intersects this cell. In contrast, we do not focus on visualizations, but aim at discovering a high-level popularity index of individual segments in a road network based on recorded movements along these segment. As such, we count each time a segment is used by a route, that is a spatial match between route and segment. The goal is the same namely measuring the intensity of a phenomenon but the basic statistical units (pixels versus segments) are different. This approach allows us to make an abstraction of the previously mentioned limitation of contemporary road databases to create an attractiveness model (e.g. missing or erroneous information) and create a sound scientific basis for future work.

2.3 Methods

2.3.1 Study area

The study area is situated in Belgium and encompasses the province of East-Flanders (see Figure 2.1 on page 27). This area covers 2991 km^2 and has one of the most dense road networks per ha in Europe (Government of Flanders, 2012). Moreover, Belgium has a cycling-savvy society which uses bicycles both as a major mode of utilitarian transport and for leisure purposes. Similar to other cycling-savvy societies such as the Netherlands and Denmark, attention is given to cycling-specific road infrastructure and creating awareness for slow traffic. Exemplary is the more than 12 000 km signposted cycling network transecting the northern parts of Belgium.

2.3.2 Data and preprocessing

The GPS traces used in this study were uploaded to the RouteYou platform and tagged as road cycling by the users sharing the traces. This sensor data originated from commercial GPS-enabled devices. Meeting the functional requirements of the device producers (e.g. Garmin, Mio, TomTom), these outdoor navigation devices track the movement of the user with a high sampling rate and a GPS accuracy of less than 15 m, on average. After basic data cleaning of the raw dataset, the final dataset consisted of 190,610 anonymized records collected between June 2013 and June 2015. For example, we removed GPS traces with an obvious absence of motion while recording (i.e. so-called tumbleweeds). Next, clear outliers in speed can also be used to remove unwanted activities from the dataset. Recorded average speed above 50 km/h is unrealistic for the activities we are modeling. Furthermore,

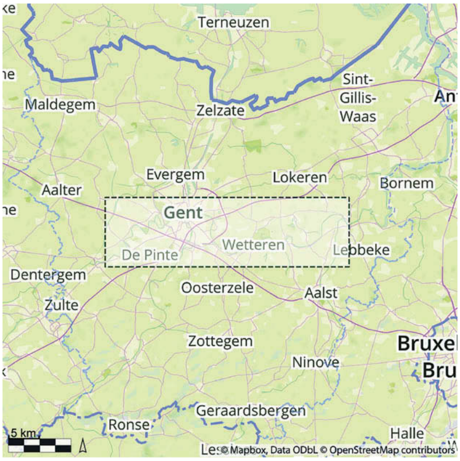


Figure 2.1: Overview of the study area with the bounding box of figure 2.3 as an inset.

erroneous tracking data can also easily be filtered out (e.g. timestamps or false points). The final dataset was created by more than 6300 unique devices. The average sampling rate of the raw sensor data was 1.98 sec with a standard deviation of 1.09 sec. While the raw data lacked metadata about horizontal accuracy and precision, comparison of the tracks with ground truth confirmed the technical specifications. Figure 2.2 on page 28 gives a detail of the traces. Inherent to many crowdsourced information, we lacked quality metrics of the tagging. However, mode and activity detection did not lie within the scope of this use case. Therefore, the tag road cycling was deemed correct and unambiguous to describe a group of cyclists with the same interest. Further exploratory analysis of the dataset showed that these self-selected road cyclists had an average speed of 26 km/h with a standard deviation of 7.03 km/h. The majority of the trips had a length between 60 and 100 km. This underpins our assumption that a significant part of the GPS traces encompass recorded movements of road cyclists. Figure 2.3 on page 28 shows a snapshot of the selected traces around Ghent and indicates that these traces had a good distribution across the study area. In the remainder of this section, we discuss the performed procedures to convert the GPS traces to detected movements along road segments in a real-world network.

Cognitive mapping research (e.g., Golledge, 1999; Tolman, 1948) has shown that proximity to well-known places improves the detail in these parts of a mental map and, as a consequence, results in a monotony in route choice decisions near these places. Additionally, the number of route choices close to a specific point in a real-world network are often limited. As a result,

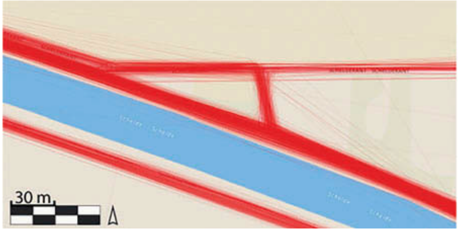


Figure 2.2: Detail of the GPS traces along the Scheldt, east of Ghent.

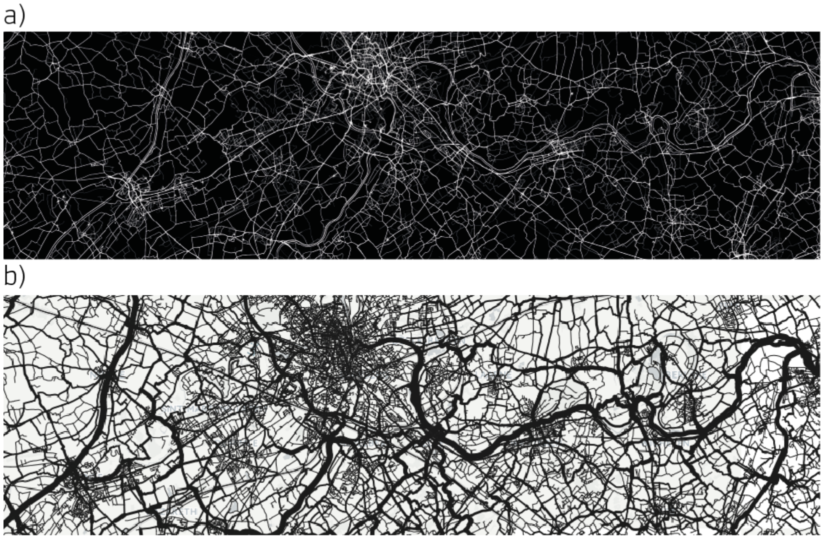


Figure 2.3: The top image (a) shows an extract of the GPS traces around Ghent. Pixels vary in shades of gray, varying from black (no traces) to white (high trace count); the bottom image (b) shows the road network in the extract with the thickness of the lines proportional to the cumulative count variable z after the map matching of the GPS traces.

near clusters of start and end points in our GPS traces, the same choice decisions were often preferred. However, this monotony in behavior lead to a network bias which was of no real significance to the leisure activity we were modeling. To avoid this, we chose to omit GPS points near both start and end of the traces. A generic rule-of-thumb was created, based on the trip distance distribution in our dataset. This heuristic was used to strike a balance between removing unwanted patterns and preserving valuable information. First, we selected all GPS points in a trace within 10% of the trip distance from start and end. Subsequently, we omitted all selected points within 5 km of start and end. All distances in this heuristic were measured along the GPS traces. This approach is approximate, however, due to the lack of empirical data on this complex and context-specific behavior. Figure 2.4 on page 29 shows the impact this preprocessing step had on the cumulative counts of movements along the road network of a single user.

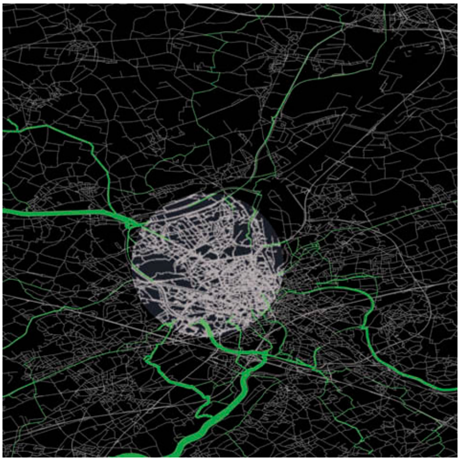


Figure 2.4: Impact of the preprocessing steps on the match results of the GPS traces of one single user. Similar to figure 2.3, the thickness of the line is proportional to z. The circle approximates the area where bias was seen by similar route choices near start and end.

For this research, an OSM road network extract was used. This extract consisted of the OSM data primitives nodes, ways and relations. The primary feature key $highway$ allowed a straightforward filtering of all roads, streets and paths in our study area (Zielstra et al., 2013). Although particular tags allow filtering between, for example, car-accessible and bicycle-only way segments (Hochmair et al., 2015), we chose to incorporate all segments in our analysis. This ensured that exclusion of certain road types did not impact the final shortest-path alternative based on the conceived attractiveness model in this use case.

To detect route trajectories along the streets, a map-matching approach was necessary. More specifically, an incremental curve-to-curve map matching algorithm was implemented to reconstruct the flow distribution of the cyclists in the road network (Newson and Krumm, 2009; Quddus et al., 2007). The road network of the study area was attributed with the cumulative number of movements along the matching edges. To reduce the errors of commission and omission within the matches, two pragmatic choices were made. First, we converted the OSM road network extract into a graph. To have a more intuitive representation of the road network, an OSM node used by multiple OSM ways became a node in the graph. Next, nodes with a valency of two were then removed from the graph and adjacent edges were joined. Secondly, to detect a movement along an edge in a trace, a minimum match length between both edge and trace was introduced. This match criterion was introduced as a heuristic for a significant experience along an edge, thus improving the robustness of the model to mismatches or missing roads in the graph. Within our approach, we used a threshold of 50%. This relative length threshold proved highly efficient in removing errors in omission and commission from our matching results. Figure 2.3 on page 28 shows a snapshot of the OSM extract around Ghent with a proportional line thickness to the cumulative count.

Following the classification of Okabe and Sugihara (2012), all movements in the GPS traces were converted to undirected network-constrained events. This approach has methodological constraints and limitations in representing the real paths in the graph (Bierlaire and Frejinger, 2008): missing connectors or over-representations of certain paths in the network can lead to underestimation or lack of counts on edges in the road network. Nevertheless, due to the high sampling rate and the assumption of spatial randomness of errors in the raw dataset we assume our map-matching result a good approximation of reality (Quddus et al., 2007). This resulted in an undirected graph $G = (N, E)$, a set N of 71 801 nodes and a set E of 101 865 edges. Each edge e is associated with a variable z, the cumulative count, and a weight w, the real world length of the edge.

2.3.3 Attractiveness modeling

In this section we document our attractiveness modeling, that is a normalization procedure of the z variable in the graph. We introduce the key concepts in this procedure to calculate a per-edge attractiveness index based on the number of times an edge was used in the route set.

Definition. *The **attractiveness index** is a proxy for the local popularity of an edge and, hence, its activity-specific fitness for use. It is a standardized score between* $[0, 1]$ *based on the number of times an edge was used in a route set.*

Different normalization procedures are applicable to attain a standardized score. Quercia et al. (2014) used maximum score standardization to convert a road network criterion to a standardized score. This is a normalization which scales all values in the full range of values from network criterion. Additional transformations can be used to remove very skewed distributions from the dataset, such as log transformation. However, these network-wide normalizations underestimate the local character of the phenomenon we are modeling. Pigram and Jenkins (1999) advocated that the response to time-distance, connection, and network bias plays an important factor in forming patterns in recreational travel. Interaction is more likely between connected edges than between unconnected or distant edges. Parts of the network can be ill connected due to human or natural barriers, among other things, such as rivers or highways. Hence, flows of cyclists on nearby edges can be compared more easily than between distant edges. We introduce a neighborhood of an edge to incorporate this spatial connectedness or association in our normalization procedure.

Definition. *The **edge neighborhood** is a distance band delineating the space within which edges have a context-specific spatial association.*

To further integrate the effect of distance in our normalization procedure, we also incorporate a distance decay function. Distance decay functions found widespread applications to model spatial interactions between places as the distance between them increases (Fotheringham, 1981). As such, within the edge neighborhood we further state that certain values of our network criterion are more related than others and this relation can be approximated by the distance between them.

Equation 2.1 presents our attractiveness index (AI) of edge e proportional to the z variable at edge e. We normalize each z value in a range of distance-weighted z variables based on all edges g in the neighborhood of edge e.

$$AI = \frac{z_e}{\max z_g f(D_{eg})} \tag{2.1}$$

With D_{eg} the distance between edge e and g and $f(D_{eg})$ a distance decay function weighting the count on edge g. As mentioned before, the distance decay function reflects the impact distance has on comparability of flow counts. As a result, it delineates a region of influence. As the flow of cyclists was constrained to the graph and influenced by the connectedness in the graph, we used the shortest-path distance as metric in the model. For simplicity and tractability of the model, distances between edges are calculated between the midpoint.

The above attractiveness model is conditioned by (i) the size of the neighborhood of an edge and (ii) the modeling of the spatial interaction within this neighborhood. However, we lack empirical data or stated preferences about what the neighborhood of an edge could be for road cyclists. As a proxy within our methodology, the trip distances of a sample set of routes from the RouteYou platform (249,942 routes) with the same tag (i.e. road cycling) was used. Figure 2.5 on page 32 shows the frequency bar plot of the trip distances in this sample set. Visual analysis of this plot, shows that Q3 can be used as a rule-of-thumb to separate head from tail in this distribution. Next, if we assume that a cyclist is willing to ride half his trip distance to reach the furthest point from his starting point, the distance band D^N around a start point can thus be approximated by half the trip distance. If we take the 75 percentile as rule-of-thumb to delineate the neighborhood, our distance band is 35.52 km.

Next, the choice of decay function within the distance band played an important role in the distance-weighting of the counts in equation 2.1. An inverse distance or negative exponential function has often been the function of choice in gravitational models. However, as Steenberghen et al. (2010) stated this type of function focuses attention more locally and makes distance a strong deterrent factor for interaction. On the contrary, equally weighting all edges within the distance band will maybe underestimate the impact of distance in the proposed model. A linear decay $f = 1 - D_{eg}/D^N$ was chosen for its comprehensibility and simplicity; weight of a count decreases equally with distance. Outside the predefined neighborhood f equals 0.

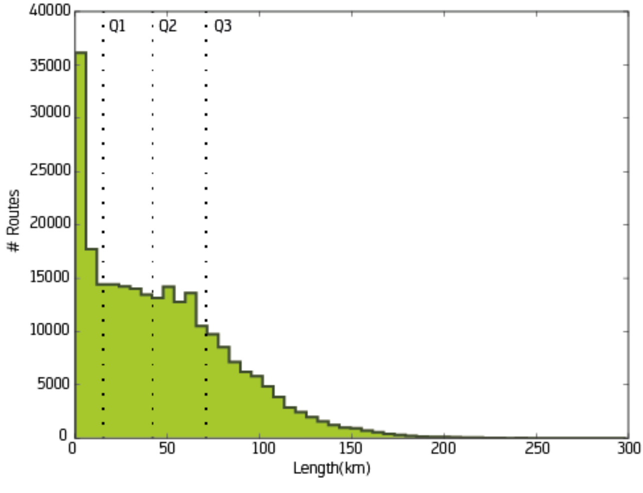


Figure 2.5: Frequency plot of the trip distance in a sample set (249 942 routes) of road cycling routes from RouteYou.

2.3.4 Path generation

To generate alternative paths in shortest-path algorithms a multiplication factor f of the geometric length of every edge in a graph is used. In other words, a detour from the shortest path will be deemed a valid alternative if the reduction in weight makes up for the gain in path length (Hochmair and

Navratil, 2008) (see Figure 2.6 on page 33). In our approach f is a function of the inverse of AI. A linear transformation of AI^{-1} to f between $[1, \beta]$ with $\beta > 1$ was used. Weights of zero are undesired in our approach because this would implicate that there is no travel expense in incorporating an edge in an alternative. As distance still plays an important role in wayfinding, we defined the weight of an edge scoring maximally in AI as his original geometric length. Furthermore, by incorporating length in the cost function, increasing β will have an impact: a higher β results in longer paths. To analyze the impact of β on the generated alternatives with this linear transformation, a Monte Carlo simulation was done. We iteratively (at least 50 times) chose 100 randomly selected start/target couples and generated the shortest path with β varying between [2,50]. To quantify the quality of the alternative we use the average AI and the proportional gain in length. Figure 2.7 on page 34 shows the scatter plot of the average improvement of AI and average proportional gain in length with increasing β over the 50 iterations. The error bars in the plot indicate the standard deviation through the 50 iterations. This shows that with increasing β the gain in average AI of a path is not proportional to the gain in geometric length. Furthermore, we see that variations of β in the interval [0,5] has the strongest impact. Next, for brevity, we will only report on results gathered with a β of 5.

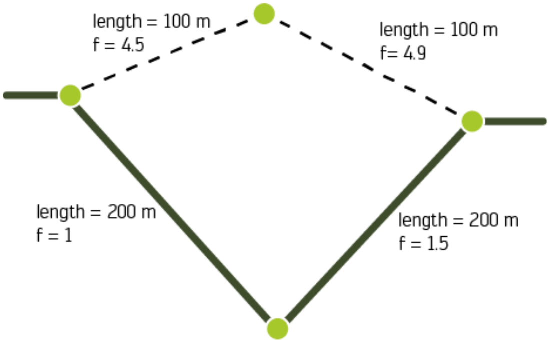


Figure 2.6: Hypothetical network explaining the multiplicator factor. In this situation, the green path is a valid alternative to the shortest path. An object is prone to choose the green path because the perceived travel expense ($(200 * 1) + (200 * 1.5) = 500$) is lower than the shorter path ($(100 * 4.5) + (100 * 4.9) = 940$).

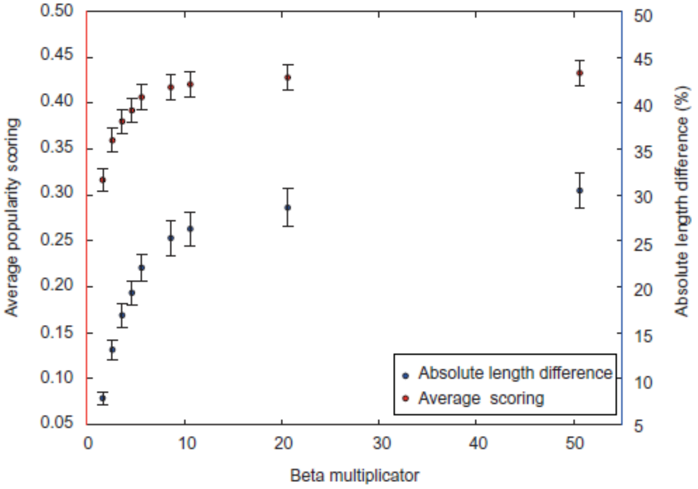


Figure 2.7: Plot showing average improvement of AI (red) and average gain in length (blue) with increasing β over the 50 iterations.

2.4 Results

2.4.1 Exploratory spatial data analysis

Before addressing the attractiveness model evaluation, an exploratory spatial data analysis of the z variable is presented. We analyzed (i) the distribution of z and (ii) the cross-correlation of the z in the network space. The rank-size plot of z in Figure 2.8 on page 35 clearly shows a heavy-tailed distribution of the cumulative count in the network. This highly imbalanced distribution indicates few edges have a large count, many edges have a low count. Next, an assessment of the similarity of cumulative counts on contiguous edges in our graph can be done with network autocorrelation (NAC) measures (Okabe and Sugihara, 2012). To calculate the NAC, the Moran's I interpretation of Black (1992) was used. A binary weight matrix was used to define contiguous edges in this analysis: if w_{ij} = 1 then i and j are contiguous edges, and 0 otherwise. More specifically, if the start or end node of j lies on a Dijkstra shortest path p_{ij} with a distance d_{ij} smaller than the threshold network distance D^N, j and i are contiguous. The origin of all shortest paths from edge i was the midpoint of this edge. Figure 2.9 on page 35 shows a network visualization of the binary weight matrix around a randomly chosen edge i in our graph. We performed the above procedure with $D^N = 1km$, $5km$ and $10km$. All resulting autocorrelation measures ($I_1 = 0.1833$, $I_5 = 0.1229$ and $I_{10} = 0.0799$) were low and indicate no significant network autocorrelation.

Hence, similar counts do not tend to occur on all contiguous links in our network. This data analysis underpins that location-specific movement in the graph is not random and there is a strong network bias defining the flow of road cyclists in our study area.

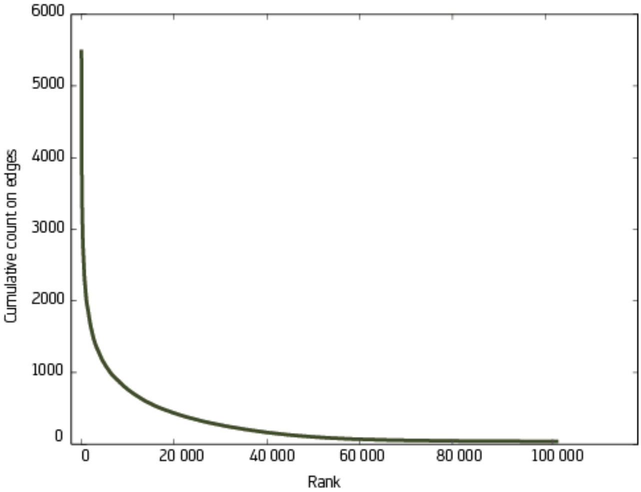


Figure 2.8: Rank-size plot of the cumulative count in the network.

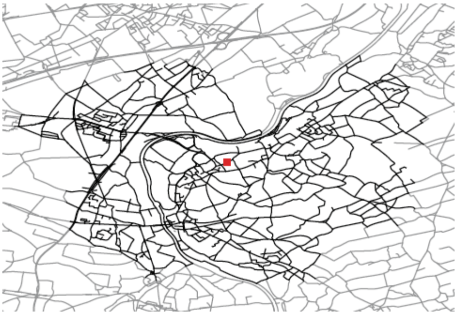


Figure 2.9: Network visualization of binary weight matrix around the midpoint of a randomly chosen edge with D^N = 5 km; contiguous edges are black, non-contiguous edges are gray.

2.4.2 Model evaluation

We designed an experiment to evaluate (i) the general characteristics of the proposed paths for the specific activity (i.e. road cycling), (ii) how our scoring is distributed in shortest-path alternatives generated by popular routing engines for cyclists and (iii) how the proposed path generation approximates the shortest path and its alternatives. We chose four transects in our study area for further analysis: northeast-southwest, southeast-northwest, north-south and west-east, respectively, with a corresponding shortest-path length of 84.8 km, 60.2 km, 82.0 km and 65.2 km. Figure 2.10 on page 37 visualizes the four alternative paths together with the shortest path and the path generated with the attractiveness model. Following routing engines where chosen as comparison:

- RouteYou: RouteYou offers several customized shortest-path alternatives for slow traffic modi (*Race cycling - nicest* profile);
- Strava: Strava offers a shortest-path alternative modeled on a popularity approximation based on the routes shared on their web environment (*Ride* profile with popularity use);
- Google Maps: Google Maps offers shortest-path alternatives for several general slow traffic modi like cycling (*Bicycle* profile);
- Brouter: Brouter offers customizable bike routing (*fastbike* profile).

2.4.2.1 Path characteristics

Prior to the comparison of paths, we report on the path characteristics in the four transects generated with the attractiveness model. Figure 2.11 on page 38 presents a proportional length-based comparison of the values in the primary feature key $highway$. These key-value pairs can be used as a general proxy of road types and characteristics used in the paths. As a proxy of road condition, we also report on the part of the OSM way segments having a key-value pair indicating a bad road condition (i.e. unpaved). We clearly see that the majority of the paths consist of $Tertiary$, $Unclassified$, $Residential$, $Cycleway$ and $Track$. This highlights that all paths, except for the W-E transect, have approximately 90% of their length tagged with these key-value pairs. It is also noteworthy that very few parts have a value $Path$, $Steps$, $Service$ or $Footway$. The N-S transect clearly stands out with a low length (<1 %) in $Cycleway$. In all paths, parts with a key-value pair indicating a bad road condition are sparse. In-depth analysis of the paths and terrain verification indicated that the majority of the detected unpaved segments were due to very local non-optimal choices or errors in the OSM attribution.

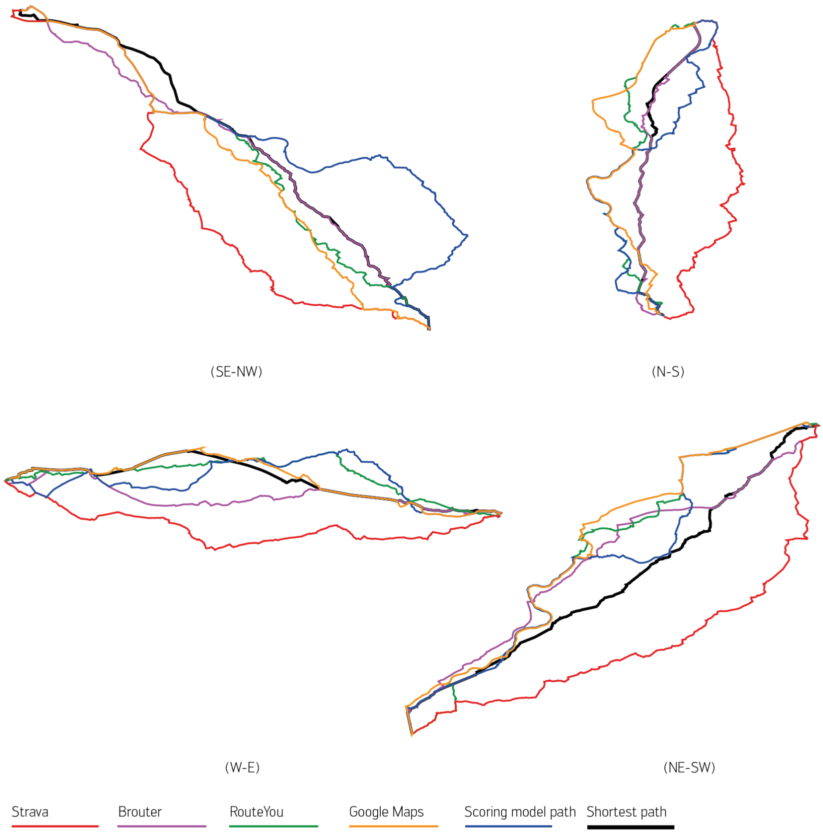


Figure 2.10: Visualizes the different paths generated with the different routing engines along the four transects.

2.4.2.2 Comparison of paths

For fair comparison of the spread in the attractiveness scoring along the presented alternatives, we reduced the spread of the scoring within the paths to an averaged value AI/km. To morphologically compare the alternatives, we refer to three robust measures of similarity between curves:

- Absolute length difference: the difference in geometric length between two curves. This is often defined as the efficiency of one path in respect to another path. A shorter path is more efficient than a longer path;
- Coincidence: the distance two curves coincide;
- Fréchet distance: similarity measure which takes the continuity of two curves into account and is a good measure for detour size between two curves.

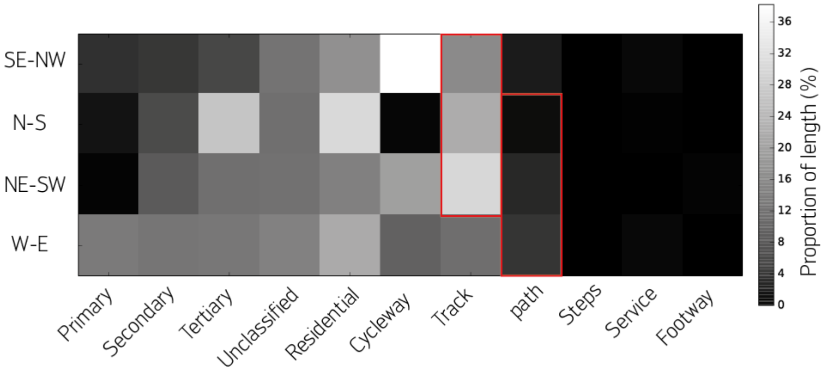


Figure 2.11: Proportional length-based summary of the values in the primary feature key *highway* in the four paths generated with the attractiveness model. Matrix cells vary in shades of gray, varying from black (no proportional length) to white (high proportional length). Cells outlined in red indicate that a part of the OSM way segment has a bad road condition. This analysis was based on specific key-value pairs in these segments. Unpaved segments had a cumulative length of $unpaved_{SE-NW} = 322\,m$, $unpaved_{N-S} = 1336\,m$, $unpaved_{NE_SW} = 1430\,m$, $unpaved_{W-E} = 811\,m$.

Figure 2.12 on page 39 presents the averaged scoring in function of the absolute length difference (km) between both the alternatives and the shortest path. Theoretically, we would want the slope of the trend line to approach zero. Thus, a gain in AI/km with a minor gain in length. This is often defined as the optimality of a route. In reality, however, there are few circumstances where this theoretical optimality of a path is possible. The gradient in the graph is used to depict the theoretical zones of maximal (white) and minimal (black) optimality. The trend line drawn in the plot, a first degree polynomial fit of the path sample set, gives a theoretical foundation to divide our set of paths in optimal (below trend line) and less optimal paths (above trend line). A reduction in efficiency always leads to a gain in average scoring, but sometimes the alternative become less optimal.

To evaluate the morphological differences creating the trends in Figure 2.12, Figure 2.13 plots the coincidence (km) in a function of the Fréchet distance (km) and shows the paths of four extrema in the plot. Both parameters in this plot are a comparison between the shortest path and the alternatives along the four transects. As a result, this presentation successfully shows the morphological differences which lead to the trade-off in efficiency of the alternatives.

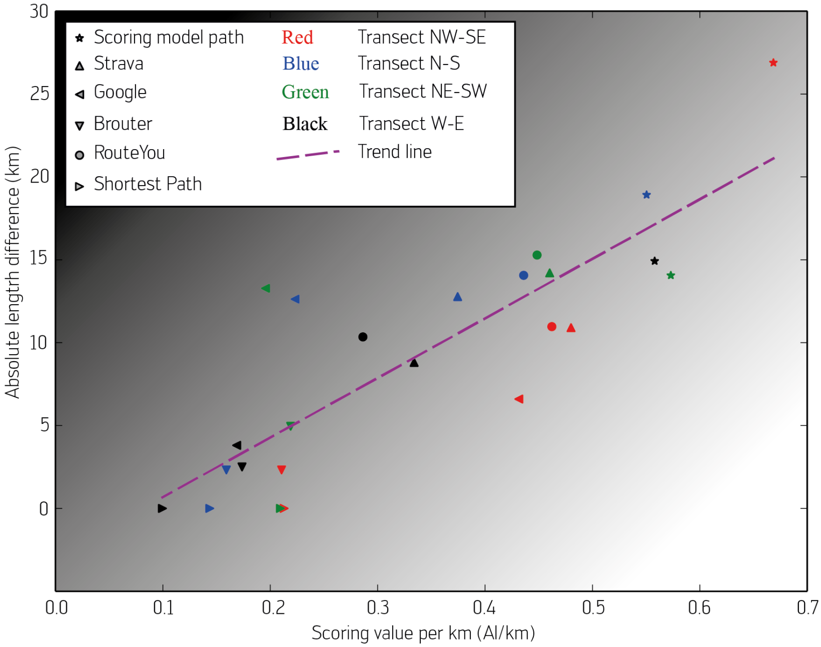


Figure 2.12: Scatter plot with the scoring value per km of shortest-path alternatives for the four transects as x-axis and the absolute length difference between the alternatives and the shortest path as y-axis; the gradient in the graph visualizes the theoretical zones of maximal (white) and minimal (black) optimality

Working as designed, the attractiveness model path exhibits the highest values on the index per kilometer. However, we also see that these routes maximize the absolute length difference with the shortest path. Similar trends are seen with the alternatives generated on RouteYou and Strava. They achieve similar AI/km, albeit with a different path. We see that Strava and RouteYou have a similar absolute length difference, but Strava paths show in all transects a lower coincidence with the shortest path and a higher Fréchet distance. It is fair to say both are designed to generate recreational paths for road cyclist, however both are still able to improve the local optimality of choices. Brouter tends to approximate the shortest-path distance with high coincidence. Thus, Brouter has a similar tendency in direction of the shortest path, but locally makes different path decisions. It is clear that the used profile of Brouter tries to generate a shortest suitable path for road cyclists, but not a recreational path. The paths generated by Google Maps are more difficult to explain. Two of the four transect paths can be deemed less optimal solutions (NE-SW and N-S). The W-E transect tends to give a suitable shortest path, while SE-NW seemingly succeeds in generating a recreational path.

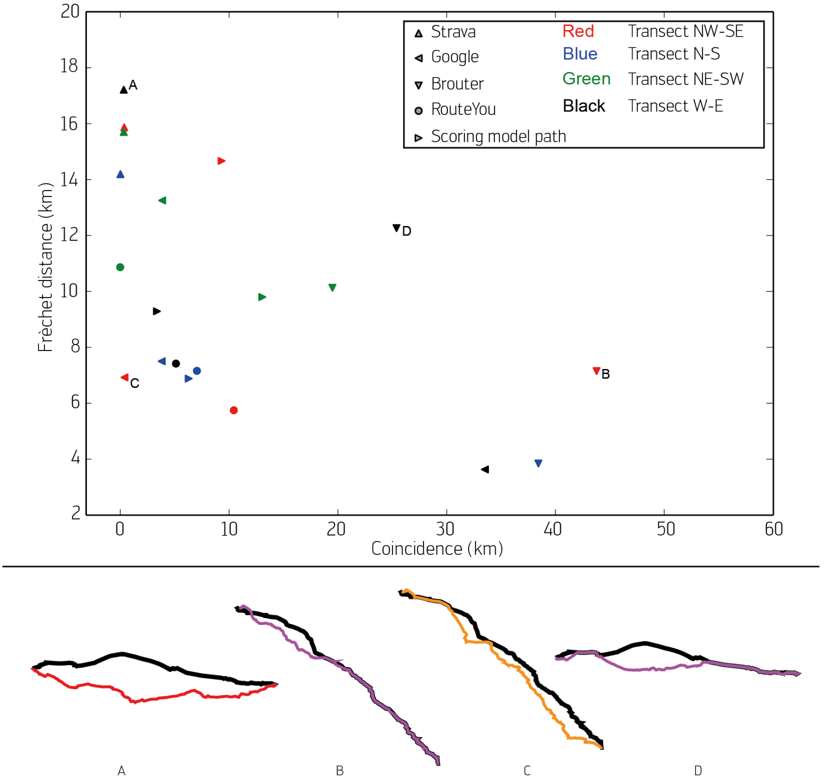


Figure 2.13: Scatter plot with as x-axis the coincidence (km) and as y-axis the Fréchet distance (km); the black line in the examples is the shortest path

2.5 Discussion

2.5.1 Data selection

Boen et al. (2011) described the population of recreational cyclists active in Flanders as a continuum from competitive cyclists to purely recreational cyclist. The first are highly achievement-oriented cyclists and the latter give more attention to the experience, surroundings and group interactions. While specific activity and mode detection were put beyond the scope of this use case, the high standard deviation around the mean speed in the raw dataset indicates our road cycle routes also mirror this continuum. Not only does this imply the presence of sub-types of road cyclists in our dataset, it also hampers the suitability and adoption of our approach as an approximation of road cyclist behavior in general. It must be noted, however, that this does not impede the validity of the presented model, but future work should focus on three parts. First, it should address if our model holds for other (sub)types of leisure or utilitarian cyclists. Second, it is imperative to understand how the parameters in the model are influenced by this change of behavior. Finally, future work will also have to address how this cyclist typology can be detected in the raw data.

To face these challenges, cyclist typology and, consequently, detecting determinants for this classifications will be imperative. Recent work of Bergström and Magnusson (2003) and Damant-Sirois et al. (2014) on cycling typology indicated that, next to road condition, precipitation and temperature are considered key factors to get people to cycle. While the avoidance of bad road condition is already visible in the results of the path characteristics, integrating historical weather data and a temporal dimension in our data selection and preprocessing would certainly improve our approach. For example, differentiating between days with or without precipitation could already lead to a different type of cyclist in the model. Similarly, season, day of the week and time of day could be used to further improve a typology of road cyclists.

As mentioned in the introduction of this chapter, using the GPS traces as a proxy for attractiveness of a road, we wanted to address the spatial heterogeneity of tagging behavior. As with other crowdsourced information, however, the collection of GPS traces also grapples with this spatial heterogeneity. Thus, some regions will always be under-represented in comparison with other regions. Additionally, the lack of census data in larger geographical regions about recreational slow traffic complicates the modeling and normalization of the flows in a road network. The introduction of the neighborhood of an edge in the model is an attempt to tackle this problem. What is crucial, though, is the critical mass of the collected data set. Drawing parallels between our work and research in sociology and marketing, critical mass refers to the theoretical tipping point where extra GPS traces will not change the observed results. Two characteristics give strong indications of critical mass in our dataset. First, Jiang et al. (2009), among other research, argue that characterizing heavy-tailed (i.e. power-law like) distributions are inherent to human mobility patterns. Human movements tend to converge in salient corridors. We observed a similar distribution in our map-matching results which highlight the representativeness of our dataset for real-world human behavior. However, to our knowledge, few work has been done to formally test this hypothesis for leisure activities, such as cycling. Secondly, network autocorrelation is a powerful tool to understand network-constrained movements(Okabe and Sugihara, 2012). We found that the recorded movements are not random, which again highlights the representativeness of the GPS traces. However, this gives few insights in how to reach the required level of performance (i.e. the number of GPS traces) to achieve reproducibility and scale. Future work should address this question of critical mass. In general literature, iterative statistical analysis of changes in the model with increasing number of data is often proposed as method to identify this equilibrium. The aforementioned characteristics are a valid starting point for this iterative approach: (i) "How does the distribution of counts vary with increasing number of traces" and (ii) "How does network autocorrelation vary with increasing number of traces?". Another future research track should focus on the ability to replicate the presented approach over time. The calculated index we devised was based on static data and, as a result, the inferred information quickly becomes *stale* information. Equally important to the data set size, is a constantly growing amount of data to *learn* from. This need and how it can be integrated in the presented approach will be imperative in future research.

Finally, we want to discuss the viability of shared traces as a proxy of local expertise and knowledge. Because of the burgeoning use of navigation services in leisure activities, it is doubtful that all followed paths were solely based on the user's mental map. While literature on this specific problem remains, to our knowledge, rather sparse, the simplifying assumption that humans always follow the least effort path supports our approach (Golledge, 1999). In other words, users of the above services will deviate from a suggested path if the suitability of an alternative is perceived higher (i.e. the resistance is perceived lower). Following this as a rationale, we can assume that GPS traces approximate route choices of a cyclist and are a proxy of the attractiveness of a path.

2.5.2 Model parameters

To further put the above results into perspective, it is also necessary to discuss some of the methodological choices made in our approach. First, the delineation of the neighborhood of an edge and the modeling of the spatial interaction within these neighborhoods plays an important role in detecting local extrema in z. We asked the question: "which edges and their respective z variable are comparable?". Because of the absence of solid theoretical foundations regarding these parameters in leisure activities, the choices were based on pragmatic choices. However, some considerations for future work can be proposed. Based on the heavy-tailed distribution of the z variable in the graph, smaller neighborhoods will increase the number of local maxima in our popularity index, whereas a larger area will decrease the possibility to detect local maxima in the overall tendencies in z. More local maxima will certainly result in more dissimilarity within all shortest-path alternatives in our graph. On the contrary, fewer local maxima will achieve the inverse creating more salient corridors in the graph. Further research and data collection is necessary to come to a more formal description of the neighborhood and its impact on the morphology of paths.

Second, we discuss the use of a multiplication to introduce our scoring value in a cost function. We have to be aware that the use of the β value is an ad-hoc solution and is not only influenced by the spread of scoring value in the road network. Sparsity of a road network plays an important role in the validity of the choice of β. Our hypothesized β value holds for dense networks, but can fail for sparser networks. Further analysis is necessary to understand the impact of network density on the morphology of the suggested paths. However, as Figure 2.7 shows, the used methodology makes our scoring function highly adjustable and configurable. For example, if for a certain mode it is more desirable to have a smaller absolute length difference with the shortest path, a smaller β could be a possible solution. Furthermore, we have to be aware of the skepticism towards single shortest-path algorithms to maximize a criterion value (Hochmair and Navratil, 2008; Mooney and Winstanley, 2006). Eppstein (1994) proposes a k shortest path approach to evaluate the spread of a criterion in multiple shortest paths. In future research attention should be given to this approach and how this relates to our approach.

Last, and maybe foremost, is the lack of stated preference of cyclists in our approach. However, the created framework within this study will certainly help to design more cost-effective user studies. Figures 2.12 and 2.13 show that the popular route generation approximates contemporary routing engines in morphology and scoring value per km. Additionally, Figure 2.11 successfully shows that the suggested paths approximate the expected road types of a leisure cycling route as shown in RCM studies for cycling. However, we have no indication of the actual local optimality of the differences in the paths. This clearly indicates that future work will have to focus on validating this optimality with local expertise and knowledge. It should also be noted that the lack of academic validation of many commercial routing engines makes assumptions about and comparison between the suggested paths precarious. Further research is necessary in collecting stated preferences of cyclists and comparing them with our attractiveness model in order to better understand the real-world validity of the methodology and the suggested alternative based on our scoring.

2.6 Conclusion and main contributions

This use case set out to develop a methodology to improve the adoption of a user's perspective in contemporary routing engines designed for specific leisure activities (i.e. road cycling). We devised a methodology to elicit a high-level attractiveness scoring of every road in a network based on homologous movements in a large route set (RQ1). In doing so, we created an alternative to popular expert-based models to create activity-specific navigation services, reducing the impact of missing or out-dated information in a creative and cost-effective way (RQ2). A set of 190 610 GPS traces collected on the route-sharing platform RouteYou was used as a proxy of movements of cyclists along the network. Next, this attractiveness model was embedded in an experiment design to evaluate the general path characteristics and compare contemporary shortest-path alternatives for cyclists (RQ3).

RQ1 - This questions focused on the methodological approach of the route set to address the goal of this use case. Given enough routes, the presented methodology successfully exploits the geographical context within this route set. We devised an alternative way to maintain a navigation service with an activity-specific model. Especially noteworthy in this conclusion are two methodological choices we introduced during the preprocessing steps. First, apart from general data cleaning and smoothing, which are well-studied in current academic literature, we focused on removing unwanted bias from a crowd-based route set. Underlying geographical drivers, such as urban and rural regions, generate clusters of contributors. While some bias will always be obscured by a lack of user- and activity-specific information linked to individual routes, the impact of clusters of start and end points is clearly observable. We proposed a 10%-rule as a heuristic to remove this bias. Second, we used a state-of-the-art map-matching approach to convert the route set to network-constrained events. This paper did not focus on documenting a step-by-step matching approach. Of particular interest, however, is a second methodological choice which impacts the contextual analysis. We introduced a half-way threshold as a proxy for a significant, network-constrained experience along an edge. This allowed us to remove noise and bias from the resulting counts of homologous movements along all edges. In conclusion, exploratory analysis of these counts highlighted observable characteristics of real-world human mobility patterns: (i) heavy-tailed distribution and (ii) no significant network autocorrelation on all contiguous edges. This conclusion has a double-edge nature: it highlights the value of this research but at the same time emphasizes the need for further analysis in how these proxies can be reused to understand the observed patterns.

RQ2 - By map matching and locally remodeling the correlations seen in the route set, we focused on condensing this correlation into an understandable object attribute of a road network, making it readily available for further valorisation activities within the services hosted on route-sharing platforms. Of particular interest during this remodeling is the network-constrained distance weighting of the counts and β-transformation of the resulting attractiveness index. Both methods were introduced to integrate a user-centered remodeling, focusing on how the end-user perceives alternatives. This perception is impacted by both distance and natural or human barriers (i.e. network connectedness). Especially noteworthy is the reproducibility and flexibility of the documented methods to alleviate the impact of spatial difference (e.g. Country, region, rural/urban) in road networks and their characteristics.

RQ3 - Finally, we focused on an evaluation of the presented shortest-path alternative. While future work is still necessary to understand the local optimality of the generated route choices, the attractiveness model showed its potential as criterion to help understand shortest-path alternatives for the aforementioned type of cyclists within the study area. For example, the clear difference in the calculated scoring value per km within different shortest-path alternatives from specific routing engines underpins the predictive performance of the model. The generic evaluation setup and visualizations presented to analyze and compare path characteristics and route optimality also prove to be a valuable tools. These can prove to be very important in future valorisation.

We can conclude that this use case presented valuable insights in how route-sharing platforms can exploit the potential of routes shared in their communities to streamline efforts to improve the fitness of navigation services. It provided a novel way to maintain these services and focused on integrating a real-world suitability for a specific activity in a cost-effective way. The largest potential for cost reduction presented by this approach lies in future work combining both data-driven and expert-based models, thus mitigating the impact of missing information on the current expert-based models and, hence, reducing cost in bridging the *reality gaps* in maps.

References

Alivand, M., Hochmair, H., and Srinivasan, S. (2015). Analyzing how travelers choose scenic routes using route choice models. *Computers, Environment and Urban Systems*, 50:41–52.

Bakillah, M., Lauer, J., Liang, S., Zipf, A., Jokar Arsanjani, J., Loos, L., and Mobasheri, A. (2014). Exploiting big vgi to improve routing and navigation services. In Karimi, H. A., editor, *Big Data Techniques and Technologies in Geoinformatics*, pages 177–192. CRC Press, Boca Raton, Florida.

Bergström, A. and Magnusson, R. (2003). Potential of transferring car trips to bicycle during winter. *Transportation Research Part A: Policy and Practice*, 37(8):649–666.

Bierlaire, M. and Frejinger, E. (2008). Route choice modeling with network-free data. *Transportation Research Part C: Emerging Technologies*, 16(2):187–198.

Black, W. R. (1992). Network autocorrelation in transport network and flow systems. *Geographical Analysis*, 24(3):207–222.

Boen, F., Lagae, W., and Scheerder, J. (2011). *Vlaanderen Fietst! Sociaalwetenschappelijk onderzoek naar fietssportmarkt.* Academia Press, Ghent.

Broach, J., Dill, J., and Gliebe, J. (2012). Where do cyclists ride? a route choice model developed with revealed preference gps data. *Transportation Research Part A: Policy and Practice*, 46(10):1730–1740.

Damant-Sirois, G., Grimsrud, M., and El-Geneidy, A. M. (2014). What's your type: A multidimensional cyclist typology. *Transportation*, 41(6):1153–1169.

Dhakar, N. and Srinivasan, S. (2014). Route choice modeling using gps-based travel surveys. *Transportation Research Record: Journal of the Transportation Research Board*, 2413:65–73.

Eppstein, D. (1994). Finding the k shortest paths. In *Proceedings of the 35th Anual Symposium on Foundations of Computer Science*, pages 154–165. IEEE.

Fotheringham, A. S. (1981). Spatial structure and distance-decay parameters. *Annals of the Association of American Geographers*, 71(3):425–436.

Golledge, R. G. (1999). *Wayfinding behavior: Cognitive mapping and other spatial processes.* JHU Press, Baltimore.

Government of Flanders (2012). *Flanders in 2050: Human scale in a metropolis? - Spatial Policy Plan.* RWO, Brussels.

Haklay, M. (2010). How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and planning. B, Planning & design*, 37(4):682–703.

Haklay, M. (2013). Citizen science and volunteered geographic information: Overview and typology of participation. In Sui, D., Elwood, S., and Goodchild, M., editors, *Crowdsourcing Geographic Knowledge*, pages 105–122. Springer Netherlands, Dordrecht.

Haklay, M. and Weber, P. (2008). OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, 7(4):12–18.

Hochmair, H. H. and Navratil, G. (2008). Computation of scenic routes in street networks. In Car, A., Griesebner, G., and Strobl, J., editors, *Geospatial Crossroads@ GI_Forum'08: Proceedings of the Geoinformatics Forum Salzburg*, pages 124–133. Wichmann Verlag, Heidelberg.

Hochmair, H. H., Zielstra, D., and Neis, P. (2015). Assessing the completeness of bicycle trail and lane features in openstreetmap for the united states. *Transactions in GIS*, 19(1):63–81.

Hood, J., Sall, E., and Charlton, B. (2011). A gps-based bicycle route choice model for san francisco, california. *Transportation Letters*, 3(1):63–75.

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

Huang, H., Klettner, S., Schmidt, M., Gartner, G., Leitinger, S., Wagner, A., and Steinmann, R. (2014). Affectroute – considering people's affective responses to environments for enhancing route-planning services. *International Journal of Geographical Information Science*, 28(12):2456–2473.

Jiang, B., Yin, J., and Zhao, S. (2009). Characterizing the human mobility pattern in a large street network. *Phys. Rev. E*, 80:021136.

Kessler, F. (2011). Volunteered geographic information: A bicycling enthusiast perspective. *Cartography and Geographic Information Science*, 38(3):258–268.

Kimerling, A. J., Buckley, A. R., Muehrcke, P. C., and Muehrcke, J. O. (2009). *Map use: reading and analysis*, volume 6. ESRI Press, Redlands, California.

Luxen, D. and Vetter, C. (2011). Real-time routing with openstreetmap data. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 513–516. ACM, New York.

Menghini, G., Carrasco, N., Schüssler, N., and Axhausen, K. (2010). Route choice of cyclists in zurich. *Transportation Research Part A: Policy and Practice*, 44(9):754–765.

Mooney, P. and Corcoran, P. (2012). The annotation process in openstreetmap. *Transactions in GIS*, 16(4):561–579.

Mooney, P., Rehrl, K., and Hochmair, H. (2013). Action and interaction in volunteered geographic information: a workshop review. *Journal of Location Based Services*, 7(4):291–311.

Mooney, P. and Winstanley, A. (2006). An evolutionary algorithm for multicriteria path optimization problems. *International Journal of Geographical Information Science*, 20(4):401–423.

Newson, P. and Krumm, J. (2009). Hidden markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 336–343, New York, NY, USA. ACM.

Okabe, A. and Sugihara, K. (2012). *Spatial Analysis Along Networks: Statistical and Computational Methods*. Statistics in Practice. Wiley, Chichester, West Sussex.

Oksanen, J., Bergman, C., Sainio, J., and Westerholm, J. (2015). Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. *Journal of Transport Geography*, 48:135–144.

Pigram, J. J. J. and Jenkins, J. M. (1999). *Outdoor recreation management*, volume 5. Psychology Press, London.

Pippig, K., Burghardt, D., and Prechtel, N. (2013). Semantic similarity analysis of user-generated content for theme-based route planning. *Journal of Location Based Services*, 7(4):223–245.

Popescu, A., Grefenstette, G., and Moëllic, P.-A. (2009). Mining tourist information from user-supplied collections. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1713–1716. ACM, New York.

Prato, C. G. (2009). Route choice modeling: past, present and future research directions. *Journal of Choice Modelling*, 2(1):65–100.

Quddus, M. A., Ochieng, W. Y., and Noland, R. B. (2007). Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312–328.

Quercia, D., Schifanella, R., and Aiello, L. M. (2014). The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 116–125, New York, NY, USA. ACM.

Rattenbury, T., Good, N., and Naaman, M. (2007). Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 103–110. ACM, New York.

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. (2010). Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks*, 6(2):1–27.

Rubin, J. and Chisnell, D. (2008). *Handbook of usability testing: how to plan, design and conduct effective tests.* John Wiley & Sons, Indianapolis, Indiana.

Snizek, B., Nielsen, S., Alexander, T., and Skov-Petersen, H. (2013). Mapping bicyclists' experiences in Copenhagen. *Journal of Transport Geography*, 30:227–233.

Srivastava, M., Abdelzaher, T., and Szymanski, B. (2012). Human-centric sensing. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 370(1958):176–97.

Steenberghen, T., Aerts, K., and Thomas, I. (2010). Spatial clustering of events on a network. *Journal of Transport Geography*, 18(3):411–418.

Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, 1(2):85–99.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, 55(4):189–208.

Verstockt, S., Slavkovikj, V., Potter, P. D., Slowack, J., and de Walle, R. V. (2013). Multi-modal bike sensing for automatic geo-annotation geo-annotation of road/terrain type by participatory bike-sensing. In *Proceedings of the International Conference on Signal Processing and Multimedia Applications (SIGMAP), 2013*, pages 39–49. IEEE.

Zheng, Y., Zhang, L., Xie, X., and Ma, W.-Y. (2009). Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 791–800. ACM, New York.

Zheng, Y.-T., Yan, S., Zha, Z.-J., Li, Y., Zhou, X., Chua, T.-S., and Jain, R. (2013). Gpsview: A scenic driving route planner. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 9(1):3:1–3:18.

Zielstra, D., Hochmair, H. H., and Neis, P. (2013). Assessing the effect of data imports on the completeness of openstreetmap–a united states case study. *Transactions in GIS*, 17(3):315–334.

3

Cultural heritage routing: a recreational navigation based approach in exploring cultural heritage

"Already know you, that which you need"
Yoda, Star Wars: Episode VI - Return of the Jedi

Similar to the previous chapter, we again focus our attention on a crowd-based source featured on route-sharing platforms and how this can be used to maintain and improve an experience-oriented and thematic navigation service. We provide an overview of state-of-the-art protocols and procedures, garnering insights in how POI can create ancillary data to address the growing complexity of queries to activity and theme-specific navigation services. First, we address the lack of data in this context. Despite the current trend of making information location aware (i.e. geotagging (r)evolution), valuable information remains location unaware and, hence, useless within spatial knowledge-discovery techniques and GIS, central in this dissertation. As such, we present a method to automatic collection and multimodal enrichment of thematic information and condense this information in POI. Next, we present a spatial aggregation procedure of this location-based information and present a proof-of-concept of a path suggestions in a thematic navigation service based on this aggregated information. Within this use case, we highlight the potential of this approach for cultural heritage exploration, more specifically World War I battlefield exploration from a cyclist's perspective.

The work presented in this chapter is based on following publication. The authors contributed equally to this work:
Kevin Baker and Steven Verstockt. (2017). Cultural heritage routing: A recreational navigation-based approach in exploring cultural heritage. *ACM Journal on Computing and Cultural Heritage*, 10(4):1-20.

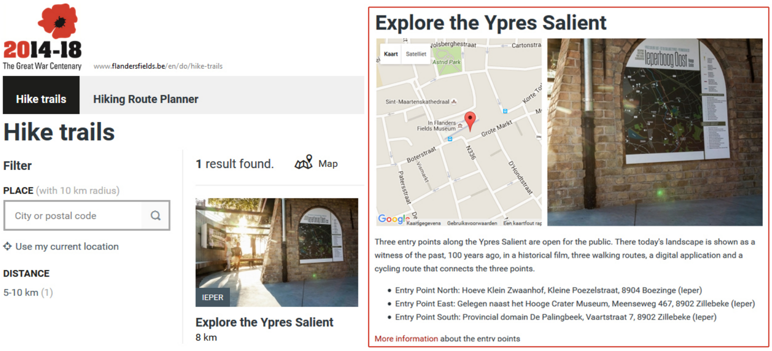
3.1 Introduction

A large proportion of cultural heritage resources, such as monuments, photos, and place descriptions, are geographically referenced, and, thus, can be identified by search terms that refer to a location. Several websites and mobile applications, not surprisingly, already publish cultural heritage content on a map (Kauppinen et al., 2011). The problem, however, is that the majority of these applications is still rather static and limited to well-known map exploration. In order to tackle this issue, geographic multimodal enrichment and thematic routing are seen as emerging trends in making sense of georeferenced cultural heritage data. Both topics have the potential to present valuable insights linked to the second research question in this dissertation and can be condensed in two additional research questions, specific for this use case:

- RQ1: How does the current **state-of-the-art** in **geographic enrichment** fit the use case of **cultural heritage**?
- RQ2: How can **contextual analysis** of **POI** be used to enrich navigation services?

First, geographic multimodal enrichment is the process of annotating and linking media items based on their location metadata and their content similarity. The location metadata or geotag reflect where media was collected or which place the media item describes. This transforms the media object, such as an image, video or text, into a point of interest (POI). As a result, these objects can be accessed with their geographical coordinates. This allows the georeferenced object (and geotagged media libraries in general) to be queried and analyzed in a spatial context and opens up a new world of possibilities for geographic related research and applications. An overview of recent research and applications on online georeferenced media is given in Luo et al. (2011) and Zheng et al. (2011). Both works address key concepts in this research topic and introduce several techniques to extract location information from text and images using geographic entity recognition (GER) and computer vision techniques.

Second, mode-specific thematic routes connect natural or artificial attractions with a certain theme (Nagy, 2011). For example, a World War I battlefield route, the test case described throughout this use case, connects topical POIs. A growing body of research focuses on thematic path generation algorithms. For a more in depth review of similar techniques, we refer to Souffriau and Vansteenwegen (2010). These algorithms model a graph (network) taking into account the location and the relevance of the geographically enriched POIs that are collected from a variety of web resources related to this specific theme. Once the thematic graph is constructed, the routing itself can be performed with a variety of techniques with different accuracy and computational properties. For general information on routing algorithms and an introduction on different routing techniques, we refer to Sanders and Schultes (2007). Considering sustainability, thematic routes serve education and leisure at the same time. They are a special opportunity for heritage tourism, targeting new groups by additional programs and attractions, making them more interesting, attractive and diversified (Nagy, 2012). Furthermore, they play an important role in common European programs too, like the Cultural Routes Program of the European Council [1].


2014-18
The Great War Centenary
www.flandersfields.be/en/do/hike-trails

Hike trails | Hiking Route Planner

Hike trails

Filter

PLACE (with 10 km radius)

City or postal code

Use my current location

DISTANCE

5-10 km (1)

1 result found. Map

IEPER

Explore the Ypres Salient
8 km

Explore the Ypres Salient

Three entry points along the Ypres Salient are open for the public. There today's landscape is shown as a witness of the past, 100 years ago, in a historical film, three walking routes, a digital application and a cycling route that connects the three points.

- Entry Point North: Hoeve Klein Zwaanhof, Kleine Poezelstraat, 8904 Boezinge (Ieper)
- Entry Point East: Gelegen naast het Hooge Crater Museum, Meenseweg 467, 8902 Zillebeke (Ieper)
- Entry Point South: Provincial domain De Palingbeek, Vaartstraat 7, 8902 Zillebeke (Ieper)

More information about the entry points

Figure 3.1: Generic thematic route for World War battlefield exploration in Ypres, Belgium.

Within this use case, we focus on recreational cultural heritage exploration and navigation, i.e., a challenging subdomain of thematic routing. Both geographic multimodal enrichment and thematic routing are further discussed in following sections.

3.1.1 Recreational cultural heritage exploration and navigation

As already mentioned, thematic routes have long been recognized as a key component in democratizing heritage for tourists. Long-distance hiking trails through natural heritage sites or sightseeing routes along a city's cultural heritage are only a few examples in contemporary tourism marketing and promotion strategies. Creating a deeper engagement with the visitors is the foremost reason to interweave and link leisure activities, such as cycling or

[1] http://culture-routes.net/cultural-routes

walking, with regional heritage. However, as interest in these leisure activities and tourism is still growing in policy and planning, so has the demand for cost-effective tools to popularize and market this type of experience-based tourism. To face this challenge, Economou (2015) stated that heritage is increasingly transferred to the digital sphere. Online and mobile applications offer opportunities to create a higher degree of differentiation and added value for tourists who encounter, explore and interact with heritage.

In this context, personalization in route planning applications for leisure activities receives much attention (Gao et al., 2010). In contrast to more generic thematic routes, such as the one shown in Figure 3.1 on page 53, the general idea is to provide users with ad-hoc created routes tailored to their specific leisure activity, location, preferences and interests. The dynamic and situational relevance of these personalized information retrieval services is a driving factor in their burgeoning use to engage a user's attention. Similar to well-known route planning applications for automotive traffic, a detailed and mode-specific annotation of digital road networks is essential to facilitate this type of route planning. However, modeling behavior in leisure and tourist activities is hampered by random agents and complex factors, such as scenery, affordance of the streetscape, popularity or group dynamics. Hence, a growing body of research focuses both on annotating and analysing road networks with specific, contextual information.

A brief overview of relevant literature in this research field is given below. First, Huang et al. (2014) and Quercia et al. (2014) collected people's in-situ affective response and their perception of cityscapes to enhance urban path suggestion in routing engines. Similarly, photo-sharing platforms have also been used to discover popular or attractive places (e.g., Popescu et al., 2009; Rattenbury et al., 2007; Zheng et al., 2009,1). Additionally, research has been conducted to semantically link georeferenced Wikipedia concepts to these places to create theme-based routes (Pippig et al., 2013). Planar-space potential surfaces and kernel density estimation are popular approaches in this growing body of research. Next, route choice modeling research focuses on the same topic but from a different perspective, modeling the trade-offs in path choices. For example, Alivand et al. (2015) analyzed how travelers choose scenic routes using route choices models. Finally, the advent of IT-mediated crowds sharing their leisure and tourist activity data, such as GPS tracking logs or POIs, has been creating new opportunities and possibilities. Knowledge discovery and data mining techniques successfully elicit and collect new information from these data sources to improve routing and navigation services (e.g. Bakillah et al., 2014; Oksanen et al., 2015; Reddy et al., 2010; Verstockt et al., 2013b).

There are two dominant trends that have a big impact on the success of thematic routing research: geotagging and the mobile sensing (r)evolution. Both trends are significant drivers in the collection of accurate, detailed, multimedia-rich information about thematic routes and points-of-interest (POIs) in a particular route search area. Based on these data, appropriate route recommendations can be generated.

3.1.2 Geotagging and mobile sensing (r)evolution

Over the last decade, geographic metadata annotation has become increasingly popular; that is, the popularity of the geotag is on the rise. On the one hand, this is caused by the increased use of internet-capable mobile devices with built-in GPS functionality. This results in a user-driven geographical enrichment of multimedia data. For example, this allows visitors of cultural heritage locations to tweet, share and log their visits. On the other hand, web technologies such as address geocoding and geographic entity recognition (Silva et al., 2006) can be used to automatically perform spatial enrichment on non-spatial data. In addition to the aforementioned user-driven annotation, a data-driven annotation receives a growing attention. With these data-enrichment technologies, content of cultural heritage websites can easily be mapped for spatial exploration. Furthermore, a trend is being observed that a growing number of data providers start to extend their cultural heritage databases themselves with aforementioned geotags, transforming their data into POIs. This information is crucial to easily integrate and query these databases in geographical applications.

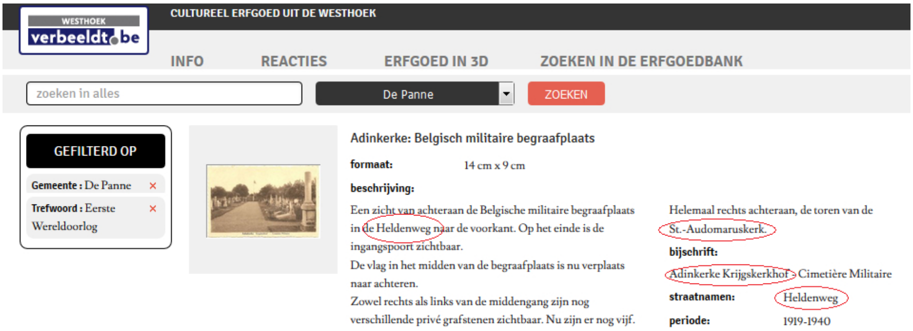
The geotag itself is a form of metadata which marks a multimedia object, such as an image, video or text message, with its location information (longitude and latitude coordinates). The majority of recent capture devices are able to automatically assign these kinds of tags. The huge benefit of geotagged media is that it allows multimedia objects to be browsed and arranged geographically. Photo-sharing websites such as Flickr[2] and Panoramio [3], for example, provide millions of geotagged images contributed by people from all over the world. In order to retrieve the multimedia data that is related to cultural heritage at a specific location, one can choose from several social media web services that support geo-based queries, such as the PANORAMIO geo-picture service and the DBPedia-based FlickrWrappr service (Becker and Bizer, 2009). The drawback of these media sharing platforms, however, is the limited control over the shared items and their quality. For this reason, we focus on more qualitative web resources targeted to a specific cultural heritage topic.

[2] www.flickr.com

[3] www.panoramio.com

Within this use case, the selected cultural heritage topic is Flanders Fields, i.e., a common English name of the World War I battlefields. Flanders Fields is particularly associated with battles that took place in the Ypres Salient, including the Second Battle of Ypres and the Battle of Passchendaele. Nowadays, this is still visible in the landscape and heritage in this region, comprising WWI relicts, warefare material remains and commemorative places. A good example of a Flanders Fields related webpage[4] that can easily be geotagged is shown in Figure 3.2.

Figure 3.2 shows the result of a query for World War I related data in De Panne, West Flanders. The retrieved pictures contain additional textual location information and a specific address element. Both can be analyzed using the geographic entity recognition and address geocoding that are discussed in the next sections. After successful geo-enrichment, the picture, its location and its description can easily be fed to a thematic routing engine or other types of cultural heritage applications, such as the mobile cultural heritage guide described in Aart et al. (2010).



Gemeente : De Panne ×

Trefwoord : Eerste Wereldoorlog ×

Adinkerke: Belgisch militaire begraafplaats

formaat: 14 cm x 9 cm

beschrijving:

Een zicht van achteraan de Belgische militaire begraafplaats in de Heldenweg naar de voorkant. Op het einde is de ingangspoort zichtbaar.
De vlag in het midden van de begraafplaats is nu verplaats naar achteren.
Zowel rechts als links van de middengang zijn nog verschillende privé grafstenen zichtbaar. Nu zijn er nog vijf.
Helemaal rechts achteraan, de toren van de St.-Audomaruskerk.

bijschrift:

Adinkerke Krijgskerkhof - Cimetière Militaire

straatnamen: Heldenweg

periode: 1919-1940

Figure 3.2: Cultural heritage website showing results for WWI query in De Panne, West Flanders. Geo-enrichment can be performed on address element and textual location descriptions.

Geotagging is not the only trend that facilitates the thematic routing data collection process. The mobile data collected by a large amount of recreational cyclists and hikers also contains a wealth of valuable information. Their mobile phones and GPS devices have increasingly evolved in functionality, features and capability over the last decade. With the continuous improvement in sensor technology built into these devices, and web services to aggregate and interpret the logged information, people are able to create, analyze and share information about their daily activities (Srivastava et al., 2012).

[4] www.westhoekverbeeldt.be

Within the mobile sensing (r)evolution, users act as sensor operators, that is, they contribute sensor measurements about their activities or the places they visit as part of a larger-scale effort to collect data about a population or a geographical area. This is the idea behind participatory or human-centric sensing. Recently, this tendency has also started to occur in the domain of geographic information systems (GIS). Where the process of mapping the Earth has been the task of a small group of people (surveyors, cartographers, and geographers) for many years, it starts to become possible now for everyone to participate in several types of collaborative geographic projects, such as OpenStreetMap and RouteYou (Schroth et al., 2011). These projects are built upon user-generated geographic content, so called volunteered geographic information (VGI). VGI makes it easier to create, combine, and share maps and supports the rapid production of geographic information.

Within the VGI data-sharing landscape, user-logged activity analysis (Biagioni and Krumm, 2013) is booming, with several platforms, such as Garmin Connect and RouteYou, which provide web services to query the activities of a particular user. Contrary to the majority of these activity loggers, RouteYou does not focus on performance analysis, but on finding routes that match a specific user and his needs. In order to perform this task, a large dataset of user routes is analyzed and queried, taking into account different user-specific query parameters such as the type of activity, the duration, and a thematic parameter, e.g., a particular cultural heritage topic (i.e., the focus of this use case). Based on all these parameters, the best matching routes are suggested. Thematic route suggestions for cultural heritage exploration, however, are not yet fully supported due to several data- and routing-related problems which are addressed in this use case.

3.1.3 Problem description

A major issue in thematic routing is the limited POI coverage and quality to perform accurate routing for user experience maximization. In order to tackle this *lack of data* problem, a methodology is needed to automatically collect cultural heritage POIs and improve their (meta)data quality. Furthermore, research is needed to optimize the weighting of a cultural heritage POI in routing algorithms. For example, the distance to the closest street, the type of recreational activity (e.g., biking, hiking or horse riding) are parameters that will have an impact on this weighting procedure and need to be taken into account. Finally, contextual routing strategies are needed to fulfill the overall user satisfaction in recreational cultural heritage exploration. In this use case we address each of these problems and present different building blocks to solve these issues.

As mentioned in the previous section, this use case takes World War I battlefield exploration as a specific cultural heritage topic (Winter, 2011) and investigates the POI coverage and quality in existing routing databases. Moreover, Belgium has a cycling-savvy society which uses bicycles both as a major mode of utilitarian transport and for leisure purposes. Similar to other cycling-savvy societies such as the Netherlands and Denmark, attention is given to cycling-specific road infrastructure and creating awareness for slow traffic. Exemplary is the more than 12 000 km signposted cycling network transecting the northern parts of Belgium. In the context of our use case, we use this network in our study area to apply our methodology and convert this to a thematic network for cyclists focused on WWI battlefield exploration. This makes our study area a valid and interesting case study for our methodology.

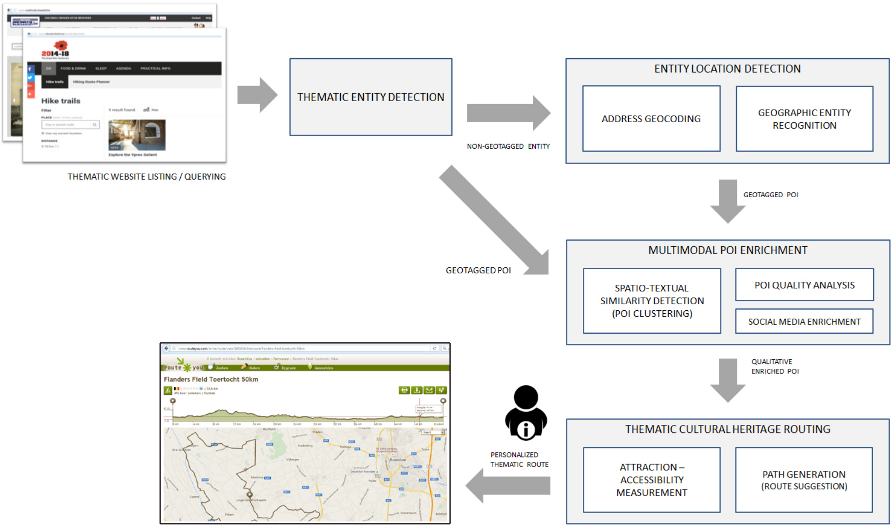

2014-18
Hike trails
THEMATIC WEBSITE LISTING / QUERYING
THEMATIC ENTITY DETECTION
NON-GEOTAGGED ENTITY
ENTITY LOCATION DETECTION
ADDRESS GEOCODING
GEOGRAPHIC ENTITY RECOGNITION
GEOTAGGED POI
GEOTAGGED POI
MULTIMODAL POI ENRICHMENT
SPATIO-TEXTUAL SIMILARITY DETECTION (POI CLUSTERING)
POI QUALITY ANALYSIS
SOCIAL MEDIA ENRICHMENT
QUALITATIVE ENRICHED POI
Flanders Field Toertocht 50km
PERSONALIZED THEMATIC ROUTE
THEMATIC CULTURAL HERITAGE ROUTING
ATTRACTION – ACCESSIBILITY MEASUREMENT
PATH GENERATION (ROUTE SUGGESTION)

Figure 3.3: Proposed workflow for cultural heritage POI collection, (meta)data enrichment and recreational cultural heritage routing.

A general overview of the proposed setup is shown in Fig. 3.3 on page 58. The remainder of this use case is organized in a similar manner. After introducing the procedure to select thematic entities, we present the geographic annotation and media enrichment techniques that are used in our cultural heritage POI collection. We subsequently discuss our proposed tool set. Next, we introduce our novel thematic routing algorithm and compare its performance with traditional routing methods. Finally, we conclude this use case with and point out directions for future work.

3.2 Cultural heritage POI collection

In order to perform heritage-based thematic routing, it is important to have a representative coverage of cultural heritage locations within the study area. Currently, however, the number of cultural heritage POIs and their (meta)-data quality in dedicated databases is still a fraction of the large amount of cultural heritage multimedia that can be found on the web. Figure 3.4 on page 60, for example, shows the current POIs of RouteYou that are labeled as CH - WWI compared to the WWI heritage dataset of the Flemish organization for Immovable Heritage[5]. Analyzing the heatmaps in Figure 3.4 shows the added value of incorporating POI data from thematic websites. In order to extend the thematic RouteYou POI sets, we suggest an automatic collection mechanism for retrieving locations and (meta)data of thematic cultural heritage POIs. Given a website listing of sources for geotagged and non-geotagged thematic entities (POI names), a dataset of location-aware POIs can be garnered that can be fed to the multimodal enrichment building block. We focus on two important building blocks which receive a burgeoning attention in both academic literature and commercial applications: (i) address geocoding and (ii) geographic entitiy recognition. We give an overview of current state-of-the-art and emphasize their value in our specific use case.

3.2.1 Collection of geotagged thematic cultural heritage POIs

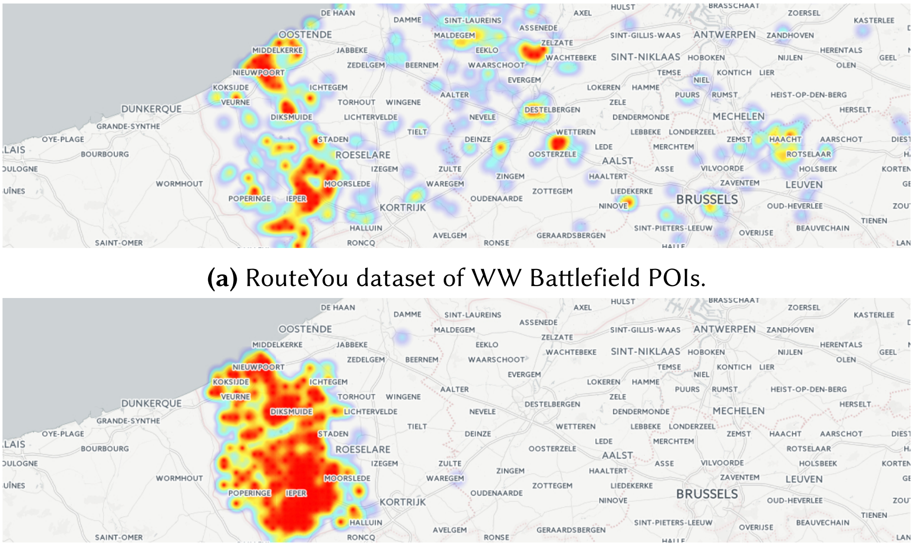
The first step in analyzing the website listing of sources for thematic geotagged and non-geotagged entities (POI names) is to extract or detect the exact location of the different entities. If the website is well structured, e.g. with HTML hCard-address microformats and tags[6], address geocoding can be used. For non-structured web documents, we propose to use geographic entity recognition techniques.

3.2.1.1 Address geocoding

Address geocoding is the process of determining an estimated latitude and longitude position for the location of a street address. In the example given in Figure 3.2 on page 56, address geocoding can be used to convert the address element “Krijgskerkhof, Heldenweg, De Panne” into the coordinates (51.0757098,2.6019398) of this place.

[5] https://inventaris.onroerenderfgoed.be/

[6] http://www.htmlandcssbook.com/extras/introduction-to-hcard/



(a) RouteYou dataset of WW Battlefield POIs.

(b) WW Battlefield POIs in dataset of Flemish organization for Immovable Heritage.

Figure 3.4: Comparison between current WW Battlefield POIs in a) RouteYou database and b) the dataset of the Flemish organization for Immovable Heritage.

First of all, a parser (such as the Microformats Parsing API[7]) will break down the address element into a number of components. Then, address standardization identifies each address component (e.g., street number, street name, city and zip code) and places them in order. Finally, the values for each address component are matched to the reference database and an estimate of the spatial location is given. Several matching problems can occur during this process, such as misspelled street names, outdated reference data, and incorrect numbers.

In order to automatize the address geocoding process, several address geocoding web services can be used, such as the ArcGis Geocoder [8] and the Mapquest Geocoding API [9]. The former one is integrated in ESRI's geospatial processing programs and can be used, for example, to automatically geocode a table of addresses. It is important to mention, however, that a large amount of cultural heritage websites do not (yet) contain structured address elements. This, of course, limits extensive use of address geocoding for the geotagging of thematic cultural heritage POIs. For non-structured websites, we propose the use of geographic entity recognition (GER).

[7] http://www.alchemyapi.com/products/alchemylanguage/microformats-parsing

[8] http://geocode.arcgis.com/arcgis/index.html

[9] http://www.mapquestapi.com/geocoding/

3.2.1.2 Geographic entity recognition

Named Entity Recognition (NER) labels sequences of words in a text belonging to predefined categories, such as the names of persons, organizations, and locations. NER plays a significant role in many application domains, such as information extraction, summary generation, document classification and internet search optimization. In our work, NER is used to create a geographical representation of a text or web page, based on the geographic entities that can be detected using a specific set of NER techniques, i.e., the GER techniques.

In broad terms, two main types of GER techniques can be distinguished: knowledge-based and learning-based GER. Knowledge GER techniques use regular expressions, rules and context patterns to detect a particular entity type. In general, this type of NER technique is very precise and only needs small amount of training data. The drawbacks of Knowledge NER, however, are its expensive development cost and domain dependency. Learning systems, on the other hand, have a higher recall and do not need grammars, but require a lot of training data. For Geographic Entity Recognition (GER), learning-based systems have proven to perform best (Mikheev et al., 1999; Silva et al., 2006).

An important aspect in the success of GER are the languages that are supported by the NLP tool. Depending on the language of the input text, different NLP tools can be used. However, correct language detection will be needed in order to select the most appropriate NLP tool in an automatic way. The language detection itself can be performed using the language identifier of the FP7 OpeNER project[10].

For English texts we have evaluated the entity extraction demo of the Dandelion API[11], which performs well in extracting historical concepts and global location entities, such as city and country names. An example of the output that the Dandelion API generates is given in Figure 3.5 on page 63. For accurate location/address extraction, however, further research is needed. Combining this GER approach with a regular expression based approach like Pyap[12] will be part of our future work. Since a lot of Flanders Fields related texts are written in Dutch, we have also evaluated different Dutch-based GER tools, such as Frog[13] (van den Bosch et al., 2007), Namescape [14], and iRead+ (Paulussen et al., 2014). The latter tool extracts Dutch entitites/lo-

[10] http://opener.olery.com/language-identifier

[11] https://dandelion.eu/semantic-text/entity-extraction-demo/

[12] https://pypi.python.org/pypi/pyap

[13] http://languagemachines.github.io/frog/

[14] http://ner.namescape.nl/namescape/tagger

cations out of structured and unstructured text documents and presents geographic features with metadata. The geocoding is done with help of an OpenStreetMap (OSM) gazetteer of Flanders, i.e. an existing list of entities that automatically can be generated from OSM data.

Important to remark is that geographic entities are prone to temporal change both in toponymy and geographical extent. As such, a GER algorithm should also be able to take into account the temporal dimension of place names and their geographic properties at that moment. However, the major online world gazetteers still ignore time. They provide millions of place names, but lack the tracking of name changes over time. They mainly focus on actual geographic references and contain only a small amount of historic place name information and spelling variants of places. As suggested by Berman (2008), gazetteers should be extended with place names over time (and their date of validity information and location description), turning them into spatio-temporal gazetteers and enabling Geo-Temporal Information Retrieval. Time-location (meta)data should be included and can be collected in various ways, e.g. from online mapping services/spatial databases providing or sourced initiatives (like HeuristScholar [15] and the Perseus Project [16]). Several initiatives in building historical gazetteers are being undertaken, like the Edinburgh geoparser Grover et al. (2010) used in the Google Ancient Places project [17], the Iberian historical gazetteer presented in Hibberd and Owens (2015), and the work of Blank and Henrich (2015) that investigates the geocoding of place names from historic route descriptions. For Dutch texts, we did not yet find an appropriate gazetteer that takes into account the "naming during time" aspect. As such, we currently only focus on actual geographic references collected by the above mentioned Dutch gazetteers.

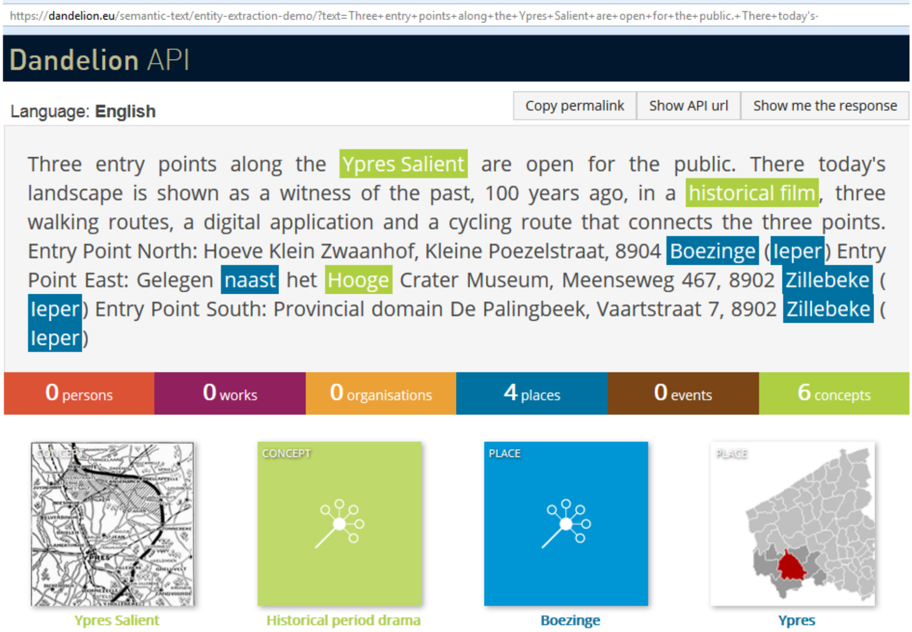
3.2.2 Multimodal enrichment of the POI

Based on the geotag of the POI (retrieved by address geocoding or GER), its entity name and related entities that appear in the text (detected by NLP), we can perform the multimodal enrichment of the POI and improve its (meta)data quality. First of all, we construct the geo-textual media object o, which is represented as a tuple $o =< S, p, t, M >$, where S is a set of text annotations, p is a location, t is a time indicating the last object modification, and M is the list of available media files. The creation time is used to sort information and analyze/visualize the POI history. In order to detect t, we use the last-modified header which can be found in the metadata of the web document(s).

[15] http://heuristscholar.org/

[16] http://www.perseus.tufts.edu/

[17] https://googleancientplaces.wordpress.com/about/



https://dandelion.eu/semantic-text/entity-extraction-demo/?text=Three+entry+points+along+the+Ypres+Salient+are+open+for+the+public.+There+today's-

Dandelion API

Language: English

Three entry points along the Ypres Salient are open for the public. There today's landscape is shown as a witness of the past, 100 years ago, in a historical film, three walking routes, a digital application and a cycling route that connects the three points. Entry Point North: Hoeve Klein Zwaanhof, Kleine Poezelstraat, 8904 Boezinge (Ieper) Entry Point East: Gelegen naast het Hooge Crater Museum, Meenseweg 467, 8902 Zillebeke (Ieper) Entry Point South: Provincial domain De Palingbeek, Vaartstraat 7, 8902 Zillebeke (Ieper)

Figure 3.5: Dandelion API for named entity recognition.

The media files M are found by document element analysis in the document object model (DOM). Important to mention, however, is that this DOM-based technique can only be used if the web document is well structured, i.e., using HTML5 image/video/audio elements. For less structured websites, multimedia hyperlink scraping techniques can be used. Next, when all object tuples have been constructed, we perform a spatio-textual similarity clustering over the entire set of POI tuples in our dataset. The spatio-textual similarity clustering groups objects that are spatially close and textually similar (Bouros et al., 2012). This grouping forms the basis for our multimodal-enrichment approach.

3.2.2.1 Spatio-textual similarity clustering of POIs

Given a collection R of geo-tagged objects with associated textual descriptors, the spatio-textual similarity clustering problem is to identify all pairs of similar objects that are close in data and distance (Rao et al., 2014; Zheng et al., 2010). In general we want to find $(o1, o2)|o1, o2 \in R$, where $overlap$ $(o1, o2) \geq t$ and $o1 <> o2$. Different technologies can be used in order to measure this overlap and different weightings for S,p and M can be taken into account. Text similarity of text annotations in S, for example, can be measured using the Dandelion API's text similarity scoring tool, which estimates the textual semantic overlap. For multimedia similarity in M, deep learning-based semantic analysis seems the most appropriate technique in state-of-the-art research. In Karpathy and Fei-Fei (2017) for example, they generate natural language descriptions (i.e., textual image labels) in order to represent and cluster the images. Important to mention is that it is out of the scope of this use case to give an overview of all similarity metrics that can be used. A minimal set of techniques is selected based on our POI dataset evaluation. However, in the proposed framework additional or alternative metrics, such as those discussed in Bar et al. (2012) and Bell and Bala (2015), can easily be investigated and evaluated.

The example in Figure 3.6 discusses the result of the Dandelion text similarity API [18] for two short segments of text describing the "Last Post in Ypres" POI. The segments were extracted from two different web resources. The presented approach calculates a semantic and syntactic similarity. The latter summarizes the grammatical resemblance of both segments. Semantic similarity on the other hand is driven by the question "How much is term X related to term B?". Considering both text segments, their syntactic similarity is low (29%), but have a very high semantic similarity of 80%. Furthermore, when comparing the semantic concepts that are generated from the images that appear alongside the texts, we get an image similarity of 62.5% (based on the cosine similarity [19] of the semantic image tags). Averaging both the textual and image similarity results, gives an overall similarity of 71% between the two POI instances. In order to generate the semantic image concepts, we used the MIT places CNN for scene recognition (Zhou et al., 2014), and the Cafe deep learning framework for object recognition (Jia et al., 2014). Figure 3.6 only shows object tags detected by Cafe, but similar scene tags are retrieved with the MIT places CNN.

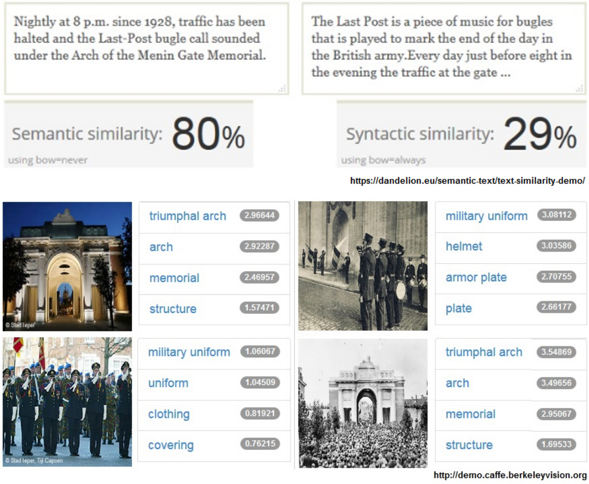
3.2.2.2 Social media querying for additional cultural heritage multimedia

In order to retrieve additional multimedia data (i.e., images, video, text, and music) that are related to the location of the selected POI, we can feed the POI and its coordinates to a set of social media web services that support geo-based queries. Our geo-based DBpedia Flickr service, for example, makes use of DBpedia, SPARQL and the Flickr API to perform this task (Verstockt et al., 2013a). DBpedia [20] is a crowd-sourced community effort to extract structured

[18] https://dandelion.eu/semantic-text/text-similarity-demo/

[19] https://bioinformatics.oxfordjournals.org/content/suppl/2009/10/24/btp613.DC1/bioinf-2008-1835-File004.pdf

[20] http://wiki.dbpedia.org/



Nightly at 8 p.m. since 1928, traffic has been halted and the Last-Post bugle call sounded under the Arch of the Menin Gate Memorial.

The Last Post is a piece of music for bugles that is played to mark the end of the day in the British army.Every day just before eight in the evening the traffic at the gate ...

Semantic similarity: 80%
using bow=never

Syntactic similarity: 29%
using bow=always

https://dandelion.eu/semantic-text/text-similarity-demo/

triumphal arch	2.96644
arch	2.92287
memorial	2.46957
structure	1.57471

military uniform	3.08112
helmet	3.03586
armor plate	2.70755
plate	2.66177

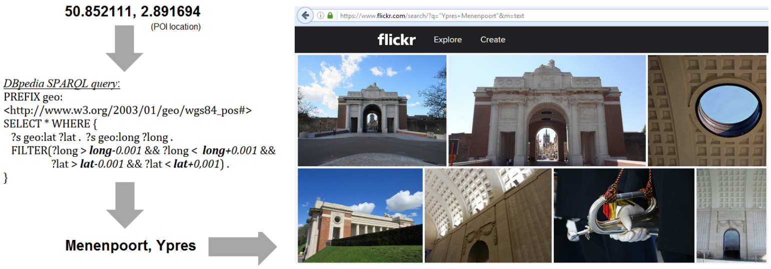
military uniform	1.06067
uniform	1.04509
clothing	0.81921
covering	0.76215

triumphal arch	3.54869
arch	3.49656
memorial	2.95067
structure	1.69533

http://demo.caffe.berkeleyvision.org

Figure 3.6: Semantic text and image similarity of two POI instances of the "Last Post in Ypres".

information from Wikipedia and make this information available on the Web. SPARQL [21] is similar to SQL and is a query language to select data from DBPedia. Our approach is split up in 2 steps, as shown in Figure 3.7. First, we extract relevant DBpedia entities using a location-based SPARQL query. In the second step, we use the retrieved DBpedia entities as parameters in a query to the Flickr API[22]. This process (illustrated in Figure 3.7) can be used for both location- and entity-based search.



DBpedia SPARQL query:

```
PREFIX geo:
<http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT * WHERE {
  ?s geo:lat ?lat .  ?s geo:long ?long .
  FILTER(?long > long-0.001 && ?long <  long+0.001 &&
            ?lat > lat-0.001 && ?lat < lat+0,001) .
}
```

Figure 3.7: DBPedia based Flickr Querying for additional Cultural Heritage Multimedia.

[21] https://www.w3.org/2009/sparql/wiki/Main_Page
[22] https://www.flickr.com/services/api/

3.2.3 Quality estimation of cultural heritage POIs

Despite the wide agreement on the need to produce high quality metadata, there is less consensus on what high quality means and even less on how it should be measured. In many instances, POIs are produced with inadequate metadata or, in the worst instance, no metadata at all. Producing qualitative metadata can be time consuming. However, it is important for thematic routing (like for other data-driven multimedia products) to have reliable, accurate, and coherent information (Cartwright et al., 2007). Consistent with this growing demand, we devise a quality estimation based on four measures that have been proposed in Ochoa and Duval (2009) and Bruce and Hillmann (2004), that is, the completeness, coherence, provenance and accuracy metrics. The combination of all these metrics generates a score for each resulting cultural heritage POI from the above described building blocks. These scorings can be used by the thematic routing proposed in the next section. Important to mention is that it is a set of metadata metrics based on the same quality parameters used by human reviewers but with the difference that they can be calculated automatically.

3.2.3.1 Completeness metric

A metadata instance of a cultural heritage POI should describe the POI as fully as possible and metadata fields should be filled in for the majority of the POI descriptors in order to make sense for a thematic routing service. Furthermore, there is certain information, that, due its nature, should be present and is more important than other metadata, i.e., not all metadata elements are equally relevant to all contexts. Different weighting parameters can be defined to express the importance of a metadata element. The calculation of the completeness metric $Q_{comp}$is given in Eq. 1, where α_i is the relative importance of the ith field, $P(i)$ is 1 if the ith field has a no-null value, and 0 otherwise, and N is the number of fields defined in the metadata standard.

$$Q_{comp} = \frac{\sum_{n=1}^{N} \alpha_i * P(i)}{\sum_{n=1}^{N} \alpha_i} \tag{3.1}$$

3.2.3.2 Accuracy metric

The information provided about the POI in the metadata instance should be as correct as possible. Several metadata errors, e.g., broken links, spelling or typographical errors and inaccurate technical properties (such as size or format metadata errors), affect this quality dimension. In order to measure these errors, we use easy to calculate accuracy metrics, such as those proposed in Moen et al. (1998). Moen et al. count the number of "visible" errors in each record (e.g., spelling or typographical errors, file formatting errors, or incorrect date formats) using spelling, format and regular expressions checkers and express the derived accuracy metrics $g(x_i)$ of the i-th metadata field as 0 if an accuracy issue is detected and 1 if no problem is found. The combined POI metadata accuracy Acc_{meta}, based on Bellini and Nesi (2013), produces a weighted average across all the N_{fields} metadata fields and becomes:

$$Acc_{meta} = \frac{\sum_{i=1}^{N_{fields}} g(x_i) \times w_i}{\sum_{i=1}^{N_{fields}} w_i} \tag{3.2}$$

where w_i is a weighting factor for the i-th field. Relevant fields to be taken into account in the evaluation of the quality assessment are the date, file format/size, language, subject, title and type, as proposed by Bellini and Nesi (2013). A similar approach for calculating the metadata accuracy is discussed in Ochoa and Duval (2006).

In addition to the Acc_{meta} metric, we introduced a spatial accuracy component DEV_{loc} (shown in Eq. 3) which focuses on the spatial differences in the retrieved set of POI locations loc_i. DEV_{loc} indicates the geographically diverse distribution of each POI and is calculated by computing the standard deviation of all its N geographical locations loc_i with the mean geographical location $\overline{loc}$ of the POI. Similarly as in Hughes et al. (2012), all DEV_{loc} values are normalized in the range 0 - 1. If huge differences are observed in the POI locations, DEV_{loc} will be close to 1, otherwise it will be close to zero.

$$DEV_{loc} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(loc_i - \overline{loc})^2} \tag{3.3}$$

The final Q_{acc} score for each POI (shown in Eq. 4) is a combination of its Acc_{meta} and the geographical variation of its locations (DEV_{loc}). This can be formally defined as:

$$Q_{acc} = Acc_{meta} \times (1 - DEV_{loc}) \tag{3.4}$$

3.2.3.3 Coherence metric

The POI coherence is related to the degree to which all metadata items describe the same cultural heritage object in a similar way. All items should describe the same resource and if the information is a combination of data from different web resources, the coherence in the data should be high. For text items, we perform a correlation-based estimation of the coherence. The semantic distance is calculated between the different text items and the average semantic distance is used as a measure of the coherence quality. This method is based on Vector Space model techniques used in Information Retrieval (IR) to calculate the distance between texts, but can also be used on semantic descriptions of other media types. A more detailed description of this technique can be found in Ochoa and Duval (2009). A low coherence value for a considerable number of POI instances could be the signal of poor titles or descriptions. More complicated coherence metrics, e.g. based on semantic co-occurrence graphs (Sonawane and Kulkarni, 2014), will be investigated in future work. For images and other types of multimedia content, coherence can be measured using semantic similarity metrics (Fang and Torresani, 2012).

3.2.3.4 Provenance metric

Provenance quality measures the trust in the web resource of the metadata instance, i.e., it measures the quality of the web resource from which the POI is extracted, not the quality of the POI itself. Knowledge about who created the POI instances, the level of expertise, the methodologies used at data collection and the transformations the metadata has passed through, could provide insight into the quality of the instance. However, a scalable way to estimate the provenance of a source of metadata is to analyze the quality values of all its N instances. First, we obtain an average quality (Q_{avg}) for each POI_i of this web resource based on the above mentioned metrics, and afterwards, the average of all these Q_{avg} values of a particular web resource are calculated using Eq. 5. Once this provenance quality Q_{prov} of the source has been obtained, it is assigned to each of its POIs.

$$Q_{prov} = \frac{\sum_{i=1}^{N} Q_{avg}(POI_i)}{N} \tag{3.5}$$

Important to mention is that Q_{prov} can only be calculated once the other quality metrics have been calculated. Furthermore, each time a new cultural heritage POI object is imported, the reputation of its web resource should be recalculated and the Q_{prov} of all its objects needs to be updated. As such, the provenance of a source is not static, but a dynamic changing value over time.

Each of the above mentioned metrics are combined in the POIs final quality score Q_{POI}, which is a weighted average of the completeness, accuracy, coherence and provenance scores. We leave it up to developers to assign the weights. Depending the type of application they are focusing on they can decide themselves which weights they want to give to each metric. In our experiments, we currently use equal weights for all metrics.

3.3 Routing in cultural heritage environments

To facilitate thematic routing, we now focus our attention on conveying and aggregating the previously discussed location-based information and its quality index on a network. We follow a similar workflow as described in the previous chapter to remodel the count variable on the network. First, we focus on an aggregation approach of the location-based information and calculate a thematic index on the network. We direct our attention on well-known attraction-accessibility measures (AM) to condense potential interactions between recreational cyclists and cultural heritage along the network. Second, we remodel the index to a cost function to make it readily applicable in performance demanding shortest-path problems. As a result, we have a thematic and activity-specific graph.

3.3.1 Attraction-accessibility measure

Following the classification of Okabe and Sugihara (2012), we define cultural heritage sites as alongside-network events. The basic idea in this methodology is to model the spatial interaction between the cycling network and the alongside-network events and calculate an AM for every location on the network. Previous studies (see for example Miller, 1999) have proposed several methods of calculating an AM. For brevity, we highlight one method. Assuming that both the number of events and the accessibility of the event have a positive impact on our measure, we use an addition of all distance-weighted event scores within a certain distance band. Similar to other network analysis methods, we divided the network in basic spatial units (BSU) and calculate AM for these units. We considered 100m long non-overlapping segments. Consistent with the work of Steenberghen et al. (2010), we used a moving-segment analysis. This approach is similar to well-known moving-window analysis in raster toolsets, doing calculations on local subsets of the data. In doing so, we calculated the AM for every segment in our network based on a subset of cultural heritage sites within an influence distance of the specific segment. Eq. 3.6 presents AM for an event e based on all segments n in our network.

$$AM_e = \sum_{g=1}^{n} W_{eg} e_{score} f(D_{eg}) \tag{3.6}$$

We introduce D_{eg} as a distance metric between event e and segment g. Because we are modeling movements limited to a road network and constrained by time, distance and connectedness, we used the Dijkstra shortest path on the underlying road network as a heuristic for these constrained movements. Previous research (e.g. Miller, 1999; Okabe and Sugihara, 2012; She et al., 2015; Steenberghen et al., 2010) has already proven that network distance is a more realistic and more generic approach in spatial clustering of location-based events. For simplicity, distance metric D_{eg} is calculated between the event point and the midpoint of the segment. Next, The weight W_{eg} represent a binary weight: if w_{eg} = 1 then event e lies within the influence distance of segment g, and 0 otherwise. Figure 3.8 on page 71 represents AM for one segment (red dot) in the cycling network. All shortest paths within the influence distance of a this segment are shown as black lines. Finally, $f(D_{eg})$ is a distance decay function weighting the event score. This decay function reflects the impact distance has on the relevance of the event for segment g. As time-distance is a strong deterrent factor in valuing leisure activities (Pigram and Jenkins, 1999), we use an inverse-distance decay function and introduce a distance band of 5 km. Figure 3.9 on page 71 serves as an exploratory visualization of the spatial distribution of the cultural heritage sites and our AM. In addition, we also want to highlight the impact of a planar-space AM with as-the-crow-flies distance and the resulting overestimation of the impact of an event on a location. When we compare the devised network-space AM with a planar-space AM, we see an overestimation. Figure 3.10 on page 72 presents a comparison between both in our study area. The cluster around zero on the y-axis is a result of segments with few or no sites in their influence distance and is clearly visualized in Figure 3.11 on page 73.

3.3.2 Path exploration and generation

In the following section we address the generation of a shortest-path alternative based on the attraction-accessibility measure. We compare two methods that are often used to generate alternative shortest paths based on a criterion. A first method implements the k shortest paths algorithm to maximize this criterion (Eppstein, 1994). This algorithm generates the shortest path between a source-destination couple and subsequently produces the shortest-

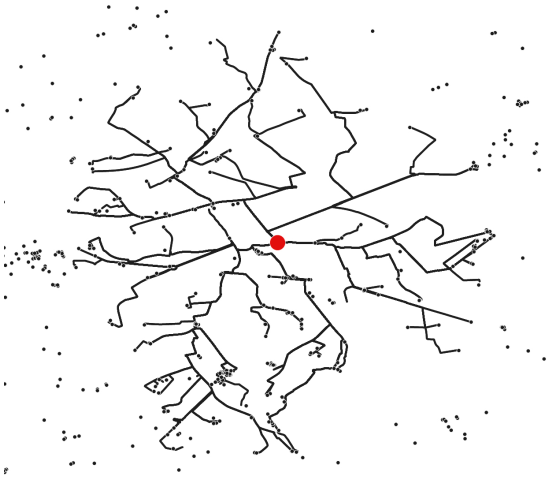


Figure 3.8: Mapping of the parameters in the attraction-accessibility measure for one segment (red dot) in the cycling network. Black dots represent the cultural heritage sites in the study area and the black lines represents the distance metric (i.e. Dijkstra shortest paths on the underlying road network within the influence distance of a segment).

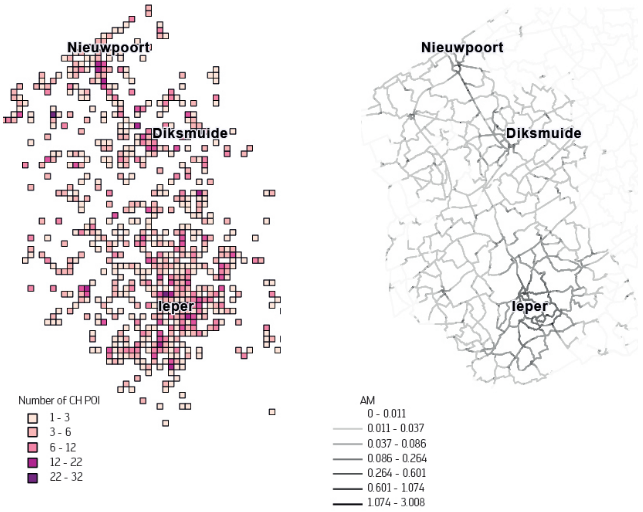


Figure 3.9: The right side image visualizes the spread of the network-space AM. As a reference, a grid-density plot of the cultural heritage sites is added on the left side.

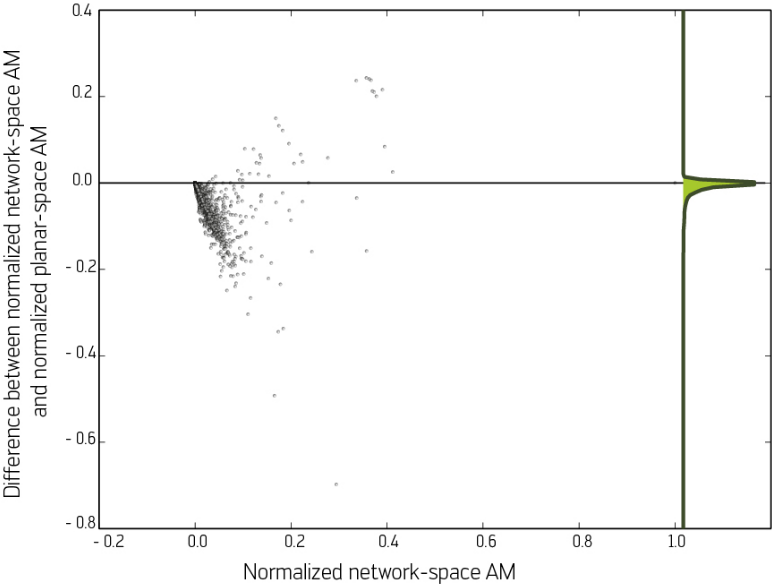


Figure 3.10: Visualization of the normalized network-space AM in a function of the difference between the normalized network-space AM and the normalized planar-space AM. This plot clearly shows the overestimation in the planar-space AM to the network-space AM. Next, we also see that a low number of segments receive a higher AM while using the shortest path as distance metric. The green line represents the density estimation of the plotted points.

path alternatives with increasing length between this s-d couple. In the context of our approach, this allows the exploration of the calculated AM in different paths in the cycling network between source and destination. This method also allows the analysis of the trade-off between increasing length of the shortest-path alternatives and AM maximization.

As efficiency and tractability of route planning applications becomes more important, an often preferred second method is the implementation of the Dijkstra shortest path with an alternative cost function. From a theoretical perspective, the cost function involves the approximation of the theme-specific travel expense of every segment in a network. While the use of a single-cost method has often been described as a naive approach to generate optimal paths (Mooney and Winstanley, 2006), many state-of-the-art routing algorithms in performance-demanding applications still use this method (Luxen and Vetter, 2011). In order to facilitate thematic routing in these algorithms, we propose a cost function based on our AM. First, we invert and normalize

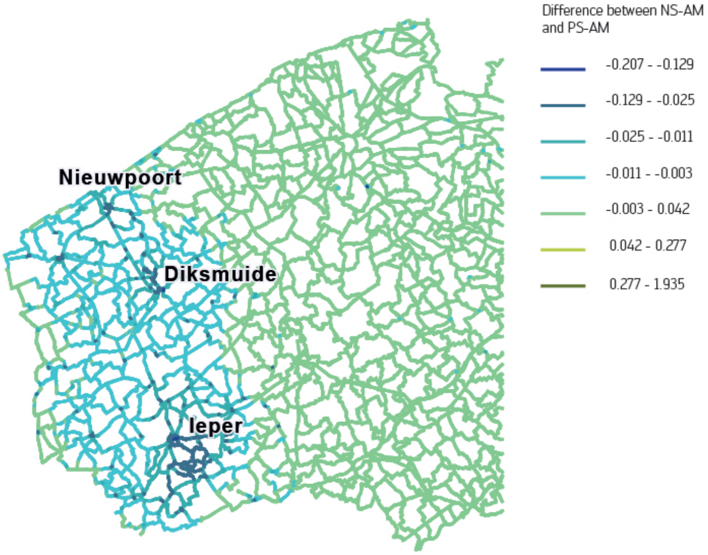


Figure 3.11: This figure gives a spatial impression of the overestimation seen in Figure 3.10 on page 72.

our AM resulting in a cost c for every segment i for heritage exploration. 0 means a low cost for heritage exploration, thus a high relevance for a thematic shortest-path alternative, and 1 means a low relevance. Because all BSU in our network analysis are approximately equal in length, it is not necessary to incorporate segment length in our cost function. This cost c_i, however, does not sufficiently emphasize the relevant segments in the sparse cycling network to generate an alternative path. As a second step and similar to the research of Pippig et al. (2013) and Hochmair and Navratil (2008), we incorporate a scaling factor a to increase the reduction in cost by incorporating relevant segments to make up for the gain in path length. Best results were generated with $a = 10^9$ in Eq. 3.7.

$$\text{segment cost} = a^{c_i} \quad \text{where} \quad c_i = 1 - \frac{AM_i}{max(AM_i)\forall_i} \tag{3.7}$$

In the following paragraph, we present the resulting paths generated by the above methods. This comparison allows us to evaluate the optimality of the shortest-path alternatives generated by both methods. In our context, optimality of a path is defined as the maximization of the sum of AM along our path with a minimal increase in length to the shortest path. Figure 3.12 on page 74 shows two randomly chosen couples. The left plot indicates that the second method approaches the most optimal choice, while the right plot depicts a situation where our alternative cost function chose a less efficient path (i.e. a longer path) while maximizing the sum of AM.

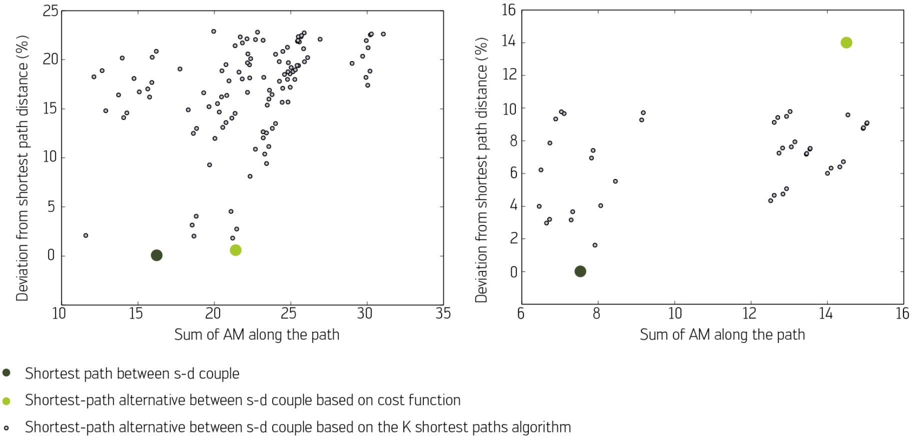


Figure 3.12: Comparison of the resulting paths generated by the k shortest path algorithm and the alternative cost approach. Both plots represent a randomly chosen s-d couple with gray dots representing k ($k = 100$) shortest-path alternatives, the blue dot representing the shortest path and the green dot the chosen shortest-path alternative based on our cost function with $a = 10^9$

Next, we analyze all shortest paths and their thematic alternative based on our cost function within a specific geographical region. We chose to use the geographic region of the Ypres salient, simulating users planning a route in this region by bike in the context of battlefield exploration around Ypres. Exploratory analysis of this case study showed that the sum of AM along the path on average increases by 6.52 with a standard deviation of 4.72. Length on average increases by 2140 m with a standard deviation of 2320 m. Additionally, Figure 3.13 and Figure 3.14 on page 76 indicates that our alternative cost function successfully identifies salient corridors for battlefield exploration and, thus, thematic routing applications. The higher number of times the edges are used around the city of Ypres, fulfills our expectations of more alternative paths along locations with a higher AM.

3.3.3 Discussion

While the above results underpin the feasibility and effectiveness of our approach, some points of discussion have to be taken into account. As already mentioned, we probably underestimate the complexity and the concurrent

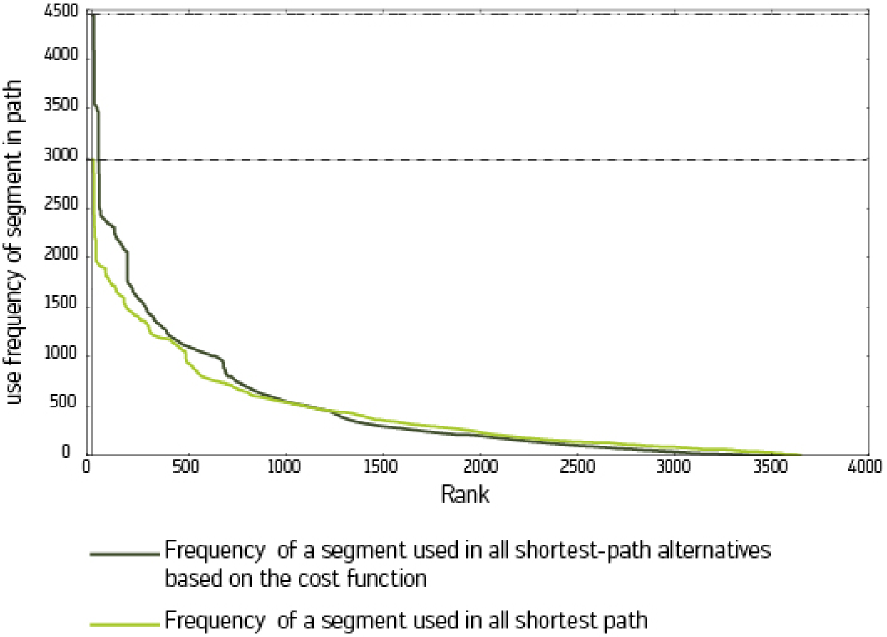


Figure 3.13: This image shows the ranksize plot of number of times a segment was used both in all shortest paths (light green line) and all shortest-path alternatives (dark green line) in the region of interest. This plot shows that our proposed shortest-path alternative successfully creates salient corridors for battlefield exploration (i.e. same ranks receive a higher number of times used).

objectives of a cultural heritage tourist. Real-life routing problems are most often not a single-objective problem. In this context, Jozefowiez et al. (2008) reviewed several multi-objective optimization techniques to reach a Pareto optimality between a set of observed objectives. Similarly, Tarapata (2007) proposed a number of solutions for multi-objective shortest-path problems, such as mathematical optimization. For example, Maervoet et al. (2013) proposed a heuristic to incorporate both cost minimization and POIs of some given types to calculate an optimal closed path for an outdoor tour. While our approach incorporates a more holistic view on the set of cultural heritage locations in our network, these techniques describe a combinatorial optimization of a set objectives in a network. Although both techniques have their merits, a more in depth comparison, evaluation and integration of both approaches will be imperative in future work.

From a cyclist's perspective, the increase in length between the shortest path

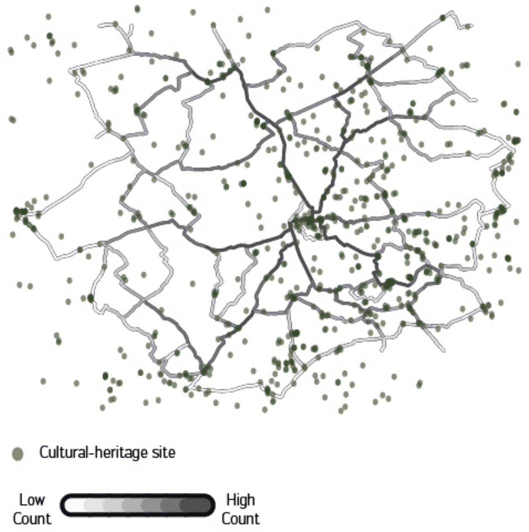


Figure 3.14: This image illustrates the cycling network in the region of interest shaded by number of times (count) used in all shortest-path alternatives. As a reference, a dot-density map of each site for battlefield exploration is added to the map.

and the shortest-path alternative seems acceptable to achieve a better experience during a leisure activity. However, we lack the data to do a formal evaluation of the perceived difference. User satisfaction will be of vital importance in this evaluation. While an assessment of this user satisfaction is considered as future work, the remainder of this section addresses some of the possible research tracks to source the necessary information. The most straightforward one is the collection of stated preferences. Huang et al. (2014) proposed the use of mobile applications to collect on-site knowledge. In a similar manner, a location-based questionnaire could allow the collection of unique information about our path suggestion (Zhao and Han, 2016). In this context, design, user experience and conduciveness of the user interface will be important drivers for success. Nielsen (2006) noted that contribution should be made a side effect of the experience. Another research track could be the analysis of human-mobility patterns during cultural heritage activities. Prato (2009) stated that aforementioned route choice modeling studies are essential to understand needs and trade-offs in path suggestions. While these analyses are well documented in literature, the data collection prior to this analysis is often problematic. Future work will have to address the collection and composition of this data set to allow more in depth evaluation of our proposed approach.

3.4 Conclusion and main contributions

The results in this use case proved that the geographical context of POI can be used beyond well-known map exploration of location-aware information. Given enough POI, we are able to create ancillary data to improve and personalize navigation services. We mainly focused on (i) the automatic collection and multimodal enrichment of thematic cultural heritage POIs and (ii) how an attraction-accessibility measure can be used to aggregate these enriched POIs, making new ancillary data readily available in navigation services. We used this use case to introduce two more specific research questions linked to overarching second research question in this dissertation.

The devised workflow gave a detailed overview of how the current state-of-the-art in geographic enrichment can be used in a topical/thematic context (**RQ1**). While the majority of the proposed steps are generic and readily applicable, thematic nuances should always be taken into account. For example, the impact of temporal changes in place names and their geographical space can be significant in a cultural-heritage context. Next, we focused on one specific type of contextual analysis which can be used to model the potential interaction of location-based information on an individual along a network, namely attraction-accessibility measures (**RQ2**). A moving-segment analysis was used to calculate this measure on basic spatial units of the network. This approach proved to be comprehensible, elegant and powerful to create a user-centered model. The use of network space to model potential interaction proved its performance. At the same time, individual weighing of every POI makes it a highly generic approach. In the current design we used a quality index of each POI similar to the introduced quality estimation. In future work, more intricate weighting of individual POI can easily be added, which makes it a flexible and extensible tool.

The proposed technological building blocks will facilitate cultural heritage exploitation in tourism in various ways. Applications ranging from spatio-temporal exploration to thematic routing will benefit from our collection, enrichment and routing tools. Furthermore, due to its generic architecture, the proposed framework can easily be used in other thematic routing and exploration applications too, making it a widely applicable approach. Our approach is not limited to recreational exploration of cultural heritage as such. Future work will mainly focus on the overall evaluation of the proposed methodology, more specifically, to investigate its conformance to user expectations (which has not been tested yet) and to evaluate the impact of the type of road network and POI coverage.

References

Aart, C., Wielinga, B., and Hage, W. R. (2010). Mobile cultural heritage guide: Location-aware semantic search. In Cimiano, P. and Pinto, H. S., editors, *Proceedings of Knowledge Engineering and Management by the Masses: 17th International Conference, EKAW 2010, Lisbon, Portugal, October 11-15, 2010*, pages 257–271, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alivand, M., Hochmair, H., and Srinivasan, S. (2015). Analyzing how travelers choose scenic routes using route choice models. *Computers, Environment and Urban Systems*, 50:41–52.

Bakillah, M., Lauer, J., Liang, S., Zipf, A., Jokar Arsanjani, J., Loos, L., and Mobasheri, A. (2014). Exploiting big vgi to improve routing and navigation services. In Karimi, H. A., editor, *Big Data Techniques and Technologies in Geoinformatics*, pages 177–192. CRC Press, Boca Raton, Florida.

Bar, D., Zesch, T., and Gurevych, I. (2012). Text reuse detection using a composition of text similarity measures. In *proceedings of 24th International Conference on Computational Linguistics December 8th -15th December 2012, Mumbai, India*, pages 167–184.

Becker, C. and Bizer, C. (2009). Exploring the geospatial semantic web with dbpedia mobile. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(4):278 – 286.

Bell, S. and Bala, K. (2015). Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics*, 34(4):98:1–98:10.

Bellini, E. and Nesi, P. (2013). Metadata quality assessment tool for open access cultural heritage institutional repositories. In Nesi, P. and Santucci, R., editors, *Information Technologies for Performing Arts, Media Access, and Entertainment: Second International Conference, ECLAP 2013, Porto, Portugal, April 8-10, 2013, Revised Selected Papers*, pages 90–103, Berlin, Heidelberg. Springer Berlin Heidelberg.

Berman, M. L. (2008). Georeferencing historical placenames and tracking changes over time. In *CGA Conference: Challenges of and solutions for assigning geographic location to digital information - a cross-disciplinary problem*, pages 1–9.

Biagioni, J. and Krumm, J. (2013). Days of our lives: Assessing day similarity from location traces. In Carberry, S., Weibelzahl, S., Micarelli, A., and Semeraro, G., editors, *proceedings of User Modeling, Adaptation, and Personalization: 21th International Conference, UMAP 2013, Rome, Italy, June 10-14, 2013*, pages 89–101, Berlin, Heidelberg. Springer Berlin

Heidelberg.

Blank, D. and Henrich, A. (2015). Geocoding place names from historic route descriptions. In *Proceedings of the 9th Workshop on Geographic Information Retrieval*, GIR '15, pages 9:1–9:2, New York, NY, USA. ACM.

Bouros, P., Ge, S., and Mamoulis, N. (2012). Spatio-textual similarity joins. *Proc. VLDB Endow.*, 6(1):1–12.

Bruce, T. R. and Hillmann, D. (2004). The continuum of metadata quality: defining, expressing, exploiting. In Hillman, D. and Westbroooks, E. L., editors, *Metadata in practice*, Chicago. American Library Association.

Cartwright, W., Peterson, M. P., and Gartner, G. (2007). *Multimedia Cartography*. Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg.

Economou, M. (2015). Heritage in the digital age. In Logan, W., Craith, M. N., and Kockel, U., editors, *A Companion to Heritage Studies*, pages 215–228, Chichester, West Sussex. John Wiley & Sons Inc.

Eppstein, D. (1994). Finding the k shortest paths. In *Proceedings of the 35th Anual Symposium on Foundations of Computer Science*, pages 154–165. IEEE.

Fang, C. and Torresani, L. (2012). Measuring image distances via embedding in a semantic manifold. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C., editors, *Proceedings of Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012*, pages 402–415, Berlin, Heidelberg. Springer Berlin Heidelberg.

Gao, M., Liu, K., and Wu, Z. (2010). Personalisation in web computing and informatics: Theories, techniques, applications, and future research. *Information Systems Frontiers*, 12(5):607–629.

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., and Ball, J. (2010). Use of the edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1925):3875–3889.

Hibberd, R. and Owens, J. B. (2015). Before highway maps: Creating a digital research infrastructure based on sixteenth-century iberian places and roads. *Bulletin for Spanish and Portuguese Historical Studies*, 40(1):19–43.

Hochmair, H. H. and Navratil, G. (2008). Computation of scenic routes in street networks. In Car, A., Griesebner, G., and Strobl, J., editors, *Geospatial Crossroads@ GI_Forum'08: Proceedings of the Geoinformatics Forum Salzburg*, pages 124–133. Wichmann Verlag, Heidelberg.

Huang, H., Klettner, S., Schmidt, M., Gartner, G., Leitinger, S., Wagner, A., and Steinmann, R. (2014). Affectroute – considering people's affective responses to environments for enhancing route-planning services.

International Journal of Geographical Information Science, 28(12):2456–2473.

Hughes, M., Jones, G. J. F., and O'Connor, N. E. (2012). A study into annotation ranking metrics in community contributed image corpora. In Nürnberger, A., Stober, S., Larsen, B., and Detyniecki, M., editors, *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation: 10th International Workshop, AMR 2012, Copenhagen, Denmark, October 24-25, 2012, Revised Selected Papers*, pages 147–162, Cham. Springer International Publishing.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 675–678.

Jozefowiez, N., Semet, F., and Talbi, E.-G. (2008). Multi-objective vehicle routing problems. *European Journal of Operational Research*, 189(2):293 – 309.

Karpathy, A. and Fei-Fei, L. (2017). Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676.

Kauppinen, T., Paakkarinen, P., Mäkela, E., Kuittinen, H., Väätäinen, J., and Hyvönen, E. (2011). Geospatio-temporal semantic web for cultural heritage. In Lytras, M., Ordóñez de Pablos, P., Damiani, E., and Diaz, L., editors, *Digital Culture and E-Tourism: Technologies, Applications and Management Approaches*, pages 48–64, Hershey, Pennsylvania,USA. IGI Global.

Luo, J., Joshi, D., Yu, J., and Gallagher, A. (2011). Geotagging in multimedia and computer vision—a survey. *Multimedia Tools and Applications*, 51(1):187–211.

Luxen, D. and Vetter, C. (2011). Real-time routing with openstreetmap data. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 513–516. ACM, New York.

Maervoet, J., Brackman, P., Verbeeck, K., De Causmaecker, P., and Vanden Berghe, G. (2013). Tour suggestion for outdoor activities. In Liang, S. H. L., Wang, X., and Claramunt, C., editors, *proceedings of Web and Wireless Geographical Information Systems: 12th International Symposium, W2GIS 2013, Banff, AB, Canada, April 4-5, 2013*, pages 54–63, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mikheev, A., Moens, M., and Grover, C. (1999). Named entity recognition without gazetteers. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pages 1–8, Stroudsburg, PA, USA. Association for Computational

Linguistics.

Miller, H. J. (1999). Measuring space-time accessibility benefits within transportation networks: Basic theory and computational procedures. *Geographical Analysis*, 31(2):187–212.

Moen, W. E., Stewart, E. L., and McClure, C. R. (1998). Assessing metadata quality: findings and methodological considerations from an evaluation of the us government information locator service (gils). In *Proceedings of IEEE International Forum on Research and Technology Advances in Digital Libraries. ADL 98*, pages 246–255, Los Alamitos, California. IEEE Computer Society.

Mooney, P. and Winstanley, A. (2006). An evolutionary algorithm for multicriteria path optimization problems. *International Journal of Geographical Information Science*, 20(4):401–423.

Nagy, K. (2011). Methodology of heritage-based tourism product development - thematic routes as new and special possibilities. *Doktorandusszok Fóruma, University of Miskolc*, 1(1):104–110.

Nagy, K. (2012). Heritage tourism, thematic routes and possibilities for innovation. *Theory Methodology Practice (TMP) - Faculty of Economics, University of Miskolc*, 8(1):46–53.

Nielsen, J. (2006). The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities.

Ochoa, X. and Duval, E. (2006). Quality metrics for learning object metadata. In Pearson, E. and Bohman, P., editors, *proceedings of ED-MEDIA 2006–World Conference on Educational Multimedia, Hypermedia & Telecommunications*, pages 1004–1011, Orlando, FL USA. Association for the Advancement of Computing in Education (AACE).

Ochoa, X. and Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, 10(2):67–91.

Okabe, A. and Sugihara, K. (2012). *Spatial Analysis Along Networks: Statistical and Computational Methods*. Statistics in Practice. Wiley, Chichester, West Sussex.

Oksanen, J., Bergman, C., Sainio, J., and Westerholm, J. (2015). Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. *Journal of Transport Geography*, 48:135–144.

Paulussen, H., Capdevilla, F., Debevere, P., Perez, M., Vanbrabant, M., De Neve, W., and De Wannemacker, S. (2014). Building an nlp pipeline within a digital publishing workflow. *Computational Linguistics in the Netherlands Journal*, 4(december):71–84.

Pigram, J. J. J. and Jenkins, J. M. (1999). *Outdoor recreation management*, volume 5. Psychology Press, London.

Pippig, K., Burghardt, D., and Prechtel, N. (2013). Semantic similarity analysis of user-generated content for theme-based route planning. *Journal of Location Based Services*, 7(4):223–245.

Popescu, A., Grefenstette, G., and Moëllic, P.-A. (2009). Mining tourist information from user-supplied collections. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1713–1716. ACM, New York.

Prato, C. G. (2009). Route choice modeling: past, present and future research directions. *Journal of Choice Modelling*, 2(1):65–100.

Quercia, D., Schifanella, R., and Aiello, L. M. (2014). The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 116–125, New York, NY, USA. ACM.

Rao, J., Lin, J., and Samet, H. (2014). Partitioning strategies for spatio-textual similarity join. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, BigSpatial '14, pages 40–49, New York, NY, USA. ACM.

Rattenbury, T., Good, N., and Naaman, M. (2007). Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 103–110. ACM, New York.

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. (2010). Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks*, 6(2):1–27.

Sanders, P. and Schultes, D. (2007). Engineering fast route planning algorithms. In Demetrescu, C., editor, *proceedings of Experimental Algorithms: 6th International Workshop, WEA 2007, Rome, Italy, June 6-8, 2007*, pages 23–36, Berlin, Heidelberg. Springer Berlin Heidelberg.

Schroth, G., Huitl, R., Chen, D., Abu-Alqumsan, M., Al-Nuaimi, A., and Steinbach, E. (2011). Mobile visual location recognition. *IEEE Signal Processing Magazine*, 28(4):77–89.

She, B., Zhu, X., Ye, X., Guo, W., Su, K., and Lee, J. (2015). Weighted network voronoi diagrams for local spatial analysis. *Computers, Environment and Urban Systems*, 52:70 – 80.

Silva, M. J., Martins, B., Chaves, M., Afonso, A. P., and Cardoso, N. (2006). Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, 30(4):378 – 399.

Sonawane, S. S. and Kulkarni, P. A. (2014). Graph based representation and analysis of text document: A survey of techniques. *International Journal of Computer Applications*, 96(19):1–8.

Souffriau, W. and Vansteenwegen, P. (2010). Tourist trip planning functionalities: State–of–the–art and future. In Daniel, F. and Facca, F. M., editors, *Current Trends in Web Engineering: 10th International Conference on Web Engineering ICWE 2010 Workshops, Vienna, Austria, July 2010, Revised Selected Papers*, pages 474–485, Berlin, Heidelberg. Springer Berlin Heidelberg.

Srivastava, M., Abdelzaher, T., and Szymanski, B. (2012). Human-centric sensing. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 370(1958):176–97.

Steenberghen, T., Aerts, K., and Thomas, I. (2010). Spatial clustering of events on a network. *Journal of Transport Geography*, 18(3):411 – 418.

Tarapata, Z. (2007). Selected multicriteria shortest path problems: An analysis of complexity, models and adaptation of standard algorithms. *Int. J. Appl. Math. Comput. Sci.*, 17(2):269–287.

van den Bosch, A., Busser, B., Canisius, S., and Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for dutch. In Eynde, F. V., Dirix, P., Schuurman, I., and Vandeghinste, V., editors, *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting, Leuven, Belgium*, pages 99–114.

Verstockt, S., Slavkovikj, V., De Potter, P., Vandersmissen, B., Slowack, J., and Van de Walle, R. (2013a). Automatic geo-mashup generation of outdoor activities. In *Proceedings of International Conference on Advances in Mobile Computing & Multimedia*, MoMM '13, pages 100:100–100:103, New York, NY, USA. ACM.

Verstockt, S., Slavkovikj, V., Potter, P. D., Slowack, J., and de Walle, R. V. (2013b). Multi-modal bike sensing for automatic geo-annotation geo-annotation of road/terrain type by participatory bike-sensing. In *Proceedings of the International Conference on Signal Processing and Multimedia Applications (SIGMAP), 2013*, pages 39–49. IEEE.

Winter, C. (2011). Battlefield visitor motivations: explorations in the great war town of ieper, belgium. *International Journal of Tourism Research*, 13(2):164–176.

Zhao, Y. and Han, Q. (2016). Spatial crowdsourcing: current state and future directions. *IEEE Communications Magazine*, 54(7):102–107.

Zheng, Y., Fen, X., Xie, X., Peng, S., and Fu, J. (2010). Detecting nearly duplicated records in location datasets. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 137–143, New York, NY, USA. ACM.

Zheng, Y., Zhang, L., Xie, X., and Ma, W.-Y. (2009). Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 791–800. ACM, New York.

Zheng, Y.-T., Yan, S., Zha, Z.-J., Li, Y., Zhou, X., Chua, T.-S., and Jain, R.

(2013). Gpsview: A scenic driving route planner. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 9(1):3:1–3:18.

Zheng, Y.-T., Zha, Z.-J., and Chua, T.-S. (2011). Research and applications on georeferenced multimedia: a survey. *Multimedia Tools and Applications*, 51(1):77–98.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 487–495.

4

Knowledge sharing in a route-sharing community to improve navigation services

"Help Me Help You"
Jerry Maguire, Jerry Maguire

The aim of this chapter is to gain a better understanding in how active user interaction can be used within route-sharing platforms to improve their services. In this context, route-sharing communities have a great potential to harvest local knowledge and expertise. These communities have engaged users that create a vast amount of routes and POI. This user-generated content is strongly interwoven with these indiviuals' cognitive map. This third and final use case evaluates a new low-level strategy to query these individuals' cognitive map beyond their user-generated content to infer road-network updates. We provide both theoretical grounding of this strategy and its practical implementation in the online route-sharing community of RouteYou. A web-based feedback tool is designed and embedded in this online community. More specifically, we combine (i) road-network attributes and (ii) specific route information to create personalized tasks where user feedback is required. To evaluate and underpin the validity, technical feasibility and effectiveness of this new strategy, we present experimental results and study the contribution patterns of 325 active contributors. The results prove the feasibility of the proposed strategy. Several future research tracks are discerned to further analyze and improve the results and the user experience of our experiment.

The work presented in this chapter is based on a working paper prior to submission: Kevin Baker, Kristien Ooms, Steven Verstockt, Pascal Brackman, Philippe De Maeyer and Rik Van de Walle. (2017). Knowledge Sharing in a Route-sharing Community to Improve Navigation Services.

4.1 Introduction

A growing number of individuals no longer rely on their cognitive map for wayfinding. Rather, they follow instructions from online and mobile navigation services to find their way (Vanclooster et al., 2016). As the number of guidance services is growing, so has the demand for ad-hoc created routes for specific sports and leisure activities, such as cycling or hiking. While these services solve a lot of knowledge-based wayfinding problems on unfamiliar routes (Golledge, 1999), flawed representation of road infrastructure in the map data impair guidance performance (Burns, 1998). Because these services become more personalized, user-centered and interwoven in daily life (Gao et al., 2010), many of these mistakes are still being found by individual users.

Giving users the means to share their local expertise, low-level error-reporting tools are often implemented in online maps and mobile services. Coleman et al. (2009) stated that these tools are still considered key components to get users involved in mapping projects. However, managing and solving these bugs quickly becomes a laborious and high-cost task. Previous research has already highlighted some of the difficulties with these tools. First and foremost, Bettenburg et al. (2008) found that similar bug-reporting tools often result in ill-defined or fuzzy information. They noted that a complete, clear and correct description is imperative to facilitating solving reports. Next, the tedious work of triaging reports still hampers bug solving and is an often-studied aspect of error-reporting tools in literature (e.g. Anvik et al., 2006; Hooimeijer and Weimer, 2007). Finally, Haklay (2016) also voiced that lowering the hurdles to participate in mapping projects is not always the key to success. Moreover, Nielsen (2006) stated that most users in online communities just do not contribute.

As a result, recent years have seen a paradigm shift toward data-driven approaches to infer quality assurance. Both knowledge-discovery and data-mining techniques are used to assure a correct, detailed and mode-specific annotation of digital road networks based on ancillary data sources harvested within crowds (e.g. routes recorded as GPS traces). Basiri et al. (2016) highlighted the value of these data sources to detect anomalies and abnormalities based on rules and patterns seen in clusters of routes. As a result, errors and bugs can be detected with a certain degree of trust. Despite this new self-healing map paradigm, however, specific local knowledge still resides with the individuals within these IT-mediated crowds. Sourcing this unique knowledge is of vital importance to improve new data-driven knowledge inference and, hence, the aforementioned navigation services. As such, the overarching theme of this use case is the possible synergy between both error-reporting tools and data-driven knowledge discovery. This use case is devised around two specific questions linked to the third research question in this dissertation:

- RQ1: How can **routes** created on route-sharing platforms be used by a **task-recommendation mechanism** to foster **knowledge sharing**?
- RQ2: Is active user interaction outperformed by current data-driven approaches (cfr. Chapter 2)?

First, this use case presents a detailed description of a new low-level strategy to collect knowledge from online route-sharing communities. While this use case focuses on routes shared on the RouteYou [1] platform, our approach is applicable in other route-sharing communities where leisure activities are shared and classified according to type of activity, such as road cycling or mountainbiking. More specifically, we study how error-reporting tasks can be restructured to increase awareness of possible errors in the underlying map data and encourage knowledge sharing by linking it to a user-specific context, that is, a user's route.

This approach in doing so addresses one of the major challenges in current community-driven mapping project See et al. (2016): "how can citizens be encouraged to map an area that has already been mapped in the last few years or be more actively engaged in change detection mapping? (p 18)". We combine the route's information and meta-information shared on this platform with OpenStreetMap (OSM) road network information to create and present context-aware and location-based human information tasks to detect errors, omissions, or other inaccuracies in map data. Thus, we convert the user's actions of creating, classifying and annotating an error report to an easier and user-specific task.

In order to evaluate this new strategy, we designed a prototype of a web-based feedback tool and embedded this tool in RouteYou's web platform. We chose to focus our attention on errors of commission in road condition tagging in OSM, based on routes with a type tag cycling in this route-sharing community (e.g. road cyclist, leisure cyclist). This was done for two reasons: (i) OSM has a very intricate tagging structure for road condition and (ii) previous research (Bergström and Magnusson, 2003; Damant-Sirois et al., 2014) showed that road condition is a key factor in cyclist typology and, hence, very suitable for the proposed methodology to detect conflicts.

Haklay (2016) also noted that for a crowd-sourcing project to become useful, it is imperative to understand the contribution patterns within this project.

[1] http://www.routeyou.com

These patterns have value to understand the shared data and streamline efforts to increase both user experience and user engagement. As such, a second objective is to gather insights on the contribution patterns during the rollout. For this purpose, we present both a general overview and an analytical approach to these contribution patterns addressing the user lifetime.

Finally, if successful, the above-described harvested verification of errors has the potential to provide valuable information for knowledge-based pattern- and rule-mining techniques. Hence, the third objective is to provide an exploratory analysis of the rules and patterns seen in clusters of routes linked to these location-aware errors in map data. The shared routes within RouteYou allow us to compare both user and type diversity in these clusters as well as the size, that is the number of routes, of these clusters against the harvested verification. In doing so, we aim at opening the door to new ways of understanding the synergy between both error-reporting tools and data-driven knowledge discovery and, hence, foster future advances of the presented low-level strategy to improve navigation services.

This use case is structured as follows. Section 4.2 addresses related work shaping the frame of reference of this use case. Subsequently, we address the experiment design in Section 4.3. We elaborate on the active contributors, the back- and front-end design of the web-based experiment and the procedure to create and answer tasks. In Section 4.4, we present the results linked to contribution patterns in the experiment. Section 4.5 presents the harvested verification during this experiment design. Finally, the exploratory data-driven analysis in described in Section 4.6. We conclude with a discussion of the presented approach and future work.

4.2 Context

4.2.1 Knowledge sharing in mapping projects

In the last decade, community-driven mapping projects such as OSM have been receiving a burgeoning interest. Similar to low-level error-reporting tools, these projects give individuals the means to share their local knowledge and expertise. The merits of these tools are clear: accessibility and transparency in these online user communities empowers these individuals and draws them together, resulting in a virtuous circle. Not surprisingly, however, these communities are not a generic task force. To create an awareness of possible errors and facilitate contribution, this mapping community also creates a multitude of tools, such as bug-reporting, error-detection, visualization, monitoring or assistant tools[2]. However, participation inequality, in time, space and quality, still shapes this and similar approaches (Haklay, 2016). Consequently, there has been a growing research trend towards understanding the on-boarding process in online mapping communities and modeling the participation inequality and quality in taxonomies, ranging from novice to power users (e.g. Coleman et al., 2009; Flanagin and Metzger, 2008; Haklay, 2013; Neis and Zielstra, 2014; Neis and Zipf, 2012). However, Panciera et al. (2010) also noted that research in online communities is often hampered by what researchers cannot see. Motivating factors such as altruism, professional or personal interest or intellectual stimulation are imperative in information-sharing communities or distributed problem-solving and production models. Wang and Fesenmaier (2003) also proposed social-economic status and personality of the participant as important drivers in contribution behavior. Furthermore, game-design elements and rules are a popular way to increase participation and encourage contributors (Matyas et al., 2011; Vyron and Schlieder, 2014; Yanenko and Schlieder, 2014). For example, Martella et al. (2015) reviewed several gamification approaches to gather volunteered geographic information, such as maproulette.org, using several game-design elements to improve OSM. However, Ferrara (2013) also pointed out that integrating game elements, such as leaderboards, rankings or badges, is not a guarantee for engagement. Wang and Fesenmaier (2003) and Deterding et al. (2011) stated that the quality of the user experience and the conduciveness of the user interface are maybe more important drivers for success.

Furthermore, not all users are contributors. As the quality of these online maps improves, *map-seekers* start using these maps for a number of activities. For example, we see a growing use of OSM for navigation services for leisure activities. This clearly shows the unique position and popularity of OSM within these user groups (Kessler, 2011). However, these users often lack intrinsic motivators to become an OSM contributor. To lower the overhead, error-reporting tools such as OSM Notes allow users to report and discuss location-based flaws or omissions in the map. Similarly, many leisure activity communities using OSM incorporate this or similar reporting tools (e.g. RouteYou, Strava, Bikemap or Runkeeper), but many possibilities to collect new information from these communities remain, to our knowledge, unexplored. Zhao and Han (2016) noted that this type of social networking could provide a valuable context to further crowdsourcing. Instead of providing specific tasks to a generic crowd-sourcing task force, such as Amazone's Mechanical Turk workers[3], it is more reasonable to assign tasks concerning

[2] http://wiki.openstreetmap.org/wiki/Quality_assurance

[3] https://www.mturk.com/mturk/welcome

cycling to a social group tagged "cyclists". Doan et al. (2011) argued that how to recruit contributors is one of the biggest challenges in crowd-sourcing approaches. For a more in depth overview of techniques of personalized task-recommendation mechanism we refer to the literature review of Geiger and Schader (2014). To our knowledge, the use of embedded procedures to harvest local knowledge (i.e. a *piggyback* approach (Doan et al., 2011)) in a route-sharing community has not yet been investigated in academic literature. As such, our use case has the potential of providing valuable insights in the opportunities, challenges and limitation in using a route-sharing community in a task-recommendation approach.

4.2.2 User contribution and activity

A growing body of research focuses on Human Computer Interaction (HCI), user-centered design and usability to improve the sharing behavior in online communities (Preece, 2000). However, we already highlighted that, despite these efforts, participation inequality is inherent in these communities. Participation inequality is know to have a theoretical distribution of 90 % consumption (lurkers) and 10 % contributors in online communities (Haklay, 2016). While many efforts have been made to achieve a more equitable lurker distribution, this inequality is believed to be an inherent characteristic of sharing behavior (Nielsen, 2006). Secondly, research in online communities and webgraphs (Broder et al., 2000; Kumar et al., 1999) noted that power law behavior is also often seen in web-based platforms. This behavior model states that a small fraction of contributors create the majority of contributions and a long tail of non-loyal contributors provide the additional contributions. In addition, analytical work on contribution patterns and user lifetime in knowledge-sharing oriented online communities, such as Guo et al. (2009), also stated that contribution patterns in online communities show strong daily and weekly patterns. Research including, but not limited to, the work of Guo et al. (2009); Oentaryo et al. (2012) also highlight the negative impact a user's communal lifetime can have on keeping users active in these knowledge-sharing communities. While it is not within the scope of this use case to provide an in-depth analysis of intrinsic and extrinsic factors influencing the contribution patterns in our experiment, we want to provide a general overview of these contribution patterns to guide future work in addressing possible factors hampering contribution.

4.2.3 Context-driven knowledge discovery

As introduced above, a growing body of research also uses more data-driven approaches to infer knowledge from routes. Applied techniques focus on detecting spatial, temporal or spatio-temporal correlations or clusters in this data to infer insights (e.g. Biagioni and Eriksson, 2012; Kasemsuppakorn and Karimi, 2013; Liu et al., 2012; Schroedl et al., 2004). However, the value of meta-information linked to these traces is, in our opinion, underestimated in current literature. Many route-sharing platforms provide tagging systems to link a route to a type, theme, group or characteristic. In doing so, users link valuable information which can be of interest for contextual analysis and information retrieval. The presented approach explores the potential of this type tagging to infer novel information. Besides the efforts to create personalized task-recommendation mechanism in this use case, we also want to address the possible value of this meta-information in future data-driven approaches. Basiri et al. (2016); Zheng (2015) noted that aforementioned knowledge-discovery techniques mainly focus on detecting travel modes, group movement patterns or unusual behavior detection. As such, we approach this knowledge discovery from a different angle moving away from purely data-driven to a more context-driven knowledge discovery.

4.3 Experiment design

4.3.1 Active contributors

Participants were sourced from the online route-sharing platform RouteYou. In total, 325 active contributors were included in the evaluation within this use case. Contributor acquisition was conditioned by a phased rollout and user's community involvement. The first phase was used to evaluate the experiment with highly-involved users in the RouteYou community and ran in the first 435 days of the experiment. As a proxy for a user's community involvement, we used the individual user score generated by the user reputation algorithm in RouteYou. Similar to other reputation algorithms in online communities, this algorithm uses the number of contributions made by a specific user and the contributions' quality and ratings. The resulting metric is scaled between 0 and 1, ranging from novice to expert users. This resulted in a limited group (86) of users with a low variance in the user score. The second phase (last 325 days of the experiment) was not conditioned by user score filtering and featured a broader rollout to all users in the community of RouteYou. The left-hand plot of Figure 4.1 presents the evolution of the variance in user score throughout the experiment. The right-hand plot presents the user scores in the second phase of the experiment. All participants took part in this study on a voluntary basis. Although no direct reward incentives were linked to task completion, all routes with user answers were prioritized in RouteYou's semi-automatic content rating system. As a result, a recognition and appreciation incentive was introduced in this study.

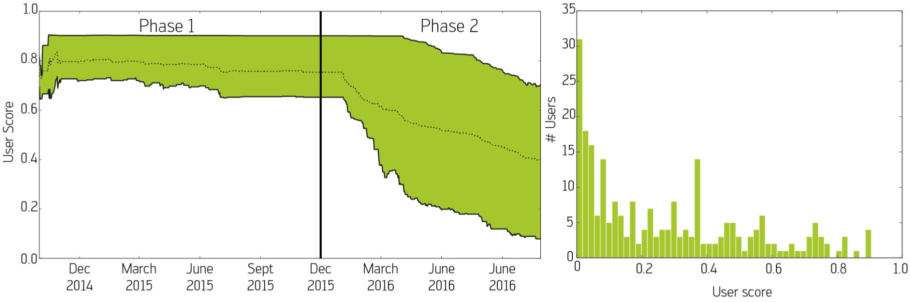


Figure 4.1: This plot summarizes the distribution of the participants' user score throughout the experiment. In the left-hand plot, the upper and lower black line represent the 90 and 10 percentile in the user score. The dashed line represents the average. The right-hand plot presents a histogram plot of the user scores in the second phase of the experiment.

4.3.2 Procedure

The objective here is to provide a detailed view on the proposed procedure to create and answer a task set linked to a user's route (see Figure 4.2). This procedure constitutes three main parts: (i) questions, (ii) route segmentation and (iii) linking questions to segments.

The experiment featured three main questions. The first question (Q_1) addressed the overall user experience of the route. The user is asked to rate his experience with the route on a scale from 1 (really bad experience) to 10 (great experience). The second question (Q_2) was used to ask the user whether a segment is accessible for the activity type of his route. A dichotomous answering scale (yes/no) was linked to this question. The third question (Q_3) requested feedback on the road condition of a segment. More specifically, we ask 'Is this segment paved?'. Keeping in mind the complex interaction between road condition and type of cyclist, we applied a nominal answering scale featuring four categories: (i) paved good condition; asphalt, concrete or tiles in good condition (pg), (ii) paved bad condition; paved but with lots of vibrational discomfort during cycling (pb), (iii) unpaved good condition; gravel or heavily compacted road surface (ug) and (iv) unpaved bad condition;

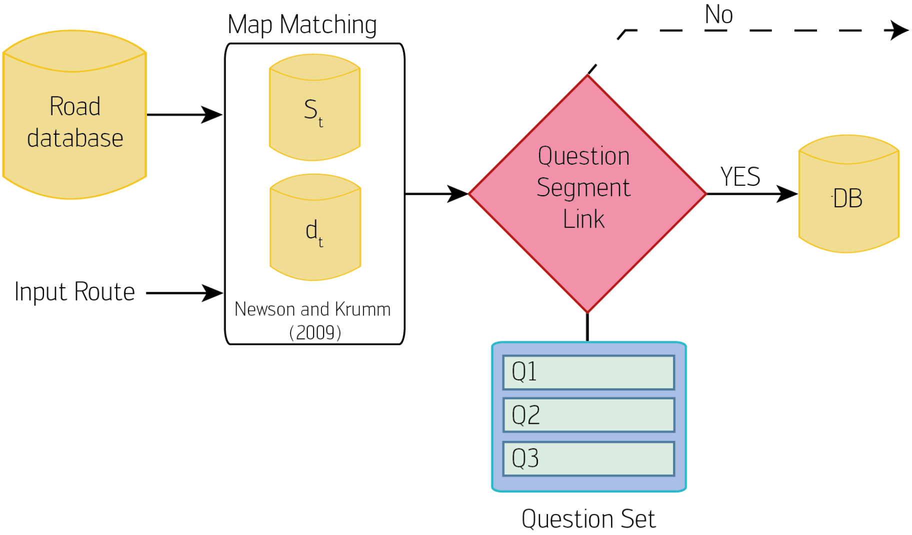


Figure 4.2: This flowchart presents the main steps in the procedure to garner local expertise from RouteYou users

soft materials with substantial rutting (ub). In the analysis of the results further on in this use case, the latter answer scale was mapped to binary responses, paved or unpaved, to avoid bias.

Trajectory segmentation is a well-studied procedure, dividing a route into comprehensible parts to reduce computational complexity and mine richer knowledge (Zheng, 2015). Stay and turning points or other semantic-change points (e.g. mode, long/short stay) are often used to split a trajectory in these parts of interest. In our approach, we use the underlying OSM road network as driver for a semantic segmentation. A segment is defined as a part of a route that matches a specific road in this road network. This step is necessary to unambiguously link a set of road-specific object attributes to parts of the route. This segmentation is illustrated in Figure 4.3 and is based on a map-matching approach. Map matching is a procedure that uses spatio-temporal input of a moving object (i.e. geographical longitude and latitude and time of recorded object location) and a road network to provide an enhanced positioning output (Quddus et al., 2007). While this approach has already proven its value in on-vehicle navigation systems to alleviate noise while determining which road a tracked object is traversing (Newson and Krumm, 2009), this procedure is equally valuable to mitigate noise while post-processing network-constrained sequences of locations as seen in routes. In doings so, we can redefine a segment as a sequence of locations matching a specific road object in the OSM road network. We applied the hidden Markov matching algorithm implemented in the Open Source Routing Machine project (Luxen and Vetter, 2011). This results in a set of route segments s_t of route t matched on the OSM road network. Every route segment has a segment length l and an OSM identifier[4]. We also generate a set of segments d_t containing the parts of a route which did not match with the road network. This can be seen as a difference between the OSM road network and the route (see Figure 4.4 on page 95). While we are aware of possible errors in the results of map-matching algorithms due to complex road lay-outs or dense urban networks (Quddus et al., 2007), the analysis of these errors is put beyond the scope of this use case.

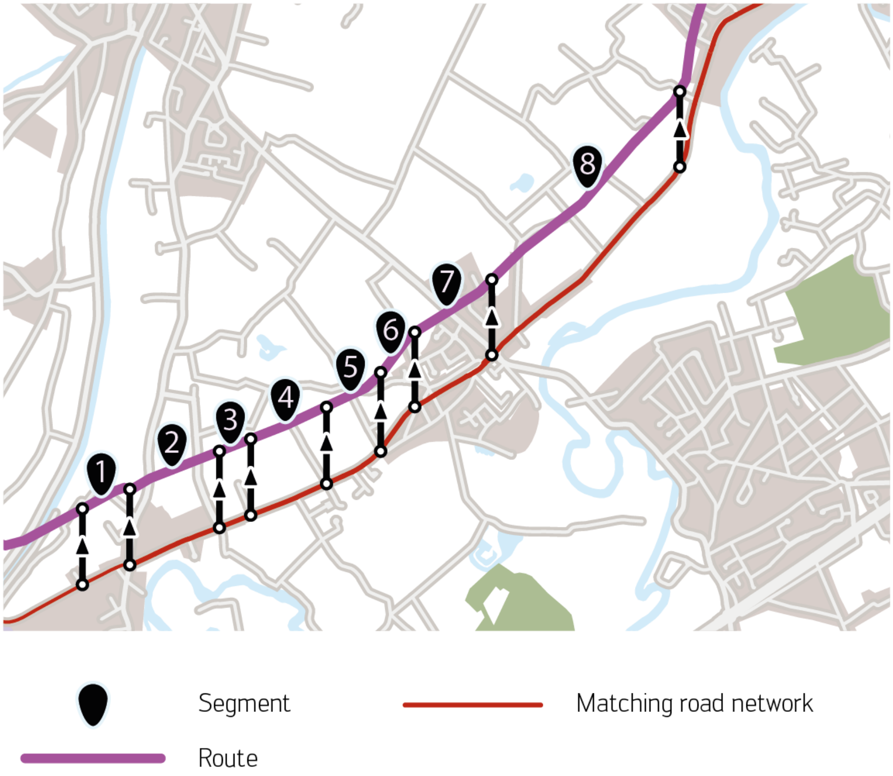


Figure 4.3: This image shows the segmentation procedure of a route. The detail shows the segmentation result of a part of a route (segment 1- 8). For visualization purposes the original route is given a vertical offset.

[4] Every object in OSM has a unique resource identifier, for example: http://www.openstreetmap.org/way/165289583

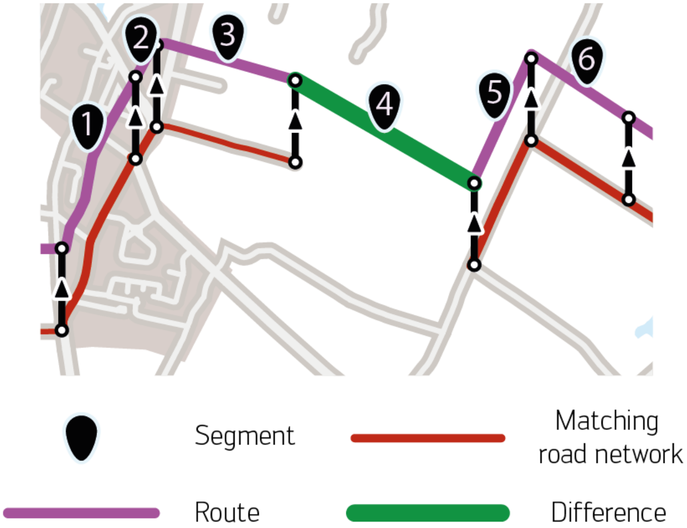


Figure 4.4: This image shows an example of a segment as a result of difference between a route and the road network. For visualization purposes the original route is given a vertical offset.

Finally, a decision tree links questions to segments. From a practical perspective, this step creates records in a relational database. Figure 4.5 on page 96 represents the entities and relationships in this database. A problem is described as a segment-question pair. In the remainder of this paragraph, we present the decision rules to link a problem to a task list. Without exception, Q_1 was linked to a task list. The next decision rule featured missing road geometry, linking Q_2 to a segment. This task was linked to a segment if it was part of the d_t set and was longer than a length threshold. A second decision rule was introduced to detect conflicts between the surface tag of an OSM Way and the activity type tag of a route. Similar to other route-sharing platforms, Routeyou has a hierarchical tagging system allowing users to link activity type or sub-type tags to a route. For example, a user can tag a route as a cycling activity, which is a general type tag. However, the user can also link a sub-type tag, such as mountainbike or road cycling, to a route if suited. Previous research in Chapter 2 has already highlighted the importance of road condition during these activity (sub-)types. Using this as a heuristic, we created links between segments and Q_3 where an OSM highway or surface tag[5] indicating an unpaved road condition or a paved but bad road condition

[5] a tag is the popular term of an attribute linked to an object within the OSM community. Tags are defined as key-value pairs. An non-exclusive overview of possible tags is listed on

co-occurred with the activity type tag cycling or sub-type tags road cycling or leisure cycling. Inherent to many tagging systems in web-based platforms, we do not have quality metrics of the tagging. However, these quality issues were placed beyond the scope of this article. Therefore, the tags were deemed correct and unambiguous.

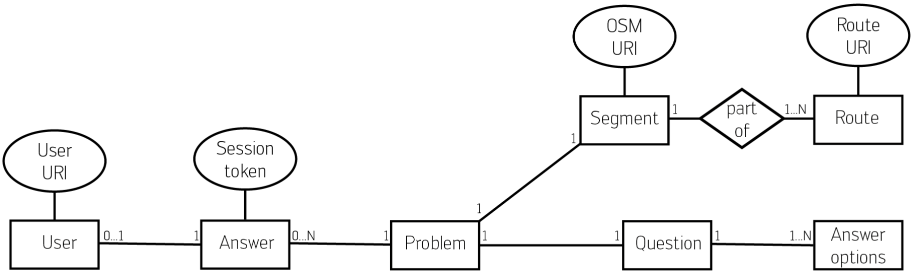


Figure 4.5: Entity Relationship Diagram describing how the different elements are related in the proposed database model.

While we also do not address the full extent of the expert-based mapping of OSM highway and surface tags to our road condition classes, previous research (e.g. Hochmair et al., 2015) has already shown the feasibility of similar approaches. The mapping uses the characteristics of individual tags as well as the co-occurence of certain tags on a road object in OSM to classify roads in the aforementioned four road condition classes. For example, a tag *highway:track*[6] has a high probability of being unpaved. If there are no additional tags specifying the quality of the track, we define a track as ’paved bad condition’. *highway:residential*, as a second example, has a high probability of being paved. Unless there are surface tags indicating otherwise, we define a residential road in OSM as ’paved good condition’. To give a general idea of the mapping of the surface tags, Tabel 4.1 on page 97 presents a selection of the most popular surface tags and how we mapped them to our four classes.

If a user submits an answer on a task, the answer is stored and linked to a problem. In addition, meta information of the answering session is stored. If the user was logged in during the answer session, the answers are linked to a user’s unique resource identifier (URI) and a session identifier of a unique IP address. A session can encompass answers on multiple task sets. Both user and session identifiers are imperative to facilitate and broaden the analysis of user contributions and their contribution patterns.

http://wiki.openstreetmap.org/wiki/Map_Features.

[6] http://wiki.openstreetmap.org/wiki/Tag:highway%3Dtrack

Surface tag	road condition class
asphalt	pg
unpaved	ug
paved	pg
ground	ub
gravel	ug
concrete	pg
dirt	ub
paving_stones	pb
grass	ub
compacted	ug
sand	ub
cobblestone	pb

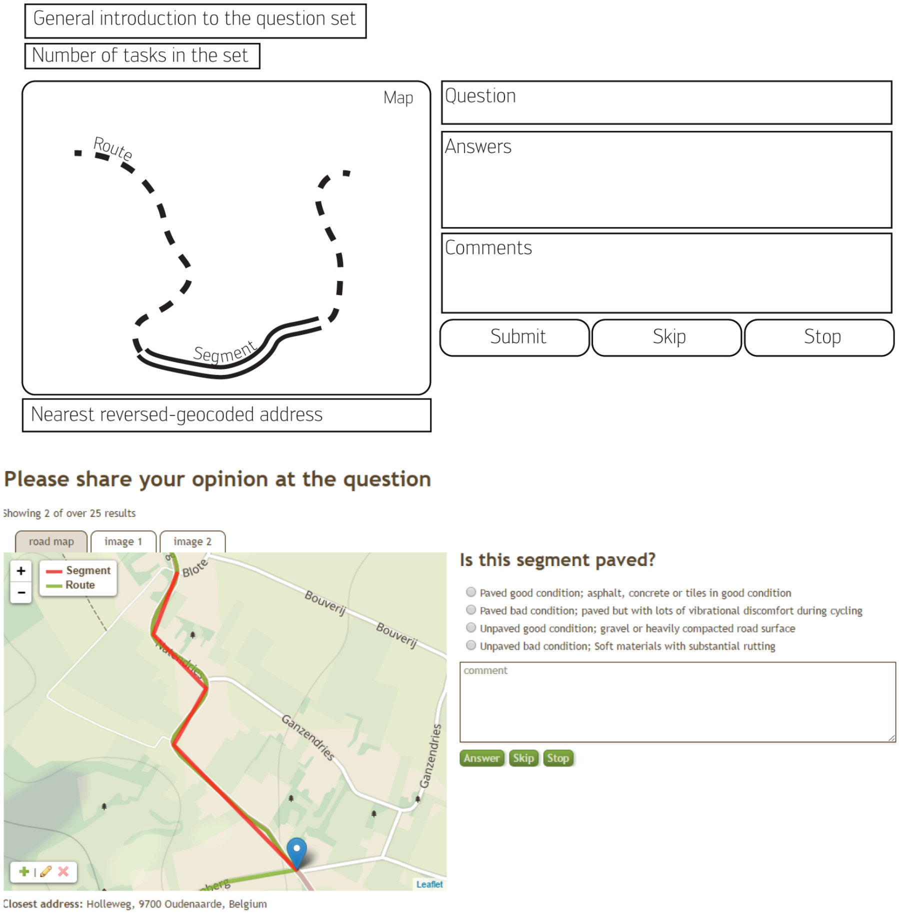
Table 4.1: This table lists an extract of the mapping between OSM surface tags and our four classes, (i) paved good condition (pg), (ii) paved bad condition (pb), (iii) unpaved good condition (ug) and (iv) unpaved bad condition (ub)

4.3.3 Apparatus and stimuli

As mentioned earlier, a minimal-complexity user interface was built around the task sets (see Figure 4.6 on page 98). A task set is linked to a route and features its problems. To provide maximal flexibility to the end users, we also introduced three action buttons: (i) 'Submit' saves the answer to the *answered* database table (see Figure 4.5), (ii) 'Skip' allows the user to skip the presented task and (iii) 'Stop' allows users to quit the current set. Furthermore, to increase the user experience and avoid duplicate answers from the same user, we checked if a user was already linked to an answer on a specific problem. If this was the case, the task was removed from the set.

As described in the previous section, the online experiment comprised two phases. While the design of the user interface remained the same in both phases, users were introduced to the experiment in a different manner. In the first phase users were invited to contribute in the experiment by direct contact. They were asked if they were willing to participate in the experiment and were given the possibility to exclude their activity on RouteYou from the online experiment. Next, they received a notification via email that a new task set was linked to their route. In the second phase, the email notification was replaced by the notification- and comment-system in the RouteYou platform. A comment was generated on every route with a task set, showing general information about the route and the link to a question set. Figure 4.7

on page 99 presents an actual example of a comment.



Please share your opinion at the question

Showing 2 of over 25 results

road map | image 1 | image 2

Is this segment paved?

○ Paved good condition; asphalt, concrete or tiles in good condition
○ Paved bad condition; paved but with lots of vibrational discomfort during cycling
○ Unpaved good condition; gravel or heavily compacted road surface
○ Unpaved bad condition; Soft materials with substantial rutting

comment

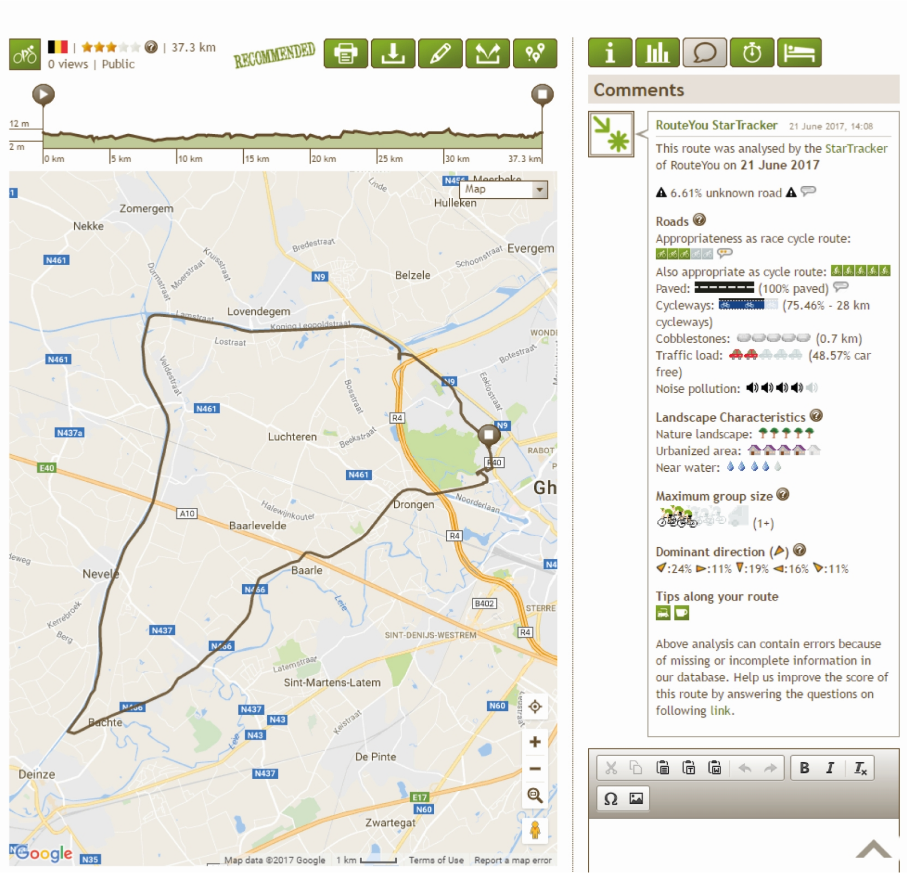
Answer | Skip | Stop

Closest address: Holleweg, 9700 Oudenaarde, Belgium

Figure 4.6: This image shows both a mock-up with a task's general structure and the appearance of the user interface in the online experiment.

4.4 Contribution patterns

As mentioned in Section 4.3.1, the experiment had 325 active contributors. These contributors submitted 4816 answers during 1605 sessions. 1339 answers on Q_1, 2091 on Q_2 and 1386 on Q_3. From all the answers on Q_3 only 115 problems received multiple answers. The number of answers per session has a median of 2. Figure 4.8 on page 100 presents the active contributors per week as well as the first-time contributors to the experiment in this



Comments

RouteYou StarTracker 21 June 2017, 14:08

This route was analysed by the StarTracker of RouteYou on 21 June 2017

⚠ 6.61% unknown road ⚠

Roads
Appropriateness as race cycle route:
Also appropriate as cycle route:
Paved: (100% paved)
Cycleways: (75.46% - 28 km cycleways)
Cobblestones: (0.7 km)
Traffic load: (48.57% car free)
Noise pollution:

Landscape Characteristics
Nature landscape:
Urbanized area:
Near water:

Maximum group size
(1+)

Dominant direction (▸)
◂:24% ▸:11% ▾:19% ◂:16% ▾:11%

Tips along your route

Above analysis can contain errors because of missing or incomplete information in our database. Help us improve the score of this route by answering the questions on following link.

Figure 4.7: An actual example of a comment as stimuli to start contributing in the online experiment. These comments were embedded in the comment- and notification-system within RouteYou.

week[7]. Visual analysis of this figure already highlights both clear global trends and horizontal and vertical variations. First, we see a clear start-up peak and a subsequent fall back. Second, the aforementioned phases in the online experiment are distinctly visible. As the second phase opened up the experiment to a broader group of potential contributors, we see a larger part of first-time contributors in the active contributors per week. Third, we also see periodic variation when contributors start (vertical variations) and stop (horizontal variations). For example, we see a clear horizontal variation in the first phase. For example, from December 2014 to March 2015 there is a noticeable lower contribution rate. Another observable horizontal trend is the decline in contributors as the first phase nears its end.

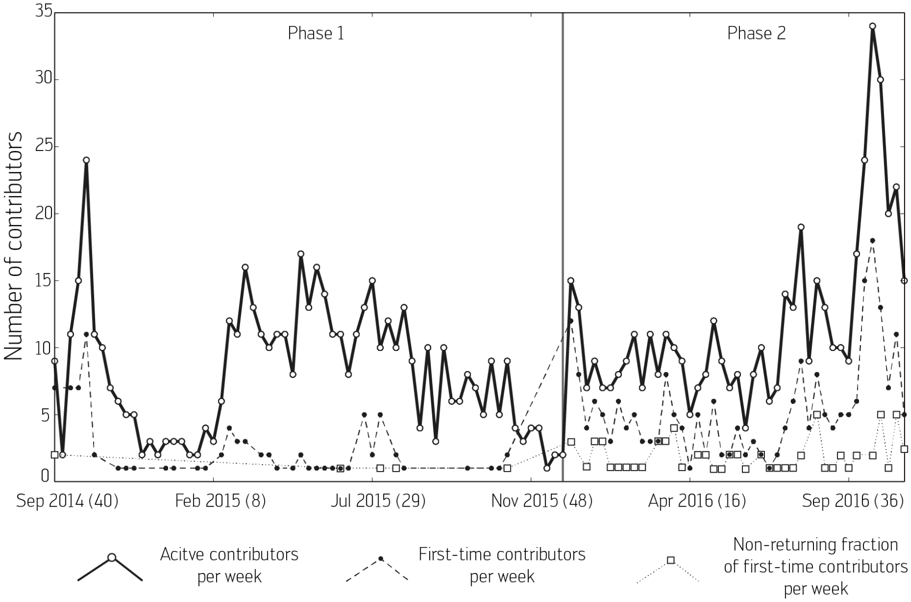


Figure 4.8: This plot presents the active contributors per week (white circle) during the run time of the online experiment. The black circles represent the first-time contributors to the experiment in the particular week and the squares represent the non-returning fraction of these first-time contributors.

Next, we report on additional measures of the contribution pattern. Conversion rate or acquisition rate can be defined as the fraction of users who answered at least one task linked to a route. As mentioned earlier 325 users answered, which is 1.65 % of the total users (19 725) who were linked to a task set. In contrast to the conversion rate, the returning or retention rate models

[7] An updated version of this plot at the moment of writing can be found in Appendix B

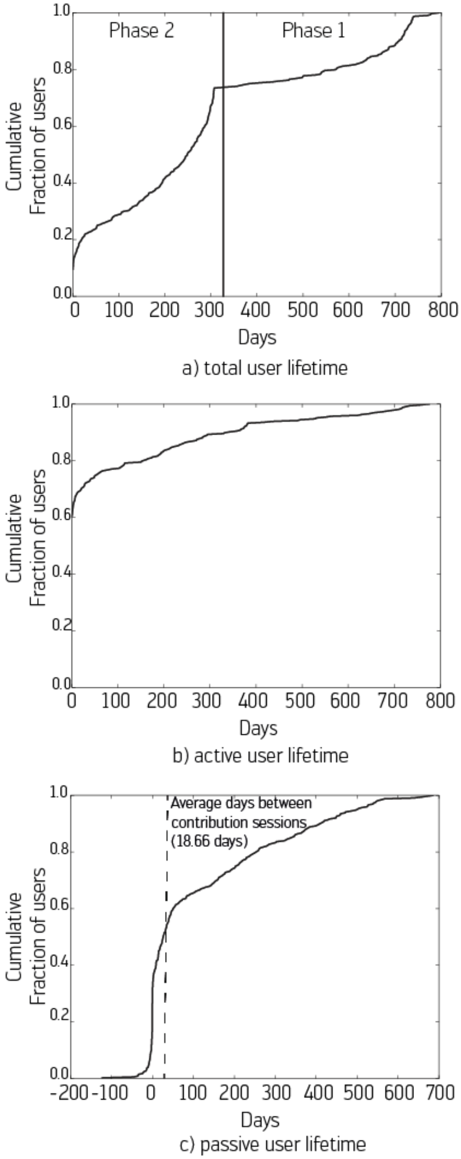
the users in the learning curve of the project (i.e. returning after their first-time contribution). This metric can be generated by counting the occurrence of a user URI in the answered table with a different session token. 39.23 % returned during the experiment. Furthermore, 10 % of the users started more than 11 answering session and the users in the top 1 % started more than 66 answering sessions. Furthermore, Figure 4.8 on page 100 shows that the majority of non-returning visitors contributed in the second phase of the experiment. These non-returning contributors have an average individual user score of 0.26. During the total run time of the experiment, we observe an average days between contributions of 18.66.

4.4.1 User lifetime

Because the experiment was embedded in an existing online community, we did not have user lifetime parameters such as profile creation date and last login information for our experiment. We replaced this by first and last task set linked to a user's route to model his lifetime in the experiment. We apply three metrics to model a specific user's lifetime within the experiment:

- Total User Lifetime (TUL) - defines the period (days) between the first and last task set linked to a user's route,
- Active User Lifetime (AUL) - defines the period (days) between the first and last answering session,
- Passive User Lifetime (PUL) - defines the period (days) between the last answering session and the last creation date of a task set linked to a user's route. The number of days is negative or zero when the last created task set was completed in the user's last session. In contrast, the number of days is non-negative when there are unanswered task sets linked to a user.

Figure 4.9 on page 102 shows the cumulative frequency plots of all active contributors for the above lifetime metrics. The plot at the top of Figure 4.9 clearly shows the impact of both phases on the total user lifetime. We see that less than 30 % of participants have a total lifetime of 325 days or more. The center plot in Figure 4.9 confirms the fact that many active users are non-recurring contributors. In addition, we also see that a fraction of the active contributors are long-term contributors. The bottom plot in Figure 4.9 also indicates that 29 of the contributors still have a high probability of returning if we use 'average days between contributions' (18.66 days) as a benchmark.



a) total user lifetime

b) active user lifetime

c) passive user lifetime

Figure 4.9: These cumulative frequency plots present the user lifetime metrics analyzed in this use case. All plots show these metrics for the active participants in the online experiment.

As the PUL of contributors further increases above this benchmarck, we expect a diminishing probability of returning.

To further our understanding of the contributor types in the experiment, we explore the relationship between our lifetime metrics and explanatory variables in the online experiment. For brevity and comprehensibility, we only present these results for the second phase of the experiment. While correlation can have limited explanatory value in online experiments due to heavy-tailed distributions or power-law behavior, they can be used to explore expected patterns in contributions (Lang and Wu, 2013). We explore the relationship between three user-engagement parameters in route-sharing communities: (i) number of routes, (ii) user score and (iii) number of sessions.

First, we test the direct correlation between the number of routes per user linked to a task set and our lifetime metrics. TUL shows a relatively strong correlation with the number of routes per user (0.54). in contrast, we see a weaker correlation between number of routes per user and AUL (0.19). Finally, PUL shows a similar weak correlation (0.25). Next, in order to compare users' involment in the RouteYou community and their lifetime metrics in our experiment, we have also analyzed the correlation between AUL, TUL and PUL and the individual user score. There is a significant but relatively weak positive correlation between AUL and the user score (i.e. a higher user score results in a higher AUL) (0.38). On the other hand, TUL shows a stronger positive correlation (0.48). PUL and the user score are more difficult to interpret as the correlation between both is rather weak (0.20). Finally, we can also compare the number of sessions per user with the defined lifetime metrics. We only see a relatively strong correlation between AUL and number of session (0.55). In contrast, TUL shows a weaker positive correlation with the number of user sessions (0.19). PUL has no significant correlation with this variable. Furthermore, correlation between the number of routes and sessions is also weak (0.33).

4.5 Experimental results

In order to assess the validity of the presented experiment design and the value of contributions, we now focus our attention on a descriptive overview of data collected during the experiment's run time. We also present a sample of the harvested data to study the nature of these data. Figure 4.10 on page 105 gives a grid overview of the full segment set linked with Q_3. With increasing detail, we first present the full spatial coverage of this segment set. Next a detail of a specific region in Belgium and finally a detailed grid with a segment overlay is presented. Tiles with at least one segment running through it are black and tiles with segments which have answers on Q_3 linked to it are green. The tiles have an approximate size of 1.5 km^2. As expected, the spatial distribution of segments is spatially correlated with included routes in the experiment. It is also clear that answers are locally clustered and will be linked to a contributors region of interest.

highway	answer fraction
unclassified	12.0%
secondary	0.5%
track	40.7%
footway	4.4%
cycleway	11.0%
service	0.7%
living_street	0.1%
tertiary	3.4%
path	19.4%
residential	5.8%
pedestrian	0.8%
bridleway	0.6%
road	0.1%

Table 4.2: This table gives an overview of the highway tag values where segments occur and their fraction of the answers linked to them

Table 4.2 indicates that the largest number of answers where collected on the highway tag value track, path and cycleway. Approximately 60% of the answers were collected on roads with an unpaved road condition class. Moreover, 47% of this answer fraction contradicted the unpaved road condition classes. Figure 4.11 on page 106 presents a more detailed view on five randomly selected answers in the segment set. Next to the OSM road object on the left-hand side of the image, we also show a Google StreetView image along the segment to give a general impression of the segment. On the right-hand side, we list the OSM URI, tags linked to this road object and the submitted answer and remark linked to the segment. The first (OSM Way 73905335) and fifth (OSM Way 24497283) example are situations where the harvested answer is an added value to the current state of the OSM road object, specifying that the road is paved. These answers are in contrast with the general characteristics of a track. The Google StreetView images confirm the answer given by the contributors. The second (OSM Way 112129773), third and forth example (OSM Way 367047242) are a verification of the information-bearing tag *surface*.

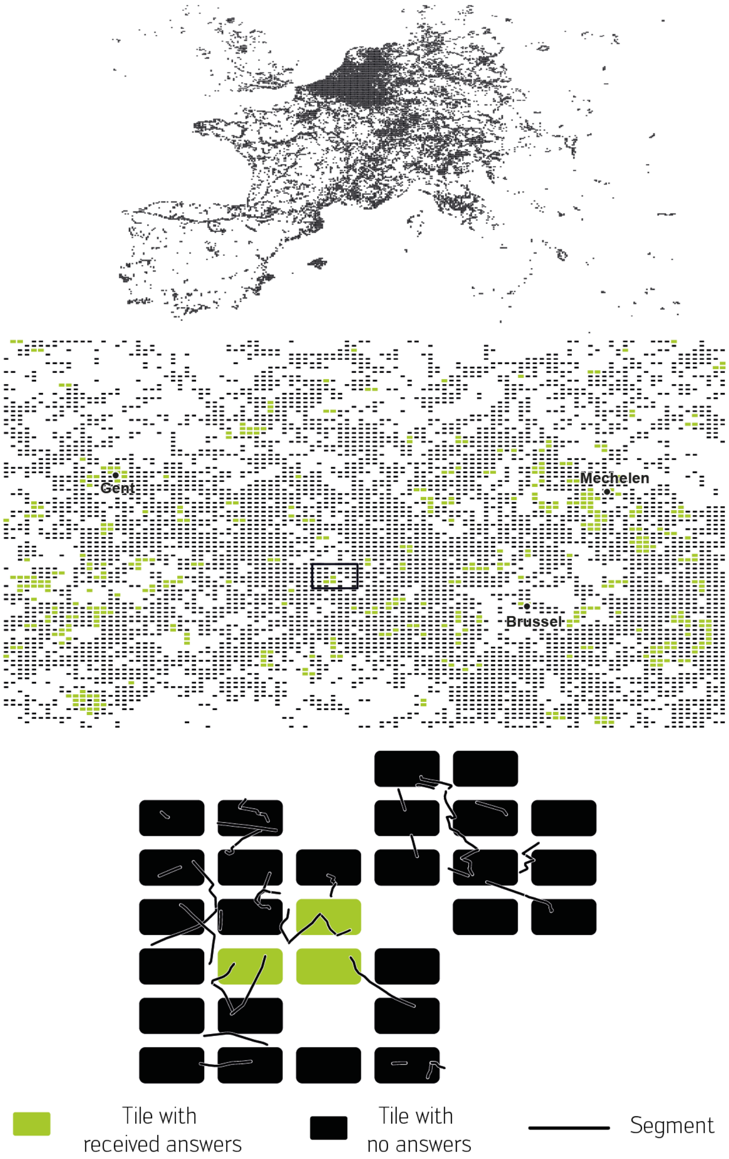
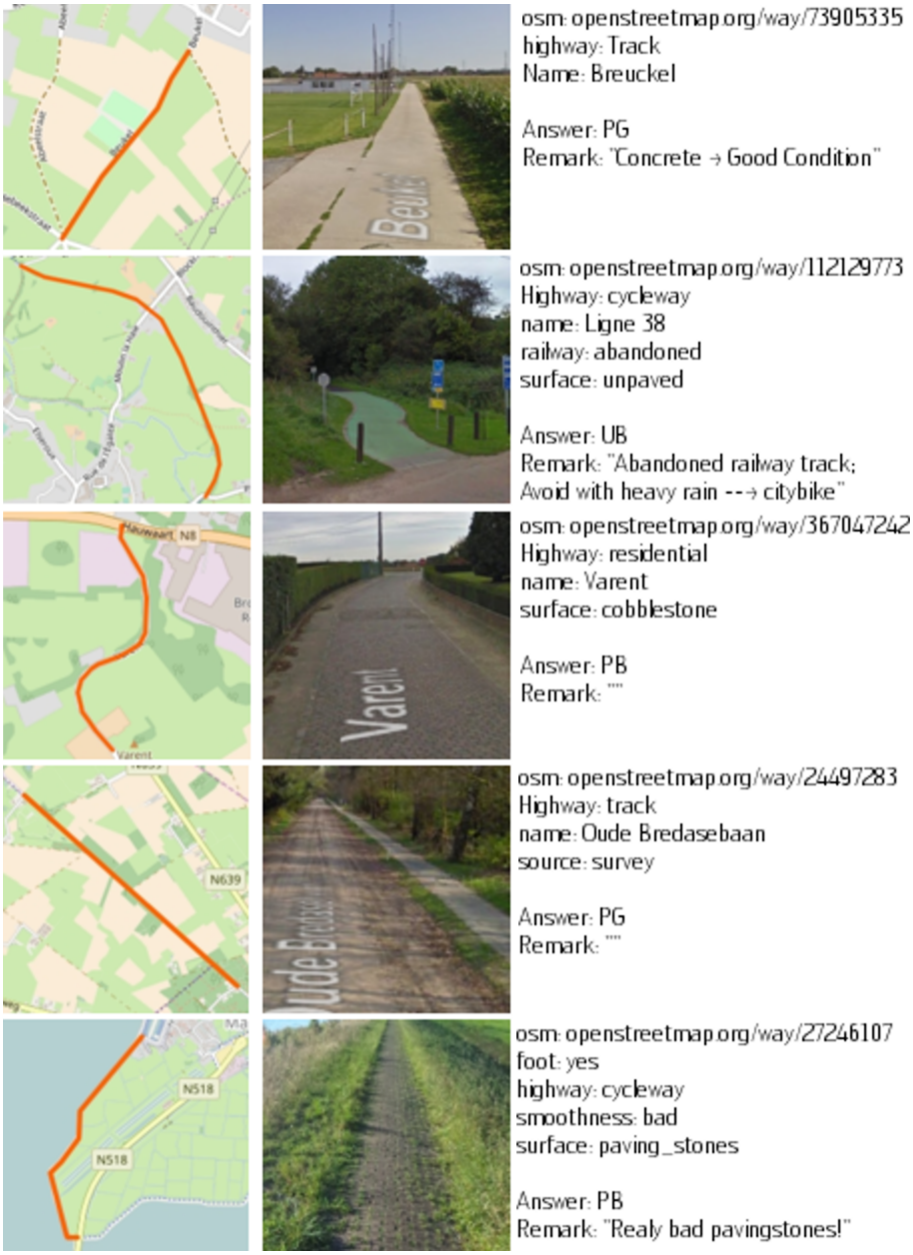


Figure 4.10: This overview presents tiles (approximately 1.5 km^2) with segments (black) and tiles with answered question on $Q3$ linked to a specific segment (green). The bottom image is a detail with a segment overlay.



osm: openstreetmap.org/way/73905335
highway: Track
Name: Breuckel

Answer: PG
Remark: "Concrete → Good Condition"

osm: openstreetmap.org/way/112129773
Highway: cycleway
name: Ligne 38
railway: abandoned
surface: unpaved

Answer: UB
Remark: "Abandoned railway track;
Avoid with heavy rain --→ citybike"

osm: openstreetmap.org/way/367047242
Highway: residential
name: Varent
surface: cobblestone

Answer: PB
Remark: ""

osm: openstreetmap.org/way/24497283
Highway: track
name: Oude Bredasebaan
source: survey

Answer: PG
Remark: ""

osm: openstreetmap.org/way/27246107
foot: yes
highway: cycleway
smoothness: bad
surface: paving_stones

Answer: PB
Remark: "Realy bad pavingstones!"

Figure 4.11: Detail of five harvested answers accompanied by their OSM road object (left-hand side) and its characteristics (right-hand side) and a Google StreetView image along the segment. Remarks are translated and paraphrased.

4.6 Context-driven knowledge discovery

In this section we study the possible redundancy of a user task compared to underlying rules and patterns based on the routes linked to a task's segment. Within this evaluation we again focus on the tasks linked to Q_3 and use the answers as ground truth. This exploratory analysis is of importance to future work, streamlining community efforts to answer questions.

We define a group of routes linked to a segment as a cluster. If we analyze the presence of these clusters and the contributed answers we see that 82% of the included routes in this analysis co-occur with a paved road condition. Furthermore, the experiments' database (see Figure 4.5) allows us to study three more specific features within these clusters. We define the presence of specific activity types in a cluster as a first feature. As described in the experiment design, the decision to link certain questions to segments was already based on this co-occurrence. Hence, we study both the predictive value of a route's activity type in future work as well as the validity of the co-occurrence in our experiment design. Secondly, we evaluate the size of the cluster (i.e. the number of routes linked to a segment) as a predictive feature. Previous work (Baker et al., 2017) already highlighted the value of location-based route clusters to infer a suitability or attractiveness of a road or path for a specific leisure activity type, such as road cycling. As such, we study if this cluster size also has a predictive value for the surface type of a road. Thirdly, we study user diversity in these clusters. Within our experiment design, user diversity can be defined as the number of unique owners of routes linked to a segment. As such, we evaluate if the popularity of a road or path within the social group "cyclist" or, more specifically, "road cyclists" has a predictive value for the surface type of a road. Three clear and comprehensible association rules can hence be proposed based on this feature set:

rule 1 - $\{A\} \implies \{Paved\}$: This association rule suggests that certain activity type tags constitute a strong argument for a paved road condition class.

rule 2 - $\{R_1, R_2, ..., R_n\} \implies \{Paved\}$: This association rule suggests that a higher number of routes linked to a segment constitutes a strong argument for a paved road condition class

rule 3 - $\{U_1, U_2, ..., U_n\} \implies \{Paved\}$: This association rule suggests that a higher number of different users on a segment constitutes a strong argument for a paved road condition class

The predictive value of a rule can best be described by its confidence, support and lift (e.g. Agrawal et al., 1993; Hipp et al., 2000). If we apply these performance metrics to our rule set, support represents the occurrence of the rule in the clusters on all segments. Confidence, on the other hand, represents the number of times the right-hand side of the rule is true in comparison to the total times the left-hand side occurs. Finally, lift indicates whether the support of the rule is higher than what could be expected if the left-hand side of the rule was independent from the contributed answers in our experiment. Hence, if lift is higher than 1, the value of the rule is perceived higher. Table 4.3 lists the confidence, support and lift of the first rule with specific cycling type tags. It is clear that the applicability and relevance of our first rule with more general type tags "Cycling" and "Leisure Cycling" is hampered. In contrast, a single route with the specific type tag "Road Cycling" has a predictive value. The upper three plots in 4.12 on page 109 show the confidence, support and lift of rule 2. A clear impact of increasing number of routes can be observed in this rule. The confidence reaches its upper limit as all segments with a cluster size above 120 routes always results in the paved road condition class. The support has an exponential decay. The plots linked to rule 3 in 4.12 can be interpreted in a similar manner as the trends seen in rule 2; the observed trends are very much alike. With a higher number of different users on a segment there tends to be a higher lift and confidence together with exponential decay in support. This exploratory analysis shows that our three rules have a predictive value in the harvested answer set.

Type	Confidence	Support	lift
Cycling	0.61	0.14	0.74
Leisure Cycling	0.55	0.04	0.83
Road Cycling	0.94	0.63	1.13

Table 4.3: This table lists the confidence, support and lift of rule 1 with specific cycling types

4.7 Discussion

4.7.1 Participation inequality

As mentioned before, participation inequality is known to have a theoretical 90/10 distribution (Haklay, 2016). The conversion rate in our experiment shows a strong participation inequality. Unquestionably, the user lifetime metrics suggest that there is room for improvement in both user experience

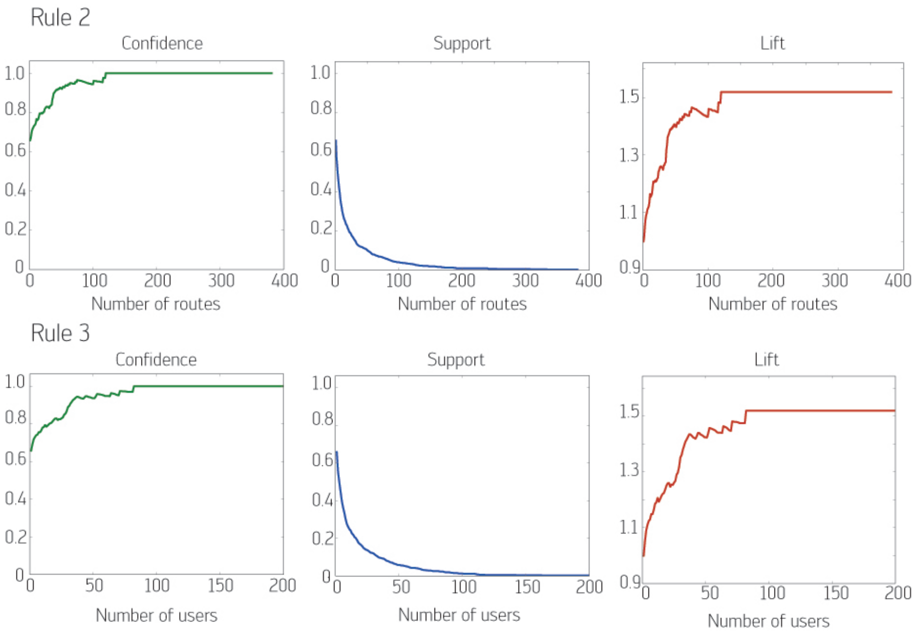


Figure 4.12: This image presents an overview of the perfomance metrics as seen in rule 2 and 3 respectively with increasing number of routes on a segment and increasing number of unique route owners (users) in the route set on a segment.

and user engagement. In an ideal situation, a user's TUL should be similar to his AUL and his PUL should be reduced to a minimum. Our results reflect a more complex interaction between these metrics. This underpins the statement that contribution in online communities is not solely explained by use time and frequency, but is also influenced by a degree of serendipity. We discuss some of the possible trends shaping the contribution patterns.

User life cycle in the route-sharing platform has an impact on our experiment. Other work in route-sharing platforms and mobile sports tracking applications (see for example Oksanen et al., 2015) already noted the impact of periodic and non-periodic events, such as seasons or weather, on contribution rates. This hypothesis is supported by the horizontal variations in the first phase. However, a similar impact is more obscured in the second phase by other drivers of contribution. As expected, the number of routes shows a relatively strong positive correlation with TUL. The fact that this correlation is not stronger can be explained by two factors. First, the previously mentioned periodic and non-periodic events can result in surges of route creation and, hence, create outliers of low TUL and higher number of routes. Similary, the correlation between user score and TUL can be influenced by these surges. Secondly, we also see a power law behavior in the experiment results. The proposed Pearson coefficient lacks robustness to model these distributions and handle the inherent extreme outliers, for example a high TUL and very high number of routes. Next, all correlations with AUL are clearly impacted by the high number of non-returning contributors in the second phase. The results also showed that non-returning contributors were found in all ranges of user types in the route-sharing community. However, we also observed that if users got involved for a longer time it was more likely that they returned more often. Evaluating, analyzing and monitoring the proposed metrics in the future will certainly give valuable insights in the evolution of the performance of this and similar experiments. Especially noteworthy is the possibility to further characterize contributors and focus on more in-depth analysis of specific characteristics. Figure 4.13 highlights one contributor group that could be of particular interest for further analysis. This arbitrary group of users have a high TUL and high AUL, which make them valuable subjects to study why the stayed engaged in the experiment. This example highlights the value of the discerned life time metrics in future research.

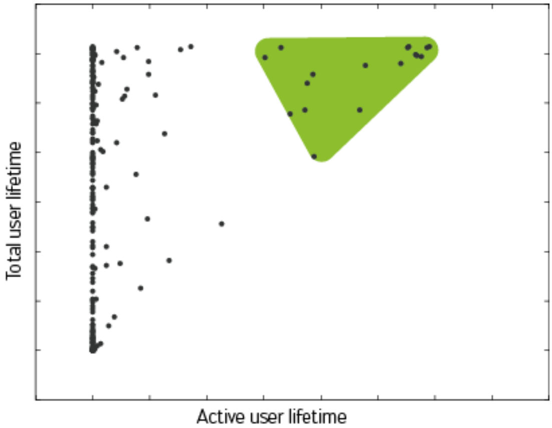


Figure 4.13: This scatterplot presents total user lifetime in function of active user life time. The green area highlights contributor types which could be of particular interest for further analysis.

There is also a noticeable difference between both phases. Next to the already mentioned explanations, the methods we used to reach contributors have an impact on the contribution pattern. Based on the results of our experiment, some research theses can be discerned for future work. In the last decades, a growning body of research addressed the current *email overload* (Whittaker and Sidner, 1996). This was also the reason why we moved away from this method to reach users in the second phase of our experiment. However, email contact in the first phase showed its merits to source returning contributors. Based on these results, direct contact could be deemed more personal and suitable to reach highly-involved users. Future work will have to address if, how and when direct contact can be used to lower the number of non-returning visitors. However, while future work should certainly study these drivers and reveal the stimulation needed to encourage contribution, they will always be nuanced by what cannot be seen or measured in proxies.

Finally, the validity of a user's routes as proxy for his local knowledge and expertise should be discussed. After all, users also use RouteYou's tools to explore regions with little or no local knowledge. While the results clearly indicated that an active contributor consciously chooses which task sets to answer, future work will be necessary to underpin our approach with an assessment of relevance and trust of an answer. Due to the low fraction of multiple answers to the same problem, we choose not to evaluate the correspondence between answers on the same problem. However, a commonly used technique to elicit trust and relevance of a crowd-sourced answer is the evaluation of corresponding answers on the same problem. A higher number of contribution and a larger overlap in the local expertise of contributors will be imperative to introducing similar techniques. A second important relevance factor is the time and geographical place that a contributor is asked for feedback. In the current design, a task set is linked to a route on the creation date of the particular route and the contributor is free to choose when to answer the set. However, the local knowledge of a user changes in time and space. For example, it is better to ask feedback after an actual activity on a particular route. Similar work of Huang et al. (2014) proposed the use of mobile applications to collect on-site knowledge. Finally, further, more in depth, user study of the experiments' active contributors can certainly give insights in the validity of the assumption that a route encompasses a part of his mental map.

4.7.2 Rule mining

We chose to analyze three clear and comprehensible rules. In the following section we discuss some of the insights gained from this analysis. The performance metrics of these rules indicated co-occurrences which can not merely be explained by chance. Furthermore, it is also clear that our rules are, to a certain degree, interwoven with each other. As described in the experiment design, the decision to link certain questions to segments was already based on this co-occurrence. However, the results in Section 4.6 proved the validity of this approach, but also showed that a more intricate decision tree is possible to guide users to more specific task which need local knowledge. Future work will certainly need to address the applicability of more intricate data-driven approaches to infer new decisions on route sets and study their possible consequences. Furthermore, Machine-learning literature (e.g. Bishop, 2006), has already noted the risk of overfitting an inferred model, resulting in algorithmic *blind spots*. The exploratory analysis of rules was based on the answers gathered in our experiment. As such, the spatial generality of the inferred rules will be a necessary subject to avoid these blind spots in future work. As mentioned in the previous section, trust and relevance metrics will be crucial to support these data-driven approaches. A quality indication of both the answers and the meta information of a route, such as type tag, could greatly improve trust in the presented approach. More in depth verification of the answers given in our experiment could also result in a unique ground-truth dataset.

4.8 Conclusion and main contributions

In this use case we introduced an online experiment using a personalized task-recommendation mechanism. This is an example of a piggyback approach of a crowd-sourcing system. We studied the value of active user interaction to improve the problem-solving ability of maps and used the content (i.e. routes) shared within route-sharing communities as driver for this task-recommendation mechanism (RQ1). Next, we studied the validity of the inferred knowledge both as a single source of new information and within more context-driven knowledge discovery (RQ2). The later was studied by comparing a set of features found in route sets co-occurring with the harvested answers.

RQ1 - This question focused on the methodological approach to create personalized tasks based on shared content. The route segmentation approach proved its value to create semantically coherent pieces of information upon which could be reasoned to create a crowd-sourcing task. In this specific use case, a task-recommendation mechanism was used to address the generic surface classification in OSM. Due to the generic nature of object attribution in OSM and its proneness to errors (Challenge III - *dirty* nature of the collaborative data sources), expert-based classifications can fail and hamper activity-specific navigation services. The design succeeded to engage users in change detection mapping in this context. Consider the 47% of the harvested answers contradicting the expert-based classification. In addition, several benchmarks can be recommended to evaluate the design performance in the future. Next to well-known measures such as conversion and returning rate, the devised design allowed us to evaluate three more specific user lifetime metrics. First, conversion rate is an important measure in online behavior. Despite the high number of generated tasks, the fraction of solved task sets is low, that is, a low conversion rate. While this can be a sign of a hampered user experience, participation inequality is inherent in crowds and their contribution behavior. Yet a more equitable distribution is desirable. However, the increase will be asymptotic. Secondly, the users in the learning curve of the experiment should increase. A low returning rate, low active user lifetime and high passive user lifetime are signs that users lack incentives to keep contributing. A stronger linear relation between the total user lifetime and the active user lifetime will be an important benchmark in future evaluations. Furthermore, returning contributors are essential to incorporate trust and relevance metrics in the design. Streamlining the efforts of contributors will be imperative and can be addressed by the second research question in this use case.

RQ2 - We found that a single route has a predictive value. Yet it is hampered by the generality of certain types, such as 'Cycling', and obviously prone to misclassification. The predictive value increases as the size of the route set and its user diversity increases. We can conclude that the presented exploratory analysis of actionable rules showed potential to streamline the efforts of contributors. Hence, there is a clear synergy between the data-driven approach devised in Chapter 2 and active user interactions. Exploiting this synergy further will be imperative. More insights in what, when and where users are willing to help are essential. This will both increase the performance of our design and the quality of the shared information.

References

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93, pages 207–216, New York, NY, USA. ACM.

Anvik, J., Hiew, L., and Murphy, G. C. (2006). Who should fix this bug? In *Proceedings of the 28th International Conference on Software Engineering*, ICSE '06, pages 361–370, New York, NY, USA. ACM.

Baker, K., Ooms, K., Verstockt, S., Brackman, P., Van de Walle, R., and De Maeyer, P. (2017). Crowdsourcing a cyclist perspective on suggested recreational paths in real-world networks. *Cartography and Geographic Information Science.*

Basiri, A., Amirian, P., and Mooney, P. (2016). Using crowdsourced trajectories for automated osm data entry approach. *Sensors*, 16(9):1510.

Bergström, A. and Magnusson, R. (2003). Potential of transferring car trips to bicycle during winter. *Transportation Research Part A: Policy and Practice*, 37(8):649–666.

Bettenburg, N., Just, S., Schröter, A., Weiss, C., Premraj, R., and Zimmermann, T. (2008). What makes a good bug report? In *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, SIGSOFT '08/FSE-16, pages 308–318, New York, NY, USA. ACM.

Biagioni, J. and Eriksson, J. (2012). Map inference in the face of noise and disparity. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '12, pages 79–88, New York, NY, USA. ACM.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. In *Proceedings of the 9th International World Wide Web Conference on Computer Networks : The International Journal of Computer and Telecommunications Networking*, pages 309–320, Amsterdam, The Netherlands, The Netherlands. North-Holland Publishing Co.

Burns, P. C. (1998). Wayfinding errors while driving. *Journal of Environmental Psychology*, 18(2):209–217.

Coleman, D. J., Georgiadou, Y., Labonte, J., et al. (2009). Volunteered geographic information: The nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research*, 4(1):332–358.

Damant-Sirois, G., Grimsrud, M., and El-Geneidy, A. M. (2014). What's your type: A multidimensional cyclist typology. *Transportation*, 41(6):1153–1169.

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., and Dixon, D. (2011). Gamification. using game-design elements in non-gaming contexts. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 2425–2428, New York, NY, USA. ACM.

Doan, A., Ramakrishnan, R., and Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Commun. ACM*, 54(4):86–96.

Ferrara, J. (2013). Games for persuasion: Argumentation, procedurality, and the lie of gamification. *Games and Culture*, 8(4):289–304.

Flanagin, A. J. and Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*, 72(3):137–148.

Gao, M., Liu, K., and Wu, Z. (2010). Personalisation in web computing and informatics: Theories, techniques, applications, and future research. *Information Systems Frontiers*, 12(5):607–629.

Geiger, D. and Schader, M. (2014). Personalized task recommendation in crowdsourcing information systems — current state of the art. *Decision Support Systems*, 65:3 – 16.

Golledge, R. G. (1999). *Wayfinding behavior: Cognitive mapping and other spatial processes*. JHU Press, Baltimore.

Guo, L., Tan, E., Chen, S., Zhang, X., and Zhao, Y. E. (2009). Analyzing patterns of user content generation in online social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 369–378, New York, NY, USA. ACM.

Haklay, M. (2013). Citizen science and volunteered geographic information: Overview and typology of participation. In Sui, D., Elwood, S., and Goodchild, M., editors, *Crowdsourcing Geographic Knowledge*, pages 105–122. Springer Netherlands, Dordrecht.

Haklay, M. (2016). Why is participation inequality important? In Cristina, C., Cristina, C., Muki, H., Haosheng, H., Vyron, A., Juhani, K., Frank, O., and Ross, P., editors, *European Handbook of Crowdsourced Geographic Information*, pages 35 – 44. Ubiquity Press, London.

Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2000). Algorithms for association rule mining &mdash; a general survey and comparison. *SIGKDD Explor. Newsl.*, 2(1):58–64.

Hochmair, H. H., Zielstra, D., and Neis, P. (2015). Assessing the completeness of bicycle trail and lane features in openstreetmap for the united states. *Transactions in GIS*, 19(1):63–81.

Hooimeijer, P. and Weimer, W. (2007). Modeling bug report quality. In *Proceedings of the Twenty-second IEEE/ACM International Conference on Automated Software Engineering*, ASE '07, pages 34–43, New York, NY, USA. ACM.

Huang, H., Klettner, S., Schmidt, M., Gartner, G., Leitinger, S., Wagner, A., and Steinmann, R. (2014). Affectroute – considering people's affective responses to environments for enhancing route-planning services. *International Journal of Geographical Information Science*, 28(12):2456–2473.

Kasemsuppakorn, P. and Karimi, H. A. (2013). A pedestrian network construction algorithm based on multiple GPS traces. *Transportation Research Part C: Emerging Technologies*, 26:285–300.

Kessler, F. (2011). Volunteered geographic information: A bicycling enthusiast perspective. *Cartography and Geographic Information Science*, 38(3):258–268.

Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Trawling the web for emerging cyber-communities. *Computer Networks*, 31(11–16):1481 – 1493.

Lang, J. and Wu, S. F. (2013). Social network user lifetime. *Social Network*

Analysis and Mining, 3(3):285–297.

Liu, X., Biagioni, J., Eriksson, J., Wang, Y., Forman, G., and Zhu, Y. (2012). Mining large-scale, sparse gps traces for map inference: Comparison of approaches. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 669–677, New York, NY, USA. ACM.

Luxen, D. and Vetter, C. (2011). Real-time routing with openstreetmap data. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 513–516. ACM, New York, USA.

Martella, R., Kray, C., and Clementini, E. (2015). A gamification framework for volunteered geographic information. In Bacao, F., Santos, M. Y., and Painho, M., editors, *AGILE 2015: Geographic Information Science as an Enabler of Smarter Cities and Communities*, pages 73–89, Cham. Springer International Publishing.

Matyas, S., Kiefer, P., Schlieder, C., and Kleyer, S. (2011). Wisdom about the crowd: Assuring geospatial data quality collected in location-based games. In Anacleto, J. C., Fels, S., Graham, N., Kapralos, B., Saif El-Nasr, M., and Stanley, K., editors, *Entertainment Computing – ICEC 2011: 10th International Conference, ICEC 2011, Vancouver, Canada, October 5-8, 2011. Proceedings*, pages 331–336, Berlin, Heidelberg. Springer Berlin Heidelberg.

Neis, P. and Zielstra, D. (2014). Recent developments and future trends in volunteered geographic information research: The case of openstreetmap. *Future Internet*, 6(1):76–106.

Neis, P. and Zipf, A. (2012). Analyzing the contributor activity of a volunteered geographic information project — the case of openstreetmap. *ISPRS International Journal of Geo-Information*, 1(2):146–165.

Newson, P. and Krumm, J. (2009). Hidden markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 336–343, New York, NY, USA. ACM.

Nielsen, J. (2006). The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities.

Oentaryo, R. J., Lim, E.-P., Lo, D., Zhu, F., and Prasetyo, P. K. (2012). Collective churn prediction in social network. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ASONAM '12, pages 210–214, Washington, DC, USA. IEEE Computer Society.

Oksanen, J., Bergman, C., Sainio, J., and Westerholm, J. (2015). Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. *Journal of Transport Geography*, 48:135–144.

Panciera, K., Priedhorsky, R., Erickson, T., and Terveen, L. (2010). Lurking? cyclopaths?: A quantitative lifecycle analysis of user behavior in a geowiki. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1917–1926, New York, NY, USA. ACM.

Preece, J. J. (2000). *Online Communities: Designing Usability and Supporting Sociability*. John Wiley and Sons, New York, USA.

Quddus, M. A., Ochieng, W. Y., and Noland, R. B. (2007). Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312–328.

Schroedl, S., Wagstaff, K., Rogers, S., Langley, P., and Wilson, C. (2004). Mining GPS traces for map refinement. *Data Mining and Knowledge Discovery*, 9(1):59–87.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.-Y., Milčinski, G., Nikšič, M., Painho, M., Pődör, A., Olteanu-Raimond, A.-M., and Rutzinger, M. (2016). Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information*, 5(5):55:1 – 55:23.

Vanclooster, A., Van de Weghe, N., and De Maeyer, P. (2016). Integrating indoor and outdoor spaces for pedestrian navigation guidance: A review. *Transactions in GIS*, 20(4):491–525.

Vyron, A. and Schlieder, C. (2014). Participation patters, vgi and gamification. In *Proceedings of the Seventeenth AGILE Conference on Geographic Information Science, Geogames and Geoplay Workshop*.

Wang, Y. and Fesenmaier, D. R. (2003). Assessing motivation of contribution in online communities: An empirical investigation of an online travel community. *Electronic Markets*, 13(1):33–45.

Whittaker, S. and Sidner, C. (1996). Email overload: Exploring personal information management of email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '96, pages 276–283, New York, NY, USA. ACM.

Yanenko, O. and Schlieder, C. (2014). Game principles for enhancing the quality of user-generated data collections. In *Proceedings of the Seventeenth AGILE Conference on Geographic Information Science, Geogames and Geoplay Workshop*.

Zhao, Y. and Han, Q. (2016). Spatial crowdsourcing: current state and future directions. *IEEE Communications Magazine*, 54(7):102–107.

Zheng, Y. (2015). Trajectory data mining: An overview. *ACM Trans. Intell. Syst. Technol.*, 6(3):29:1–29:41.

5

General discussion and conclusion

"Never send a human to do a machine's job"
Agent Smith, The Matrix

5.1 Summary

There are several important findings that stem from the specific use cases presented in this book. It is the aim of this section to provide a link between these findings and create a broader context. In addition, more specific research questions were introduced in each chapter to guide the reader and highlight the chapter's major contributions. Figure 5.1 presents a summary of these topical research questions. Next, we draw upon the introduced overarching research questions to reassess and summarize these findings.

RQ1: How can large route sets maintained on route-sharing platforms be used to improve activity-specific navigation services?

Chapter 2 and 4 demonstrated that an activity-specific route set has a predictive value to characterize roads and paths in the models supporting navigation services. Both chapters showed how to capture and analyze new information from a route set using a novel contextual analysis. We moved beyond statistical exploration of road object attributes in a route set to improve expert-based models (i.e. Route Choice Modeling (e.g. Prato, 2009)). By eliciting and using the context within activity-specific route sets as a proxy for certain characteristics of a road or path, we were able to study how navigation services for leisure and sports can be improved. We also want to highlight that routes, even if they lack a temporal dimensions, can be used for more than popular trajectory pattern mining or anomaly detection (i.e. missing roads) as discussed

Chapter	RQ	Materials and Topics	Main contributions and highlights
Chapter 2: Crowdsourcing a cyclist perspective on suggested recreational paths in real-world networks	RQ1	Materials: routes - GPS tracking logs collected through user uploads Topics: - How can movements described in large route sets maintained on route-sharing platforms be used to model specific features or qualities of a road? - Can this model be used to create a route suggestion for specific leisure activities? - How can path suggestion be evaluated?	- Step-by-step preprocessing steps of a route set - Remodelling procedure of network-constrained phenomena (e.g. movements along a road network in rouite sets) into a cost function (i.e. appreciation index) - Generic evalutation set-up to analyse and compare path characteristics and route optimality
Chapter 3: Cultural heritage routing: a recreational navigation based approach in exploring cultural heritage	RQ2	Materials: POI - theme-specific set of POI Topics: - How does the current state-of-the-art in geographic enrichment fit the use case of cultural heritage? - How can contextual analysis of POI be used to enrich navigation services?	- Detailed overview of how the current state-of-the-art in geographic enrichment can be used in a topical/thematic context - Moving-segment analysis to calculate a segment-based attraction-accessibility measure
Chapter 4: Knowledge sharing in a route-sharing community to improve navigation services	RQ1;RQ3	Materials: Routes and user feedback- all routes created during the runtime of the presented experiment and the user feedback collected in this experiment Topics: - How can routes created on route-sharing platforms be used by a task-recommendation mechanism to foster knowledge sharing? - Is active user interaction outpreformed by current data-driven approaches (e.g. Chapter 2)?	- Route segmentation approach proved its value to create semantically coherent pieces of information upon which could be reasoned to create a crowd-sourcing task - Engaging users in change detection mapping - A single route has a predictive value., but its reuse is hampered by the generality of certain types - A clear synergy between the data-driven approach devised in Chapter 2 and active user interactions

Figure 5.1: Summary of the topical research questions presented in each chapter

by Zheng (2015) and Basiri et al. (2016) or novel visualization approaches such as heat maps (e.g. Oksanen et al., 2015) to reveal new insights. We readdress certain aspects within the specific use cases.

Map matching is a well-studied tool to alleviate noise while recording location of moving objects in a road network (e.g. humans, vehicles) (Newson and Krumm, 2009; Quddus et al., 2007). It also proved to be an important preprocessing step in our contextual analysis. Besides noise reduction, Zheng (2015) hinted at its importance as an alternative to complex trajectory clustering based on similarity in feature vectors. Our research together with other recent work (such as, but not limited to, Bergman and Oksanen (2016)) support this hypothesis. It allowed us to create a semantic route segmentation, linking sets of road object attributes and clusters of routes and their meta information. While this map-matching approach is constrained in complex road lay-outs or dense urban networks and hampered by inaccuracies in digital road networks (Bierlaire and Frejinger, 2008; Quddus et al., 2007), the map-matching results presented in both use cases proved their value in exploiting the potential in routes in our research context. As such, it succeeded in simplifying the complex task of trajectory clustering and studying spatial patterns in route sets to more comprehensible analyses of network-constrained events.

However, there are several limitation for future valorisation. First and foremost, the constrainedness to a network is both its strength and weakness. Due to the high degree of freedom in both the route-plannning tools and recorded movements in GPS traces during sports and leisure activities, map-matching approaches will always be constrained in identifying all real-world movements. Furthermore, we also want to note that map matching large route sets can be a time-demanding task despite the current high-performance algorithms and increased computational power of hardware. Finally, the matching results are a static snapshot of a road network which becomes invalid when the digital record of the road changes. For example, the simple task of splitting a road object makes the link between route and road network useless. Future challenges lie in devising new data management techniques fostering efficient storage, retrieval and mining without losing the advantages of map matching.

Especially noteworthy is the predictive value of a **cluster size** or route count on a specific road object as a result of above-described preprocessing step. Chapter 2 documented a localized remodeling of this count to infer an attractiveness index of every road object in a network. We analyzed the spread of the derived popularity score in four shortest-path alternatives of popular routing engines for this activity. This analysis successfully discriminated these shortest paths based on the scoring value and three morphological parameters of the path. However, the robustness of the model should be improved to ensure the viability of the proposed approach in future work. More specifically, further research on the local optimality of the route choices will be imperative.

A distinguishable benchmark in cluster sizes is crucial to achieve reproducibility and scale in future work. We need to know how much routes are necessary to attain a certain degree of predictive performance, that is a critical mass. We propose the use of across-network cluster size variation and network autocorrelation as powerful tools to study this in future research tracks.

From another viewing angle, the latter was also studied in Chapter 4. We evaluated the predictive value of a segment's cluster features, such as route count, in relation to the harvested answers. We found that a relatively small route count has a high predictive performance if combined with specific activity types, such as road cycling. A single route tagged as road cycling, for examples, co-occurs in 94% of route-segment pairs with an answer indicating a paved road condition. This not only emphasizes the potential to streamline the efforts of contributors by using a more intricate decision tree based on segment's cluster features in the presented task-recommendation approach. It also underpins the value of the route count in Chapter 2 to improve activity-specific navigation services. Yet, the impact of mis matches in type tagging should be studied in more depth.

Furthermore, the value of content-rating systems in route-sharing platforms or more intricate use or popularity proxies ,such as downloads or views, could provide additional predictive value to the feature set of a cluster. For example, it can be expected that two highly-rated routes have a higher predictive value than two routes without any rating score. However, we have to be aware that proxies of reality, such as routes and their meta information, are inherently fallible. This does not undermine the value of the presented work, but emphasizes the need for a combination of expert-based and data-driven explanations to exploit the full value of route sets to improve activity-specific navigation services. By integrating both explanations in model-based logic, we can condense knowledge in these models which can not be encompassed in clear causalities or association rules such as "When a path has a characteristic X and Y, it is (highly/not) suitable for race cyclist".

Another interesting finding for future valorisation trajectories is the value of **a route to create a context** within which an individual is able to share his local knowledge. Finding the right contributors is one of the main challenges in crowd-sourcing approaches (Doan et al., 2011; Geiger and Schader, 2014). A task-recommendation mechanism and a web-based feedback tool were devised in Chapter 4 based on the aforementioned route segmentation. In doing so, we tried to search through the localized knowledge collection of end-users and connect this knowledge with specific data patterns in route segments. Almost half of the harvested answers at least questioned the quality of road object attributes describing the road condition. It underscored that given the right tools at the right time, valid and actionable knowledge can be collected from a variety of user types based on their routes. This finding is also related to RQ3 and will be discussed further in the following sections linked to this questions.

RQ2 - Which opportunities do POIs provide to enrich current navigation services beyond well-known map exploration?

Points of interest or POIs are an important aspect to enable storytelling within route-sharing communities. They have the potential to create a richer and personalized experience for a long tail of smaller niche markets within tourism and leisure activities. A cycling round trip in the countryside, for example, can attract different, but very specific user groups depending on the location-based information, such as pictures, text or videos, linked to the route. Consider, for example, recreational cyclists interested in geology, industrial buildings and history. Automating this route- and content-generation has the potential to attract a large and diverse user group with a decreasing marginal cost. Using the multidisciplinary view point within this dissertation, Chapter 3 documented and discussed a generic toolset to address this need.

First, we focused on content creation. There is a current trend of making information location aware. Yet valuable information remains location- and context-unaware. This makes this aspatial information useless within spatial knowledge-discovery techniques and GIS, central in this dissertation. We presented different building blocks to condense information in space beyond what is user generated, creating trustworthy and actionable knowledge; for example a user-generated geotag in contrast to the applied geographic entity recognition in our approach. An important aspect in automated content generation is quality propagation of meta data. As such, a quality metric was devised based on the work of Moen et al. (1998), Ochoa and Duval (2009) and Bruce and Hillmann (2004) to mitigate trust and credibility problems while reusing the harvested information. While further testing of this framework is necessary to come to a formal evaluation, it provides valuable insights to build on during future valorisation trajectories. Understanding the quality of a POI, can tell us how and when it can become valuable for a user group.

While the opportunities presented by the growing availability of open APIs and other software-as-a-service are vast, data management will become an important challenge in future valorisation. Performance-demanding services are constrained in condensing the necessary information on the fly. Pre- and post-processing steps will be necessary to keep the actionable data available and fresh. Furthermore, the commercial context in future valorisation can limit the availability of certain resources and increase costs to acquire the necessary information.

Next, we reused a set of quality-controlled poi in a network-constrained accessibility measure. Concurrent with the state-of-the-art in clustering events on a road network (e.g. Okabe and Sugihara, 2012; She et al., 2015), we documented an approach to measure the thematic utility of an activity-specific network to increase storytelling-functionalities of current navigation services in a cost-effective way. It catches a location-specific *attractiveness* and guides the user's route choices to maximize the utility of the generated route within a specific sightseeing theme (i.e. World War I battlefield exploration). In this sense, it differs from the more holistic attractiveness index in chapter 2 and RouteYou's *attractiveness* model. While we lack formal end-user validation, the experimental results prove the potential of our attraction-accessibility measure to detect salient corridors in activity-specific networks and its value as a travel-cost alternative in theme-based services. We concluded that the largest valorisation potential lies in using the presented measures to improve the results of combinatorial routing problems such as described by Maervoet et al. (2013). By incorporating more contextual measure such as ours in path suggestion algorithms, the optimized route connecting a set of theme-specific POIs can be enriched with additional cultural value and historical significance.

RQ3 - Do route-sharing platforms have engaged end-users that are able to help improve the problem-sovling ability of maps? If so, what is the value of active user interaction?

Considering the vast amount of created content on route-sharing platforms, these communities have engaged members which spend significant time to create routes and poi. The results linked to RQ1 and RQ2 already highlighted that as a route-sharing platforms and their communities mature (i.e. more engaged users that share content, attracting other engaged users), the results harvested from the adapted methods have the potential to also improve. Services can benefit from this network effect. The performance to present real-world situation and collect this information in a time-effective manner increases. In this scenario, user-generated content and provided services are deemed reciprocal in the customer relationship. Users receive a basic set of tools to manage their content. In exchange, service providers create value on top of the created content by attracting new users, by data-driven improvements of services, by targeting advertising or by popular freemium models (e.g. advanced features behind a paywall).

Yet these data-driven approaches underestimate the full value of an engaged end-user. Coleman et al. (2009) noted that many map-based services recognize this value and incorporate error-reporting tools to facilitate low-level knowledge sharing about potential errors, ommisions or inaccuracies. However, we stated that these tools are maybe not the right tool for different levels of engaged users in route-sharing communities. Chapter 4 focused on the steps to foster continuous user contributions in a specific problem-solving type of crowdsourcing. Especially noteworthy is that given a well-defined context, we found that specific information can be collected from individual users without any financial reward. Not unexpectedly, however, we also saw clear constraints, consistent with other research (e.g. Haklay, 2016; Panciera et al., 2010; She et al., 2015), in what can be collected within a given time frame (e.g. participation inequality, power-law behavior and limited spatial density). It is clear that future work will have to focus on finding the right contributor in the right contextual setting, such as time, place and medium. If successful, it has the potential to not only harvest new knowledge, but also increase the involvement and engagement of end-users of route-sharing platforms.

5.2 General discussion

The aim of this section is to present additional topics of discussion and remaining issues. With the benefits of hindsight, the following sections (re-)addresses some aspects which present challenges for future valorisation. We use both the insights summarized in the previous section and those gathered during four years of project-based research focused on improving navigation services for leisure activities to bring this dissertation to a more general outlook on this research topic.

5.2.1 Generalizability of methods

This dissertation describes topical use cases within a specific frame of reference. Several of the proposed methodologies were evaluated within one route-sharing community (i.e. RouteYou), in a confined region of interest (e.g. regions in Belgium) and with very specific types of leisure activities (e.g.

race cycling, cycling in a sightseeing context). Yet the value of the different building blocks discussed throughout this dissertation is not limited to these specific set-ups. The implementation of the presented tools, methods and ideas in the thriving community of RouteYou allowed us to evaluate these building blocks within other set-ups.

Several opportunities and challenges were already hinted upon throughout the different use cases. Country-specific road network layouts and characteristics, for example, were already discussed as a challenge when scaling and reproducing the introduced methods. But this is hardly a new truth. Local or regional road network characteristics also hamper current expert-based models. Functional road classes and their characteristics are defined network- and continent-wide, but often lack subtle local changes which can be very important for activity-specific navigation services. Incorporating a high granularity in tools will be necessary in future work, balancing between a coarse-grained model or a too ad-hoc solution. In the remainder of this section, we (re-)assess additional drivers in the generalizability of our methods.

User contributions in route-sharing communities can also have a very regional character and, hence, condition the applicability of our methods. The community size and country focus of a route-sharing platform can impact this regional character. Furthermore, this *regionality* also varies between types of leisure activities. Consider the activity space of a hiker. In general, a hiker's activity space is much smaller and has a much stronger local character than, for example, a race cyclist (Chapter 2). As such, the applicability of the presented methods is not only conditioned by the size of the active community, but also by the activity space of the modeled recreational pastime. Hence, this conditioning factor has to be taken into account when transposing the gathered insights on other leisure activities.

The in-house research character of this dissertation had clear benefits. It made it possible to directly interact with an active community project and understand the different aspects when sourcing specific information. Furthermore, it presented hands-on insights which could not have been conceived without this context. However, as already mentioned throughout the different use cases, all methods are reproducible on other route-sharing communities and their crowd-based sources. Yet understanding when and why certain data and meta data are collected remains important and emphasizes why these sources can not just be seen as a commodity. Consider the type tagging of routes. Platform-specific incentives to link correct types to shared content can have an impact on the quality of the tag. Some platforms make the type tagging embedded in the content-creation process, while other platforms make this type tagging optional. Understanding these nuances while reusing crowd-based sources impacts the value of derived ancillary data and their ability to improve maps.

5.2.2 OpenStreetMap

In the last five years, a growing number of commercial service providers use the free-for-use[1] map data from the OpenStreetMap project to increase performance and reduce costs. RouteYou, the industrial partner within this dissertation, did the same thing. As the community behind this mapping project is reaching a critical mass and commercial geographic data and service providers such as Mapzen[2],Mapbox[3] or Telenav [4] invest in improving this map and its community, the value for activity-specific navigation services also increases (Kessler, 2011). However, as the project matures, a growing body of research focused on this specific crowd-sourcing project still highlights the impact of participation inequality in space and time (Haklay, 2016). This results in a regional-dependent (in)completeness and (in)accuracy in both road geometry and object attributes. Similar to other community-driven projects, mappers can lack experience, intrinsic motivators or incentives, knowledge, time or attention to fix inaccuracies or omissions. At the same time, a regional lack of mappers results in *under-mapped* places in time and space.

See et al. (2016) discerned change detection mapping as one of the most important future challenges of community-driven mapping projects. The results in chapter 4, for example, underpinned that statement. 47% of the harvested answers on unpaved roads or paths in OpenStreetMap contradicted this classification. This highlights the value of the presented work in this dissertation to improve machine-readable maps, and more specifically those based on OpenStreetMap, for activity-specific navigation services. Future research will have to focus on how this synergy can be further exploited and which role different stakeholders can play. Especially interesting are possible overlapping and conflicting interests in this updating process. The question arises if the goals of the different stakeholders, such as commercial data/service providers or individual community members, will remain the same as this project further matures.

[1] OpenStreetMap is open data, licenced under an Open Data Commons Open Database License

[2] https://mapzen.com

[3] https://www.mapbox.com

[4] http://www.telenav.com

5.2.3 Personalization and privacy

Personalization is an important aspect in every facet of social media platforms (Gao et al., 2010). As such, the presented work can have a big potential in all facets of a route-sharing platform. Navigation services, the main focus of this dissertation, is only one example. But this is just the start. Personalization can also, for example, help you find your coming Sunday morning ride, making it a personalized and user-centered experience. An inferred user-specific geographical context and its characteristics can help attain this type of service. A simple place to start this personalization is a user's shared and used content, such as routes, and their geographical context. Based on a user's content of interest a (spatial) area of interest can be defined.

However, concurrent with the increasing commercial interest in personalized services, the public concern about privacy is also growing. There are obvious ethics in re-using shared content. Many commercial companies focusing on this personalization underestimate the value of an emotional customer relationship to mitigate these privacy concerns. Of particular interest in this context is the value of active user interaction. Aside from knowledge collection, it can be used to sensitize end-user to the potential value of (personal) data collection and how it is used to improve services for everyone. Creating a transparency in data re-usage is imperative. In addition, doing this in a cost-effective way is equally important.

5.2.4 Maintaining up-to-the-minute navigation services

As discussed in the introduction, the success of navigation services lies in the combination of the right machine-readable data and recent online and mobile technological advances. We presented three use cases to infer new information to improve this machine-readable data to create up-to-date navigation services. In this section, we want to (re-)address why a *reality gap* can persist through time despite *fresh* information.

Within the use cases, we already referred at the importance of choosing the right routing algorithm. The most suitable algorithm to solve a specific query to a navigation service can change both for reasons of performance and optimality of the given solution. The work of Bast et al. (2016) gives an overview of several state-of-the-art routing algorithms in performance-demanding applications and their strengths and weaknesses. Certainly when maintaining continent-size services, as RouteYou does, algorithm choice can have a major impact on perceived performance by end-users (i.e. latency), but also on service-provider side in preprocessing efforts and space usage and, hence, costs. As such, even if up-to-the-minute information becomes available in trustworthy and actionable data structures, the capacity to include this information can be limited by available resources. Furthermore, providing a customized navigation service also entails maintaining a spatial database storing both the geographic objects (lines and points) and their intricate attribute structure. While speed-up techniques increase the performance of necessary data transformation, specific manipulations are still hampered when working on world-wide or continent-size road networks. As such, available resources can again limit the speed with which new information can be integrated in these navigation services. While this does not undermine the validity of the presented use cases, future work will certainly focus on further valorization of the research results within this context.

5.2.5 Data graveyards

There is a disincentive to trade as each side will worry that it is getting the short end of the stick

The Economist
"Fuel of the Future: Data is Giving Rise to a New Economy"

The crowd-based sources used in this research were explicitly shared within the community of RouteYou: (i) routes and (ii) POI. However, there are other sources which could provide valuable insights in the context of our research. Up until a decade ago private firms have been able to collect a great deal of information about our leisured self (e.g. Griffin and Jiao, 2015; Lupton, 2016; Silk et al., 2016; Stragier et al., 2016). However, many of these sources are still stored in so-called *data silos* and linked to single-source applications. Consider brand-based platforms such as Garmin Connect, collecting all tracking logs of Garmin devices. The aforementioned disincentive to trade still hampers harnessing the full potential of these sources, resulting in *data graveyards* from the viewpoint of our research. Searching for cross-correlations between platforms could prove to be a very valuable research track for future work. On example is measuring the reuse of shared content. A community of a route-sharing platform thrives on reuse of shared routes. However, this reuse is most often not explicitly shared within the community. Highly-engaged users will use feedback mechanisms such as up- and down-votes or other comment systems, but a lot of users lack this engagement. Cross-correlations could provide further contextual information improving the presented analyses in this dissertation. Especially noteworthy in this context is the growing availability of web-enabled applications and APIs with end-user authentication, fostering a cross-platform sharing of content. While these data sources can prove to be valuable in a research context, guidelines on how to use these APIs in a commercial context are at the least fuzzy and hard to interpret.

Another potential source in this research are other sources of probe data. Probe data is best described as the spatio-temporal residue of the aforementioned mobile sensing revolution (Srivastava et al., 2012); traces of locations and time stamps lumped in logs as a result of the burgeoning amount of location-based services and location-aware devices. Many commercial mapping agencies such as Mapbox or Google already collect this data through free and reusable software development kits (SDK) for application builders. However, to our knowledge, few efforts have been made to use this data to address the main objective of this dissertation. The competitive advantage of collecting this data is also in this situation inseparable from the disincentive to trade. The question rises if it is up to route-sharing platforms to create a new *data silo* with self-collected probe data from a dedicated applications to improve their services.

5.3 Conclusion and further research avenues

We started this dissertation by stating that maps and their problem-solving ability have limits as information is inevitably left out. Within the bounds set by their business model, mapping companies fail to capture specific information. *Reality gaps* exist in maps and their underlying data. As such, missing, erroneous or out-dated information in machine-readable maps hamper navigation services and their model-based logic. Because of these *reality gaps* they fail to unravel the full complexity of route choices and parametrize certain factors. This limits their current use in a burgeoning amount of location-based services such as popular navigation services for leisurely sports or recreational pastime activities (e.g. cycling, sightseeing).

Considering both the crowd-based sources managed on route-sharing platforms and the crowds creating these data, we directed the presented research to how crowds and their leisured data doubles can be used to cross these gaps in a creative and cost-effective way. In doing so, we based three use cases on two specific data sources managed within these communities: (i) routes and (ii) POI. The research specifically focused on their capacity to create ancillary data and streamline a community's efforts to further activity-specific navigation services.

Overall, all three use cases showed their merits to gauge the intrinsic value of specific paths and roads for these services. Route-sharing communities and their crowd-based sources are a resource and a tool. The presented use cases amply illustrate their potential to improve and personalize navigation services. Furthermore, the adapted methods have the potential of doing this in a more cost- and time-effective manner. It is, however, recognized that the created ancillary data from a single source of content (e.g. a route set or topical POIs sourced from thematic websites) is not perfect. The ensuing discussion highlighted both specific and more general insights together with important issues which require further research in the future. We briefly summarize these insights and give an outlook on future work.

As mentioned before, research in online communities is often hampered by what researchers cannot see or measure. There is a necessity to assume a certain user behavior. However, the different use cases showed how these used assumption have a validity to source actionable ancillary data and attain the goal of this dissertation. The presented use case can prove to be a good starting point to posit new and more differentiated research question to gather new insights in future research.

Secondly, we again want to highlight the fact that these sources should not just be seen as a commodity. Understanding the biasing factors in crowd-based sources such as routes is challenging, but is necessary to move beyond "what" can be seen and comprehending "why" something is seen. Making the feedback loop more interwoven in communities such as those managed on route-sharing platforms has a clear potential in this process, but closing this loop has proven to be challenging. Future work will have to create deeper interpretation of what, when and where users are willing and able to give valuable feedback. Furthermore, the strengths of both active and passive information collection will have to be more intricately combined to attain this goal.

From a broader perspective, the major challenges hence still lie in the combination and fusion of different community-driven or authorative (spatial and aspatial) data sets and their derived ancillary data. Linking different sets of ancillary data onto a specific contextual spatial element, such as a road network segment or a location with a rich context, has the potential to create holistic characteristics of these spatial elements. This insight offers new ways of looking at geographical data sources and emphasize their value for a wide variety of applications.

References

Basiri, A., Amirian, P., and Mooney, P. (2016). Using crowdsourced trajectories for automated osm data entry approach. *Sensors*, 16(9):1510.

Bast, H., Delling, D., Goldberg, A., Müller-Hannemann, M., Pajor, T., Sanders, P., Wagner, D., and Werneck, R. F. (2016). *Route Planning in Transportation Networks*, pages 19–80. Springer International Publishing, Cham.

Bergman, C. and Oksanen, J. (2016). *Estimating the Biasing Effect of Behavioural Patterns on Mobile Fitness App Data by Density-Based Clustering*, pages 199–218. Springer International Publishing, Cham.

Bierlaire, M. and Frejinger, E. (2008). Route choice modeling with network-free data. *Transportation Research Part C: Emerging Technologies*, 16(2):187–198.

Bruce, T. R. and Hillmann, D. (2004). *The continuum of metadata quality: defining, expressing, exploiting*. American Library Association, Chicago.

Coleman, D. J., Georgiadou, Y., Labonte, J., et al. (2009). Volunteered geographic information: The nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research*, 4(1):332–358.

Doan, A., Ramakrishnan, R., and Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Commun. ACM*, 54(4):86–96.

Gao, M., Liu, K., and Wu, Z. (2010). Personalisation in web computing and informatics: Theories, techniques, applications, and future research. *Information Systems Frontiers*, 12(5):607–629.

Geiger, D. and Schader, M. (2014). Personalized task recommendation in crowdsourcing information systems — current state of the art. *Decision Support Systems*, 65:3 – 16. Crowdsourcing and Social Networks Analysis.

Griffin, G. P. and Jiao, J. (2015). Where does bicycling for health happen? analysing volunteered geographic information through place and plexus. *Journal of Transport & Health*, 2(2):238 – 247.

Haklay, M. (2016). Why is participation inequality important? In Cristina, C., Cristina, C., Muki, H., Haosheng, H., Vyron, A., Juhani, K., Frank, O., and Ross, P., editors, *European Handbook of Crowdsourced Geographic Information*, pages 35 – 44. Ubiquity Press, London:.

Kessler, F. (2011). Volunteered geographic information: A bicycling enthusiast perspective. *Cartography and Geographic Information Science*, 38(3):258–268.

Lupton, D. (2016). Foreword: lively devices, lively data and lively leisure studies. *Leisure Studies*, 35(6):709–711.

Maervoet, J., Brackman, P., Verbeeck, K., De Causmaecker, P., and Vanden Berghe, G. (2013). *Tour Suggestion for Outdoor Activities*, pages 54–63. Springer Berlin Heidelberg, Berlin, Heidelberg.

Moen, W. E., Stewart, E. L., and McClure, C. R. (1998). Assessing metadata quality: findings and methodological considerations from an evaluation of the us government information locator service (gils). In *Proceedings of IEEE International Forum on Research and Technology Advances* *in Digital Libraries. ADL 98*, pages 246–255, Los Alamitos, California. IEEE COmputer Society.

Newson, P. and Krumm, J. (2009). Hidden markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 336–343, New York, NY, USA. ACM.

Ochoa, X. and Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, 10(2):67–91.

Okabe, A. and Sugihara, K. (2012). *Spatial Analysis Along Networks: Statistical and Computational Methods*. Statistics in Practice. Wiley, Chichester, West Sussex.

Oksanen, J., Bergman, C., Sainio, J., and Westerholm, J. (2015). Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. *Journal of Transport Geography*, 48:135–144.

Panciera, K., Priedhorsky, R., Erickson, T., and Terveen, L. (2010). Lurking? cyclopaths?: A quantitative lifecycle analysis of user behavior in a geowiki. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1917–1926, New York, NY, USA. ACM.

Prato, C. G. (2009). Route choice modeling: past, present and future research directions. *Journal of Choice Modelling*, 2(1):65–100.

Quddus, M. A., Ochieng, W. Y., and Noland, R. B. (2007). Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312–328.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.-Y., Milčinski, G., Nikšič, M., Painho, M., Pődör, A., Olteanu-Raimond, A.-M., and Rutzinger, M. (2016). Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information*, 5(5):55:1 – 55:23.

She, B., Zhu, X., Ye, X., Guo, W., Su, K., and Lee, J. (2015). Weighted network voronoi diagrams for local spatial analysis. *Computers, Environment and Urban Systems*, 52:70 – 80.

Silk, M., Millington, B., Rich, E., and Bush, A. (2016). (re-)thinking digital leisure. *Leisure Studies*, 35(6):712–723.

Srivastava, M., Abdelzaher, T., and Szymanski, B. (2012). Human-centric sensing. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 370(1958):176–97.

Stragier, J., Abeele, M. V., Mechant, P., and Marez, L. D. (2016). Understanding persistence in the use of online fitness communities: Comparing novice and experienced users. *Computers in Human Behavior*, 64:34 – 42.

Zheng, Y. (2015). Trajectory data mining: An overview. *ACM Trans. Intell. Syst. Technol.*, 6(3):29:1–29:41.

Appendices

A

Accomplishments

A.1 Peer-reviewed publications

A.1.1 Accepted for publication

Baker, K., Ooms, K., Verstockt, S., Brackman, P., De Maeyer, P. and Van de Walle, R. (2017). Crowdsourcing a Cyclist Perspective on Suggested Recreational Paths in Real-world Networks. *Cartography and Geographic Information Science*, 44(5):422-435.
doi: https://dx.doi.org/10.1080/15230406.2016.1192486

Baker, K. and Verstockt, S. (2017). Cultural Heritage Routing: a Recreational Navigation based Approach in Exploring Cultural Heritage. *ACM Journal on Computing and Cultural Heritage*, 10(4):1-20.
doi: https://doi.org/10.1145/3040200.

A.1.2 In review

Baker, K., Ooms, K., Verstockt, S., Brackman, P., De Maeyer, P. and Van de Walle, R. (2017). Knowledge Sharing in a Route-sharing Community to Improve Navigation Services. *working paper.*

A.2 Other publications

Maertvoet, J., Baker, K. and Vande Berghe, G. (2013). Route planning enhancement through collective intelligence. LICT '13

Baker, K., Brackman, P., De Maeyer, P. and Van de Walle, R. (2013). Reconstructing movement traces through a hybrid map-matching algorithm.

Pre- AGILE conference workshop – Understanding urban cycling: a data challenge

Baker, K., Brackman, P., De Maeyer, P. and Van de Walle, R. (2013). From passive to active crowd-sourcing using location-based questionnaires: Google Street View to rebuild a cyclists' experience. Mobile Ghent '13

Baker, K., Alivand, M., Slavkovikj, V., Mannens, E. and Verstockt, S. (2014). Multi-modal route enjoyment prediction. W3C Linking Geospatial Data Workshop (LGD14)

Slavkovikj, V., Baker, K., Alivand, M., Colpaert, P., Mannens, E. and Verstockt, S. (2014). Personalized time-dynamic points of interest. W3C Linking Geospatial Data Workshop (LGD14)

Baker, K., Alivand, M., Slavkovikj, V., Mannens, E. and Verstockt, S. (2014). Crowd-sourcing trail popularity for online route planning. GISSCIENCE 2014 – Workshop on Role of Volunteered Geographic Information in Advancing Science: Effective Utilization

Snizek, B., Vanclooster, A., Yeboah, G., Barkow, B., Sick Nielsen, T., Skov-Petersen, H., Van de Weghe, N. and Baker, K. (2014). CopenhagenABM: an agent based model of cyclists' travel. GISSCIENCE 2014 – Workshop on analysis of movement data

Verstockt, S., Slavkovikj, V., Baker, K. and Van de Walle, R. (2014). Map-based linking of geographic user and content profiles for hyperlocal content recommendation. ACM SIGSPATIAL GIS 2014

Verstockt, S., Slavkovikj, V. and Baker, K. (2015). Mapbased linking of geographic user and content profiles for hyperlocal content recommendation. 17th International Conference, HCI International 2015, Proceedings. pp. 53–63

Baker, K. (2016). A Digital Ecosystem to Collect Recrational Route Choice Decisions. Belgian Geography Days

Verstockt, S., Baker, K., De Mey, K. and Stragier, J. (2016). Gamification-based feedback app for crowdsourced monitoring of recreational cycling and running loops. Science and Engineering Conference on Sport Innovations.

A.3 Bachelor's and master's theses (advisor)

Academic year	Student	Title
2014-2015	Bram Van Impe	Generatie van trainingsschema's voor wielertoeristen met verrijkte kaartdata
2014-2015	Karel Geiregat	Active Crowdsourcing Issue Detector
2016-2017	Mathijs Raats	Ontwikkeling van een Methode om de Tagkwaliteit van Routes Na te Gaan
2016-2017	Lennert Teugels	Optimalisatie van het SRTM-model op basis van fitness sensor data
2016-2017	Gert Blanchaert	Route Matching op basis van Community Data met Kwaliteitsverificatie als doel
2016-2017	Thomas Van de Weghe	Dynamische routeringsalgoritmes

A.4 Internships (advisor)

Academic year	Student
2014-2015	Tim Baert
2014-2015	Karel Geiregat
2015-2016	Mathijs Raats
2016-2017	Gert Blanchaert

A.5 Student projects (advisor)

Academic year	project
2015-2016	Concor
2016-2017	Runamic Ghent
2016-2017	Running in Ghent 2.0

B

Contribution pattern

This should be seen as an appendix to Chapter 4 and, similar to Figure 4.8 on page 100, shows the contribution patterns in the devised experiment at the time of writing.

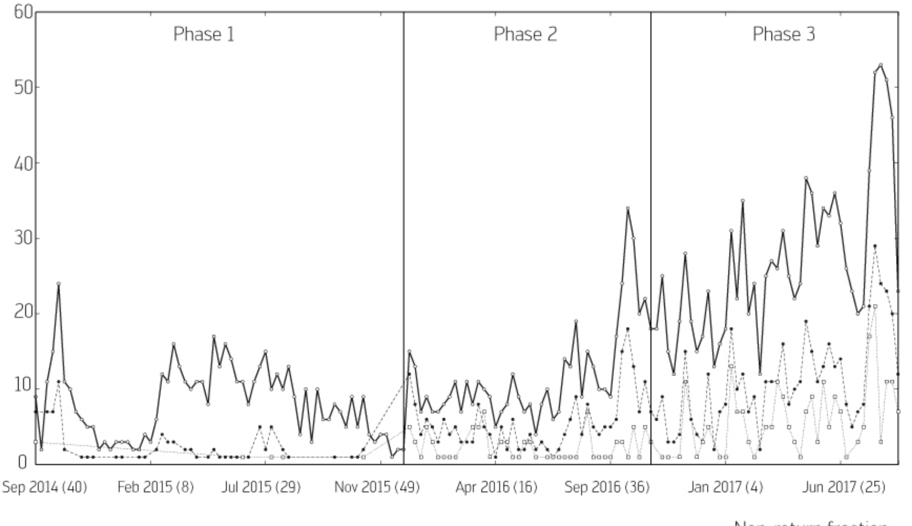


Figure B.1: This plot presents the active contributors per week (white circle) during the full run time of the online experiment. The black circles represent the first-time contributors to the experiment in the particular week and the squares represent the non-returning fraction of these first-time contributors.

Nederlandstalige samenvatting - summary in Dutch

Probleemstelling

Kaarten vormen een belangrijk hulpmiddel om complexe ruimtelijke problemen op te lossen. Dit proefschrift focust op één van deze problemen: navigatie. Kaarten hebben immers de mogelijkheid om dit probleem op een duidelijke, snelle en begrijpelijke manier op te lossen.

Echter, de volledigheid waarmee de wereld in kaart gebracht wordt, en dus ook het probleemoplossend vermogen van kaarten, heeft zijn beperkingen; informatie wordt immers onvermijdelijk weggelaten. Eén reden hiervoor is duidelijk: rekening houdend met de schaal en het doel van het kaartproduct, moet bepaalde informatie gegeneraliseerd worden. Een tweede oorzaak vinden we in het bedrijfsmodel van kaartproducenten. Door economische en budgettaire keuzes wordt namelijk bepaalde informatie niet (tijdig) verzameld.

Efficiënt ontbrekende of onvolledige informatie inwinnen, zowel in kost als in tijd, is bijgevolg één van de belangrijkste uitdaging voor kaartproducenten om aldus aan de groeiende vraag voor specifieke ruimtelijke informatie te voldoen. Deze groei dient niet alleen gezien te worden in aantal vragen; ook de complexiteit neemt toe. Navigatie is daarenboven al jaren niet meer beperkt tot analoge kaarten. Digitale navigatiediensten gaan machine-leesbare kaarten *interpreteren* om eindgebruikers te helpen hun weg te vinden tijdens verschillende activiteiten.

Bovenstaande uitdaging staat centraal in dit proefschrift, waarbij we de focus leggen op de specifieke nichemarkt van digitale recreatieve navigatiediensten. Het voorbije decennia zijn er verschillende technologische vooruitgangen geboekt binnen deze sector zoals mobiele, locatiebewuste toestellen en applicaties die navigatietoepassingen tijdens vrijetijdsbesteding zoals fietsen of wandelen mogelijk maken. Echter, voor vele van de kaartproducenten is juist deze informatie moeilijk te verzamelen om verschillende redenen. Zo is een reden de beperkte toegankelijkheid van de plaatsen die van belang zijn voor recreatieve navigatie, zoals smalle paden en lokale wegen. Hoewel in het laatste decennia grootschalige karteringscampagnes veel efficiënter zijn geworden (bijvoorbeeld door *mobile mapping*), blijven deze plaatsen kostelijk om op te nemen in deze campagnes. Daarnaast zijn, zoals reeds vermeld, de verwachtingen van de dienstverlening van applicaties in recreatieve navigatie hoog. Men verwacht correcte, gepersonaliseerde informatie, wat vandaag de dag echter nog niet vaak verzameld wordt tijdens karteringscampagnes.

In dit boek zoeken we een antwoord op deze uitdaging in de huidige stroom van geografische informatie gegenereerd door een grote groep individuen. Afgelopen jaren zijn er immers verschillende online platformen ontstaan waar gemotiveerde eindgebruikers hun persoonlijke interesses, ervaringen en gevoelens over hun vrijetijdsbesteding kunnen delen. Een van deze platformen is RouteYou, de industriële partner binnen dit proefschrift. We onderzoeken hoe deze nieuwe, door gebruikers gegenereerde, geografische informatie kan ingezet worden om recreatieve navigatiediensten te verbeteren op een kostefficiënte manier. Centraal hierbij staan twee types van informatie: (i) routes en (ii) bezienswaardigheden.

Dit proefschrift beoogt twee doelstellingen. Ten eerste beschouwen we drie verschillende methodologische benaderingen om deze nieuwe bronnen te gebruiken om navigatiediensten te verbeteren. Ten tweede willen we inzichten verwerven in zowel de opportuniteiten als de uitdagingen in deze context. Naast bovenvermelde opportuniteiten zijn de vier voornaamste uitdagingen: (i) hoe willen eindgebruikers helpen om diensten te verbeteren, (ii) de geschiktheid van deze bronnen voor onderzoek, (iii) 'ruwheid' van deze bronnen en (iv) begrijpbaarheid en verklaarbaarheid van de gegenereerde informatie. Deze doelstellingen worden verder uitgediept in drie specifieke onderzoeksvragen:

- OV1: Hoe kunnen beschikbare sets van routes beheerd binnen een online community gebruikt worden om navigatiediensten voor specifieke activiteiten beter te maken?
- OV2: Welke opportuniteiten bieden bezienswaardigheden om navigatiediensten te verrijken zodat ze verder gaan dan huidige toepassingen?
- OV3: Hebben online communities waarbinnen routes worden gedeeld geëngageerde eindgebruikers die bereidt zijn te helpen het probleemoplossend vermogen van kaarten te verbeteren? Indien wel, wat is de waarde van actieve intereactie met deze gebruikers?

Overzicht van belangrijkste bevindingen

In **hoofdstuk 2** beschrijven we een benadering die de focus legt op het integreren van een uniek gebruikersperspectief in navigatiediensten. Daarbij wordt een specifieke activiteit uitgelicht, namelijk racefietsen. We ontwikkelden een methodologische benadering die een appreciatie-index berekent voor iedere weg in een wegennetwerk. We baseren ons hiervoor op een clustering van identieke bewegingen beschreven in een grote set routes (**OV1**). De belangrijkste bijdragen hieromtrent kunnen samengevat worden in drie delen.

De methodologische stappen vormen een eerste bijdrage. We geven een overzicht van de uitdaging bij het gebruiken van routes om identieke bewegingen in een wegennetwerk te bepalen. Hiervoor gebruiken we een *map matching* algoritme, waarbij we verschillende heuristieken introduceren om de herberekening van bewegingen langsheen het wegennetwerk te verbeteren (cfr. 10%-regel, halfweg-regel). Analyse van de *match* resultaten toont aan dat geclusterde bewegingen op wegsegmenten karakteristieken hebben van menselijke bewegingspatronen, zoals een verdeling met een dikke staart en geen significante netwerk autocorrelatie.

Een tweede bijdrage zit in de modellering van de clusters op de wegsegmenten in het netwerk waarbij het gebruikersperspectief centraal staat. Dit perspectief wordt beïnvloed door afstand en netwerkconnectiviteit. We gebruiken netwerkgebaseerde afstandsweging om dit perspectief te integreren in onze modellering. Daarnaast biedt de β-transformatie een generieke oplossing om de appreciatie-index om te zetten naar een kost in een kortste pad algoritme. Een laatste bijdrage zit in de evaluatieprocedure om kortstepad-alternatieven te vergelijken. De voorgestelde procedure gebruikt nieuwe visualisaties en analyses om verschillende alternatieven tegenover elkaar af te wegen.

Tenslotte concluderen we in dit hoofdstuk dat belangrijke inzichten verworven worden over hoe routes gebruikt kunnen worden om navigatiediensten te verbeteren. Het grootste potentieel in toekomstige valorisatie zal gevonden kunnen worden in het combineren van dergelijke datagedreven benaderingen en expertgebaseerde modellen. Hierdoor kan de impact van ontbrekende of onvolledige informatie opgevangen worden en kan op een kost-efficiënte manier aan de groeiende vraag voldaan worden.

Hoofdstuk 3 is gelinkt aan **OV2** en documenteert een benadering om de geografische context van bezienswaardigheden te gebruiken om navigatiediensten te verrijken. Hierbij focussen we op twee bijdrages: (i) het automatisch vergaren en verrijken van thematische bezienswaardigheden van cultureel erfgoed en (ii) het gebruiken van toegankelijkheidsmetrieken om een set bezienswaardigheden te aggregeren op een activiteitspecifieknetwerk.

Ten eerste toont het beschreven generieke stappenplan aan dat de huidige *state-of-the-art* in geografische verrijking gebruikt kan worden in een thematische context. Het leggen van thematische nuances is hierbij evenwel nodig om het potentieel volledig te benutten. Bijvoorbeeld, de temporele veranderingen in plaatsnamen krijgen weinig aandacht in de huidige technologische bouwstenen om geografische entiteiten te detecteren in teksten, hoewel deze een belangrijke impact kunnen hebben in thematische beschrijvingen van cultureel erfgoed.

Vervolgens wordt een contextuele methodiek opgezet om de potentiële interactie tussen een netwerk en de gegeorefereerde informatie, met name bezienswaardigheden, te modelleren. Hiervoor gebruiken we een bekende benadering, namelijk toegankelijkheidsmetrieken (*attraction-accessibility measures*). Deze benadering wordt gebruikt binnen een segmentfunctie die deze potentiele interactie berekend voor kleinere delen van het netwerk (m.n. segmenten van 100 m). Deze benadering bewijst zijn begrijpbaarheid, elegantie en kracht om dit doel te bereiken.

We concluderen dat de beschreven technologische bouwstenen nieuwe opportuniteiten blootleggen om actieve toeristen dichter bij cultureel erfgoed te brengen. Daarnaast biedt de generieke architectuur van deze bouwstenen ervoor dat deze zich niet alleen beperken tot cultureel erfgoed, maar ook kunnen gebruikt worden voor andere activiteitsspecifieke bezienswaardigheden.

De derde en laatste voorgestelde benadering in **hoofdstuk 4** spitst zich toe op **OV1** en **OV3**. Hierbij ligt de focus op actieve interactie met eindgebruikers op basis van hun gedeelde routes op online platformen zoals RouteYou. We beschrijven een experiment dat een gepersonaliseerd taakallocatiemechanisme gebruikt om deze actieve interactie te verbeteren. Dergelijke benaderingen worden ook wel *piggyback*-benaderingen genoemd omdat ze een bestaande applicatie en hun geëngageerde eindgebruikers inzetten om nieuwe informatie te verkrijgen via *crowdsourcing*. Hierbij stellen we twee doelen voorop. Ten eerste willen we actieve interactie gebruiken om het probleemoplossend vermogen van kaarten te verbeteren. Daarnaast willen we analyseren hoe deze informatie ook kan gebruikt worden om meer datagedrevenbenaderingen, zoals beschreven in hoofdstuk 2, te verrijken. Deze analyse vergelijkt de verzamelde antwoorden met een specfieke *feature set* verkregen uit een route set.

We gebruiken de actieve interactie om specifieke vragen te stellen over het verhardingstype van een weg. Deze vraagstelling is gestuurd door een conflictdetectie tussen het activiteitstype gelinkt aan routes (bv. racefietsen) en attributen van het onderliggende wegennetwerk beschreven in OpenStreetMap. Bijvoorbeeld, racefietsers wensen routes die niet over onverharde wegen gaan. Op basis van een routesegmentatie-algoritme, een beslissingsboom en een webgebaseerde feedbackapplicatie linken we een specifieke vragenset aan een individu.

De experimentele resultaten geven aan dat we de lokale kennis van eindgebruikers kunnen inzetten om specifieke informatie te verkrijgen. Zo geven 47% van de verkregen antwoorden een contradictie aan met het verwachte verhardingstype op basis van wegattributen.

Hoewel de resultaten waardevolle informatie bevatten, zijn er ook beperkingen in wat kan verwacht worden van dergelijke *crowdsourcing*-benaderingen. Zoals aangegeven in de academische literatuur rond actieve interactie op het web, zal altijd maar een kleine fractie van de beoogde populatie deelnemen en zal het overgrote deel van de contributies komen van een nog kleinere fractie. We moeten dus op een intelligente manier omspringen met de beperkte waardevolle informatie.

De vergelijking van de verkregen antwoorden met de *feature set* verkregen uit een routeset toont aan dat er verschillende mogelijkheden zijn om de opgestelde beslissingsboom krachtiger te maken. Zo zien we dat zelf één route al een voorspellende waarde kan hebben voor een antwoord van een eindgebruiker. We kunnen concluderen dat er een duidelijke synergie is tussen een datagedrevenbenadering en actieve interactie om de uitkomst van *crowdsourcing*-applicatie te versterken. Deze versterking zal zowel leiden tot een meer performant design alsook een verhoogde kwaliteit van de bijdrages van eindgebruikers.

Relevantie en verder toepassingen

We concluderen dat de voorgestelde benaderingen een onmiskenbaar potentieel hebben om recreatieve navigatiediensten beter te maken. Daarenboven is het duidelijk dat de verworven inzichten niet alleen een meerwaarde hebben om het probleemoplossend vermogen van kaarten te verbeteren; personalisatie van *social media*-platformen, zoals een platform waarop recreatieve routes gedeeld worden, is belangrijk in elk facet. Zo is het ruimtelijke aspect in de beschreven contextuele analyses is een aspect dat kan gebruikt worden om de interactie tussen gebruiker en platform persoonlijker te maken. Ook het aanbieden van content gefocust op een regio waarin een gebruiker reeds actief is, kan voor een sterkere emotionele relatie tussen gebruiker en platform zorgen. Verschillende technieken die beschreven zijn in dit proefschrift kunnen gebruikt worden om deze *region of interest* te bepalen. Daarnaast willen we ook wijzen op het potentieel van actieve interactie om eindgebruikers te sensibiliseren waarvoor hun gedeelde informatie gebruikt wordt. Er is een toenemende vraag naar privacy en transparantie tijdens het hergebruiken van persoonlijke gegevens. Benaderingen zoals beschreven in hoofdstuk 4 kunnen hierbij een belangrijke rol spelen.

Curriculum vitae

Kevin Baker (°1987) obtained his master's degree (*magna cum laude*) in Geography in 2010. Before starting his PhD in 2013, he worked at the company RouteYou focusing on geomatics and geographic information systems. He received a Baekeland mandate funded by Flanders Innovation & Entrepreneurship (VLAIO) and RouteYou to start his PhD at the Department of Geography (Ghent University). During his PhD, he was involved in several project-based work both in academia and industry. This resulted in a broad and hands-on experience in solving technological problems with a geographical context.