MASTER

Can social media be used to map points of interest?

a case study of HEREs points of interest in Austria and the Netherlands

van Krugten, L.T.F.

Award date:
2017

Can social media be used to map points of interest?

A case study of HEREs points of interest in Austria and the Netherlands

In partial fulfillment of the requirements for the degree of
Master of Science
In Innovation Sciences

Eindhoven University of Technology

Submitted by:
L.T.F. van Krugten (0740976)

Supervisors:
Dr. B.M. Sadowski (Innovation Sciences, TU/e)
Dr. A. Nagy (Product manager, HERE)
Dr. Z.O. Nomaler (Innovation Sciences, TU/e)
Dr. C. Castaldi (Innovation Sciences, TU/e)

Eindhoven, 31 January 2017

Title:
Can social media be used to map points of interest? A case study of HEREs points of interest in Austria and the Netherlands

Keywords:
Big Data, Social Media Data, Value, Points of Interest

Submitted by:
L.T.F. van Krugten (0740976)

Supervisors:
Dr. B.M. Sadowski (Innovation Sciences, TU/e)
Dr. A. Nagy (Product manager, HERE)
Dr. Z.O. Nomaler (Innovation Sciences, TU/e)
Dr. C. Castaldi (Innovation Sciences, TU/e)

Place, date:
Eindhoven, 31 January 2017

Preface

This report presents my master thesis project for 'Innovation Sciences' at Eindhoven University of Technology. During my study, I became interested in innovations and the process that enables them. Specifically, the opportunities of data have gained my interest. I am happy that I had the opportunity to conduct my master thesis on this topic.

I am using this opportunity to express my deepest gratitude to Dr. Bert Sadowski who has guided me through this project. I really appreciate his support, advice and useful feedback at all steps of this research. Also, his research ideas and input were essential for the development of this study.

Furthermore, my deepest gratitude to Dr. Akos Nagy who has always provided useful input and guidance on the project. I really appreciate the efforts to provide me with all the resources and especially with such interesting data. Additionally, his scientific and empirical insights were vital for combining theoretical and empirical results.

I would also like to thank Dr. Onder Nomaler for his feedback on the thesis and research design particularly. Moreover, thanks to Dr. Carolina Castaldi for the input on this study.

Furthermore, I would like to thank the product management department of HERE. Specifically, thanks to Ms. Sowmya Gopal, Mr. Stan Heuvink, Mr. Huib Kalb, Ms. Mai Nguyen, Mr. Martijn van de Runstraat and Mr. Emil van de Ven. Thanks for their support and good time at the company and department.

I also really appreciate the research opportunities and internship provided by HERE. Apart from my thesis, they gave me the opportunity to gain practical experience as well. I see this as very valuable for my professional development.

At last, I would like to thank my friends and family for always supporting me. Particularly, I would like to thank my parents, who always supported me and gave me the opportunity to complete two master programs.

I am glad that I had the opportunity and support for learning so much in such an interesting area. So dear reader, I hope you will have fun by reading this report and that you will start to explore the interesting world of data.

Thanks,

Laurie van Krugten

Abstract

Technological developments have enabled fast processing of Big Data. Several researchers have recognized the large potential of data in solving societal problems, such as education or health. Some even state that there will be a data revolution, such as the industrial revolution. Hence, it is forecasted that data can add high value to society and businesses. Especially, the usage of data beyond its initial purpose is an emergent field that is expected to add high value. Despite the high expectations, there is limited theoretical and empirical insight into this value addition of data. Studies have addressed the value of data for providing information, but the highest value expectation is for data as a product. However, this value is not explored yet. This research contributes by providing insights into the added value of data as a product, which can help to unleash the value potential of data.

The aim of this research is to provide a value framework that can assess the added value of data as a product. Additionally, it aims to provide empirical support for this framework. For this support, the study focuses on the value of social media data for mapping points of interest (POIs). Particularly, the study uses HEREs points of interest and Yellowstone as the focus. HERE is an established location service provider and Yellowstone is a confidential name for a well-established social media source. Social media data is expected to contain value for this product due to the usage beyond its initial purpose and the volume, velocity, variety and veracity of the data.

The value framework is developed by a theoretical study that indicates that the product advantage determines the value addition. The outcome of this study shows that value is closely linked to customer value and therewith product advantage. For POIs, this product advantage consists of the combination of quality aspects and usage. Specifically, quality consists of five determinants: coverage, positional accuracy, thematic accuracy, freshness and richness. The quality improvement by aggregating Yellowstone and HEREs places data is explored by means of descriptive statistics. Thereafter, three binary logistic models are used to indicate the relation between quality improvements and usage of HEREs POIs.

Yellowstone data largely adds value to HEREs places data by confirming the correctness of attributes at highly used places. The improvement of Yellowstone is particularly large at Eat & Drink and Going Out-Entertainment places. Additionally, the city center of Linz, The Hague and Utrecht are largely consistent between Yellowstone and HERE. To a lesser extent, Yellowstone data provides value by coverage and richness improvements of low used places. Coverage and richness can mainly be improved at areas surrounding the city center. Overall, Yellowstone adds value to HEREs places data by improving coverage, positional accuracy, thematic accuracy and richness.

This study shows how social media data delivers value for location service providers in mapping points of interests. Therewith, it shows that the usage of data beyond its initial purpose can add value to a product by improving the product advantage. The additional value of data can be obtained by reusing and aggregating data beyond industry boundaries.

The main theoretical implication is the method used in this study to assess the added value of data. This study clearly shows how the product advantage can be used to determine the value of data as a product. Future research can extend this study by assessing cases with other data and products. Furthermore, a practical implication is that HERE is recommended to aggregate Yellowstone with its places data because it adds value. HERE should not use it as a substitute but rather as an addition to its current places and attributes. Therewith, coverage and richness can be improved. HERE can also use Yellowstone to confirm its current geographical coordinates, address, city and category. The method that matches POIs and combines information of data is key to this process. Overall, reusing and aggregating data are value extraction techniques that unleash the additional value of data.

Table of Contents

List of figures

List of tables

Glossary

Attributes
Human-defined information about a POI, such as name, address and category.

Coverage
Completeness of the database.

Freshness
Speed at which changes of the ground truth are captured in the (digital) map.

Ground truth
The observed world.

HEREs places data
The database of HERE with its points of interest.

Location-based services (LBS)
Software-service that consider the geographical position in providing features.

Location data
Data that contains all aspects of mapping geographical locations (e.g. POIs, roads and lakes).

Location service providers
Companies that provide services based on locations (e.g. HERE)

Places data
Data that contains points of interest.

Point of Interest (POI)
Geographical point location that can be geographically displayed and is identified by human-defined attributes.

Positional accuracy
Correctness of mapped geographical location compared to the ground truth.

Quality
The extent to which customers are satisfied according to their expectations.

Richness
Breadth of attributes.

Social media data
Online technologies that enable sharing of all kinds of information via virtual networks.

Thematic accuracy
Correctness of attributes compared to the ground truth.

Value
The importance, worth, or usefulness of something.

Volunteered geographical information (VGI)
Geographical data that is produced by volunteers that are not necessarily geographical experts.

Yellowstone
Confidential name for a well-established social media.

1 Introduction

1.1 Current situation

Current technological developments have resulted in an enormous increase in the volume of data. The Internet, sensors and Internet of Things (IoT) are generating more data than ever. It is estimated that the Internet of Things (IoT) will have 26 billion units installed by 2020, which can all deliver data (Gartner, 2014). Technological developments have enabled easy and cheap storage of data, fast transportation and information computation techniques are developed for analyzing this large volume (Hilbert, 2016). The large increase of data is referred to as Big Data. Although, there is no agreement on the definition of Big Data yet, it is often defined by three V's; volume, velocity, variety (Gandomi & Haider, 2015). Others also define veracity and value as characteristics and thus data can be characterized by five V's. Volume concerns the size of the data, velocity the speed at which data is obtained and analyzed, variety the heterogeneity, veracity the trustworthiness and value the usefulness of the data. The value of data is essential for the success of Big Data in solving societal problems and creating competitive advantage for businesses.

An emergent field that generates much data is the Web 2.0 technology. Web 1.0 is characterized by a one-to-many relationship (Wyrwoll, 2014) where organizations can publish their information (Chen, Chiang, & Storey, 2012). In contrast, Web 2.0 is characterized by a many-to-many relationship where user-generated content data is created. Although a clear definition of user-generated content data is missing in literature, it is clear that the content is an interaction of and published by users. For individuals, this has shifted web usage from consuming information to publishing data. Additionally, the data reveals information about customers' needs, behaviors and helps identifying new business opportunities (Chen et al., 2012).

Social media data, which is considered as user-generated content data, largely increases the volume of generated data. Social media digitizes our relationships, attitudes, sentiments and provides extensive knowledge about human behavior (Mayer-Schonberger & Cukier, 2013). It is evident from the following facts that the V's of Big Data can also characterize social media data. Social media data has a large volume. Facebook has 1.13 billion daily users with 233 million daily users in Europe (Smith, 2014). More than ten percent of the world population uses this social media channel (Mayer-Schonberger & Cukier, 2013). The velocity of social media data is also high. Twitter has roughly 6000 tweets posted every second (Internet Live Stats, 2016). Furthermore, social media has a high variety. It is often unstructured (Polous, Freitag, Krisp, Meng, & Singh, 2015), meaning that it is not organized in a pre-specified manner. Additionally, it contains heterogeneous information, such as time, location, visual and textual information (Polous et al., 2015). The veracity of social media could be considered as low since social media can and is also used to spread rumors (Papadopoulos, Bontcheva, Jaho, Lupu, & Castillo, 2016). Hence these first four V's of the definition are clearly assessed. However, it is unclear what value social media data can add to society and businesses.

1.2 Problem statement

As can be concluded from above, social media data has largely increased. However, its added value is unknown, while this is crucial for success. It is expected that Big Data contributes to societal problems in areas such as health care, climate control, energy consumption, clean water and education (Irving, 2014). Especially, it is expected that location based data has a significant value (Lin et al., 2016) and that social media could contribute to this data. The importance of location data and particularly points of interest is elaborated below.

The importance of location has increased in many industries. A clear-cut example is the mobility industry. Mobility is an important area for economic development (European Commission, 2012) and connectivity of people (Wulfhorst & Klug, 2016). Despite of the importance, transportation is an increasing concern for society due to produced emissions, increased population in cities and safety. Hence, location content can significantly contribute to solutions. Location content can contribute to emission reduction by mapping efficient routes including public transit. This route could consider real-time information, such as traffic jams, accidents and events. Additionally, many organizations believe that location-based data contains value. Additionally, the volume of location-based data is increasing (Lin et al., 2016).

Points of interests (POIs) are essential elements of location content, nevertheless maintenance is a challenging task. A POI is physical point location that can be displayed geographically and is identified by human-defined attributes, such as organization name, building name, address and category (Chuang & Chang, 2015). POIs contribute to location data in searching and moving to a preferred spot. Ying, Kuo, Tseng, & Lu (2014) state that POI services can contribute to improving smart urban living, which concerns the problems mentioned above. Even though POIs gain increasing importance, its maintenance is essential but challenging. Maintaining the quality of POIs is a challenging task since real-world changes should be timely and accurately mapped. The ground truth of POIs might change due to openings, moving, or closing of POIs (Chuang & Chang, 2015). Location service providers, such as HERE, license data to map and maintain POIs. This data, which is mainly not user-generated data, is well suited for permanent structures without frequent changes. However, it is less suitable for ephemeral places, such as restaurants or hotels (Kelm et al., 2013). Therefore, quickly mapping POI transformations is a challenging, but essential task for location data.

Hence, location data is important for societal developments. Currently, communities are leveraged to maintain location data. For example, HERE introduced Map Creator where everyone can edit their location data. However, these tools are not in place yet and the trustworthiness is arbitrary. Therefore, other sources are still important for maintaining location data. Kelm et al. (2013) state that social media data can contribute to licensed, non-user generated data by improving coverage and reducing biases. Additionally, McKenzie, Janowicz, & Adams (2014) state that aggregating data, such as location-based social media data, can improve the number of POI attributes.

Even though social media data is expected to be valuable, there is no general framework to define the added value of data yet. Value is often measured with financial measures such as ROI, profit contribution or sales volume (Booz, Allen, & Hamilton, 1982; Griffin & Page, 1996). However, these measures are not available before transforming data into a product or service. Therefore, there is currently a lack of understanding of the added value of data and an assessment method of the value is missing.

Concluding, value is an essential aspect for the success of data in solving significant problems. Particularly, social media and location data are expected to have a high value. Social media data is expected to significantly contribute to map POIs, which are an essential aspect of location data. However, there is no method to assess the added value of data yet and an understanding of the value is lacking.

1.3 Research questions

Based on the current situation and problem description described above, this research examines the value of social media data for mapping and maintaining POIs. The study explores an empirically observed trend and aims to gain a deeper understanding of the current phenomenon. Therewith, it combines empirical observations with theoretical developments on this subject. The theoretical insights are empirically examined by applying them to a case that is introduced below. In summary, the main question is answered by means of four sub questions that each provide insight into an aspect of the main question. The first two sub question theoretically underpin the main question and the last two sub question produce empirical results.

Main question:
How can value of social media data be obtained for location service providers in mapping points of interest?

Sub Questions:
1. Why can social media data deliver value for location service providers in mapping points of interest?

This sub question elaborates on the reason for conducting this research. It is largely believed that data is highly valuable but theoretical and empirical support are lacking. This sub question combines empirical observations with theoretical developments and links it with traditional literature about data analysis, new product development and value. Therewith, it provides an understanding of why social media data can add value for mapping points of interest (see section 2).

2. How can the value of social media data for mapping points of interest be measured?

Since a framework for assessing the added value of data is missing, sub question two aims to provide this framework. This question uses the theoretical insights developed by sub question one for the development of the conceptual framework. Section 3 introduces and explains the conceptual framework.

3. To what extent does social media data change quality of HEREs places data for mapping points of interest?

Sub question three applies the first part of this conceptual framework to the case of this study. Therewith, it measures the extent of quality improvement by social media data. Namely, quality appears to be part of the value assessment framework and therefore an elaborated overview of this aspect is provided in section 5.

4. To what extent does social media data deliver value to HEREs places data for mapping points of interest?

Sub question four extends sub question three by considering the relation between quality improvements and usage of POIs. This combination represents value according the conceptual model. Section 6 elaborates on the extent of the added value of social media data to HEREs places data in mapping points of interest.

Together, these sub questions answer the main question by enlightening all essential aspects of this question. The first two questions result in a value assessment framework that can indicate where social media data adds value for mapping points of interest. This framework is scientifically underpinned by literature. The framework is applied in a two-step approach. First, quality improvements are examined and thereafter these results are related with the usage to reveal the added value. These questions create insight into where value is added. Concluding, value can be obtained by implementing the social media data in valuable areas that are revealed by the developed value assessment framework.

1.4 Aims and scope

Firstly, this research aims to develop a framework that can measure the added value of data for a product. In current literature, a framework that measures the value of (social media) data is missing and this research contributes to the body of literature by developing this framework. This framework can create an understanding of the value addition of data. This insight is helpful for unleashing the value potentials of data.

This research focuses on the aggregation of data as a product since it is expected that this method results in high value. The aggregation of data can deliver a higher value than the value of its parts (Mayer and Schonberger, 2014) and enables information to cross the traditional industry boundaries (Parmar, Mackenzie, Cohn, & Gann, 2014). It is also stated that aggregation can increase POI attributes (McKenzie, Janowicz, & Adams, 2014). Additionally, value of data as a product instead of information retrieval is explored since this is expected to result in a higher value (Huberty, 2015). Thus, this study focuses on the value addition by aggregating data for a product.

Secondly, an important aim is to create empirical support for this framework. Literature provides statements about value but theoretical and empirical support are lacking, while this is essential for exploring the value. Therewith, it also aims to provide valuable business insights. The framework can reveal useful insights for the product strategy of POIs. Particularly, it can help to address the maintenance challenges by showing the contribution of social media data and assessing the quality of POIs. These quality insights reveal the quality change for categories and geographical locations, which are important inputs for developing a strategy. Therewith, this study also contributes to practical challenges of location service providers.

The value assessment framework is applied to a case study of Yellowstone and HEREs places data in Austria and the Netherlands. Yellowstone is a confidential name for a well-established international social media website. As shown previously, social media data is rapidly increasing and the value is unknown yet but it is expected that it has high value with great potentials. HERE is a market leader in the global mapping industry. Since location based data is important for current societal problems (see section 1.2) and expect to be valuable for many organizations, it is used for this research. Particularly, points of interest have an important role, while their maintenance is challenging. Furthermore, the study focuses on two countries in West-Europe since data was available for these geographical areas. The two countries were selected based on local knowledge and business value. An elaborated introduction of the case study is provided in section 4.2.

1.5 Conclusion

The volume, velocity and variety of social media data has largely increased due to technological developments. Social media data is expected to have a high value addition for society and businesses. In general, researches announce a Big Data era with enormous societal impacts. However, there is no framework that measures the value of data as a product yet and theoretical and empirical support is lacking in literature. This gap is addressed by this research since it develops a value assessment framework and empirically applies this model. The model is applied to the value of social media data for mapping points of interest. Explicitly, it is assessed whether Yellowstone can add value to HEREs places data for mapping points of interest. Point of interest mapping is a data intensive and challenging task that is essential for location data. Many organizations prospect that location data is valuable and increasing and hence it is a relevant case study. Concluding, results provide a value assessment framework and insights into how value can be obtained from social media data for location service providers in mapping points of interest.

2 Theoretical study

2.1 Introduction

It is expected that location service providers can obtain value from social media data in mapping points of interest (POIs) and therefore this study is conducted. This section elaborates on theoretical support for this expectation and theories that help to understand and develop the research. The theoretical study aims to explore the general accepted facts, identify and understand established models and identify unresolved problems on the subject of the study (Remenyi, Williams, Money, & Swartz, 1998). The first part of this section reviews the state of the art by exploring studies that use social media data for places data and studies that define value assessment models. The second part elaborates on the theoretical background in which this research is embedded. The theoretical frame consists of three research areas: (1) data analysis, (2) new product development, and (3) value concepts. These three areas are explored and used as scientific support of this research. The subsequent chapter applies the theoretical background to points of interest (POIs) and presents the conceptual model.

2.2 Previous research

Social media data and point of interests

Matching and blending is the first essential step of the process in successfully using social media data for location-based services. The outcome of the aggregated data is depended on this step. Matching is the process of combining POIs in different files that represent the same real-world entity. Blending is the process of merging attributes to increase the richness of a POI. Several studies have focused on optimizing matching and blending algorithms. However, these studies mainly focus on the success of matching and blending algorithms rather than the created value of the aggregated data. This process is expected to enrich, validate semantics and improve coverage of POIs (McKenzie et al., 2014).

Furthermore, some studies have addressed the usage of social media data for location-based services (LBS). The LBS market is an emergent market that is currently growing (Ying et al., 2014). These services leverage user-generated Web 2.0 data and particularly social media data by using its heterogeneity and freshness. The data is heterogeneous since it includes temporal and geographical information in a visual and textual format (Polous et al., 2015). Additionally, social media data is considered as real-time due to the high update frequency. Even though it is described that social media can deliver benefits, an empirical assessment of these characteristics resulting in value is lacking in literature.

Applications in the LBS industry show some practical examples of value resulting from social media data. Event detection is an example of a location-based service. An event is defined by a specific time and place (Polous et al., 2015). Hence, in contrast to a place, an event is temporary. Events are detected and mapped by clustering algorithms that search for unusual high numbers of published messages at a location (Polous et al., 2015). Additionally, disasters, such as hurricane's or traffic accidents, can be identified with social media data. However, information credibility is not assured and therewith inaccurate information and events can also be detected. Generally, inaccurate events can be distinguished (Castillo, Mendoza, & Poblete, 2011). Hence, social media data can contribute to identifying and mapping events. Furthermore, several researches explore the usage of social media data for recommending POIs (e.g. Li, Xu, Chen, & Zong, 2015; Yin & Cui, 2016; Ying, Kuo, Tseng, & Lu, 2014). Providing relevant recommendations is the most challenging part of this service. There is a difference between hometown and out-of-time preferences (Yin & Cui, 2016) and it is important to include social, preferential and popularity aspects when recommending POIs (Ying et al., 2014). Furthermore, solely considering ratings does not result in optimal outcomes and therefore the numeric order of ratings by a user should be considered (Li et al., 2015). Apart from these challenges, it is also recognized that only a fraction of visited places are published online (Sang, Mei, & Xu, 2015a). POI recommendation is still an evolving research field that aims to develop the most effective and efficient algorithm to recommend POIs to users.

These studies provide some insight into the usage of social media data for POIs but no study (according to the knowledge of the author) has researched the value of social media data for creating an accurate and complete map of POIs. This POI map is essential for location-based services. For example, a POI cannot be recommended when it is unknown to the recommending system.

Social media can be categorized as volunteered geographical information (VGI) and some studies have researched the quality of VGI to determine the usefulness of this data. VGI is geographical data that is produced by volunteers that are not necessarily geographical experts (Chuang & Chang, 2015). For example, Open Street Map (OSM) or social media data is VGI data. Recently, the developments of Web 2.0 have resulted in a large increase of VGI (e.g. Arsanjani, Barron, Bakillah, & Helbich, 2013; Goodchild, 2007). Studies that assess the quality of VGI are discussed in section 4.2. Nevertheless, these studies use an extrinsic evaluation method that compares VGI with an established source, such as governmental or HEREs data, to determine its quality. The established source is considered as the ground truth, which is the situation in the real world. This has clear limitations. Even though location service providers and governments have high-quality location data, this data is still not flawless. Additionally, VGI data is largely improved and more equal in quality to established sources than earlier. Therefore, the assumption that established sources represent the ground truth is not valid anymore (Antoniou & Skopeliti, 2015). Additionally, collecting information about the ground truth is a very time-consuming and expensive task that needs an alternative. Only a few studies (e.g. Barron, Neis, & Zipf, 2014) have used an intrinsic assessment method that does not use reference data (Antoniou & Skopeliti, 2015). However, a well-developed intrinsic quality assessment framework is far from being established (Barron et al., 2014), which is a gap in the current literature. Additionally, few studies have assessed multiple quality indicators or multiple countries in the same study (Antoniou & Skopeliti, 2015). Furthermore, these studies only consider quality without indicating the value of VGI data for mapping points of interest. This clearly indicates a gap in current literature.

Hence, several recently developed location-based services have obtained value from social media data. Additionally, some researches have defined value aspects of social media data. It is stated that heterogeneity and freshness provide opportunities and social media data aggregation can enrich, validate semantics, and improve coverage. Contrary, only a fraction of places is published on social media. However, these researches clearly lack empirical support of their statements. There is still no insight into the magnitude of these value aspects. Additionally, studies that cover multiple countries and quality measures are lacking. Therefore, it is still unclear whether and to what extend social media data contains value for mapping POIs.

Value assessment methods

In order to understand the value of social media data, a value assessment method is required. Value assessment methods for (Big) data are examined because no method for social media data is established yet.

Value can be retrieved from the primary and secondary use of data (Mayer-Schonberger & Cukier, 2013). A special characteristic of data compared to traditional products is its non-rivalrous usage. Data can be used for its intended purpose and for a secondary use. The secondary use is a use of the data for which it was initially not collected. Data can be used multiple times without affecting the value of previous usages. Therefore, data has infinite application opportunities, which generally differs from traditional products. These opportunities create many possibilities for customer value by satisfying multiple customers' needs and delivering product advantage. However, it is hard to assess the value of data for secondary use because the use-cases are not defined in advance.

Even though secondary use can provide value, data should be suitable to solve specific problems to deliver value. Huberty (2015) claims that second- and third-order value does not create the high expected value from data, while the first-order of data will. Second-order value is resulting from improved insights into customers and third-order value from advertisement. Most studies have examined the value chain for second-and-third order value. Value can be retrieved from a value chain that transforms data into information and therewith creates knowledge for understanding a phenomenon. This improved knowledge and understanding is considered as value retrieved from data (Amankwah-Amoah, 2016). However, as explained above, this is defined as second-order value that does not provide the largest potential. The first-order is purposefully collected to solve a specific problem and is assumed to have the largest value (Huberty, 2015). This contrasts to the statements of Mayer-Schonberger & Cukier (2013) since they state that significant value can be retrieved by using data differently than the initially intended usage. This research examines whether a first-order value that solves a specific problem can be captured from a secondary use. Social media data is initially not collected for mapping points of interest, but is expected to create value for a relevant problem instead of solely capturing value from customer insights and advertisement.

Apart from the relevant distinction between primary and secondary usage, there are distinct methods for extracting value from data. Mayer-Schonberger & Cukier (2013) define three methods for gaining value from data. Firstly, data can be reused for a different purpose than initially collected. For example, search terms are used to discover preferences and trends. Secondly, the aggregation of data can result in value. For example, Zillow combines data, such as housing prices and neighborhood characteristics, to estimate the value of housing real estate (Inc, 2016). The authors state that the combination of data has a higher value than the sum of its parts. Thirdly, data collection scope can inexpensively be extended for a secondary purpose. For example, Google collected data for Google Streetview and meanwhile collected geographical information for updating their maps and WiFi network names. The value of data is defined as the sum of these three methods. Parmar et al. (2014) define five methods to create customer value from data. These methods are often combined by businesses. Firstly, companies can use connected products that provide information to improve processes and services. Secondly, assets, especially in the creative industry such as books and music, can be digitized and this results in new services and business models. Thirdly, businesses can combine data within and across industries, which is consistent with the second method described by Mayer-Schonberger & Cukier (2013). The authors mention that this pattern provides opportunities for industry border crossing. Fourthly, companies can sell their data to other companies within or across the industry. This is often combined with the third pattern. Lastly, companies can sell their IT systems or software systems as an additional source of revenue (Parmar et al., 2014).

Concluding, authors indicate value possibilities of the primary and secondary usage and several methods for creating value from data. The value chain of data to insights is researched, but this does not have the highest value expectations. Value can also be retrieved from data as a product that solves specific problems instead of only generating information for improved decision-making. This is clarified by the methods of both Mayer-Schonberger & Cukier (2013) and Parmar et al. (2014).

Apart from the value extraction methods, the combination of new product development (NPD) and data is interesting since this study assesses value from data as a product. This is a little explored research area. It is stated that a data-driven environment will affect NPD (Bharadwaj & Noble, 2015). According to Bharadwaj & Noble (2015) the three V's, by which Big Data is often defined, will affect NPD. The large volume raises challenges regarding processing and transforming the data into valuable products. The velocity requires new methods that faster analyze and process data and reduces the product life cycles. Currently, the speed of innovation is increasing in importance for competitive advantage and survival (Nambisan, 2010). Additionally, the variety poses challenges regarding processing the heterogeneity of social media data (Bharadwaj & Noble, 2015). Furthermore, Roberts & Candi (2014) study the value of social network sites for NPD. Results show that it does not create value for marketing research or market growth. It is likely that the full benefits are not captured yet. Social media data provides much information about customers' experiences and can be used to extract value. However, it is not largely researched what information is valuable. Firms do leverage value from the data for the innovativeness of products and launch decisions. The innovativeness can be improved by involving customers in the design process with social media. Chen et al. (2012) marks the field of e-commerce and market intelligence as a critical and high-impact area.

Even though social media data might improve decision-making in the NPD process, NPD of products or services existing of data is little explored. Results focus on information that social media data can provide and therewith improves the decision-making process of NPD. They examine the value chain of data to insights, which is second-order value, instead of developing data products. This study does not consider data as information that improves NPD, but explores the value of data as a product. A product can solve specific problems, which is expected to result in larger a value than information retrieval (Huberty, 2015). According to the knowledge of the author, there is no research that has addressed this topic yet.

Concluding, value assessment is difficult due to the secondary use and several methods for value extraction that do not diminish the value of other usage. Additionally, information creation only results in a small portion of the value, while product development creates a larger value. Studies have examined the value of information retrieval, but there is a gap for value creation from data as a product. It is stated that this will create high value, but scientific and empirical support is lacking. Therefore, this research contributes to current literature by creating a value assessment framework that is theoretically and empirically supported. The value of social media data as a product can be assessed by means of this framework. Therewith, the study provides insight into the contribution of social media data for mapping places and contributes to POI maintenance challenges.

2.3 Theoretical framework

To understand supporting theories for the value assessment framework, this research links three research fields: (1) data analytics, (2) new product development, and (3) value concepts. Data analytic tools are fundamental to process data from a raw material into a finalized product. Data analysis tools are largely researched to retrieve information from data. Only after analysis, data provides information that can be used for several purposes. Therefore, data analysis is the first research field that is covered by the theoretical framework. This research is also theoretically embedded in new product development (NPD) because it focuses on leveraging value from data in a product. The NPD process examines and manages the strategic and tactical implementation of new products. The process is a clear, stage-wise process that transforms ideas into valuable products. It is largely researched and can be used for both incremental and radical innovation. As data can be considered as a raw material, it can be processed into a valuable product by means of this process. Therefore, it is essential to understand the process and success determinants of NPD. Additionally, a link with the broader concept 'value' provides an overview of the embeddedness of success determinants within this broader notion. Together, these research fields provide a theoretical embedded understanding of value. This section elaborates on each of the fields and the implications for this study.

From data to information

Data is a raw material that needs to be processed to create a product or service. Data can be processed by means of two sub-processes, namely data management and analytics. These sub-processes include collecting, extracting, aggregating, modeling, analyzing and interpreting data (Gandomi & Haider, 2015). Often these two sub-processes are not distinguished but considered as data analytics. Since the format is essential for deciding upon tools and analysis methods, the explorative phases of the process can show the format of the data. This format can be structured, semi-structured or unstructured (Gandomi & Haider, 2015). Structured data is presented in relational tables, while unstructured data is characterized by its heterogeneity.

Data analytic tools have transformed from traditional to big data tools due to the volume increase of data (Boyd & Crawford, 2012; Mayer-Schonberger & Cukier, 2013). Earlier, it was often not feasible to collect data about the population and therefore random sampling techniques were developed. These techniques accept or reject predefined hypotheses. Mayer-Schonberger & Cukier (2013) state that Big Data includes the full population instead of the traditional sample. However, others have rejected this notion since the online world does not equal the offline world (Boyd & Crawford, 2012; Hilbert, 2016; Huberty, 2015). The new tools enable research without pre-defining hypotheses. Furthermore, data is often valued for its ability to predict (Mayer-Schonberger & Cukier, 2013). However, information retrieved today does not necessarily represent tomorrows' information (Huberty, 2015). These statements clearly show the transition of traditional to modern analytic tools, but also show the disagreement of authors on these modern tools.

The transition of analytical tools based on the characteristics of data has resulted in several analytical streams. These streams help to position this research into the current developments. Chen et al. (2012) define three large streams in the evolution of Business intelligence and analytics (BI&A). The BI&A field has evolved from BI&A 1.0, which is considering structured content often stored in relational database management systems (RDBMS), to BI&A 2.0, which analyses unstructured data present in Web 2.0. Social media data can be classified in this stream. However, social media data is increasingly mobile and sensor-based, which is assigned to BI&A 3.0. This stream is concerned with analyzing location-aware, person-centered and context-relevant operations. The analysis techniques for these high-volume and fluid mobile and sensor data are underdeveloped and offers large opportunities for research (Chen et al., 2012). This research clearly operates in the BI&A 3.0 stream since it is highly location-aware. Some general data analytic tools exist, but often tools are depended on application area and data characteristics. Mainly (Big) data analytics (e.g. RDBMS data warehousing, regression and association analysis) and mobile analytics (e.g. mobile social networking) are used for this research. Furthermore, this research spans several high-impact application areas. As explained before, location data is essential for many industries and application areas. Hence, (business) analytic tools are largely developed in the past years to obtain the expected value from data in the high-impact areas (see Figure 2-1).

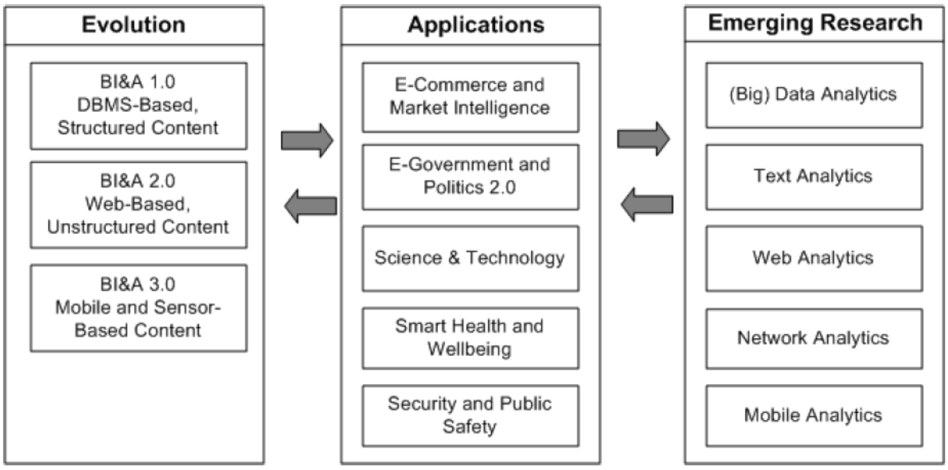


Figure 2-1 BI&A overview: Evolution, Applications and Emerging Research (Chen et al., 2012)

Thus, to transform data into a product, analytics tools are essential. Tools can be selected according objective, format, structure and content of the data. Recently, these tools have developed and largely changed and this research uses several of these developed methods. Namely, suitable analysis tools provide the opportunity to transform data into a successful product. Section 4 elaborates on the analytical tools used in this research.

Aggregating data in a new product

The new product development process explores and manages new products from ideas to profitable streams in several phases. New products can be categorized by their internal and external newness. Internal newness is characterized by newness-to-the-firm and external newness by newness-to-the-market (Booz et al., 1982). For example, an improvement of an existing product is also defined as a new product. The NPD process transforms ideas into value. In this study, the idea of aggregating data is transformed into value. Authors have described several process phases of NPD. Some researches identify three phases (e.g. Eling, Langerak, & Griffin, 2013; Ernst, Hoyer, & Rübsaamen, 2010); idea and concept development, product development and commercialization. Crawford & Benedetto (2014) describe five phases. Although phases differ among sources, most processes are aligned with the structure defined by Booz, Allen, & Hamilton (1982) (Bhuiyan, 2011). These authors define seven phases constructed from interviews (see Figure 2-2). Another well-established NPD process is the Stage-Gate process described by Cooper (2008). This process consists of stages and gates. During the stages information is gathered to reduce uncertainties and risks, and deliverables are produced. Deliverables are assessed during the gates where the project can be stopped or flow to the next stage (Cooper, 2008). Hence, although the division of phases differs between studies, the core of the process is consistent among authors.

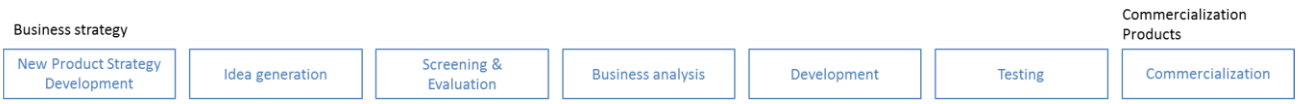


Figure 2-2 New product development process as defined by Booz et al. (1982)

The success and therewith value of the product can be measured at different phases of the NPD process. It is important for this research to examine what determines success and creates value. Mostly financial metrics, such as return on investment (ROI), profit contribution or sales volume (Booz et al., 1982; Griffin & Page, 1996) are used to assess value. However, these metrics are not available before commercialization, which is problematic for assessing value. Additionally, value cannot be fully captured by financial value. Therefore, several other determinants of NPD success are developed. Examples of such measures are the degree of customer involvement (e.g. Roberts & Candi, 2014), team composition (e.g. Ali E. Akgün, Gary S. Lynn, & John C. Byrne, 2004; Sivasubramaniam, Liebowitz, & Lackman, 2012), inter-functional coordination (e.g. Ernst et al., 2010) and speed (e.g. Cankurtaran, Langerak, & Griffin, 2013; Eling et al., 2013).

Measures related to customer satisfaction are often indicated as most significant. Griffin & Page (1996) distinguish between the new product categories according newness-to-the-firm and newness-to-the-market. For all categories, customer satisfaction or degree of acceptance are important success measures. Henard & Szymanski (2001) identify product advantage (superiority over other competitive offerings) and meeting customer needs (perceived satisfying desires/needs of customer) as major measures. Cooper & Kleinschmidt (1987) show that product advantage and particularly product superiority is the most significant contributor to product success. Product advantage includes delivering benefits, product quality, reducing customers' costs, providing innovativeness, product superiority according customers' perspective, and the product should solve a problem. To deliver the largest value, it is essential that the product advantage is relative to competitive offerings. Hence, literature emphasizes that product advantage, including product superiority, is a key determinant of successful new products.

A clear statement about the product advantage is developed in the Fuzzy Front End (FFE). The FFE are the first phases of NPD process showed in Figure 2-2. These phases are crucial for a successful new product (Cooper, 2008; Eling et al., 2013; Khurana & Rosenthal, 1998; Cooper & Kleinschmidt, 1987). The FFE is characterized by a high uncertainty, large flexibility, and few allocated resources (Kim & Wilemon, 2002). The FFE consists of three steps: (1) problem definition based on market insights and opportunity identification, (2) idea generation, and (3) concept development (Eling et al., 2013). These phases are executed before the development phase is entered (Khurana & Rosenthal, 1998; Kim & Wilemon, 2002). After passing a threshold level regarding uncertainty and risk, the process can enter the development phase where more resources are allocated. The development phase results in a product, while the FFE generates a blueprint for NPD. According to Kim & Wilemon (2002), this blueprint consists of time, people and product dimension. The outcome of the product dimension is the identification of opportunities and defining a clear product concept. The product concept clarifies the benefits of the product and the positioning in the market (Cooper & Kleinschmidt, 1987). It is essential for success that the product advantage is clearly defined before the development phase is entered (Cooper & Kleinschmidt, 1987).

Hence, the new product development process is a structured process that consists of several phases. A successful process results in value. During the process, the most significant success measure of products is product advantage, which is defined in the FFE. Product advantage consists of (1) product benefits, (2) product superiority, (3) product innovativeness, (4) ability to solve a relevant problem, (5) product quality, and (6) reducing customers' costs. Therewith, this research acknowledges that a clear formulated product advantage in the FFE is essential for value creation by new product development.

Value concepts presented in literature

Even though success determinants of NPD are explored, the value concept is still not clearly defined by these studies. Value is a broad subject in literature and many concepts are developed that represent value. Concepts can be distinguished by their objective. For example, technology road mapping is a tool that aims to create value by aligning market and technological developments with product or business strategy (Phaal, Farrukh, & Probert, 2004). Other concepts focus on defining customer value, such as the value proposition (e.g. Hassan, 2012) and utility function (e.g. Lin et al., 2016). Apart from differences in objectives, there are also differences regarding analysis level. Value networks assess the value created by interactions between firms, hence this is value on an ecosystem level. Business models are used to create value for the firm (Chesbrough & Rosenbloom, 2002). At last, value can be assessed for products, such as the value proposition and product marketing strategy (e.g. Zott & Amit, 2008). This research studies the value on product level because it concerns aggregated data as a product.

On product level, customer value is generally identified as an important value concept (e.g. Hassan, 2012; Henard & Szymanski, 2001; Stahl, Matzler, & Hinterhuber, 2003; Woodall, 2003). There are several definitions of customer value in literature. Woodall (2003) distinguishes between concepts where value flows from customers to the company (e.g. Customer Lifetime Value), and the company to customers (e.g. value of use). Additionally, four temporal forms that cover from before until after the purchase are defined. In general, customer value is defined as the equation of customer's perceived benefits minus customer's perceived sacrifices (Hassan, 2012; Ivanović, America, & Snijders, 2013). Customers' perceived benefits depend on its superiority and ability to meet customers' need (Hassan, 2012). Additionally, sacrifices can be defined as costs (Ivanović et al., 2013) and as price, effort and risk (Hassan, 2012). Thus, this definition addresses the perceived value when using the product. Customer value can be defined and communicated with the value proposition. This proposition is a statement that clarifies the value that firms deliver to its customers. The statement helps to communicate the value of a product to customers and stakeholders (Hassan, 2012).

Customer value is also positively correlated with shareholder value. Shareholder value is the cash flow produced in a period and the residual value of the firm after this period. Customers actively generate shareholder value by purchasing products or services, attracting new customers and providing information about customers expectation (Stahl et al., 2003). Additionally, product superiority, which is part of the customers' perceived benefit, results in competitive advantage (Hassan, 2012). Hence, there is a positive relation between customer and shareholder value.

Concluding, customer value is the most important value concept on product level. It is positively related to shareholder value and therewith also creates value for the firm. Customer value is an equation of the customers' perceived benefits, represented by product superiority and customers' fulfilled needs, and customers' perceived sacrifices, represented by price, effort and risk. These results are consistent with the NPD success measures identified from literature. The primary success measure of NPD is product advantage, which is related to the customers' perceived benefits and needs. Additionally, product advantage aims to reduce customers' costs. This yields both parts of the equation of customer value. Therefore, both theoretical streams clarify that product advantage is an important determinant for customer and shareholder value.

2.4 Conclusion

Studies state that Big Data can provide high value by its first-order value and its secondary usage. Its first-order value is obtained by solving a relevant customer problem. Both information and products can solve problems, but data is more valuable as a product than as an information source. Therefore, this study focuses on data as a product instead of solely an information provider. This is an important distinction from current literature. Many researches have studied the value chain from data to insights for decision-making, which is the second-order value. However, the first-order value is expected to provide a higher value. This value can be obtained by the primary usage, but the secondary usage provides vast opportunities. By a secondary use, data is used for which it was initially not collected. Literature describes several value extraction methods. However, there is currently a gap in literature that investigates the magnitude of value for data as a product.

Value concepts show that value on a product level is mainly determined by customer value. The product advantage and particularly product superiority are determinants for customer value. These determinants for customer value are also the main success determinants of NPD success. Hence, value is mainly represented by customer value, for which product advantage is essential. An increased product advantage is related to a higher customer value and therewith provides insight into the value magnitude of social media data.

Since this research focuses on points of interest of HERE, outcomes of the literature study indicate that HEREs customer value is essential for this study. HERE is mainly a business-to-business company in the automotive industry but is currently transforming to an open location platform company that serves many industries and consumers. Therewith, it serves consumers of their website and application, platform users and the automotive industry. Generally, customers license a combination of products (e.g. map, navigation and points of interest), but the usage of the website and application are for free. This study focuses on points of interest and the added value of social media is solely explored for this product. Even though HERE has several customers, the product advantage of POIs is mainly consistent among customers. Since product advantage also contains product superiority, HERE should have a product advantage that exceeds its competitors. An elaborated discussion of the product advantage is provided in section 3.3.

Furthermore, literature explains several methods that can be used to extract value from data. HERE uses several of these methods. Traditionally, location based services intentionally collected information to accurately map POIs. People collected information by walking around or driving cars. However, this is not a sufficient method for quickly capturing changes, which is the main challenge of maintaining POIs (see section 1.2). This traditional method has shifted to a secondary usage of data. POIs are digitally mapped and maintained by data that provides information about places. Often this data was initially not collected with this purpose and thus this is a secondary usage of the data. Hence, HERE has digitized its product and reuses, aggregates and sells data, which are all value extraction techniques. The aggregation of Yellowstone fits within the current approach of HERE.

It is expected that social media data can add value to HEREs places data due to several reasons indicated by this literature study. First, POIs have the potential to solve relevant customer problems, such as locating places or exploring nearby places that fulfill needs. Additionally, the social media data is used differently than its primary use for a product rather than information retrieval. Hence, the former indicates that it concerns first-order value of secondary usage of data as a product. Overall, literature states that the highest value from data can be obtained by these methods. Furthermore, several value extraction methods elaborated in literature are used. The social media data is reused and aggregated beyond industry boundaries. Yellowstone could enlarge its value to HERE by extending its current data. In summary, these arguments provide abstract arguments that social media adds value for mapping POIs.

Practically, the V's of (social media) data indicate that it can add value for mapping POIs. Social media data has a large volume and includes many places that could increase the number of mapped POIs. It is stated that the aggregation can enrich, validate semantics and improve coverage. Additionally, the velocity can improve the speed at which changes are captured. This is the main challenge of POI maintenance (see section 1.2). The variety can improve the heterogeneity of captured information. However, the veracity poses challenges since the trustworthiness of the data is doubtful. Thus, these statements indicate that value can be added by social media data for mapping POIs.

Apart from challenges regarding the veracity, challenges associated with the V's of social media data should be properly managed. For this case study, HERE should be able to manage the challenges. The large volume results in challenges regarding processing and transforming the data and the velocity aims for fast analysis. Additionally, processing the heterogeneity due to the variety can be challenging. However, there is potential to obtain value from social media data for location service providers in mapping POIs.

Concluding, it is expected that social media can add value for mapping POIs due to the abstract and practical arguments obtained from literature. This research explores whether this expectation can be confirmed by assessing the added value of Yellowstone to HEREs places data. Therewith, the value magnitude of social media data as a product is explored, which is currently a gap in literature. The magnitude is defined by the enhanced product advantage of POIs. Hence, this research aims to provide an understanding about the added value of data as a product and contributes to literature. Additionally, the scale of the research regarding social media data and geographical area extends current research.

3 Conceptual framework

3.1 Introduction

The theory described in the previous section is well applicable for this research. This section presents the conceptual framework that is developed from constructs and variables presented in literature. This framework provides the model for this research and results in seven main hypotheses. The model is used to assess the added value of social media data to HEREs places data, which is an objective of this study. The section starts with an introduction of the framework and thereafter underpins the model by scientific and empirical studies. At last, hypotheses for this study are introduced and discussed.

3.2 Value assessment framework

As discussed in the theoretical study, the product advantage is essential for assessing customer value (see section 3.3). Namely, an enhanced product advantage increases value. Therefore, the conceptual model of this study is a value assessment framework developed from a review of current studies regarding the product advantage of POIs. Since it concerns added value, HEREs places data is the benchmark and the added value to this base is assessed. The review shows that the product advantage of POIs consists of three factors: (1) quality measures (coverage, positional accuracy, thematic accuracy, and freshness), (2) richness and (3) usage. These three factors are the constructs used in the conceptual model (see Figure 3-1). Particularly, the value is assessed by relating the added quality and usage as shown in the figure. Quality and usage are inseparable because never retrieved POIs do not deliver customer satisfaction. Explicitly, frequently used POIs deliver often customer satisfaction. Therefore, quality improvement and usage are related in assessing customer value.

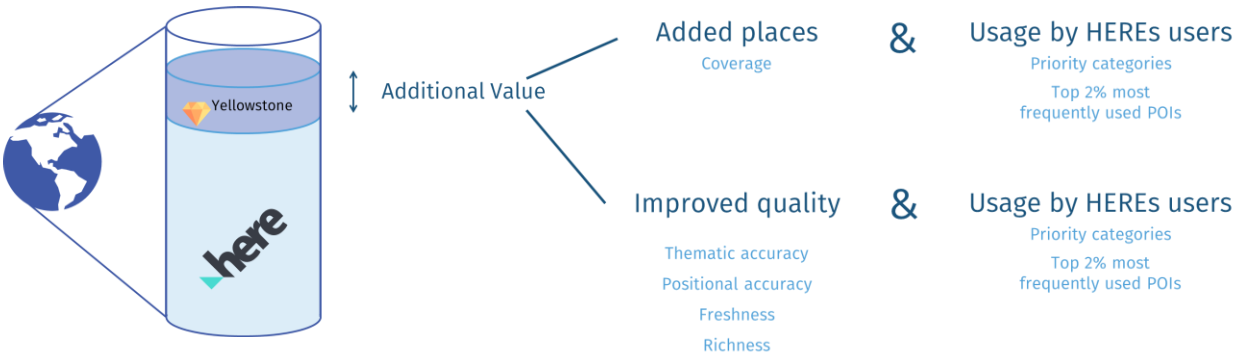


Figure 3-1 Value assessment framework

Table 3-1 clarifies the constructs of the conceptual model. These constructs are obtained from literature (see section 3.3) but are also consistent with the business perspective. HERE uses four quality measures for its POIs: accuracy, completeness, freshness and richness. From a business perspective, positional and thematic accuracy are combined since these have the same assessment approach. However, they measure different aspects and therefore are not combined for this study. Furthermore, literature uses temporal quality, while HERE defines this as freshness. This research is largely empirically focused and therefore the terms defined by the business are used (see Table 3-1). Hence, scientific and empirical findings are consistent in quality measures and confirm each other.

To measure the defined constructs for each POI, they are transformed to quantitative variables by means of literature. The empirical assessment reveals that most studies use an extrinsic assessment method for these quality constructs. An extrinsic method qualifies data by comparing it with the ground truth. Many studies consider an established source, such as HERE, as the ground truth. However, this assumption is not well grounded since no source can perfectly represent reality on a large scale yet. A direct comparison with the ground truth is time-consuming, expensive and often not feasible as was discussed in the literature study. Therefore, an emergent research field studies intrinsic variables that reveal the quality without comparing directly to the ground truth. Intrinsic variables for the presented constructs are showed in Table 3-2. These intrinsic variables reveal the added quality of Yellowstone to the established HERE places data. For every construct, the variable is elaborated and discussed below.

Table 3-1 Definitions of constructs that are used in the value assessment framework

Construct	Definition
Coverage	Completeness of the database
Positional accuracy	Correctness of geographical location
Thematic accuracy	Correctness of attributes
Freshness	Speed at which changes are captured
Richness	Breadth of attributes
Usage	Importance of POIs based on their usage frequency

Table 3-2 Value constructs and variables of the value assessment framework. Furthermore, the right column presents the sources that justify the variables.

Construct	Variable	Source
Coverage	Added places of social media to HEREs data	Barron et al., 2014
Positional accuracy	Geodesic distance between social media and HEREs data	Barron et al., 2014; Mullen et al., 2014
Thematic accuracy	Matching rate attributes of social media and HEREs data	McKenzie et al., 2014
Freshness	Days until last update of social media and HEREs data	Girres & Touya, 2010
Richness	Added attributes of social media to HEREs data	Barron et al., 2014
Usage	Priority categories and high database retrieval frequency areas of HEREs POIs	-

Coverage

Coverage or completeness is the extent to which reality is captured, meaning that missing and redundant POIs are identified (Girres & Touya, 2010). For extrinsic evaluation methods, this is often measured by comparing matched and unmatched items of VGI data with an established source and therewith calculating a completeness ratio (Chuang & Chang, 2015; Girres & Touya, 2010). The error of commission indicates redundant POIs, while the error of omission indicates missing POIs (Barron et al., 2014). Redundant POIs are places that should not be in the data (e.g. a closed restaurant).

Barron et al. (2014) developed an intrinsic assessment for OSM data. A large decrease in the number of additions could indicate that the data reaches completeness. Furthermore, an increasing number of points indicates that the data is closer to completion (Barron et al., 2014). Since the first indicator is not applicable for social media data, the second indicator is used in this study. This represents the addition of places to the benchmark (HEREs places data).

Hence, a binary variable is created that indicates whether a place in Yellowstone is a new place in HEREs data. Generally, coverage is a quality aspect of the data instead of a single POI. However, the proposed variable can indicate the coverage improvement on a POI level, which was preferred for this study. New places do not necessarily represent an increased coverage. Increasing POIs without tracking and deleting closed POIs does not indicate a coverage improvement. If the data is fresh and therewith captures these changes, an increase is a satisfying proxy for coverage improvement. Since HEREs data is maintained, this proxy is used as an indicator.

Positional accuracy

A new place that is added by social media data always increases the remaining quality measures. Therefore, proxies for these quality aspects are only used for places that are present in both Yellowstone and HEREs data. Hence, these quality measures are improvements rather than new place additions (see Figure 3-1).

Positional or geometric accuracy is the correctness of the geographical location compared to the ground truth (Girres & Touya, 2010). This measure is often used in studies and measured by calculating the distance between the VGI data and an established source. The ideal measure to assess the added quality would be the changed distance to the ground truth by aggregating Yellowstone. A decreased distance would indicate a positional accuracy improvement. This can be calculated for example by a buffer analysis (Arsanjani et al., 2013) or Euclidean distance (Mullen et al., 2014). However, these measurement methods need a source that correctly represents the ground truth, which is challenging and expensive.

An intrinsic proxy is the developments of polygons or coordinates over time (Barron et al., 2014). In this study, the developments over time are not available but the difference between sources is used as an indication of the positional accuracy. Just as developments over time can verify each other, sources can also confirm each other. Mullen et al. (2014) used the Euclidean for which a transformation of coordinates is required for the spheroid surface of the Earth. This study uses the geodesic distance that does not require this transformation. Both methods measure the same distance. To indicate the added value of social media data, the distance between this data and HEREs data is calculated as a proxy for positional accuracy. Ideally, one would like to have three sources for verification since two consistent sources can also be incorrect (see Table 3-3). However, only two sources are available for this research.

The geodesic distance between Yellowstone and HEREs places data is used as a proxy for positional accuracy. This indicator defines whether sources verify each other. Small inconsistencies do not indicate incorrectness since points represent polygons of places. Therefore, several distance buffers were assessed. Another limitation is that the benchmark could be incorrect. Inconsistent records were also examined to identify patterns. McKenzie et al. (2014) state that the aggregation of sources can validate attributes. Therewith, the proxy for positional accuracy indicates the likelihood of correctness by considering the confirmation among sources.

Table 3-3 Verification of information with two sources

	HERE: (X_1, Y_1)
Yellowstone: (X_1, Y_1)	Both are right/wrong
Yellowstone: (X_2, Y_2)	One is wrong, other is right; both are right/wrong

Thematic accuracy

Semantic, thematic and attribute accuracy is the correctness of attributes compared to the ground truth (Girres & Touya, 2010). For example, the correctness of categorization (Arsanjani et al., 2013; Girres & Touya, 2010) or name (Arsanjani et al., 2013; Jonietz & Zipf, 2016) is often assessed by comparing the attribute to a ground truth source. Ideally, the added value of social media data can be assessed by measuring the change in agreement to the ground truth. The correctness can be measured by the Levensthein distance (Girres & Touya, 2010; Jonietz & Zipf, 2016), which represents the steps to change one name into the other. However, it is debatable whether for example 'Restaurant Tol' or 'Tol Restaurant' is the correct name. In many use cases both are correct, but they largely differ in Levensthein distance. Therefore, the Levensthein distance is not used in this study. Additionally, this is an extrinsic assessment, which is disadvantageous.

An intrinsic indicator proposed by Barron et al. (2014) is the attribute completeness, which measures the number of attributes compared to a pre-defined list. However, this does not indicate the accuracy of these attributes, but only the richness. Therefore, this study uses this indicator for the richness rather than the thematic accuracy. The aggregation of sources can validate attributes (McKenzie et al., 2014). According this validation statement and the arguments introduced at the positional accuracy sub section, the attributes of social media and HEREs data are compared to assess the thematic accuracy. Since a correct name highly depends on the use case as explained above, only address, city and category are assessed in this study.

Hence, the matching rate of address, city and category between Yellowstone and HEREs data is used as an indicator for the thematic accuracy. As with the positional accuracy, a limitation is the lack of a third validation source. However, the variable still indicates the likelihood of correctness by considering the confirmation among sources.

Freshness

Temporal quality, which is identical to freshness, is the quickness of capturing changes of the ground truth (Girres & Touya, 2010). In general, there are only a few studies that examine the freshness of VGI data (Antoniou & Skopeliti, 2015). Chuang & Chang (2015) developed a semi-supervised model to detect outdated POIs and therewith assess the temporal quality. This model is developed by an extrinsic assessment method. Ideally, the improved time to capture changes of the ground truth in the database would be an extrinsic measure. Quickly detecting changes in the ground truth and calculating the time that is needed to capture this change in the data would indicate the freshness. However, this method is not feasible due to the high cost and difficulty in detecting changes in the ground truth.

For OSM data, an intrinsic method is to identify POIs surrounded by updated POIs. Since the POI is not changed and its neighbors are, it can be assumed that the updater of surrounding POIs confirms the existence of the POI (Barron et al., 2014). However, this approach is not suitable for social media data or location service providers since POIs are not updated according this method. Girres & Touya (2010) measures the evolution of objects in the OSM data. They correctly note that these evolutions mainly include additions rather than updates. Therefore, they use the capture date and version of objects to assess the freshness of objects. The general assumption is that a more up-to-date object has a higher freshness. This study uses this assumption to create an intrinsic indicator that assesses the changed update frequency by adding social media data.

Hence, a change in the last update date by adding Yellowstone is used as an intrinsic indicator of freshness. This proxy has some limitations. The timestamp representing the last update date does not indicate whether changes are captured. It could be updated with information that is not consistent with the ground truth. However, a high update frequency does indicate the ability to capture changes quickly. Another limitation is that the last update date does not represent update frequency. It could be coincident that a not frequently updated source is just update before the assessment. However, this timestamp still indicates the time until the last update. Based on the assumption of Girres & Touya (2010), this measure is used as a proxy for freshness.

Richness

Richness is the breadth of attributes and therewith closely relates to thematic accuracy. Barron et al. (2014) suggest an intrinsic indicator that measures the number of attributes compared to a pre-defined list. Based on this, the richness improvement is measured by the number of attributes added by social media data. Since this construct concerns a comparison of richness between two sources, a pre-defined list that excludes some obsolete attributes for categories is not required.

Usage

In literature, usability represents the fitness of location data for use cases (Girres & Touya, 2010). Jonietz & Zipf (2016) state that quality should be assessed based on the functionality of the POIs. For example, online check-ins mainly require high semantic accuracy. Also Girres & Touya (2010) mention the importance of considering use cases in assessing the quality and Barron et al. (2014) emphasize the fitness-for-use for intrinsic evaluation methods.

Usability for location service providers is very broad since POIs are used for many cases. The main customers of HERE are automotive businesses that use the data for in-car navigation, customers that use the Open Location Platform and website or application consumers. Apart from current use cases, places data should enable product and service innovations. Namely, location service providers are the base for location services and therewith should enable every location service. Quality objectives based on pre-defined use cases would limit the quality of POIs for location service providers. Therefore, all quality constructs are important for this study.

Instead of defining use cases, the usage of POIs is used as a determinant of the usefulness of a POI. Usage is a significant factor in determining the added value of social media data. Namely, POIs that are never retrieved, do not deliver customer satisfaction since they do not fulfill a customers' need. The quality improvement is not perceived by any customer. An essential aspect of quality is delivering customer satisfaction to users (Jaccard, 2013). Therefore, product advantage is only delivered when the product, hence POI, is used. POIs are used when they are retrieved from the database. For example, a POI is used when it is searched. There are large differences between the usage frequency and importance of POIs for location service providers. Well-known places, such as monuments or airports, have a higher user frequency than relatively small places, such as local fire stations. Apart from the usage frequency, some POI categories might be essential for certain users. For example, these local fire stations could be essential for the fire department using the map to plan their routes. Therefore, HERE has established priority categories that are important for their users. The usability of a POI is the combination of the usage frequency and importance of the category. Hence, an ordinal variable is used that distinguishes low, medium and highly used places.

3.3 Theoretical support

Value proposition of POIs

The theoretical study clearly states that product advantage is essential in value assessment. This statement is used to develop the conceptual model. In this sub section, the product advantage of POIs is elaborated, which provides a justification for the introduced conceptual model.

The product advantage consists of: (1) ability to solve a relevant problem, (2) product benefits, (3) reducing customers' costs, (4) product quality, (5) product superiority, and (6) product innovativeness (Cooper & Kleinschmidt, 1987). Each of these aspects is elaborated below for POIs.

Location data delivers product advantage in many location-based technologies. Location-based technologies, such as smart phones have largely increased (Chuang & Chang, 2015). Already in 2011, the number of wearable devices exceeded the number of computers. Additionally, the introduction of Internet-enabled devices with RFID tags and internet connection have increased the importance of location (Chen et al., 2012). Currently, also other industries, such as the health care industry, are increasingly location-based due to the increase of location-based technologies (Chen et al., 2012).

POIs are crucial elements in fulfilling customers' location-based needs. A POI can help to locate oneself and reduce customers' cost for locating and navigating. For example, POIs can be used as landmarks that help to relate an online map with the offline situation. Additionally, consumers search for POIs to explore and navigate to the place. When buying a product, consumers search for businesses to locate the address, retrieve the phone number or find businesses with a certain product or service (Stirling, 2014). Furthermore, POIs can naturally guide users to a place by using POIs in navigation. Moreover, POIs can be used to share a location with friends. For example, Twitter users can geotag their photos. For all these use cases, it is essential that the POI can be found effectively and efficiently in order to reduce the cost of the customer for searching and retrieving the POI. The use cases clearly show that POIs are essential in many location-based needs and meet the first three points of product advantage.

Hence it is clear that POIs solve relevant problems and contribute to fulfill in customers' location-based needs. However, POIs should have a sufficient quality to deliver product advantage. Quality is defined as the extent to which customers are satisfied according to their expectations (Jaccard, 2013). Thus, quality also captures the previous parts of product advantage since it clearly relates to customers' needs. If features of a product are improved without increasing customer satisfaction (nonconformity), the quality of the product is not improved (Jaccard, 2013). Therefore, quality measures are related to solving problems and providing product benefits. Quality expectations differ between use cases (Girres & Touya, 2010; Jonietz & Zipf, 2016). However, HEREs location based data is used for HEREs Open Location Platform that aims to enable any location-based service in any industry. Therefore, the quality of POIs should aim to meet all location-based needs.

Apart from delivering quality according to the customers' needs, product advantage is offered by exceeding competitors' offerings. The product should deliver more benefits and a higher quality than its competitors and therewith become superior to competitive offerings. In the case of HERE, main competitors are Google Maps and TomTom. Therefore, location service providers should aim to deliver the highest feasible quality. Additionally, more benefits can be delivered. This is associated with the innovativeness of the product, which is point six of product advantage. POIs are an established product that is mapped for years and therefore its innovativeness is low. However, incremental innovation of POIs is vital. As explained in section 1, the maintenance of POIs is challenging. Especially, the speed of capturing changes in the ground truth is challenging, but essential for real-time opportunities. Therefore, incremental innovations of the maintenance process add product advantage. Product advantage can also be improved by offering new features. New attributes could be considered as a new feature and hence as an incremental innovation. Thus, apart from quality measures, richness of a POI also creates product advantage.

The exploration of radical new location-based products or services from social media data is beyond the scope of this research.

Concluding, the six aspects of product advantage are explored for POIs and define clear insights into the customer value. POIs evidently deliver benefits in solving relevant problems of customers by enabling location-based services. Additionally, a superior quality and larger richness results in a higher product advantage of HEREs POIs compared to competitors. These findings are clearly represented by the three important factors of the conceptual model: (1) quality measures, (2) richness and (3) usage. Quality is an essential aspect for product advantage and therewith customer value of HEREs POIs since the definition of quality in this study closely relates to customers’ needs and benefits.

Quality dimensions of POIs

The quality of volunteered geographical information (VGI), which is also social media data, is assessed in multiple studies. VGI has largely increased consumer involvement and has changed map making efforts (Goodchild, 2007). The increased involvement of large communities provides substantial benefits for mapping places. Nevertheless the quality of VGI is a substantial concern (Jonietz & Zipf, 2016). Many authors have studied the quality of Open Street Map (OSM) data as an assessment of VGI quality. Even though motivation differs between generating OSM and social media data (Mullen et al., 2014), quality measures are consistent. Namely, quality measures do not measure motivation but customer satisfaction of POIs. Therefore, quality measures are consistent among these sources.

Numerous studies about VGI data have used the quality measures introduced by ISO 19157, which describes the quality of geographic data (International Organization for Standardization , 2013). This norm identifies six quality measures: (1) completeness, (2) logical consistency, (3) positional accuracy, (4) temporal quality (5) thematic accuracy and (6) usability. Many studies have used these measures with extrinsic evaluations (e.g. Arsanjani, Barron, Bakillah, & Helbich, 2013; Girres & Touya, 2010; Mullen et al., 2014) as is elaborated in section 3.1. Extrinsic evaluations assess the quality by comparing VGI with a well-trusted reference source, such as HERE, that is assumed to be equal to the ground truth (Jonietz & Zipf, 2016).

These recurrent six quality measures are well established in literature, but not all are applicable for this study. Logical consistency is the degree of modeling consistencies, such as usage of integrity constraints (Girres & Touya, 2010). This measure is often not used in studies since it is a characteristic of the processing system rather than the input data (Antoniou & Skopeliti, 2015). It is also not used in this study since it investigates the process rather than quality of a POI. Yellowstone and HEREs data are both ingested according the same logical consistency since only ingested data is used. Namely, logical consistency is a quality measure concerning the design of the database system. In general, location service providers have a good logical consistency due to their strict input requirements. VGI generated maps have a lower consistency due to the higher input flexibility. For example, POIs can be tagged instead of categorized according a fixed, predetermined system. Hence, this quality measure is not on POI level but rather one about the database design and process. Therefore, the other five well-established quality measures are used in this study and represented in the conceptual model (see Figure 3-1).

3.4 Hypotheses

This sub section introduces five hypotheses that are tested in this study. Firstly, expectations about the similarity of popularity variables in the Yellowstone data are described, which is relevant for understanding the social media data and used in the last part of this study. Thereafter, hypotheses that make statements about the relation among quality measures and usage of HEREs POIs are introduced. These are assessed in section 6. Therewith, these hypotheses cover the main research area of this study. Since this research spans an embryonic research area, some hypotheses were developed based on the expertise of the authors.

Yellowstone popularity factors

Yellowstone data contains several variables that indicate the popularity of a place. Therefore, it is expected that these variables are correlated and represent underlying factors. It is useful to reveal those underlying factors because they can be used as control variables in the relation between quality changes and usage. Since no research has addressed this topic, an exploratory factor analysis was conducted and a clear defined hypothesis was not used. Table 3-4 shows solely expectations regarding factors and item loadings based on the reasoning below.

Users of Yellowstone can show their interest, share their location and share their satisfaction of a place. These actions all reflect the activity from users at a certain place on Yellowstone. It is expected that often visited places have a high number of all these variables because some visitors might like the place, while others check-in. Therefore, it is assumed that the number of likes, check-ins, reviews and ratings are positively correlated. These variables are expected to reflect the activity of a place.

Variables are also expected to reveal the activity frequency of users at a page. The number of monthly likes and check-ins, and weekly likes and check-ins is expected to represent activity frequency as the underlying factor. The total number indicates the activity of a place since the place was launched on Yellowstone, while the monthly and weekly numbers indicate the current activity level at the place.

Furthermore, the popularity of its neighbors is expected to be an underlying factor of popularity. The popularity of a place can be predicted by the popularity of its neighbors (Lin et al., 2016). Therefore, it is expected that the popularity of an area is represented by the number of check-ins and number of check-ins within the neighborhood.

Table 3-4 Expectations of item loading on underlying factors of the Yellowstone data

	Activity level	**Activity frequency**	**Popularity area**
Likes	X		
Check-ins	X		X
Reviews	X		
Rating	X		
Monthly likes		X	
Monthly check-ins		X	
Weekly likes		X	
Weekly check-ins		X	
Total neighborhood check-ins			X
Average neighborhood check-ins			X

Coverage

McKenzie et al. (2014) state that matching of several sources can improve coverage. In this study, Yellowstone is matched and blended with HEREs places data. Additionally, Yellowstone can include many places due to its volume. Hence, it is expected that the coverage will improve. In general, it is obvious that an addition is likely to add at least a few new places. Coverage improvement could be related to the category and geographical location of a place. POIs are divided into several categories and due to the difference in nature it is expected that there are differences between categories. Businesses can leverage social media by presenting themselves and communicating with customers (e.g. Roberts & Candi, 2014). Due to these benefits, it is expected that businesses are well presented on social media and coverage is improved in this category. Additionally, places were people gather are well presented on social media since social media aims to connect people. Therefore, larger coverage improvements are expected for these categories. Furthermore, geographical differences of social media usage and current coverage produce different improvements in geographical areas. Areas with a large social media usage could have a larger coverage improvement. Therefore, it is expected that social media usage is higher in city centers and that there is a larger improvement in these areas than in surrounding neighborhoods.

Even though coverage improvement might differ among categories and geographical areas, it is expected that social media significantly improves the current coverage of HEREs places data.

Specifically, a positive relation between coverage and usage is expected. High usage areas of HERE are generally located in city centers. Since it is expected that city centers have a higher coverage improvement, a positive relation between coverage improvement and usage is hypothesized. This would represent a high added value since large improvements are related to highly used places.

H1: Large coverage improvements by aggregating social media data with HEREs places are more likely at highly used POIs.

Positional accuracy

The positional accuracy is a challenging quality attribute of POIs and therefore it is expected that social media data can improve the accuracy by confirming HEREs places data. Social media data has multiple geographical inputs of a POI, which together can indicate the geographical position of a POI. Even though the information credibility of social media data is questionable, it is expected that many geographical inputs provide a good positional accuracy. Therewith, it is expected that categories where people can socially gather and businesses will most largely improve the positional accuracy since these have many check-ins. Check-ins require physical presence and therefore provide geographical locations. Additionally, businesses have advantages of a good positional accuracy because they can leverage social media data by customer communication and advertisement (Roberts & Candi, 2014). The contribution in city centers is expected to be lower than in other areas. City centers have in general more out-of-town visitors than surrounding neighborhoods. These visitors have less geographical knowledge of POIs, which can cause errors in the positional accuracy of Yellowstone.

Overall, social media data is expected to have a good positional accuracy that can confirm HEREs places data. This confirmation results in an improved positional accuracy since the likelihood of correctness is increased. The relation with the usage is discussed below at the thematic accuracy.

Thematic accuracy

Matching and blending data can validate and therewith improve the semantics of POIs (McKenzie et al., 2014). Therefore, it is hypothesized that the thematic accuracy is improved by matching Yellowstone with the current places data of HERE. It is expected that businesses have the highest thematic accuracy on social media since they benefit largely by providing accurate information (see positional accuracy). Specifically, it is expected that categories with less user intervention are more accurate. Socially gathering categories have many user interventions and users do not necessarily have accurate information, while businesses are expected to have this information. In contrast to position, attributes can be changed by users without a physical presence and therewith could be incorrect. Therefore, user intervention is expected to diminish thematic accuracy. Furthermore, a lower thematic accuracy improvement in city centers is expected. In general, there is less local knowledge in city centers than in surrounding neighborhoods as argued above.

In summary, the matching and blending of social media data is expected to improve thematic accuracy by confirming HEREs places data.

Furthermore, it is expected that highly used places are mainly consistent regarding position and information between Yellowstone and HEREs places data. Highly used places are likely to be well maintained due to their importance for several data sources. Additionally, these places are likely included in multiple sources of HERE and therewith have a good overall positional and thematic accuracy. It is also expected that these places have a sufficient accuracy on Yellowstone since they are expected to be frequently used on social media as well. Therefore, the two sources are expected to largely match. Hence, a positive correlation between accuracy and usage is expected.

H2: Positional accuracy confirmations by aggregating social media data and HEREs places data are more likely at highly used POIs.

H3: Thematic accuracy confirmations by aggregating social media data and HEREs places data are more likely at highly used POIs.

Freshness

Social media data is considered as fresh data that is frequently updated (Polous et al., 2015). It has a high velocity as presented in section 1.1. Therefore, the freshness of HEREs places data is likely to be improved by matching it with social media data. POI categories differ on their decay rate and update frequency on social media. Categories that have a high decay rate change fast and it is more challenging to capture changes quickly. However, a clear insight into the evolution of POIs is currently lacking (Lu et al., 2016). This study does not consider the decay rate since there is no scientific or empirical indicator for the rate. For the update frequency, it is assumed that gathering places have a higher update frequency because a high-usage of these places on social media is expected. It is also expected that city centers have a higher update frequency improvement than surrounding neighborhoods because these areas have a higher activity level. Additionally, gathering places are often located in city centers.

Hence, social media data generally improves HEREs places data since social media data has a considerably high update frequency.

However, highly used places of HERE are likely to be included in multiple input sources of HERE and therefore have generally a higher freshness. Since these places are expected to have a high freshness, it is expected that Yellowstone can improve less at these places. This would be indicated by a negative correlation among freshness and usage.

H4: Large freshness improvements by aggregating social media data with HEREs places are less likely at highly used POIs.

Richness

McKenzie et al. (2014) state that data matching can enrich attributes. Aggregating data can match attributes of both sources and enrich the description of POIs with this process. Therewith, the richness can be improved by aggregating Yellowstone and places data of HERE. It is expected that businesses ensure a total profile with many attributes as was explained by the thematic accuracy. Therefore, these places are expected to have many attributes. Therefore, it is expected that the richness improvement is larger for the categories Business & Services and Shopping. Furthermore, it is expected that the richness improvement is not correlated to a geographical location. Thus, city centers are expected to be similar to surrounding neighborhoods.

Overall, social media data can improve richness by adding new attributes to matched places.

Specifically, a negative relation between richness improvement and usage is expected. Highly used places are expected to have many attributes in HEREs places data since multiple sources of HERE are likely to capture these places. Therefore, social media data cannot add many attributes to these places and a negative relation is expected.

H5: Large richness improvements by aggregating social media data with HEREs places are less likely at highly used POIs.

3.5 Conclusion

Based on the literature study, the product advantage and therewith customer value was explored for POIs and established the value assessment framework. The value assessment framework clearly elaborates that value is a combination of quality changes and usage of a POI. To leverage the highest added value of social media data, the most frequently used places or important categories should be improved. Most studies have used an extrinsic evaluation method that assesses the quality of OSM data. This research assesses social media data by means of an intrinsic evaluation method and contributes to the gap in current literature. The intrinsic variables are deduced from empirical studies. Therewith, the framework consists of three important factors: (1) quality measures, (2) richness and (3) usage. Literature shows that five well-established quality measures can be used to assess the quality of POIs. By means of these variables, several relevant hypotheses about the relation between quality changes and usage are introduced. Figure 3-2 shows an overview of the hypotheses tested in this study. Hence, the framework enables a assessment of the product advantage improvement and therewith value addition.

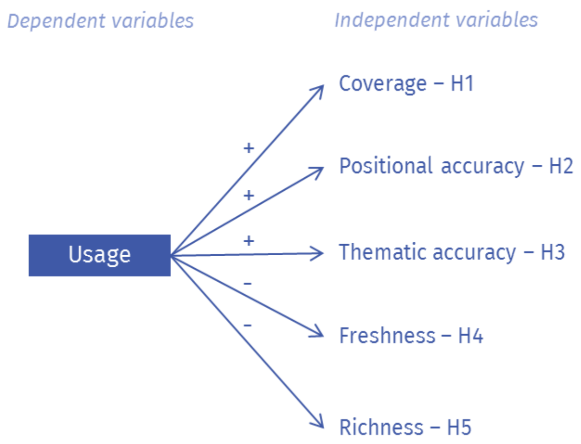


Figure 3-2 Hypotheses of the expected relation among quality changes and usage of POIs.

4 Research method

4.1 Introduction

This section elaborates on the method for revealing the added value of social media data for location service providers in mapping point of interests. The value is revealed by means of a quantitative study that is introduced in this section. Furthermore, the data preparation, sample statistics and statistical tools are presented and discussed.

4.2 Case

This research focuses on places data to reveal the added value of social media data. Therefore, an established location service provider and social media source were selected. These were explored for a geographical area that is significantly larger than current literature, but still achievable in the timeframe of this research. The location service provider, social media and geographical area are introduced below.

HERE is a market leader in the global mapping industry and the location service provider for this study. HERE recently developed the Open Location Platform, which aims to leverage the power of location in our increasingly inter-connected lives. This platform assembles innovators, customers and companies for ingesting, processing and analyzing location data. The location data is provided by their website and application. Additionally, the automotive industry is an important customer that uses location data for their in-car navigation systems and developments of autonomous cars. Due to the increasing online accessibility options of location data, accelerating the development of real-time maps is an important aspect for location service providers. For the development of accurate and real-time maps, the company aggregates data from different sources into a single cohesive unit. Hence, HERE is a well-suited location service provider due to this aggregation approach and the current position in the location industry.

Social media data is retrieved from Yellowstone, which is a confidential name for a well-established international social media. The data has the characteristics of Big Data and spans a large geographical area, which is essential for this study.

Austria and the Netherlands are the selected geographical areas for this study. The Netherlands was selected for the availability of local knowledge. Thereafter, Austria was interesting from a business perspective and geographically differs from the Netherlands. Austria has roughly a population of 8,5 million, while this is 17 million in the Netherlands. Additionally, the density of population is roughly four times higher in the Netherlands. However, Austria hosts more tourist since there were 135 million overnight stays in 2015 (Statistics Austria, 2016), while this is 66 million in the Netherlands (CBS, 2016). Furthermore, the landscape of Austria and the Netherlands largely differ since the Netherlands is flat and Austria has mountains and hills, which affects the livability of areas. At last, the Internet and social media usage differs among countries. In 2015, 84% of the population in Austria had access to the Internet, while this was 93% in the Netherlands. Additionally, 29% of the population in Austria had a fixed broadband Internet subscription, while this was 42% in the Netherlands (The World Bank, 2016). Yellowstone usage is roughly similar among the two countries (Yellowstone, 2016). These differences can result in contrasting outcomes among the countries. Results are assessed for both countries and a country comparison is conducted. Moreover, the five largest cities were examined in each country. Table 4-1 shows that Vienna is significantly larger than the cities in the Netherlands, while the other Austrian cities are significantly smaller. These differences among countries and cities are considered in analyzing outcomes of the statistical analyses.

Table 4-1 The five largest cities of Austria and the Netherlands. These cities are examined in this study.

Country	City	Population
Austria	Vienna	1,863,881
	Graz	280,200
	Linz	200,839
	Salzburg	150,887
	Innsbruck	130,894
Netherlands	Amsterdam	834,713
	Rotterdam	619,879
	The Hague	510,909
	Utrecht	330,772
	Eindhoven	221,402

4.3 Creation of constructs

The precedent section elaborates on the conceptual construct and variables used in this study (see Table 3-2). For each construct, an elaborated description of the created variable is provided below. Each construct is measured on a POI level.

The coverage variable was constructed by a binary variable that indicates whether it is a new place for HERE. Places solely retrieved from Yellowstone and with no other input source of HERE are indicated as new places.

The positional accuracy was determined by the difference between the latitude and longitude of Yellowstone and HEREs places data. Since HEREs places data consists of the aggregation of multiple input sources, the average of these is considered as the position of the places data. This position was compared with the geographical coordinates of the Yellowstone records. By means of the geodesic distance, the distance between the coordinates was calculated. This distance is the value associated with the positional accuracy.

The thematic accuracy was determined by the match rate on address, city and category of Yellowstone with HEREs places data. If one of the other suppliers of HERE contained the same attribute, the attributes matched. Matching of address and city are represented in a binary variable that indicates whether there is a match or not. Since a POI can have multiple categories, the number of verified categories is used as a match rate. For example, a POI with three categories in HEREs places data of which two are verified by Yellowstone obtained a matching score of 2/3. A new category added by Yellowstone is considered as an increase of richness and hence considered by richness. The total thematic accuracy score is calculated by summing the three variables. Therefore, a score between zero and three can be obtained for each POI. This score is used as a proxy for the thematic accuracy.

The freshness was determined by calculating the number of days between file retrieval (07-11-2016) and timestamp. The days of Yellowstone were compared with the days of HEREs places data. Since HERE has multiple input sources with different timestamps, Yellowstone was compared to the minimum number of days for each place. The difference between Yellowstone and HEREs places data is the freshness improvement of Yellowstone to the places data.

The richness was determined by checking whether attributes of Yellowstone were already present in HEREs places data. The address, city, state, postal, category, phone, website, email, opening hours on Friday (to capture both businesses on weekdays and weekends), display latitude and longitude, and routing latitude and longitude were considered for this comparison. Each attribute that was present in Yellowstone but not in the places data gained a value one. The sum of all values is the richness improvement by Yellowstone.

For the usage construct, high-usage areas and top priority categories were used. The high-usage areas were developed by means of the top two percent most frequently used POIs of HEREs website and application. Since new records of Yellowstone are not included in this top two percentage, buffers of these places were created. Buffers of 100 meter were used since Lin et al. (2016) showed that this radius is most suitable to determine the popularity of a place based on its neighbors. Although Lin et al. (2016) studied social media data, it is expected that the popularity of POIs from HERE is also related to this distance. The buffers were spatially joined with the sample and places within a buffer received a point for the usage value. Additionally, priority categories were considered because these represent the high-usage places in the automotive industry. Usage within this industry is not included in the top two percent of most frequently retrieved places and therefore used as an addition to the top frequency. Every POI that was categorized as one of the priority categories obtained one point for usage. The usage of POIs was determined by the sum of these two values (see Table 4-2).

Table 4-2 Usage groups

Value	Definition	Meaning
0	Low used	Priority category NOR high-usage area
1	Medium used	Priority category OR high-usage area
2	Highly used	Priority category AND high-usage area

4.4 Data collection and preparation

This study uses and aggregates data of three large sources: 1) places data of HERE, 2) social media data from Yellowstone, and 3) usage data of HEREs POIs.

The places data is obtained in an XML format from HEREs database at 07-11-2016. The extracted XML data was flattened to CSV files to analyze the data with programs as ArcGIS and Microsoft Access. The data contained roughly 550,000 records for Austria and 1,2 million for the Netherlands. Some records contained a delimiter, which resulted in import flaws of shifted cells. These records were omitted from the data (371 rows for Austria and 825 for the Netherlands).

The places data of HERE is constructed by the aggregation of several input sources. Hence, multiple source suppliers can include similar POIs and therefore the places data can contain a POI multiple times. Additionally, a POI can have several categories and therewith be represented in multiple records. Each POI is distinguished by an unique ID. The combination of the ID of the place, source and category defines unique tuples. Therefore, the primary key is the combination of place, source and category ID.

The places data contained duplicates due to several import time stamps and categories assigned to a place. Duplicates can influence the analysis since some records will be considered several times. Therefore, only the newest record is retained for each duplicate caused by diverse time stamps. The newest record is expected to best represent the data of the source. These filters result in Austria and the Netherlands in 526,748 and 1,155,402 unique tuples, 410,099 and 895,580 unique places and sources (unique place and source ID) and in 313,487 and 741,100 unique places (unique place ID), respectively. This data was used as the places data of HERE.

Other data was retrieved from Yellowstone. Some Yellowstone pages were ingested in the places database of HERE and hence present in the previous dataset. However, this is not the complete

Yellowstone data. Yellowstone data was extracted in a XML format at three timestamps: 23-09-2016, 16-10-2016 and 23-10-2016. These files were flattened by means of a Python script. The latest data contained 366,461 unique pages of Yellowstone in Austria and 934,558 in the Netherlands, distinguished by an ID. Since an overview of attributes over time was desired, the tuples of the three files were joined. Only tuples present in all three files were retained since other tuples would have missing values. Therefore, an inner join on their ID was conducted, which resulted in 342,680 tuples in Austria and 901,605 in the Netherlands.

Yellowstone was ingested in HEREs places data and this set was used as the sample. HEREs places data does not include all Yellowstone data since some Yellowstone records are not POIs. Some tuples represent for example pages of private houses or events. An algorithm developed by HERE ingested and matched only those pages that are presumably POIs. Since pages that not represent POIs should be omitted from this research, only Yellowstone pages ingested in the places data are studied. This sample selection also enables a simple comparison with other supplier sources, but largely depends on this matching algorithm since POIs might be excluded (not randomly) and mismatches can occur. Although the algorithm can be too restrictive or ingest too many pages, it is assumed that this algorithm is sufficient for this study.

This Yellowstone sample has 90,862 records in Austria and 262,468 in the Netherlands. These records represent 85,707 and 249,682 unique places for Austria and the Netherlands, respectively. There are slightly more records than unique places because some places have multiple Yellowstone pages. This sample lacks information about Yellowstone characteristics such as likes because this is not ingested in HEREs places data. Therefore, the POIs were matched by source ID to the initial Yellowstone data described above. In total, 13,002 tuples in Austria and 7279 tuples in the Netherlands were not matched. These tuples were not present in the Yellowstone data from the selected timestamps. From an initial assessment, 29 out of 50 places were private or outdated pages that are not publicly accessible. Therefore, it is likely that these values are not missing at random, but related to properties of the place. Since only public and currently existing POIs should be present in the sample, these tuples were omitted from the sample. Additionally, 426 and 3017 records were not present in every Yellowstone timestamp. This resulted in missing values on weekly and monthly likes and check-ins. An assessment of 25 records revealed that 17 pages were created by solely check-ins without any additional information. These are unofficial places meaning that they are not related to a person or sponsor. Hence, these missing values are likely to represent public POIs and not randomly missing. Missing values seem to be related to the number of likes, reviews and ratings, but a thorough analysis is missing. A sufficient method for replacing missing values is lacking. Since pages largely differ on their number of likes, check-ins, reviews and ratings, the average is not a sufficient replacement. Additionally, values of zero would indicate that there was no activity, which might not be valid. Due to the missing values, these records cannot be included in the statistical analysis and were omitted for the sample. This omission enables a comparison between the sub questions since the sample is consistent among questions. Limitations regarding the omitted results are discussed in section 7. The process resulted in a sample of 77,434 records for Austria and 252,172 for the Netherlands.

Furthermore, 5,555 records in Austria and 17,959 in the Netherlands had missing values for the number of reviews, ratings and average rating. In Austria, five out of 25 assessed records had a review and/or rating. An assessment of 25 records from the Netherlands showed that all these records did not have any reviews or ratings. Therefore, missing values of the number of ratings and reviews were replaced by zero. However, an average rating of zero would not be representative for the place and therefore missing values were replaced by the mean of the average rating (4.5849 for Austria and 4.5575 for the Netherlands).

Apart from the total and monthly/weekly check-ins, the check-ins of the neighborhood were also needed to determine the popularity of the area (see section 4.3). However, this was not included in the Yellowstone data and was constructed. A radius of 100 meter defines the neighborhood area. Lin et al. (2016) tested several radius distances within one kilometer that best represent the neighborhood. Results show that 100 meter provides the best results. Additionally, results of this study show that check-ins of neighbors better predict popularity than likes, hence check-ins of neighbors were used. By means of a spatial analysis with ArcGIS, neighbors for each POI were identified. The total number of neighbors is calculated by counting the number of unique POIs within the radius. So POIs with multiple Yellowstone pages were counted as one. The total and average number of check-ins in the neighborhood was calculated by summing the Yellowstone pages within the radius. These variables were added to the sample.

Within this sample, several missing values and outliers were identified that were omitted before conducting the statistical analyses. Namely, the regression is conducted with a list-wise omission and the factor analysis and regression are sensitive to outliers.

Apart from the missing values described above, the positional accuracy variable had eight missing values in Austria and eight in the Netherlands. Distances larger than 10,000 meters cause these missing values because these were transformed to missing values in the construction process of the positional accuracy. Since these missing values are essentially outliers, these eight records were removed from the sample. This resulted in a sample of 77,426 records for Austria and 252,164 for the Netherlands.

Even though the other quality measures did not have any outliers, other Yellowstone variables had outliers. The number of likes had one large outlier (KLM with 11,405,372 likes) in the Netherlands, which was at least ten times higher than other pages. In Austria, three large outliers were detected (KTM with 904,333 likes, Hofer with 634,646 likes and OFID with 596,937 likes). For the number of check-ins, Vienna Airport (1,123,069 check-ins) and Amsterdam Airport (5,397,599 check-ins) were clearly outliers. Amsterdam Airport had even more than eight times the number of check-ins compared to others and Vienna airport two times. The Efteling, which is an amusement park in the Netherlands, had at least two times more reviews than other records in the sample (The Efteling has 17,073 reviews) and was omitted. In Austria, the Redbull Arena (6536 reviews) and Schloß Schönbrunn (5395 reviews) were outliers with more than two times the number of reviews. The number of likes in a week had one outlier in each country (Tante Fanny with 3,257 likes in Austria and Just Wellness with 22,844 likes in the Netherlands) and the number of check-ins as well (Ski Weltcup Opening Sölden with 2661 check-ins in Austria and Walibi with 12,142 check-ins in the Netherlands). Furthermore, the monthly likes had an outlier in the Netherlands. Shimano road had in one month a diminish of 113,263 likes, which is hundred times higher than any other record. In Austria, Wok Sushi Bar (19,965 likes) and Wiener Wiesn-Fest (12,037 check-ins) were outliers. These outliers were omitted from the sample.

The final sample of 77,416 records in Austria and 252,158 records in the Netherlands was used for the analysis of quality measures.

4.5 Data profiling

This section provides insight into the places included in the sample. Figure 4-1 shows an overview of the sample sizes elaborated above. This figure presents a distinction between new and existing places since positional and thematic accuracy, freshness and richness are only relevant for existing places. Hence, the correlation presented in section 6 only considers new places as is explained below.

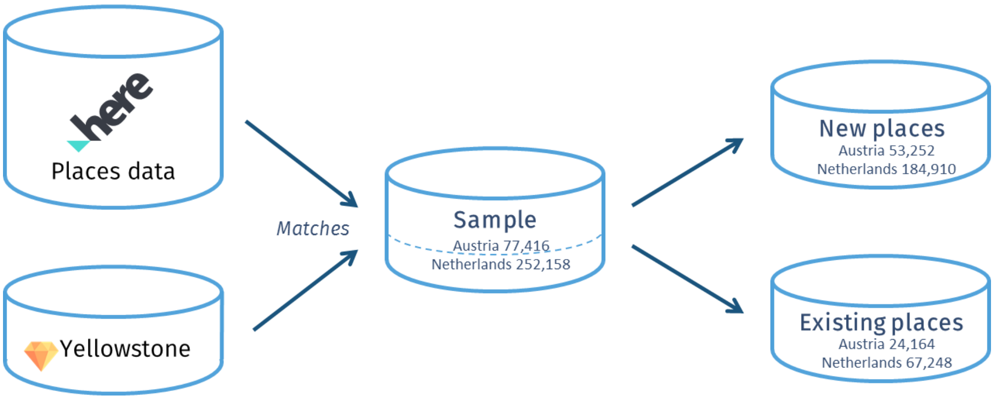


Figure 4-1 Overview of sample

The sample includes roughly a quarter of all places in HEREs places data for Austria and roughly one third for the Netherlands. Therewith, Yellowstone does not match with all places of HERE.

Yellowstone includes many Businesses & Services and least Administrate Areas & Building places. HERE has defined eleven main categories that are showed in Figure 4-2. Apart from the eleven main categories, HERE has two other more detailed levels of categories. However, this research mainly presents the eleven main categories to provide a clear overview and uses the detailed levels for a deeper understanding. Roughly 35% of the places is categorized as a Business & Services of which half are commercial services. This latter is a rest category meaning that these places cannot be categorized in one of the other categories. It is notable that the Business & Services places are also often categorized as Eat & Drink ($perc_{AUT}$=22% and $perc_{NLD}$=20%), Shopping ($perc_{AUT}$=18% and $perc_{NLD}$=28%) and Facilities ($perc_{AUT}$=19% and $perc_{NLD}$=23%) places. Therefore, this category has a large scope and can include many types of places. Both countries have many places in this category, but they largely differ in the shopping category. Roughly a quarter of the sample is categorized as shopping in the Netherlands, while this is only 12% in Austria. However, both countries have roughly 20% Consumer Goods and roughly 20% Hair and Beauty places within this category. Additionally, Austria has relatively more Accommodation and Going Out-Entertainment places than the Netherlands. The accommodation places in Austria are mainly hotels (75%), while the Netherlands has also a significant share of Bed and Breakfasts (15%) and Campgrounds (17%) compared to Austria. For the Going Out-Entertainment places, half of Austria's places is defined as Bar-Pub-Stube-Biergarten, while this is less for the Netherlands. Overall, Yellowstone has relatively many Business & Services, Eat & Drink and Shopping places.

Geographically, Yellowstone data is spread among the two countries. However, cities generally have a higher density of places and this is also represented in the sample.

The analysis also showed some notable things of the data. Some places have several Yellowstone pages representing the place. For example, the Keukenhof has seventeen Yellowstone pages.

Since it has multiple pages, this unique place is several times represented in the sample. However, Yellowstone and quality variables do differ and therefore the several pages are included. Another notable aspect is the negative number of monthly or weekly likes or check-ins. A clear explanation for these notable values was not found but the database design could have created those negative values. A logarithm function would omit these negative values, which is considered below.

Furthermore, Yellowstone attributes were analyzed and compared between countries for additional insight. Generally, pages have the highest number of likes and less check-ins, ratings and reviews. On average, places were rated with 4.5 stars, which is considerably high. Additionally, a comparison between countries showed differences regarding characteristics of pages. Austria has a higher mean of likes, check-ins, reviews and ratings, while the weekly and monthly activity is on average lower than the Netherlands. This could indicate that Yellowstone is generally more often used in Austria, while currently it is more used in the Netherlands. Austria also has higher average check-ins in the surrounding area than the Netherlands, but lower POI density and total check-ins in the neighborhood. This might indicate geographical differences.

All Yellowstone indicators were highly positive skewed and did not have a normal distribution. The variables have a long-right tail and the mean is higher than the median. Since this skewness can affect outcomes of the factor analysis, the logarithm of these variables was constructed. The logarithm produced a stronger transformation than the square root. Additionally, the square was less suitable since this mainly reduces negative/right skewness. For the logarithm transformation, every value was increased by one to include values of zero. The logarithm function cannot transform negative values, which resulted in missing values. However, all Yellowstone indicators should have positive values and therefore this transformation omitted unclear values. Although the skewness and kurtosis came closer to normality with the transformation, the Kolmogorov-Smirnov test still indicated that all variables were not normally distributed. Therefore, both variables were considered in the statistical analyses.

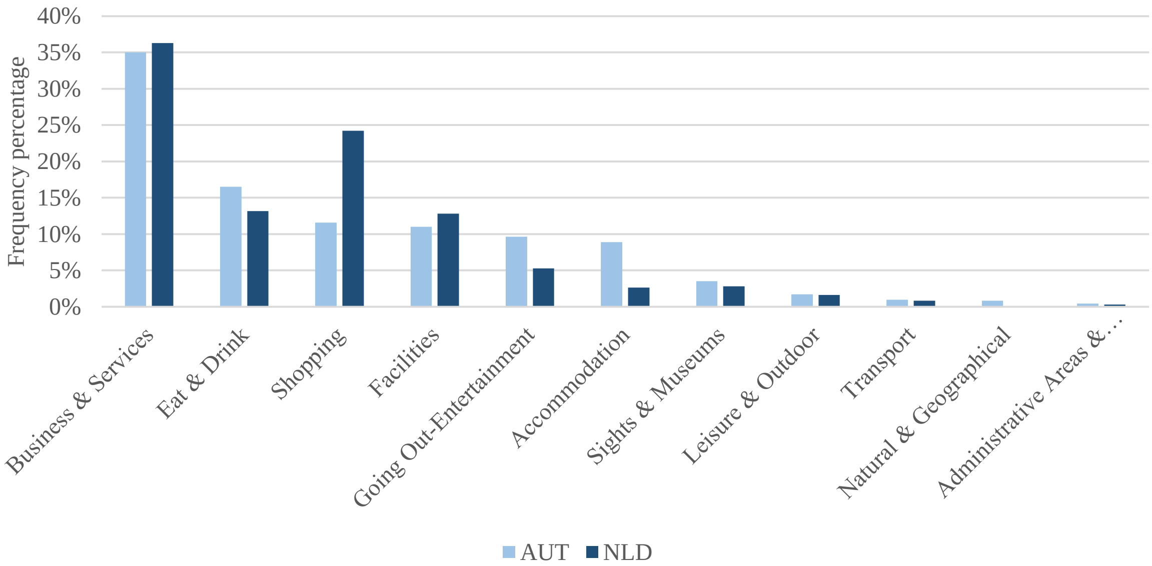


Figure 4-2 Category share of the sample. Sample sizes are presented in Table I-1

4.6 Analytical tools

Sub question one and two were answered by means of a literature study and results are described in section 2 and 3.

Sub question three, which reveals the added quality by social media data, was analyzed with descriptive statistical tools, hot spot analyses and Kruskal-Wallis tests. In general, descriptive statistics were used to reveal the added quality. Apart from a general quality assessment, differences regarding categories and geographical areas were explored. Statistical significant differences between categories were revealed by means of a Kruskal-Wallis since a non-parametric test was preferred. For this analysis, the eleven main categories were used because more detailed categories would reduce the clarity of results. For the geographical analysis, the five largest cities of Austria and the Netherlands were assessed (see Table 4-1). A hot spot analysis was used to reveal statistically significant differences between geographical areas. Assessing distances between 100 and 1000 meters determined the distance band since this scale fits the preferred analysis radius. Eventually, the distance with the highest z-value was selected. A fixed distance band between quality measures for each country was preferred for comparison and an analysis showed that 700 meters for Austria and 900 meters for the Netherlands was most suitable. Furthermore, quality improvements and usage of POIs were examined. To reveal statistical differences between the mean of usage groups, a Kruskal-Wallis was used. Hence, this analysis resulted in a clear overview of the added quality, which answers sub question three.

For sub question four, the relation between usage and quality improvements was assessed to reveal the added value of Yellowstone to HEREs places data. A regression model examined the relation. Usage is the dependent variable and the quality measures are the independent variables since the relation between those variables was examined. The popularity of places on Yellowstone was used as a control variable so the regression controls for the popularity of a place, which likely influences usage. Additionally, the model controls for the country, category and whether a place is part of a chain. Namely, the latter one could influence usage because chains can be used more often due to brand awareness and their quality could be better due to chain sources that HERE uses.

Solely the positional accuracy, thematic accuracy, freshness and richness improvement were included as independent variables, while coverage was excluded. Coverage is a full addition rather than a quality improvement and therewith the listwise omission of records would omit all new places. These records have missing values for the other quality improvements. Therefore, only the four other quality improvement variables were included. Eventually, freshness was also omitted because the data did not provide insight into this quality aspect.

To determine the popularity factors that are used as control variables, a factor analysis was used to identify these factors. Since this study aims to understand the explanation of the structure of the data, a factor analysis was selected. In general, Principal axis factoring (PAF) and maximum likelihood factor analysis (MLFA) are the most popular methods (Winter & Dodou, 2012). PAF is preferred for factors with a few variables (Winter & Dodou, 2012). This is the case for this study (see Table 3-1) and therefore PAF was used. Both oblique and orthogonal rotation were tested. In total, four models were tested with either the transformed variables or untransformed variables (transformation by the logarithm) presented in Table 3-4. All models use the transformed variables for likes, check-ins, reviews and ratings since the transformation of these variables resulted in a significantly closer fit to normality. Model one included the other variables without transformation. Model two and three replaced untransformed variables by the transformed weekly and monthly variables or transformed neighborhood variables, respectively.

Model four uses only transformed variables. Transformed variables were initially preferred because they are closer to normality but also fitness to hypothesized outcomes was considered. Based on these considerations, the best model was chosen. Variable values associated with factors were obtained by using the regression method. These variables were included in the regression as control variables.

A multinomial regression model was preferred since the dependent variable is ordinal. Unfortunately, this regression was not possible due to the processing time of the data. Therefore, three separate binary logistic regressions were conducted according a one-to-one schema. One regression assessed the low and medium used places, another medium and highly used places and the last one low and highly used places. All assumptions for a binary logistic regression are met by the data. Namely, the depended variable is binary, a linear relation between the covariates and log odds is expected, observations are independent and the sample size is sufficient. At last, all relevant variables must be included for an unbiased model to prevent the omitted variable bias. As discussed above, the model likely includes all relevant variables. The coefficients of the regression model were analyzed and a general conclusion about a relation was formed.

4.7 Conclusion

This section clearly elaborates on the method of this research. A theoretical study was used to develop the value assessment framework. For the empirical test of this framework, HEREs places data and Yellowstone were examined for Austria and the Netherlands. This section elaborates on the data preparation and profiles the sample. Furthermore, it explains the statistical tools used in this study. The quality improvements were measured by means of descriptive statistic tools, hot spot analyses and Kruskal-Wallis tests. For the value assessment, three binary logistic regressions were conducted that together provide an insight into the added value of Yellowstone to HEREs places data.

5 Quality change by aggregating social media data

5.1 Introduction

This section elaborates on the changed quality by aggregating Yellowstone with HEREs places data for mapping points of interest. Based on the conceptual model, five quality measures are assessed and each is discussed below in a separate subsection. Additionally, the hypotheses introduced in section 3.4 are examined and discussed.

For each quality measure, the quality change among usage groups was assessed. As explained in Table 4-2, places were distinguished in three groups ranging from low to high usage. Countries, categories and geographical areas differ between their average usage. Roughly a quarter of the sample is located in high-usage areas for both countries. However, Austria has relatively more places in the high-priority categories. In Austria, roughly 40% of the records is categorized as one of the high-priority categories, while this is 26% in the Netherlands. Namely Austria has more accommodations and less shopping places compared to the Netherlands (see Figure 4-2). This category distribution results in relatively more medium used places in Austria than in the Netherlands (see Table 5-1).

Table 5-1 Frequency table of usage groups.

	AUT	% AUT	NLD	% NLD	Total
Low	2678	15%	18767	36%	21445
Medium	11031	60%	20561	39%	31592
High	4725	26%	12847	25%	17572
Total	18434		52175		70609

Figure 5-1 shows that the average usage among categories is similar between Austria and the Netherlands. Eat & Drink places have a significantly higher usage than other categories. Thereafter, Accommodation, Going Out-Entertainment and Transport categories contain often used places. Eat & Drink, Accommodation and Going Out-Entertainment places are frequently located in a high-usage area. Often these places are located in city centers where the density is generally high. In these high-density areas, spatial buffers of highly used places include more surrounding POIs. Additionally, city centers are generally popular areas for using places data since it attracts many visitors. Geographical maps show that highly used places are mainly located in city centers (see Appendix III). This location of categories and high-usage areas could explain the categorical and geographical distribution of usage groups. These distribution insights into usage is used in the following subsections and in section 6.

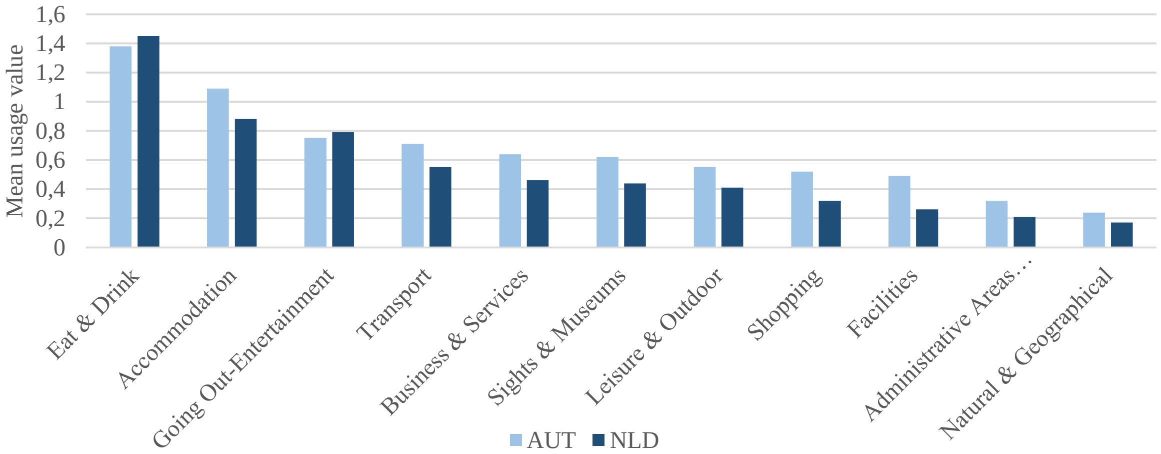


Figure 5-1 Mean of usage among categories. Sample sizes are presented in Table I-1.

5.2 Coverage

Yellowstone improves the coverage of HEREs places data since the data adds 53,252 new places in Austria and 184,910 in the Netherlands. Therewith, results support the expectation regarding coverage. Even though coverage is improved, Yellowstone covers only a small portion of HEREs places data. Roughly 240,000 and 500,000 unique places of HEREs places data are not present in the Yellowstone data in Austria and the Netherlands, respectively. Therefore, results show that Yellowstone and HEREs places data are complements since Yellowstone adds new places to HERE, while it only contains a fraction of the places of HERE. Hence, Yellowstone improves coverage, but is not a sufficient substitute for the places data.

In general, businesses and social gathering places contribute relatively less to the coverage improvement of HEREs places data as indicated by the mean, which contrasts to expectations. The mean indicates which part of the sample is a new place and which part is an existing place. For example, a mean of .8 indicates that 80% of the sample is a new place. Therewith, it shows the efficiency in a certain category. The part that improves coverage for each category is roughly consistent between the two countries (see Figure 5-2). However, there are differences among categories. Businesses and social gathering places have relatively a lower coverage improvement (dark colored bars) than other categories. The categories Accommodation, Eat & Drink, Going Out-Entertainment and Sights & Museums have a significantly lower coverage improvement than all light categories (Kruskal-Wallis, $p<.05$). The mean of the Natural & Geographical category is relatively high, which indicates that the few places in Yellowstone are mainly new places for HEREs places data.

Even though social gathering places have a low percentage, they add many new places. From a business perspective, absolute numbers are interesting. For both countries, most places are added to the Business & Services, Shopping and Facility categories (see Table II-1 for a complete overview). Generally, these categories also have a high sample size.

Concluding, Yellowstone adds most places to the Business & Services, Shopping and Facilities categories, while it is most efficient in the Natural & Geographical, Facilities and Administrative Areas & Buildings categories since relatively most places are added in these categories.

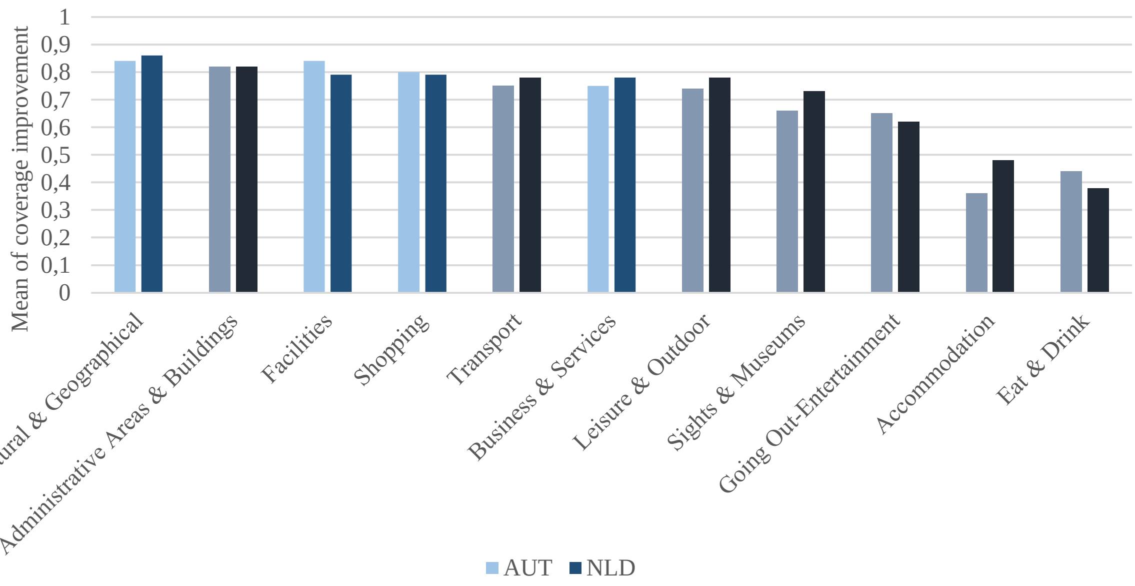


Figure 5-2 Mean of the coverage improvement for each category. Dark blue bars represent businesses and social gathering places. Sample sizes are presented in Table I-1.

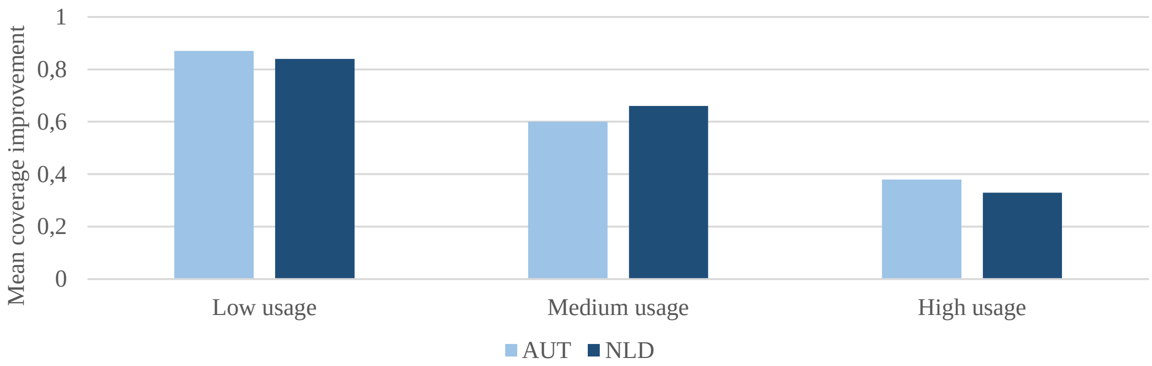


Figure 5-3 Mean of coverage improvement for the three usage groups (low usage N_{AUT}=51,932 and N_{NLD}=149,663, medium usage N_{AUT}=67,016 and N_{NLD}=78,038, high usage N_{AUT}=20,608 and N_{NLD}=24,457).

Even though city centers might have a higher social media usage, they are relatively less improved compared to surrounding neighborhoods. The hotspot analysis indicates significant differences despite of the low geographical correlation (Morans Index of 0.061194 for Austria and 0.044483 for the Netherlands). All cities have a cold spot in the city center indicating a lower mean than average and scattered red spots in surrounding neighborhoods indicating a higher mean than average (see Appendix III). An assessment of the red spots in Amsterdam and Eindhoven showed that these spots are mainly industrial/business areas. However, an assessment of Vienna and Linz does not support this idea. Generally, HEREs places data could have a higher coverage in city centers or Yellowstone mainly contains the same places as HEREs places data in these areas. Hence, coverage is relatively more improved at surrounding neighborhoods even though Yellowstone includes many places in the city center.

Even though coverage is improved, new places are mainly categorized as low used places (see Figure 5-3). Medium and highly used places have a statically significant lower coverage improvement than the low used places (Kruskal-Wallis, $p<.5$). Highly used places are mainly located in the city center, while coverage is mainly improved at the surrounding areas (see Appendix III). Additionally, Eat & Drink has a low coverage improvement, while this category has on average a high usage. The categorical and geographical distribution of usage groups and coverage improvement as discussed above explains the relatively large improvement of low used places. Thus, low used places have on average the highest coverage improvement.

5.3 Positional accuracy

On average, the geodesic point distance between Yellowstone and HEREs places data is 21.4 meter in Austria and 13.6 in the Netherlands. Roughly 70% of the social media records are within five meter and roughly 85% within 30 meter in Austria. These percentages are slightly higher for the Netherlands because roughly 80% is within five meter and roughly 90% within 30 meter. Since places are polygons represented by points, a small distance does not indicate incorrectness with the ground truth. However, as explained in the methodology, it is not clear which source is correct in case of larger inconsistencies. A brief analysis of records with distances larger than 30 meters with the ground truth does not reveal clear similarities or a consistently incorrect source. In the ten assessed records, all sources have a position that significantly differs from the ground truth. Therewith, inconsistencies do not indicate incorrectness or a lower accuracy of Yellowstone. Therefore, this study does not assess the opportunity of decreasing the distance to the ground truth by aggregating Yellowstone. However, Yellowstone improves positional accuracy by verifying geographical locations and increasing the likelihood of correctness. HEREs places data could be considered as the benchmark and therewith results show that Yellowstone is more consistent with HEREs places data in the Netherlands than in Austria. Nevertheless, more than 85% of the Yellowstone records is consistent with the benchmark in both countries. Therewith, results support the expectation regarding positional accuracy since the aggregation of social media data has increased the certainty of correctness regarding the mapped geographical location.

The positional accuracy is not significantly more confirmed at social gathering categories, which was expected. Figure 5-4 presents the mean score for each category, which indicates the average distance in meters between Yellowstone and HEREs places data. The category Eat & Drink and Going Out-Entertainment have a significantly lower mean (Kruskal-Wallis, $p<.05$) than other non-socially gathering categories. The categories Natural & Geographical and Leisure & Outdoor have the largest differences between Yellowstone and HEREs places data. The places in these categories generally have a large footprint size and this can result in larger differences between geographical points. For example, a mountain is significantly larger than a restaurant. Results seem to indicate that the average size of a place is correlated to the difference between sources. However, this measure does not consider the requirement since generally larger positional inaccuracies are accepted for larger places. Despite this exclusion, the measure still indicates which records confirm HEREs places data and therewith improve the positional accuracy. The shopping category differs among countries since Austria has a higher distance than the Netherlands. The shopping category has a relatively smaller sample size in Austria. So, Austria has less shopping pages on Yellowstone and those pages are less consistent with HEREs places data. Overall, the positional accuracy is least improved at Administrative Areas & Buildings, Leisure & Outdoor and Natural & Geographical places.

Results show conflicting outcomes regarding the confirmation of sources between city centers and surrounding areas. Although the transformed variable indicates a higher correlation than the untransformed one, the correlation between geography and improvement is low (Morans Index of 0.068491 for Austria and 0.046184 for the Netherlands).

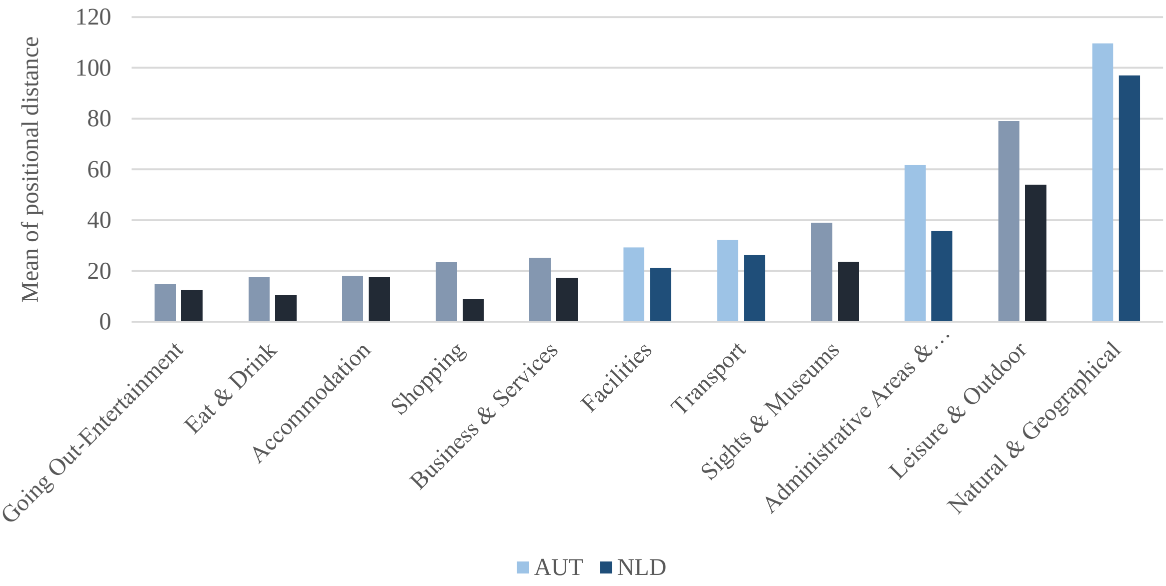


Figure 5-4 Mean of the distances between HEREs places data and Yellowstone for each category. Dark bars represent social gathering places. Sample sizes are presented in Table I-2.

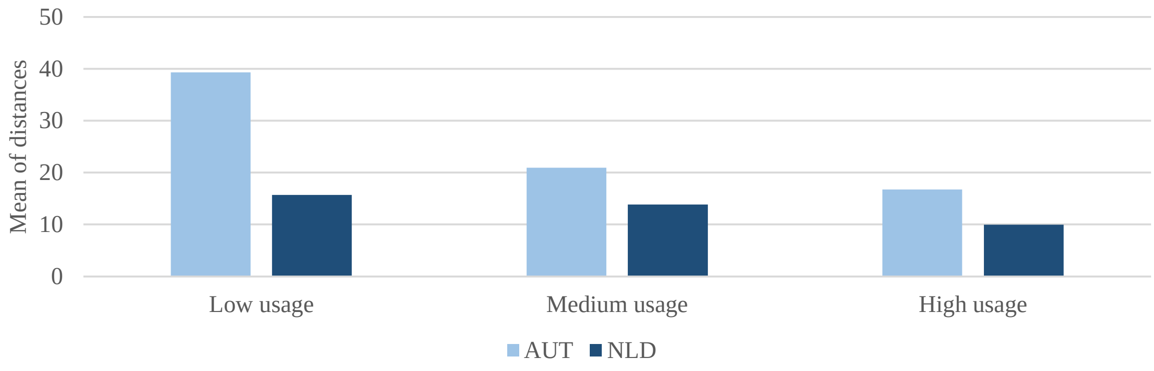


Figure 5-5 Mean of distances between HEREs places data and Yellowstone for the three usage groups (low usage N_{AUT}=6735 and N_{NLD}=24,599, medium usage N_{AUT}=26,742 and N_{NLD}=26,372, high usage N_{AUT}=12,703 and N_{NLD}=16,277).

Despite of the low correlation, the city center of Vienna, Linz, The Hague and Utrecht have hot spots, while the city center of Amsterdam, Rotterdam and Eindhoven are cold spots (see Appendix III). The other cities in Austria did not have significant spots at the city center. It was expected that city centers have cold spots because of less local knowledge due to out-of-town visitors. However, results do not support this idea. The reason for differences among cities is not clear from this analysis, although it can be caused by local sources used in HEREs places data. However, this is not clear and further analysis could reveal the underlying reason. Hence, the positional accuracy is mainly confirmed in the city center of Vienna, Linz, The Hague and Utrecht and least confirmed in the city center of Amsterdam, Rotterdam and Eindhoven.

The positional accuracy is mainly confirmed at highly used places. Figure 5-5 shows that highly used places have the lowest positional distance indicating that the distances between the places data and Yellowstone are significantly less than other categories. The Eat & Drink, Accommodation and Going Out-Entertainment categories have in general a high-usage and a large positional accuracy confirmation. This could explain the large positional accuracy consistency of highly used places.

5.4 Thematic accuracy

The thematic consistency is relatively high since two or three attributes are on average consistent among Yellowstone and HEREs places data (the mean is 2.3 for both countries). Less than one percent did not have any match and it is notable that nearly all records match on city in Austria. In general, city names are most often matched and categories least often. A place can have multiple correct categories, while it can only have one correct place and address. Therefore, unmatched categories do not necessarily indicate incorrectness. From the analysis, it is not clear which of the unconfirmed variables is correct. Therefore, attributes cannot be improved by correcting them, which is similar to the positional accuracy. However, the confirmation of attributes can improve the thematic accuracy by increasing the likelihood of correctness. Therewith, the expectation regarding thematic accuracy is supported since 19,360 records (80%) in Austria and 52,088 (77%) in the Netherlands have a match on address and city and 6,483 (27%) and 26,765 (40%) on all attributes, respectively.

Businesses have a large benefit by providing correct information. However, attributes are not significantly more confirmed for these categories. The category Eat & Drink mostly confirms HEREs places data since it has a significantly higher mean than all other categories (Kruskal-Wallis, $p<.05$). Figure 5-6 shows that averages are roughly consistent between the two explored countries. It is notable that the distribution of categories is largely consistent with the distribution of the positional accuracy showed in Figure 5-4 (section 6.2 elaborates on the correlation of

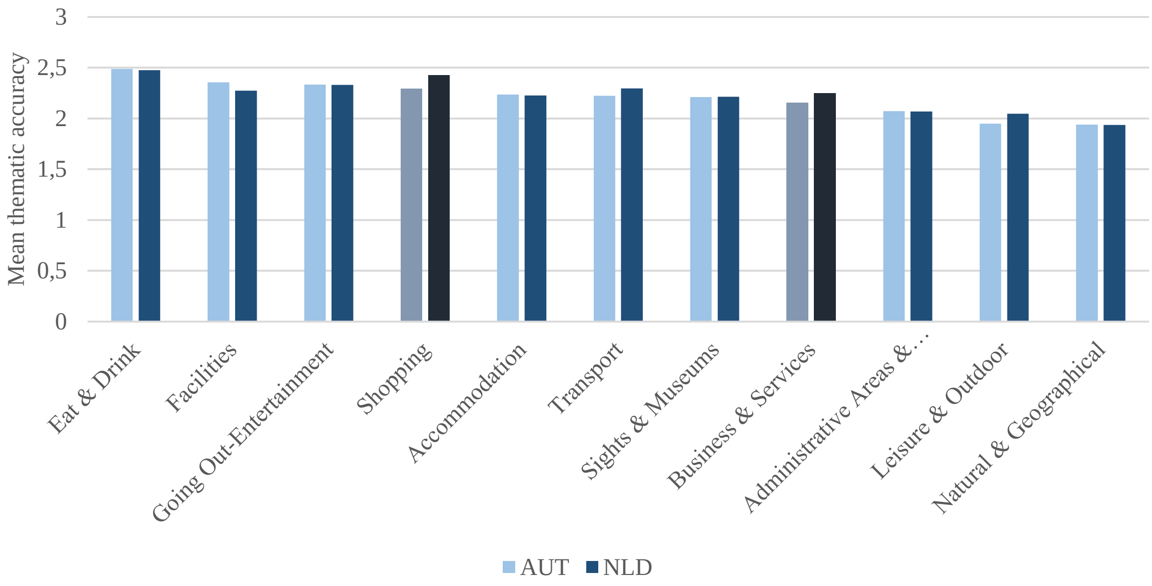


Figure 5-6 Mean of the thematic accuracy improvement for each category. Dark colored bars represent businesses. Sample sizes are presented in Table I-2.

quality measures). For the thematic accuracy, two geographical aspects were compared, which might have larger inconsistencies when footprint size of places increases. This explanation was used to clarify categorical distribution of the positional accuracy. However, results could also indicate that highly consistent categories have a good accuracy in both Yellowstone and HEREs places data and therefore confirm each other. Overall, Eat & Drink, Facilities, Going Out-Entertainment and Shopping places are largely consistent among Yellowstone and HEREs places data.

The thematic accuracy confirmation is not consistent between city centers and surrounding neighborhoods among cities. The correlation between geography and thematic accuracy is low (Morans Index of 0.039592 for Austria and 0.041741 for the Netherlands). Linz, Innsbruck, The Hague and Utrecht have a hot spot area in the city center, indicating that the thematic consistency is high in these areas. However, Vienna and Amsterdam have a cold spot in the city center. As was stated in section 3.4, out-of-town visitors, such as international tourist, could decrease the thematic accuracy of Yellowstone. Vienna and Amsterdam both have generally more out-of-town visitors than the other cities, which might explain the cold area in the city center of Vienna and Amsterdam. However, this idea is not confirmed with the results of the positional accuracy. Hence, attributes are mainly confirmed at the city center of Linz, Innsbruck, The Hague and Utrecht and less confirmed at the city center of Vienna and Amsterdam.

The thematic accuracy confirmation is roughly consistent among usage groups (see Figure 5-7). Highly used places have on average slightly more matches but the difference with other usage groups is small. Therefore, the thematic accuracy is roughly evenly confirmed for high, medium and low used places.

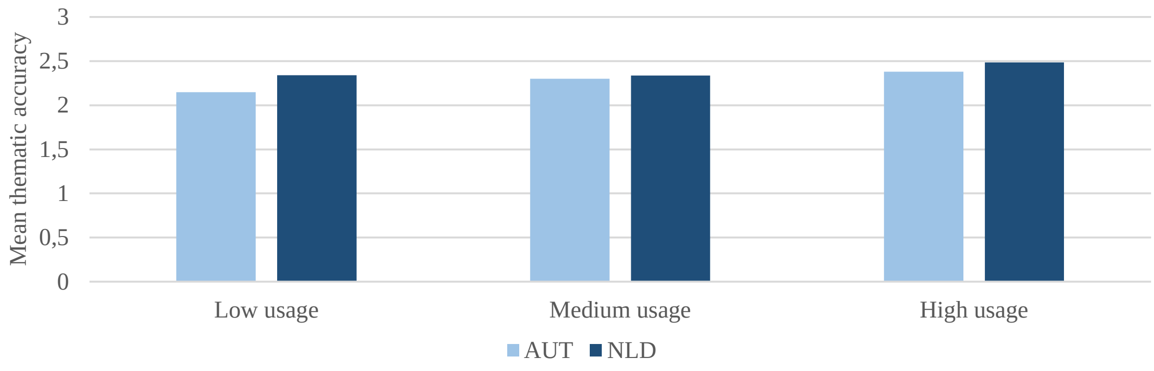


Figure 5-7 Mean of thematic accuracy improvement for the three usage groups (low usage N_{AUT}=6735 and N_{NLD}=24,599, medium usage N_{AUT}=26,742 and N_{NLD}=26,372, high usage N_{AUT}=12,703 and N_{NLD}=16,277).

5.5 Freshness

Literature states that social media data has a large freshness but results cannot reject nor confirm this statement. For Yellowstone, roughly 98% of the records was updated between 82 and 96 days before the data was retrieved. Social media data has many updates each day and therefore this indicates that data is loaded in batches rather than real-time refreshment. Additionally, this is also the case with the other sources that are ingested in HEREs places data. Therefore, the last update date is a characteristic of the ingestion process rather than of the POI or data source. Therewith, data does not present the freshness of a POI and freshness cannot be studied in this research. Hypothesis four cannot be supported nor rejected.

5.6 Richness

Yellowstone data improves the richness of HEREs places data. More than three quarter of the records did add one or more attributes to the places data. The majority of additions was an addition of a category, namely 26,067 categories in Austria and 69,878 categories in the Netherlands were added (see Table 5-2). Additionally, records that have added nine or more attributes added at least four categories. HEREs places data contained all cities, postal codes and routing latitude and longitudes and therefore Yellowstone did not add anything to these attributes. Yellowstone did also not add any emails or opening hours although they are available on the social media site. This is likely due to the ingestion restrictions of HERE. Additionally, no state information is added in Austria, while this information is added in the Netherlands. It was expected that Yellowstone improves richness, which is confirmed by these results.

Table 5-2 Addition of new attributes by Yellowstone to HEREs places data (N_{AUT}=24,164 and N_{NLD}=67,248).

	Address	Category	City	Postal	State	Email	Opening	Phone	Website	Display Latitude	Display Longitude
Austria	97	26067	0	0	0	0	0	152	3932	2559	2559
Netherlands	133	69878	0	0	11361	0	0	2953	20658	11354	11355

In general, Natural & Geographical and Administrative Areas & Buildings places have most added attributes (see Figure 5-8). It is notable that these results are roughly contrasting to positional and thematic accuracy improvements. This is analyzed in section 6.2. Additionally, many new places are added in these categories as well. The high number of additions in these categories might be caused by a low number of initial attributes. It is likely that HEREs places data did not have input sources with a high richness yet. Hence, Yellowstone largely complements HEREs places data in Natural & Geographical and Administrative Areas & Buildings places and least complements Eat & Drink and Accommodation places.

In general, Yellowstone adds less attributes in city centers than at surrounding areas in the Netherlands. All city centers in the Netherlands have a cold spot indicating significantly less added attributes than average. Only the city center of Eindhoven did not show significant results. In Austria, only Vienna has a significant hot spot in the city, while the other cities did not have significant spots. This outcome could indicate that Austria differs from the Netherlands in the richness improvement in city centers. However, this spot in Vienna is small and therefore does not provide a clear result. The lower addition in city centers in the Netherlands might be caused by less available attributes in Yellowstone data or many available attributes in HEREs places data at city centers. Yellowstone might have fewer attributes in the city center due to less knowledgeable (out-of-town) visitors in this area. However, this idea is not supported nor rejected by the data of Vienna. The richness is least improved in Eat & Drink and Accommodation places, which are often located in city centers. The cold spot might be explained by the places category present in city centers.

Results of the geographical analysis indicate a negative correlation between thematic accuracy and richness as is also indicated by the category distribution. The city center of The Hague and Utrecht have a high thematic accuracy, while the richness improvement is low. Additionally, the thematic accuracy is low and richness improvement are contrasting in several areas of Vienna (see Appendix III). However, the thematic accuracy and richness improvement is both low in the city center of Amsterdam. This area also showed contradicting results for the thematic accuracy. Other cities did not have statistically significant hot or cold areas that overlap between the quality measures.

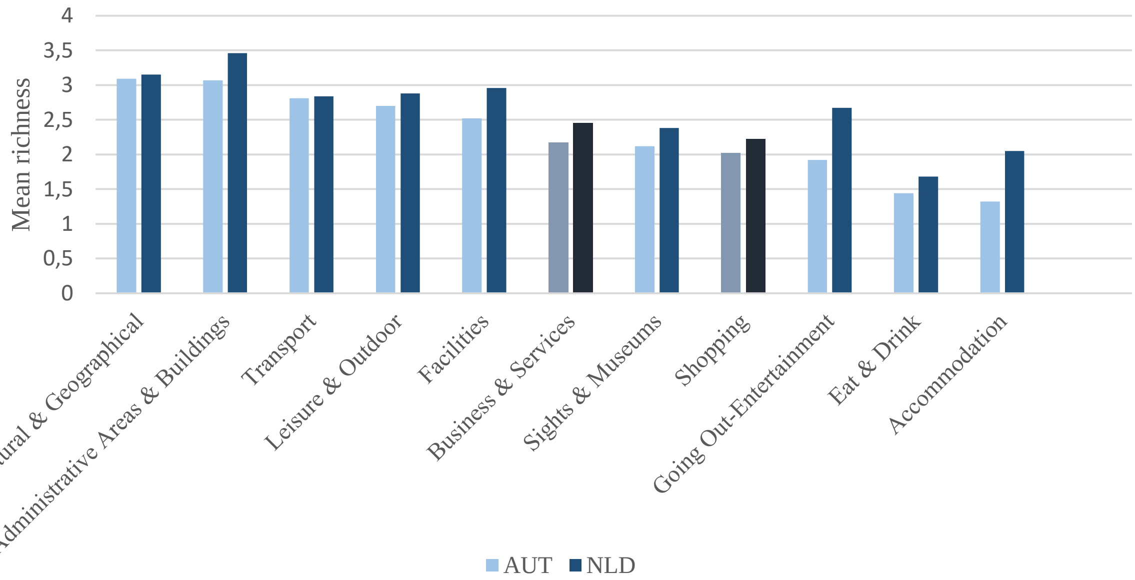


Figure 5-8 Mean of the richness improvement for each category. Darker colored bars represent businesses. Sample sizes are presented in Table I-2.

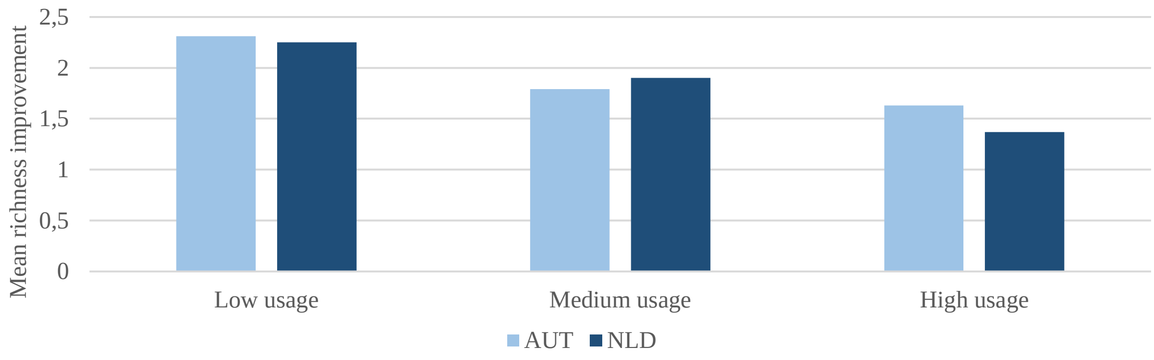


Figure 5-9 Mean of richness improvement for the three usage groups (low usage N_{AUT}=6735 and N_{NLD}=24,599, medium usage N_{AUT}=26,742 and N_{NLD}=26,372, high usage N_{AUT}=12,703 and N_{NLD}=16,277).

Yellowstone adds on average more attributes to low used places than to highly used places (see Figure 5-9). The richness is less improved in the Eat & Drink, Accommodation, Going Out-Entertainment categories, while these are mainly highly used. Additionally, Natural & Geography and Administrative Areas & Buildings are mostly improved, while these are mainly low used category. Furthermore, richness is generally less improved at city centers than at surrounding neighborhoods, while city centers are mostly highly used areas. Hence, the richness is mainly improved at low used places.

5.7 Conclusion

This section presents the added quality of aggregating Yellowstone with HEREs places data. Results show that Yellowstone improves coverage by adding more than 230,000 places in Austria and the Netherlands. The positional and thematic accuracy are also improved by confirming the current places data. More than 70% of the sample differs less than five meters from HEREs current location. Additionally, more than 75% has consistent address and city attributes. Apart from the confirmation of current information, more than 75% of the sample adds one or more attributes to HEREs places data and therewith improve its current richness. Currently, opening hours and email address are not ingested, which could even enlarge the richness improvement by Yellowstone. Unfortunately, the data does not indicate whether Yellowstone can also improve freshness due to the ingestion process of HERE. However, Yellowstone improves all other quality aspects (see Table 5-3). The social media source is not suitable for replacing HEREs places data although it might replace other sources that HERE uses. Overall, HERE can benefit from aggregating Yellowstone with their current places data since quality is improved.

Yellowstone relatively adds most places and attributes to Natural & Geographical and Administrative Areas & Buildings places, while positional and thematic accuracy attributes are least consistent in these categories. The positional accuracy might be least consistent due to the large footprint size of places. Additionally, this could also lower the thematic accuracy in these categories. Even though the accuracy is generally low at these categories, results do not indicate the accuracy of new places or attributes. For the thematic accuracy, only address, city and category were assessed, while address and city are generally not the newly added attributes. Therefore, HERE can improve its coverage and richness largely in these categories.

Relatively more places and attributes are added at surrounding neighborhoods than at city centers. Only results in the Netherlands support the latter, while Austria does not provide significant results. Therewith Yellowstone better complements HEREs places data at surrounding neighborhoods. This could be influenced by a low coverage and richness of HEREs places data and/or a high coverage and richness of Yellowstone in these areas. Yellowstone might have less attributes in the city center due to less knowledgeable (out-of-town) visitors in this area. Overall, HERE can improve its coverage and richness largely at surrounding neighborhoods.

Yellowstone mainly confirms positional and thematic information for Eat & Drink and Going Out-Entertainment places, while these places have relatively low additions of places and information. The positional and thematic consistency are both high for these categories. This might be related to the small footprint size or good current coverage and richness of HEREs places data. Even though the relative coverage might be low in these categories, the absolute coverage improvement is high. Many places are added in these categories. Hence, HERE can confirm its positional and thematic information in these categories.

The accuracy is mainly improved in the city center of Linz, The Hague and Utrecht. The position of places is also largely consistent in the city center of Vienna, but attributes are not largely confirmed. It was expected that city centers would have a lower accuracy due to less local knowledge and many out-of-town visitors. However, results are contradicting and show that Yellowstone can also largely confirm places in city centers. Therefore, there is no clear geographical area where HERE can mainly confirm its current places data.

Even though Yellowstone largely improves quality, HERE should aggregate Yellowstone with caution since the data also contains flaws. This study does not assess the accuracy of newly added places or attributes. Additionally, some places have attributes that do not match with HEREs places data. If Yellowstone is incorrect and these attributes are used, it decreases the current quality of HEREs places data. Furthermore, quality improvement is largely dependent on the matching and blending algorithm.

With the current developed insights, the subsequent section elaborates on the added value of Yellowstone by relating quality improvements with usage. This section already provides some insights, but an extensive analysis is conducted in the following section.

Concluding, Yellowstone increases quality of HEREs places data on all quality aspects. Even though all categories and geographical areas benefit, some have larger improvements than others. New additions of places and attributes are relatively largest in Natural & Geographical and Administrative Areas & Buildings places. Additionally, surrounding neighborhoods have generally a higher coverage and richness improvement than city centers. The accuracy is largely improved for Eat & Drink and Going Out-Entertainment places. Furthermore, the city center of Linz, The Hague and Utrecht show a large consistency between Yellowstone and HEREs places data.

Table 5-3 Categorical means for sample size share, usage and quality measures.

	Share in Percent	Mean usage	Mean coverage	Mean positional accuracy	Mean thematic accuracy	Mean richness
Austria						
Eat & Drink	16%	1.38	0.44	17.32	2.49	1.44
Going Out-Entertainment	10%	0.75	0.65	14.76	2.33	1.92
Sights & Museums	3%	0.62	0.66	38.79	2.21	2.12
Natural & Geographical	1%	0.24	0.84	109.61	1.94	3.09
Transport	1%	0.71	0.75	32.19	2.22	2.81
Accommodation	9%	1.09	0.36	18.01	2.24	1.32
Leisure & Outdoor	2%	0.55	0.74	78.91	1.95	2.70
Shopping	12%	0.52	0.80	23.29	2.29	2.02
Business & Services	35%	0.64	0.75	25.11	2.16	2.17
Facilities	11%	0.49	0.84	29.26	2.36	2.52
Administrative Areas & Buildings	0%	0.32	0.82	61.70	2.07	3.07
Netherlands						
Eat & Drink	13%	1.45	0.38	10.41	2.47	1.68
Going Out-Entertainment	5%	0.79	0.62	12.51	2.33	2.67
Sights & Museums	3%	0.44	0.73	23.52	2.21	2.38
Natural & Geographical	0%	0.17	0.86	97.04	1.94	3.15
Transport	1%	0.55	0.78	26.18	2.30	2.84
Accommodation	3%	0.88	0.48	17.39	2.23	2.05
Leisure & Outdoor	2%	0.41	0.78	54.01	2.05	2.88
Shopping	24%	0.32	0.79	8.89	2.43	2.22
Business & Services	36%	0.46	0.78	17.23	2.25	2.45
Facilities	13%	0.26	0.79	21.14	2.27	2.96
Administrative Areas & Buildings	0%	0.21	0.82	35.74	2.07	3.46

6 Value of social media data for points of interest

6.1 Introduction

The preceding chapter explored the added quality of Yellowstone to HEREs places data. However, this does not reveal insights into the added value since value is a combination of quality and usage. Therefore, this section examines the relation between usage and quality by using three binary logistic regression models. The model includes improved quality of existing HEREs places meaning that positional and thematic accuracy and richness are considered. By means of this model, the relation between each of these quality aspects and usage is clarified.

This section starts with a description of the correlation among quality measures. Thereafter, popularity factors are obtained from the Yellowstone data by a factor analysis. These factors are used in the regression model, which is described in the last subsection.

6.2 Relation among quality measures

In section 5, quality measures were separately described and explored. Results indicate that there might be a relation between positional and thematic accuracy. Therefore, correlations are examined and presented in Table 6-1. The positional distance is considered instead of positional accuracy since the distance can be measured. Hence, a larger positional distance between Yellowstone and HEREs places data represents a lower positional accuracy.

*Table 6-1 Correlation among quality measures (N_{AUT}=24,164 and N_{NLD}=67,248). *** is significance at 0.01 level (2-tailed).*

		Positional distance	Thematic accuracy	Richness
Positional distance	AUT	1	-.357***	.166***
	NLD	1	-.400***	.272***
Thematic accuracy	AUT	-.357***	1	-.265***
	NLD	-.400***	1	-.439***
Richness	AUT	.166***	-.265***	1
	NLD	.272***	-.439***	1

The positional and thematic accuracy have a moderate, positive correlation. This indicates that a high positional consistency between Yellowstone and HEREs places data is related to a high thematic consistency. Results of the previous chapter show that the accuracy is mainly confirmed in the Going Out-Entertainment, Eat & Drink and Shopping category, but not in a common geographical area. Regarding geography, some areas have a strong consistency on both the positional and thematic accuracy as is described in section 5.7. However, the city center of Vienna and Amsterdam show contrary results. The quality measures only have a moderate relation, which indicates that the relation is not perfect.

Results of the previous section indicate that coverage and richness are improved in the same categories, which suggests a relation among those quality measures. Unfortunately, these measures cannot be correlated since new places do not have an indication of richness improvement. However, the correlation of the mean coverage and richness improvement provides an indication (the means are presented in Table 5-3). This indicates that there is a positive correlation between the variables. However, it is notable that the shopping category has a large coverage improvement, but a relatively low richness improvement. Additionally, Leisure & Outdoor places have contrasting results. This implies that these variables do not have a perfect correlation. Contrasting results of the positive correlation might be caused by quality differences of input sources of HERE.

The accuracy variables have a weak to moderate negative correlation with the richness improvement. This indicates that a high positional and thematic consistency is related to a lower richness improvement. This negative correlation is also visible in the results of section 5. Categories with a high accuracy improvement have a low richness improvement and contrary. Additionally, accuracy is mainly improved in city centers, while richness is mainly improved at surrounding neighborhoods. Therefore, this correlation supports the idea that Yellowstone generally improves either the accuracy or the richness.

Places seem to have either a large accuracy improvement or richness improvement. However, places with a low accuracy might have a large potential for correcting HEREs places data. This study assesses the accuracy by considering the consistency among sources. This does not show any insight into Yellowstone's opportunities for correcting incorrect information of HEREs places data. Therefore, inconsistent places could correct HEREs places data. However, this study does not indicate which source is corresponding to the ground truth and an initial assessment did not indicate consistencies of correctness (see section 5.3). Therewith, the Natural & Geographical, Leisure & Outdoor and Administrative Areas & Buildings category and places in the city center of Vienna and Amsterdam have the potential to improve the accuracy by correcting incorrect attributes. These records have a high potential for adding value to HEREs places data, but could also largely decrease value when they contain incorrect information.

Concluding, the positional and thematic accuracy are positively correlated, while these are negatively correlated with richness improvements. Results of section 5 already showed that this is also visible in the improvements of categories and geographical areas. Even though this provide clear implications for improving quality, it does not reflect the value obtained from social media since usage is not considered. Therefore, a follow-up analysis is conducted in section 6.4.

6.3 Popularity indicators of Yellowstone

Places on Yellowstone differ in their popularity and therefore underlying factors were retrieved. The factors are revealed by an explanatory factor analysis. This analysis has several assumptions that need to be satisfied. The theory does hint three factors, linear correlation is expected, the sample size is sufficient, outliers were removed and all variables are scale level of measurement. Furthermore, the correlation matrix can indicate whether factors can be retrieved. The variables transformed by a logarithm function have a higher correlation among each other than the non-transformed variables. The correlation matrix shows that there are several variables with a sufficient correlation for factor analysis (>.3). However, variables violated normality. The analysis is robust to minor violations of normality. Therewith, all assumptions except normality are satisfied. Since the transformed variables approach normality, these variables are preferred for the exploratory factor analysis.

The factor analysis that includes all variables transformed by a logarithm function provides the best results. Model one result in an ultra-Heywood case for both countries and model three for Austria (models are described in section 4.6). Therefore, these models are not selected. The fourth model that includes only transformed variables is preferred because outcomes correspond most with the hypothesized factors. Additionally, the transformed variables are closer to normality, which is an assumption of the factor analysis. Therefore, the factor analysis that considers the logarithm variables is preferred. The results for both countries have a sufficient anti-image correlation matrix (the diagonal is larger than .5), communalities are above .5, the Bartlett's test of sphericity is significant ($p<.05$) and Kaiser-Mayer Olkin (KMO) value is sufficient (.742 for Austria and .741 for the Netherlands). By transforming the variables, negative values that are present in the weekly and monthly variables were altered to missing values. This resulted in an omission of 19,038 records in Austria and 59,357 records in the Netherlands that had unexplainable values (N_{AUT} = 58,378 and N_{NLD} = 192,801).

Three popularity factors can be identified from the analysis. There are some differences with the hypothesized item loadings even though results correspond to three hypothesized factors (see Table 3-4). The item loadings are presented in Table 6-2, which shows that all items have a high loading on one of the factors (>.7). Both orthogonal and oblique rotations show similar factors. The item loadings on the popularity of the area is as expected. It was also expected that the number of likes, check-ins, reviews and ratings would load on an activity level factor. However, check-ins do not load on this factor. Additionally, the weekly and monthly likes do load on this factor, which is unexpected. The total, weekly and monthly number of check-ins are loaded on a separate factor. Based on these results, the correlation between likes and check-ins might not be linear as was expected. This indicates that likes and check-ins cover different aspects of popularity, which is important for understanding the three factors.

Table 6-2 Overview of item loadings on the three factors obtained from a principal factor analysis with Direct Oblimin rotation. The numbers on the left are obtained from Austria and numbers on the right from the Netherlands.

	Information popularity	Visiting popularity	Popularity area
Likes	.970 / .948		
Likes per week	.709 / .708		
Likes per month	.810 / .806		
Review	.798 / .742		
Rating	.786 / .732		
Check-ins		-.755 / -.795	
Check-ins per week		-.816 / -.838	
Check-ins per month		-.989 / -.984	
Total check-ins neighbors			.987 / .848
Average check-ins neighbors			.985 / .862

The number of likes and check-ins are not linearly correlated but have an L-shaped relation. The total likes and check-ins differ between categories (see Figure 6-1). For shopping places, there is a notable large difference between likes and check-ins, which could be explained by the nature of these actions. Namely, a like indicates that the person publicly supports the place and would like to receive updates from this place. Check-ins are shared with connections and are included in place suggestion for friends. Therewith, people need to be physically present for a check-in, while this is not the case for likes. Hence, the motivation and objective of both actions differ, which could explain the missing linear correlation among the variables.

Furthermore, it was expected that social gathering categories as used in section 3.4 and 5 have a high number of check-ins. People are physically present at these places and sharing their location with friends is consistent with the social gathering function. However, Accommodation, Eat & Drink, Going-out – Entertainment, Leisure & Outdoor, and Sights & Museums do not have significantly more check-ins than other categories (see Figure 6-1).

A country comparison between Austria and the Netherlands showed that there are notable differences in the Leisure & Outdoor, Natural & Geographical, Transport, Going Out-Entertainment, Accommodation and Eat & Drink category. Generally, places in Austria have more likes and check-ins than places in the Netherlands, which is also visible at most categories. However, some categories show contrasting results. The number of likes is generally less for the categories Sights & Museums and Natural & Geographical in Austria than in the Netherlands. This would indicate that these places are less publicly supported and of less interest in Austria. Especially, the difference for Natural & Geographical places is notable. Austria has significantly more places in this category since the sample is four times larger. For this category, the Netherlands could have a few very popular places, while Austria has also less popular places included on Yellowstone. Additionally, half of the places in Austria are mountains or hills and a quarter are lakes, while roughly three quarters in the Netherlands are lakes. These variations of sample size and category can result in differences regarding the mean. Furthermore, the number of check-ins is on average lower for the categories Going Out-Entertainment, Accommodation and Eat & Drink in Austria. Hence, people in the Netherlands check-in more often at these categories. It could be that these categories have more visitors in the Netherlands. Overall, results show that there are differences between countries among the likes and check-ins provided by users on Yellowstone.

Hence, the correlation and factor analysis show that likes and check-ins represent different aspects of popularity and therefore they are loaded on separate factors. Initially, the three factors were identified as activity level, activity frequency and popularity of the area. However, the first two factors have different item loadings, which indicates that they represent something else. The first factor contains likes, reviews and ratings, which are mainly associated with exchanging information. Liking a page results in receiving information about the page and information is provided by ratings and reviews. Therefore, this factor was identified as information popularity. The second factor contains variables associated with check-ins. Physical presence is required for check-ins and therefore this factor was identified as visiting popularity. The last factor was

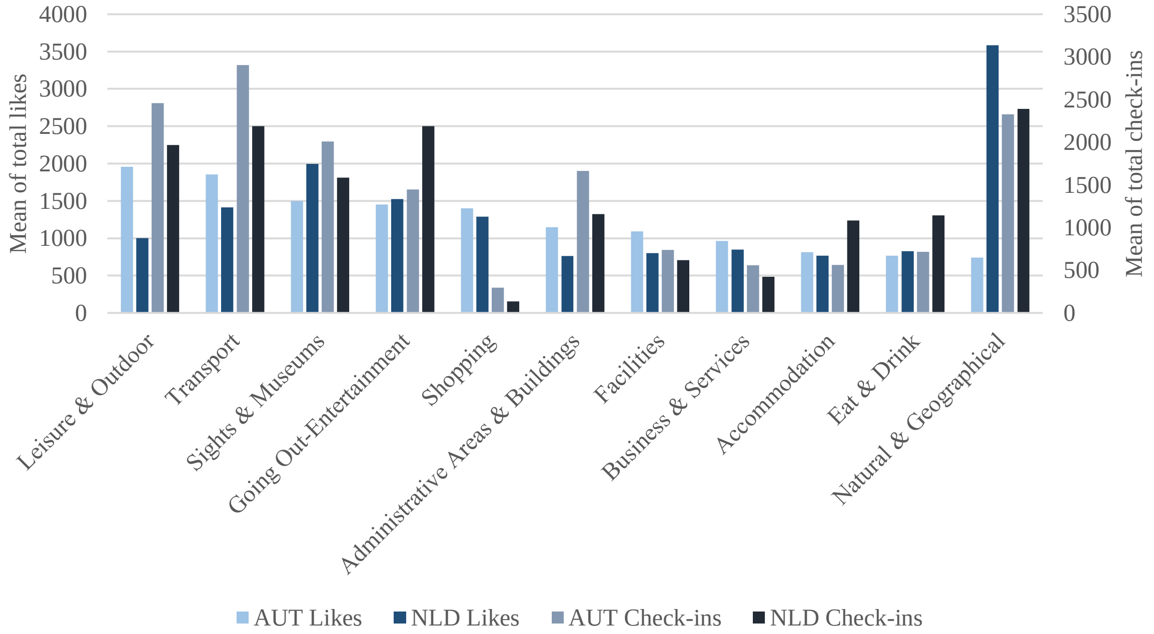


Figure 6-1 The mean number of likes and check-ins of the eleven POI categories. Sample sizes of categories are presented in Table I-1.

identified as area popularity since it considers the popularity of neighbors within a radius of 100 meter.

The three identified factors are positively correlated among each other. Since an oblique rotation was used, factors can correlate. The visiting and information popularity have a moderate correlation (r_{AUT}=.446 and r_{NLD}=.448) indicating that a higher information popularity is associated with a higher visiting popularity. The area popularity has a weak positive correlation with the other popularity factors. Hence the three popularity factors represent different aspects of popularity but are positively related among each other.

Concluding, Yellowstone has three factors that represent the popularity of a place on social media. The factors represent information, visiting and area popularity and are positively correlated with each other.

6.4 Added value model

To assess the added value, the quality measures and usage of HEREs POIs are considered as was presented in Figure 3-1. Therewith, the relation of the quality measures with usage can indicate whether and what value is added by Yellowstone to HEREs places data (see section 4.6).

Quality measures are significantly correlated with usage. As was discussed in section 6.2, quality measures are significantly correlated with each other. Positional and thematic accuracy are positively correlated, while they are negatively correlated with the richness improvement. This clustering of variables is also visible in the correlation with usage, which is presented in Table 6-3. The positional distance is negatively correlated with usage, which indicates that an improved accuracy is positively correlated with a higher usage. This is consistent with the thematic accuracy, while it is contrasting with the richness improvement. However, this correlation does not provide insight into the relation of each quality measure separately with usage. A regression model provides this insight since the relation of each quality measure is explored disregarding the effect of the other quality measures and control variables.

Table 6-3 Correlation of usage with the quality measures

	Usage
Positional distance	-.017**
Thematic accuracy	.081**
Richness	-.196**

The sample size for the regression analysis was reduced due to the omission of newly added places. Newly added places were omitted because these places have missing values on the other quality measures. Hence, this analysis covers only records that improve existing places of HEREs places data and therefore hypothesis one cannot be supported nor rejected. Since three separate binary logistic regressions were conducted, the sample size is related to the distribution of places among the usage groups (see Table 5-1).

Results show that the model does not have multi-collinearity problems. The included variables, as described in section 4.6, are elaborated in Table 6-4. With these variables, the unrestricted model is:

$$Pr_i\ usage = \alpha + \beta_1\ Positional\ distance + \beta_2\ Thematic\ accuracy + \beta_3\ Richness + \beta_{4-6}\ Popularity\ factors + \beta_7\ Country + \beta_{8-18}\ Catagories + \beta_{19}\ Chain + \mathcal{E}$$

Table 6-4 Description of the variables in the model

Variable	Description	Explanation	N	Mean	Standard Deviation
PostionalDis	Distance between HERE and Yellowstone	In meters	91412	15.6408	131.0353887
ThematicAcc	Consistency between HERE and Yellowstone	In number of attributes	91412	2.356943	0.630086511
Richness	Number of added attributes by Yellowstone	In number of attributes	91412	1.78	1.657
InformationPop	Popularity on social media regarding information	Construct	70609	0.230896	1.047840157
VisitingPop	Popularity on social media regarding visiting	Construct	70609	0.497761	1.162124045
AreaPop	Popularity of the neighborhood on social media	Construct	70609	0.289664	0.950662535
Country	Country (Austria or the Netherlands)	Binary (0=AUT; 1=NLD)	91412	0.74	0.441
Eat & Drink	Category of HERE	Binary (0=no; 1=yes)	91412	0.42	0.494
GoingOut	Category of HERE	Binary (0=no; 1=yes)	91412	0.12	0.33
SightsMuseums	Category of HERE	Binary (0=no; 1=yes)	91412	0.04	0.189
NaturalGeo	Category of HERE	Binary (0=no; 1=yes)	91412	0.00	0.05
Transport	Category of HERE	Binary (0=no; 1=yes)	91412	0.01	0.10
Accommodation	Category of HERE	Binary (0=no; 1=yes)	91412	0.13	0.339
LeisureOut	Category of HERE	Binary (0=no; 1=yes)	91412	0.02	0.146
Shopping	Category of HERE	Binary (0=no; 1=yes)	91412	0.22	0.413
BusinessServ	Category of HERE	Binary (0=no; 1=yes)	91412	0.44	0.496
Facilities	Category of HERE	Binary (0=no; 1=yes)	91412	0.13	0.338
AdminArea	Category of HERE	Binary (0=no; 1=yes)	91412	0.00	0.061
Chain	Part of a business chain	Binary (0=no; 1=yes)	91412	0.08	0.264

From the three analyzed regressions, the medium and highly (MH) used model performs least, while the low and highly (LH) used model performs best. Table 6-5 presents an overview of the results and measures of fit. All models have a better fit than the intercept only model. However, the MH model has least improvement. It is a moderate model (based on Nagelkerk R^2) and has the lowest overall prediction. Even though this model has the lowest measure of fit, it is still a sufficient model indicated by the Nagelkerk R^2, Log Likelihood and Chi-Square. The low and medium (LM) used model has a better fit than the MH, but the LH performs even better. The LH model is a strong model that can predict 96.7 percent correctly.

In all models, nearly all variables are significant. This significance in combination with the measures of fit imply that the model has a sufficient predictive power in assessing the relation between quality improvement by Yellowstone and the usage of POIs. The coefficients of the model indicate the likelihood that a place is a medium or highly used place by a one-unit increase of the independent variable holding constant all other variables.

All models show that a higher positional accuracy does not increase the likelihood that a place is highly used. Therewith, hypothesis two is rejected since the positional accuracy change is not positively correlated with the usage of POIs. Hence, a better positional accuracy does not increase the likelihood of being a highly used place.

The thematic accuracy has contrasting signs in the models. A lower thematic accuracy improvement increases the chance of being a medium used place. However, the other models show a significant positive relation indicating that it is a positive determinant of usage. Therewith, a place with a high thematic accuracy is more likely to be a highly used place compared to low or medium used place. For example, one extra confirmed attribute increases the

Table 6-5 Logit estimates of low, medium and highly used places

	Low used / medium used (LM)	Medium used / highly used (MH)	Low used / highly used (LH)
PostionalDis	0.000** (.000)	0.000 (.000)	0.000* (.000)
ThematicAcc	-0.052** (.021)	0.106*** (.023)	0.227*** (.056)
Richness	-0.07*** (.008)	-0.051*** (.010)	-0.155*** (.023)
InformationPop	-0.028* (.016)	-0.014 (.018)	-0.277*** (.041)
VisitingPop	0.188*** (.017)	0.263*** (.016)	0.653*** (.038)
AreaPop	1.113*** (.017)	1.722*** (.022)	1.854*** (.048)
Country	-1.09*** (.032)	-0.326*** (.028)	-1.332*** (.071)
Eat & Drink	5.848*** (.114)	2.464*** (.037)	7.939*** (.141)
GoingOut	-0.211*** (.050)	-0.255*** (.036)	-0.806*** (.118)
SightsMuseums	-0.159*** (.056)	-0.413*** (.074)	0.38*** (.116)
NaturalGeo	-1.074*** (.219)	-0.32 (.371)	-2.06*** (.482)
Transport	0.466*** (.114)	1.38*** (.120)	2.353*** (.166)
Accommodation	2.253*** (.051)	0.905*** (.041)	3.253*** (.093)
LeisureOut	-0.746*** (.077)	-0.076 (.098)	0.154 (.133)
Shopping	-0.639*** (.038)	-0.461*** (.046)	-0.564*** (.105)
BusinessServ	-0.063** (.029)	0.086*** (.030)	-0.258*** (.075)
Facilities	-0.88*** (.042)	-0.218*** (.067)	-0.433*** (.101)
AdministrativeArea	-1.245*** (.180)	-0.516* (.304)	-1.866*** (.375)
Chain	1.143*** (.042)	.586*** (.053)	1.781*** (.105)
Observations	53037	49164	39017
Sign.	.009	.000	.000
Nagelkerke R2	.614	.468	.926
-LL2	39372.038	43591.072	7769.588
Chi-Square	283.578	635.8	47.039
Prediction	82.6	80	96.7

Notes:***=significant at1%;**=significant at5%;*=significant at10%.Standard errors are in parentheses. The estimates are robust maximum-likelihood.

log odds chance of being a highly used place with .227 compared to low used place. Specifically, one extra confirmed attribute increases the odds ratio of being a highly used place by 1.255. Hence, it increases the odds by 26 percent holding constant all other variables. Concluding, a high thematic accuracy improvement increases the likelihood of being a highly used place. This statement confirms hypothesis three and it supports the idea that highly used places have a good thematic accuracy in both Yellowstone and HEREs places data.

All models show that a richness improvement by Yellowstone decreases the likelihood of being a highly used place. The three models have a significant negative coefficient. Hypothesis five is supported since a richness improvement negatively affects the likelihood of being a highly used place. This supports the idea that highly used places already have many attributes in HEREs places data and thus Yellowstone can add less attributes.

Based on these insights of the models, places with a high accuracy and large richness are most likely to be highly used. Namely, the thematic accuracy is highest at these places and richness is least improved, which could imply that richness was already high.

Hence, the model shows that thematic accuracy and richness are significant determinants of usage. In general, a positive coefficient indicates a high added value since this represents a high quality improvement at highly used places. Thematic accuracy improvement is related with a high usage and hence creates high value. Less value is created by richness improvements. The richness improvement is negatively related to usage and therefore if none of the other variables changes, a higher richness improvement is associated with a lower usage.

HERE can use Yellowstone if it aims to largely improve the richness of low used places or verify the thematic accuracy of highly used places. If HERE aims to improve highly used places that are both important for consumers and the automotive industry, mainly the thematic accuracy can be improved by a verification of attributes. A lower thematic accuracy of low and medium used places could indicate that Yellowstone is more often incorrect at these places. Therewith, one could assume that new attributes added by Yellowstone are incorrect at these places. However, mainly different attributes are added than the ones tested by the thematic accuracy (see Table 5-2) so this idea is not supported by results of this study. Additionally, an inconsistency between Yellowstone and HEREs places data does not necessarily indicate an incorrectness of Yellowstone. This study uses HEREs data as benchmark, but this benchmark is not necessarily correct, which is a limitation. An elaborated discussion is presented in section 7.2. Hence, Yellowstone is well-suited for major richness improvement of low used places and thematic accuracy confirmation of highly used places.

Even though most value is associated with a thematic accuracy improvement, other improvements also add value. Lower improvements also result in customer value since it still improves product advantage. Additionally, an improvement of low used places is valuable since these records are still used. These places are used but are not one of the top-used records by the website/application or categorized in a priority category for automotive users. However, customers can be satisfied with the increased product advantage. Hence, a low improvement and improvement of low used places still results in value.

6.5 Conclusion

This section elaborates on the added value of Yellowstone to HEREs places data by relating quality improvements and usage. The preceding section provides insights into quality improvements and the improvements of categories and geographical areas. The additional value of social media is explored with these insights and understandings of this section.

An important insight for the added value model is the three popularity factors obtained from Yellowstone. Yellowstone includes information popularity, visiting popularity and popularity of the area. The information popularity indicates the eagerness of people for receiving information from a place and providing information. The likes, reviews and ratings are loaded on this factor. The second popularity factor, visiting popularity, represents the popularity of physically attending a place. Check-ins are highly loaded on this factor. Hence, likes and check-ins do not have a strong linear correlation as was expected. The last factor represents the popularity of the neighborhood by considering the total and average check-ins of neighboring places. Overall, these factors represent the popularity of places on Yellowstone and are positively correlated. These factors are used as control variables to assess the relation between usage and quality measures without the effect of popularity.

This relation between quality improvements and usage is examined by three sufficient binary logistic regressions. The independent and control variables are generally significant indicating a good model fit. The models confirm hypothesis three and five and reject hypothesis two. It shows that the thematic accuracy is positively related and richness negatively related to usage, which confirms the hypotheses. The positional accuracy is not positively nor negatively related to usage, which rejects hypothesis two.

Most value is added by the improvement of the thematic accuracy. A high thematic accuracy improvement increases the likelihood of being a highly used place. Highly used places are often used by consumer and automotive customers of HERE. The thematic accuracy is assessed by the consistency between Yellowstone and HEREs places data. A confirmation of sources results in an increased likelihood of correctness, which is important for accuracy. Accuracy is essential for product advantage and therewith customer value. Thus, Yellowstone adds value to HEREs places data by confirming attributes of their highly used places.

The thematic accuracy is positively correlated with the positional accuracy. This indicates that places that have a high thematic consistency generally also have a high positional consistency. However, the positional accuracy is not positively correlated with usage, which is the case for thematic accuracy.

Yellowstone also adds value by richness improvements of low used places. This adds value to a lesser extent since these places are less used by consumers and automotive customers of HERE. Even though this creates less value, every improvement delivers value. Specifically, the extent of the added value differs among usage groups.

Concluding, Yellowstone adds high value by thematic accuracy confirmation of highly used places and to a lesser extent by major richness improvement of low used places.

7 Conclusion, discussion and implications

This section provides the overall conclusion that can be drawn from outcomes of this study. It reflects on the main question and sub questions introduced in section 1.3. It also reflects on the outcomes and the methodology by describing limitations and considering the generalization of results. Furthermore, theoretical and practical implications are provided. The scientific world can contribute from the theoretical implications and recommendations for future research. HERE can mainly benefit from the results and practical implications described in this section.

7.1 Conclusion

This study shows that social media data can deliver value for location service providers in mapping points of interest. As is often the case for product improvements, financial indicators cannot be used to determine value since the product is not launched yet. Customer value is a sufficient indicator for these cases. Since customer value is mainly determined by product advantage, this study explores whether social media data improves the product advantage of point of interests (POIs). For POIs, this product advantage consists of the combination of quality and usage. Specifically, quality consists of five determinants: coverage, positional accuracy, thematic accuracy, freshness and richness. Therewith, this study shows the value magnitude by presenting the product advantage improvement of POIs by social media data in Austria and the Netherlands.

Yellowstone data, which is the social media data for this study, largely adds value by confirming the correctness of attributes at highly used places. Large improvement of places that are highly used results in a product advantage for many customers. The confirmation of attributes is considered as a thematic accuracy improvement, which is part of the product advantage of POIs. The improvement is particularly large at Eat & Drink and Going Out-Entertainment places. Additionally, the city center of Linz, The Hague and Utrecht are largely consistent between Yellowstone and HEREs places data. Therefore, the value extent differs between quality aspects, categories and geographical areas.

Yellowstone data provides value to a lesser extent by coverage and richness improvements of low used places. Thus, Yellowstone is a better complement for low used places of HERE. It adds on average many attributes to Natural & Geographical and Administrative Areas & Buildings places. Generally, Yellowstone least complements HEREs places data in city centers because coverage and richness are more improved at surrounding areas.

These outcomes show that the value magnitude differs among quality aspects although social media data provides value on all aspects in Austria and the Netherlands. A replacement of places data with social media data is not sufficient since it only includes a fraction of all places. However, the aggregation of social media with places data improves the product advantage of POIs. Social media data can improve the coverage, sufficiently verify the position and thematic accuracy of the places data and improve the richness by mainly adding category and website information. Hence, social media data can improve the product advantage and therewith delivers value for mapping POIs.

The obtained value is clearly related to the characteristics of social media data defined by the five V's. Results of this study show that the volume results in value by improving the coverage and richness. Unfortunately, a freshness improvement by the high velocity of social media data could not be showed by this study because the data did not provide insight into this improvement. However, Yellowstone has a high weekly activity suggesting opportunities for a freshness improvement. The variety of the data resulted in richness improvements. Furthermore, this study indicates that the veracity of social media data for mapping POIs is high. There is a large consistency of position and attributes between social media and places data indicating that the social media data has a high trustworthiness for mapping POIs. At last, results clearly show that social media data can add value.

This research provides insight into the additional value of data since value potentials are extensively discussed within the current Big Data area, but there is no clear consensus yet. Especially, the usage of data beyond its initial purpose is an emergent field that is expected to deliver high value. The value of data as a product is expected to result in high additional value for businesses and society. For example, location service providers have largely shifted from data collection to the aggregation of data for their products. Even though industries are shifting to data-intensive industries, theoretical and empirical research of the value magnitude is lacking. Therefore, this study provides insight into the additional value of data as a product.

Outcomes show that the usage of data beyond its initial purpose can create high value for a product. This research examines reuse and aggregation as a value extraction method that unleashes the potential of data. Social media data was used to map POIs, which was not the initial purpose when creating the data. Therefore, this study clearly involves a secondary usage. Additionally, first-order value, which has the largest value potential, is created because a specific problem was addressed. The maintenance of POIs is a challenging problem, while POIs are essential for location data. Results clearly elaborate that social media contributes in addressing the maintenance challenge by improving the quality of the current places data. It is shown that data can add value by using it as a product. Therewith this research also contributes to the current challenges of mapping POIs and addresses a specific problem. Thus, this research reveals that the secondary usage of social media data can result in a first-order value.

Concluding, the study shows how social media data delivers value for location service providers in mapping points of interests. Value can be obtained by reusing data and aggregating it beyond industry boundaries. It can add value to products by improving the product advantage. Specifically, results indicate high potential categories and geographical areas for mapping POIs.

7.2 Discussion and limitations

Statements about the value of social media data for mapping POIs in previous studies are empirically supported by this research. Yellowstone includes only a fraction of places, which supports the statement that only a part of the visited places are published online (Sang, Mei, & Xu, 2015). Results clearly show that social media data can improve coverage, which confirms the statement of Kelm et al. (2013). Furthermore, McKenzie et al. (2014) state that social media data can enrich, validate semantics and improve coverage. These statements are empirically confirmed since coverage and richness are improved and attributes are confirmed by the social media data. Furthermore, the secondary usage of data can result in value by reusing and aggregating data, which confirms the statement of Mayer-Schonberger & Cukier (2013). Even though these statements are confirmed, literature showed a gap in assessing the value of data as a product. This research contributes by supporting statements with empirical evidence, providing a method to assess additional value of data as a product and providing insight into the magnitude of this value. Though, some limitations are discussed below.

This study uses customer value and therewith product advantage as value, but other value concepts exist. Customer value is often identified as an important value concept (e.g. Hassan, 2012). It is the equation between customer’s perceived benefits minus customer’s perceived sacrifices (Ivanović et al., 2013). However, results of this study are not verified by other value measures, which is a limitation of the current study. For example, there is no insight into financial value, such as ROI, profit contribution or sales volume (Griffin & Page, 1996). A financial perspective of value could provide enhanced insight into social media data value.

In the determination of value, this study does not consider requirement differences among categories. For example, the acceptable distance between the digital map and ground truth can differ among categories (e.g. mountain versus shop). Additionally, the definition of a place can differ among sources. For example, a parking lot could be considered as part of a place. These differences in customer requirements and place definition are not considered, while those aspects influence the correctness of positional and thematic accuracy. These considerations could provide a more granular insight into the value of social media data for mapping POIs.

The sample of this study includes places of social media data that were ingested in HEREs places data by means of a matching and blending algorithm. The matching and blending algorithm pre-selected data. This pre-selection omitted records that are not places and determined which places were matched. This has two consequences for this research. Firstly, valuable records could have been omitted, which could influence results. Secondly, the value of social media data is depended on the correctness of the algorithm. Incorrect matches and blends could indicate incorrect coverage and richness improvements. However, the algorithm was developed by experts and therefore assessed as sufficient. Examining the algorithm could provide additional insight in the consequences for this study.

This research examined Yellowstone and HEREs places data for mapping POIs, which has some limitations for the generalization of outcomes. Therefore, the generalization to Europe, worldwide, other data, other products and over time is discussed. Results have mainly showed two important implications: Yellowstone adds value to HEREs POIs and the extent of the value is revealed. The first implication can be generalized. It can be concluded that data can deliver first-order value by a secondary usage. However, this study does not claim that this is true for all data or all cases, but it supports the statement that data as a product can deliver value by reusing and aggregating data beyond industry boundaries. The extent of the value should be generalized with more caution. The study only assesses Yellowstone as social media data and HEREs places data, while there are many social media sources that include places and multiple location service providers. This could change the extent to which social media data contains value for location service providers. Furthermore, results cannot be extended without caution to other countries. For example, HERE has some focus-countries in West-Europe where quality is generally higher. Additionally, this research has time dependency meaning that the added value is subject to diminishing returns over time. Developments, such as the Open Location Platform strategy of HERE or the ingestion of other sources could diminish returns. Also, daily data could deliver less value than the bulk load that was tested in this study.

At last, this study does not provide insight into the separate quality of Yellowstone and HEREs places data. In case of an improvement, it is not clear whether Yellowstone has a significantly good quality or HEREs places data had a significantly bad quality. Since there is no direct comparison with the ground truth, this insight is not provided by this study.

7.3 Implications

Theoretical implications

The main theoretical implication is the method used in this study to assess the added value of data as a product. By linking traditional literature about new product development and value, the product advantage was used to examine the added value of data. This study clearly shows how the product advantage can be used to determine the added value of data as a product. This method can be used for other products as well. Therefore, this study provides a clear insight into a value assessment method of data as a product.

Another theoretical implication is the identified gap in current literature. Despite the large expectations of Big Data, there is no extensive research area that examines this potential. Researchers do not agree on the impact degree of data yet. This research provides a first insight, although a comprehensive understanding is still missing.

Practical implications

Since Yellowstone adds value to HEREs places data, it is recommended that HERE aggregates Yellowstone with its places data. Yellowstone improves all quality aspects that are assessed in this study. However, HERE should not substitute its current places data by Yellowstone since only a fraction of places is included in Yellowstone. Hence, it should aggregate Yellowstone with its places data.

Particularly, HERE should use Yellowstone as addition to its current places and attributes. Therewith, coverage and richness can be improved. Yellowstone will improve the surrounding neighborhoods and the Natural & Geographical and Administrative Areas & Buildings places most efficiently. However, it will add most places to the Business & Services, Shopping and Facilities places. HERE can obtain extra value when phone numbers and addresses are aggregated as well.

Furthermore, HERE should use Yellowstone to confirm its current geographical coordinates, address, city and category. Especially, the confirmation of attributes adds high value to its places data. This is most valuable for Eat & Drink and Going Out-Entertainment places and places in the city center of Linz, The Hague and Utrecht.

Apart from Yellowstone, HERE can attempt to create value from other social media sources as well. HERE announced its partnership with TenCent, which is a social media website in China (HERE, 2016). This could supplement Yellowstone since Yellowstone is not accessible in China. However, Tencent differs from Yellowstone because it is also a well-established provider of chat, music and game services. Despite those differences, HERE can use the approach of this study to assess the added value of Tencent data. Apart from Tencent, HERE can use the method for other sources as well.

Generally, reusing and aggregating data are good value extraction techniques of data but HERE should pay large attention to the matching and blending algorithm. HERE has transformed to a business that largely extracts its value from aggregating data. Therewith, the algorithm is essential for value extraction of data and should be extensively examined before data is aggregated with its current places data.

7.4 Recommendations for future research

Future research can extend this study by assessing other data and products. This research can examine the added value for other products and therewith confirm or reject findings. Additionally, it can test the proposed method that is used to assess value. The identified gap in literature can be further reduced by these studies that provide insight into the (potential) value of data.

Furthermore, the accuracy of added places and attributes, and the potential for correcting attributes by social media data can be examined. This study only assesses the positional and thematic accuracy by considering the consistency between Yellowstone and HEREs places data.

However, the accuracy of newly added places and attributes is uncertain. There is also no insight into the potential of Yellowstone in correcting HEREs places data. The places data can be incorrect compared to the ground truth. This study does not provide an assessment of correcting these incorrect attributes in the places data. Future research can examine the correctness of both sources in case of inconsistencies. A comparison of the data with the ground truth can reveal correct attributes. However, this assessment is time-consuming and should be thoroughly planned. Despite the time, it would create additional insight into the value of social media data for mapping POIs.

At last, future research can enhance the current the model by detailing the product advantage statement. The product advantage can be researched by means of qualitative methods, such as customer interviews. For example, requirement differences between categories can be included in the product advantage. Therewith, the product advantage can be detailed, which improves insight into the added value of data as a product.

8 References

Ali E. Akgün, Gary S. Lynn, & John C. Byrne. (2004). Taking the guesswork out of new product development: how successful high-tech companies get that way. *Journal of Business Strategy*, *25*(4), 41–46. https://doi.org/10.1108/02756660410547395

Amankwah-Amoah, J. (2016). Emerging economies, emerging challenges: Mobilising and capturing value from big data. *Technological Forecasting and Social Change*, *110*, 167–174. https://doi.org/10.1016/j.techfore.2015.10.022

Antoniou, V., & Skopeliti, A. (2015). Measures and Indicators of VGI Quality: An Overview. In *ResearchGate* (Vol. II-3/W5). https://doi.org/10.5194/isprsannals-II-3-W5-345-2015

Arsanjani, J. J., Barron, C., Bakillah, M., & Helbich, M. (2013). Assessing the Quality of OpenStreetMap Contributors together with their Contributions. Retrieved from https://core.ac.uk/display/23652698

Barron, C., Neis, P., & Zipf, A. (2014). A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, *18*(6), 877–895. https://doi.org/10.1111/tgis.12073

Bharadwaj, N., & Noble, C. H. (2015). Innovation in Data-Rich Environments. *Journal of Product Innovation Management*, *32*(3), 476–478. https://doi.org/10.1111/jpim.12266

Bhuiyan, N. (2011). A Framework for successful new product development. *Journal of Industrial Engineering and Management*, *4*(4), 746–770. https://doi.org/10.3926/jiem.334

Booz, Allen, & Hamilton. (1982). *New Products Management for the 1980s*. Booz, Allen and Hamilton Inc.

Boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Cankurtaran, P., Langerak, F., & Griffin, A. (2013). Consequences of New Product Development Speed: A Meta-Analysis. *Journal of Product Innovation Management*, *30*(3), 465–486. https://doi.org/10.1111/jpim.12011

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 675–684). New York, NY, USA: ACM. https://doi.org/10.1145/1963405.1963500

CBS. (2016, December 5). CBS StatLine - Overnight accommodation; guests, overnight stays, occupancy, key figures. Retrieved January 9, 2017, from http://statline.cbs.nl/Statweb/publication/?DM=SLEN&PA=82058ENG&D1=4,7-14,22-42&D2=36-47&LA=EN&HDR=G1&STB=T&VW=T

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS Quarterly*, *36*(4), 1165–1188.

Chesbrough, H., & Rosenbloom, R. S. (2002). The role of the business model in capturing value from innovation: evidence from Xerox Corporation’s technology spin-off companies. *Industrial and Corporate Change*, *11*(3), 529–555. https://doi.org/10.1093/icc/11.3.529

Chuang, H.-M., & Chang, C.-H. (2015). Verification of POI and Location Pairs via Weakly Labeled Web Data (pp. 743–748). Presented at the Proceedings of the 24th International Conference on World Wide Web, ACM. https://doi.org/10.1145/2740908.2741715

Cooper, R. G. (2008). Perspective: The Stage-Gate® Idea-to-Launch Process—Update, What’s New, and NexGen Systems*. *Journal of Product Innovation Management*, *25*(3), 213–232. https://doi.org/10.1111/j.1540-5885.2008.00296.x

Crawford, C. M., & Benedetto, C. A. D. (2014). *New Products Management* (10 edition). New York, NY: McGraw-Hill Education.

Eling, K., Langerak, F., & Griffin, A. (2013). A Stage-Wise Approach to Exploring Performance Effects of Cycle Time Reduction. *Journal of Product Innovation Management*, *30*(4), 626–641. https://doi.org/10.1111/jpim.12019

Ernst, H., Hoyer, W. D., & Rübsaamen, C. (2010). Sales, Marketing, and Research-and-Development Cooperation Across New Product Development Stages: Implications for Success. *Journal of Marketing*, *74*(5), 80–92.

European Commission. (2012). Road Transport: A change of gear. Retrieved September 20, 2016, from http://ec.europa.eu/transport/modes/road/doc/broch-road-transport_en.pdf

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gartner. (2014, March 19). Gartner Says the Internet of Things Will Transform the Data Center. Retrieved September 2, 2016, from http://www.gartner.com/newsroom/id/2684616

Girres, J.-F., & Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, *14*(4), 435–459. https://doi.org/10.1111/j.1467-9671.2010.01203.x

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

Griffin, A., & Page, A. L. (1996). PDMA success measurement project: Recommended measures for product development success and failure. *Journal of Product Innovation Management*, *13*(6), 478–496. https://doi.org/10.1016/S0737-6782(96)00052-5

Hassan, A. (2012). The Value Proposition Concept in Marketing: How Customers Perceive the Value Delivered by Firms– A Study of Customer Perspectives on Supermarkets in Southampton in the United Kingdom. *International Journal of Marketing Studies*, *4*(3), 68. https://doi.org/10.5539/ijms.v4n3p68

Henard, D. H., & Szymanski, D. M. (2001). Why some new products are more successful than others. *JMR, Journal of Marketing Research*, *38*(3), 362–375.

HERE. (2016, December 27). HERE gets new investors and plans expansion into China. Retrieved from http://360.here.com/2016/12/27/here-gets-new-investors-to-expand-into-china/

Hilbert, M. (2016). Big Data for Development: A Review of Promises and Challenges. *Development Policy Review*, *34*(1), 135–174. https://doi.org/10.1111/dpr.12142

Huberty, M. (2015). Awaiting the Second Big Data Revolution: From Digital Noise to Value Creation. *Journal of Industry, Competition and Trade*, *15*(1), 35–47. https://doi.org/10.1007/s10842-014-0190-4

Inc, Z. (2016, November 27). Zillow: Real Estate, Apartments, Mortgages & Home Values. Retrieved November 27, 2016, from http://www.facebook.com/zillow

International Organization for Standardization (ISO). (2013). *ISO 19157:2013 - Geographic information - Data quality*. Geneva, Switzerland: ISO. Retrieved from http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32575

Internet Live Stats. (2016, September 17). Twitter Usage Statistics - Internet Live Stats. Retrieved September 18, 2016, from http://www.internetlivestats.com/twitter-statistics/

Irving, L. (2014, February 6). This is How Big Data Can Solve the World’s Biggest Issues. Retrieved September 2, 2016, from http://www.bigdatalandscape.com/blog/big-data-can-solve-worlds-biggest-issues

Ivanović, A., America, P., & Snijders, C. (2013). Modeling customer-centric value of system architecture investments. *Software & Systems Modeling*, *12*(2), 369–385. https://doi.org/10.1007/s10270-012-0235-2

Jaccard, M. (2013). *The Objective is Quality: An Introduction to Performance and Sustainability Management Systems*. Lausanne, Switzerland: EPFL Press.

Jonietz, D., & Zipf, A. (2016). Defining Fitness-for-Use for Crowdsourced Points of Interest (POI). *ISPRS International Journal of Geo-Information*, *5*(9), 149. https://doi.org/10.3390/ijgi5090149

Kelm, P., Murdock, V., Schmiedeke, S., Schockaert, S., Serdyukov, P., & Laere, O. van. (2013). Georeferencing in Social Networks. In N. Ramzan, R. van Zwol, J.-S. Lee, K. Clüver, &

X.-S. Hua (Eds.), *Social Media Retrieval* (pp. 115–141). Springer London. Retrieved from http://link.springer.com.dianus.libr.tue.nl/chapter/10.1007/978-1-4471-4555-4_6

Khurana, A., & Rosenthal, S. R. (1998). Towards Holistic “Front Ends” In New Product Development. *Journal of Product Innovation Management*, *15*(1), 57–74. https://doi.org/10.1111/1540-5885.1510057

Kim, J., & Wilemon, D. (2002). Focusing the fuzzy front–end in new product development. *R&D Management*, *32*(4), 269–279. https://doi.org/10.1111/1467-9310.00259

Li, X., Xu, G., Chen, E., & Zong, Y. (2015). Learning recency based comparative choice towards point-of-interest recommendation. *Expert Systems with Applications*, *42*(9), 4274–4283. https://doi.org/10.1016/j.eswa.2015.01.054

Lin, J., Oentaryo, R., Lim, E.-P., Vu, C., Vu, A., & Kwee, A. (2016). Where is the Goldmine? Finding Promising Business Locations through Facebook Data Analytics. In *ResearchGate*. https://doi.org/10.1145/2914586.2914588

Lu, X., Yu, Z., Sun, L., Liu, C., Xiong, H., & Guan, C. (2016). Characterizing the life cycle of point of interests using human mobility patterns (pp. 1052–1063). ACM Press. https://doi.org/10.1145/2971648.2971749

Mayer-Schonberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. Hodder And Stoughton.

McKenzie, G., Janowicz, K., & Adams, B. (2014). A weighted multi-attribute method for matching user-generated Points of Interest. *Cartography and Geographic Information Science*, *41*(2), 125–137. https://doi.org/10.1080/15230406.2014.880327

Mullen, W. F., Jackson, S. P., Croitoru, A., Crooks, A., Stefanidis, A., & Agouris, P. (2014). Assessing the impact of demographic characteristics on spatial error in volunteered geographic information features. *GeoJournal*, *80*(4), 587–605. https://doi.org/10.1007/s10708-014-9564-8

Nambisan, S. (2010). The Role of Information Technology in Product Development: An Introduction. In S. Nambisan (Ed.), *Information Technology and Product Development* (pp. 1–16). Springer US. Retrieved from http://link.springer.com.dianus.libr.tue.nl/chapter/10.1007/978-1-4419-1081-3_1

Papadopoulos, S., Bontcheva, K., Jaho, E., Lupu, M., & Castillo, C. (2016). Overview of the Special Issue on Trust and Veracity of Information in Social Media. *ACM Trans. Inf. Syst.*, *34*(3), 14:1–14:5. https://doi.org/10.1145/2870630

Parmar, R., Mackenzie, I., Cohn, D., & Gann, D. (2014, January 1). The New Patterns of Innovation. Retrieved September 6, 2016, from https://hbr.org/2014/01/the-new-patterns-of-innovation

Phaal, R., Farrukh, C. J. P., & Probert, D. R. (2004). Technology roadmapping—A planning framework for evolution and revolution. *Technological Forecasting and Social Change*, *71*(1–2), 5–26. https://doi.org/10.1016/S0040-1625(03)00072-6

Polous, K., Freitag, A., Krisp, J., Meng, L., & Singh, S. (2015). A General Framework For Event Detection From Social Media. In F. Harvey & Y. Leung (Eds.), *Advances in Spatial Data Handling and Analysis* (pp. 85–105). Cham: Springer International Publishing. Retrieved from http://link.springer.com/10.1007/978-3-319-19950-4_6

R. G. Cooper, & Kleinschmidt, E. J. (1987). New products: What separates winners from losers? *Journal of Product Innovation Management*, *4*(3), 169–184. https://doi.org/10.1016/0737-6782(87)90002-6

Remenyi, D., Williams, B., Money, A., & Swartz, E. (1998). *Doing Research in Business and Management: An Introduction to Process and Method*. SAGE.

Roberts, D. L., & Candi, M. (2014). Leveraging Social Network Sites in New Product Development: Opportunity or Hype? *Journal of Product Innovation Management*, *31*, 105–117. https://doi.org/10.1111/jpim.12195

Sang, J., Mei, T., & Xu, C. (2015a). Activity Sensor: Check-In Usage Mining for Local Recommendation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *6*(3), 41. https://doi.org/10.1145/2700468

Sang, J., Mei, T., & Xu, C. (2015b). Activity Sensor: Check-In Usage Mining for Local Recommendation. *ACM Trans. Intell. Syst. Technol.*, *6*(3), 41:1–41:24. https://doi.org/10.1145/2700468

Sivasubramaniam, N., Liebowitz, S. J., & Lackman, C. L. (2012). Determinants of New Product Development Team Performance: A Meta-analytic Review. *Journal of Product Innovation Management*, *29*(5), 803–820. https://doi.org/10.1111/j.1540-5885.2012.00940.x

Smith, C. (2014, March 13). By the Numbers: 400+ Amazing Facebook Statistics. Retrieved September 13, 2016, from http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/4/

Stahl, H. K., Matzler, K., & Hinterhuber, H. H. (2003). Linking customer lifetime value with shareholder value. *Industrial Marketing Management*, *32*(4), 267–279. https://doi.org/10.1016/S0019-8501(02)00188-8

Statistics Austria. (2016, December 28). Arrivals, Overnight Stays. Retrieved January 9, 2017, from http://www.statistik.at/web_en/statistics/Economy/tourism/accommodation/arrivals_overnight_stays/index.html

Stirling, G. (2014, April 9). Study: 78 Percent Of Local-Mobile Searches Result In Offline Purchases. Retrieved October 27, 2016, from http://searchengineland.com/study-78-percent-local-mobile-searches-result-offline-purchases-188660

The World Bank. (2016). World Bank Group. Retrieved January 17, 2017, from http://www.worldbank.org

Winter, J. C. F. de, & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*(4), 695–710. https://doi.org/10.1080/02664763.2011.610445

Woodall, T. (2003). Conceptualising “Value for the Customer”: An Attributional, Structural and Dispositional Analysis. *Academy of Marketing Science Review*, *2003*, 1.

Wulfhorst, G., & Klug, S. (2016). Sustainable Mobility in the Metropolitan Region of Munich: An Introduction. In G. Wulfhorst & S. Klug (Eds.), *Sustainable Mobility in Metropolitan Regions* (pp. 1–15). Springer Fachmedien Wiesbaden. Retrieved from http://link.springer.com.dianus.libr.tue.nl/chapter/10.1007/978-3-658-14428-9_1

Wyrwoll, C. (2014). User-Generated Content. In *Social Media* (pp. 11–45). Springer Fachmedien Wiesbaden. Retrieved from http://link.springer.com.dianus.libr.tue.nl/chapter/10.1007/978-3-658-06984-1_2

Yellowstone. (2016). Audience Insights. Retrieved January 17, 2017, from https://www.yellowstone.com/ads/audience-insights/geography?act=601428306600033&age=18-&country=AT-NL

Yin, H., & Cui, B. (2016). Location-Based and Real-Time Recommendation. In H. Yin & B. Cui, *Spatio-Temporal Recommendation in Social Media* (pp. 65–98). Singapore: Springer Singapore. Retrieved from http://link.springer.com/10.1007/978-981-10-0748-4_4

Ying, J. J.-C., Kuo, W.-N., Tseng, V. S., & Lu, E. H.-C. (2014). Mining User Check-In Behavior with a Random Walk for Urban Point-of-Interest Recommendations. *ACM Transactions on Intelligent Systems and Technology*, *5*(3), 1–26. https://doi.org/10.1145/2523068

Zott, C., & Amit, R. (2008). The fit between product market strategy and business model: implications for firm performance. *Strategic Management Journal*, *29*(1), 1–26. https://doi.org/10.1002/smj.642

I. Appendix – Sample size of categories

Table I-1 Sample size of places in each category

Categories	**Sample size Austria**	**Sample size the Netherlands**
Accommodation	12388	11302
Administrative Areas & Buildings	600	1330
Business & Services	48863	154994
Eat & Drink	23007	56236
Facilities	15339	54803
Going Out-Entertainment	13451	22532
Leisure & Outdoor	2368	6913
Natural & Geographical	1162	333
Shopping	16147	103407
Sights & Museums	4884	11971
Transport	1347	3550

Table I-2 Sample size of existing places in each category

Categories	**Sample size Austria**	**Sample size the Netherlands**
Accommodation	7897	5880
Administrative Areas & Buildings	109	240
Business & Services	12164	34864
Eat & Drink	12840	34869
Facilities	2424	11750
Going Out-Entertainment	4719	8458
Leisure & Outdoor	625	1528
Natural & Geographical	181	46
Shopping	3226	21354
Sights & Museums	1655	3261
Transport	340	780

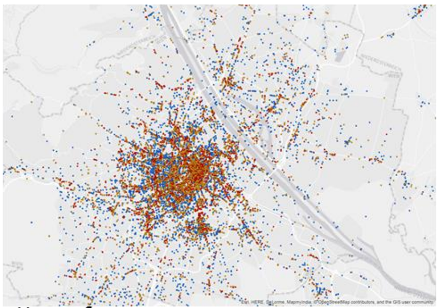
II. Appendix – Absolute coverage improvement

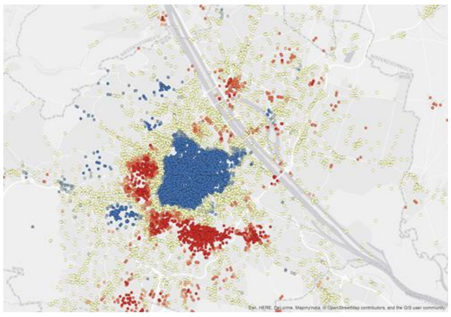
Table II-1 Overview of absolute added places by Yellowstone to HEREs places data

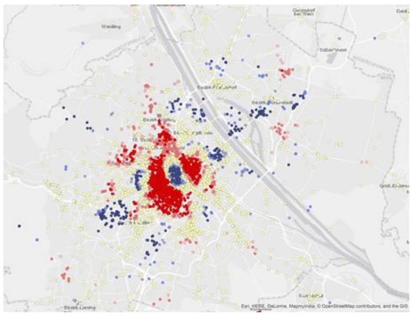
Category name	AUT - Number of added places	NLD - Absolute number of added places
Business & Services	36699	120130
Shopping	12921	82053
Facilities	12915	43053
Eat & Drink	10167	21367
Going Out-Entertainment	8732	14074
Accommodation	4491	5422
Sights & Museum	3229	8710
Leisure & Outdoor	1743	5385
Transport	1007	2770
Natural & Geographical	981	287
Administrative Areas & Buildings	491	1090

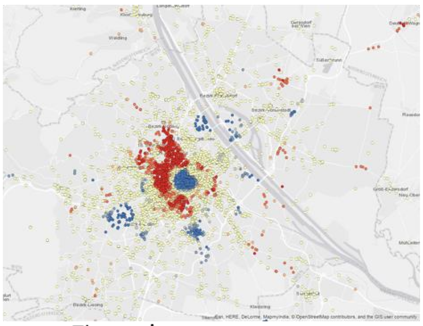
III. Appendix – Geographical analysis of quality measures and usage

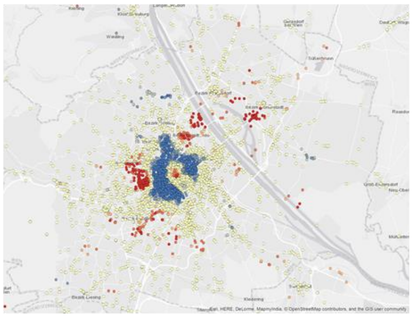
Austria – Vienna


Usage frequency

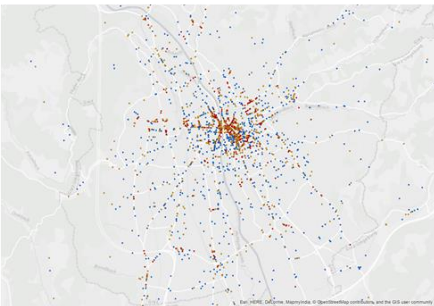

Coverage

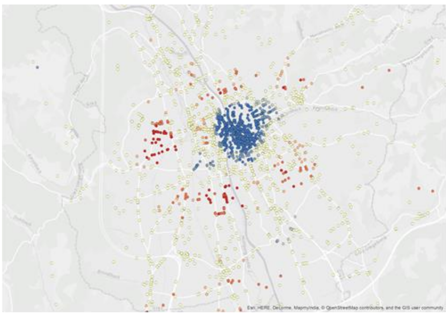

Positional accuracy

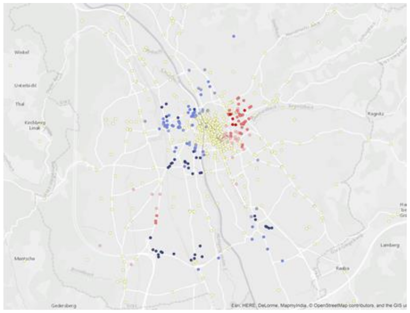

Thematic accuracy

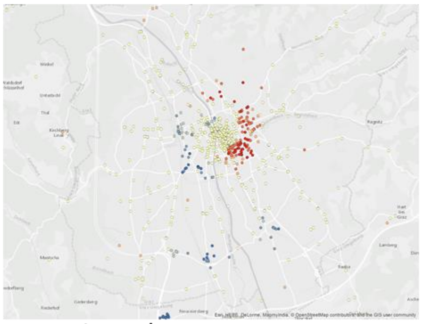

Richness

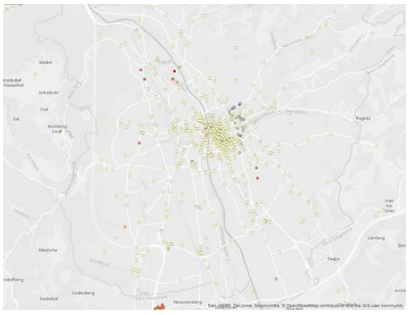
Austria – Graz


Usage frequency

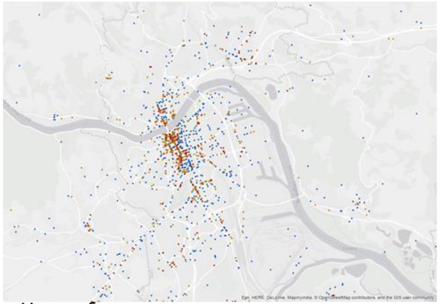

Coverage

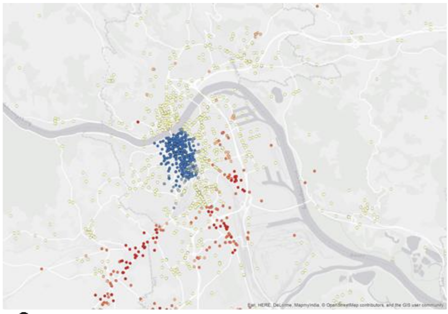

Positional accuracy

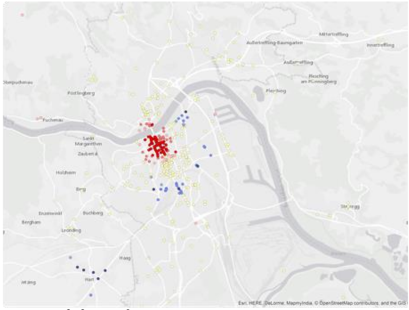

Thematic accuracy

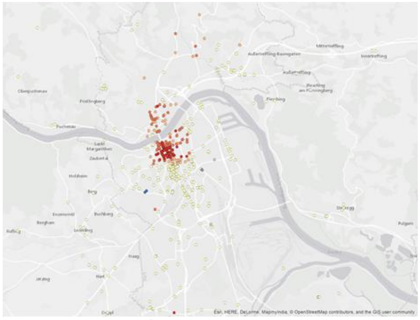

Richness

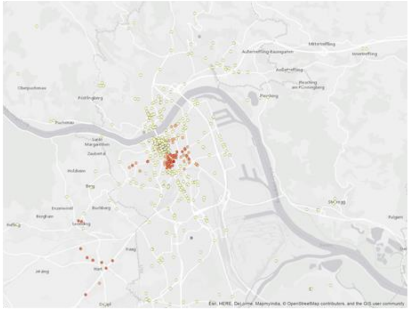
Austria - Linz


Usage frequency

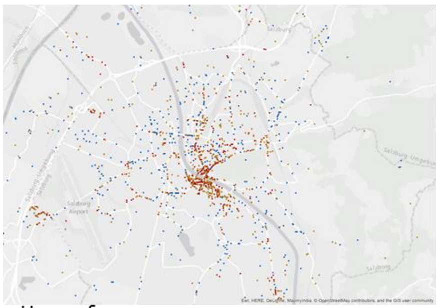

Coverage

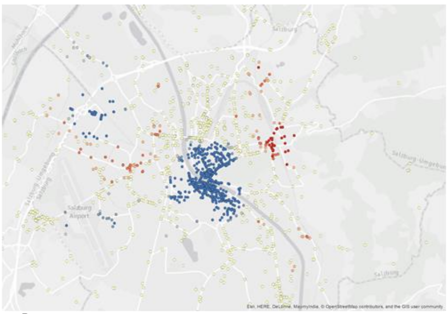

Positional accuracy

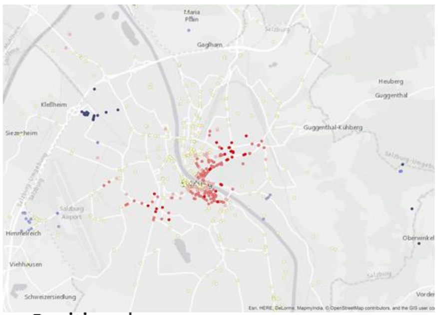

Thematic accuracy

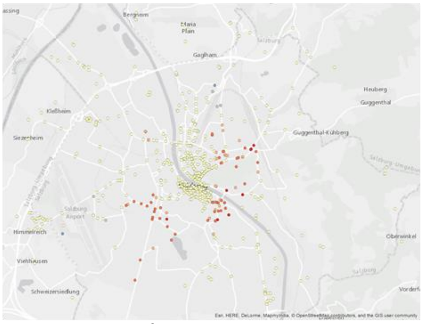

Richness

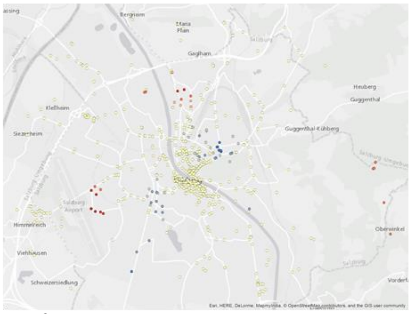
Austria – Salzburg


Usage frequency

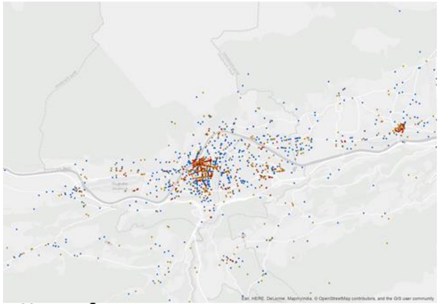

Coverage


Positional accuracy

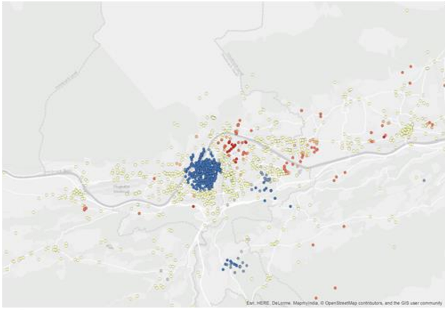

Thematic accuracy


Richness

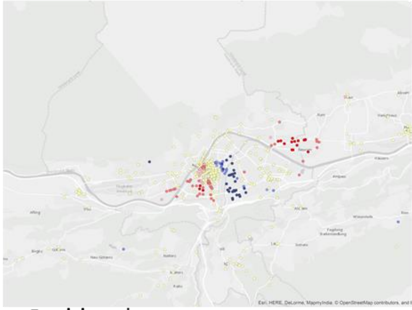
Austria – Innsbruck



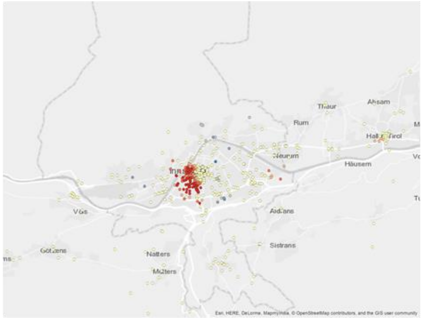
Usage frequency



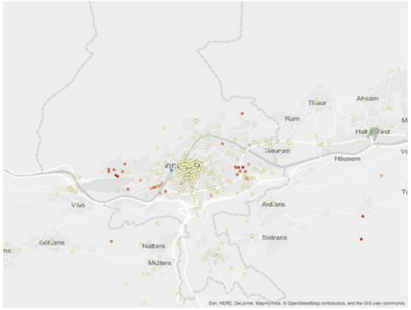
Coverage



Positional accuracy

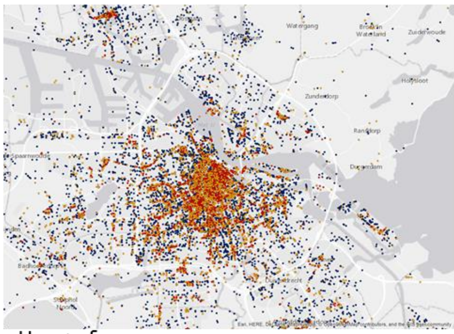


Thematic accuracy

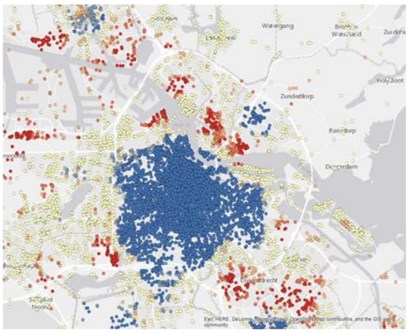


Richness

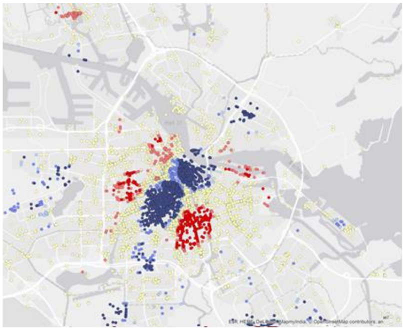
Netherlands – Amsterdam



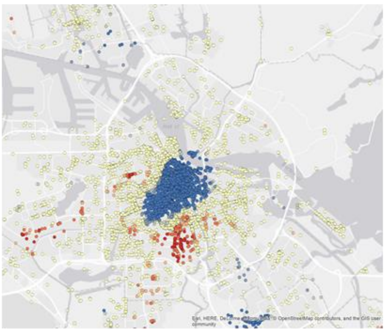
Usage frequency



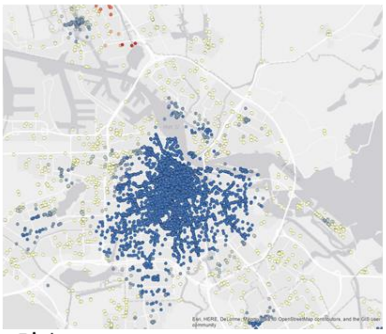
Coverage



Positional accuracy

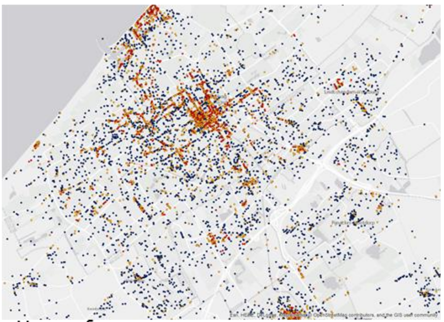


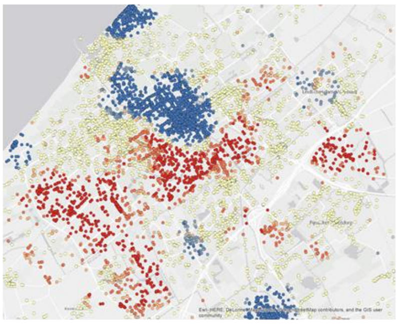
Thematic accuracy

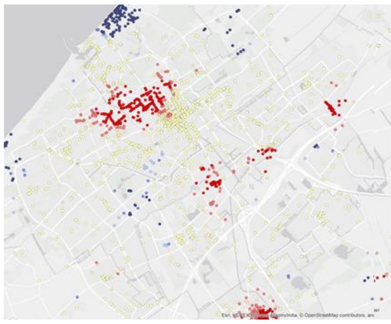


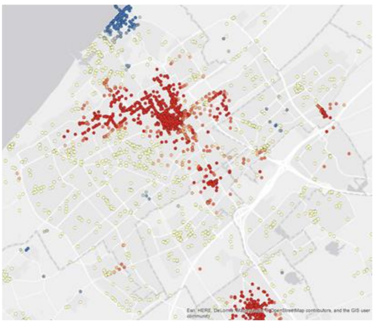
Richness

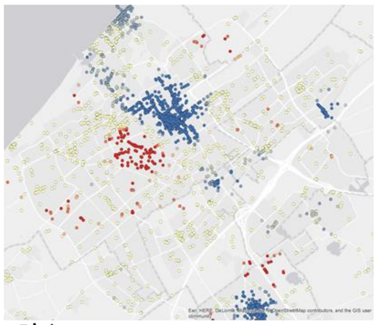
Netherlands - The Hague


Usage frequency

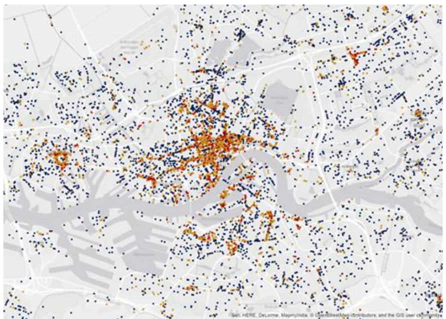

Coverage

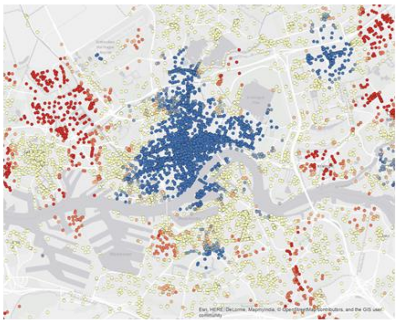

Positional accuracy

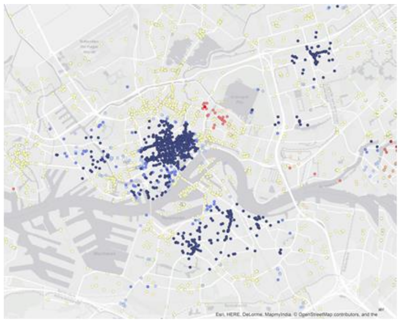

Thematic accuracy

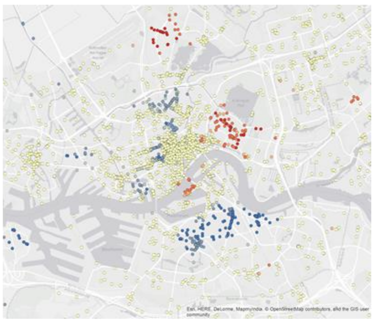

Richness

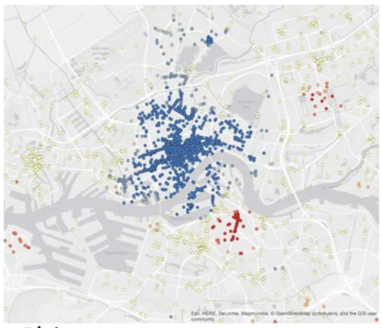
Netherlands - Rotterdam


Usage frequency

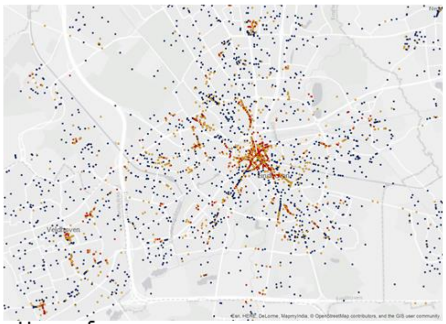

Coverage

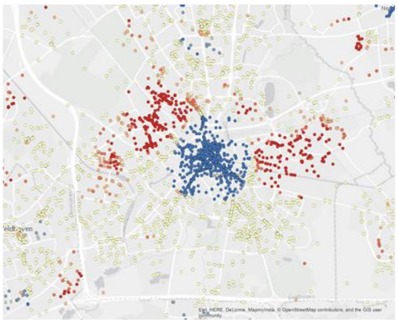

Positional accuracy

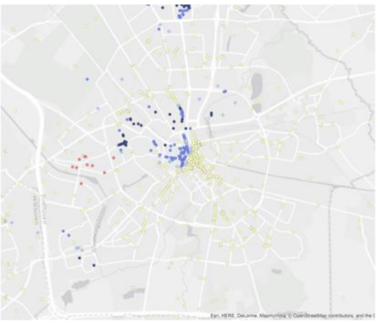

Thematic accuracy

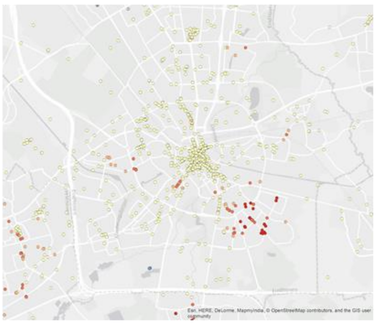

Richness

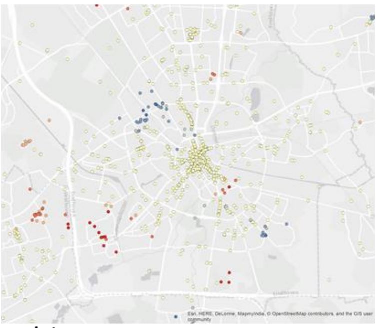
Netherlands - Eindhoven


Usage frequency

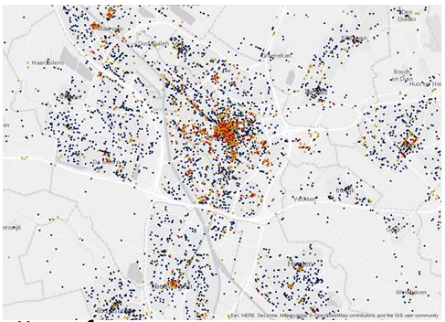

Coverage

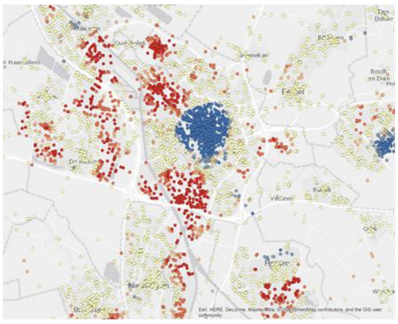

Positional accuracy

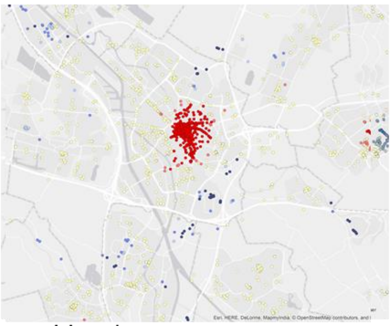

Thematic accuracy

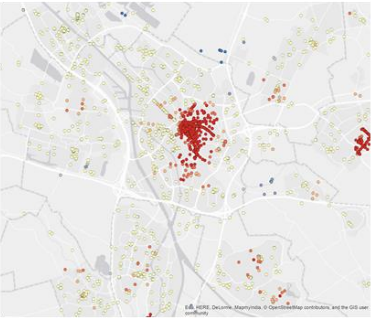

Richness

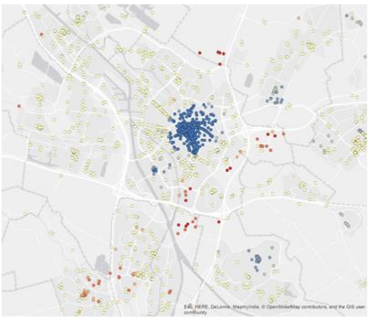
Netherlands - Utrecht


Usage frequency


Coverage


Positional accuracy


Thematic accuracy


Richness