Janja Jeznik, BSc

Evaluating geo-tagged Twitter Data to analyse Tourist Flows in Styria

MASTER'S THESIS

to achieve the university degree of
Master of Sciences
Master's degree programme: Geospatial Technologies

submitted to
Graz University of Technology

Advisor: Johannes Scholz, Ass.Prof. Dipl.-Ing. (FH)

Institute of Geodesy
Graz, March 2019

EIDESSTATTLICHE ERKLÄRUNG

AFFIDAVIT

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Datum / Date	Unterschrift / Signature

Abstract

Within our master thesis a broad selection of analyses were implemented, all in order to confirm adequacy of Twitter data in statistical, geospatial and semantic aspect. Only due to the possibility of acquiring spatial related Twitter data it is possible to implement researches related to a particular spacial scale or to a particular region. We were working on a very fine spatial scale, namely state Styria in Austria. Data was acquired with Twitterscraper API on a municipality level and evaluation was performed on a district level. Extracted tweets in the time period from 2008 till middle of 2018 were stored and submitted to extensive filtering process within noSQL database MongoDB. 80% of our dataset was filtered out in the process of determining tweets with relation to tourism. As main implementations within these thesis count spatial-temporal and semantic analyses. Spatial distribution and clustering patterns were examined with Hotspot analysis and Kernel Density estimation method. The final evaluation based on correlation of amount of extracted tweets through the years from 2008 and 2017 in Styria and also on district and seasonal level. Statistically significant correlations between our and reference data confirmed connection between the data and usability of Twitter for researches on such a fine spatial scale for tourism purposes. Sentiment analysis proved additional value of Twitter data, since we were able to collect user's opinions about their touristic destinations. Apart from that we dealt also with usability of official Twitter REST API and further text mining methods.

Zusammenfassung

Im Rahmen dieser Masterarbeit wurde eine breite Auswahl an Analysen implementiert, um die Plausibilität und Eignung von Twitter-Daten in Bezug auf statistische, räumliche und semantische Aspekte zu überprüfen. Nur aufgrund der Möglichkeit, räumlich verortete Twitter-Daten zu erfassen, können Untersuchungen durchgeführt werden, die sich auf eine bestimmte räumliche Skala oder eine bestimmte Region beziehen. Wir fokussierten uns auf eine sehr feine räumliche Skala und zwar auf das Bundesland Steiermark in Österreich. Die Daten wurden mit Twitterscraper API auf Gemeindeebene erfasst und auf Bezirksebene ausgewertet. Extrahierte Tweets im Zeitraum von 2008 bis Mitte 2018 wurden gespeichert und einem umfangreichen Filterprozess in der NoSQL-Datenbank MongoDB unterzogen. 80% des Datensatzes wurden bei der Ermittlung von Tourismus relevanten Tweets herausgefiltert. Als Hauptimplementierungen innerhalb dieser Arbeit zählen räumlich-zeitliche und semantische Analysen. Die Hotspot-Analyse und die Kern density estimation Methode dienten zur Untersuchung der räumlichen Verteilung und zur Identifikation von räumlichen Clustern. Die abschließende Bewertung basiert auf der Korrelation der Menge der extrahierten Tweets in den Jahren von 2008 bis 2017 in der Steiermark sowie auf Bezirksebene unter Betrachtung der ganzjährigen Entwicklung, der Winter- und Sommersaison. Statistisch signifikante Korrelationen zwischen unseren Daten und den Referenzdaten bestätigten den Zusammenhang zwischen diesen und somit den sinnvollen Nutzen von Twitter für Forschungen auf solch feiner räumlicher Skala für touristische Zwecke. Die Sentiment-Analyse wurde als zusätzlicher Wert für Twitter-Daten hinzugezogen, um die Meinungen der Nutzer über ihre touristischen Ziele zu betrachten. Daneben befassten wir uns sowohl mit der Nützlichkeit der offiziellen Twitter REST API als auch mit weiteren Text-Mining-Methoden.

Acknowledgements

I would like to express my very great appreciation to my supervisor Mr. Johannes Scholz for your great assistance. Thank you for your prompt answers as well. Thank you also to all the colleges who accompanied me through my studying years.

I am also particularly grateful for extraordinary support of my parents who stand at my side both in good and bad moments.

For my parents:

Zelo sem hvaležna tudi mojim staršem, ki so mi stali ob strani tako v dobrih kot tudi slabih trenutkih.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Social media such as Twitter enables user communication and the sharing of their state of mind, behaviour or their activities. In addition, there is also possibility to provide a current position or location of each tweet. From a geographical aspect, Twitter conveniently provides real time geo-data directly from the users, unlike the official data collections with postponed availability. Various spatial analyses may be performed with collected geographical data, such as identifying trends within the data to obtain information on locations with most twitter users or analysing their movement flows. Within this master thesis we intend to take advantage of the freely available twitter data and analyse the tourist flows in Styria, a state in Austria. Since 2006 [1] when geolocation of tweets was enabled by Twitter, Twitter users have contributed with their real time geo-data, which has led to a big selection of Twitter-related research to take place, especially those on global scale. In contrast to many research on global scale, our study focuses on the reliability of Twitter data on a regional scale.

The investigation of obtained data will focus on tourist spatial patterns recognition and evaluation of geo-tagged Twitter data reliability and adequacy in contrast to the available data of the official statistics. Additionally, text mining techniques are used in order to perform a sentiment analysis and determine tourist attitudes to the visited locations in Styria. The plausibility and accuracy of results obtained with Twitter data are evaluated against official statistics on tourism [2], [3], which play a role of ground truth data. The results were visualized with maps and presented in graph forms.

1.1 Research goals

The main goal of this thesis is to be able to determine the plausibility and accuracy of the Twitter data on the target topic of tourism on a regional scale. In addition to the main goal, the following set of task had to be achieved:

- Literature analysis of data acquisition, processing, spatial and semantic

analysis of Twitter data.

- Development of a methodology for spatial and semantic analysis of Twitter data for tourism purposes.
- Application of developed methodology on selected test area.
- Evaluation, Visualization and Interpretation of results.

Due to research on regional scale, the selected area of application is an administrative region in Austria - Styria, which has an area of 16,388 km^2. It is located in south-east Austria with approx. 1.2 million inhabitants [4].

1.2 Research questions

Within the evaluation of the obtained geo-tagged data, a selection of research questions was determined. These were focused on the spatial and semantic analysis of tourist flows, determined via geo-tagged Twitter data. Additionally, a research question of filtering methodology at the beginning of the analysis and an evaluation methodology at the end had to be taken into consideration. All research questions refer to our target topic tourism and are listed in the following two subsections. All the results were visualized, interpreted and evaluated in comparison against to official statistics on tourism.

Furthermore, it has to be stressed, that a variety of applications are offered in order to analyse tourism flows, from simple spatial distributions to more complex ones such as identifying mobility trajectories of observed tourists. Some of the possibilities are introduced in the following literature research chapter. However, due to the scale of our master thesis a limited selection of the most important research questions were chosen to be taken into consideration.

1.2.1 Research questions - spatial analysis

Spatial-statistical analyses were focused on the following investigations:

- Determining most visited locations or sites.
- Generation of a Heat Map and performing a Hot Spot analysis.
- Growth of the available data from 2006 to August 2018.

- Seasonal comparison of available data through the years.
- Evaluation, Visualization and Interpretation of results.
- Usability of Twitter data acquired with official Twitter REST API.

1.2.2 Research questions - semantic analysis

To perform semantic text-analysis the first step was to filter-out tourism unrelated data. The second step, the main part of the analysis, evaluated the sentiment of tourists expressed in their tweets. An analysis of the spatial distribution of their positive or negative attitude within or towards the region was performed as well. Additionally, we also discussed the strength of topic modelling algorithms when dealing with Twitter data.

- Filtering the tweets on the target topic of tourism.
- Sentiment determination of each individual tweet.
- Investigation of spatial sentiment patterns within the region.
- Usability of Topic Modelling with a selection of algorithms in order to determine the types of tourism.

1.3 Methodology

Development of the methodology for spatial and semantic analysis for tourism purposes is one of the goals of this master thesis. In this chapter the methodology of each step of our research is briefly introduced only, since it is going to be explained into more details in the theoretical section. First, our chosen methods to data acquisition, processing and filtering are introduced, followed by employed methods for spatial and semantic analysis as well as for visualisation, interpretation and evaluation. We would like to stress, that this research did not follow any existent predetermined methodological workflow. On the contrary, the most appropriate methods were chosen within the working process after a closer look and a simple evaluation.

1.3.1 Data acquisition

Our data acquisition procedure consisted of two approaches. First, the official Twitter REST API for open source data acquisition was taken into

consideration. Due to its official restrictions on achieving historical tweets older than 7 days [5] a Python package Twitterscraper [6] was selected as a very adequate alternative, since it enables collecting historical tweets from Twitter's establishment in 2006 onwards. However, although REST API was not selected as our data acquisition mediator, we still connected to it each week to request weekly data and collected the data from mid August 2018 to the end of December 2018. With the acquired data we explored and discussed the quantity of data delivery collected via this REST program application interface and therefore its indirect usability for data collection for research on regional scale.

In order to request historical tweets from the beginning of its conception in 2006 but not ordering them at Twitter organization itself and causing us high economical expenses, we therefore decided for a Twitterscraper. Twitterscrapper is a Python package available for free distribution on GitHub [6]. It enables to request tweets based on spatial location, but no exact coordinates are returned. To overcome this drawback, we decided to request and to analyse the data on municipality-level. Tweets are in our case scraped based on several arguments, such as location and various tourism-related keywords. Additional possible scraping arguments would be for instance start-date and end-date, user name or limit of the tweets [6].

1.3.2 Data processing and filtering

For storing purposes of the extracted data, we used a non-relational MongoDB database, based on NoSQL query language with collections and documents within it [7]. In our case a document corresponds to a tweet. Due to the fact that the tweets on the municipality scale were returned without the coordinates, we updated each document the coordinates of the municipality centroid. Such geo-coded data are ready for further spatial analysis. Next, the tweets with no relevance to tourism were filtered out. For the filtering purposes the tweets were distributed into three collections - namely into three groups of tweets according to the estimated level of their tourism relevance. Exact filtering approach is explained into detail in the experiment chapter.

1.3.3 Spatial analysis

Spatial analyses of processed data were implemented with the help of geographic information systems - GIS. We used a leading GIS software ArcGIS from ESRI and open source QGIS. The data was captured on a munici-

pality scale, therefore any spatial analysis undertaken was using the same spatial resolution. However, within each municipality, tweets were randomly distributed, since we assigned coordinates of municipality's centroids in the first place. Spatial distribution, density and clusters were investigated with the help of heat map and Hot Spot analysis. Heat map was generated with the Kernel density estimation method. Furthermore, the spatial distribution of our established three groups of tweets was also taken into consideration in order to analyse the patterns relative to each other and within them.

1.3.4 Semantic text analysis

Text analysis helped us at the final filtering in order to extract tourism-related tweets. We used text mining for a word frequency determination and visualized them in a word cloud with the R programming language. With the help of more key-words dictionaries and according to our professional knowledge we determined the final dataset of used tourism relevant key-words. Within semantic text-analysis a sentiment of each tweet was extracted in order to determine the attitude of tourists towards visited sites. Polarity of inputs was obtained within the open source software for data mining "Orange" with the VADER method, which was chosen as the most appropriate since it works well with multilingual datasets and also determines the intensity of a sentiment [8]. Categorized results as positive, negative or neutral attitude may be used to point out the problematic tourist regions or unsatisfactory tourism supply in the municipalities in Styria.

Additionally, we discussed also Topic Modelling techniques and experimentally implemented it within the Orange software. With the topic modelling we aimed to categorize tourism types expressed in the tweets. With this approach we wished to evaluate if the obtained types of tourism in particular municipality correlate with the represented types according to the official statistical data. However, Topic Modelling was only an additional experiment and was not part of our main objectives. With our dataset, topic modelling did not provide any tangible results, but was nevertheless discussed in the practical part of this thesis.

1.3.5 Visualization and interpretation of results

For mapping and visualisation purposes the GIS Software ArcGIS and QGIS were used. The results are represented with the help of cartographic visualisation methods such as a choropleth map, dot-distribution maps and maps with proportional symbols. To display numerical data, column graphs, line

graphs and scatter plots were applied.

1.3.6 Evaluation

The accuracy of results obtained with Twitter data were evaluated with official statistics on tourism from Statistics Austria [2] and Styrian Statistics [3], which play a role of ground truth data. We used the data between 2008 and 2017, as our Twitter data from 2018 is only available until August 2018 and not the whole tourism year. The correlation of the spatial distributions within the region was investigated in order to determine the suitability of Twitter data for studies on a target topic of tourism on a regional scale. Pearsons correlation coefficient was used for determining association between our data set and official statistical data. We evaluated the data at a district level, as a result of the community structural reform that took place in-between 2010 and 2015, that grafually reorganized 542 municipalities into 287 [9]. Furthermore, not all municipalities are included in the official accommodation statistics, since only those of at least 1000 arrivals are obligated to report the data [10] and the number of reporting communities is therefore varying from year to a year.

1.4 Literature overview

For this master thesis a broad literature research had to take place since our topic includes different fields of research. In general, the topics of Geography of Tourism and Geographic information Systems (GIS) had to be studied. VGI – Volunteered geographic information – had to be discussed in detail regarding open social media data. However, during our literature research we were primarily focused on the methodology of data acquisition and their applications. We examined other existing studies discussing possible spatial and semantic analysis both in tourism as well as in other research fields. Therefore, in this chapter of the literature overview, mostly the literature on Twitter application and Twitter geo-data are presented. More about theoretical principles of social media and open geo-data will be discussed within the following chapter exploring theoretical principles.

Since the launch of Twitter in 2006 [1] a big collection of research regarding the usage of Twitter data has taken place. A recent and extensive study on Twitter Data acquisition and analysis is the work from 2018 "Twitter as Data" of Zachary and Steinert-Threlkeld [11]. Another comprehensive work on Twitter data acquisition and applications is from Fabian Pfaffenberger "Twitter as basis of scientific studies". It also includes samples of scripts for

Twitter API and data storing within MongoDB database [12].

The perspective of Twitter data has been studied in different fields. Broad literature and approaches of spatial analysis for tourism purposes are well presented in the article "Spatial analysis: A critical tool for tourism geographies" [13]. Methodologies for semantic analysis may be found in "Tourism, travel and tweets: Algorithmic text analysis methodologies in tourism [14]. Also the article "London2012: Towards citizen-contributed urban planning through sentiment analysis of twitter data" [15] enabled us a very deep insight into possible methods for a sentiment analysis and its value for further planning-related investigations.

On a global scale Hawelka, Sitko, Beinat, Sobolevsky, Kazakopoulus and Ratti (2014) studied the perspective of "Geo-located Twitter as the proxy for global mobility patterns" in order to compare mobility characteristics of different nations and to uncover global mobility patterns. Mobility profiles were determined according to the user's residence country and tweets from other visited countries with the aspect to the Twitter penetration rate in each country, which defines the ratio between the number of Twitter users and the population of the country. As mobile users, only the users tweeting from at least one foreign country were considered. Spatial concentration of mobility was captured through the average radius of gyration of the users – a model which defines the spread of user's locations around their usual one. To validate the network of Twitter user flows, a gravity model was used. Positive correlation between the number of Twitter users, economic prosperity of the country and average travelled distance was established. The results were evaluated with the official statistics on international tourism. As a result, geo-located Twitter was validated as an objective and freely accessible proxy of global mobility behaviour. However, they were aware that the next step is translating the potential of geo-located Twitter into finer spatial scales [16]. Another study example of global mobility patterns is "Analysing refugee migration patterns using geo-tagged tweets" by Hübl, Cvetojevic, Hochmair and Paulus (2017) with the aim to identify and visualize refugee migration patterns from the Middle East and North Africa to Europe in 2015. Data analysis included spatial and semantic analysis with the refugee trajectory extraction and aggregation and detection of topical clusters along migration routes. Movement patterns and semantic clusters were extracted with the help of SQL filtering and V-Analytics software with the OPTICS method [17].

Also Zagheni, Grimella and Weber (2014) evaluated migrations in their study "Inferring international and internal migration patterns from Twitter Data", since the official data of migration flows are internationally inconsistent, outdated or non-existent. For estimation of recent internal and international trends in mobility rates, they proposed a methodological "dif-

ference in differences approach". Developed estimator enabled an estimation of relative changes in trends when new Twitter data are included or the data time frame is changed [18]. In "Moving on Twitter: Using episodic hot spot and drift analysis to detect and characterise spatial trajectories", Senaratne, Bröring, Lehle and Schreck (2014) were working on detection of trajectories through hot spot analysis and characterising trajectories through drift analysis. The study was applied on a concert route determination of a singer Lady Gaga. To detect a hot spot cluster of Twitter activities and derive spatial trajectories Kernel Density Estimation was used [19].

Another possibility of using Twitter geo-data is presented in "Twitter as Sentinel in Emergency Situations: Lessons from the Boston marathon explosions" (Cassa, Chunara, Mandl and Brownstein, 2013), where they examined the role of social media messages in emergency events and their early recognition. On the example of posted messages immediately after the Boston marathon bombing, it was recognized that that keywords can spread the news prior to official public safety or news media reports [20]. A further application on emergency events and taking advantage of real-time Twitter data is a study titled "Earthquake shakes Twitter Users: Real-time event detection by social sensors" by Sakaki, Okazaki and Matsuo from 2010. Using real-time notification semantic analysis, they were monitoring Twitter messages and developed an application for earthquake detection and notification. Semantic analysis was based on keywords, number of words and their context. For location estimation Kalman filtering and particle filtering was applied. As a result, the notifications of upcoming earthquakes were delivered much faster than announcements from official emergency agency [21].

Among the social media there are, apart from Twitter, also other applications offering free user-generated data. Another example of successful data integration is the usage of public available data from photo-sharing system Flicker. In "Predicting human mobility through the assimilation of social media traces into mobility models" by Beiro, Panison, Tizzoni and Cattuto (2016), they improved gravity and radiation models, which commonly in use for mobility modelling but still with room for improvement. Supported through additional integrated real-time Flicker data these models for spatial mobility patterns came at high performance even under low official data requirements and at different spatial scales [22]. Among other notable publications, the outstanding and for our work suitable studies dealing with analysis of Twitter data are also the studies "Tourist site attractiveness seen through Twitter" [23], "Text-based Twitter user geo-location prediction" [24] and "Mobility in Cities: Comparative analysis of mobility models using geotagged Tweets in Australia" [25].

Literature research provided us with a broad overview of possible studies

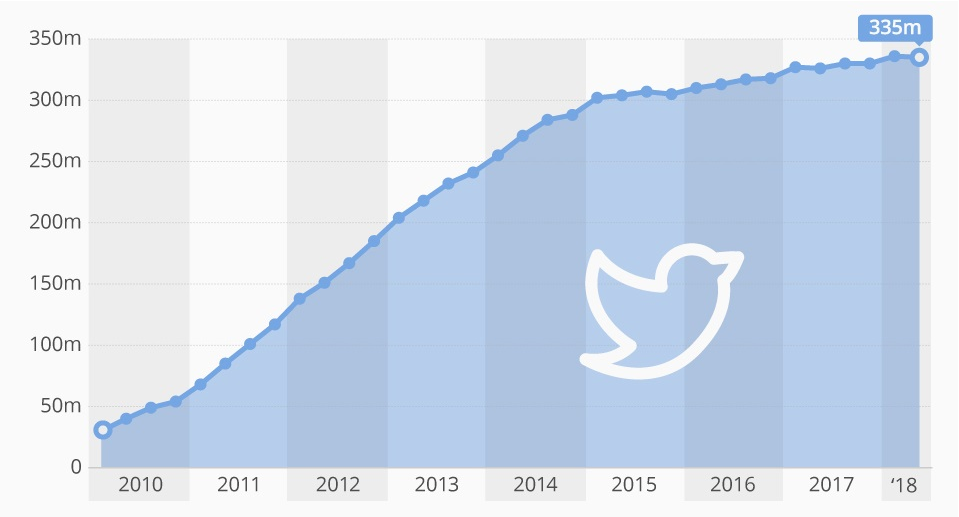
with Twitter data and confirmed our assumption on many existent studies on a global scale, but much less available studies within a finer spatial area. Therefore, our study will hopefully scope out new opportunities on Twitter data usage on fine spatial scales.

Chapter 2

Theoretical principles

Twitter is a micro-blogging social network established in 2006 in USA in Florida with over 310 million of worldwide monthly active users who produce over 500 million of tweets [26]. The number of users has been growing steadily from its beginning, even though with a noticeable deceleration since 2015 [27]. It offers information on "what is happening in the world and what are people talking about right now", because "when it happens, it happens on Twitter" [28]. Twitter is a so-called volunteered geographic information system (VGI) producing tremendous source of spatial-temporal data on a nearly real-time basis and that can bring notable benefits to diverse analyses and applications ranging from trend detection, urban management and marketing, to early disaster warning [19].

Figure 2.1: Twitter user growth 2010 - 2018 [27]



2.1 Crowdsourcing (Social Media)

In the world of volunteered geography M.F.Goodchild refers to citizens as sensors [29]. He reviewed the explosion of geographic information provided by individuals on a voluntary basis already in 2007, just after the expansion of social media. He discussed the characteristics voluntary geographic information with the aspect to the fact, that participants have varying expertise and all they need is motivation and a smart device with an internet connection. He also questioned the drivers, accuracy, individual privacy and correspondence of these data to conventional sources [29]. Such citizen generated content is defined as voluntary geography information - VGI. VGI are classified according to the provided spatial reference and the type of volunteering. Explicit VGI, where volunteers contribute with explicit mapped spatial reference offer portals such as for instance Openstreetmap or Wikimapia representing services that focus on elaborate representations of the Earth's surface. On another hand, a citizen can also implicitly refer to a location through mentioning it, having location service switched on when posting something or sharing their usual location under their profile. Such implicit voluntary geographic information is offered by users on most of the social media networking sites - for example Twitter or Flicker [30].

On another hand, Senaratne et. al [19] differentiate between types of VGI according to the methods used to capture the data: Map-based VGI, Image-based VGI or Text-based VGI. These are currently most common VGIs. Examples of Image-based VGI are Flickr, Instagram or Panoramio platforms for posting images taken with smart devices and containing geospatial information. Map-based VGI refers to VGI sources that include basic geometries of a map - points, lines and polygons. Open Street Maps, Wikimapia, Google Map Maker and Map Insights are only a few more known examples. Finally, text-based VGI are typically microblogs such as Twitter or Reddit, where a spatial reference can be revealed through text itself or through a geo-tag [19].

A geotag is a crucial part of both VGI and also of our extracted dataset of Tweets we are working with, since without that we would be unable to uncover a geographical location that a tweet refers to. A geotag is "an electronic tag that assigns a geographical location to a photograph, video or a posting on a social media website, etc." [31] defined by English Oxford dictionary or, explained into more detail by Goodchild [29], is "a standardized code that can be inserted into information in order to note its appropriate geographic location". However, technology that enabled evolution of VGI in the first place is definitely Web 2.0 with a combination of broadband communication, which expanded from a one-way relationship between user and a Web page into a two-way fast exchange of information. With Web 2.0, new protocols

enabled users to interact with the information stored in database on a server and not only accessing it but also adding or changing the content. Another important service that prompts VGI is georeferencing within user's device. Cellphones or cameras are equipped with tools for coordinates identification such as built-in GNSS systems, that enable creating geographic data to general citizens and, if connected to internet, easy sharing of geo-tagged content on social media or other VGI based services - for instance in order to create a map of walking or riding a bike [29].

2.2 Big data driven Geography

Because over 320 million of Twitter users generate over 500 million tweets per day [26]and 6000 in every second, Tweet analytics is viewed as a fundamental principle of Big data stream [32].

The term big data refers to a complex collection of very large, unstructured or time sensitive datasets where traditional database management tools can not be used for data processing [33]. Relational databases, statistics or visualization tools cannot reveal the meaning behind big data due to their high complexity. Important insights can only be uncovered if handling with data as a whole. Four concepts distinguish big data from traditional Business Intelligence Concepts [34]: Volume, Velocity, Variety and Veracity. Volume and Velocity due to the vast number of data generated, delivered and analysed every second. Data is flooding both from traditional data acquisitors as well as from different modern sources such as mobile devices, social media and sensors. Furthermore, data comes in various forms - structured, unstructured, semistructured, streamed or non-streamed data etc. identifying their Variety as a third concept. Finally, Veracity refers to the questioned certainty in the data - lack of accuracy and quality, which may be overlooked due to the vast involved volumes of data [34].

Big data and accompanied technologies in the modern world are reflected even in everyday information such as weather prognosis, which is generated from attributes of big data. Real-time processing and machine generated inputs, that are of a great importance in this process in order to analyse a massive amount of data and to derive an information and value from it. Raw data has no value, which is the inherent problem of big data. They reveal their usable insight and gain on value only after processing it from their native object format. However, processing is very costly, time consuming and technological demanding [33].

As Goodchild [35] discusses in his work "GIS in the Era of Big Data", in the field of geography our capacity to acquire great amount of data started already in the early 1970s with the launched satellite Landsat. However, more data was obtained than could be stored, examined, analysed, visualized. Geographic information had to be generalized by ignoring spatial detail or creating uniform regions. Yet, due to increase in computing speed and storage capacity the advantages of Big Data became of a big importance also in geography. Since fast analysing at finer spatial scale became possible, this enabled even the entire planet to be analysed efficiently. Traditionally, geographers had to rely on authoritative data sources, but nowadays, data acquisition from satellites, ground sensors or social media are available in near-real time. Additionally, average citizens create extra spatial data with the help of GPS-enabled devices and open source software and basemaps. In the Era of Big Data, both new and more frequent data sources are available, and now it is even possible to process and extract information from all of them [35].

2.2.1 Quality issues

The first three positive concepts of Big Data - Velocity, Velocity, Variety can of course also be observed in Geography. However, also Veracity or the quality as the fourth, not positive property should be questioned,as inherent uncertainties lie within the data, such as lacking metadata, documentation and unclear provenance [36]. Yet, an interesting example in Goodchild's discussion [35] underlined that this uncertanty may rather origin from too broad of a choice of sources to be compared with each other. Since nowadays there are more sources monitoring the same phenomenon, the results are reasonably distinguishable from each other. We used to trust authorities and their officially published data. Therefore, Big Data allows the questioning of official statistics and helps integrating data from various different sources. Also Sui [37] argued that, in the era of Big Data a priority for quantitative geography should be the synthesis of available data from different sources and not focusing on individual data sources.

Yet, every interested citizen, even without expertise or training can share his or her part in providing VGI and establishing rapidly growing data collections. How much information is spam and how much is credible? A study by Gupta and Kumaraguru in 2012 [36], discussed "Credibility Ranking of Tweets during High Impact Events". Among Tweets collected for event analysis 14% were spam. 30% referred to situational awareness information whereas only 17% were credible.

Reasons for quality issues are many. Let's discuss some regarding text-

based VGI, as they are most relevant to our research. Apart from GPS errors, bigger quality issues are caused by individual citizens as content contributors, either due to the lack of spatial knowledge or location settings within the portal or medium used. Also the spatial resolution can be insufficient, since they might specify the location only at city or even at state level. Errors within the data can also occur in case a contributor is at the moment of writing not exactly at the location he or she is posting about. From 2007 onwards, many research about quality assessments were undertaken considering the uncertainty of VGI data and quality assessment, methods and techniques to determine the credibility of collected data. Senaratne et al. [19] did a comprehensive study about VGI quality assessment research examined between 2007 and 2015. They took 56 research under consideration, that were dealing with VGI quality assessment. As a result of their study they determined 30 different quality measures and indicators that were resolved within them. As a result, a variety of indicators that have to be taken into consideration were outlined: Positional, Thematic, Temporal, Geometric and Semantic accuracy, Topological consistency, Completeness, Lineage, Usage, Credibility, Trustworthiness, Content quality, Vagueness and Local knowledge, Experience, Recognition and Reputation of the contributor. The authors of the study categorized the types of approaches and methods for VGI quality assessment grouped on four contents: Geographic, Social, Crowd-sourcing and Data mining. The results are represented in Figure 2.2.

2.3 Data acquisition

According to the planned study or analysis that requires Twitter data, there are different methods for data acquisition. Four primary ways include retrieving from the Twitter public Application programming interface (API), finding an existing Twitter dataset which suits your field of research, directly purchasing from Twitter organisation or finally an access or purchase from a Twitter service provider [38]. However, acquisition of the data depends on the requirements of a particular project, such as the need of historical or areal-time data, data amount and demand of complete or sampled dataset. Therefore, the most appropriate means of acquiring data can be selected with respect to available funding, technical skills and intended use of data-processing tools [38].

As already explained in the Methodology chapter, the official Twitter REST API was not applicable for our data acquisition due to its limitation of historical data requesting. For our main analysis we used a Python package Twitterscraper [6]. However, because we continued with requests via REST API in order to check the quantity of returned data no older than 7 days,

Figure 2.2: Type of approaches and methods for VGI quality assessment[19]

Type of approaches and methods	
Compare with reference data	Geographic
Line of sight	
Formal specifcations	
Semantic consistency check	
Geometrical analysis	
Intrinsic data check	
Integrity constraints	
Automatic tag recommendation	
Geographic proximity	
Time between observations	
Automatic scale capturing	
Geographic familiarity	
Manual inspection	Social
Manual inspection / annotation	
Manual annotation	
Comparing limitation with previous evaluation	
Linguistic decision making	
Meta-data analysis	
Tokens achieved, peer reviewing	
Applying Linus law	Crowdsourcing
Possibilistic truth value	Data Mining
Cluster analysis	
Latent class analysis	
Correlation statistics	
Automatic detection of outliers	
Regression analysis	
Supervised classification	
Feature classification	
Provenance vocabulary	
Heuristic metrics / fuzzy logic	

both data acquisition interfaces are explained into detail in the following two subsections.

2.3.1 Twitter API

Two public APIs for Twitter data are available – REST API and streaming API. Both are only adequate for researchers with basic programming knowledge, since Twitter only offers the application programming interface and no graphic user interface. Application programming interface (API) is defined by Oxford dictionaries [39] as "a set of functions and procedures allowing the creation of applications that access the features or data of an operating system, application, or other service". In our case, using Tweepy Python library allowed the access of the Twitter API, which permitted the inclusion of various parameters and return responses [40] - Tweets that match our criteria.

Streaming API allows real-time data acquisition, while REST API provides the user with the ability to search Tweets up to 7 days. REST stands for Representational State Transfer, which allows the server to transfer a representation of the state of the requested resource - eg. returning a state of a tweet including user, timestamp, ID etc, to a client - eg. us mining for tweets [5]. Conveniently for studies based on geo-tagged Twitter data, filters based on language and latitude and longitude pairs are possible. Apart from that, REST API also offers access to particular user's tweets, specific tweets via their ID number or obtaining user's followers and friends[11].

For the purpose of this master thesis, we were collecting the data weekly using REST API for tweets published in the last 7 days that meet certain parameters on tourism within the state of Styria. Access to API enables the use of Tweepy, which is a Python based library [41]. The library requires user authorization and provides access to both REST and streaming API in order to communicate with the Twitter platform, to retrieve tweets from a particular account, stream and filter real-time data or to access to desired tweets within the last week and confined to certain spatial or semantic parameters. However, due to Twitter API limitations such as limited access to historical tweets and current tweets, returning only a sample of tweets [38], significant adjustments have to take place within a particular research. No research based on historical or tweets older than 7 days days can take place with the use of these data acquisition techniques. When searching for tweets with REST API, 180 requests per 15 minutes are possible, returning only 100 tweets per request[11]. Hence we had to search after alternatives, because for our research, it was not possible to overcome the limitation of no return of historical tweets older than a week.

2.3.2 Twitterscraper

A developer [6] of Twitterscraper states the limitations of REST API's provided by Twitter as a big motivation to develop an open source Python package to scrape Tweets and make it available on GitHub. By using Twitterscraper there are no hourly and historical limitations but only the internet speed when mining and the number of instances [6]. After installation of the package, the request can be conveniently written within the command prompt. Arguments play a significant role of each request and may consist both of semantical and spatial parameters: User, language, start and end-date and for us crucial keywords and location. Request is based on Python programming language, for example on logical operators in order to determine the combination of keywords we are mining for. The package returns information such as username, fullname, Id, URL, text, html, timestamp and number of likes, replies and retweets per each extracted Tweet [6].

2.3.3 Geospatial Twitter data

Researchers or other business analysts often take advantage of a user location or of a location of a particular tweet. We differ between three different kinds of geospatial data which can be extracted from Twitter metadata: Tweet exact position, mentioned location within the tweet, or profile location [42].

Geographical position of current tweeting location may be revealed via geo-tagged tweets with exact coordinates. This geographic information comes at the highest precision level and requires no language processing. However, only 1-2 % of Tweets are geo-tagged. Additional drawbacks are very large target areas with the lack of precision, for example user tagging the country of his or her position and not the precise location [42]. A second source is mentioning current location of presence, which is semantically cited by user within the tweet text. It requires additional post-processing of a Tweet message in order to extract desired location among heterogeneous location descriptions within the tweet. This is a location indicator of low accuracy since the location filtering is implemented via keywords or phrases. Finally, each user optionally reveals also their profile location. However, it can be misleading since it can be described inccurately by any phrase and does not always fit the official geographical denomination of the living area. Profile location does not identify the current position or location of the particular tweet but only the user's city or area of residence[42].

2.4 Data processing

Encyclopaedia Britannica [43] classifies the term Data processing in the field of Computer Science and simply defines it as "manipulation of data by a computer". This process focuses on "conversion of raw data to machine-readable form, flow of data through the CPU and memory to output devices, and formatting or transformation of output. Any use of computers to perform defined operations on data can be included under data processing". [43]. Collins English Dictionary [44] simplifies Data processing as "the series of operations that are carried out on data, especially by computers, in order to present, interpret, or obtain information".

The data processing workflow, usually performed by a data scientist can be in general described in five stages annotated below [45] [46]. However, a forestep of every study has to be a clearly defined research problem.

- Data acquisition - collecting raw data acquired from trustworthy data sources.
- Data storage - storing in digital databases.
- Preprocessing - filtering, sorting, cleaning errors and duplicates. Providing the output file in a format needed for an analysis.
- Data analysis - applying algorithms on preprocessed data in a selected Software according to the individual study.
- Data interpretation and presentation - interpretation of results and presentation in form of graphs, maps, tables etc.

2.4.1 Databases - NoSQL

Modern applications demand database technologies that overcome traditional rigid relational databases, e.g. very broadly used MySQL, Postgres, Microsoft SQL Server, Oracle Database etc. Modern applications must be always on and accessible from different devices. They encompass working with scale, relational databases were not designed for. They involve volumes of new and changing data types, either structured, semi-structured, unstructured or polymorphic data. Furthermore, relational databases were developed in 1970s to deal with first data storage applications. They were, therefore, not designed to take advantage of processing power and commodity storage available today. As an answer NoSQL as "not only SQL" databases were developed in the first decade of 2000s. NoSQL databases MongoDB,

Cassandra, HBase and Neo4j are among more known ones [7]. A big key difference is in structure type, since SQL databases come of one type - always table-based where data is stored in columns and rows. Columns correspond to records and rows to elements within one record. Storing related data in separate tables and then joining together requests more complex queries. On another hand, different NoSQL databases are structured differently - either document-based, key-value pairs, graph databases or wide-column stores [47].

Another two key differences between SQL and NoSQL databases are in used language and their scalability. Language is fundamentaly different since relational databases use structured query language - SQL, which is both extremely powerful but also very restrictive. It is versatile and widely-used, which makes it a safe choice and great for complex queries. However, it requires prefixed schemas [47] - structure and data type, which all data has to follow and has to be determined in advance. In case of storing information of another data type, the entire database must be put offline and altered [7]. A change in the structure would therefore be disruptive for the whole system. A NoSQL differs from SQL with its dynamic schema for unstructured data and various storing ways of the data. A user creates documents without having to first define their structure. This flexibility allows the creation of documents without requiring to define their structure and can contain varying syntax from database to database with the ability to add fields on the fly [47].

Scalability, as another key difference corresponds to managing a server power in order to deal with increased demand. SQL databases are vertically scalable meaning it is possible to increase a single server power, but only through additional engineering, for instance adding RAM [47]. An alternative is spreading SQL databases over many servers but in this case core JOINs, transactions or other relational features usually get lost. A better solution for the demands of modern applications that work with large and changing datasets is by scaling horizontally. Horizontally scaled NoSQL databases can handle more data traffic by adding additional commodity servers or cloud instances. In order to achieve faster query times, sharding is also a possible approach [7], since a large database gets separated into data shards which are defined as smaller, faster, more easily manageable parts [48].

MongoDB:

Our choice for data storing and preprocessing within this master thesis was a non-relational database MongoDB. It is an open-source and document-based

NoSQL database management system (DBMS). In 2006 it got established as an answer on needs of building big data applications and rigid relational database models. Since 2009, when they started to offer an open source model there were already over 10 updated versions [49]. In 2016 also a cloud database service MongoDB Atlas was launched. As it is a NoSQL database it is made up of collections and documents instead of tables and rows as is in the case of relational databases [50]. Due to storing data in flexible JSON documents, fields can vary from document to document. Also data structure can be changed over time. With its capabilities it meets the demands of modern apps. With all these advantages of MongoDB various intelligent operational data platforms can be built [51].

Data administration within MongoDB can be done within mongo shell or by using Graphical User Interface. Mongo shell, a standard component of MongoDB, is an interactive interface based on JavaScript. It enables all operations such as querying or updating data and administrative operations [50]. In order to visually explore data with MongoDB it comes wit a GUI - graphic user interface called Compass. Compass enables interaction with the data through full CRUD functionality - e.g. creating, reading, updating and deleting data, working with document structure, index data etc [52].

2.4.2 Data format

Our data was acquired and stored in a database in a JSON format. JSON, an abbreviation for Java Script Object Notation, is a human and machine-readable open standard for exchanging data. Its design is easily understood by humans and easily parsed by machines. Even though it is based on JavaScript, it is language independent. It supports all basic data types such us numbers, strings, boolean values, arrays and hashes. JSON notation is popular for responses through RESTful web services and other web frameworks or mobile devices with slower connections, since it works well even at limited network speeds. JSON files comes with .json extension [53].

JSON is defined by universal data structure name/value pairs or ordered list of values [53]. Name/value pairs is in various languages realized as an object, record, struct, dictionary, hash table, keyed list, or associative array. Ordered list of values is realized as an array, vector, list, or sequence [54].

A JSON is build-up of following forms [54]:

- Object - unordered set of name/value pairs, beginning and ending with curly braces "{ }". Names are followed by a colon ":" and pairs are separated by commas ",".

- Array - ordered collection of values of any type, separated by a commas "," and beginning and ending with a bracket "[ ]".
- Value - value can be a string (a sequence of Unicode characters notated by double quotes), a number, null, boolean true or false, an object or an array. They can also be nested.

Object on example 2.3 consists of many fields, including nested objects and arrays. Field -id is numerical, nested object name consists of first and last name, in array "contribs" are four strings and in "awards" there are two further nested arrays.

Figure 2.3: Example of JSON document [55]

```
{
    '_id' : 1,
    'name' : { 'first' : 'John', 'last' : 'Backus' },
    'contribs' : [ 'Fortran', 'ALGOL', 'Backus-Naur Form', 'FP' ],
    'awards' : [
        {
            'award' : 'W.W. McDowell Award',
            'year' : 1967,
            'by' : 'IEEE Computer Society'
        }, {
            'award' : 'Draper Prize',
            'year' : 1993,
            'by' : 'National Academy of Engineering'
        }
    ]
}
```

2.5 The Power of Geographic Information Systems

Geograpphic Information System (GIS) is "a system to collect, analyze, store, manipulate, manage, share and visualize spatial information" [35]. Organizations in very various fields are using it to identify problems, monitor change, manage and respond to events, perform forecasting, understand trends etc. Making maps that communicate, perform analysis, share information and solve complex problems is contributing to a change also in our everyday life. Since "Modern GIS is about participation, sharing, and collaboration" [56], significant elements defining GIS are Maps, Data, Analysis and Apps. Maps can be referred to as a geographic container for data with a spatial component and analytics that bring GIS to everyone through integrating it on Apps accessible to everyone possessing a smart device [56].

Applications of GIS for tourism purposes are very wide. Using it as a tourist product for conventional purposes such as producing tourist maps is one of the most obvious applications, accompanied by using it for Apps development in order to plan a sightseeing route or attraction selection. Furthermore GIS is very promising a step before and after tourism is "happening" - namely for tourism planning in order to boast it and tourism analysis in order to document it. In this chapter we focus on spatial analysis techniques in order to evaluate our dataset of Twitter geodata.

2.5.1 Spatial analysis

Spatial analysis is intended for the investigation of geographic patterns in spatial data. It examines locations and attributes and studies relationships between features in order to determine spatial patterns and trends. The analysis is implemented through layer overlay and other analytical techniques [57], within GIS software. The analysis helps to investigate a correlation between spatial characteristics and to extract new information from spatial data. Within the spatial analysis, the concept of nearness and relatedness are highly significant [58]. Shorty, through evaluating suitability and capability, estimating and predicting, interpreting and understanding spatial analysis leads to better decision-making [56].

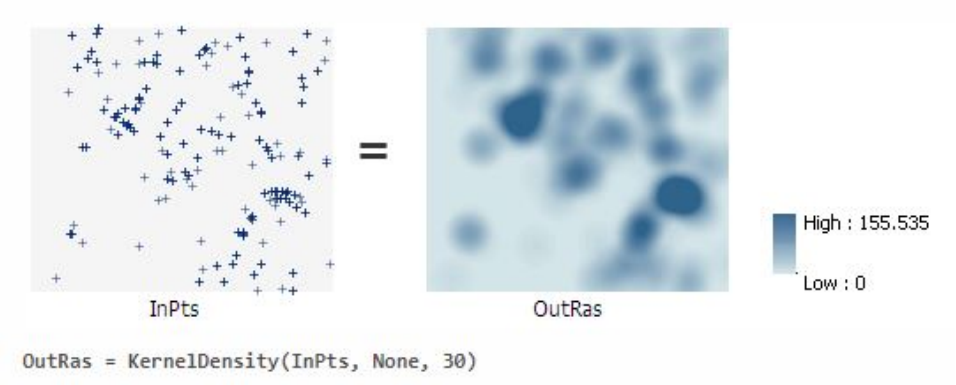
Kernel Density Estimation:

For us, a very promising method for density estimation of Twitter data is a Kernel density estimation (KDE), since it belongs to a non-parametric class with no fixed structure and depends on the point data [15]. With KDE,

a magnitude-per-unit area is calculated from point features, hence in this Study from Twitter point data - meaning from positions of the Tweets. By calculating the density, we are spreading the input values over a raster cell. If implementing an analysis within a leading GIS software ArcGIS, we receive an output raster with determined density value for each cell according to the optionally user-given or default circular search area within which a density should be calculated [59]. If a detailed result should be gained, smaller search radius values have to be set up. Contrary, in case of using a large search radius, the output would therefore result in more generalized density raster. However, the search radius may be automatically calculated according to the spatial configuration and number of input points. Neighbourhoods without features within them return cells assigned as NoData. Since the density calculations depend on accurate distance and area calculations, the data has to be projected accordingly [60].

Kernel density estimation can be used either for point or line features. Kernel density example with point data type as input is represented in Figure 2.4. It represents unweighted features that have a value of 1, since no population field was chosen, and a search radius set to 30 units. In case that a population field would be chosen, the feature's value would determine the number of times to count the feature [61]. The population field therefore plays a role of a feature's weight.

Figure 2.4: Kernel density [60]



For point density it is also possible to get each cell's density value by simply summing up the points within the search area (neighbourhood) and dividing by the search area size. The difference to the Kernel density estimation method is that here the known quantity value (a population field) of a feature is spread out from the point location based on a quadratic kernel function [62] originally described by Silverman [63] in 1986 in a book Density Estimation for Statistics and Data Analysis. Due to this formula the highest value at the point's center is decreasing up to the search radius border

where it reaches the value of zero. Output cells returned at the end of the analysis indicate the kernel density as a sum or total number of overlying spread surfaces accumulated from individual point (or line) features [61].

Results of the Kernel density estimation or any other density methods can be visualized by a heat map. Heat map refers to feature concentration from a geographic, namely spatial, perspective. From point or line data an interpolated surface is created in order to indicate the density of occurrence. It is a way of visualizing a density surface by coloured gradient in order to easily identify locations of higher densities or clusters of geographic features [64].

In our case we are using KDE in order to identify density of tweets. Other very common applications using cases are indicating houses or some service density (e.g. doctors in some area), for crime reports or traffic accidents in a case of point input features. For line features an example could be analysing the densities along the roads or utilities lines and identifying their their interdependencies and influence to a town [61].

2.5.2 Spatial statistics

A very important field of study within spatial analysis are spatial statistics. There are similarities in concepts and objectives between traditional and spatial statistics, however the latest was design for working with geographic data [65]. It refers to statistical methods that implement mathematical computations by incorporating spatial characteristics of the data such as distance, area, volume, length, connectivity, orientation, centrality, etc. To include the concepts of space and spatial relationships is very significant for analyses such as statistical comparison of spatial data, statistical modeling and prediction of spatial interaction, for pattern or shape analyses, for surface modelling or surface prediction, spatial regression etc. [66].

Statistical toolsets offered by ArcGIS are annotated below. If necessary, a user has also a possibility to modify the source code of the tools in order to meet the needs of a particular analysis. With different methods they allow to analyse patterns and clusters, measuring geographic distributions, modelling spatial relationships and so on [65]:

- Analysing Patterns - Evaluating spatial pattern of features. Features may form clustered, dispersed or random spatial pattern.
- Mapping Clusters - Identifying statistically significant hot spots or identifying and grouping features of similar characteristics.

- Measuring Geographic Distributions - Addressing the centre, shape, orientation or dispersion of features.
- Modelling Spatial Relationships - Implementing regression analyses or constructing spatial weights matrices in order to model data relationships.
- Utilities - A toolbox of various tools for computing areas, collecting coincident points, assessing minimum distances, exporting variables and geometry or converting files.

2.5.3 Mapping clusters

Among many statistical tools, mapping clusters are of a particular interest for our research. Mapping clusters is in close relation to the "first law of geography", defined by Tobler in 1979: "Everything is related to everything else, but near things are more related than distant things". Spatial clustering is expressed as a positive spatial autocorrelation, revealing a relationship between the values of features in different locations. Similar values are therefore spatially clustered together. In case of spatial dispersion of similar features, we talk about negative spatial autocorrelation [67]. It is of particular importance to identify spatial clustering within the data as patters help reveal information, especially if the incidents are exposed to similar underlying impacts or geographic processes [68].

The chosen methods for identifying and describing spatial clusters depend on the particular research question and the extent of information we are looking for. Mapping clusters identifies the presence, the location and the intensity of any existing clusters. Some of the main methods for cluster identification are MORAN I, Geary C, NNI and Ripley's K and for determining cluster location Getis-Ord's GI* and Anselin's LISA [68]. In ArcGIS, available tools for cluster identification are listed below. With these tools it is possible, apart from identifying the location, also pointing out the most similar features or spatial outliers [69].

According to ArcGIS toolbox for Mapping Clustes [69]:

- Cluster and Outlier Analysis - Identifying statistically significant hot spots, cold spots, and spatial outliers according to a given set of weighted features. Method: Anselin Local Moran's I statistic.
- Density-based Clustering - Identifying clusters based on spatial distribution of point features surrounded by noise.

- Hot Spot Analysis - Identifying statistically significant hot spots and cold spots using the Getis-Ord Gi* statistic and with a given a set of weighted features. Since we made use of Hot Spot analysis within out research it's process is also explained into more detail in the following subsection.
- Multivariate Clustering - Natural clustering based on feature attribute values.
- Optimized Hot Spot Analysis - Similar to Hot Spot Analysis with the difference of automatic evaluation of the input feature class characteristics.
- Optimized Outlier Analysis - Similar to basic Cluster and Outlier Analysis with the difference of automatic evaluation of the input feature class characteristics.
- Similarity Search - Based on feature attributes it identifies most similar and dissimilar candidate features.
- Spatially Constrained Multivariate Clustering - Based on feature attribute values and cluster size limits the tool identifies spatially contiguous clusters.

Hot Spot Analysis:

Hot Spot analysis is a method of cluster identification and visualization that we took advantage of within our research. Hot Spot indicates statistically significant spatial clusters of high values as hot spots and of low values as cold spots. It is based on Getis-Ord Gi* statistic and optionally also on a set of weighted features [69]. Explained in a simple way, a hot spot indicates an area of point incidents of unusually high occurrence. It is based on a relative concept of high concentration, since hot spots are identified in immediate surroundings, even though their absolute concentration may not stand out if compared to entire study area. A transformation from point observation into area measurement is possible. Using Hot Spot analysis for area measurement purposes, an area of high quantity or intensity of observed feature values can be identified [68]. We used Hot Spot analysis in the same way for identifying locations of high Tweets concentration.

Table 2.1: Hotspot analysis output values[71]

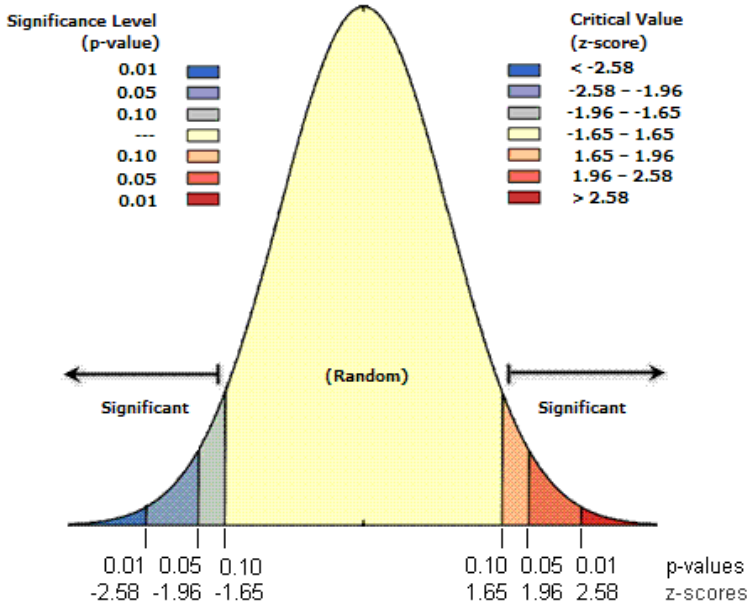
z-score (Standard Deviations)	p-value (Probability)	Confidence level
$<$ -1.65 or $>$ +1.65	$<$ 0.10	90%
$<$ -1.96 or $>$ +1.96	$<$ 0.05	95%
$<$ -2.58 or $>$ +2.58	$<$ 0.01	99%

Maps are subjective and with the help of Hot Spot analysis, the spatial patterns may be revealed that would be left unnoticed if identifying with the human eye. The end result of a Hot Spot analysis is less subjective due to statistically significant hot spot areas. The designation of an area as being a hot spot is therefore expressed in terms of statistical confidence. Spatial dependencies within our feature is determined through a regression workflow. If feature and its neighbourhood is statistically higher than the rest of the study area, it is going to be determined as a hot spot. A hot spot is therefore defined within the context of neighbouring features. A Hot Spot analysis produces values identifying statistically significant clusters of high and low values and values determining the probability that spatial clustering in data is not due to random chance [70]. The tool within ArcGIS returns a new output with 3 important values: Standard deviations as z-score, probability as a p-value and confidence level bin Gi-Bin. A null hypothesis for the pattern analysis tools is Complete Spatial Randomness (CSR). With the help of z-scores and p-values we can determine if the null hypothesis could be rejected or not. If rejected, then the features exhibit statistically significant clustering or dispersion. Confidence levels of z-scores and p-values are associated with the standard normal distribution (Figure 2.5 and Table 2.1). In order to reject the null hypothesis we therefore strive for higher z-values and lower p-values [71].

According to ESRI scores definition [71]:

- Z-scores - Z-scores are standard deviations: A score of +1,5 would mean 1.5 of standard deviations.
- P-value - P-value refers to a probability if the spatial pattern was generated due to some random process. Small p-values therefore mean small probability and rejection of the null hypothesis.
- Gi-Bin - Gi-Bin refers to confidence level. The end result of the analysis is visualised according to this value of this field.

Figure 2.5: Significance values of Hot Spot analysis [71]



Optimized Hot Spot Analysis:

An alternative to Hot Spot analysis with manual settings is Optimized Hot Spot Analysis. With the use of Optimized Hot Spot Analysis the characteristics of the input data are evaluated automatically in order to extract the parameters that are needed for an optimal result. After aggregation of the incident data into weighted features an appropriate scale of analysis is determined. The statistically significant values of the end result are adjusted multiple times. Adjustment of the values is performed according to the testing and spatial dependencey with the help of the False Discovery Rate (FDR) correction method [72].

2.5.4 Correlation as an evaluation tool

For evaluation purposes, we worked with Pearson correlation coefficient. Correlation coefficient revealed us the relationship between our dataset of Tweets and official statistics. We evaluated the absolute and seasonal quantity of Tweets between the years 2008 and 2017 at a district spatial scale. In order to determine Twitter geo-tagged data as plausible for tourism-related studies we look for association between our data and reference data. The evaluation result should in this case be a correlation coefficient indicating

Table 2.2: Pearson's correlation coefficient values[75]

Qualitative description of the Strength	Value of Correlation coefficient (R)
no association	0 - less than 0,4
weak association	0.4 - less than 0.6
moderate association	0.6 - less than 0,8
strong association	0,8 - less than 1
complete association	1

positive values. As defined by Howitt & Cramer [73] a correlation coefficient is "a numerical index which indicates the strength and direction of a relationship between two variables". Since we deal with metric variables and linear relationship between them, we used Pearson Correlation coefficient. It is very recommendable to check the relationship between the two observed variables by a simple scatter plot, since the results of Pearson coefficient may be misleading at curved relationship or outliers may exist that distort the end result. In case of many extreme scores affecting the result a Spearsman coefficient comes as a good alternative, since it is less affected by outliers [73]. Even though the Pearson coefficient is the most common one, there are different variants of correlation coefficients such as Spearman's rho or Kendall's tau for ordinal variables and Cramér's V or Contingency coefficient for nominal data [74].

Pearson coefficient, also referred to as a Bravais-Pearson Product Moment Correlation is a measure of a linear relationship between features. It is a result of the standard deviations and standardized covariance of observed variables [75]. Correlation values (R) range from -1.00, indicating a negative correlation up to +1.00, indicating a positive correlation. Positive correlation takes place if the score of one variable increases according to an increased score of another variable. A negative correlation is determined if one variable is increasing while another one is decreasing. A value of 0.00 means no association at all and randomly scattered values in scattergram that do not follow any linear line. There are different definitions of minimum correlation that is still expressing an association or relationship between two variables. From the value 0.5 on we can talk of moderate enough relationship between values [73], but should definitely interpret the value within the context of the particular study and not make any generalized conclusions. In natural sciences the correlation of 1 or a perfect association and complete relationship could only exist when dealing with physical matrices [75]. In such case the values of both variables on a scattergramm would follow a straight line completely.

Similarly, as already explained P-value in subsection 2.5.3 about Hot Spot analysis and returned vales, correlation analysis using SPSS software also gives information about statistical significance in a form of Sig.(2-tailed)

value. The bigger Sig.(2-tailed) is, the higher is probability that relationship between observed variables is random and therefore based on null hypothesis of complete spatial randomness. Hence, we strive for small Sig.(2-tailed) values in order to confirm statistically significant correlation between variables. If the probability of being under null hypothesis is under 0.05 [76], as is generally accepted, the correlation is statistically significant.

Pearson correlation coefficient is a very suitable indicator of association of our dataset of tweets and statistical data issued by official statistical offices. However, it is important to keep in mind, that correlation only displays the kind / direction and strength / degree of the association between two variables. It does not give us an information how do both variables influence each other. To determine which variable has an influence on the other, it is necessary to do a regression analysis [75]. Yet, this study focuses primarily on identifying correlations between variables acquired from different sources but representing the same topic. Therefore, we do not seek to identify which variable has an influence on the other. A correlation analysis is therefore the most suitable one in order to determine the data plausibility of our alternative source.

Figure 2.6: Value of correlation coefficient corresponding to the point cloud and regression line [75]

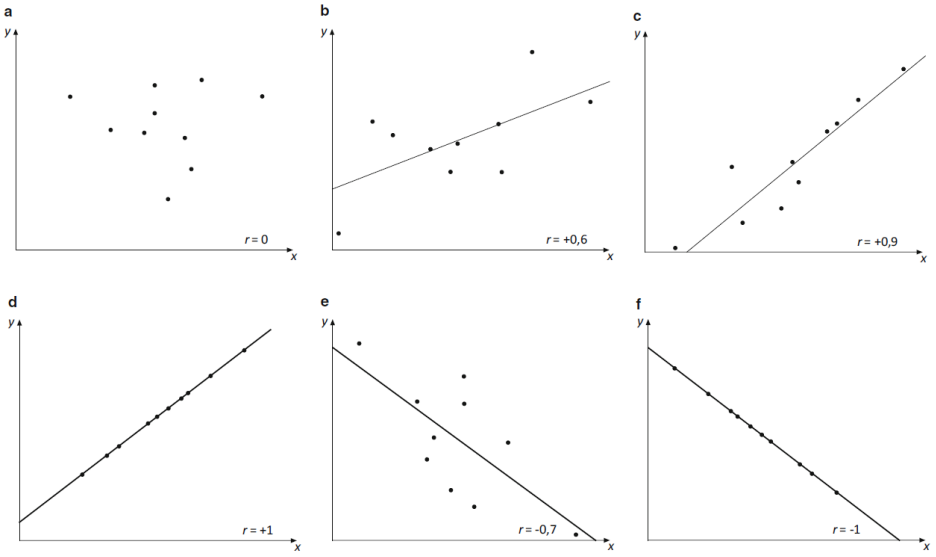
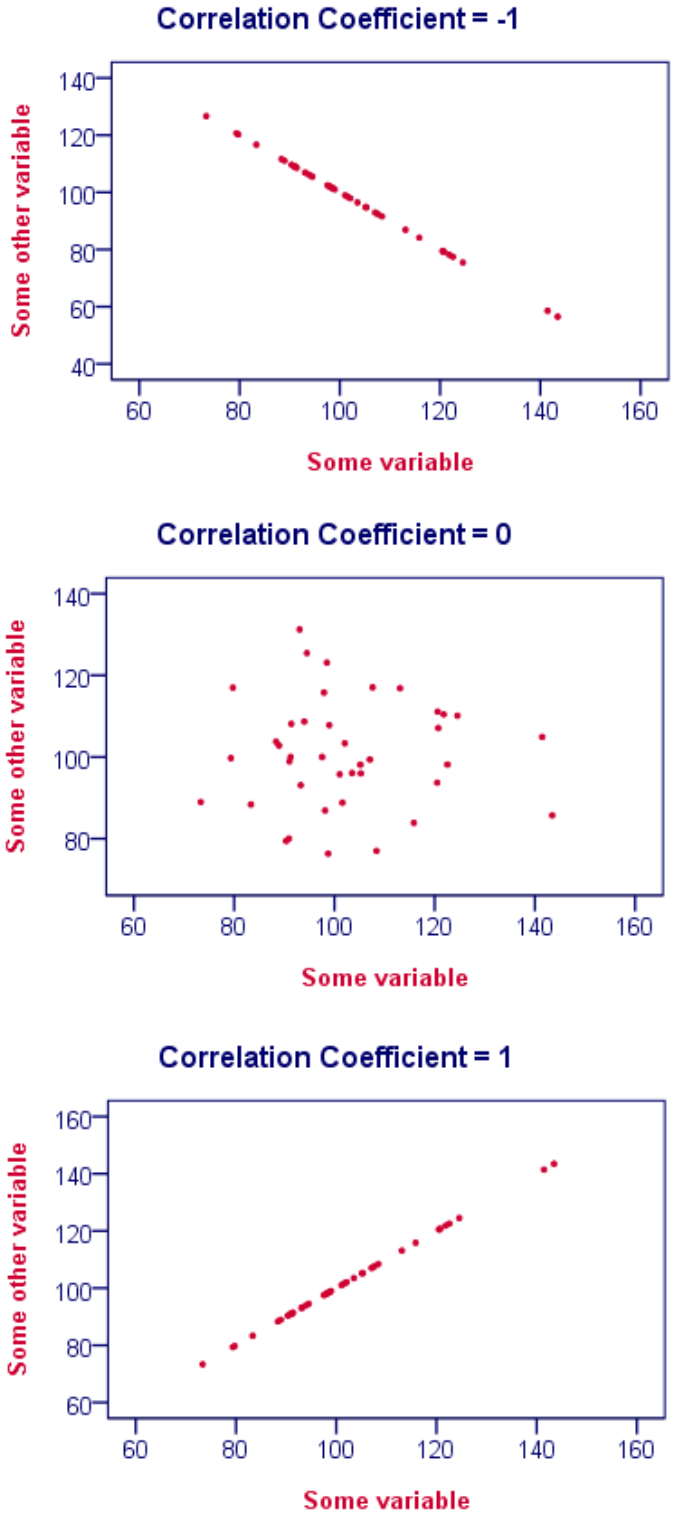


Figure 2.7: Correlations at perfect or no linear relation [74]



2.5.5 Thematic maps

Our research results are represented by means of thematic mapping. Thematic maps convey information about spatial pattern of a single topic or theme. They focus on a particular geographic quantity or quality rather than on information of different sorts. Due to their emphasis to a particular subject exclusively, they are clear, direct and accurate. Thematic maps are generally classified to qualitative or quantitative maps according to the type of represented data. Quantitative data are most commonly represented on choropleth, dot-distribution, proportional symbols maps and also as isoline or flow maps [77]. For our research possible cartography visualization methods are explained below. Within this master thesis we mostly used choropleth maps in order to represent the quantitative data (e.g. number of tweets) within the municipality.

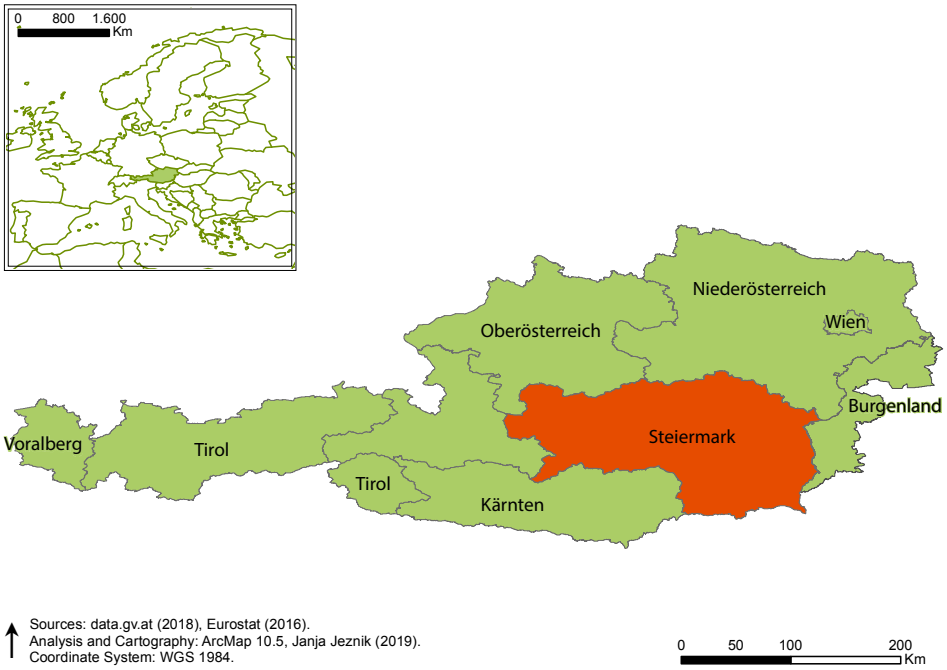
- Choropleth maps - A thematic map where classed values of a geographical phenomen are displayed by distinctly coloured or shaded patterns [78]. It is common to use choropleth maps for both absolute values or for spatially averaged ratio data, e.g. population density. However, choropleth maps should better not be used to display absolute quantities that are not area-related. Classification of the data and choosing right number of classes is of great importance. Even though there is room for creativity it should not override the goal to create a map with a clear message. Usually around six division classes are applied [77].
- Dot-distribution maps - Maps where a dot represents a quantity of some geographic feature. A ratio between feature quantity and number of dots is commonly generalized according to the scale of the base map. In a case of ratio of one-to-many, a number of occurrences are grouped and represented by one dot. An appropriate degree of generalization should be tested out so no blackout or dot-merging occours due to too high dot density [77]. In our case, a dot - distribution map can be a good solution in order to distribute all the tweets within one municipality, since the same position of municipality's center was assigned to the tweets originating from the same municipality. In this case we would use a one-to-one ratio, where each dot would represent one tweet. With such an approach, good insights into arisen clusters can be provided [79]. Another option is to apply non-uniform but geographically appropriate positions of the dots - e.g. dasymetric plotting which refers to clustering dots - for example representing population more densely around cities and sparse in rural or remote areas [77].

- Maps with proportional symbols - Proportional symbol refers to a symbol whose size differs according to the mapped phenomenon [80]. Data is not classified and the legend displays graduated size of symbols of few values from smallest to largest [81]. Proportional symbols may be scaled linearly, which is best suited for data of a small range, or non-linearly in case of big value differences between data. Non-linear proportions are necessary since otherwise a symbol of the highest value would be countless times bigger of the symbol representing the smallest value - for instance when representing cities with multiple millions of inhabitants and smaller towns on the same map. Simple symbols such as squares and circles are most practical in this regard since the eye can easily differentiate small size differences [77]. Within the group of maps employing proportional symbols, graphs or statistical summary graphics can also be applied. Bar graphs or pie diagrams are a simple means to depict both a component quantity and category. Pie diagrams can, in addition, even be scaled proportionally to allow for the total quantity to be displayed. Another possibility is to also apply a histogram to reveal a frequency distribution [77].

2.6 Tourism

The area of application is a state Styria in Austria, with an area of 16,388 km^2 [4] and with 1.240.214 inhabitants, according to the census in 2018, which accounts to 14% of the Austrian population [82].

Figure 2.8: Austrian states



Styria's topographic diversity differs from high-alpine regions in Upper Styria over the hills in Eastern and Western Styria extending to the Hungarian lowlands in the southeast. Big river basins of the rivers Mur, Feistritz, Enns and the Raab form the landscape, together with numerous bathing lakes. The biggest lakes (Grundlsee and Altausser See) are located in the northwest of the state. A wide variety of natural and cultural landscapes and scenic diversity is available in one national park under special protection and in seven nature parks. In winter, Styrian cross-country trails and over 700 km of skiing slopes attract around 2,200 visitors. Other popular destinations include state capital city of Graz, many thermal baths and culinary thematic roads such as wine and apple roads. In addition to the numerous excursion destinations and attractions, diverse yearly events also attract thousands of both foreign and local visitors [83]. Some tourist destinations in Styria have very long tourism tradition. The Ausseerland was a popular recreational and summer resort already in the 19th century and a spa resort in Bad Gleichenberg was already formed in 1834 [83].

2.6.1 Data acquisition and tourism statistics

Official collection of tourism data is performed by authorities of Statistics Austria [2]. Municipalities have to follow the rules for data acquisition that are determined in the "Guide to the accommodation statistics". Statistics Austria (or former Austrian Central Statistical Office) collects Data on Austrian overnight tourism since 1875, when the first record on "health and beauty tourism" ("Kurtourism") was written down. The most important records are so-called accommodation statistics that consist of the arrivals and overnight stays and differ between the types of accommodation and countries of origin, and between the capacities of private and commercial accommodation [10].

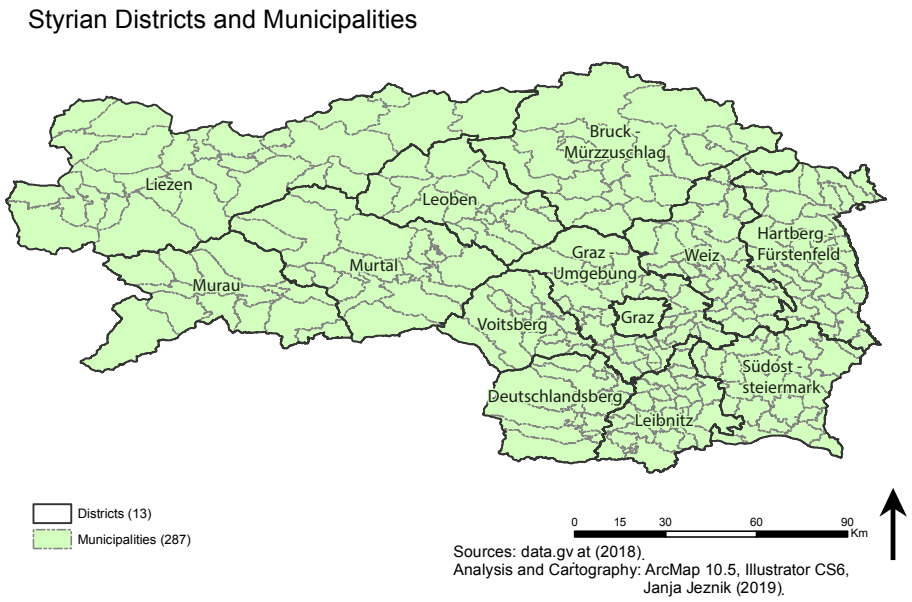
Annual period for statistics on tourism lasts from 1. November to 31.October. The data is reported monthly from the municipalities with more than 1000 overnight stays per year. These are so-called tourism municipalities. Valid for 2018, 218 of 287 Styrian municipalities were classified as tourism municipalities, which accounts for 76% of them. Due to data security, only these belonging to reporting communities are monthly obliged to report the data. Before the community structural reform started in 2010 and finished in 2015 there were 542 municipalities [9].

The Guide for statistics on tourism also defines which type of guest is included in the reported data. According to EU Regulation, "Tourism is the activity of persons who travel to a main destination outside their usual environment and spend less than one year there for any main purpose, including business, vacations or any other personal reason other than employment at the visited place" (Regulation (EU) No. 692/2011, Registration Act 1991, §9, §10, §19). Guests staying in a tourist accommodation for a longer period than one year are not included into statistics. Also arrivals and overnight stays of guests staying in their private second housing properties are not reported [10].

Very important for the purposes of our thesis is that only overnight guests are reported. Therefore, the number of arrivals may never be greater than that of overnight stays, since arrivals are noted at the first overnight stay [10]. Data acquisition via Twitter enables collecting data on tourists visiting on a single day and not necessarily overstaying. Collecting day tourists and real-time data is as such in this case a crucial advantage of Twitter data.

Throughout the year there are 2-summit seasons in Styria with one peak in August with 14% and another one in February with 11% of annual overnight stays. Current annual arrivals (58% in 2017) and overnight stays (55% in 2017) are suggesting a slight seasonal tendency towards the warm half of the year from May to October. Compared with Austria-wide

Figure 2.9: Districts and municipalities of Styria from 2015 onwards



overnight stays, Styria marks a slightly weaker winter share and slightly higher shares in the secondary seasons - eg. In October, November and April, May. The municipalities with the greatest amount of overnight stays include Ramsau am Dachstein, Schladming, Bad Radkersburg, Sankt Georgen am Kreischberg and Hohentauern [83].

The five-year tourism development from 2012 and 2017 shows a positive trend of continuous growth of +17.6% regarding arrivals and +13.0% regarding overnight stays. The only exception with a weak decline is in the winter seasons 2012/13. However, growth continued already in the next season in spite of an extremely warm winter. However, there is a continuous decline in terms of length of overnight stays. In the tourism year 2003, an average overnight guest still overnighted 3.7 days but only 3.1 days in 2017 [83].

Figure 2.10: Tourist arrivals to Styria in tourism years 2008 - 2018 [83]

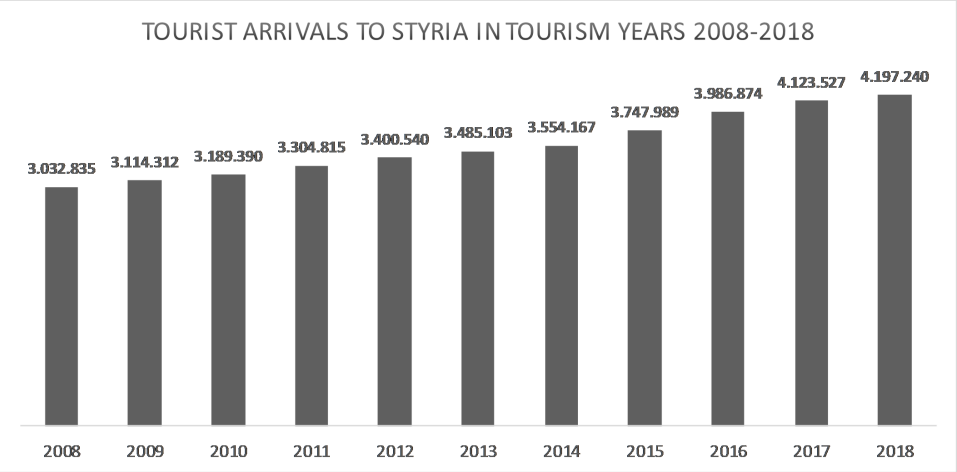
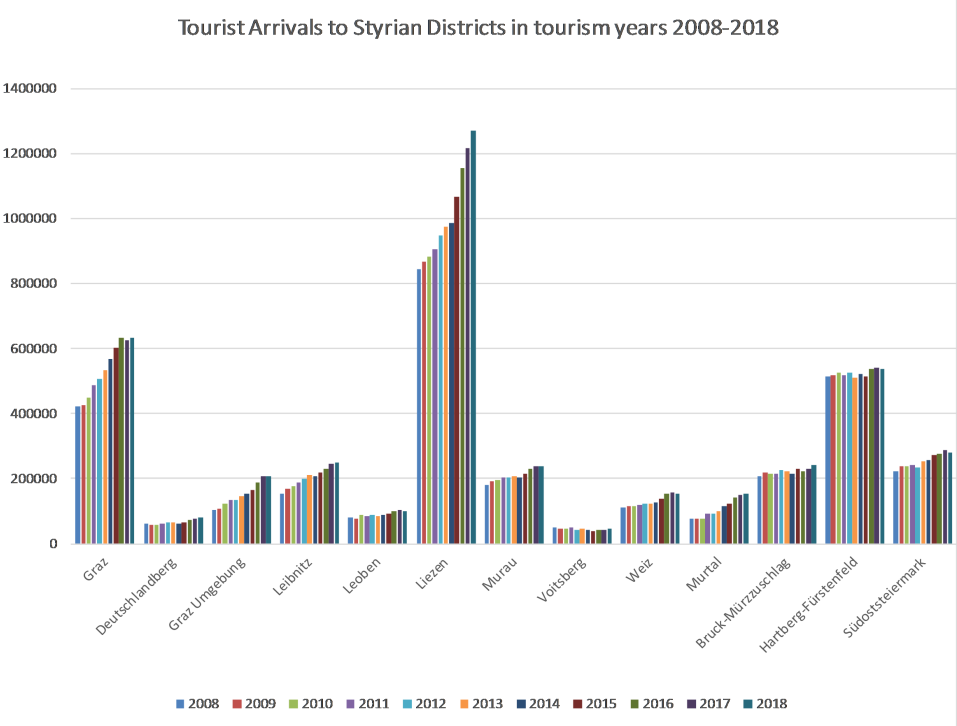


Figure 2.11: Tourist arrivals to Styrian districts in tourism years 2008 - 2018 [83]



2.6.2 Overview of Tourism regions

Due to a very heterogeneous landscape and the variety of tourist attractions throughout Styria, the state is divided into 9 tourist regions. Tourist regions are combined of different districts and municipalities in order to connect them and enhance tourist opportunities. They are described into detail in the following pages.

Figure 2.12: Styrian tourist regions - a graphical representation [84]

Ausseerland-Salzkammergut:

Ausserland-Salzkammergut includes 4 tourism municiplities: Grundlsee, Altaussee, Bad Aussee and Bad Mitterndorf. Due to its mountainous landscape, the region offers a variety of attractions in both the summer and winter season. The landscape is characterised by numerous lakes, such as Grundlsee, Altaussee, Toplitzsee, Kammersee or Ödensee. There are about 260 km of cross-country trails and 76 km of skiing slopes, with the largest ski resorts Tauplitz in Bad Mitterndorf and Loser in Altaussee. Two spas are also located in Aussee. Apart from hiking and bathing, there are also festivals and music events in the summer season, that attract many visitors. Other very visited destinations in the region are the scenic route Loser Panoramastrasse and the saline mine Altaussee [85].

Table 2.3: Districts in tourism regions

Tourism region	Included districts or municipalities
Ausseerland-Salzkammergut	Municipalities Altaussee, Grundlsee, Bad Aussee and Bad Mitterndorf.
Upper Styria (Obersteiermark)	Districts Leoben and Bruck-Mürzzuschlag. Without municipalities Pernegg an der Mur and Breitenau am Hochlantsch. Municipalities from other District (Liezen): Gaishorn am See and Wildalpen.
Region Graz	Districts Graz, Graz-Umgebung and municipalities Pernegg an der Mur and Breitenau am Hochlantsch.
Schladming-Dachstein	13 municipalities of district Liezen.
Southwestern Styria	Districts Votisberg, Deutschlandsberg, Leibnitz.
Spa & Vulcano country and Eastern Styria	Districts Weiz, Hartberg - Fürstenfeld and Südoststeiermark.
Murtal Holiday Region (Murau - Murtal)	Districts Murau and Murtal.
Gesäuse	5 municipalities of district Liezen.

Upper Styria (Obersteiermark):

Upper Styria accounts for almost 20% of the total Styrian area and represents the second largest tourism region. In the winter season, the region offers several small to medium sized skiing resorts with almost 140 km and over 60 km of cross-country ski trails. In summer there are numerous hiking trails in the Nature Park Mürzer Oberland or around the lakes such as Leopoldsteiner See and Green Lake. The major attractions of the Region is the basilica in Maria Zell, annually bringing around 700,000 visitors in that area. Among the many touristic events and destinations include summer concerts and a Christmas market in Maria Zell, the town of Erzberg with an open iron mine, the Adventure world Mautern, the castle in Oberkapfenberg, the Art Museum and spa in Leoben, the museums of the writer Peter Rosegger etc. [86].

Region Graz:

Tourism region Graz corresponds to the state capital Graz and surrounding municipalities, offering a big selection of events and sights. The old town of Graz is a UNESCO's world heritage site due to its harmonious blend of the architectural styles and artistic movements such as Cathedral and mausoleum or the Mur island [87]. Over the old town rises the Schlossberg with the Graz landmark, the Clock tower. The most known museums are contemporary art Kunsthaus Graz, Museum centre Joanneumsviertel, the Graz Opera and the Styrian armoury called Landeszeughaus. A popular desti-

nation outside of the city centre is the Schloss Eggenberg. Many excursion destinations are in the hinterland of Graz, such as the mountain Schoeckl, Lurgrotten cave, the monastery Rein, the open-air museum Stübing or the Arnold Schwarzenegger Museum. In addition, there are some bathing lakes and many cultural, art and music festivals organized as well [88].

Schladming-Dachstein:

Schladming-Dachstein has a mountainous topography and as such, if offers a variety of hiking or climbing routes, skiing and cross country skiing areas. Dachstein glacier is the only glacier ski resort of Styria. Numerous regular winter sport events such as Alpine Skiing Cups and cultural and music events and festivals attract both foreign and national visitors. Other popular attractions are Dachstein Sky Walk, the nature park Sölktäler and various adventure parks, swimming baths and Trautenfels Castle [89].

Southwestern Styria (Südsteiermark, Schilcherland & Lipizzanerheimat):

Tourism region South-Western Styria consists of districts Deutschlandsberg, Voitsberg and Leibnitz. Topography in the north, the west and southwest is characterized by Alps and hills form part of the central and eastern part. Settlements stretch across the valleys and in the plain of the Leibniz field. The South-Western Styria has a total of over 49 km of ski slopes. In summer the mountains attracts numerous hikers. In addition, the region has about 250 hectares of lake area e.g. the reservoir Soboth or the Sulmsee. The southern half of the region is covered by cultural landscape. Four wine routes lead past numerous small taverns and other visitor attractions such as the castle Gamlitz, a wine museum and an interactive culinary exhibition in Vogau "Genussregal Südsteiermark". Further culinary themes roads include Styrian pumpkin seed oil trail and Styrian Milky Way. Other popular excursion destinations include a thermal spa in Köflach, the federal stud farm with the famous horses Lipizzaners, the Hundertwasser church, the glass museum in Bärnbach, the leisure island Piberstein, Seggau Castle and Stainz Castle and animal park Preding. In addition, many guests come to the various events such as vintage and cultural festivals[90].

Spa & Vulcano country and Eastern Styria (Thermen- & Vulkanland Steiermark and Oststeiermark):

Spa country (or so-called Thermenland in German) and Eastern Sytria represents the largest tourism region. It consists of the districts Hartberg-Fürstenfeld, Südost-Styria, Weiz and two municipalities of the District Bruck-Mürzzuschlag, namely Breitenau am Hochlantsch and Pernegg on the Mur. With the exception of Bad Gleichenberg, that dates back to the 19th century, the other health resorts were opened in the 1970s after discovering thermal water while digging for oil. The six spas of the region include thermal and healing spas in Loipersdorf, Bad Waltersdorf, Bad Blumau, in Sebersdorf, Bad Gleichenberg and Bad Radkersburg. Even though tourism is focused on wellness, culinary, wine and hiking also play an important role as tourist product in the region. In addition there are also bathing lakes, the Castle Riegersburg and Pöllau, animal park Herberstein, natural park Almenland Teichalm-Sommeralm and nature park Pöllauer Tal. In the northern part of the region, there are smaller ski resorts with 38 km of ski slopes and 40 km of cross-country ski trails [91].

Murtal Holiday Region (Murau - Murtal):

The "Murtal holiday region" consists of the the Styrian districts Murtal and Murau. It is entirely located within the Alps - Low Tauern mountains, that extend in the north and the Gurktaler Alps, Seetaler Alps and Glein Alps, that cover the south. The Alpine mountaineous landscape is the most important factor for tourism. 25% of Styrian ski slopes and 37 km of cross-country trails are here in ski resorts such as Kreischberg-Murau, Turracher height Grebenzen - Sankt Lambrecht and Lachtal. Most of the population lives along the Mur, which flows from west to the wider valleys in the eastern part of the region. Other attractions include the Red Bull Ring in Spielberg and the air show AirPower in Zeltweg, which attracted around 300,000 visitors in 2016. Other destinations include a thermal spa in Fohnsdorf, the Benedictine monastery Sankt Lambrecht, the Benedictine abbey Seckau, the Austrian Air Force Museum in Zeltweg or the Holzwelt Murau [92].

Gesäuse:

Topographically, the Gesäuse region is shaped by mountainous ranges of Low Tauern and the Ennstal Alps. The majority of the population settled in the western part of the region in the valleys. The tourism-intensive municipalities in terms of overnight stays per inhabitant, include Admont, Landl, Sankt Gallen, Lassing and Ardning. In contrast to other high alpine Regions of Styria the alpine winter sports tourism here play only a subordinate role. 80% of the arrivals fall between May and October. In summer, the mountain landscape attracts hikers, climbers and mountain bikers. The Rivers Enns and the Salza are suitable for diving, rafting and kayaking. A national park Gesäuse is the only one in Styria and extends over almost 111 (km^2) . There is also 586 (km^2) comprehensive nature park Styrian Eisenwurzen with its excursion destinations such as george Wasserlochklamm in Palfau, the cave Kraushöhle in Gams near Hieflau or the water adventure park Wasser-Spielpark Eisenwurzen in Sankt Gallen. Another important tourist destinations are the Benedictine monastery Admont with the largest monastery library of the World and a natural history and an art history museum, castle Strechau or the cave Odelsteinhöhle [93].

2.7 Semantics

Semantics is defined as "the study of meaning in natural and artificial language" [94], with an objective to extract and interpret a meaning of a group of words within a certain context [95]. On Twitter, which is one of most popular micro-blogging social media platforms [32] users were recently able to communicate using messages with up to 160 signs, which was later in 2017 extended to 280 characters per tweet [96].

In order to reveal the meaning of the tweet, semantic analysis has to, apart from classical text in natural language, deal also with abbreviations, slang, ambiguity, emotions and hash-tags. Hash-tags are a user-defined hyper-linked words which may convey for example a topic, emotion or event [32]. In line with vast growth in global conversation on social media there is also enormous amount of freely accessible communication data available. Twitter promotes itself by "understand what is happening" and "build on what is happening" or "tap into what is happening" [97] and offers its data to developers and analysers who are interested in public opinions and trends such as for example market analysts [98]. However, extracted tweets about a particular topic play a significant role also in academic research.

2.7.1 Semantics in Geography

Even though semantics are of great use for text mining in voluntary geographic information, the research field semantics in geography does not rise or originate from it. "Geospatial semantics as a research field studies how to publish, retrieve, reuse, and integrate geo-data, how to describe geo-data by conceptual models, and how to develop formal specifications on top of data structures to reduce the risk of incompatibilities"[99]. Even though at first sight the study of Semantics seems irrelevant in the field of Geography, they still play an important role. Semantics in geography are closely connected with spatial data standards and GIS interoperability, that enable integration of disparate data sets. Spatial data are shared within different organizations that are working in different systems. Only due to open standards a high level of interoperability is ensured. Interpolatable geodata can be shared and used across a variety of platforms, databases, in different development languages and applications [100]. Without agreed standards this would not be possible, since geo-data comes in very heterogeneous forms from thematic maps, satellite imagery, vector data or qualitative interviews. Also defining at a glance trivial geographic terms would not be possible. Geographic features types such as mountain, rivers, lakes, city are used in everyday language and due to our familiarity with these terms it is easy to assume a shared un-

derstanding. However, most of these terms have many domainspecific and incompatible meanings and formalization of such geographic terms is a difficult task. Such geospatial terms are taken for granted but due to many local definitions, it can still lead to misunderstanding while communicating with machines. Local definitions of geospatial terms cannot be globally standardized. In order to prevent such misunderstanding, further research on semantically annotated meta-data and complex ontology alignment services should be introduced [99]. Ontologies refer to "a set of concepts and categories in a subject area or domain that shows their properties and the relations between them" [101]. Thus, only better geo-ontologies enable mapping of local ontologies and the integration and of data from heterogenous sources [99].

2.7.2 Sentiment analysis

One of the fields of text mining within semantics is opinion mining, also referred to as sentiment analysis. Sentiment analysis reveals sentiment orientation, which classifies tweets into polarity classes such as positive, negative or neutral [32], referred to as polarity format. Another form is valence-based, that takes into account also the sentiment's intensity.

For instance, the two words "good" and "excellent" would be both treated positively using the polarity approach. But when using valance-based approach, "excellent" would be recognized as more positive than "good" [102]. In order to identify the attitude of a user towards a topic, there are common sentiment analysis approaches, including natural language processing, artificial intelligence, text analysis and computational linguistics [98]. Sentiment classes can be described by keywords or hashtags. A very suitable method is classification with lexicon-based approaches, e.g. via dictionaries of opinion terms or semantics of sentiment orientation [32].

Identifying the attitude towards a topic is not easy, since semantic information has to be extracted from text, while also dealing with denial, sarcasm, ambiguity of text, context dependence of word use etc. Correct determination of sentiment is useful for many purposes For example, for announcing the success of products, election results or sociological issues research. Sentiment analysis became a necessary research tool after the start of Web 2.0 due to world-wide opinion expressing through online forums, comments on news, tweets or other social media or through reviews of products and services. Sentiment analysis can be carried out at various levels, from a high level general overview of the entire text to individual aspects or characteristics within the text. However, at the latter level, a major problem is the identification of individual aspects [103]. In our research, we focused on the

entire text.

Sentiment analysis is similar to other tasks of text classification, since we classify the text into one of three sentimental categories. Roughly, there are two established approaches - lexical and machine learning. Lexical method requires one or more lexicon of sentiment that include words and phrases with a positive and negative connotation. The most simple approach classifies the individual words and phrases in the text and according to the prevailing words the whole text is denoted as a negative, positive or neutral. For good results a comprehensive lexicon of high quality is needed. However, there are problems of ignored context of the sentence and needed lexicons in a particular language. The same word can be of course both positive and negative in different contexts. In practice lexical approach should therefore be combined with machine-learning, since such a hybrid approach produces best results. In case of machine-learning approach, it constructs a sentiment classifier with the help of training text data [103]. This approach works with training data, since machine learning is "a branch of artificial intelligence in which a computer generates rules underlying or based on raw data" [104]. It is provded with new situations to help it learn, in a similar way humans learn, e.g. via experiences and observations, self-training or analysis [105].

VADER methodology:

In the software Orange used for our text analysis, there are two sentiment analysis approaches at disposal - lexicon based Liu-Hu and hybrid (lexicon- and rule-based) Vadar. Liu-Hu approach was first introduced in 2004 when the authors presented their experiments and new technique in a paper Mining Opinion Features in Customer Reviews [106]. It is a lexicon based method that is within Orange software limited to either the English or Slovenian language, but otherwise applicable to other languages as well. On the another hand, a Vadar method developed by Hutto and Gilbert in 2014 [8] is applicable to a dataset of text of mixed languages, since nonenglish text is translated through a machine-translation web service [8]. Our dataset of Tweets consists of various languages, mostly in German and English language but also some others such as Spanish, Italian, Slovenian etc. Because of multilingual dataset and further positive aspects of Vader method we decided for this approach.

VADER or Valence Aware Dictionary for Sentiment Reasoning is a lexicon and rule-based model for sentiment analysis. In the first step, lexical features were constructed with a combination of quantitative and qualitative methods. Furthermore, general rules, that embody syntactical and grammatical conventions human beings typically use for expressing a sentiment

Table 2.4: Examples of intensity of sentiment rankings[102]

Word	Sentiment ranking
tragedy	-3.4
rejoiced	2.0
insane	-1.7
disaster	-3.1
great	3.1

intensity are applied to the list of lexical features. Within an experiment of Hutto & Gilbert (2014) to assess the sentiment of tweets, the authors found out that VADER even outperforms human individual assessors. With their result, they proved how the incorporation of a human being in the development process can aid to the improvements in computer science [8].

VADER is a valence-based approach for sentiment analysis, taking into account both the sentiment itself and its intensity. It the table below examples of words and the degree of intensity of their sentiment ranking are displayed. More positive words have higher and more negative words have lower ratings. As a result, Vader provides 3 metrics annotating the proportion of the text, that falls into the positive, neutral and negative categories. The fourth metric, called the compound score, is the sum of all of the lexicon ratings standardised between -1 and 1 [102]. The closer to 1, the more positive the sentiment is and the closer to -1 more negative the sentiment is. A score of 0 represents a completely neutral text.

VADER approach is very suitable for working with social media, since it includes a selection of social media related terms or informal writing such as emotions, acronyms or multiple punctuation marks. Without this information the result could be completely different. For example a sentence would be first rated as neutral, but when considering an additional emotion symbol, a sentiment can be determined as strongly positive or negative [8]. Even acronyms 'omg' (oh my god), or 'smh' (shaking my head) are included in the lexicon. Acronyms help to determine the intensity of positivity and negativity. Word context, which is another important aspect, is taken into consideration as well, since the algorithm recognizes capitalization and exclamation marks, increasing the intensity or enabling the words to be taken into account within context of the tweet. This helps the algorithm to normalize the resulted sentiment. The sentiment both before and after 'but' word, is analysed but weighted differently, weighting the latter part of sentence more heavily. An additional aspect of Vader, is that even modifying words are not ignored, eg: "really good" is determined as more positive as only "good" [102].

Use of Sentiment analysis in Tourism:

Why do we use Sentiment analysis within our research? Sentiment analysis is commonly used in product and market research. What is people's opinion about a product, what kind of people use it and what should be improved? In case of tourism, we can easily refer to it as a product. Communities, investors or local people offer their city, site, museum, shop, excursion, adventure opportunities as a product and it is necessary to know the market's response to it. Visitors share their opinion via social media and a sentiment analysis is a great technique to extract their feedback. According to extracted opinions, communities or investors can properly react in order to either improve tourism offer or plan their actions in terms of future tourism development.

2.7.3 Text Mining

"Text mining is the process of analysing collections of textual materials in order to capture key concepts and themes and uncover hidden relationships and trends"[107]. Within this process, it is not required to know exact words or terms used by authors but rather defining the patterns within the text. Unstructured and semi-structured information comes in the form of e-mails, open survey responses, news feeds, comments of readers of an online newspaper, forums, social media feeds etc. The question is, how to collect and examine this information. Via text mining, a computer has to understand and reveal the meaning of such unstructured text that is written without any standard rules. There are two general approaches: Linguistic and nonlinguistic [107].

According to the IBM [107] definitions:

- Nonlinguistic (Rule-based) text mining - Nonlinguistic text mining refers to automated solutions for text scanning and categorizing the key concepts faster than a human reader. It is based on statistics and neural networks. The accuracy is not as high as using linguistic approaches and the occurrence of irrelevant results or noise can quite often occur. Most systems are not flexible enough and simply calculate the statistical proximity of counted words to related key concepts. A potential solution to low accuracy, are Rule-based text mining approaches, that are, with the help of incorporated complex nonlinguistic rules, able to distinguish between relevant and irrelevant results .
- Linguistic Processing and NLP - A linguistic approach of text mining is much more accurate, since it is based on the principles of natural lan-

guage processing (NLP). Natural language processing is "a computer-assisted analysis of human languages in order to perform the analysis of words, phrases, and syntax, or structure, of text." NLP encompasses a variety of algorithmic methods and just as people recognize a variety of word forms with similar meaning and analyse sentence structure, NLP can determine the meaning of the text. Using extracted concepts, compound phrases, meaning and context of the text, NLP classifies it into related groups e.g. groups of products or organizations.

NLP methods are categorized into supervised and unsupervised. Supervised methods require a manually classified sample set of documents by researchers. According to this training data algorithms learn word association and apply this knowledge to other collections of documents. Unsupervised methods do not need any training data, but learn directly from the use of words and patterns in the documents. An example of unsupervised learning is topic modelling [108].

Term frequency analysis and Word Cloud:

Word clouds are, in the field of text mining, seen as a very visually appealing and straightforward text analysis approach. They are based on term frequency within the text - those words appearing more often are visually emphasized, being displayed bigger or in a coloured form. Technically, it corresponds to static text summarization, but can still act as a powerful tool for general test analysis tasks [109]. It is very useful for identifying words that frequently appear in documents, interviews or finding out what topics are most discussed on social media. There are many free word clouds generators enabling a simple and quick term frequency determination and a graphical representation in a form of a word cloud. Yet, basic word cloud generators application provide no means to compute the underlying text analysis and can only be transformed into a powerful tool for text analytics with additional information and interactive features [110].

However, our analysis did not focus only on a graphical representation of a word cloud but, in particular, on determining the frequency of words, that initially takes place prior the word cloud generation, and on mining after tourism-related words. In order to perform a meaningful term frequency analysis, the text should be preprocessed and all “stop-words”, punctation, extra whitespace, URL and non ASCII signs filtered out. For the purpose of this master thesis, we performed term frequency analysis in the programming language R.

2.7.4 Topic modelling

Topic modelling is a technique of text mining using algorithms with statistical modelling to scan a set of documents and discover the abstract categories or clusters or words that occur in a text and best characterize the whole corpus. Included within topic modelling are the extracted coherent categories referred to as "topic" and a set of text documents as a "corpus" [108]. Topic modelling techniques find their use in various applications, for instance in qualitative studies for opinion analysis for marketing purposes, media studies or sociological research. However, as pointed out by Nikolenko et al.[111], who also proposed new solutions to these problems, there are two main downsides: The lack of an evaluation coefficient of high quality that matches human judgement of modelling result and secondly, it is not possible to indicate the specific topics we are mining for. As also in our case, the modelling result does not return specific topics we are most interested in. Our goal of topic modelling was to determine the types of tourism revealed from Tweets, for example types of winter sports, hiking or urban sightseeing. However, topic modelling algorithms return a predefined quantity of numbered categories without any semantic name. Since a text document typically contains multiple topics in different proportions, a weight per document per topic is also returned [112].

The three topic modelling algorithms offered by Orange Data Mining software are Latent Semantic Indexing, Latent Dirichlet Allocation and Hierarchical Dirichlet Process. The only parameter accepted by Latent Semantic Indexing (LSI) and a very commonly used Latent Dirichlet Allocation (LDA) is the required number of modelled topics. Hierarchical Dirichlet Process (HDI) is intended for more advanced users, since there are many parameters that should be taken into consideration. The three possible algorithms offered by Orange software are explained below [112]:

- Latent Semantic Indexing - LSI - Latent Semantic Indexing or Latent semantic analysis (LSA) is based on the distributional semantics hypothesis and works under assumption that words of similar meaning occur in similar parts of a text. The similarity of words is evaluated from a part of text with a mathematical technique SVD - singular value decomposition matrix [113]. Within Orange Data Mining software topic modelling with probabilistic LSI provides both positive and negative weights, meaning the word is highly representative or unrepresentative of a topic [112].
- Latent Dirichlet Allocation - LDA - LDA was introduced for the use in machine learning by Blei et al. [114]. It is a generative probabilistic three-level hierarchical Bayesian model using inference techniques

based on variational methods and an EM algorithm. A document is represented by topic probabilities, since "each item of a collection is modelled as a finite mixture over an underlying set of topics and each topic is, in turn, modelled as an infinite mixture over an underlying set of topic probabilities" [114].

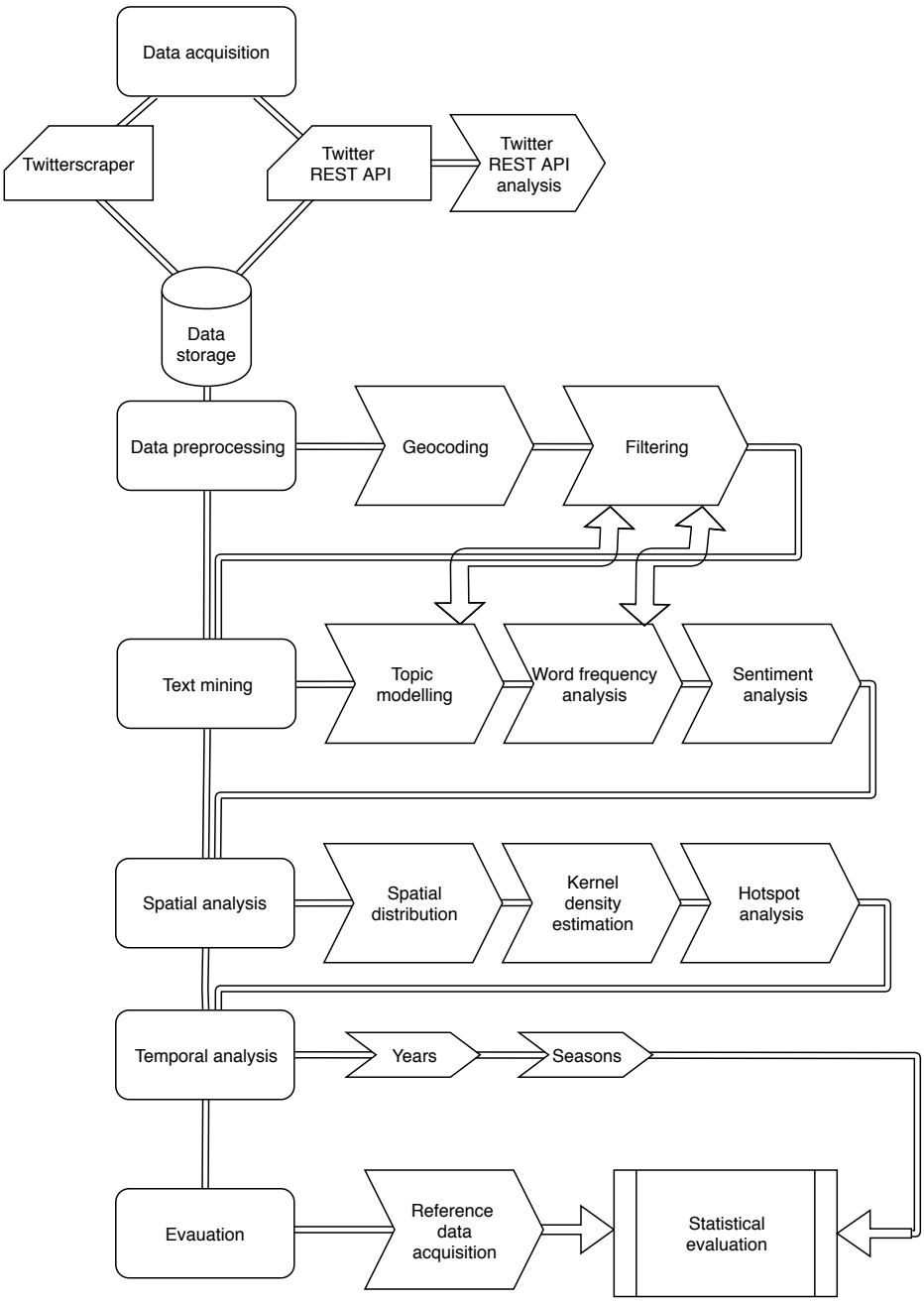
- Hierarchical Dirichlet Process - HDP - HDP is computationally a very demanding nonparametric Bayesian algorithm. With the help of Dirichlet process it groups the data and clusters these groups so that they can share their statistical strength [115]. Within Orange software there are seven parameters that have to be simulated [112].

Chapter 3

Experiment

In this chapter, the whole experimental workflow and the results is described. Our working processes started with the data acquisition with the help of Twitterscraper API. We acquired 35,234 tweets. The acquired data was stored in a MongoDB database where each document (each tweet) was geocoded - i.e. it was updated with the coordinates of corresponding municipality. Filtering was an extensive process also performed within the MongoDB database. Within this step we applied text mining in order to choose the final selection of tourism relevant words and then returned to filtering in order to extract the final dataset of our tweets. In a final dataset that consisted of 6,953 documents a sentiment analysis and topic modelling were applied in Orange Data Mining software. Spatial analyses and geographical visualizations of the results were performed in ArcGIS. Within the evaluation a reference dataset was acquired and used for correlation analysis implemented in the statistical software SPSS. Results are represented in the form of graphs and maps.

Figure 3.1: Workflow



3.1 Data acquisition

Data acquisition was performed with the Twitterscraper Python package. It was based on a spatial scale of municipalities and tourism-related keywords in English and German. Each request was composed of the municipality name, state name Styria (Steiermark) and 12 tourism-related keywords in English and 3 in German. For each of the 287 Styrian municipalities one request was performed, i.e. 287 individual requests. The keywords used are presented in the table 3.1. Tourism keywords were collected with the help of keywords dictionaries ([116], [117], [118] etc.) and our own professional expertise in the field of tourism.

Table 3.1: Predefined tourism-related keywords for data acquisition

English keywords	German keywords
Styria	Steiermark
trip	Urlaub
travel	Reise
travelling, traveling	reisen
traveller	
tourism	
tourist	
holidays, holiday	
vacation, vacations	
adventure	

Our first request sentence structure was simple and intended for mining tweets with tourism-related keywords at particular coordinates with a defined search radius by using near and within operators. The results were unsatisfying and further request-building process followed a repeated "learning by doing" approach, since we had to deal with low quantity and low location accuracy of returned tweets. By using the search radius there was also a possibility of duplicate tweets from the overlapping areas. Our solutions are explained below.

Figure 3.2: Query Syntax Example 1, based on coordinates and search radius

```
twitterscraper "trip OR travel OR travelling OR
traveling OR traveller OR tourism OR tourist OR
holidays OR holiday OR vacation OR vacations OR
adventure OR Urlaub OR Reise OR reisen"
 near:15.439337 47.07512 within:20km"
-o C:/twitterscraper/Data/Graz.json
```

We decided to work with municipality names and not with coordinates

and search radius. We tried out both options and using municipality name as a location parameter resulted in up to 90% more extracted tweets. With this approach the quantity of returned tweets was therefore increased. We also found that the location accuracy of the tweets extracted with -near parameter only was quite low. We analysed and checked the location accuracy on a selection of returned data. Location accuracy was manually compared to the location revealed in the text. As a solution we also added a corresponding municipality name as a keyword to each query. In order to increase location accuracy we therefore used a municipality name both for localization of the tweet and as a keyword.

Our query at this step is shown bellow:

Figure 3.3: Query Syntax Example 2

```
twitterscraper "Graz AND (trip OR travel OR
travelling OR traveling OR traveller OR
tourism OR tourist OR holidays OR holiday OR
vacation OR vacations OR adventure OR Urlaub
 OR Reise OR reisen) near:Graz"
- o C:/twitterscraper/Data/Graz.json
```

The quantity of returned data at this step was still too low to perform research of a high quality. Although we were aware that we were handling a very fine spatial scale, we tried to find a solution and extract further tweets that did not use any of our predefined tourism keywords but still indicated on tourism activities using other words.

In order to enlarge the quantity of returned tweets we therefore also added the name of the state Styria (Steiermark) to the primary selection of keywords. With this approach we also hoped to extract also other tourism-related tweets that otherwise do not contain any of our main 12 English or 3 German predefined tourism keywords. While looking for a solution we had to choose a keyword that also brought a possibility of returning tweets other tourism-related topics. Deciding for the state name as a suitable keyword followed the presumption that people that tweet about their tourist activities also reveal their location within the text. However, we were well aware that this approach would require much more extensive filtering in the next step in order to filter out all tourism-irrelevant documents.

The final query is represented in Figure 3.4. The municipality keyword was connected with the tourism-related keywords with a logical operator AND. The keyword Styria was connected together to other tourism-related keywords with an OR operator. The request has therefore mined for the tweets that mentioned the municipality name and any of the tourism-relevant keywords or the state name Styria (or Steiermark in German).

Figure 3.4: Query Syntax Example 3

```
twitterscraper "Graz AND (Steiermark OR Styria
OR trip OR travel OR travelling OR traveling OR
traveller OR tourism OR tourist OR holidays OR
holiday OR vacation OR vacations OR adventure
OR Urlaub OR Reise OR reisen) near:Graz"
-o C:/twitterscraper/Data/Wildon.json}
```

A further challenge we were confronted with was to achieve the uniqueness of municipality names. The names of some municipalities in Styria are not unique, since they also exist in other states or countries. Hence, when dealing with such municipalities, the request queries were adjusted with the obligation that next to the municipality name either one of the words Styria (Steiermark) or Austria (Österreich) has to be mentioned. With such obligations we aimed to extract only tweets referring to a Styrian municipality. In this case, the code example is shown in Figure 3.5.

Figure 3.5: Query Syntax Example 4

```
twitterscraper "Sankt Gallen AND (Steiermark
OR Styria OR Austria OR Österreich) AND
(Steiermark OR Styria OR trip OR travel OR
travelling OR traveling OR traveller OR
tourism OR tourist OR holidays OR holiday OR
vacation OR vacations OR adventure OR Urlaub
OR Reise OR reisen) near:SanktGallen"
-o C:/twitterscraper/Data/StGallen.json}
```

The request was applied for each of 287 municipalities. For 80 requests no tweets were found, meaning 207 municipalities were returned. All together we acquired 35,234 tweets, dating from 16.02.2008 till 22.08.2018.

Figure 3.6 displays the returned data of an example of an extracted tweet. It consists of the fields of fullname, html, id, number of likes, replies and retweets, text, timestamp, url and user. There is no location data included, so it is important to keep track of what files were returned to what location-related query. However, a nested object of coordinates with an array of latitude and longitude was added in the geocoding step.

Figure 3.6: Example of extracted tweet in .json format


```
{
    "fullname": "Lisa Marie",
    "html": "<p class=\"TweetTextSize js-tweet-text tweet-text\"
    "id": "953915758727213056",
    "likes": "0",
    "replies": "0",
    "retweets": "0",
    "text": "#BLOGGED - besides traveling also reading
             is one of my biggest passions so when i
             visited ADMONT\u2026
             https://www.instagram.com/p/BeFchRsDEkK/\u00a0",
    "timestamp": "2018-01-18T09:03:34",
    "url": "/WanderlustinLiz/status/953915758727213056",
    "user": "WanderlustinLiz"
},
```


3.2 Data preprocessing

In order to prepare the data for the analyses, it had to be geocoded and cleaned of tourism-irrelevant tweets. Since we focused on users represented by natural persons we also filtered out all tweets issued by organizations.

3.2.1 Data importing and geocoding

Each file with tweets from an individual municipality was separately imported to a common collection in the MongoDB database. Importing individually was essential in order not to loose the information of which municipality the filed belonged to, since the geocoding followed as a second step directly after the importation of tweets belonging to a particular municipality. In each geocoding step, only municipalites that still did not have the object "municipality" were updated. Each document was updated with a nested object consisting of municipality name and the corresponding coordinates. The coordinates of municipality's centroid were used. The coordinates of centroids were calculated beforehand with the help of a shapefile of Styrian municipalities and QGIS geographic information system software. The coordinates correspond to the WGS 1984 reference coordinate system and were added in a form of decimal degrees.

At the end of the importation process, the database included a collection of 35,234 documents.

Figure 3.7: Import example

```
mongoimport --jsonArray --db Gemeindeebene
--collection Gemeinden
--file C:/twitterscraper/Data/Graz.json
```

Figure 3.8: Example of geocoding

```
db.Gemeinden.update(${'municipality':
{'$exists':false}},
{'$set': { 'location' : { type: 'Point',
coordinates: [15.43933,  47.07512] },
'municipality': 'Graz'}  },
    {multi:true})
}
```

3.2.2 Filtering

Filtering was the most extensive and fundamental process within this research in order to achieve results of high quality. In order to determine a suitable filtering workflow, an extensive insight into the dataset and an overview of the tweets characteristics had to be achieved. A large selection of tweets was manually examined. Determined filtering workflow followed our predefined objectives and was in general applied in 2 steps:

- A) Applying filter on all extracted tweets.
  1. Focusing on natural persons and no organizations.
  2. Filtering out according to the tourism-irrelevant keywords.
  3. Filtering due to same username and municipality name
  4. Cleaning users reaching too many tweets.

- B) Applying adjusted filter on 3 differently categorized subcollections of tweets.

  The extracted dataset consisted of tweets with and without our predefined tourism-related words. Documents without predefined tourism words were extracted due to mentioned state Styria (Steiermark) and the relevance to tourism of such documents had to be checked through a further filtering process. Due to the popularity of mentioning personal location in Twitter, a large group in our dataset consisted of documents revealing location, such as "I am at museum in Bad Aussee".

According to these characteristics, documents were reorganized from the origin collection into 3 new subcollections. Filtering on these belonged in the filtering step B.

Yet the process was not straightforward, since in some cases the sub-steps of applying the filter on all extracted tweets had to be implemented within the 3 subcollections as well. The entire filtering process of steps A and B is explained below.

3.2.3 A) Applying filter on all extracted tweets

The first filtering step was applied to all 35,234 extracted tweets with the following techniques:

Focusing on natural persons:

Extracted tweets included both natural and legal persons, such as organizations and companies that use Twitter as an advertising medium. For the purposes of our research, all legal persons were filtered out in order to create a data set focused on private Twitter users. Documents were aggregated according to the user's fullname. We discovered that a large proportion of the documents refer to organizations such as real estate and job agencies. For instance, the user "Immoads", a real estate agency, was responsible for over 4,000 tweets.

Filtering according to the keywords irrelevant to tourism:

Within the text of our tweets, many keywords irrelevant to tourism were found. According to the keywords dictionaries ([116] [117], [118] etc.) and our expert insight on tourism, the documents including words referring to job searches, working and selling or renting items were determined as irrelevant and were deleted from the dataset

Filtering according to the individual's quantity of tweets:

In order to obtain a stable dataset of varying individuals, it was necessary to filter out the noise of individuals contributing a large number of tweets. The remaining natural persons were therefore additionally aggregated according

to their fullname and username. Apart from individual cases with over 1,000 tweets, most of the users did not exceed 10 tweets. However, because we are handling a dataset within a time frame of 12 years we filtered out only the users with over 20 tweets.

Filtering due to same username and municipality name:

Within the preprocessing of our dataset we found, that many documents were extracted due to the username or fullname of the user being the same as the municipality name. Hence, the documents where fullname or username equals the municipality name were deleted. Apart from this there were also many cases where the municipality name was mentioned in the text but actually referred to a surname of some person the user was tweeting about. These cases were also extracted via searching through the text field and were manually deleted if the municipality name referred to a person and not a location. For instance, many of these cases occurred in the municipalities Schäffern, Rohrbach, Gralla and Hartl.

3.2.4 B) Applying adjusted filter on 3 differently categorized subcollections of tweets

As a result of the first step of filtering, our dataset consisted of approximately 17,000 tweets. In order to extract as many tweets related to tourism as possible, we further built 3 groups of tweets in 3 separate subcollections and applied adjusted filters on each of them.

Documents that include any of the predefined tourism keywords:

This group of tweets was assumed to contain mostly tweets with a strong relevance to tourism, since they included any of the predefined 12 keywords in English or 3 in German. However, within this group we still found further documents issued by legal persons or that mentioned a municipality name that actually referred to a person's surname. The filter developed within this subcollection was also applied to 2 other subcollections in order to avoid the same issue.

Documents that reveal location by using "I am at ..." or "I am in ..." sentences:

In the section of the second categorized subcollection, the revealed location and its relevance to tourism had to be taken into consideration. The location is commonly revealed through sentences using "I am at ... " or "I am in ...". Many users were tweeting about being in a hotel or in a museum, being on a viewing platform or hiking on a particular hill or a mountain. On the other hand, some users revealed in which shop or restaurant they were located. Our task was to determine which locations we observe as tourism-related places and which ones not. The methodology to implement this task is explained in the text analysis section 3.3.

The rest of the documents:

The rest of the documents are the tweets that did not pass to the first two categories. These documents were extracted due to their text content which mentioned both the name of the municipality and the state of Styria (or Steiermark in German). They did not include any of our primary predefined tourism-related keywords. However, we acted on the presumption that more tourism-relevant tweets may be dug out. Hence, we had to define a further approach to indicate tourism activities. This group needed the most extensive further filtering.

The decision to mine for further tourism-related tweets was correct, since we obtained a higher number of tweets from the third subcollection. Within a manual examination of a part of the tweets, we determined that there were many tweets that were relevant and should not be lost.

However, there were also many that were completely irrelevant and therefore should had been eliminated. For this purpose, we applied a text analysis. We tried out two methods - topic modelling and extracting the word frequency within all the documents.

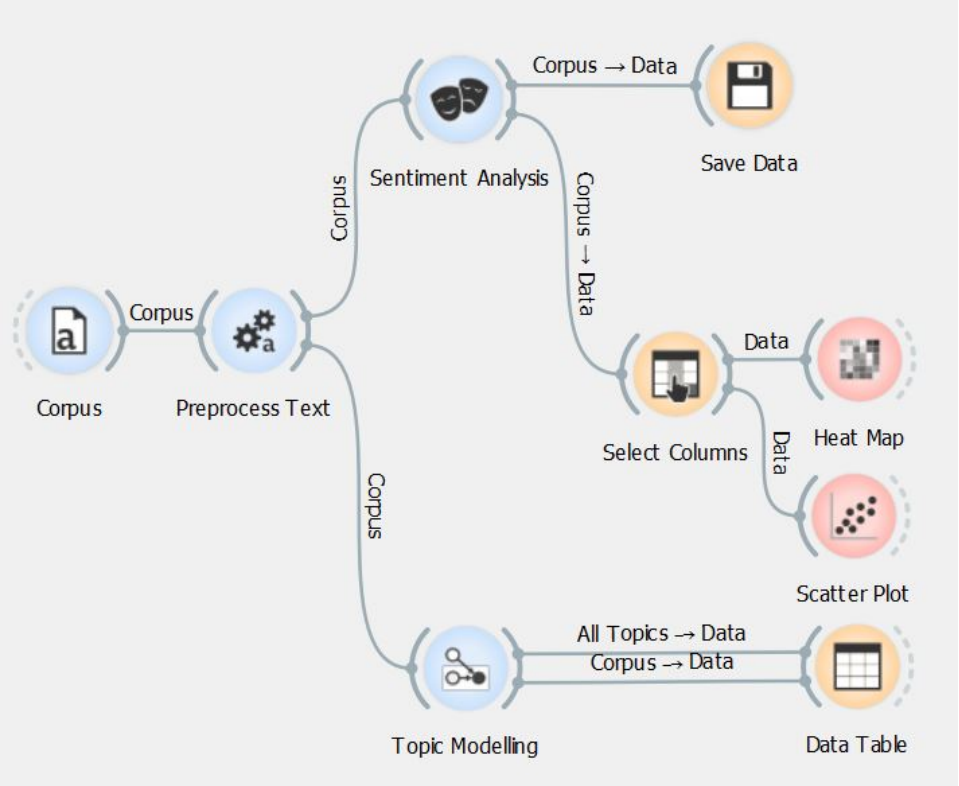
3.3 Text Analysis

Within this research, text analysis was implemented for two purposes. First, as an instrument to further determine tourism-relevant keywords within our dataset and so continue with further filtering of the second and third categorized subcollection. Second, text mining was used as a technique for sentiment analysis.

Text mining techniques used within this research:

- Topic modelling
- Term frequency analysis and Word Cloud implemented in R programming language.
- Sentiment analysis

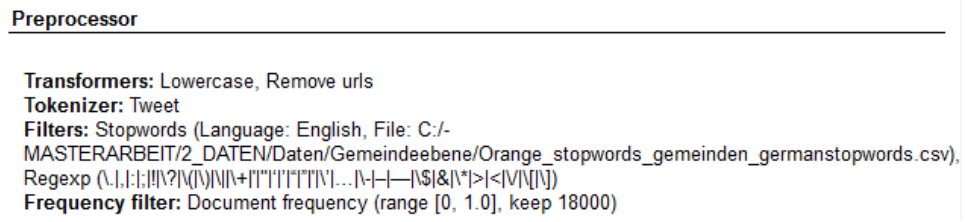
Figure 3.9: Text mining implementations within Orange software for data mining



3.3.1 Text preprocessing

As a prior step to text mining, text has to be preprocessed. the text was transformed to lower case and URL adresses were removed. Within the tokenization process the tweets are transformed into corpus, which refers to a dataset used in text mining. Tokenization breaks the text into smaller units (tokens) such as words, sentences and bigrams. As tokenization approach, a method intended for working with tweets was used. Hence, according to the pre-trained Twitter model, hashtags, emoticons and other special symbols are preserved [119]. As a result 64,864 tokens were created. In addition, stopwords in English and German were also eliminated. Because the software enables only one language at time, English was chosen as a default language for stopword elimination and German as an additional one through an added file, where all the German stopwords were listed. The German stopwords list was set up with the help of R programming language. Another setting Regexp (regular expression) removed the punctuation and brackets and other unregular signs. For frequency determination of tokens, all of the first 18,000 most frequent ones were kept.

Figure 3.10: Preprocessing within Orange software for data mining



3.3.2 Topic modelling

The objective of topic modelling was to determine groups of topics according to the words used in the tweets' text. We tried out two methodologies: Latent Semantic Indexing and Latent Dirichlet Allocation. With both approaches the words (corpus) are grouped into the predefined number of topic categories.

Topic modelling was applied for two purposes. First, in the frames of filtering process, when we used it as a technique to determine further tourism-related words from the subcollection nb. 2 (with "I am at ..." sentence structure) and from subcollection nb. 3 (with the rest of the tweets). Second, we applied topic modelling on the final dataset, where only tourism-related documents were left. This second application was implemented in order to define the types of tourism. However, the results of both applications for the

first and second purpose were unsatisfying and of low quality. We accepted the decision to reject the results of topic modelling and to find alternative techniques.

Topic modelling implementation for filtering reasons:

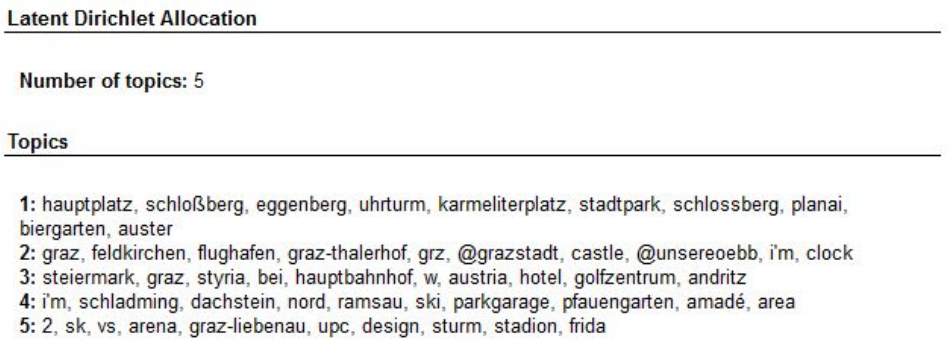
Our experiment in the field of topic modelling did not bring relevant results to our research, but still counts as an working step, that should be explained.

Through preprocessing, tokens were created and corpus was prepared for text modelling. With used approaches (Latent Semantic Indexing and Latent Dirichlet Allocation) the words are grouped into the predefined number of topic categories. We tested both methods since we expected better results from the more commonly used Latent Dirichlet Allocation. We examined the different results of the most frequent tokens using several different upper limits and several different numbers of groups, for example the results of an analysis having the first 18,000 most frequent tokens grouped into 3 and up to 15 groups. However, it didn't matter what upper limit of most frequent tokens and number of groups we set up, the results were so mixed that we were unable to determine a group of words whose meaning would be relevant to tourism. After categorizing the words, topic modelling requires a manual determination of the semantic meaning of each group. In our opinion, the words within each category of our dataset were still much too diverse to be able to point out one or more groups of tokens that tokens that seemed to be relevant to tourism.

Also, within our second application of text modelling when we imported only tourism-related documents and wanted to categorize them by tourism type, the results produced by the algorithms did not correspond to our human ability to categorize and interpret the text by tourism type. Even though the results were much better and it was possible to determine topics for the returned groups they still did not match our wish to determine tourism type.

There could be many possible reasons for the unsuitable results. We wanted to achieve results on a very fine scale. In the case of a topic modelling application on a dataset of various topics, it would be easier to categorize it for different topics, for example on well-defined categories, e.g. working and free time topics. However, we have already used a double prefiltered dataset (first filtered by data acquisition and secondly in the filtering process in MongoDB) and the tourism type categories we expected were too narrow. In addition, our final dataset may not have been large enough, and if we had more (hundreds of thousands) tweets for the algorithms to work with, the re-

Figure 3.11: Result of Latent Dirichlet Allocation method



sults might have been better. Also, we used very standard techniques. More complicated approaches such as, for example, the more advanced Hierarchical Dirichlet Process (HDI) would maybe deliver better results. However, topic modelling only played the role of an additional experimental task and we could not explore all possibilities that it offered since our main objective was to focus on geoinformatics, namely the geographical side of the extracted tweets. Yet, text mining and topic modelling are very promising techniques in informatics and machine learning in order to extract tweets that exactly correspond to a particular research topic. In such a way, further analyses in other research fields would also deliver better results.

3.3.3 Term frequency analysis and Word Cloud

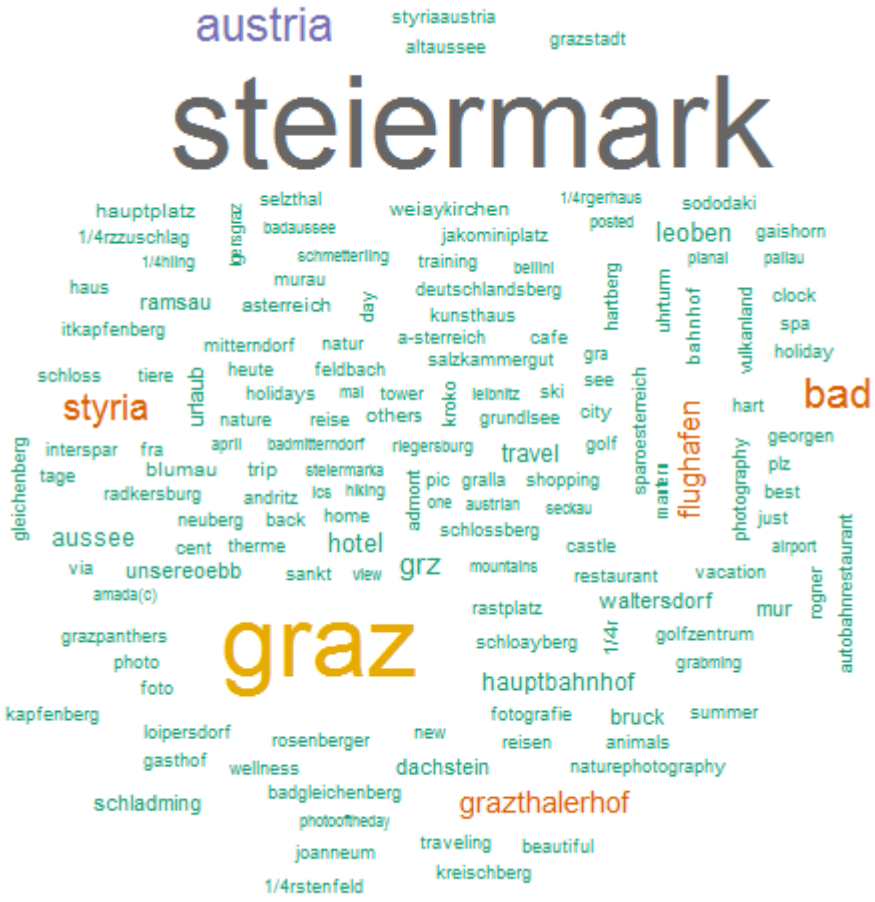
In order to determine further tourism-related keywords that we can use for further filtering of the two subcollections, term frequency analysis was applied. Term frequency analysis was implemented with R programming language. From all the tweets, we listed the most frequent words and manually determined the keywords with a high relevance to tourism due to our expert insight into the topic. To do so, a definition of the kind of tourism we are dealing with had to be determined. Since our objective was also to extract all the day tourists that are usually not included in the official statistics, we focused on a broad selection of free time activities in public places. Therefore, we targeted all keywords indicating free time public activity. We decided for such methodology in order to also include tweets from our second categorized collection where users tweet on everyday free time public activities such as being in a museum or in a cafeteria.

Text mining within R programming language has to be implemented on only one text string in order to examine a word frequency within this large string. Fields of tweet's text from all documents were therefore collapsed

into one string of words. As such, a corpus was built and prepared for further filtering. In a further text preprocessing process, the whole content was transformed to lowercase letters, punctuation was removed, extra whitespaces were stripped and numbers and stopwords in English and German were removed. We also imported an additional list of stopwords containing all the municipality names, since they occur in a large number of tweets. Finally, URL addresses and all non ASCII signs were also eliminated.

Out of this cleaned up data, a document-term matrix was built and the word frequency was determined. There were 18,000 words and we manually checked the first 5,000 of them to determine whether the word was related to tourism or not. For this task we again used the keywords dictionaries ([116] [117], [118] etc.) and our expert insight on our topic. As a result we obtained 229 words which we used for further filtering for the second categorized subcollection indicating location by "I am at ..." or "I am in ..." sentences and of the third categorized subcollection with the rest of documents which did not pass into the first two categories and were extracted due to the mention of a particular municipality and the state of Styria. Additionally, we also indicated 15 further words that clearly indicated no relation to tourism, e.g. words related to work or selling advertisements. Documents including these tourism-irrelevant words were again filtered out.
For visualization purposes we also set up a word cloud of the 150 most frequent words.

Figure 3.12: Word Cloud of the first 150 most frequent words



3.4 Final dataset

After the entire filtering process our dataset consisted of 6,953 tweets.

Table 3.2: Final count of tweets in 3 subcollections

Subcollection	Final number of tweets
Predefined tourism keywords	1,391
"I am at..." structure	2,717
The rest of the tweets	2,845
Total	6,953

Figure 3.13: Count of total extracted tweets at a district level

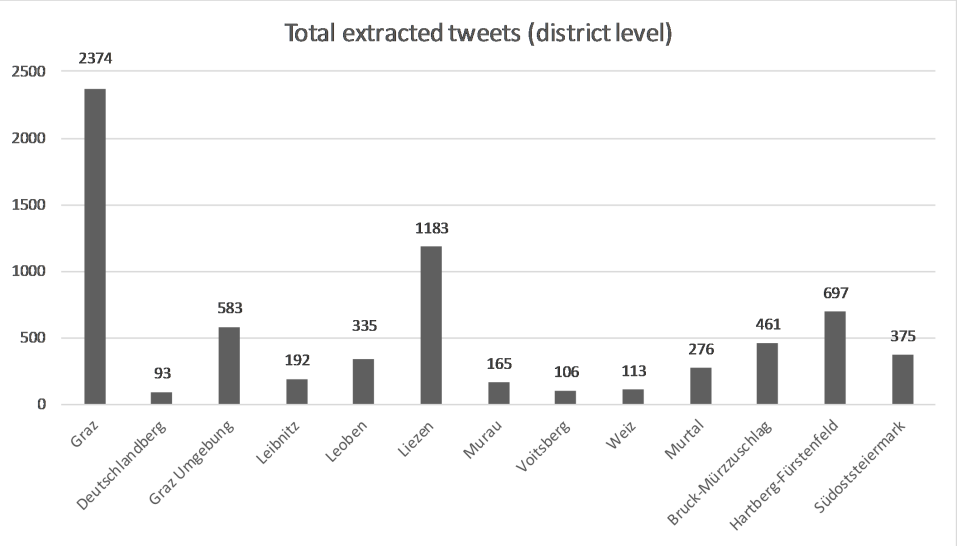
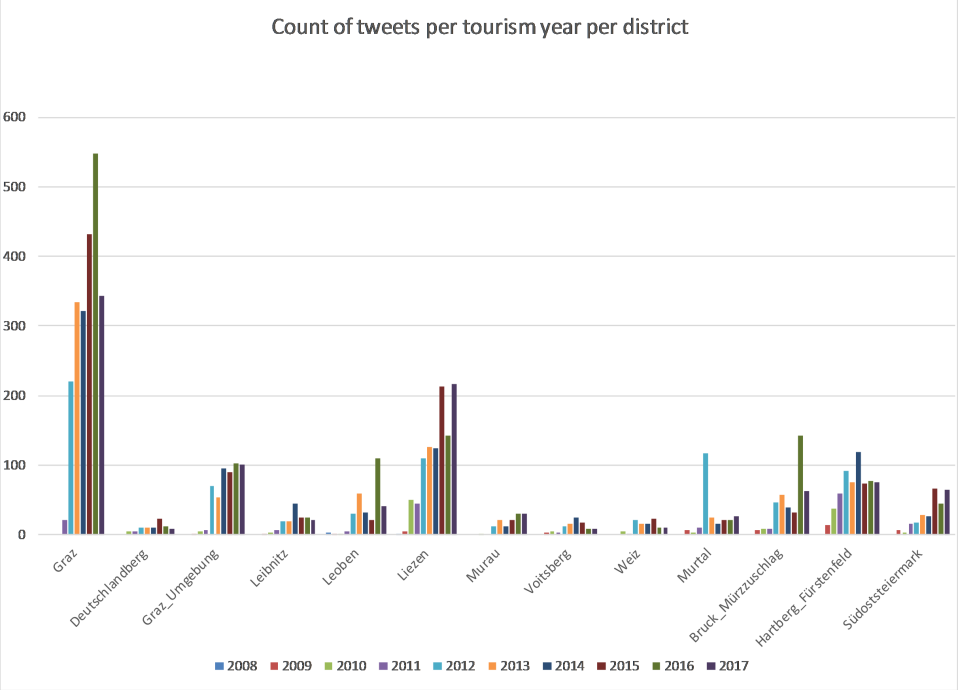


Figure 3.14: Count of total extracted tweets per tourism year at a district level



3.5 Twitter API evaluation

Parallel to our research using the data acquired with te Twitterscraper package, we also ran an analysis of Twitter REST API. We created an Python script and in the time period between 31.08.2018 and 28.12.2018 we connected multiple times to Twitter REST API in order to extract tourism-related tweets within Styria. The goal of this analysis was to determine the quantity of returned tweets via official Tweepy REST API versus Twitterscraper and to deal with its limits of returned tweets, that are no older than 7 days.

The obvious advantages of extracted data from Twitter REST API are the exact coordinates, where they exist, returned for a particular tweet. However, due to scraping with a location ID, API mostly returns tweets that contain exact coordinates. Within our work, 80 - 90% of each returned file of tweets also contains exact coordinates. Another big advantage is a large number of fields within the json object. Compared to Twitterscraper data, where only 10 fields are returned, Twitter REST API returns an extensive dataset of fields with a number of nested objects including all metadata considering location, user, tweet-related information etc.

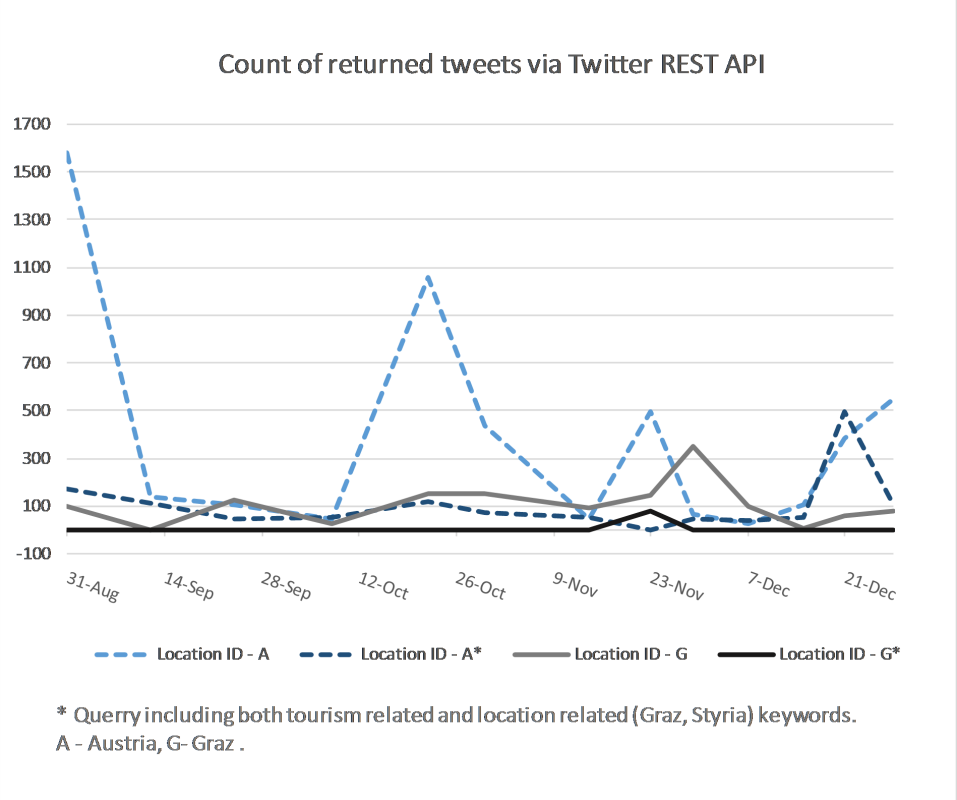
Data acquisition was performed in roughly weekly intervals within the mentioned four-months time period. Each time we requested data using two different location parameters - the country Austria and the city of Graz. Location parameters within Python syntax are implemented in a form of a place ID. Using the state of Styria as a location parameter was not possible, since no place-id exists for a spatial level of a state. Additionally, within each location we used two different syntaxes - one containing only predefined tourism-related words (the same predefined tourism keywords as in syntax used with the Twitterscraper package) and another one using both tourism-related keywords and also location-related keywords, such as Graz, and the name of the state in English and German language - Styria and Steiermark. For each date we connected to the REST API which returned four files, two with a place id of Austria and two for the city of Graz. Our final dataset consisted of 56 files related to 14 dates within a four-months time period, 28 corresponding to the place ID of Austria and 28 to the place ID of Graz.

Table 3.3: Count of extracted documents via Twitter REST API.

	Austria	Austria*	Graz	Graz*
Count	5,054	1,362	1,387	86
Count December	1,070	689	250	4

A lower count of returned tweets within our datasets marked with a star (*), was easily predicted, since documents were extracted based only on

Figure 3.15: Count of returned tweets via Twitter REST API - four different queries



predefined tourism-related keywords and no location-related keywords as for the other two datasets. However, both datasets with the Austria location ID refer to the whole country. On the contrary, we are only interested in tweets originating from the state of Styria. Similarly, both datasets with the location ID of Graz only refer to the state's capital and not to the whole of Styria. Datasets should also be exposed to extensive filtering before drawing any conclusions. Based on our experiences when dealing with Twitterscraper datasets, we filtered out 80.23% of originally extracted tweets. Furthermore, according to the official statistics, Styrian annual tourist arrivals account for 9.37% [120] of the total arrivals for Austria, and the city of Graz for 17.24% [121] of all Styrian arrivals (calender year 2018). Because we only deal with 4 months dataset, we collected statistics from September till December 2018. In this four months time period, count of tourist arrivals to Styria account for 9.7% of those in whole Austria. In the same period, arrivals to Graz account for 18.2% of the Styrian ones. As such, the whole of Styria had 550% of the Graz arrivals. In order to achieve a rough estimation, we could choose to follow a simple mathematical approach and apply the proportion of our filtered dataset (19.7% of the original one) and the proportion of

official arrivals. With such a projection we could achieve an estimation of how many tourism-related tweets in Styria it would be possible to extract via Twitter REST API in a four-month time period. However, it would be more recommendable to apply the proportions from our Twitterscraper dataset and not from official statistics. We do not posses data on the whole of Austria, but we do posses data from Graz. In the four months between September and December 2018 there were 262 tweets throughout Styria and 67 in the municipality of Graz in the final filtered dataset. As such, Styria accounts for 291% of the Graz count.

A similar projection was also applied on December's dataset acquired from Twitter Rest API. In this projection we can again follow official statistics that reveal that in 2018 Styria comprised 8.4% of all Austrian visitors. Graz comprised 17.,2% of all Styrian visitors and, as such, all Styrian visitors account for 580.9% of Graz visitors. When using statistics from the Twitterscraper analysis, there were 15 visitors in Graz in December 2018 and 59 in the whole of Styria. This accounts for 393.3% of all Graz's visitors.

These numbers were projected onto the dataset achieved by Twitter REST API. However, we did not apply filtering proportions on the dataset acquired with only tourism-related keywords. This model is very rough and should undergo further research. However, it still offers a first insight into the number of possible suitable tourism-related tweets within Styria acquired by a Twitter REST API. The results in following two tables gives us numbers of possible acquired tweets in four months or in one month. Depending on the purpose of the research, but projecting this count of tweets in a dataset over one or more years might be sufficiently large for professional use.

Table 3.4: Projected count of Twitter REST API tweets in Styria - four months

	Austria	Austria*	Graz	Graz*
Count	5,054	1,362	1,387	86
After filtering (20%)	1,010.8	/	277.4	/
Proportional four-months value for Styria (official data)	98.2	132.3	1526.4	473.2
Proportional four-months value for Styria (Twitterscraper)	/	/	1,084.6	336.3

Table 3.5: Projected count of Twitter REST API tweets in Styria - one month

	Austria	Austria*	Graz	Graz*
Count December	1,070	689	250	4
After filtering (20%)	214	/	50	/
Proportional December value for Styria (official data)	17.9	57.7	289.9	23.2
Proportional December value for Styria (Twitterscraper)	/	/	196.7	15.7

3.6 Sentiment analysis

On our final dataset of 6,953 tweets, a sentiment analysis was implemented within the open-source software from Orange for data mining. As a result of applied Vader algorithm, 5,208 tweets were determined as neutral, 646 as positive and 1,100 as negative. Sentiment was categorized according to the compound analysis result with more than 0 and up till 1 classified as positive, 0 as neutral and between less than 0 and -1 as negative sentiment. Due to the nature of our topic, 75% neutral tweets is an expected result, since many users do not reveal their feelings or attitude while posting where are they travelling or spending their free time. Further results related to sentiment analysis from a spatial perspective are explained in the spatial analysis section.

Table 3.6: Count of tweets per sentiment

Sentiment	Quantity
Positive	646
Negative	1,100
Neutral	5,208
Total	6,953

Table 3.7: Example tweets representing different sentiments

Sentiment	Text
Positive	- Good morning Graz. Thank you for being a great location for thoughtful talks, for a beautiful evening & morning and all together for a very special time. #Graz #Steiermark - back home from 3 days relaxing in "Therme Bad Waltersdorf"!! now im ready 4 the last weeks @ school till x-mas vacation XD xoxo ly all <33 - hotel Rogner Bad Blumau : Hiking holidays Austria Feel Free To Like Tag Share Beautiful Nature And...
Negative	- Traveling with 4 guitars for the first time since killed by 9v batteries disbanded! #meltdowner # killedby9vbatteries @ Graz, Austria - # Gemeinde # Pusterwald ist also die # Terrorhochburg der # Steiermark. - Death @ Stift Admont
Neutral	- Some late afternoon # planespotting @ Flughafen Graz-Thalerhof in Feldkirchen bei Graz, Steiermark) - Ich bin bei Hotel Zum Dom Graz (Graz, Styria) - Train ride through Austria on our way to Croatia! @ Gemeinde Neumarkt In Der Steiermark

Results of sentiment analysis are represented by a scatter plot and were examined by clustering using a k-means approach. Colors of incidents in the scatter plot show the compound sentiment value and are explained in the legend. Figure 3.16 displays an example of 7 clusters represented by a heatmap. Clusters were determined by allocating sentiments (positive, negative, neutral) and the compound value of every tweet to the nearest cluster, while forcing the centroids of clusters to remain as small as possible [122]. The biggest cluster of 5,228 (3) represents most of neutral classified tweets. Clusters 1 and 2 represent positive sentiments. In clusters 4, 5, 6 and 7, negative sentiment gradually increases with 4 being close to neutral and 7, close to the score of -1, as very negative.

Figure 3.16: Scatter plot of sentiment analysis

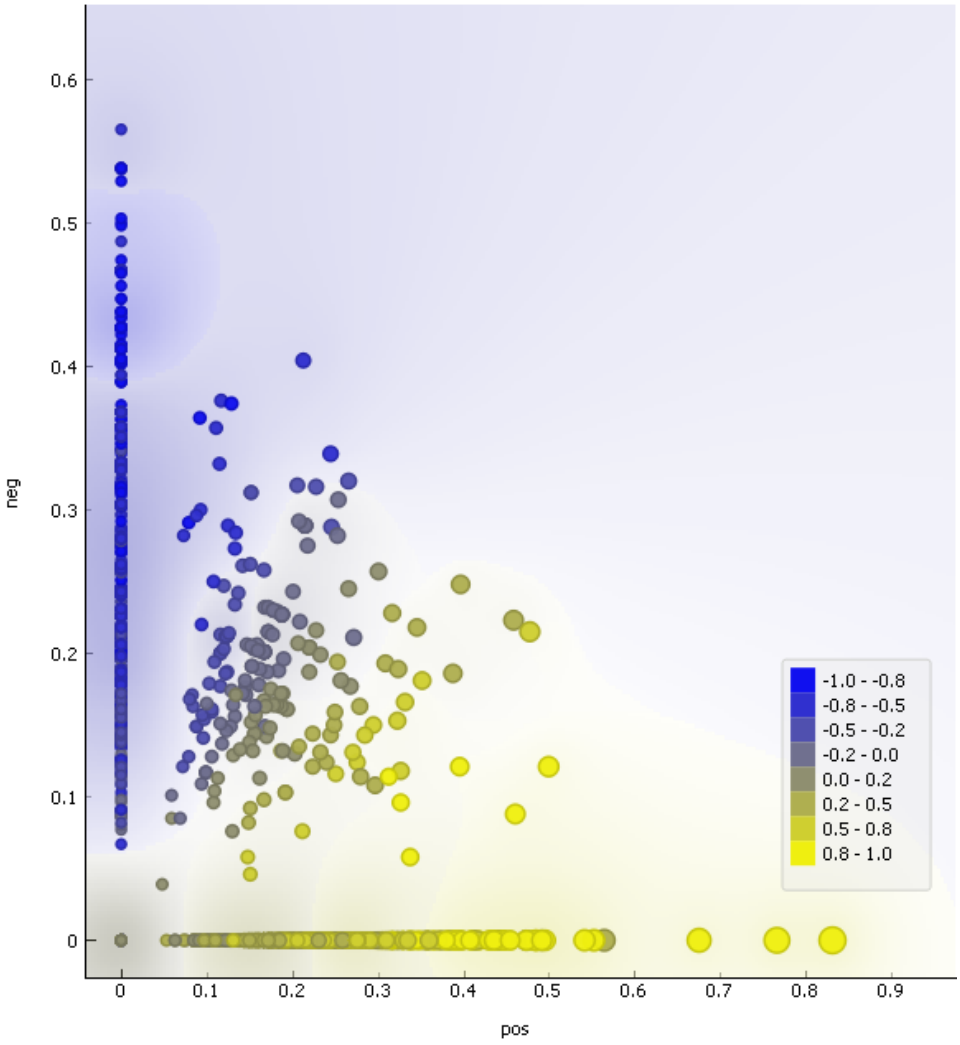
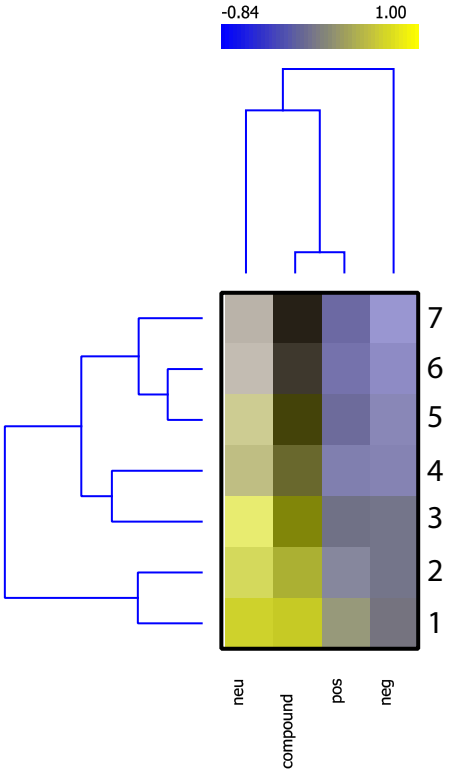


Table 3.8: Count of tweets in 7 clusters

Cluster	Quantity
1	286
2	330
3	5228
4	185
5	498
6	271
7	158

Figure 3.17: Sentiment clusters merged by k-means



3.7 Spatial analysis

The spatial analysis within this research focuses on the evaluation of spatial distribution and temporal patterns within the state of Styria as our target region. Our objectives are to statistically analyse the growth of the tweets in our dataset through the years and to compare them with regard to the seasons.

For a spatial analysis the data was exported in a .csv (comma separated values) format from the MongoDB database and was further processed and analysed in ArcGIS 10.5 software. In GIS, a .csv datasheet was imported as a shapefile according to the coordinates of the municipalities centroid. All the data corresponds to the WGS 1984 coordinate system. Imported centroids were then spatially joined with the shapefile of Styria. In this step, all the tweets of common municipalities had exactly the same coordinates - namely the municipality's centroid that was prescribed in the geocoding step within MongoDB. order to obtain a spatial image and be able to perform further spatial analyses at both municipality and state level, we used a tool to randomly distribute a corresponding number of the tweets within

each municipality. Based on this spatial distribution, further cluster-related analyses were also performed. The tweets' distribution accuracy within the municipality is therefore not absolutely correct, but this approach is a very good solution for relatively visualizing the tweets and obtaining answers of relative distribution within the whole state of Styria through further spatial analyses.

3.7.1 Spatial distribution

Figure 3.18 represents a spatial distribution of all tweets of our final dataset. 6,953 Tweets were determined as tourism-related messages and distributed in 165 municipalities in the years from 2008 and August 2018. In 80 municipalities there were already no tweets extracted via API. After filtering the total number of municipalities without any data increased to 122 municipalities. Further, 112 municipalities which are spread throughout the whole state can be categorized as areas with a very small tourism sector, since there are only 1-14 tweets. There are 15-29 tweets in a further 25 municipalities and 30-71 in 11 municipalities. The areas with the most tourists are 11 municipalities with 72-189 tweets and 5 municipalities with 190-420 tweets. These are located in the mountainous region in the north-western part of the state and in the south-eastern part rich with thermal water being exploited for tourist thermal resorts. Apart from these strong tourist areas, there are also other larger municipalities such as Leoben and Bruck an der Mur. More than one-third of all tweets (2,374) belongs to Styria's capital Graz.

Figure 3.18: Spatial distribution of all tweets after filtering

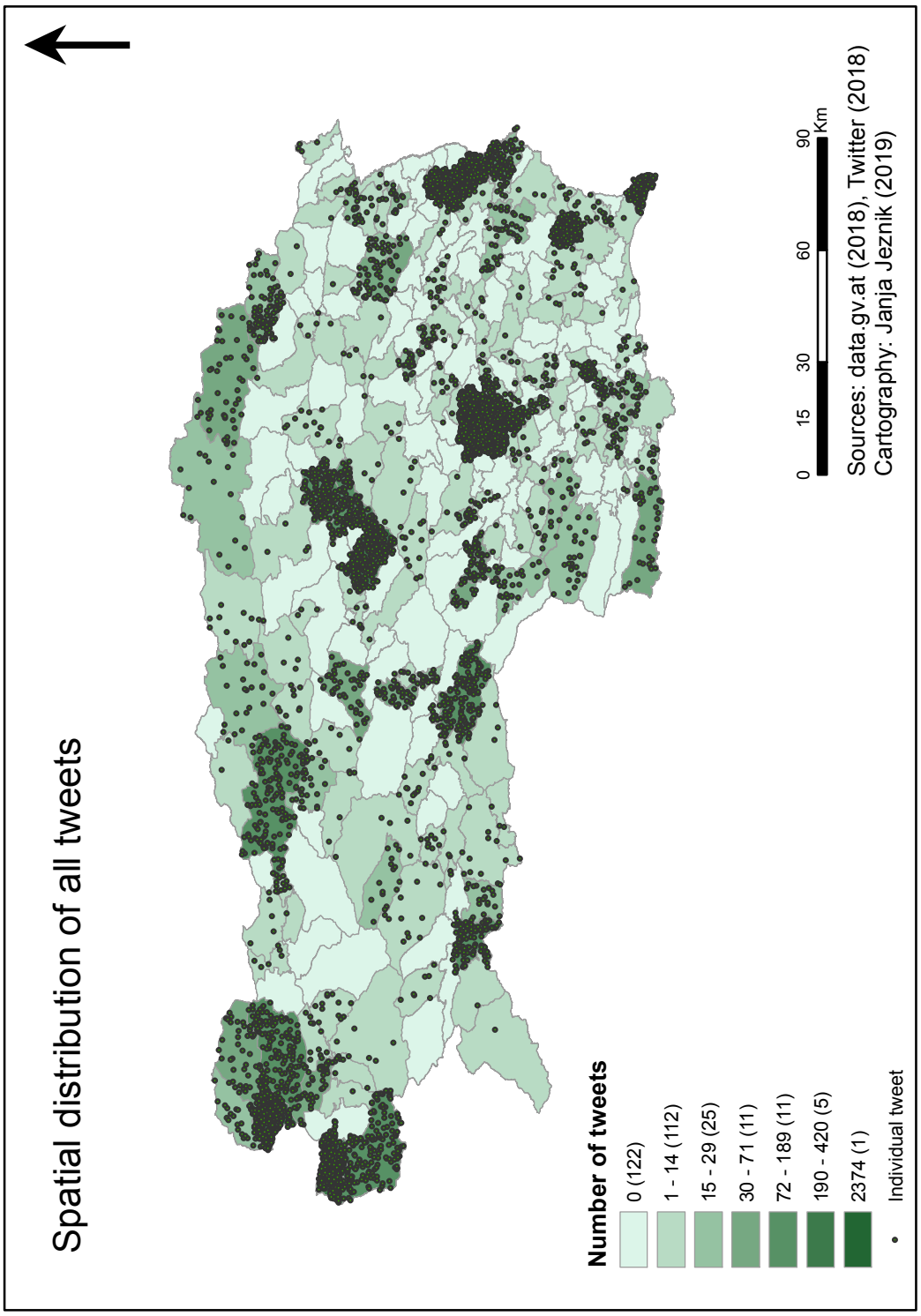


Table 3.9: Municipalities with over 100 tweets

Municipalities with over 100 tweets	Nb. of tweets
Graz	2,374
Feldkirchen bei Graz	420
Bad Aussee	274
Leoben	260
Bad Waltersdorf	235
Ramsau am Dachstein	230
Bad Blumau	189
Bad Gleichenberg	178
Schladming	162
Admont	158
Kapfenberg	154
Weißkirchen in Steiermark	144
Bad Mitterndorf	138
Bruck an der Mur	130
Bad Radkersburg	117
Fürstenfeld	113
Murau	104

3.7.2 Categories of extracted tweets

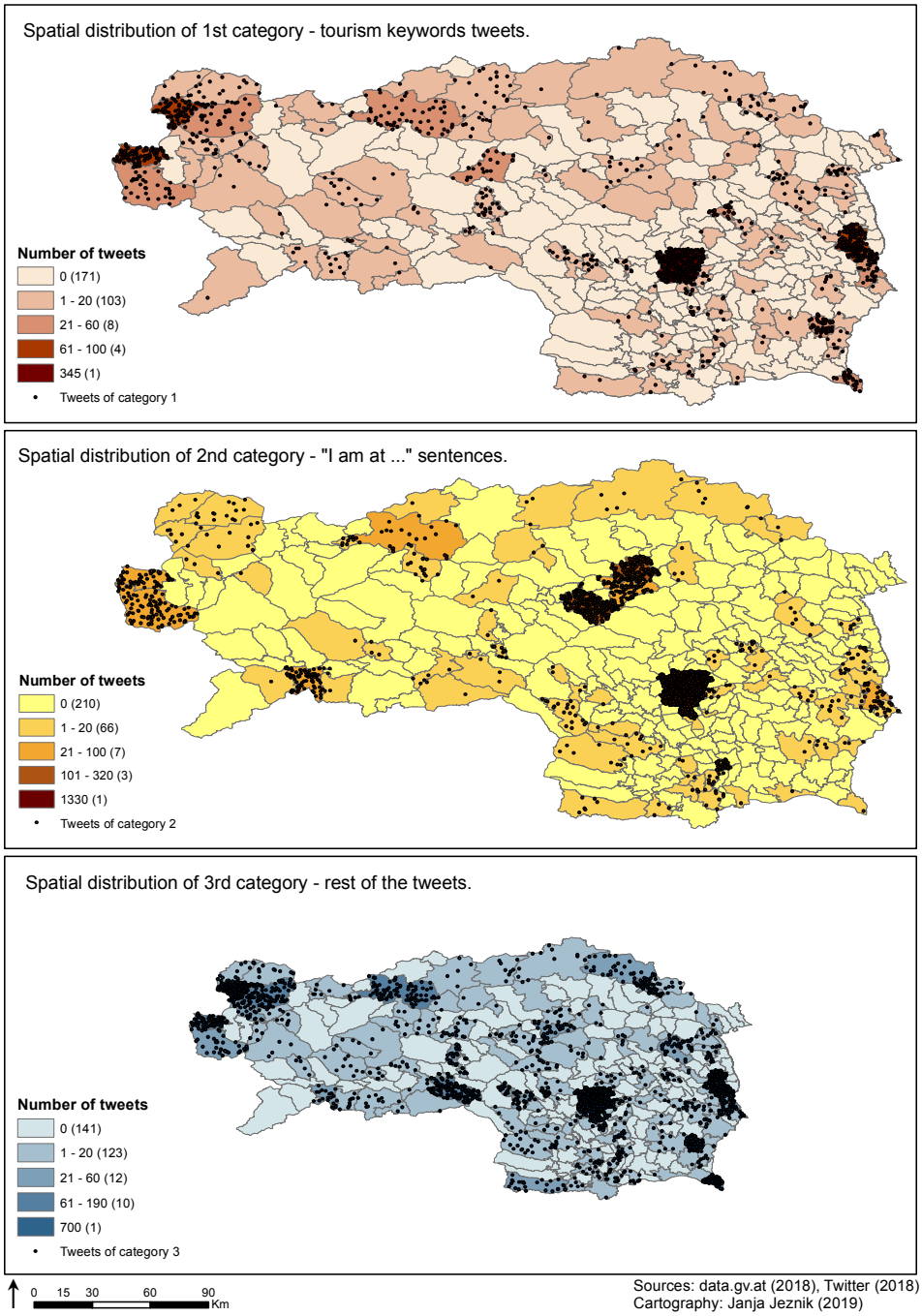
Figure 3.19 represents a spatial distribution within the 3 separate subcollections.

In the first map are shown 1,321 tweets extracted using our predefined tourist keywords. They cover the most traditionally strong tourist areas - the state's capital Graz and mountainous municipalities with lakes in the north-west on one side and thermal areas in the east on another side. The "I am at..." sentence revealing location is represented by 2,717 tweets which are more equally distributed within Styria and do not only focus on classical tourist regions but also on cities such as Leoben and Bruck an der Mur. The third and largest subcollection, with 2,854 extracted tweets, reveals the most equal distribution and covers most of the regions that are already represented in the previous two categories. This even distribution within the region is expected for this subcollection, since it consists of tweets of very heterogeneous tourism types. In contrast, the first subcollection covers most of the typical tourism regions since it contains the most typical tourism keywords.

Table 3.10: Top 10 municipalities of each subcollection

Subcollection 1	Count 1	Subcollection 2	Count 2	Subcollection 3	Count 3
Graz	345	Graz	1,330	Graz	699
Ramsau am D.	100	Feldkirchen	301	Bad Aussee	188
Bad Waltersdorf	100	Leoben	240	Wießkirchen	140
Bad Blumau	90	Kapfenberg	144	Bad Gleichenberg	121
Bad Aussee	80	Schladming	95	Bad Waltersdorf	117
Admont	53	Bruck an der Mur	86	Feldkirchen	144
Bad Gleichenberg	53	Murau	79	Bad Mitterndorf	96
Schladming	34	Ramsau am D.	50	Bad Blumau	94
Bad Mitterndorf	33	Gralla	40	Bad Radkersburg	90
Fürstenfeld	26	Fürstenfeld	28	Admont	82

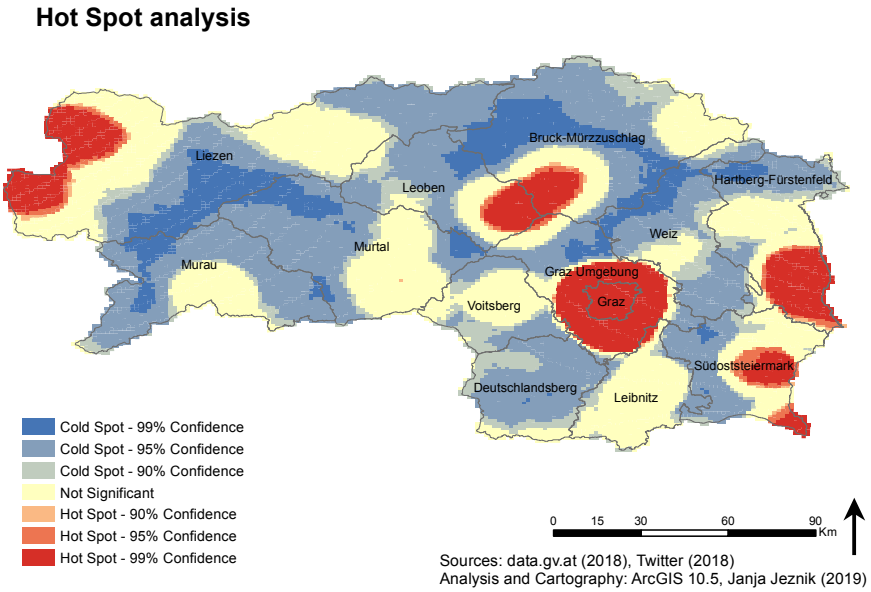
Figure 3.19: Spatial distribution of tweets within subcollections



3.7.3 Hot Spot analysis

We implemented a Hot Spot analysis with the tool Optimized Hot Spot Analysis within the ArcGIS. Analysis was implemented on a dataset with all tweets randomly distributed within a corresponding municipality. As an aggregation method, the incidents were counted within a fishnet of a cell size of one kilometre. A shapefile of the state of Styria was used as a dimension border for the analysis.

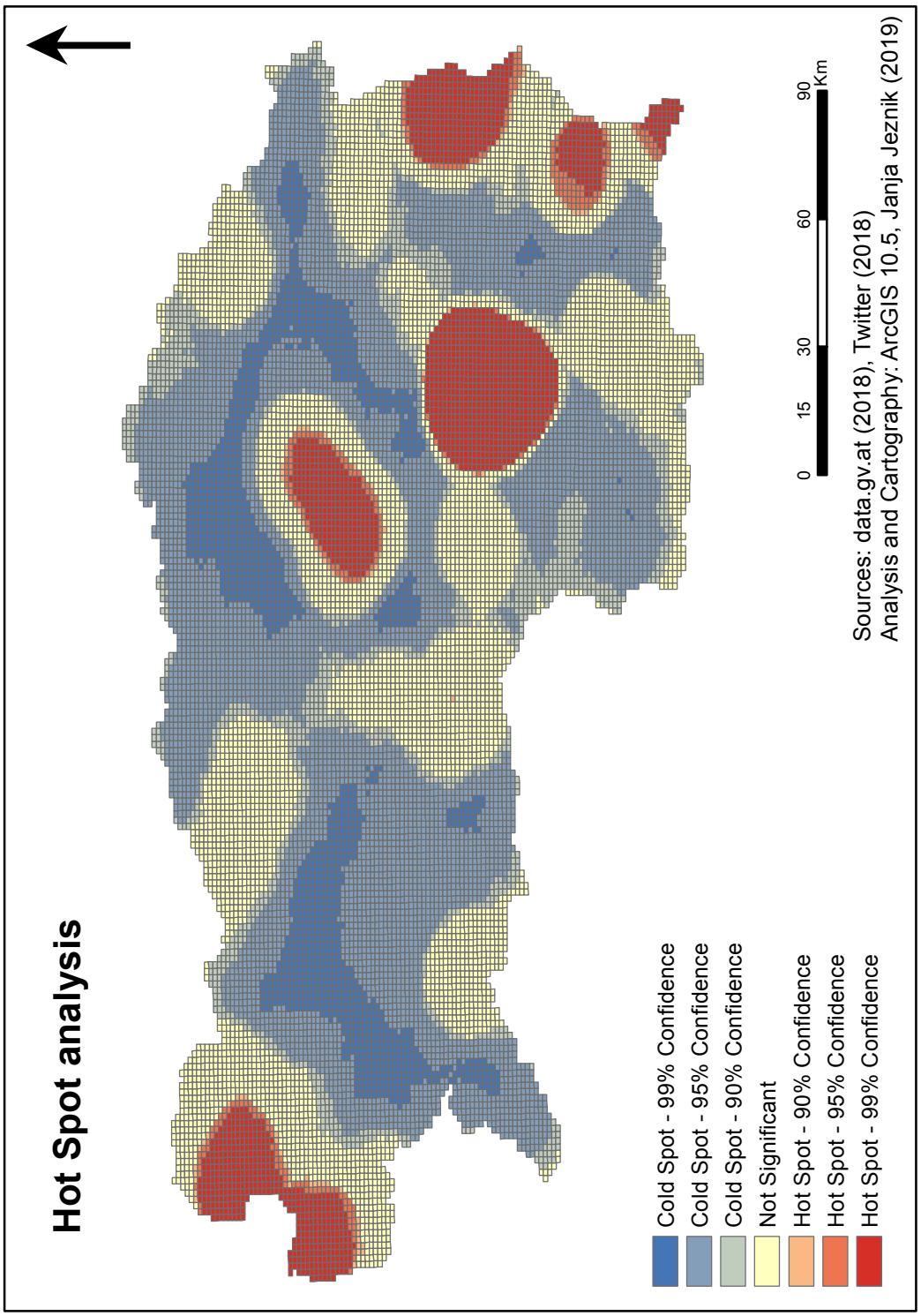
Figure 3.20: Optimized Hot Spot Analysis - districts



As a result there are six clear and evident hot spots with a confidence of 99%. As an obvious and expected result we can observe a hot spot of the state's capital city of Graz and its surroundings in a central part of southeastern Styria. Further city tourism-related hot spot is the urban agglomeration of the municipalities of Leoben, Bruck an der Mur and Kapfenberg. A hot spot in the shape of a semicircle in the northwest belongs to the tourism-intensive municipalities of Bad Aussee and Ramsau am Dachstein, offering a selection of nature- and mountains-related tourism activities, e.g. hiking, skiing or bathing in natural lakes. The last three significant hot spots with 99% confidence belong to the municipalities of health and beauty thermal tourism, namely Bad Radkersburg in the extreme south-east, Bad Gleichenberg further north and Loipersdorf bei Fürstenfeld in the east.

All not significant locations in our case refer to areas of moderate visits

Figure 3.21: Optimized Hot Spot Analysis - fishnet (1km)



with no significant spatial clusters. These not significant clusters include all hot and cold spots and cover the whole district of Leibnitz and Voitsberg, the southern part of the district of Murau and the south-eastern part of Murtal. Another cluster of moderate tourism is in the north in the district of Liezen in the area of Gesäuse National Park and the tourist-developed municipality of Admont. The clusters in the central part of the district of Hartberg-Fürstenfeld and east of Bruck-Mürzzuschlag are also not significant.

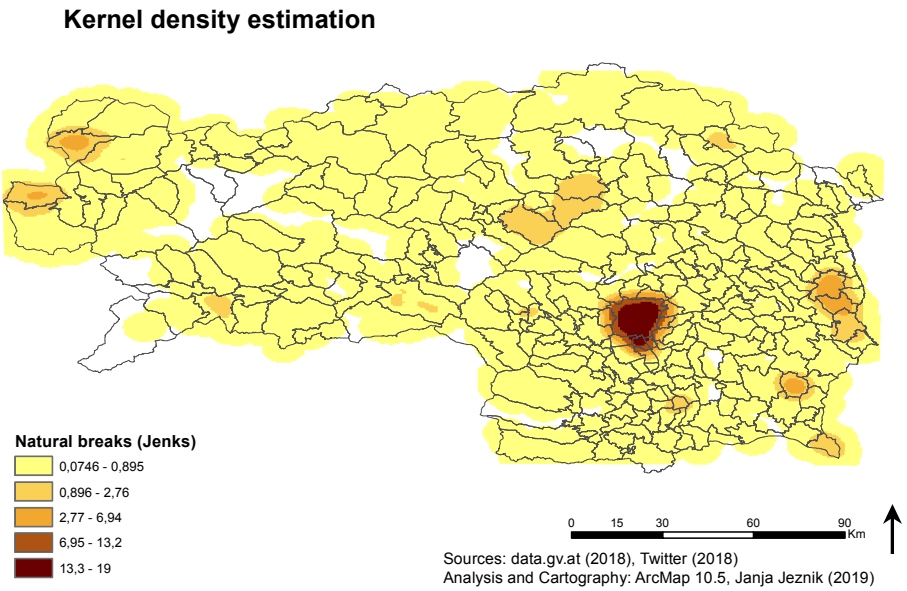
Statistically significant cold spots represent areas confronting low tourist visits and are, with the exception of the whole district of Deutschlandsberg located on the edges of more populated areas such as in the mountains or in the countryside. Cold spot clusters with 99% confidence are located in the western and northern mountainous parts of the state, mostly in the districts of Liezen and Bruck-Mürzzuschlag. However, significant cold spots are located also on the peripheral edges of all other districts (except Leibnitz).

3.7.4 Kernel density estimation

Kernel density estimation was applied with a search radius of 5 kilometres and an output cell size of 500 metres. Such search radius and output cell size were determined after analysing results of different settings ranging from a 1 to 10 kilometre search radius and a 100 up to 5,000 meter output cell size. Selected values were determined as suitable because these settings enable the creation of a convincing visualization with meaningful values. Due to big differences in resulting densities, we used a Jenks Natural Breaks classification method, which arranges values into different classes, minimizing the difference within a class and maximizing it between classes [123]. The difference within a class is minimized by minimizing the average deviation of the class from the mean of the class and maximizing the deviation of the class from the means of other groups. Uneven breaks and unusual class boundaries are typical in order to separate values of large changes in values. However, maps visualized with this method cannot be compared. Also, the number of classes may significantly affect the final result and have to be chosen with consideration [124].

Our results with Kernel Density Estimation confirm large changes in values, especially between the state's capital Graz and other areas. Most of the region corresponds to a density of under 1 tweet per search radius. White values are locations with the kernel density of 0 and were excluded from the visualization in order to stress the difference to other parts.

Figure 3.22: Kernel density estimation



3.7.5 Sentiment analysis - spatial patterns

Spatial patterns of positive, negative and neutral sentiments are represented in the maps in Figures 3.23 and 3.24. However, it is important to take a notice, that there are some municipalities with the word "Bad" in their name, such as Bad Radkersburg or Bad Aussee. In these cases, although the user was referring to the name of the municipality, the tweet was categorised as negative due to the word "Bad" which refers to thermal or swimming areas in German, but is understood by the algorithm as the English word referring to something negative. As in Figure 3.24, 6 municipalities are marked with a cross in order to point out their irrelevance to the category of very high negative sentiment percentage. There are some municipalities with a higher negative proportion of negative sentiments, due to a lower tweet count in the first place.

3.7.6 Temporal analysis

Temporal analysis was applied from a spatial-temporal aspect in order to investigate the growth of tweets from 2008 to 2017 at a district level. It was also applied as part of the evaluation process, where tweets were com-

Figure 3.23: Spatial distribution of tweets within subcollections

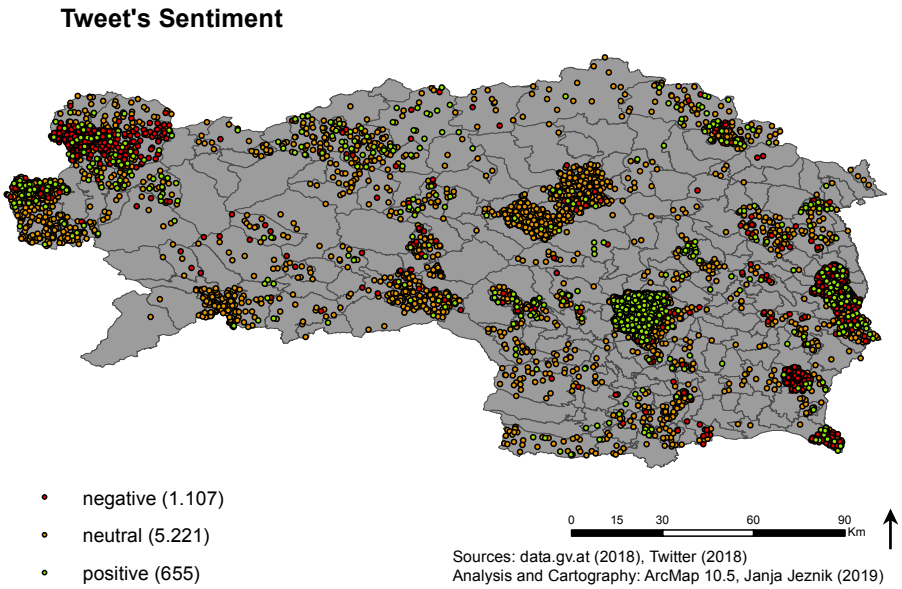
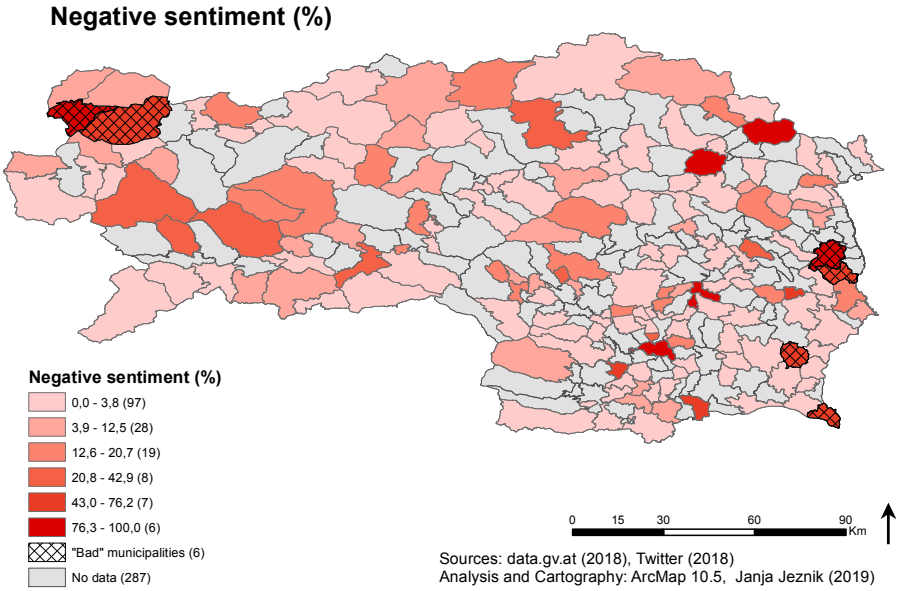
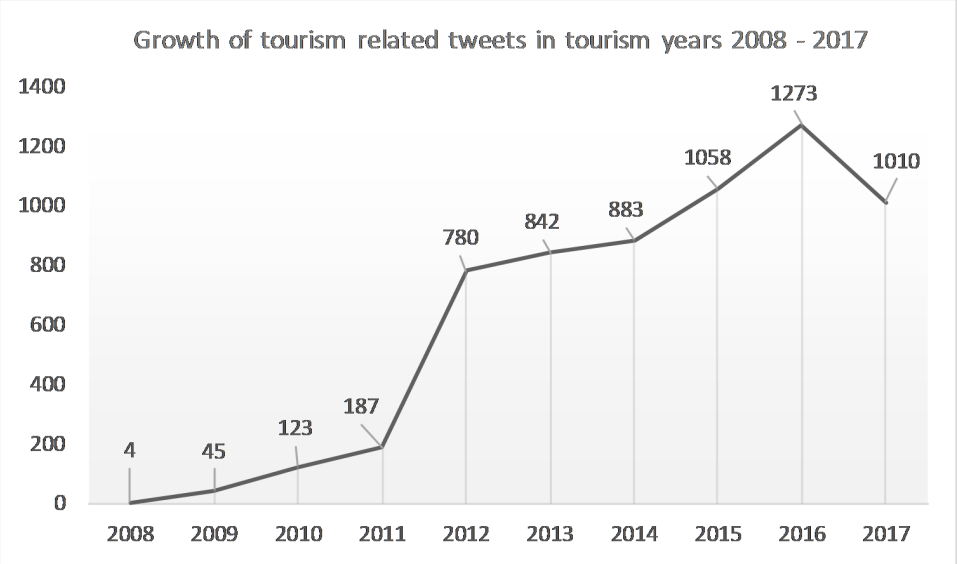


Figure 3.24: Spatial distribution of tweets within subcollections



pared through the years with our reference data. The years applied do not match with the calendar year, but with tourism year which, according to our reference data, is from 1 November till 31 October the following year. tweets come with an exact date and time and in order to reach comparability with our reference data we distributed all the tweets into the corresponding tourism year.

Figure 3.25: Growth of tourism-related tweets in tourism years 2008 - 2017



A growth of tweets is experienced across the observed years. Although we also extracted tweets of first half of 2018 they are not included in our temporal analysis due to their incomparability. As can be seen in Figure 3.26, the count of extracted tweets is growing steadily and is most markable in the districts of Graz and Liezen. The years 2016 and 2017 are the strongest in all the districts. However, in the first three years, 2008, 2009 and 2010, there is a negligible number of acquired tweets. This could be explained by a notably lower number of active Twitter users in its first years from establishment in 2006, as represented in Figure 3.27. Furthermore, a possibility to add location to the tweets only started in March 2010. Hence, the tweets within our dataset before that date could only be extracted due to the mentioned municipality or state name. From 2011 onwards, the count of our tweets is growing (as is the general number of tweets per day around the globe) and as such, the suitability for deeper investigation and research increases. Interestingly, from 2017, Twitter has struggled to maintain the growth of new users at prior level. This is seen in graph 2.1 displayed at the beginning of the Chapter 2 of theoretical principles. However, even though the count of extracted tweets also decreased after the peak in 2016, it is too early for any conclusions.

Figure 3.26: Growth of tourism-related tweets at a district level

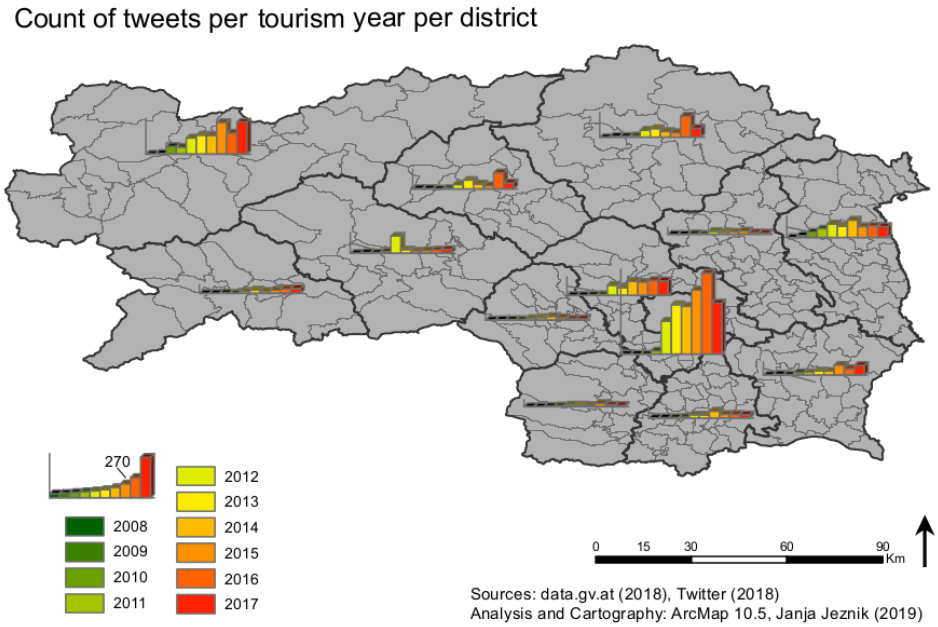
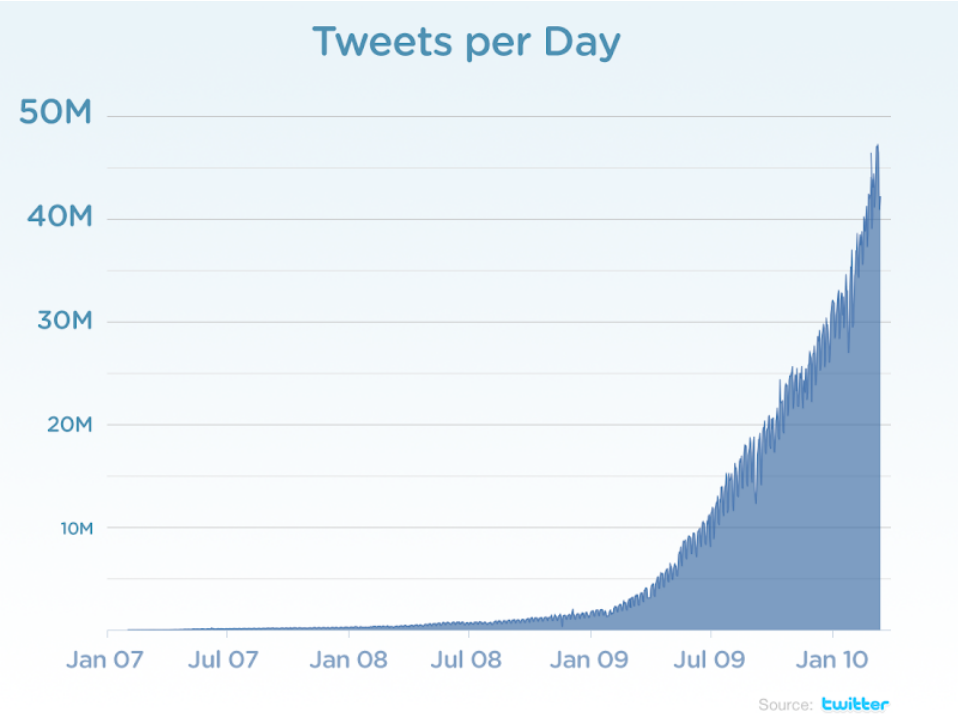


Figure 3.27: Tweets per day in Twitter's first years [125]



Seasonal analysis:

Apart from general temporal analysis, we also investigated seasonal distribution. As in our reference data, we used the summer and winter tourism season. Tweets from 1 May till 31 October are assigned to the summer season, while tweets from from 1 November till 30 April of the following year to the winter season.

In a seasonal distribution there is an obvious pattern of prevailing tweets in the warm part of the year through all the years. Figure 3.38 represents the count of tweets per year throughout Styria, while Figure 3.29 shows the distribution within the districts. For comparison purposes, Figure 3.30 contains reference data. Correlations between them are represented in the evaluation section.

Figure 3.28: Count of tweets per tourism season 2008 - 2017

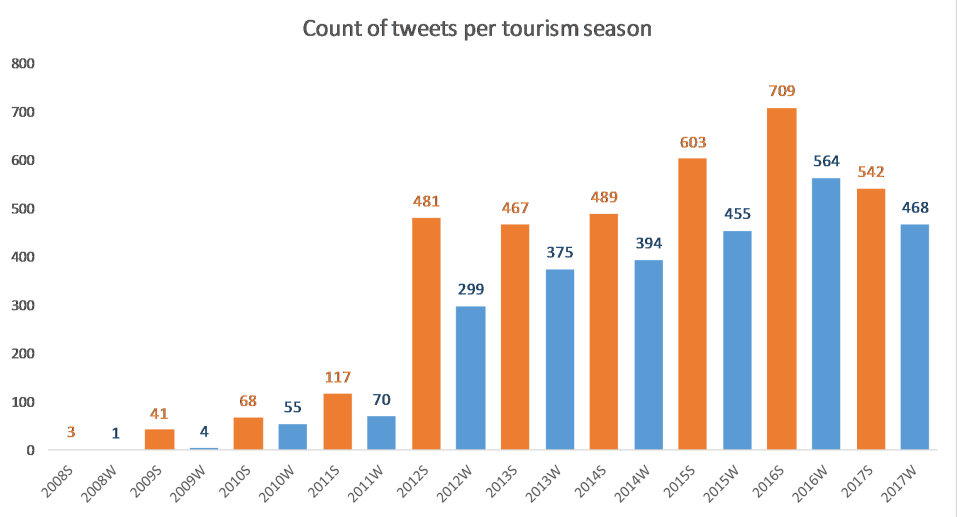


Figure 3.29: Count of tweets per tourism season at a district level 2008 - 2017

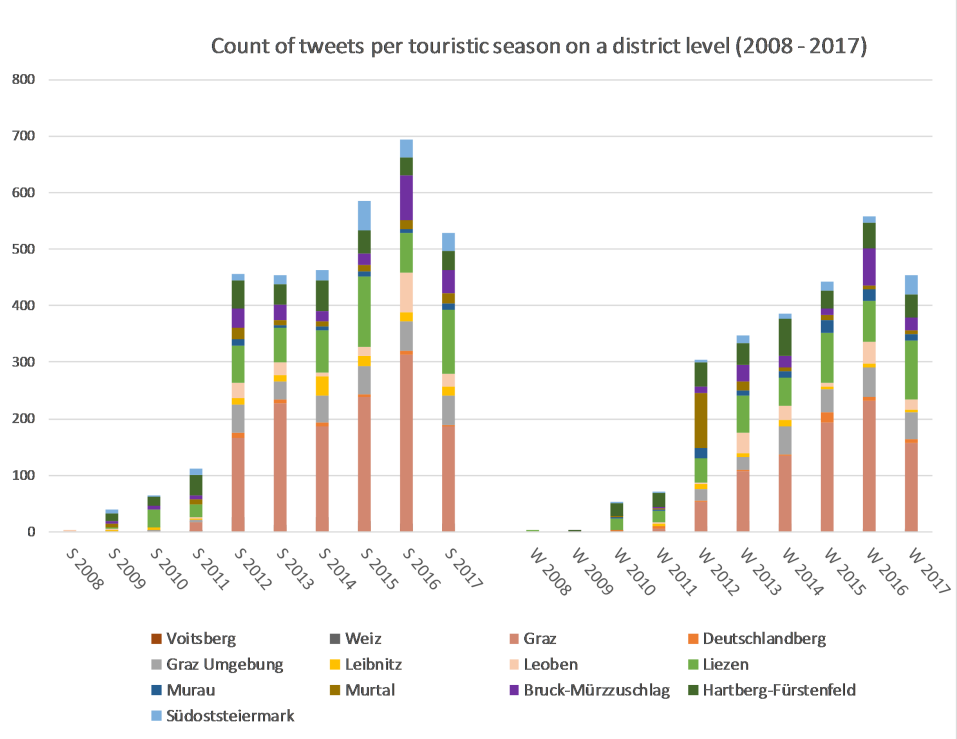
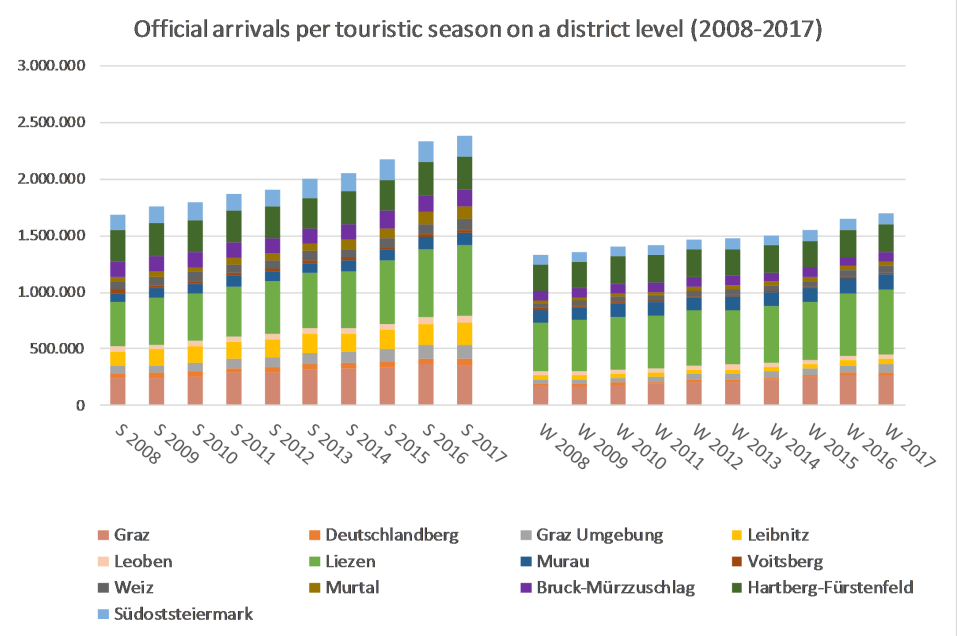


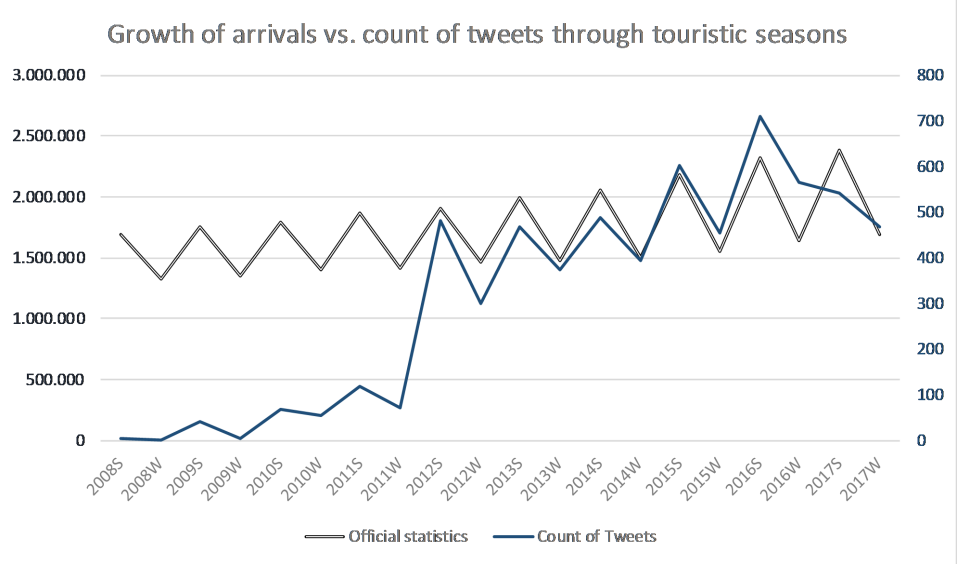
Figure 3.30: Count of offical tourist arrivals per tourism season 2008 - 2017



3.8 Evaluation

Evaluation of the entire analysis is based on a comparison between a count of all the tweets extracted through the Twitterscraper package and our reference data. Extracted data was preprocessed and filtered from tourism-irrelevant tweets. Our reference data was collected from Austrian official authorities for statistics. The evaluation approach was based on correlation determination between our data and official data, representing two scale variables. Although our graphs already visually demonstrate the relationship between growth in the collected data and the reference data (Figure 3.31), the correlation coefficient needs to be determined to confirm whether the correlation is statistically significant or insignificant. Applying in statistical programme SPSS, we used Pearson's coefficient to determine the level of correlation of data from 2008 till 2017, at both a state and a district level. The seasonal aspect was also taken into consideration.

Figure 3.31: Growth of arrivals and count of tweets throughout tourism seasons 2008 - 2017



As can be seen in Figure 3.31, there is an especially strong relationship between our data and reference data in the years from 2011 onwards. In the first years aafter the launch of the Twitter portal, the number of daily and new users was in the first phase of growing, which is also clearly visible in our dataset. In order to achieve relevant results, we applied correlation analysis to two separate datasets - one to our whole dataset from 2008 till 2018, whereas in 2018 only the data of the first half year (and also only first half year of reference data) was used. There is a significant correlation at the 0.05 level with Pearson's coefficient of 0.650. The Sig(2-tailed) value

under 0.05 confirms statistical significance of our correlation. On the other hand, taking into account only the full years between 2011 and 2017, the Pearson's coefficient increases to 0.772, which, according to the definitions in our theoretical section, is already a strong correlation. The correlation for years between 2011 and 2017 is also significant at the 0.05 level.

Figure 3.32: Scatter plot of official and Twitter data distribution 2008 - 2018

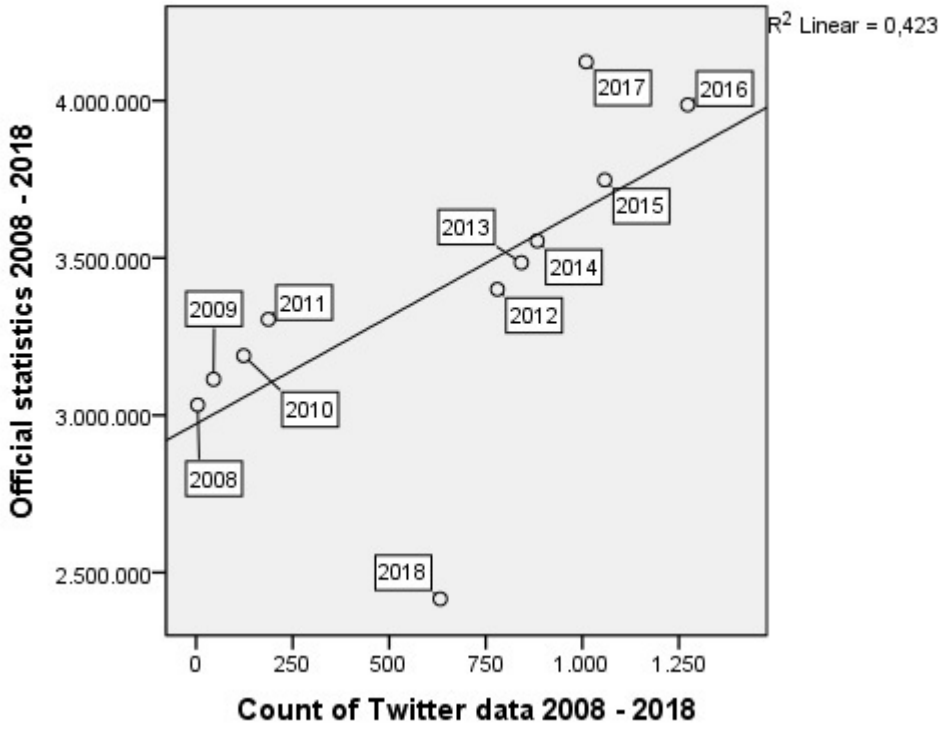


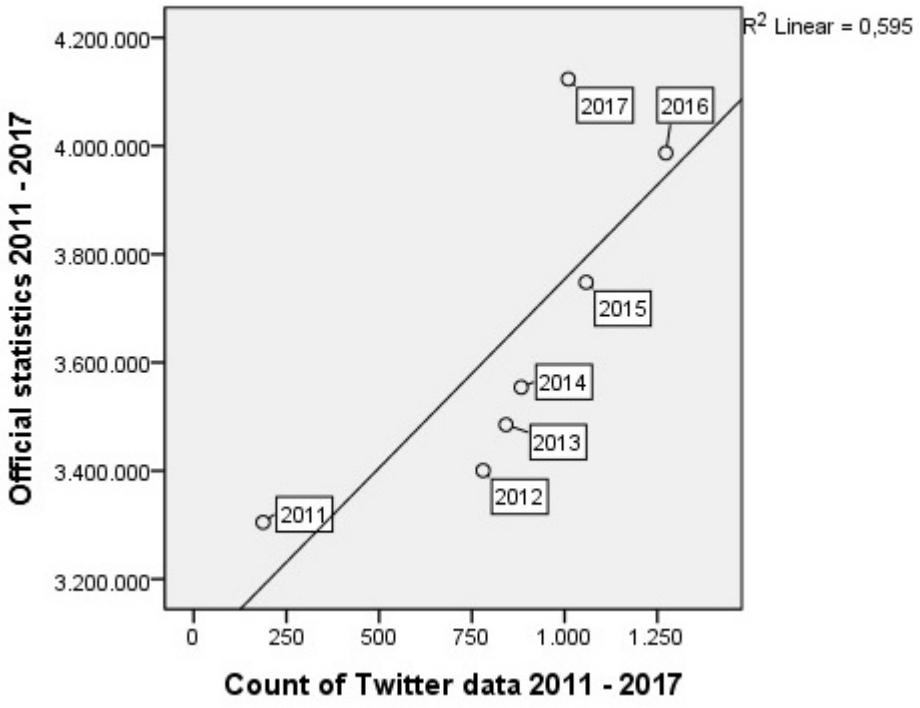
Figure 3.33: Correlation of the count of official and Twitter data between tourism years 2008 and 2018

Correlation - tourism years 2008 - 2018

		Official statistics	Twitter data
Official Statistics	Pearson Correlation	1	,650*
	Sig. (2-tailed)		,030
	N	11	11
Twitter data	Pearson Correlation	,650*	1
	Sig. (2-tailed)	,030	
	N	11	11

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 3.34: Scatter plot of official and Twitter data distribution 2011 - 2017



Further correlations were also determined at the district level for the time period from 2011 - 2017, since we believe that the low count of tweets in the years 2008, 2009 and 2010 does not reflect such low tourist visitors but

Figure 3.35: Correlation of the count of official and Twitter data between tourism years 2011 and 2017.

Correlation - tourism years 2011 - 2017

		Official statistics	Twitter data
Official statistics	Pearson Correlation	1	,772*
	Sig. (2-tailed)		,042
	N	7	7
Twitter data	Pearson Correlation	,772*	1
	Sig. (2-tailed)	,042	
	N	7	7

*. Correlation is significant at the 0.05 level (2-tailed).

rather the low number of Twitter users in general. As can be seen in Figure 3.36, there are 5 districts with positive correlation over 0.8 and are statistically significant at the 0.001 level. Murau, Liezen and Graz Umgebung have a strong correlation over 0.85 and Graz and Südoststeiermark even a very strong correlation over 0.9. With 0.694 there is also Leoben with significant correlation on 0.001 level. However, at half of the districts there is no significant correlation between our and reference data. One district, Voitsberg, even confronts negative correlation at 0.005 significance level.

Table 3.11 shows the Pearson's coefficients for the summer and winter seasons. Districts are arranged according to the correlation strength in the summer season. It can be seen that correlations are, in general, higher in the summer compared to the winter. Graz, Südöststeiermark, Graz Umgebung and Liezen are on the top with the highest values for both seasons. The other districts fluctuate - for instance in Murau there is a significant positive correlation in summer but a very weak negative relationship with no significant correlation in winter. Also in Bruck-Mürzzuschlag, Deutschlandberg, Hartberg-Fürstenfeld and Voitsberg there are positive relationships in one season and negative relationships in another season.

Figure 3.36: Correlation of the count of official and Twitter data between tourism years 2011 and 2017 - at the district level

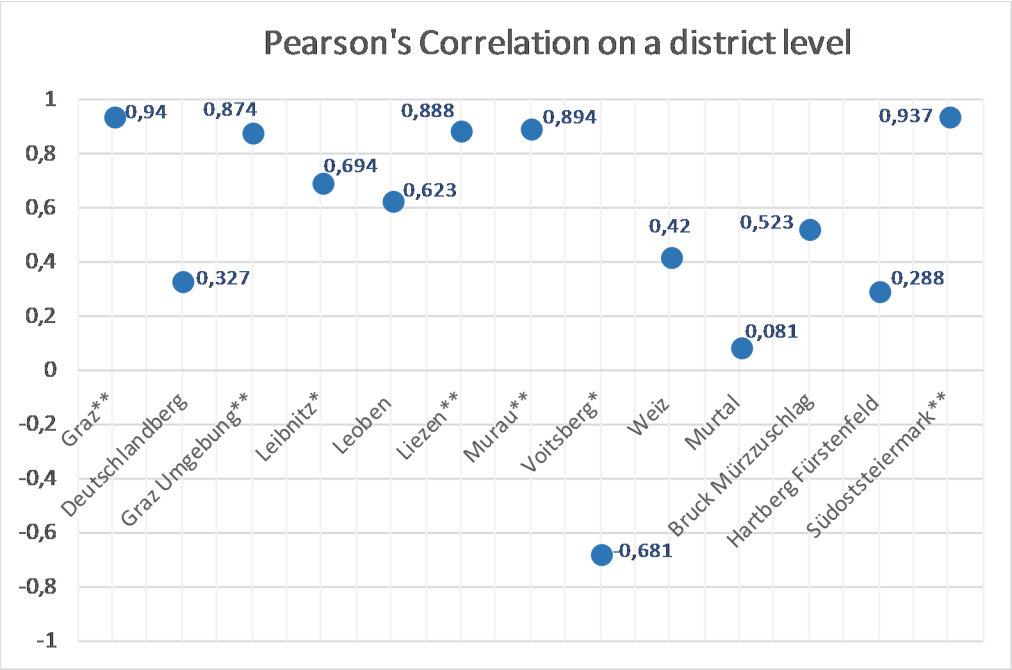


Table 3.11: Correlations at a district level - seasons

District	Pearson's coeff. Summer	Pearson's coeff. Winter
Graz	0,916**	0,948**
Südoststeiermark	0,892**	0,790**
Graz Umgebung	0,835**	0,890**
Liezen	0,839**	0,913**
Murau	0,736*	0,557
Murtal	0,720*	-0,168
Leibnitz	0,704*	0,599
Leoben	0,683*	0,582
Bruck-Mürzzuschlag	0,682*	-0,715*
Weiz	0,368	0,246
Deutschlandberg	0,310	-0,280
Hartberg-Fürstenfeld	-0,258	0,681*
Voitsberg	-0,564	-0,422

Chapter 4

Conclusion

The evaluation revealed statistically significant correlations between our data with tourism-related tweets and reference data of official statistics on tourism in Styria. Due to the low count of Twitter users in its first years there is also a neglibible amount of tweets between 2008 and 2010 in our dataset. Thus, if only the dataset from 2011 onwards is considered, the correlation would be even stronger. At a district level there are significant correlations in half the districts. The same applies to the seasonal evaluation, where half of the districts relate better to the reference data in summer than in winter. This proves that the spatial scale of a state is still an adequate one, while results at a district level do not look that promising.

Yet, our dataset has proved to have a very similar spatial distribution of tourists within the region compared to the official statistics. Spatial clusters of Twitter data cover the traditional tourist-strong municipalities in the north-west (e.g. Bad Aussee and Ramsau am Dachstein), the thermal municipalities in the south-east (e.g. Bad Radkersburg, Bad Gleichenberg, Loipersdorf) and also the cities of Graz, Leoben and Bruck an der Mur. A sentiment analysis also revealed a very important prospect for the use of Twitter data in tourism.

It is important to stress that filtering is a crucial part of each research that deals with Twitter data. Only clean datasets of tweets with high relevance to a particular topic can yield solid, objective and resistant results. The filtering approach of our research was a combination of automatic and manual determination of tourism-related tweets according to their text content. Developing a complete machine learning approach for natural language processing for extraction of tourism-related tweets would definitely be the next step in order to reach objective results of high quality.

Twitter is a modern platform offering voluntary geographic information for use in many aspects. As in other fields, it is also crucial in tourism to track user's opinions in order to offer appropriate responses and supply on tourist demand within the region. Text mining approaches for sentiment analysis applied on Twitter data are therefore of great importance, since text written in social media is very specific and is written without any general rules but

rather in an everyday language including abbreviations, typing mistakes, hashtags and emotion tags. Also, due to multilingual datasets, mistakes can arise due to inbetween translation steps, as demonstrated in the dataset of municipalities with the word "Bad" in their name being defined as having a negative sentiment. Improvements in analyses with multilingual data would therefore be of great success.

However, we only dealt with a very small dataset of 10 years. In the first 3 years there was a lack of Twitter users and consequently a lack of data. Hence, our findings and conclusions have to be taken with caution and an awareness that further data collection and research on a larger dataset should take place in order to come to more solid conclusions. Still, our research findings may be accepted as a very good introduction to the use of Twitter as a alternative for statistical data acquisition for tourism purposes and as a very adequate mean for opinion mining.

Bibliography

[1] Encyclopaedia Britannica. (Feb. 10, 2019). The histroy of Twitter, [Online]. Available: https://www.britannica.com/topic/Twitter.

[2] Statistik Austria. (Feb. 10, 2019). Tourismus, [Online]. Available: https://www.statistik.at/web_de/statistiken/wirtschaft/tourismus/index.html.

[3] Das Land Steiermark. (Feb. 10, 2019). Landesstatistik Steiermark, [Online]. Available: http://www.landesentwicklung.steiermark.at/cms/ziel/141976103/DE/.

[4] Land Steiermark. (Feb. 7, 2019). Allgemeines über Steiermark, [Online]. Available: http://www.europa.steiermark.at/cms/beitrag/10139601/3709911/.

[5] S. B. Avraham. (Jan. 15, 2019). What is REST — A Simple Explanation, [Online]. Available: https://medium.com/extend/what-is-rest-a-simple-explanation-for-beginners-part-1-introduction-b4a072f8740f.

[6] A. Taspinar. (Jul. 10, 2018). Twitterscraper, [Online]. Available: https://github.com/taspinar/twitterscraper.

[7] MongoDB. (Jan. 28, 2019). NoSQL explained, [Online]. Available: https://www.mongodb.com/nosql-explained.

[8] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in Eighth international AAAI conference on weblogs and social media, 2014.

[9] Das Land Steiermark. (Feb. 17, 2019). Gemeindestrukturreform, [Online]. Available: http://www.gemeindestrukturreform.steiermark.at/.

[10] Statistik Austria, Ed., Organisation und Ablauf der Österreichischen Beherbergungsstatistik - Ein Leitfaden für Berichtsgemeinden. Wien, Österreich: Statistik Austria, 2018, 5th ed.

[11] Z. C. Steinert-Threlkeld, Ed., Twitter as Data. Los Angeles: University of California, Cambridge University Press, 2018. [Online]. Available: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/27B3DE20C22E12E162BFB173C5EB2592/9781108438339AR.pdf/twitter_as_data.pdf.

[12] F. Pfaffenberger, Ed., Twitter als Basis wissenschaftlicher Studien. Nürnberg, Deutschland: Springer, 2016.

[13] C. M. Hall, "Spatial analysis: A critical tool for tourism geographies," THE ROUTLEDGE HANDBOOK OF TOURISM GEOGRAPHIES, pp. 163–173, 2012. [Online]. Available: http://www.gbv.de/dms/goettingen/645266957.pdf.

[14] W. Claster, P. Pardo, M. Cooper, and K. Tajeddini, "Tourism, travel and tweets: Algorithmic text analysis methodologies in tourism," Middle East Journal of Management, vol. 1, no. 1, pp. 81–99, 2013.

[15] A. Kovacs-Gyori, P. Cabrera-Barona, A. Ristea, C. Havas, and B. Resch, "# london2012: Towards citizen-contributed urban planning through sentiment analysis of twitter data," Urban Planning, vol. 3, no. 1, pp. 75–99, 2018.

[16] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti, "Geo-located twitter as proxy for global mobility patterns," Cartography and Geographic Information Science, vol. 41, no. 3, pp. 260–271, 2014.

[17] H. H. Hübl F. Cvetojevic S. and P. G., "Analyzing refugee migration patterns using geo-tagged tweets," ISPRS International Journal of Geo-Information, vol. 6, no. 10, pp. 302–325, 2017.

[18] E. Zagheni, V. R. K. Garimella, I. Weber, et al., "Inferring international and internal migration patterns from twitter data," in Proceedings of the 23rd International Conference on World Wide Web, ACM, 2014, pp. 439–444.

[19] S. T. Senaratne H. Bröoering A and L. D., "Moving on twitter: Using episodic hotspot and drift analysis to detect and characterise spatial trajectories," LBSN@SIGSPATIAL/GIS, 2014.

[20] C. A. Cassa, R. Chunara, K. Mandl, and J. S. Brownstein, "Twitter as a sentinel in emergency situations: Lessons from the boston marathon explosions," PLoS currents, vol. 5, 2013.

[21] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in Proceedings of the 19th international conference on World wide web, ACM, 2010, pp. 851–860.

[22] M. G. Beiró, A. Panisson, M. Tizzoni, and C. Cattuto, "Predicting human mobility through the assimilation of social media traces into mobility models," EPJ Data Science, vol. 5, no. 1, p. 30, 2016.

[23] A. Bassolas, M. Lenormand, A. Tugores, B. Gonçalves, and J. J. Ramasco, "Touristic site attractiveness seen through twitter," EPJ Data Science, vol. 5, no. 1, p. 12, 2016.

[24] B. Han, P. Cook, and T. Baldwin, "Text-based twitter user geolocation prediction," Journal of Artificial Intelligence Research, vol. 49, pp. 451–500, 2014. [Online]. Available: https://doi.org/10.1140/epjds/s13688-016-0092-2.

[25] S. F. Khan, N. Bergmann, R. Jurdak, B. Kusy, and M. Cameron, "Mobility in cities: Comparative analysis of mobility models using geo-tagged tweets in australia," in 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(, IEEE, 2017, pp. 816–822.

[26] Socialbakers. (Jun. 21, 2019). Twitter statistics directory, [Online]. Available: https://www.socialbakers.com/statistics/twitter/.

[27] F. Richter. (Feb. 10, 2019). Twitter fails to reignite user growth, [Online]. Available: https://www.statista.com/chart/10460/twitter-user-growth/.

[28] Twitter. (May 30, 2018). About Twitter, [Online]. Available: https://about.twitter.com/.

[29] M. F. Goodchild, "Citizens as senzors: The world of volunteered geography," GeoJournal, vol. 69, no. 4, pp. 211–221, 2007.

[30] M. Craglia, F. Ostermann, and L. Spinsati, "Digital earth from vision to practice: Making sense of citizen-generated content," International Journal of Digital Earth, vol. 5, no. 5, pp. 398–416, 2012.

[31] English Oxford dictionary. (Feb. 10, 2019). Geotag, [Online]. Available: https://en.oxforddictionaries.com/definition/geotag.

[32] V. Gupta and H. Rattikorn, "Harnessing of power of hashtags in tweet analystics," EEE Big Data Conference 2017, 2017.

[33] Techopedia. (Jan. 24, 2019). Big Data, [Online]. Available: https://www.techopedia.com/definition/27745/big-data.

[34] S. Shilpi and M. Taneja, "Big data and twitter," International journal of research in computer applications and robotics, vol. 2, no. 5, pp. 144–150, 2014.

[35] M. F. Goodchild, "Gis in the era of big data," Cybergeo : European Journal of Geography, 2016. [Online]. Available: https://journals.openedition.org/cybergeo/27647?lang=en.

[36] A. Gupta and P. Kumaraguru, "Credibility ranking of tweets during high impact events," International Journal of Geographical Information Science, pp. 2–8, 2012.

[37] D. Sui, "Mashup and the spirit of gis and geography," GeoWorld, vol. 12, pp. 15–17, 2009.

[38] J. Littman. (May 30, 2018). Where to get twitter data for academic research, [Online]. Available: https://gwu-libraries.github.io/sfm-ui/posts/2017-09-14-twitter-data.

[39] OXFORD Dictionaries. (Jan. 15, 2019). API, [Online]. Available: https://en.oxforddictionaries.com/definition/api.

[40] Tweepy Documentation. (Jan. 15, 2019). API, [Online]. Available: http://docs.tweepy.org/en/3.7.0/getting_started.html.

[41] ——, (2009). Tweepy, [Online]. Available: https://tweepy.readthedocs.io/en/3.7.0/.

[42] Twitter. (May 30, 2018). Tweet geospatial metadata, [Online]. Available: https://developer.twitter.com/en/docs/tutorials/tweet-geo-metadata.

[43] Encyclopaedia Britannica. (Jan. 25, 2019). Data processing, [Online]. Available: https://www.britannica.com/technology/data-processing.

[44] Collins English Dictionary. (Jan. 25, 2019). Data processing, [Online]. Available: https://www.collinsdictionary.com/dictionary/english/data-processing.

[45] P. Tank. (Feb. 11, 2019). Data processing, [Online]. Available: https://planningtank.com/computer-applications/data-processing.

[46] C. Ordonez and J. Garcia-Garcia, "Managing big data analytics workflows with a database system," 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), 2016. [Online]. Available: https://ieeexplore.ieee.org/document/7515752.

[47] Xplenty. (Jan. 28, 2019). The SQL vs NoSQL difference. MySQL vs MongoDB, [Online]. Available: https://medium.com/xplenty-blog/the-sql-vs-nosql-difference-mysql-vs-mongodb-32c9980e67b2.

[48] M. Rouse and J. Denman. (Jan. 28, 2019). Sharding, [Online]. Available: https://searchoracle.techtarget.com/definition/sharding.

[49] MongoDB. (Feb. 11, 2019). Release notes, [Online]. Available: https://docs.mongodb.com/manual/release-notes/.

[50] Techtarget. (Jan. 26, 2019). MongoDB, [Online]. Available: https://searchdatamanagement.techtarget.com/definition/MongoDB.

[51] MongoDB. (Jan. 26, 2019). MongoDB architecture, [Online]. Available: https://www.mongodb.com/mongodb-architecture.

[52] ——, (Jan. 26, 2019). MongoDB Compass, [Online]. Available: https://www.mongodb.com/products/compass/.

[53] Spring. (Jan. 25, 2019). Understanding JSON, [Online]. Available: https://spring.io/understanding/JSON.

[54] D. Crockford. (Jan. 25, 2019). Introducing JSON, [Online]. Available: https://www.json.org/.

[55] MongoDB. (Jan. 25, 2019). JSON and BSON, [Online]. Available: https://www.mongodb.com/json-and-bson.

[56] ESRI. (Jan. 31, 2019). What is GIS, [Online]. Available: https://www.esri.com/en-us/what-is-gis/overview.

[57] ——, (Jan. 31, 2019). GIS Dictionary - spatial analysis, [Online]. Available: https://support.esri.com/en/other-resources/gis-dictionary/term/474ae759-ebdd-49d7-ae34-7e902b57a011.

[58] E. L. Nelson and P. G. Greenough, "Geographic information systems in crises," vol. 2nd edition, pp. 312–316, 2016. [Online]. Available: https://www.sciencedirect.com/science/book/9780323286657#ancsc0025.

[59] ESRI. (Jan. 31, 2019). An overview of the density toolset, [Online]. Available: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/an-overview-of-the-density-tools.htm.

[60] ——, (Jan. 31, 2019). Kernel density, [Online]. Available: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/kernel-density.htm.

[61] ——, (Jan. 31, 2019). How kernel density works, [Online]. Available: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kernel-density-works.htm.

[62] ——, (Jan. 31, 2019). Differences between point, line, and kernel density, [Online]. Available: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/differences-between-point-line-and-kernel-density.htm.

[63] B. W. Silverman, Ed., Density Estimation for Statistics and Data Analysis. London, New York, Chapman and Hall: University of Bath, UK, Chapman and Hall/CRC, 1986. [Online]. Available: https://books.google.si/books?id=e-xsrjsL7WkC&printsec=frontcover&redir_esc=y&hl=sl#v=onepage&q&f=false.

[64] C. Dempsey. (Jan. 31, 2019). Heat Maps in GIS, [Online]. Available: https://www.gislounge.com/heat-maps-in-gis/.

[65] ESRI. (Jan. 31, 2019). An overview of the Spatial Statistics toolbox, [Online]. Available: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/an-overview-of-the-spatial-statistics-toolbox.htm.

[66] ——, (Jan. 31, 2019). GIS Dictionary-spatial statistics, [Online]. Available: https://support.esri.com/en/other-resources/gis-dictionary/term/97a2c5ae-d8cb-476f-b72c-fda1314c27f4.

[67] B. N. Boots and A. Getis, Eds., Point Pattern Analysis. Harlow, England: Sage Publications, 1988.

[68] L. Yongmei, "Spatial cluster analysis for point data: Location quotients verses kernel density," UCGIS. University Consortium of Geographic Information Science Summer Assembly 2000At: Portland, Oregon, 2000. [Online]. Available: http://dusk.geo.orst.edu/ucgis/web/oregon/papers/lu.htm.

[69] ArcGIS Pro. (Feb. 16, 2019). An overview of the mapping clusters toolset, [Online]. Available: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/an-overview-of-the-spatial-statistics-toolbox.htm.

[70] D. Attaway, "Gis analysis workshop," ESRI, Ed., 2016 GIS for a Sustainable World Proceedings, Geneva: ESRI, 2016, p. 51. [Online]. Available: http://proceedings.esri.com/library/userconf/unic16/papers/un_19.pdf.

[71] ArcGIS Pro. (Feb. 17, 2019). What is a z-score? What is a p-value? [Online]. Available: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/what-is-a-z-score-what-is-a-p-value.htm.

[72] ESRI. (Jan. 28, 2019). Optimized Hotspot analysis, [Online]. Available: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/optimized-hot-spot-analysis.htm.

[73] D. Howitt and D. Cramer, Eds., Introduction to SPSS Statistics in Psychology. Harlow, England: Pearson, 2011 5th ed.

[74] SPSS Tutorials. (Feb. 16, 2019). Pearson Correlations – Quick Introduction, [Online]. Available: https://www.spss-tutorials.com/pearson-correlation-coefficient/.

[75] S. Zimmermann-Janschitz, Ed., Statistik in der Geographie. Graz, Austria: Springer, 2014.

[76] SPSS Tutorials. Sigma Plus Statistiek. (2018). Statistical significance – what does it really mean? [Online]. Available: https://www.spss-tutorials.com/statistical-significance/.

[77] J. E. Hickin, Ed., Maps and Mapping - A cartographic manual. Burnaby, B.C. Canada: Simon Fraser University, R.S. Graphics, and Printing, 2014, 3rd edition. [Online]. Available: https://books.google.si/books?id=e-xsrjsL7WkC&printsec=frontcover&redir_esc=y&hl=sl#v=onepage&q&f=false.

[78] ESRI. (Jan. 31, 2019). Choropleth map, [Online]. Available: https://support.esri.com/en/other-resources/gis-dictionary/term/4581559f-32b3-4575-b5c9-34b120593809.

[79] S. Isaac. (Jan. 31, 2019). Carto: 5 popular thematic map types and techniques for spatial data, [Online]. Available: https://carto.com/blog/popular-thematic-map-types-techniques-spatial-data/.

[80] ESRI. (Jan. 31, 2019). Proportional symbol, [Online]. Available: https://support.esri.com/en/other-resources/gis-dictionary/term/f361d050-3f2a-4672-9045-ed2c5c375580.

[81] ——, (Jan. 31, 2019). Using proportional symbols, [Online]. Available: http://desktop.arcgis.com/en/arcmap/10.3/map/working-with-layers/using-proportional-symbols.htm.

[82] WIBIS. (Feb. 7, 2019). Einwohner, [Online]. Available: https://wibis-steiermark.at/bevoelkerung/struktur/einwohner-gesamt/.

[83] Das Land Steiermark. (Feb. 2, 2019). Steiermark - Das (Tourismus) Jahr 2017: Daten & Fakten, [Online]. Available: https://www.verwaltung.steiermark.at/cms/dokumente/12208857_117401915/3ed0f368/20180926_600_Steiermark.pdf.

[84] tierischer-urlaub.com. (Mar. 1, 2019). Uts-karte-steiermark-touristische-regionen-tierischer-urlaub-mit-hund-und-katze-hundefreundlich, [Online]. Available: https://tierischer-urlaub.com/uts-karte-steiermark-touristische-regionen-tierischer-urlaub-mit-hund-und-katze-hundefreundlich/.

[85] Das Land Steiermark. (Feb. 2, 2019). Ausseerland-Salzkammergut - Das (Tourismus) Jahr 2017: Daten & Fakten, [Online]. Available: https://www.verwaltung.steiermark.at/cms/dokumente/12208857_117401915/d6232160/20180926_21_Ausseerland-Salzkammergut.pdf.

[86] ——, (Feb. 2, 2019). Hochsteiermark. Das (Tourismus) Jahr 2017: Daten & Fakten, [Online]. Available: https://www.verwaltung.steiermark.at/cms/dokumente/12208857_117401915/d099207b/20180926_24_Hochsteiermark.pdf.

[87] UNESCO. (Feb. 10, 2019). City of Graz – Historic Centre and Schloss Eggenberg, [Online]. Available: https://whc.unesco.org/en/list/931/.

[88] Das Land Steiermark. (Feb. 2, 2019). Region Graz. Das (Tourismus) Jahr 2017: Daten & Fakten, [Online]. Available: https://www.verwaltung.steiermark.at/cms/dokumente/12208857_117401915/28e8cbde/20180926_23_Region%5C%20Graz.pdf.

[89] ——, (Feb. 2, 2019). Schladming-Dachstein. Das (Tourismus) Jahr 2017: Daten & Fakten, [Online]. Available: https://www.verwaltung.steiermark.at/cms/dokumente/12208857_117401915/3e03bdaa/20180926_22_Schladming-Dachstein.pdf.

[90] ——, (Feb. 2, 2019). Süd-Weststeiermark. Das (Tourismus) Jahr 2017: Daten & Fakten, [Online]. Available: https://www.verwaltung.steiermark.at/cms/dokumente/12208857_117401915/772ee699/20180926_27_Su%5C%CC%5C%88d-Weststeiermark.pdf.

[91] ——, (Feb. 2, 2019). Thermenland Steiermark. Das (Tourismus) Jahr 2017: Daten & Fakten, [Online]. Available: https://www.verwaltung.steiermark.at/cms/dokumente/12208857_117401915/36374be3/20180926_26_Thermenland%5C%20Steiermark-Oststeiermark.pdf.

[92] ——, (Feb. 2, 2019). Urlaubsregion Murtal. Das (Tourismus) Jahr 2017: Daten & Fakten, [Online]. Available: https://www.verwaltung.steiermark.at/cms/dokumente/12208857_117401915/6b2ca15f/20180926_25_Urlaubsregion%5C%20Murtal.pdf.

[93] ——, (May 20, 2018). Sonstige (Gesäuse). Das (Tourismus) Jahr 2017: Daten & Fakten, [Online]. Available: https://www.verwaltung.steiermark.at/cms/dokumente/12208857_117401915/7f5a1265/20180926_28_Sonstige.pdf.

[94] Encyclopaedia Britannica. (May 10, 2018). Semantics, [Online]. Available: https://www.britannica.com/science/semantics.

[95] Dictionary.com. (May 10, 2018). Semantics, [Online]. Available: https://www.dictionary.com/browse/semantics.

[96] A. Rosen. (May 20, 2018). Tweeting made easier. Twitter Blog, [Online]. Available: https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html.

[97] Twitter. (May 10, 2018). Twitter Developers. Twitter, [Online]. Available: https://developer.twitter.com/en.html.

[98] Twitter Data. (May 20, 2018). Twitter data and the financial markets, [Online]. Available: https://blog.twitter.com/official/en_us/topics/insights/2016/twitter-data-and-the-financial-markets.html.

[99] K. Janowicz, S. Scheider, and A. B., "A geo-semantics flyby," Reasoning Web. Semantic Technologies for Intelligent Data Access. In: Reasoning Web 2013. Lecture Notes in Computer Science, vol. 8067, 2013.

[100] ESRI, "Spatial data standards and gis interoperability," An ESRI white paper, 2003. [Online]. Available: https://support.esri.com/en/white-paper/436.

[101] English Oxford living dictionary. (Feb. 8, 2019). Ontology, [Online]. Available: https://en.oxforddictionaries.com/definition/ontology.

[102] J. Burchell. (Feb. 3, 2019). Using VADER to handle sentiment analysis with social media text, [Online]. Available: http://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html.

[103] K. Kadunc and M. Robnik, "Analiza mnenj s pomocjo strojnega učenja in slovenskega leksikona sentimenta," Conference on Language Technologies Digital Humanities, 2016.

[104] Collins English Dictionary. (Feb. 8, 2019). Machine learning, [Online]. Available: https://www.collinsdictionary.com/dictionary/english/machine-learning.

[105] Techopedia. (Feb. 8, 2019). Machine learning, [Online]. Available: https://www.techopedia.com/definition/8181/machine-learning.

[106] M. Hu and B. Liu, "Mining opinion features in customer reviews," In Proceedings of AAAI Conference on Artificial Intelligence, vol. 4, pp. 755–760, 2004. [Online]. Available: https://pdfs.semanticscholar.org/ee6c/726b55c66d4c222556cfae62a4eb69aa86b7.pdf.

[107] IBM Knowledge Center. (Feb. 17, 2019). About text mining, [Online]. Available: https://www.ibm.com/support/knowledgecenter/vi/SS3RA7_18.0.0/ta_guide_ddita/textmining/shared_entities/tm_intro_tm_defined.html.

[108] P. van Kessel. Pew Reserch Center. (Feb. 18, 2019). An intro to topic models for text analysis, [Online]. Available: https://medium.com/pew-research-center-decoded/an-intro-to-topic-models-for-text-analysis-de5aa3e72bdb.

[109] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on word clouds," 47th Hawaii International Conference on System Science, vol. 47, 2014. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6758829.

[110] S. Henderson, S. Evergreen, T. Jarosewich, and A. Mountain. (Feb. 8, 2019). Word cloud, [Online]. Available: https://www.betterevaluation.org/en/evaluation-options/wordcloud.

[111] S. Nikolenko, S. Koltcov, and O. Koltsova, "Topic modelling for qualitative studies," Journal of Information Science, vol. 43, no. 1, pp. 88–102, 2017. [Online]. Available: http://dusk.geo.orst.edu/ucgis/web/oregon/papers/lu.htm.

[112] Orange, Laboratory of Bioinformatics UL. (Feb. 18, 2019). Topic modelling, [Online]. Available: https://orange3-text.readthedocs.io/en/latest/widgets/topicmodelling.html.

[113] T. Dumais, "Latent semantic analysis," Annual Review of Information Science and Technology, vol. 38, pp. 188–230, 2005. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440380105.

[114] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, vol. 3, no. 4-5, pp. 993–1022, 2003. [Online]. Available: http://jmlr.csail.mit.edu/papers/v3/blei03a.html.

[115] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical dirichlet processes," Journal of the American Statistical Association, vol. 101, no. 476, pp. 1566–1581, 2006. [Online]. Available: http://www.gatsby.ucl.ac.uk/~ywteh/research/npbayes/jasa2006.pdf.

[116] Twinword ideas. (2018). Dictionary, [Online]. Available: https://www.twinword.com/ideas/graph/dictionary/.

[117] A. Beaver, Ed., A Dictionary of Travel and Tourism. Harlow, England: Oxford University Press, 2012. [Online]. Available: http://www.oxfordreference.com/view/10.1093/acref/9780191733987.001.0001/acref-9780191733987.

[118] webterm.term-portal. (2018). Webterm.term-portal. tourismus, [Online]. Available: http://webterm.term-portal.de/DEUTERM/tourismus/tourismus_e.htm.

[119] Orange3 Text Mining. (Mar. 1, 2019). Preprocess text, [Online]. Available: https://orange3-text.readthedocs.io/en/latest/widgets/preprocesstext.html.

[120] Statistik Austria. (Mar. 1, 2019). Ankünfte und Nächtigungen nach Herkunftsländern, Kalenderjahr 2018, [Online]. Available: https://www.statistik.at/web_de/statistiken/wirtschaft/tourismus/beherbergung/ankuenfte_naechtigungen/index.html.

[121] Land Steiermark. (Mar. 1, 2019). TOURISMUS - Analysen Dezember 2018. Ankünfte, Übernachtungen und durchschnittliche Aufenthaltsdauer nach Herkunftsländern, [Online]. Available: http://www.landesentwicklung.steiermark.at/cms/beitrag/12707979/141979459/.

[122] M. J. Garbade. (Feb. 24, 2019). Understanding K-means Clustering in Machine Learning, [Online]. Available: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.

[123] ArcGIS Pro. (Feb. 21, 2019). Data classification methods, [Online]. Available: https://pro.arcgis.com/en/pro-app/help/mapping/layer-properties/data-classification-methods.htm.

[124] M. J. de Smith, M. F. Goodchild, and P. A. Longley. (Feb. 21, 2019). Classification and clustering. geospatial analysis — a comprehensive guide, [Online]. Available: http://www.spatialanalysisonline.com/HTML/index.html.

[125] L. Widrich. (Feb. 28, 2019). How Twitter evolved from 2006 to 2011, [Online]. Available: https://buffer.com/resources/how-twitter-evolved-from-2006-to-2011.