Paper

Tourism statistics and the use of social media data

Shirley Ortega Azurduy
Nico Heerschap

May 2020

Contents

Summary

Social media are, generally speaking, digital platforms through which users create online communities and share personal messages, information, ideas and other content, like pictures, videos and music. Since the last decade, the penetration rate of social media platforms in society has grown enormously: more and more businesses, governments and above all persons are using social media platforms in some way. This growth is expected to continue. In the era of big data, that raises the question whether data from publicly available social media messages can be used for (new and quicker) statistics? Social media data are, in fact, unique due to their combination of content and information on geo-location and time. On the other hand, social media research also clearly has its limitations.

This paper, therefore, presents the results of an exploratory study of the use of social media as a new big data source for the domain of tourism. Before the COVID-19 crisis, tourism was a substantial industry of the Dutch economy. Growth was expected to continue. But this growth also raised questions about the limits of the development of tourism, asking for quicker and more detailed statistics. Recently, tourism was hit especially hard by the COVID-19 crisis. This did not lead to the need for less statistics, on the contrary. However, the nature of the questions has changed.

Based on a number of experimental studies it was investigated whether it is possible to use social media data to produce (new and quicker) statistics for the domain of tourism. For example, research was carried out if it is possible at all to distinguish social media messages from groups of visitors from social media messages from other groups, like residents or businesses. Or, can sentiment analysis be used to get more insight in the satisfaction of visitors about the places they have visited. To analyze sentiment, it is necessary to examine the content of the social media messages. This is in line with the possibilities to examine also other aspects of the customer journey by looking more closely at the content of these messages. Some social media messages contain a timestamp and a geo-location. This provides opportunities to look at the mobility of visitors, or separately at the mobility of excursionists or tourists. This is comparable to research based on mobile phone data, where social media have the advantage that they also offer content and through the GPS better location data. The disadvantage is that the number of contact points of social media is much lower than that of mobile phones. This means that with further detailing, the number of available cases can quickly decrease. What still lingers over the market of social media research, is the representativeness and validity of the end results for the target population. Although not a subject of this study, more research needs to be done on this issue. A more solid conceptual and methodological foundation for social media research is needed. Nevertheless, results of social media research can be used as beta-indicators and can be analyzed in relation with other big data sources. Layered geographic information systems (GIS) are the right tools for the integration of these different data sets. It not only opens up possibilities to include, combine and analyze all kinds of (external) data sources on tourism, but, on the other side, it offers essential geographical information as well and it offers especially information on the supply side of tourism, like attractions, accommodations, museums, restaurants, airports etc.

To promote social media research for tourism statistics, it is recommended to start simple, to see the results as beta-indicators and, if possible, combine these data with other data sources in a geographical information system. Visualization of the results will also certainly help. Important issues for research are, among others, representativeness, classifiers, data

cleaning, origin and destination matrices and last but not least content analysis. It is also important not to forget the concepts used in tourism research. And, finally, research should be carried out in multi-disciplinary teams.

Acknowledgments

We would like to thank the Ministry of Economic Affairs and Climate for their contribution. This research would not have been possible without their support.

1 Introduction

During the past ten to fifteen years, tourism has become an important sector for the Dutch economy. In 2018 the sector generated around 87,5 billion euros in spending, leading to an added value of around 30,4 billion euros: 4.4 percent of GDP. The expenditure of visitors resulted in a work force of 679 thousand people employed, good for 474 thousand fulltime equivalents: 6,3 percent of the total labour volume of the Netherlands. The strong growth of tourism was expected to continue in the near future. However, due to the COVID-19 crisis, tourism has been hit particularly hard recently. The number of flights has been drastically reduced, museums and restaurants are closed, events cannot take place and in some cases there are entry bans. It is unclear when, how and whether full recovery will take place. Nonetheless, these developments did not change the principles of this research.

Whether it was the growing importance of tourism or the consequences of the COVID-19 crisis, both events raised more demand for data and supporting information for policy makers and research. Until now, this information was mainly based on traditional surveys and, in a later stage, also on administrative sources. Recently, however, a new source of information has been added: so-called big data sources. These new big data sources are mainly based on 'digital footprints' that people leave behind when they surf the internet, use their mobile phone, buy goods with their credit card, walk past cameras or park their cars. This type of data is increasingly stored electronically and is in some cases available for research and statistics.

One of those new sources of big data are social media. Social media platforms are online communication channels dedicated to community-based input, interaction, content-sharing and collaboration. Examples of social media platforms are Twitter, Facebook, LinkedIn and Instagram. For tourism this involves social media messages of visitors to their family and friends up to their opinions about the places they visit or have visited. These messages can contain pictures, sound, videos or text. Often, these messages also contain a geo-location and a timestamp. This combination of content, place and time, in fact makes social media a unique data source. On the other hand, the use of social media data also has clear limitations. If these social media messages are public, they can be scraped from the internet, stored and analyzed for statistics.

Statistics Netherlands already conducted research to look at the usability of social media data for statistics in the domain of businesses (Ortega and Heerschap, 2019). The main conclusion of that study was that there are clear possibilities, but to capitalize on the results and move to the production of structural statistics more research is necessary. In fact, this is a conclusion that still accounts for many big data sources. Besides this, the representativeness of social media data is a major issue. The experiments described in this paper follow the work that has been carried out in the framework of business statistics. Partly because they use the same data set as the starting point of the analysis.

The structure of this paper is as follows. In chapter 2 the problem statement is described. Chapter 3 briefly outlines the current use of social media data in tourism research. Before presenting the main results in chapter 5, the sources that were used in this experimental study are described in chapter 4 first. The paper ends with some conclusions about the possibilities and the limitations to use social media data in the domain of tourism statistics (chapter 6).

2 Problem statement

Tourism entails the movement of people to places outside their usual environment for personal, business or other purposes for not more than one year.[1)] What belongs to the "usual environment" is left to either the respondent or to the compiler of the statistics. Criteria, that are often used, are time (> two hours), frequency (e.g. less than once a month) and distance (e.g. outside the city boundary). People who travel outside their usual environment are called visitors. Visitors can be divided into excursionists (no overnight stay) and tourists (with at least one overnight stay). Another approach is to divide visitors into Dutch and foreign visitors, respectively domestic and inbound tourism. The nationality is not the determining factor, but the place of residence. A person with a Dutch passport, but living in Germany, is therefore a foreign visitor for the Netherlands, while an Englishman living in the Netherlands is a domestic visitor. Tourism is not just about traveling for leisure, but it can also be about traveling for business or, for example, for health, study, pilgrimage or for any other personal reason, as long as it is outside the usual environment and for no longer than one year. Indicators for tourism focus on the whole customer journey from orientation, booking the trip, the travel to the destination, the stay and activities at the destination, as well as the impact on the destination. The impacts can be measured in terms of expenditure, but also, among others, in social and environmental pressure.

In 2019, almost 9 out of 10 Dutch persons used social media platforms either just for sending text messages or for maintaining social and professional networks. Already in 2017, the percentage of social media users accounted for 85 percent of the Dutch population older than 12 years (CBS, 2019). The penetration rate will differ per country. In this study, we assume that a part of the social media users also uses these platforms during their holidays or business trips while they are travelling through the Netherlands. The main question then is: can social media data be used to say something about the movement and behavior of visitors in the Netherlands? To investigate this question, in this study several experiments were conducted with the social media data which were also at the basis of a study on the use of social media data for businesses statistics. In that study social media messages and profiles from businesses were separated from a set of all social media messages and profiles. So, in fact, this study uses the other part of that data set: the social media messages and profiles from persons.

This then leads to the following problem statement:

Can social media data be used as a new source of information for the production of statistics in the domain of tourism?

This problem statement was specified in the following sub-questions:

- Can social media data from visitors be separated from social media data in general?
- If yes, is it possible to divide social media messages and profiles in excursionists and tourists?
- If yes, is it possible to perform some kind of sentiment-analysis?
- If yes, can social media data be linked to points of interest (supply side of tourism)?

1) See https://www.unwto.org/glossary-tourism-terms.

– If yes, can we visualize flows of visitors with social media data, comparable to mobile phone data?

This study must be seen as an exploratory research.

3 Tourism research using social media

Social media data are already applied in research in the domain of tourism. A first field of research is the use of social media data for **marketing and strategy** purposes. It is clear that social media have transformed the way information is generated and distributed. This also accounts for the domain of tourism. For example, it is not only important to send information, but also to encourage visitors to talk about (good) experiences on the internet when they visit a city or region. Some examples of research questions in the field of tourism marketing are: how to influence visitors through the use of social media; how to build and keep a positive brand name (brand confidence); and how to find and react to negative responses of visitors. The last two questions relate to the internet presence of tourism enterprises and aspects such as likes, shares and retweets. The field of marketing research is outside the scope of this paper.

A second field of research focuses on the **content** of the social media messages, that are posted. An example is that of sentiment analysis as a proxy for the satisfaction of visitors. Methods are developed to see if social media messages are positive, neutral or negative. A next step in this kind of research is not only to look at the sentiments expressed in social media messages, but also to research the content of the messages in more detail. This can lead, for example, to information about which topics are of interest to visitors before and during his or her visit, what kind of activities visitors undertake during their holidays and information about the background of visitors (e.g. their profiles). For the analysis of the content of social media messages methods of text mining through word frequencies, collocation and accordance or machine learning and text analytics are used. This kind of research falls within the scope of this paper.

A third field of research, that uses social media data, focuses on the **mobility and whereabouts** of visitors. This can be analyzed, because a part of the social media messages contains a timestamp and a geo-location. With this, visitors can be followed for a longer period. Depending on the type of social media platform, only a relatively small part of the messages contains a geo-location and a timestamp. However, because the big number of messages available, this is often still a substantial number of messages. This kind of research compares to the research carried out with mobile phone data. The advantage of location data from social media messages is that these are based on GPS, rather than on estimates like those of mobile phone data. On the other hand, mobile phone data are larger in numbers (more contact points) and the penetration rate in the population is much higher. Because tourism is also about mobility, the combination of content and information on location and time is a unique feature of social media data. This becomes more important when this kind of data can also be combined with data on the supply side of tourism, that is: accommodations, attractions, restaurants, airports and transportation. This kind of research falls within the scope of this paper.

A final area of research, that can be mentioned here, focuses on **methodological research**. This research is not specifically domain-orientated but broader. It is, among others, about techniques such as text mining, machine learning, sentiment analysis, geographic

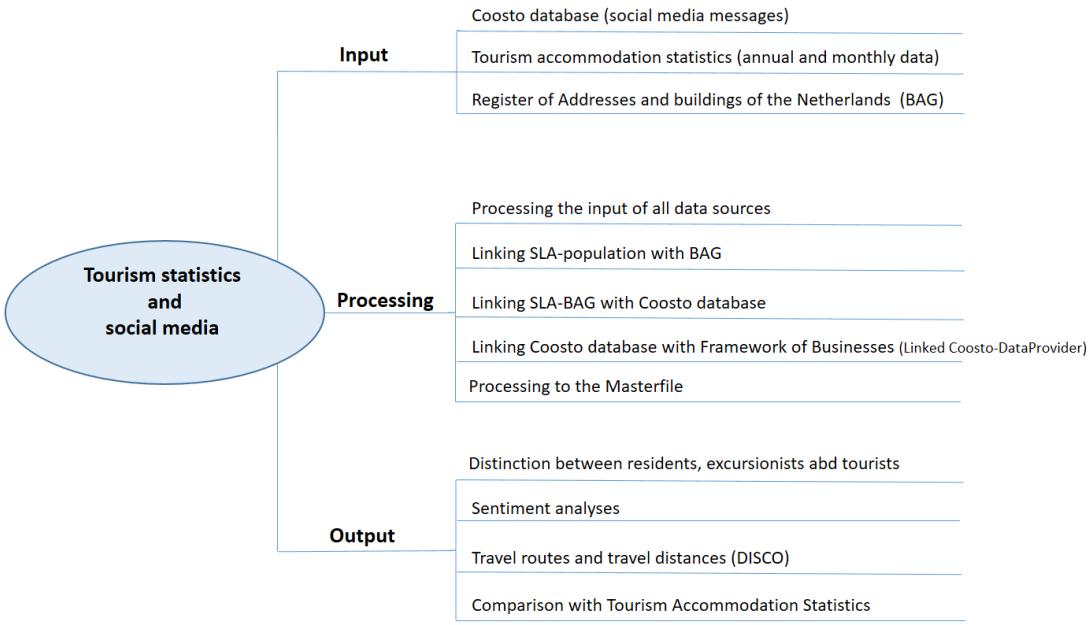
information systems (GIS), classifiers [2] and process mining. Another, in fact, much bigger methodological issue concerns the representativeness of the end results when using social media data. As with more big data sources, there is insufficient insight in the **representativeness and validity**, or in other words: does the information based on social media data represent the target population in a correct way? [3] This requires a better conceptual and methodological foundation for this kind of research. One of the ways to do so, is, among others, comparisons with the results of other (big) data sources. Research into the issue of representativeness is not part of this study. That would take more time than available for this study.

2) See for example Spinder, S.T., 2019.
3) Or can it be weighted that it represents the target population in a correct way?

4 Data sources and process flow

This chapter provides a description of the data sources, which were used in this exploratory study, and the way these data sources were processed to a Masterfile. This includes the description of all the steps that were taken to produce the basic data for this study. See Figure 4.1.

Figure 4.1 Scheme of all the steps in the process flow



Source: Statistics Netherlands

4.1 Description of the sources

In this study five data sets from four different sources were used:

1. A database with social media data from the company Coosto[4)];
2. A register with all addresses and buildings in the Netherlands (BAG);
3. The population register of the Tourism Accommodation Statistics of Statistics Netherlands (TAS, annual data);
4. The monthly figures of the number of tourists and overnight stays from the TAS;
5. The dataset from the company Dataprovider. Dataprovider scrapes on a regular basis all the websites of the Dutch internet domain. In another study this data set of Dataprovider was already linked to the Dutch Business register of Statistics Netherlands. The resulting data set was then linked to data of Coosto on the basis of the matching key “profile name”. These two linkages made it possible to distinguish between social media data of private persons and social media data of businesses (Dutch part only). This data source is not described further in this paper.

4) A company that is specialized in the scraping of publicly available social media messages from the internet.

Table 4.1 Input query specifications for the API-tool of Coosto and the number of returned records (with a geo-location)

Input query keywords	Period	Type of sample	Number of records	GPS-coordinates	Used
.nl OR .com	**21-06-2017 until 04-08-2017**	**Chronological and Random**	**1.762.000**	**1%**	**Selection 1**
strand OR beach OR Texel OR Scheveningen OR Texel OR Zandvoort OR Noordwijk OR Zeeland	30-07-2017 until 14-08-2017	Chronological	148.000	13%	Not used
Scheveningen OR Texel OR Zandvoort OR Noordwijk OR Zeeland	14-04-2017 until 14-08-2017	Chronological	140.000	12%	Not used
Scheveningen OR Texel OR Zandvoort OR Noordwijk OR Zeeland	1-01-2017 until 07-07-2017	Random	10.000	7%	Not used
Texel (OR strand OR beach)	**12-08-2016 until 07-07-2017**	Chronological	**20.000**	**>95%**	**Selection2**
Texel (photo OR video)	01-01-2015 until 17-03-2017	Chronological	80.000	9%	Not used
Ameland (photo OR video)	01-01-2015 until 31-03-2017	Chronological	80.000	12%	Not used

Source: Coosto, processed by Statistics Netherlands

Coosto-database with social media data

The Coosto-database with social media data is a source which contains almost all messages and profiles circulating publicly on social media platforms[5)], such as Instagram, Twitter and Facebook, and in forums and blogs. The social media platforms Instagram and Twitter are particularly important for this study as these social media messages often include GPS-coordinates of the sender, when he or she posts a message.

The data from Coosto are from 2017. The data were initially extracted for a study on social media data and business statistics (Ortega and Heerschap, 2019). To extract the data from the database the standard API-tool was used. The number of records, that can be extracted from the Coosto-database in one session, is limited to 10.000 messages per query. During weekdays queries could be restarted (chronological data). During weekends the extraction of the data was randomly limited to one query of 10.000 records (random data). See table 4.1 for the different periods data sets were extracted.

The extracted data were used in two different ways. First, a part of this study was based on the data, which were already processed for the research on social media data and business statistics (**selection 1**). The main advantage was that the distinction between the

5) It does seem that Coosto is more focused on social media messages with text than on social media with, for example, pictures.

social media data of persons and the social media data of businesses was already made. As said, this was done by linking the social media data with data from all websites of businesses from the Dutch internet domain (Dataprovider). Instead of using the part with social media data of businesses, the other part of the data set was used: the set of social media data of persons. A disadvantage was that this selection was, for specific reasons, limited to social media messages with an URL (.com and .nl). However, for the results of this study this limitation did not really matter. The data set consisted of around 1,7 million messages. However only 1 percent of the messages contained a geo-location and a timestamp, resulting in a data set for analysis of somewhat more than 17.000 messages.

Second, the other part of this study was based on the whole set of extracted social media data from the Coosto-database, at least after the social media data of businesses were deleted[6] (**selection 2**). This part of the study focused specifically on certain touristic areas in the Netherlands. Therefore, a series of selections of social media data based on keywords such as 'Scheveningen", "Texel", "Noordwijk", "beach" etc. was tried out. These keywords were chosen because they relate to major touristic destinations.[7] Subsequently, these selections were limited to social media messages that contained a timestamp and a geo-location.[8] The analysis was then largely carried out on the dataset with the keyword "Texel" (somewhat less than 20.000 messages). This dataset was chosen over the other possible data sets, because it contained the most messages with a geo-location and a timestamp. In addition, this data set covered an entire year.

Table 4.1 gives some indications on how the different sets of social media messages were selected through the chosen keywords, how many records were thereafter collected from the Coosto-database and what the percentages were that contained a geo-location and a timestamp. Striking was the high percentage of messages in the chosen data set of the island of Texel that contained a geo-location and a timestamp. Most likely this has to do with the fact that visitors will use Google-maps, for which they have to turn on their GPS.

The data set, which was used in the analysis, is structured as follows:

- Query
- Date
- Url
- Sentiment
- Number of views
- Author of the message
- Followers
- Influence
- GPS-Latitude
- GPS-Longitude
- Content of message
- Source of message (from which social media platform)

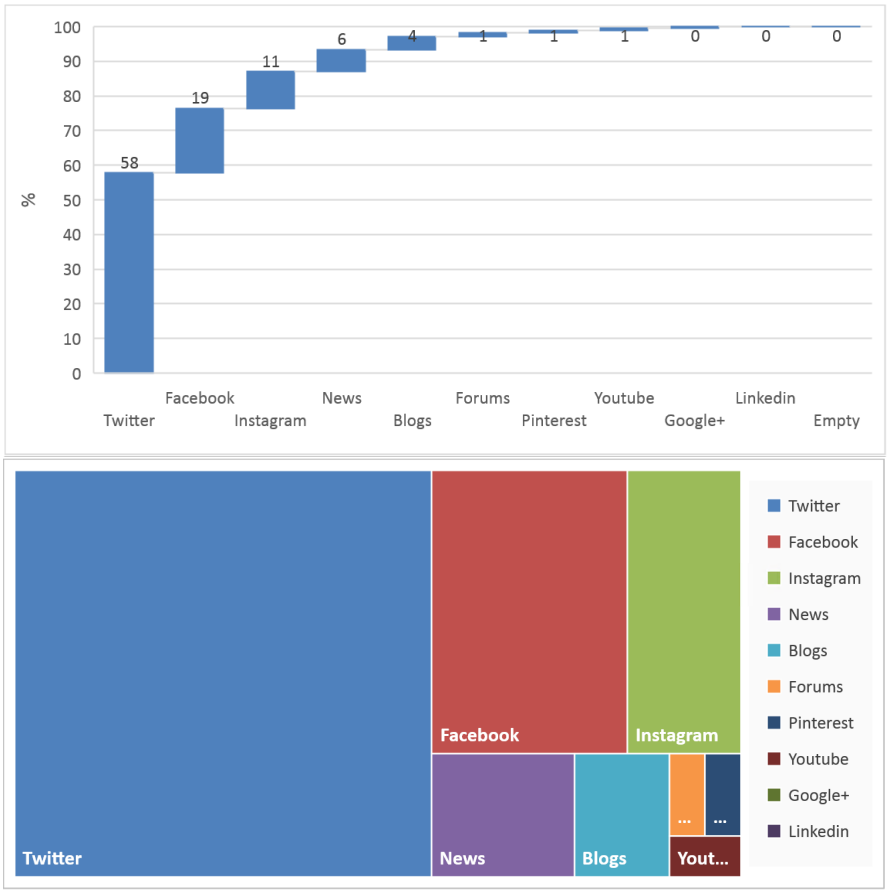
6) To this end, this step uses the linked data of the study on social media data and business statistics (Ortega and Heerschap, 2019).

7) The shares of the coastal area in terms of number of foreign visitors and their number of nights spent are 78 percent and 75 percent of the total respectively. As for the Dutch visitors shares are 43 percent for the total number of visitors and 37 percent for the total nights spent.

8) It is known that the number of persons providing "location features" (=GeoMarks) in their messages is larger than those who have their GPS turned on (= GeoTagging). In this exploratory study, however, the GeoTagging feature is preferred.

Figure 4.2 shows that social media messages with a geo-location and a timestamp were present in microblogs, such as Twitter and in social networks, like Facebook and Instagram, but less often in blogs and forums. In the study on the use of social media and business statistics (Ortega and Heerschap, 2019) Twitter messages also easily outnumbered other types of social media.

Figure 4.2 Share of the different social media messages available in the Coosto-database: the case of coastal tourism, first semester 2017



Source: Coosto, processed by Statistics Netherlands

Register with all addresses and buildings in the Netherlands (BAG)

A second source, that was used in this exploratory study, was a register with all addresses and buildings in the Netherlands. This register is called the BAG. The BAG also contains geographic coordinates (longitudes and latitudes) of all buildings and dwellings in the Netherlands.

The BAG-Extract of 2017 contained about 9 million objects (buildings and dwellings). The top 5 cities with the most objects are Amsterdam, Rotterdam, The Hague, Utrecht and Eindhoven. About 7,8 million objects have as their main function "residence" and about 120 thousand objects are registered as "tourist accommodations" ('logies'). [9)] That is about 1,4 percent of the total number of objects.

9) The BAG is, among others, used to control that objects registered as tourist accommodations submit tourism tax on their incomes fairly.

The BAG uses the so called RDNAP-system[10)] to presents its geo-locations. The geo-locations in the Coosto-database are presented in the European Terrestrial Reference System 1989 (ETRS89). So, it was necessary to translate RDNAP-geo-locations to the ETRS89-format. This was done with the conversion program RDNAPTRANS.

Establishments from the Tourism Accommodation Statistics (TAS)

The third data source, that was used in this study, is data from the Tourism Accommodation Statistics ('Logiesaccommodaties', TAS) of Statistics Netherlands. The TAS includes the annual number of tourist accommodation establishments in the Netherlands by type and, per accommodation, the number of bedrooms, bed places and, in particular, a register with addresses. This population register is based on an annual inventory, which uses different input sources, like scraped data and data from the Chamber of Commerce.

For every record in the register the following information is available: an identification number, name and address, city, touristic region, province, type of accommodation, number of beds and/or number of camping pitches, opening period, as well as other information such as number of stars and registration number of the Chamber of Commerce.

In this study the address (postal code 6, house number and house letter) was used as the matching variable to link the population of the accommodation register to the BAG-database. As this was not a perfect match, the resulting registry consisted eventually of 11,4 thousand tourism accommodation establishments. In 2017, the top-5 municipalities with the highest number of tourism accommodation establishments were Amsterdam, Schouwen-Duiveland (Zeeland), Súdwest-Fryslân (Friesland), Texel and Veere (Zeeland). It is important to realize here, that every house in a bungalow park counts as one object in the BAG.

Monthly data on number of tourists and nights spent

This data source also comes from the TAS. It contains data on the number of tourists (or guests) and nights spent per month per tourism accommodation establishment in the Netherlands. The survey is based on a sample from the accommodation register of the TAS described in the previous section. In this study monthly data were used from 2015 -2017. The data on the number of tourists and the number of overnight stays from this survey were used to compare with the estimated relative values obtained using social media data.

4.2 Process flow

After the data extraction and collection was completed, the processing and linking of the different data sets were carried out. This process consisted of three phases:

1) Reading, analyzing and editing the collected data;
2) Linking the different data sets and removing duplicates; and
3) Manipulating data frames to obtain the final Master File for the analysis.

10) Using the Rijksdriehoeksmeting (RD) and the Normaal Amsterdams Peil (NAP).

Reading, analyzing and editing of the input data

This process step includes, among others, the elimination of outliers. For example, records with a GPS-position that were outside the expected range, were deleted. In addition, the data were also cleaned from white spaces, strings were set on capital letters and non UTF-8 characters were deleted.

Linking data sources and removing duplicates

Three different linkages were carried out in this step of the process.

First, the annual data of tourism accommodation establishments from the TAS were linked to the data from the BAG. The combined matching key was the 6 figures of the postal code, house number and, if present, house letter. It is important to notice that it is also necessary to remove duplicates after this linkage. For example, bungalow parks have many dwellings with the same address, the address differs only on the house letter.[11)] The deduplication of records after the matching was a separate step in the process flow. The deduplication process reduced the number of records with about 8.000 (from 23.000 to 15.000).

Second, the resulting data source from the previous step (TAS+BAG) was linked to the selected Coosto-database. Matching was done on the basis of the BAG-geo-location. The Coosto-data are matched by vicinity, i.e. messages up to 100 meters (or 200 meters) are clustered around the closest registered accommodation in the TAS+BAG.

Third, the results of the second step were linked to the monthly data on the number of tourists and overnight stays from the TAS. This was needed because we wanted to compare estimates of the number of visitors and overnight stays from social media data with data from the TAS. The standard linking of the TAS was used here.

Processing and visualizing to compile the final Master File

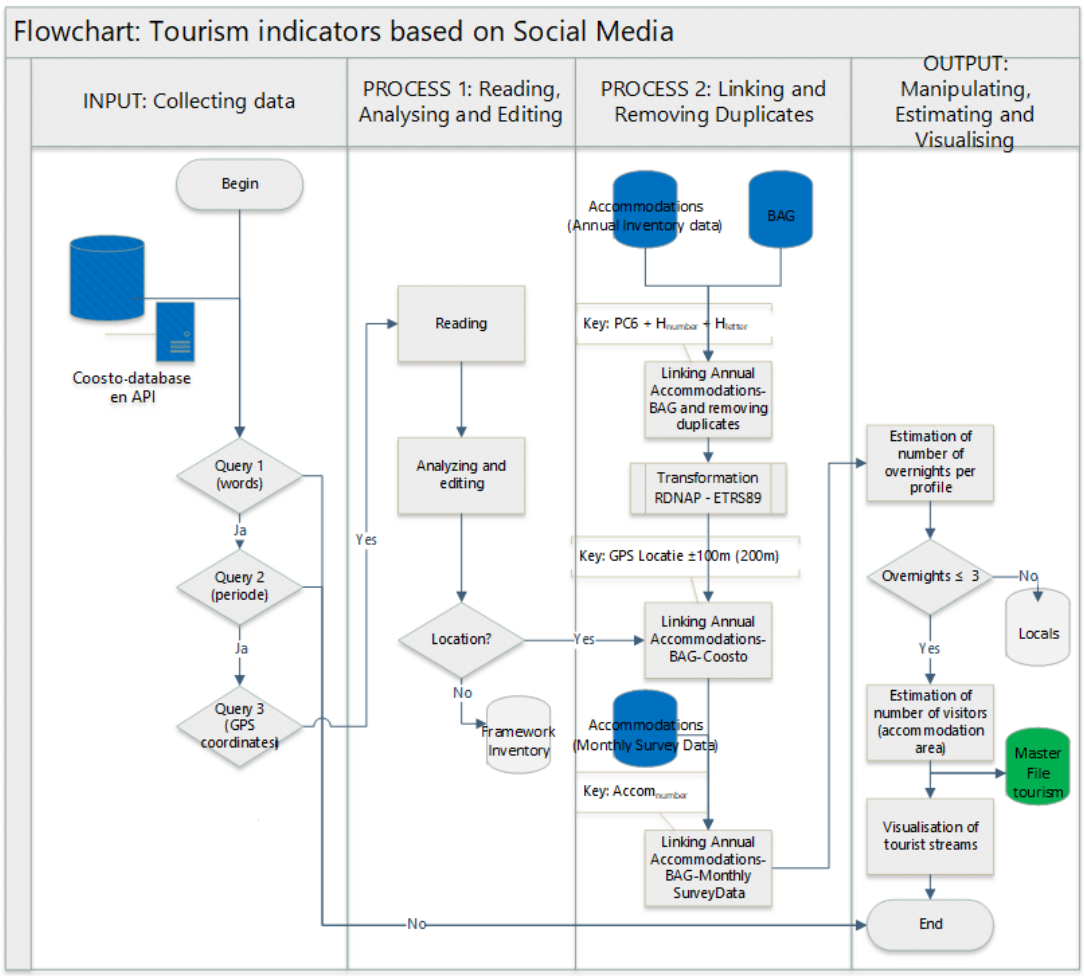
The last step in the process flow led to the Master file on which the analyses and visualizations in chapter 5 were conducted. This step included, among others, some data manipulation of names, changes in the order of columns, as well as the development of a classifier to discriminate between profiles of persons that are either residents or visitors. Moreover, this file contains also the estimations made of the number of visitors and the number of overnight stays.

The Master file has eventually three statistical units:

1) the social media messages from the Coosto-database;
2) the social media profiles from the Coosto-database; and
3) the accommodation establishments from the TAS.

11) The house letter is not presented in the TAS.

Figure 4.3 Diagram of the process flow



Source: Statistics Netherlands

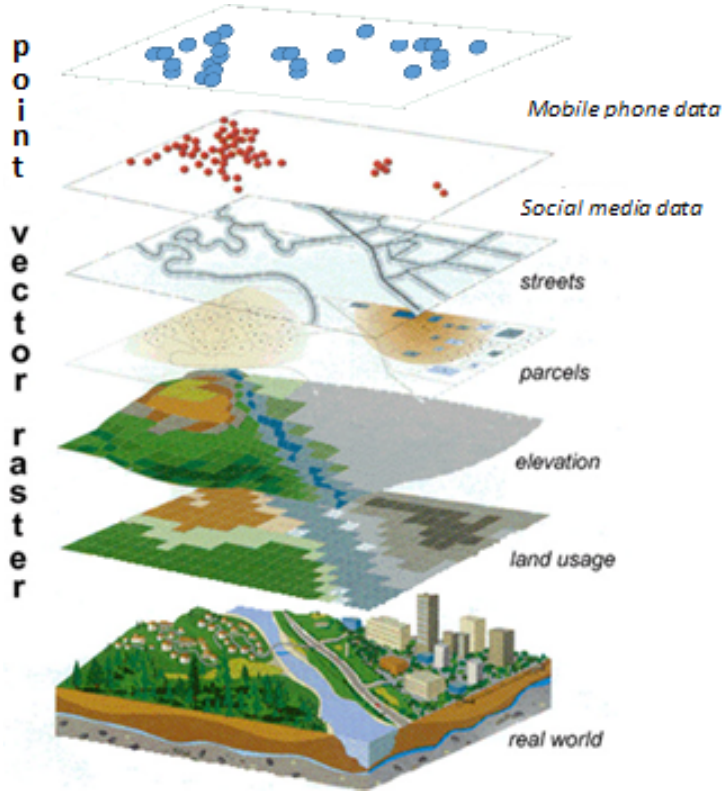
Visualizations

To visualize the data, two open source software environments were chosen: Quantum GIS and R. These visualization and analysis tools rely on a layers-approach. This approach is based on the idea that the real world can be decomposed into raster-, vector- and point-layers.

Typical examples of raster-layers are grids or matrices of cells containing information on land use, elevations or parcel information. The vector-layers include for example roads networks, whereas the point-layers refer to single points. This layers-approach is shown in Figure 4.4.

The raster- and vector-data on provinces and cities of the Netherlands come from the department of Regional statistics of Statistics Netherlands. The vector- and point-data come from the Master file and the raw data from the project itself, respectively. So, the Master file is a result of linkages between the databases from Dataprovider, Coosto, the BAG and the TAS.

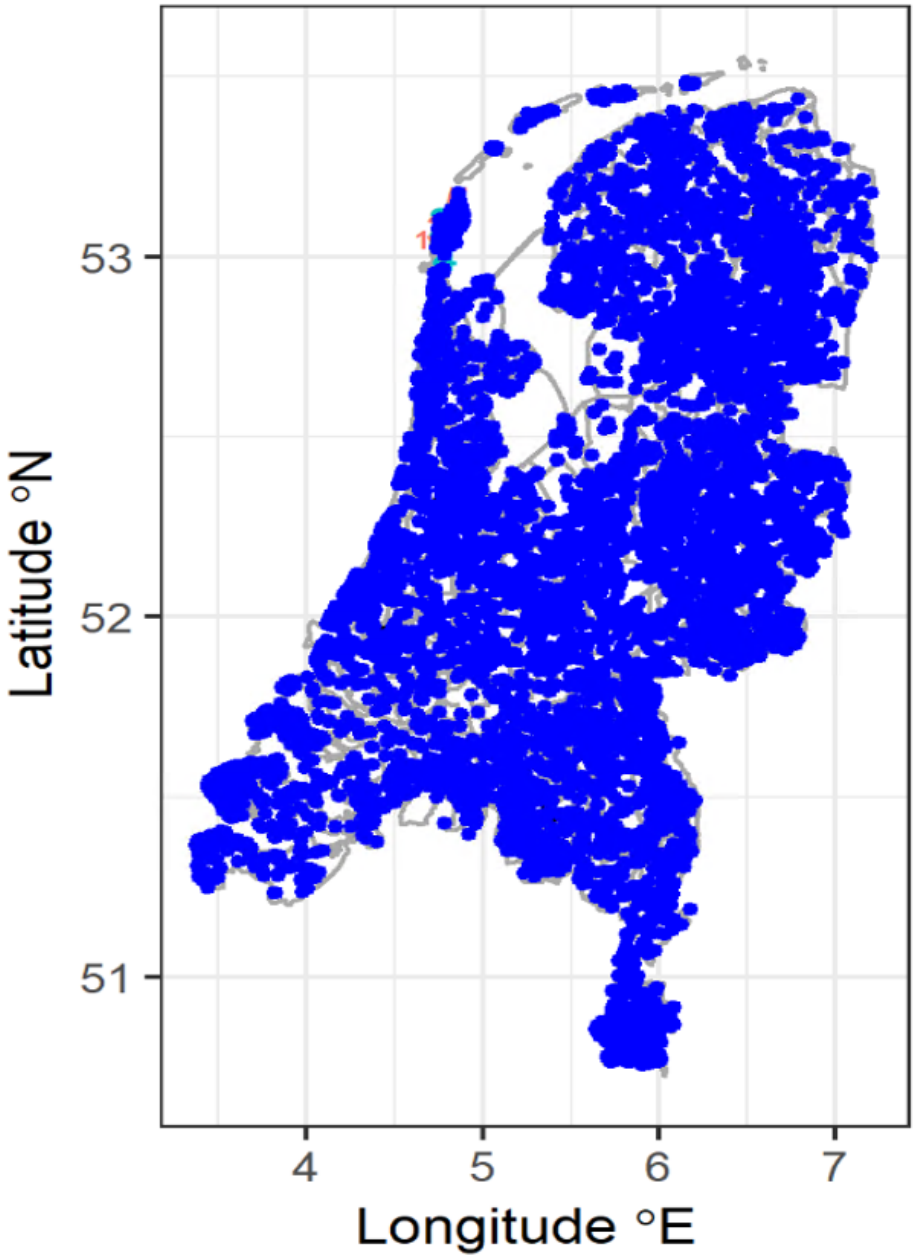
Figure 4.4 Diagram of GIS-layers extended using GPS-data from social media- and mobile phones



Source: Cloudfront.net and Statistics Netherlands

For the visualizations the software R was used. An illustration of the use of R is provided in Figure 4.5a. This map presents an overview of the population of tourism accommodation establishments in the Netherlands, consisting of hotels, holiday camps and holiday houses available in the TAS (2017). This figure shows basically that the population of tourist accommodation establishments of Statistics Netherlands is spread all over the Netherlands. No discrimination is made for the size of the accommodation (number of bed places).

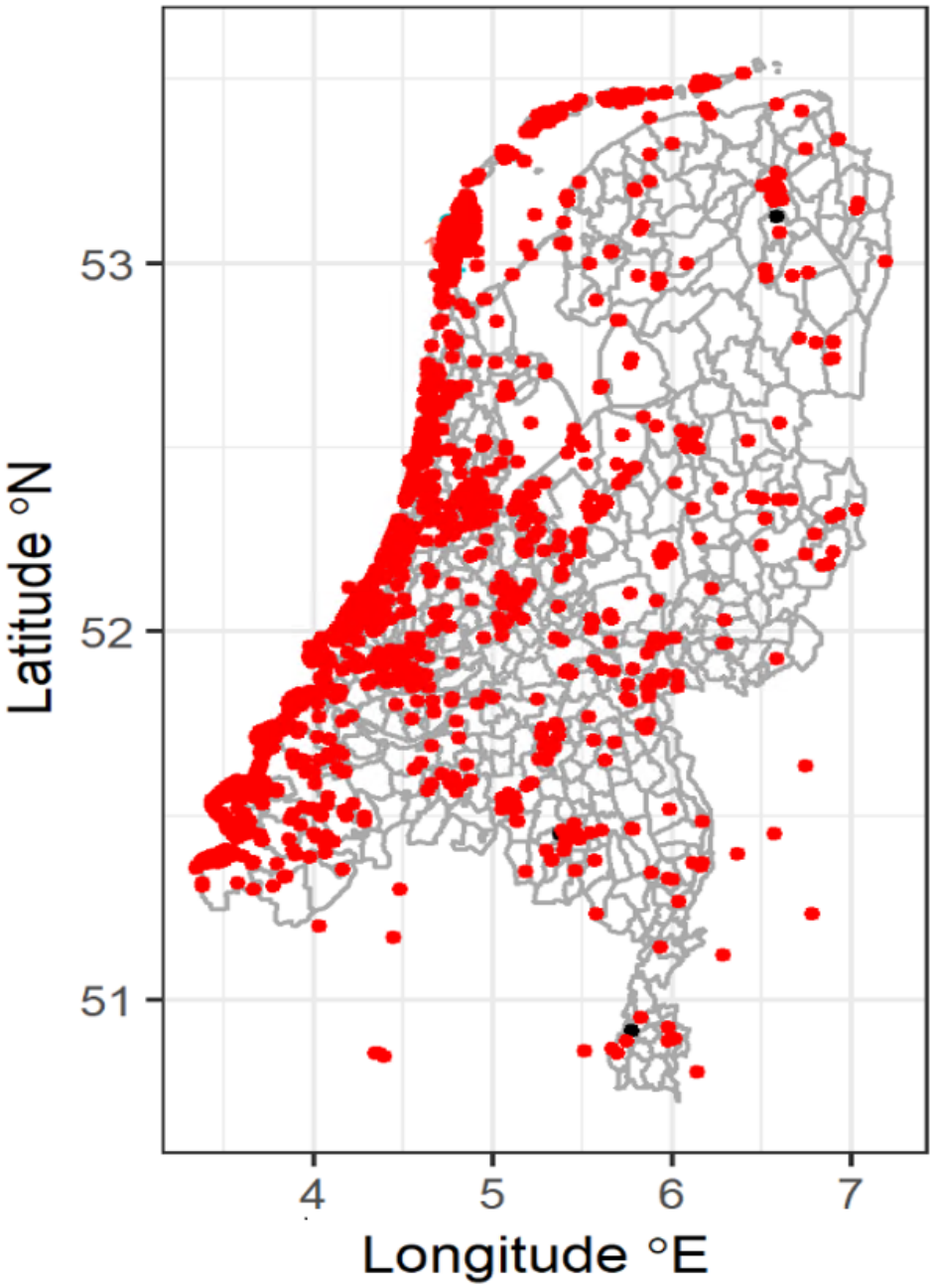
Figure 4.5a Overview of population of tourism accommodation establishments of Statistics Netherlands



Source: Statistics Netherlands

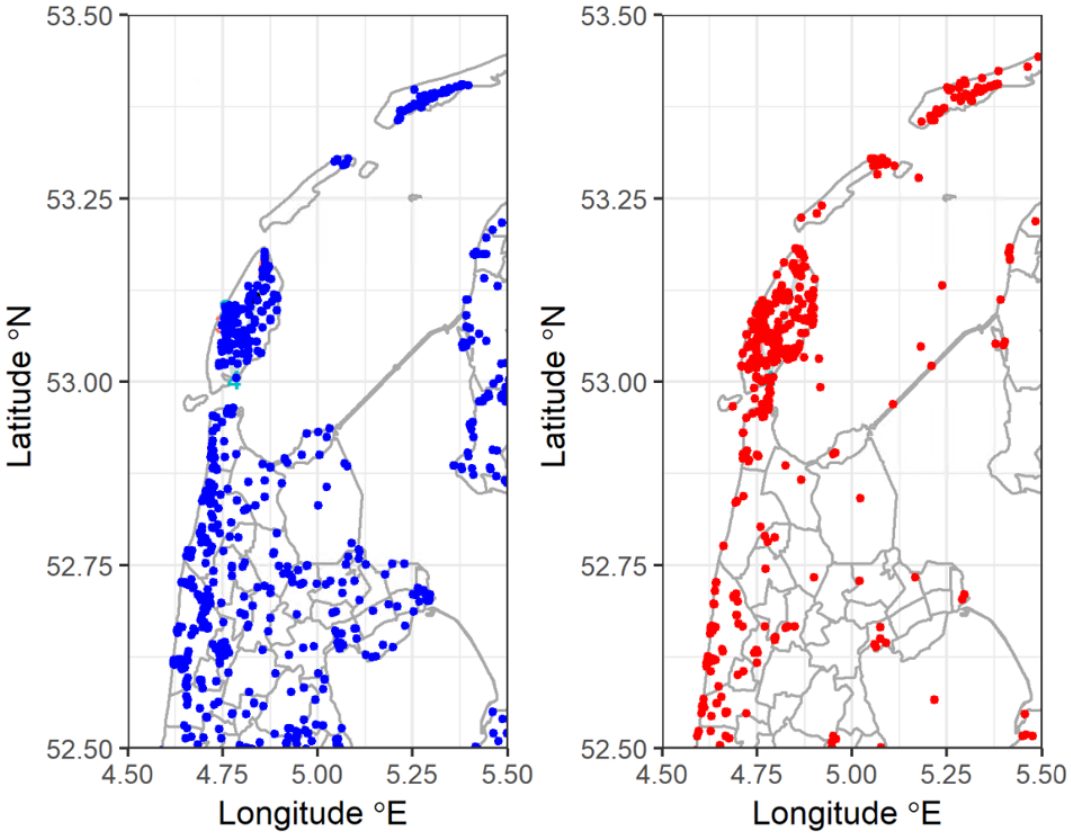
The map in Figure 4.5b also gives an overview of the geo-location of all social media messages generated in the period 12 August 2016 until 7 July 2017 which were obtained by using the filter "Texel (OR strand OR beach)" in the database of Coosto. As expected, some of the messages originate from Belgium and Germany. Figure 4.6 zooms in on the data of Figure 4.5. Figure 4.6 focusses on some of the Wadden islands (Texel, Vlieland and Terschelling), and the cities of Den Helder, Schagen and Hollandse Kroon.

Figure 4.5b Overview of social media messages with a geo-location generated using select 2 (Texel), 12 August 2016 -7 July 2017)



Source: Statistics Netherlands

Figure 4.6 Detail of Figure 4.5. of tourism accommodation establishments (left) and social media messages (right), based on the Wadden-islands and part of the province Noord-Holland

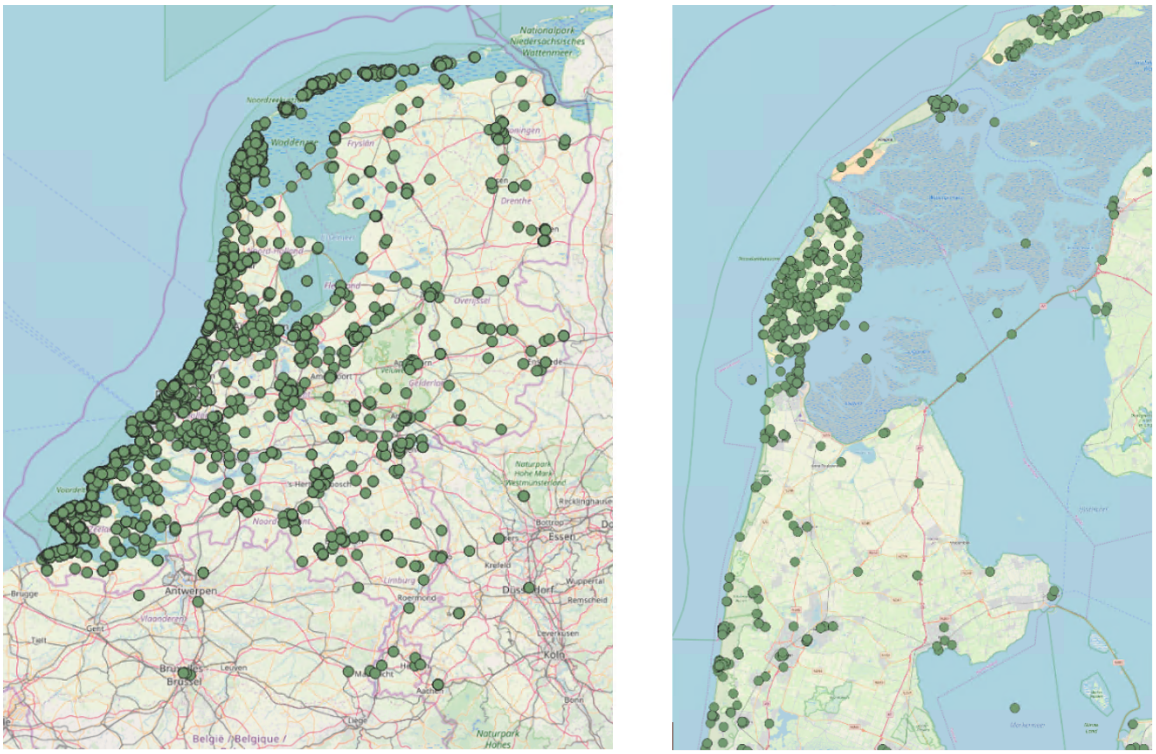


Source: Statistics Netherlands

Besides the package R, also QGIS can be used to make data-visualizations. In Figure 4.7 the raster and vector-layers of OpenStreetsMap (OSM) and the point-layer for the social media data are used. OSM-data combine raster- and vector-data to represent infrastructure information such as roads, water bodies but also other geographical limits. The combination of all these data sets provide a quick overview of the tourism accommodation establishments and social media messages in a geographical context. For instance, 4.7 (right) shows that some of the social media messages seem to come from recreational ships sailing on the Wadden and the IJsselmeer.

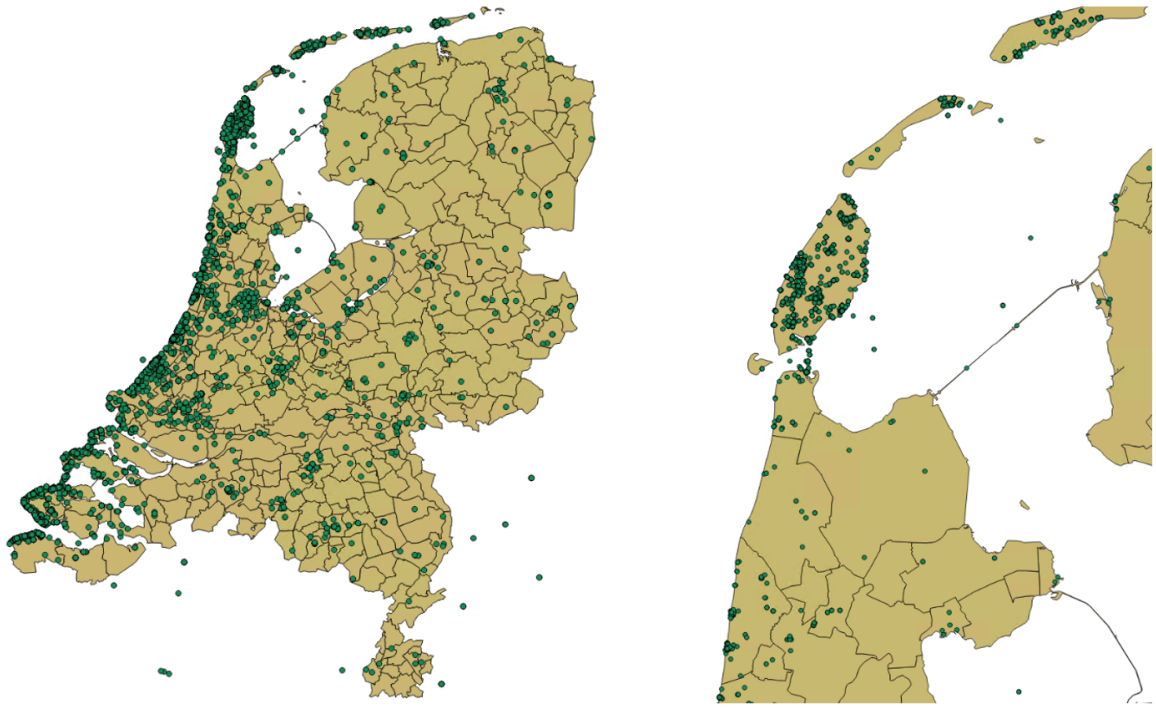
The identification and proper conversion of the data to the Coordinate Reference System (CRS) are performed using QGIS. Notice that Figure 4.8 is made using the Geo-data services of Statistics Netherlands in QGIS. This allows to control confidentiality and privacy of the data. QGIS has also filtering features to report aggregated outcomes in specific areas.

Figure 4.7 Overview of tourism accommodation establishments (left) and social media messages (right), using QGIS 3.4.10 on a laptop (OSM layer, CRS: WGS 84 / Pseudo Mercator, ID EPSG 3857)



Source: Statistics Netherlands

Figure 4.8 Overview of Figures 4.5 and 4.7 using QGIS 2.8.13 in the virtual machine of Statistics Netherlands (Geo-layer instead of OSM, CRS: Amersfoort / RD New, ID EPSG 28992)



Source: Statistics Netherlands

5 Results

In this chapter, the main results of this exploratory study are presented. Because of the mobility of visitors, special attention is given to tools with visualization capabilities. For the analysis only those data sets were used that are described in Table 4.1 in the previous chapter. A main part of the analysis was carried out on data related to the coastal area of The Netherlands, including the island of Texel (somewhat less than 20.000 social media messages). The data-mining was performed on social messages whose geo-location (GPS-coordinates) belong to this area. Social media messages which related to the coastal area, but had no geo-location were not taken into account. This means, among others, that a (substantial) part of the visitors to the coastal area were not in the data set and that therefore there may be an underestimation of the results. However, given the exploratory purpose of this study, this underestimation was accepted. More attention was given to general patterns and feasibility aspects rather than precision and bias. Nonetheless, at a later stage, outcomes were compared with the monthly results of the TAS, i.e. the number of guests and overnight stays in tourism accommodation establishments in the Netherlands.

The main outcomes of the analysis focused, among others, on:

- Distinction between residents and groups of visitors, e.g. excursionists and tourists;
- Sentiment analysis;
- Cluster analysis on points of interest (POI's);
- Travel routes and travel distances;
- A comparison with the monthly results of the TAS.

5.1 Representativeness and other methodological issues

Although not a subject of research in this study, representativeness should be mentioned here first. Maybe a clincher, but the uncertainty about the representativeness and validity of the end results, is a clear limitation of the use of social media data for tourism statistics or any other statistic. This also includes, for example, issues like the effects of retweeting, impacts when the emphasis is laid on one specific social medium, consequences when the number of cases is reduced for more detail and gaps in the information.[12)] However, when figures are structurally produced, there must be clarity about the representativeness and validity. Therefore, for the credibility of social media research, more emphasis should be placed on this issue. Research based on social media data must be put on a firmer conceptual and methodological foundation than is the case now. In the meantime, social media research has added value as additional information to other (big data) sources. But it should be clear that statistics based on social media data can only be seen as beta-indicators.

[12)] For example, persons who post social media messages do not do that all the time. So, in their travel route there can be missing information.

5.2 Distinguishing social media messages from visitors

Setting aside the issue of representativeness, a first question is whether social media messages from visitors can be distinguished from social media messages from other groups, like residents and businesses? Is this possible at all or do we have to live with a certain amount of noise in the end results? The conclusion of this study is that there are possibilities to get more grip on this issue. But it must be realized that there will always be some noise in the end results, or sometimes a lot of noise. As said, this does not make the end results less valuable. They should be seen as indications. The value of the end results can be increased when they are compared with or added to results of other (big) data sources, like mobile phone data or data collected through GPS-trackers. Noise in the end results of course already occurs when a distinction is made between social media messages from businesses and social media messages from persons. Social media messages posted by businesses can also relate to tourism, that is to business travel. When these social media messages from businesses are not taken into account, this information will not be in the data set.

A first step to distinguish better between social media messages from visitors and social media messages in general is to first distinguish between social media messages from businesses and social media messages from persons. Social media messages from businesses can be derived by linking the data to information on social media profiles of businesses. For Dutch businesses this can be done, for example, through a combination of data of a company, like Dataprovider, and the Business register of Statistics Netherlands. This method is already described in paragraph 4.1. The second step is then to distinguish between social media messages from visitors (travelling out of their “usual environment”) and social media messages from residents. This can be done if the messages of a profile can be followed for a longer time (e.g. based on origin and destination matrices). The ‘home’ or ‘usual environment’ of that person, inside or outside the area under research, can then be detected. See the next paragraph for some elaborated methodological examples. The pre-requisite is that a geo-location and a timestamp are available in the social media messages of that person. This is comparable with research based on mobile phone data.

A second, more rough, option is to simply start from the total set of social media messages and select messages based on (a combination of) keywords or geo-locations that relate to a certain region, city, attraction or accommodation and presume that the resulting selection of social media messages are from visitors. This method can be refined and followed again by looking, per profile, at the timeline of the messages posted.

A third option is by working with techniques such as text mining and machine learning. For example, a classifier can be developed on the basis of the content of the social media messages to determine with a certain chance whether the message is from a visitor, a resident or a business. A pre-requisite is that there is a big enough training set of examples that are true cases. In other areas classifiers often show good results. In order to properly develop this method, a separate study is actually desired.

A fourth option is not to look at the content of the social messages or the timeline, but more closely at the profiles of people who post social media messages. Besides the matching with other data sources with information on profiles, also here machine learning techniques and classifiers can be applied.

A final option for especially foreign visitors is simply to look at the language of the social media messages. For this exercise already software packages exist. To remove noise, such as border traffic or labor migrants, one could follow the timelines of the messages of these profiles again. A distribution by country of residence remains difficult. Many people post their messages in English. And some countries use the same language. Besides, an Englishman who lives in the Netherlands, and is a domestic visitor, will also post his messages in English. But, for example, by looking at the distinction between messages with a Dutch and non-Dutch language one can get some picture of the shares of domestic and inbound visitors.

So, in principle it is possible to get more grip on the distinction between social media messages from visitors and other groups. However, the results will never be 100 percent reliable. There will be (much) noise in the end results. Most promising is working with a timeline per profile of origin and destination matrices based on social media messages with a timestamp and a geo-location. The disadvantage of this option is that the number of cases can quickly decrease. It is also unclear what this does with the representativeness of the end results. Also promising seem machine learning techniques with classifiers or a combination of options, including the use of auxiliary information from other data sources.

As with the issue of representativeness, it is also important here to ensure a better conceptual and methodological foundation. As said, such a foundation is certainly not yet available.

5.3 Analysis based on origin and destination matrices

Following the previous paragraph, in this section, some empirical examples are presented to distinguish between the social media messages of different groups of visitors. As tourism relates to mobility, the main method is to use timelines per profile. Timelines per profile can be generated from the social media messages that contain a timestamp and a geo-location. As shown in the previous chapter this only counts for a relatively small part of the social media messages. In the literature, typical “migration pattern” plots have been used to visualize migration and tourist flows (Hort et al, 2018; Habib and Krol, 2017). Migration flows are used to show the directions of the displacements of people. For this experiment the data of selection 2 (see Table 4.1.) were used.

The timelines in this study with origins and destinations of visitors were determined by using the geo-location data and timestamps in the social media messages of the selected data set. Based on these timestamps and geo-locations, data were clustered and chronologically sorted per profile to a set of timelines (cases). See Table 5.1 for an example of a timeline with multiple trips of one profile (“Camera”).

Additionally, the average of all GPS-coordinates per profile was computed. This average (mean) was assumed to be the “home” or the “usual environment” of each set of data of one profile. Because this average is still quite arbitrary (we do not know a priori if this is true) also the first quantile of the GPS-coordinates was computed.

Table 5.1 Example of a fictitious timeline of social media messages of a profile ("Camera"): multi-trip

Profile name	Sentiment	GPS-latitude	GPS-longitude	Message	Date message posted
Camera	+	52,033298	4,433330	Schelp ¸ #vlieland #	17-6-2017 15:43
Camera	+	53,299110	5,071467	Vlieland, wat ben je	17-6-2017 14:04
Camera	NA	53,066666	4,800000	Stoer! [#texel] #lifec	13-6-2017 22:13
Camera	NA	53,066666	4,800000	Family love! #lifeof	13-6-2017 22:11
Camera	+	53,066666	4,800000	Blij op het [strand]	13-6-2017 22:10
Camera	+	53,401100	5,331400	Genoten! #terschell	28-5-2017 19:30
Camera	+	53,401100	5,331400	Alone #world #tersc	28-5-2017 11:42
Camera	+	53,401100	5,331400	Throwback to these	4-5-2017 19:17
Camera	+	53,489201	6,202200	#heimwee¸ Wat was	13-3-2017 10:40
Camera	+	53,490318	6,149781	Preview van een gev	27-1-2017 12:43
Camera	+	53,490318	6,149781	Wat is het hier wee	27-1-2017 11:43
Camera	NA	53,223591	4,866958	"Wat de diepste inc	25-1-2017 14:22
Camera	NA	52,033298	4,433330	Vlieland! Wat heb i	9-1-2017 15:17
Camera	+	53,296387	5,074056	Nog eentje @jannek	26-11-2016 17:50
Camera	NA	53,296421	5,074077	"Eens zal het weer g	25-11-2016 15:08
Camera	NA	53,296421	5,074077	Goedemorgen Vliela	25-11-2016 10:21
Camera	NA	53,295609	5,067370	Zo mooi! En ook nog	4-11-2016 15:23
Camera	NA	53,296421	5,074077	1 van m'n favoriete	4-11-2016 14:55
Camera	NA	53,398808	5,254393	#wolken #loveit #ter	30-9-2016 18:07
Camera	-	53,399601	5,262526	Sorry! Terschelling	30-9-2016 17:58
Camera	NA	53,489201	6,202200	Schiermonnikoog is	12-9-2016 13:55

Source: Coosto

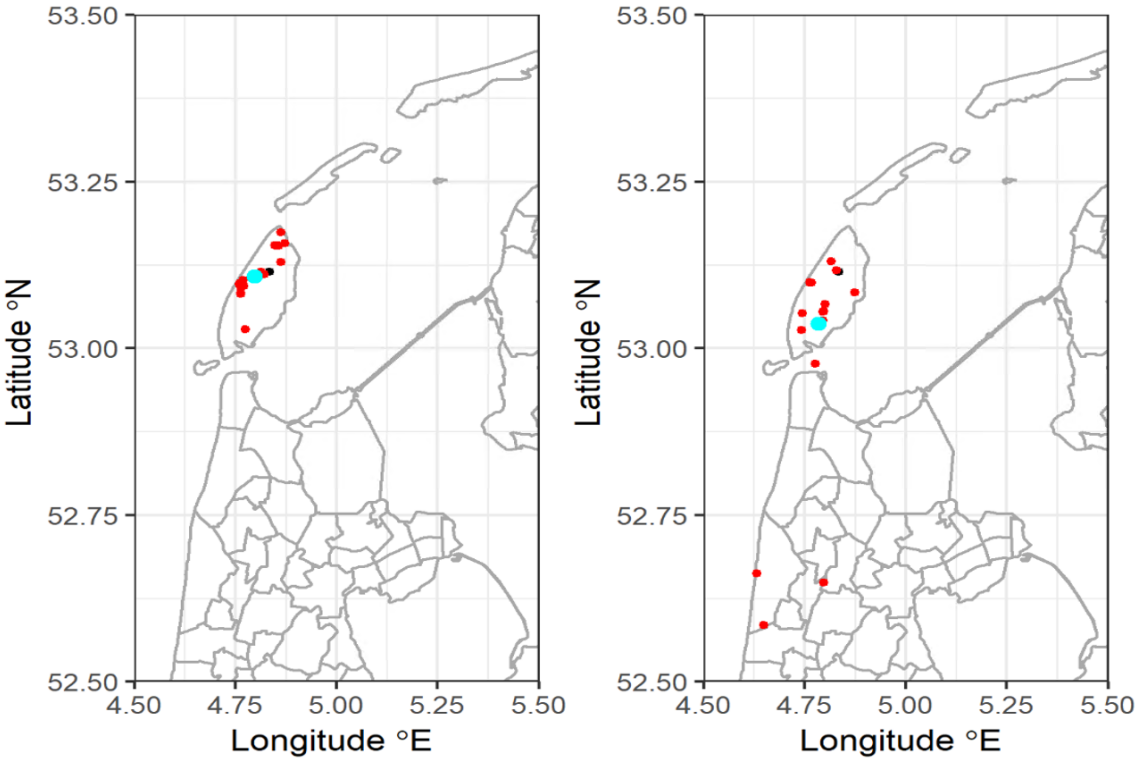
Figure 5.1 shows two examples. The figure on left shows the data that belong to a local resident of Texel. Social messages are represented with the red-colored dots and the cyan-colored dots show the average GPS-location of the social media profiles. The figure on the right shows a probable visitor, a non-resident of Texel, although the cyan dot is on the island of Texel, indicating that it could also be a resident.

To try out another method to automatically derive "the home" of a visitor, also the first quantile of the GPS-coordinates (blue circle) of the profiles was added. Figure 5.2 shows two visitors to Texel. Notice that the cyan-dot (average) on the right-side picture of Figure 5.2 falls in the sea while the first quantile (blue circle) shows a more plausible "home" of the depicted profile name.

Finally, the time between the messages was computed for each profile (from home and back and including cases with multi-trips within a year). Profiles with messages which remained within a timeframe of 18 hours were seen as excursionists and those above this limit of hours were considered tourists, at least if the messages were posted in the interval of 1 to 14 days per profile.

So, these methods and examples show some possibilities to make a distinction between residents, excursionists and tourists with at least one overnight stay. In another experiment on travel distances the "home" of a profile (person) was detected by looking at the place where messages were sent between 10:00 in the evening and 6:00 in the morning. In future research these kinds of algorithms or classifiers to detect trips outside

Figure 5.1 Residents and visitors (Locals of Texel: n = 18, Visitors of Texel: n = 74)



Source: Statistics Netherlands

the "home" or the "usual environment" must be further refined or elaborated. Here rule-based algorithms were used. Another option is to use data-driven algorithms. However, in that case training sets with true cases are needed. Much can also be learned from methodological developments in research on mobile phone data.

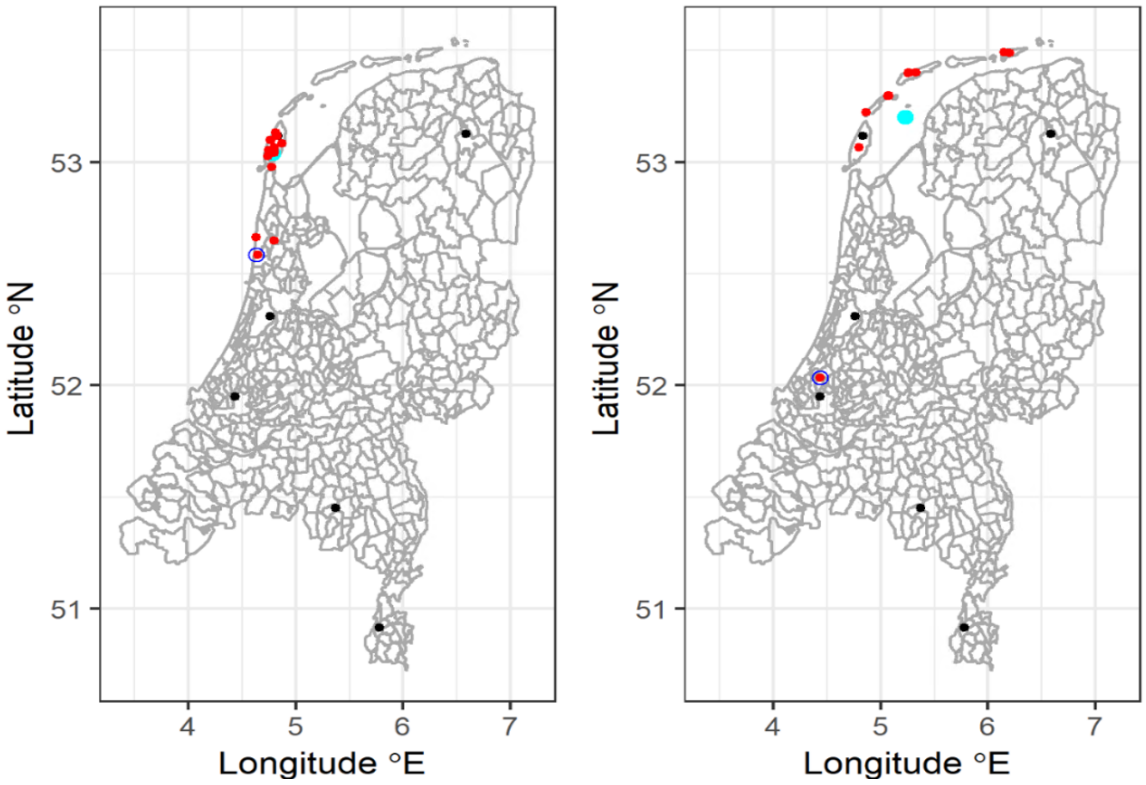
5.4 Sentiment analyses

Social sentiment analysis, the automated extraction of expressions of positive or negative attitudes from text, has received considerable attention from researchers during the past decade (see, for example, Taj, S. et al, 2019 and Alaei, A.R. et al, 2017). Sentiment analysis determines whether a person (or a profile) is messaging in a positive, a negative or a neutral way. To derive the type of sentiment, algorithms apply natural language processing (NLP) to the content of social media messages that contain positive or negative attitudes towards people, organizations, places, events and ideas. Sentiment analysis can be seen as a proxy for 'satisfaction'. To determine the sentiment mostly a set of positive or negative words (concepts) are used (lexicon-based approach).

In the first instance, in this study the sentiment indicator[13)] in the extracted data of Coosto was used. The indicator of Coosto is based on a lexicon-based approach. In this paragraph the data of selection 2 (see Table 4.1) were used.

13) For this purpose, a proprietary variant of a sentence level-based classification approach is used (for an overview, see Pang and Lee, 2008). The approach strictly determines the overall sentiment of the combination of words included in each message (source: Daas and Puts, 2014).

Figure 5.2 Excursionists and tourists (Visitors of Texel n = 74 and n = 61)



Source: Source: Statistics Netherlands

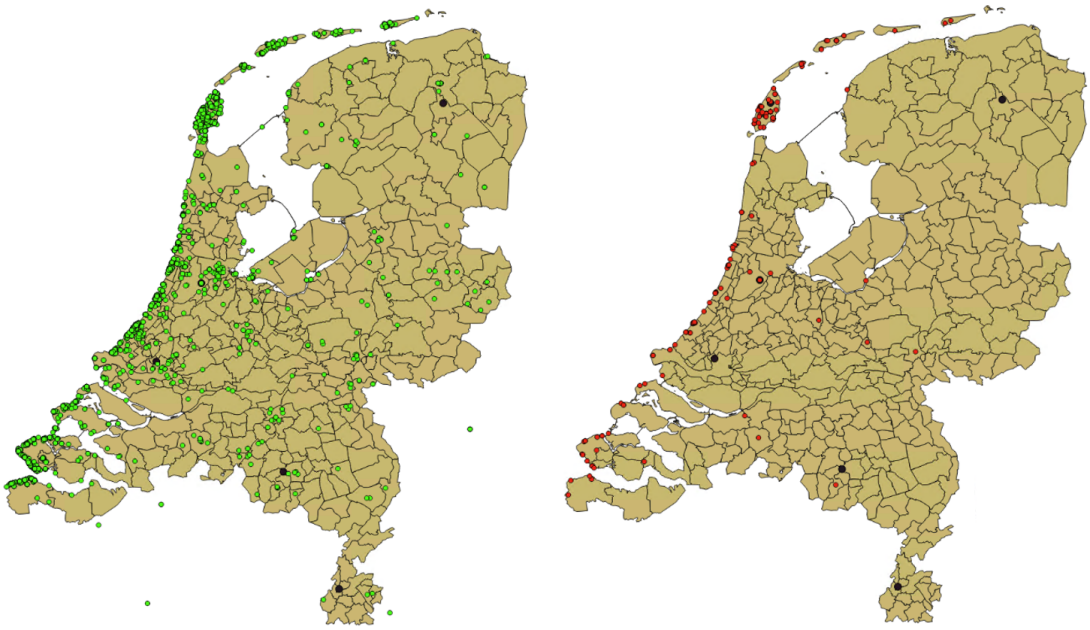
To produce counts of sentiments in social media messages for an area of interest, vector-points (social media messages with their sentiment) were linked to the desired polygons, i.e., closed vector-data representing the boundaries of cities or touristic regions.[14)] Figures 5.3a and b show the positive and negative sentiment, respectively. Messages whose sentiment was not filled in are not shown here. The share of social media messages without a sentiment indicator was rather high (82 percent). Part of these messages could be considered as neutral. However, it may occur that visitors may use sometimes emoticons only or hash tags in their posts.

Depending on the information requirements of the users, the data can be aggregated to other levels of detail, such as specific areas, provinces or related to attractions, events etc. In Tables 5.2a and b, the positive and negative sentiments per tourist region (TG) and per city (GM) are presented. The coastal region showed the highest number of positive social media messages: 96 percent of the social media messages were positive about their visit to the Dutch coastal region. In terms of cities, Texel, Veere and The Hague hold the highest number of positive messages. In fact, more than 90 percent of the messages from these cities were positive. See also Table 5.2a. As for the regions and cities with the highest number of negative messages, also the coastal region ranked first. The number of negative social media messages was, however, about 20 times lower than the number of positive messages. The same holds for Texel. The number of negative messages was about 6 times lower than the number of the positive messages. For the case of The Hague, the proportion of positive versus negative messages was 33:1. See Table 5.2b.

14) The Netherlands uses the classification “touristic regions” which has 6 (TG) groups.

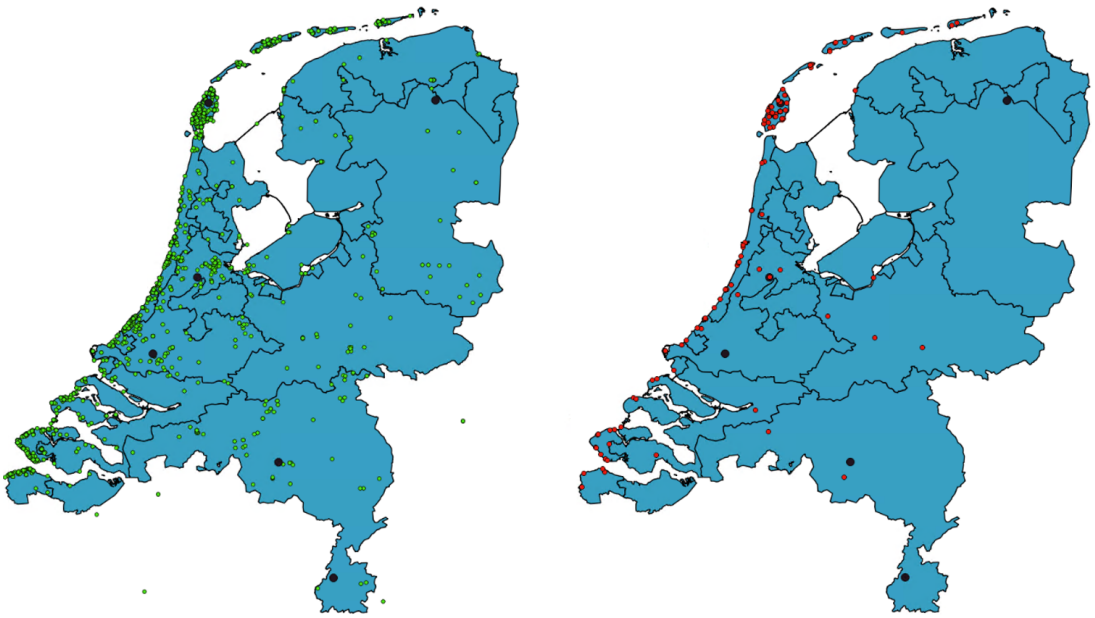
In fact, numbers of social media messages with a positive or a negative sentiment actually do not say very much. It is better to look at relative proportions, i.e. the ratio of positive and negative messages or the percentage of all the messages of that region or city (the sum of the positive, negative and neutral messages). The last indicator is presented in the last column of tables 5.2a and b. See also Figures 5.4a and b for relative percentages of positive and negative sentiments compared to those from the Netherlands as a whole.

Figure 5.3a Positive and negative sentiment per city (GM) ('o'= 5068, 'o' = 213; Total messages = 20.000, using selection 2 (Texel), 12 August 2016 - 7 July 2017)



Source: Statistics Netherlands

Figure 5.3b Positive and negative sentiment per touristic region (TG); (Total messages = 20.000, visitors of the island of Texel, 12 August 2016 - 7 July 2017)



Source: Statistics Netherlands

Table 5.2a Positive sentiment per tourist region and city (number of vector points > 150)

a) Per tourist region

Year	Code	Area name	Number of positive messages	% of all messages
2017TG1	TG1	Coastal region	3708	26%
2017TG6	TG6	Rest of Netherlands	830	25%
2017TG2	TG2	Water sport region	197	33%
2017TG5	TG5	Forrest and Moorlands of South Netherlands	71	30%
2017TG4	TG4	Forrest and Moorlands of North-East Netherlands	38	36%
2017TG3	TG3	Forrest and Moorlands of Middle Netherlands	28	27%

b) Per city

City code	Year	City name	Number of positive messages	% of all messages
GM0448	2017GM0448	Texel	1089	24%
GM0717	2017GM0717	Veere	759	26%
GM0518	2017GM0518	The Hague	563	25%
GM0575	2017GM0575	Noordwijk	305	29%
GM0473	2017GM0473	Zandvoort	191	27%
GM1676	2017GM1676	Schouwen-Duiveland	177	29%
GM1714	2017GM1714	Sluis	175	23%
GM0718	2017GM0718	Vlissingen	154	26%

Source: Statistics Netherlands

Table 5.2b Negative sentiment per tourist region and city

a) Per tourist region

Year	Code	Area name	Number of negative messages	% of all messages
2017TG1	TG1	Coastal region	170	1%
2017TG6	TG6	Rest of Netherlands	27	1%
2017TG3	TG3	Forrest and Moorlands of Middle Netherlands	3	3%
2017TG2	TG2	Water sport region	2	0%
2017TG5	TG5	Forrest and Moorlands South Netherlands	2	1%
2017TG4	TG4	Forrest and Moorlands of North-East-Netherlands	0	0%

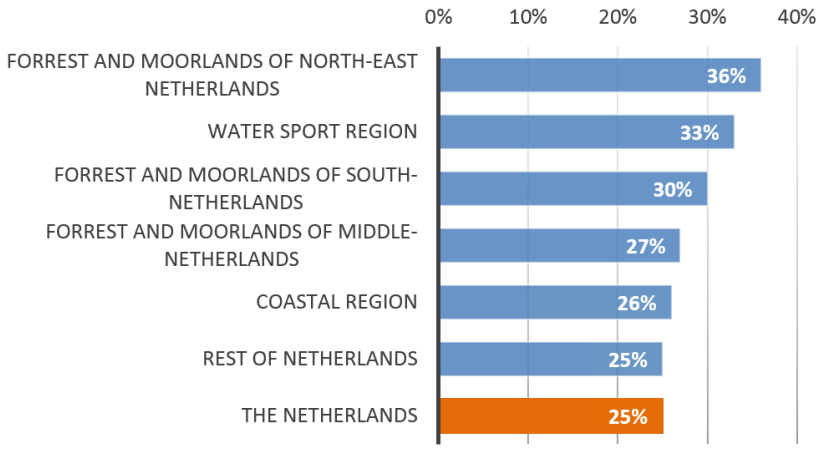
b) Per city

City code	Year	City name	Number of negative messages	% of all messages
GM0448	2017GM0448	Texel	98	2%
GM0518	2017GM0518	The Hague	17	1%
GM0093	2017GM0093	Terschelling	12	3%
GM0717	2017GM0717	Veere	11	0%
GM0473	2017GM0473	Zandvoort	9	1%
GM0575	2017GM0575	Noordwijk	7	1%
GM0718	2017GM0718	Vlissingen	5	1%

Source: Statistics Netherlands

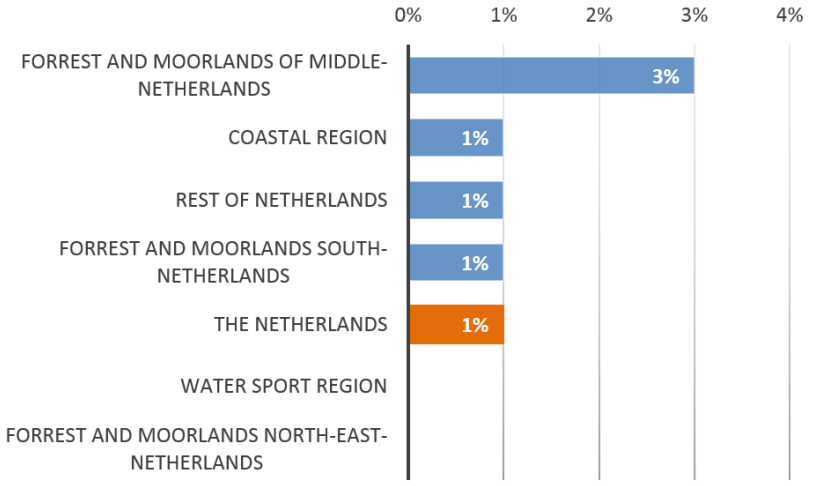
Besides a relative indicator, it is also wise to look at the content of the social media messages with a sentiment: is there anything in the messages that indicates a positive or negative sentiment? Techniques to analyze the content of social media messages can be very helpful here. Here a R-library "word cloud" was used. A word cloud is used to summarize the relative frequency of the words being used in the social media messages, in this case the messages of the Coosto-database for the Dutch Coastal area.

Figure 5.4a Positive sentiments of tourist regions compared to the total of the Netherlands



Source: Statistics Netherlands

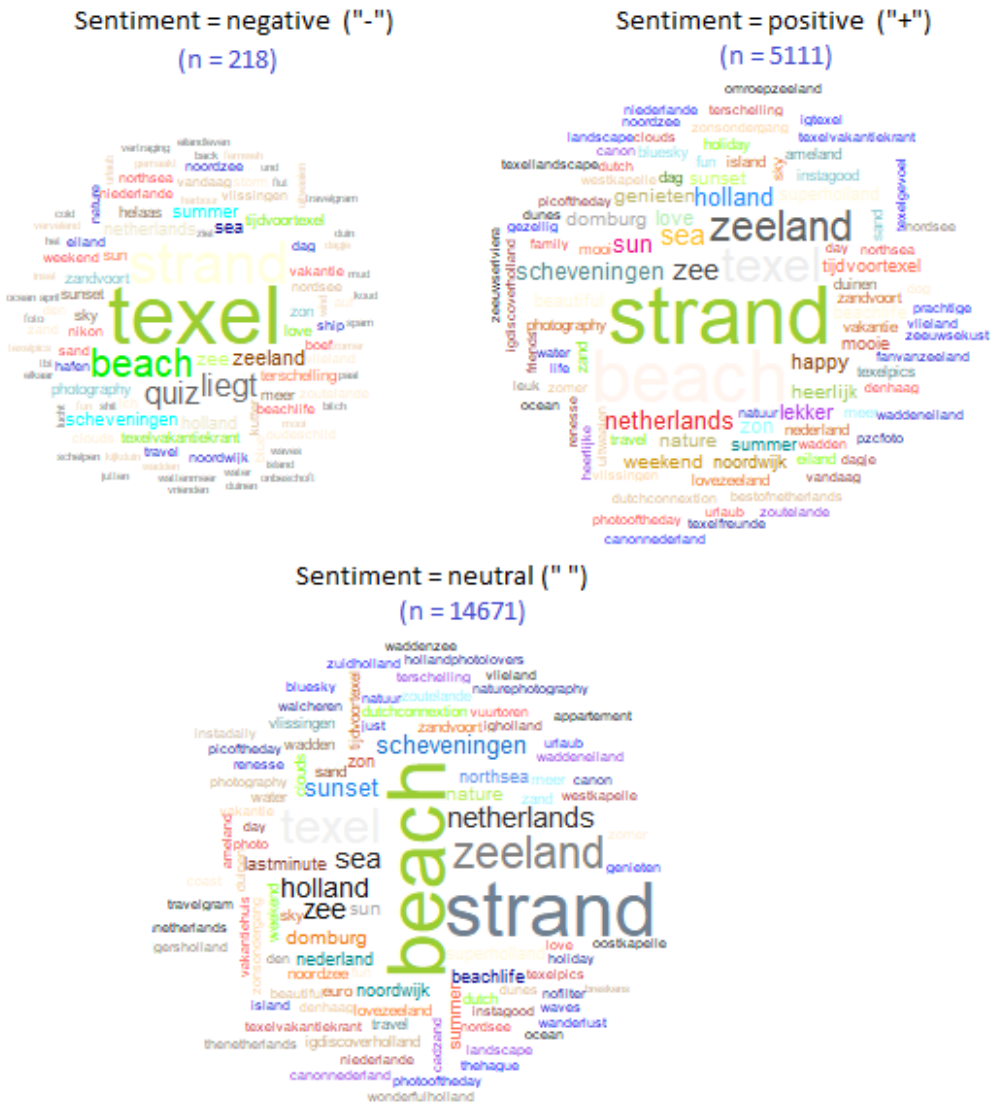
Figure 5.4b Negative sentiments of tourist regions compared to the total of the Netherlands



Source: Statistics Netherlands

Figure 5.5, for example, provides a snapshot of the words being used in the 20.000 social media messages of the Coosto-database, which were classified as "negative"(218), "positive" (5.111) or "neutral" (14.971). Also, a minimum frequency of 5 words was used to highlight the most important words. It can be observed that the number of positive messages surpasses those classified as negative with a factor 23 and that the neutral or not classified messages are 73 percent of all the messages.

Figure 5.5 Word clouds based on Coosto's sentiment indicators a) negative, b) positive and c) neutral. (relevancy of top 100 words, minimum frequency = 5)



Source: Statistics Netherlands

For the messages with a negative sentiment, the most predominant words were, as expected, the location or the names of cities (like Texel, Zeeland, Scheveningen and Zandvoort), then the attractions (like beach, sea and sun), a time reference (like summer, weekend, vacation). The sentiments expressed the words like "lying ("liegt"), "unfortunately" ("helaas") and "crook" ("boef"). This could indicate that people were cheated.

For the messages with a positive sentiment, the frequencies of attractions (like "strand/beach", "zee/sea", and "zon/sun") were more dominant than those of the locations (like "Texel" and "Zeeland") and the time reference (like "summer", "weekend", "day"). As for the sentiment, the most frequently used words were "happy", "beautiful" ("mooie" and "prachtig"), "fine" ("lekker") and "nice" ("heerlijk"). Also, words like "friends" and "freunde" (German word for "friends") appeared quite often. This could

Matrix 5.1 Experimental lexicon-based approach (extract, see appendix I)

Language			
Dutch	English	German	
bezoek	visit	besuch	
reis	trip	reise	
overnacht	sleep	nacht	
Countries	**Cities**		
Dutch	English	Dutch	English
België	Belgium	Amsterdam	Amsterdam
Duitsland	Germany	Den Haag	The Hague
Frankrijk	France	Rotterdam	Rotterdam
Positive sentiment		**Negative sentiment**	
Dutch	English	Dutch	English
zon	sun	regen	rain
warm	warm	koud	cold
lekker	tasty	ijskoud	freezing

Source: Statistics Netherlands

indicate that people travelled with friends.

The messages without a sentiment (neutral) showed the lack of related words in this context. This word cloud shows locations, attractions, accommodations ("holiday home" / "vakantiehuis") and also the time reference ("lastminute") and a specific time of the day ("sunset").

For the lexicon-based approach also our own experimental sentiment indicator was developed. Three sets of words in different languages (Dutch, English and German), that related to tourism, were used: language (blue columns), geography (orange columns) and tourism-related words that may represent positive or negative experiences (purple columns). See Matrix 5.1 and, for all words used, the table in Appendix 1.

The word clouds in Figure 5.6 show the results of the applied sentiment indicator. Figures 5.6a1 and b1 and 5.6a2 and b2 summarize the positive and negative sentiments, respectively, of the social media messages analyzed. In these two plots the sentiment is more visible than with the standard indicator of Coosto. The enumerated words mentioned in Matrix 5.1, and more elaborated in Appendix 1, seem to deal better with social media data in the domain of tourism. Striking is that the word clouds on the right (negative sentiment) show a few words that are considered positive like "enjoy" ("genieten"), "beautiful" ("mooie"), "love" and "wonderful". This shows that a sentiment indicator must be made robust for the domain in which it will be applied. Such that it can provide ultimately adaptive net value scores.

Figure 5.6 Word clouds based on own classifier based on a) language, b) geography and c) sentiment. (Relevancy top 100 words, minimum frequency = 5)

Source: Statistics Netherlands

So, sentiment analysis can say something about the 'satisfaction' of visitors. If possible, visitors can be distinguished in excursionists and tourists. However, the results of sentiment analysis are not easy to interpret directly. The number of messages with a positive and negative sentiment do not say very much. At least, the results have to be put in context (relative proportions). For a lexicon approach, it may also be important to use positive and negative words that relate to the domain under research, in this case tourism. Furthermore, it is also very informative to obtain more information by using word clouds of the social media messages that relate to positive and negative sentiments. These word clouds can also say something about the time, the subject, events, activities and actions found in the social media messages. The study on social media and business statistics showed that this analysis can be done on a quite detailed level.

In this experiment relatively simple methods were used. For further research also more advanced languages processing techniques could be researched, or techniques where the context of the words are taken into account or where the use of methods like machine learning are integrated (see, for example, Agarwal, B. and N. Mittal, 2015).

Although not carried out in the context of this study, the use of word clouds for sentiment analysis can also be replaced by research on the main activities that are carried out by visitors of a city or region. This can also be combined with routes that visitors follow (see paragraph 5.6).

5.5 Clustering based on points of interest (POI)

This paragraph zooms in on the relationship between the sentiment in social media messages and points of interest (POI's), such as attractions and tourism accommodations. In this paragraph the data of selection 2 were used again. This exercise was also carried out to get more experience with (Quantum) Geographic Information Systems (QGIS) in combination with external data, like a selection of social media messages.

Insights in the sentiment or the satisfaction about POI's is also important to understand why visitors are attracted to certain places or locations. For the location and clustering of the social media messages (and their related profiles) a buffer of 1 kilometer around every POI was used. It was assumed that the social media messages within this buffer contained information on features connected to the popularity of a POI. Table 5.3. lists the POIs chosen for this experiment.

Table 5.3 POI's involved in the selection of social media messages.

Attraction name	Code	Latitude	Longitude	City (Province)
Pretpark Duinrell	Duinrell	52.14473	4.383029	Wassenaar (Zuid-Holland)
Ecomare Texel	Ecomare	53.07819	4.744547	Texel (Noord-Holland)
Keukenhof	Keukenhof	52.27126	4.546365	Lisse (Zuid-Holland)
Anne Frank Museum	A.F.Museum	52.37522	4.883977	Amsterdam (Noord-Holland)
Red Light District	Wallen	52.37186	4.895861	Amsterdam (Noord-Holland)
Renesse	Renesse	51.73129	3.771445	Schouwen-Duiveland (Zeeland)
Zaanse Schans	Zaanse Schans	52.47099	4.809830	Zaanstad (Noord-Holland)
Giethoorn	Giethoorn	52.73716	6.073166	Steenwijkerland (Overijssel)

Source: Statistics Netherlands

After aggregating the data of the social media messages in each buffer around each POI, only three POI's seem to contain enough useful data (53 messages): Renesse (26 messages) and Ecomare (24 messages). And, also for the Red-Light District in Amsterdam a small number of messages (3) was found. Data were classified as coming from excursionists. The main data sources were Instagram (39 messages) and Twitter (14 messages). The data are generated across the year, namely spring (19 messages), autumn (16 messages), summer (13 messages) and winter (5). With caution, It could be said, that these POI's seem to be equally attractive to visit in the spring, summer and autumn, but not in the winter. Most messages were posted on Sunday (15 messages) and on Monday (13 messages).

According to the indicator of Coosto the sentiment towards the three selected POI's together was positive (9 messages) and neutral (44 messages). There were no negative messages identified. For both Renesse and Ecomare the sentiment was predominately positive. The number of social media messages related to the Red-Light District in Amsterdam were too small to say anything about it.

Figures 5.7a and b provide another example and focus on selected POI's like airports and attractions as well as on the coastal area (selection 2). Figure 5.7a also includes social media messages from profiles outside the Netherlands or in the sea (probably ships). For a comparison exercise later on in this chapter, the messages of these "foreign" profiles were filtered out, resulting in Figure 5.7.b. However, the timelines of social media messages of these types of profiles are very useful to identify foreign visitors to the Netherlands.

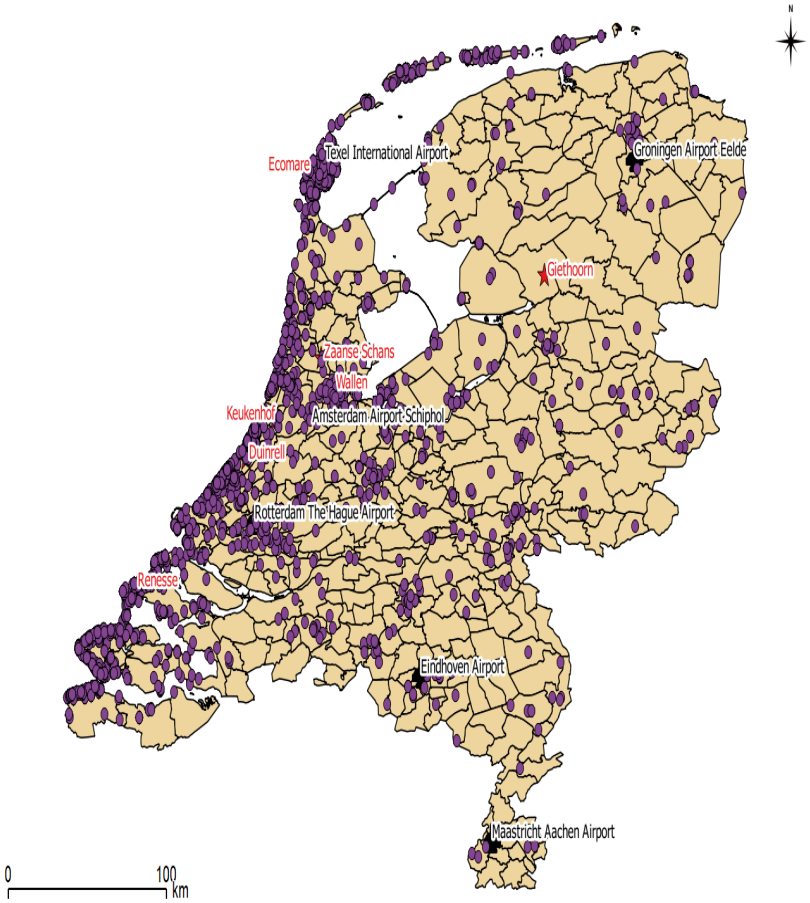
Figure 5.7a Social media messages collected based on a filter for the coastal area and others points of Interest (airports and attractions)

Source: Statistics Netherlands

QGIS permits the use of POI's to identify, filter and analyze social media messages or profiles in areas of different sizes, like 1 or 2 kilometer around our examples of Renesse, Ecomare and the Red-Light district, but also Duinrell. Figure 5.8 shows the geographical position of these POI's.

The word clouds of the content of the social media messages that are posted within the buffers of Renesse and Ecomare are shown in Figure 5.9. The right word cloud shows, among others, that the main attraction of Renesse looks to be the "beach" ("strand") and the "sea" in the "summer", also connected to the "sunset". The fact that many Germans visit Zeeland is reflected in the German words "sommer" and "sonneuntergang". The word cloud of social media messages of Ecomare has to be treated with caution as there are just a few messages available. Nevertheless, the main attraction of Ecomare are the "seals" ("zeehond"). In both word clouds there are a few qualifying adjectives included that reflect a more positive sentiment.

Figure 5.7b Selected POI's and the filtering data, after the removal of social media messages of profiles outside the Netherlands (sea and foreign)



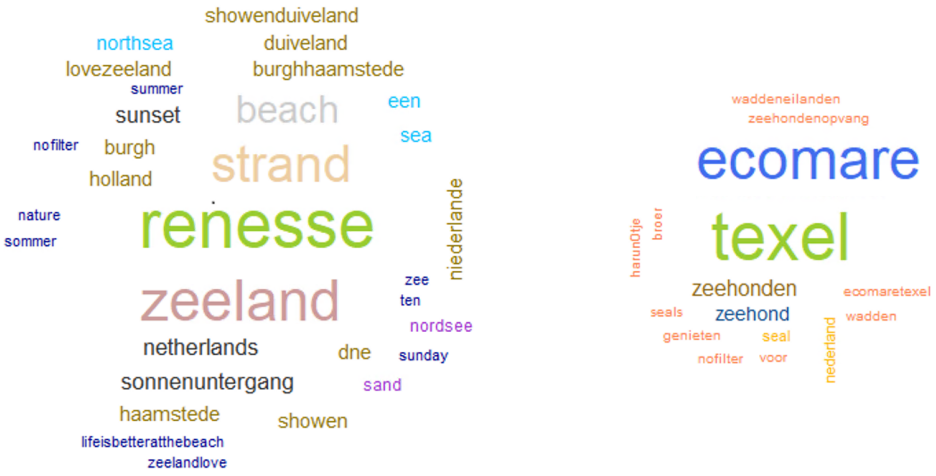
Source: Statistics Netherlands

One lesson that can be drawn from the experiments described above is that due to the further delineation of types of tourists or places, the number of social media messages and profiles can quickly decrease. This reduces the possibilities of the analysis that look quite interesting for tourism, at least at first sight. For example, also the desire to carry out analysis per week or even per day. However, a good balance should be kept between the desired detail of the analysis on the one hand and the quality of the analysis (e.g. number of cases) on the other hand. A possible solution is to look also at social media platforms that are not or insufficiently represented in the Coosto data set.[15)]

Figure 5.7a points to the possibilities to focus on social media messages from foreign visitors as a separate group, either as excursionists or as tourists. The easiest way to derive foreign visitors is to look at the language of the messages. This may reflect reality relatively well, but there are also problems that still require a solution. See also paragraph 5.2.

15) See for example Flickr.

Figure 5.9 Word clouds using the sentiment about POI's, cases: Renesse and Ecomare



Source: Statistics Netherlands

Figure 5.8 Use of buffers (purple circles) to analyze POI's: a) Renesse, b) Ecomare, c) Red-Light District and d) Duinrell

Source: Statistics Netherlands

5.6 Travel distances, visited areas and travel routes

The potential of social media messages for tourism statistics could also lead to new indicators and new output. In this paragraph, some preliminary results are presented of analysis on travel distances and travel routes. In this paragraph the data sets of selections 1 and 2 were used.

Travel distances

Travel distances can be another indicator that can be estimated with geo-location data of social media messages. First the profiles were classified into either excursionists, tourists or residents. Then the travel distances per profile were computed. The analysis was done with R-library "ggmap". Two selections were used as examples:

1. **Keyword: "Texel" (selection 2)**. This selection is already described in paragraph 4.1. It started from the 20.000 social media messages which are related to the coastal area of the Netherlands. Of this set 95 percent of the messages had a timestamp and geo-location. After removing the social media messages from Dutch businesses, a set of 19.853 messages remained. The number of messages belonging to persons posting messages in the Netherlands was 18.764. This can be messages from Dutch persons as well as foreigners. Thereafter, the messages were added to a geographic information system and linked to city polygons of the Geo-services of Statistics Netherlands, which was defined for the period of 2017.[16)] This latter step was meant to remove messages that are posted from outside the Netherlands.

   The classifier used to distinguish whether a social media profile (person) belongs to an excursionist, a tourist or a resident is in the first instance based on three practical criteria (with the area under research in mind; rule based):

   – Profiles that only send messages the same day were classified as excursionists;
   – Profiles that send messages within a period of maximum 14 days were classified as tourists;
   – Profiles that send messages longer than 14 days were classified as residents.

   Profiles which have sent only one post were removed from the data set, as these profiles did not provide enough information for this experiment. This concerned 6784 profiles. This was 35 percent of the total number of profiles (= social media messages) available in this experiment. Messages that were sent between 22:00 and 6 a.m., provided a proxy on the probable "home" of a person (or profile). This exercise resulted in the figures shown in Table 5.4.

Table 5.4 Number of excursionists, tourists and residents (coastal area of the Netherlands).

Excursionists	Tourists	Residents	Not classified
3.161	3.110	2.191	10.302
17%	17%	12%	55%

Source: Statistics Netherlands

In addition to the criteria for the first exercise, an additional rule was added to the classifier. It was also examined whether there was a displacement outside the city limits (from one city to another). In addition to time and frequency, the displacement

16) The frame of cities is set to 2017, because the social media messages belong to the period Augustus 2016 and July 2017. The frame is checked and published every year by the Geo-services unit of the CBS, given that some cities can me merged.

outside the city boundary is often used in tourism research to determine whether someone travels outside his or her "usual environment". The resulting shares of excursionists, tourists and residents of this second exercise are shown in Table 5.5.

The step from Table 5.4 to Table 5.5 shows that the number of messages available for analysis can decrease quite rapidly. The new classifier could, instead of 45 percent, only identify 20 percent of the all the messages. The other 80 percent could not be classified. See Table 5.5. An explanation for this decrease in the number of cases could be that the obligatory displacement between cities seems to strict for tourists. This may hold, in particular, for tourists that would be arriving and visiting only one city in the coastal area (e.g. Texel or The Hague). They may only post messages in that one city.

This example shows that a criterion like "outside the city boundary" cannot simply be applied to social media data to determine if a trip was outside the "usual environment" or not. There can be all kinds of "holes" in the timeline of a profile.

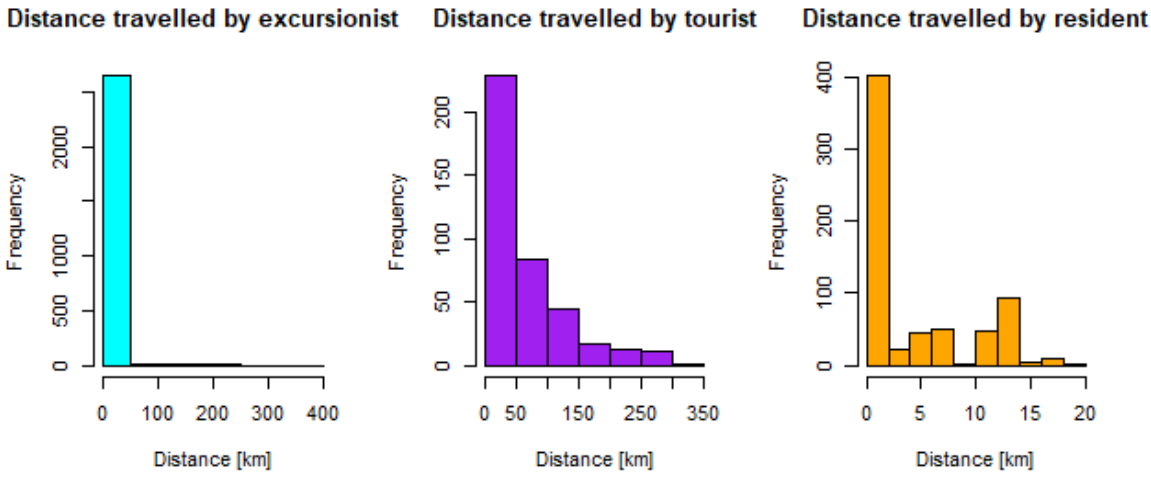
Table 5.5 Number of excursionists, tourists and residents in the of case Texel using geo-location, timestamp and displacement between cities (based on geotags matched to cities.

Excursionists	Tourists	Residents	Not classified
2.678	397	676	15.013
14%	2%	4%	80%

Source: Statistics Netherlands

Figure 5.10 shows the differences in travel distances (from home and back) by excursionists, tourists and residents. The travel distance by excursionists (2,7 thousand profiles) was less than 50 kilometers. Tourists (400 profiles) on the other hand travelled on average less than 150 kilometers. The identified residents (700 profiles) travelled less than 20 kilometers.

Figure 5.10 Distances travelled by the selected excursionists, tourists and residents (coastal areal)



Source: Statistics Netherlands

2. **Keywords: ".com" or ".nl" (selection 1)**: This selection started from the set of data from the study on social media messages and business statistics. These messages contained a URL. Filtering messages from businesses was not needed because this was already done in the study mentioned. The first selection lead to a total number of (raw) social media messages available of 1,8 million. From these messages, however, only 1 percent included messages with a timestamp and a geo-location, i.e. 17.850 messages. Hence, 17.850 messages (with a URL) belonged to profiles of persons only. After the matching, however, still 892 social media messages turned out to be from foreign companies or empty records. These messages were also removed from the data set. Eventually, the number of messages with a geo-location and a timestamp belonging to profiles of persons posting messages in the Netherlands was 16.958 messages.

   The obtained shares of excursionists and tourists from this exercise are shown in Table 5.6. This shows that the number of social media messages classified as excursionist is the highest (16 thousand). The classifier identified 95 percent of the messages as a day-tripper. It is difficult to explain why there are so less tourists and no residents in the data set. The only reason maybe the fact that messages were selected on the basis of the presence of a URL in the message.

Table 5.6 Number of excursionists and tourists and residents in the of case Texel

Excursionists	Tourists	Not classified
16.157	136	665
95%	1%	4%

Source: Statistics Netherlands

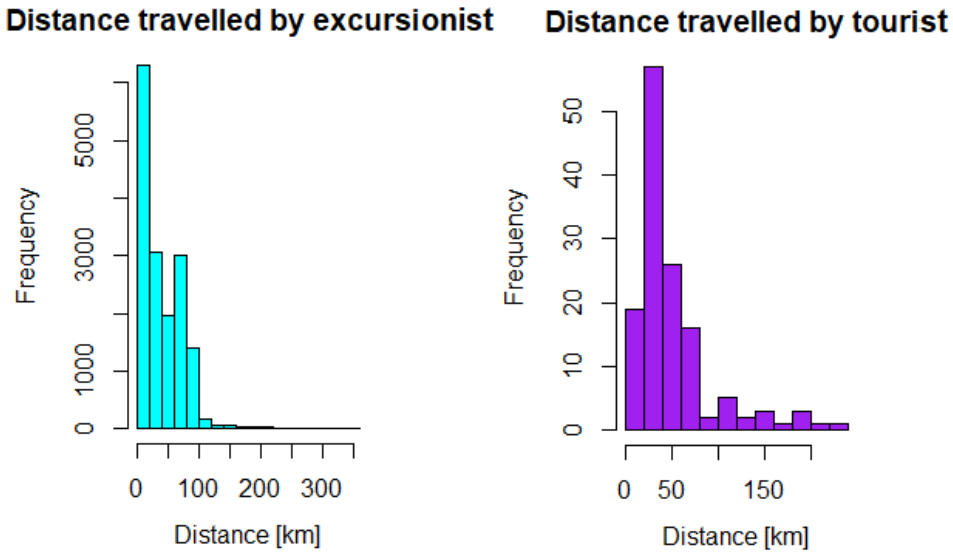
Figure 5.11 shows the differences in the travel distances by excursionists and tourists. The travel distance by excursionists (16 thousand profiles) was less than 100 kilometers. Tourists (100 profiles) on the other hand travelled on average less than 200 kilometers.

Visited areas

Next to travel distances, the research looked also into the visited areas of a set of profiles. This included among others airports. This experiment was also carried out to check roughly the performance of the classifier used to discriminate between excursionists, tourists and residents belonging to selection 2. Also, in this experiment, the analysis was carried out with R-library "ggmap".

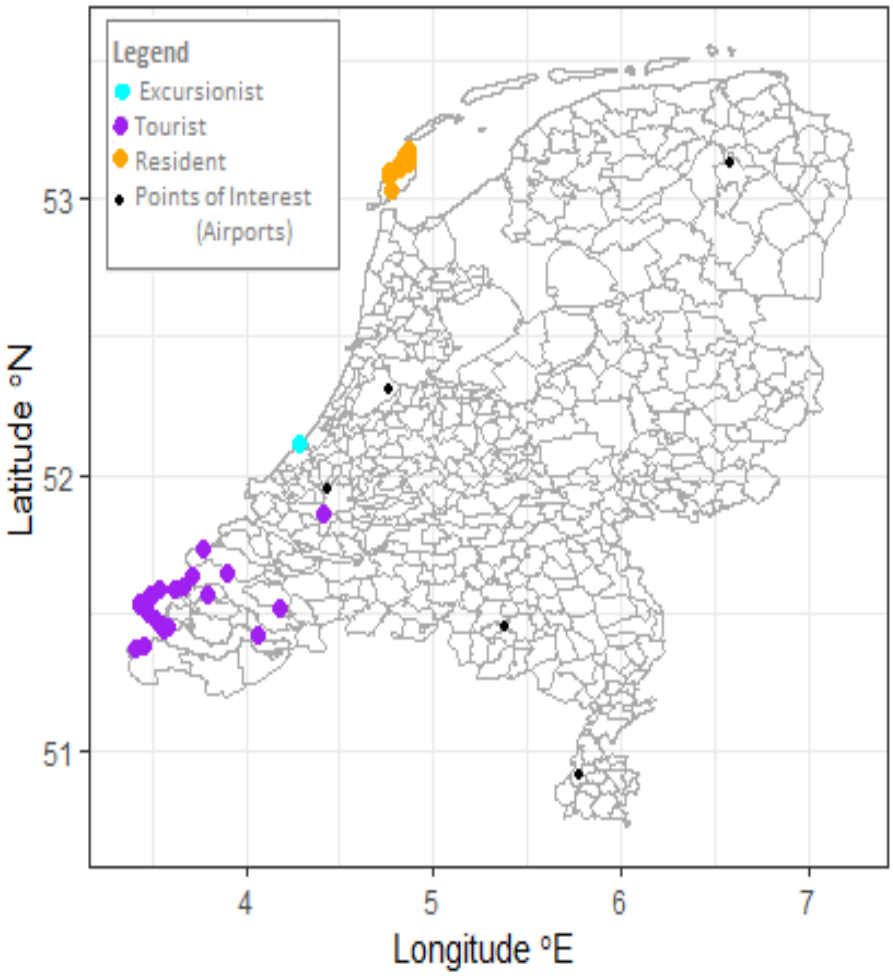
Figure 5.12 shows the geo-locations of three profiles of selection 2: excursionist ("cyan"), tourists ("purple") and residents ("orange). The excursionist remained in the same area (coastal area) almost continuously. The tourist, however, visited different locations along the coast of Zeeland. And, finally, the resident remained while posting messages on the island of Texel.

Figure 5.11 Distances travelled by the selected excursionists and tourists (keywords: ".nl" or ".com")



Source: Statistics Netherlands

Figure 5.12 Comparing an excursionist, a tourist and a resident (coastal area)



Source: Statistics Netherlands

Travel routes

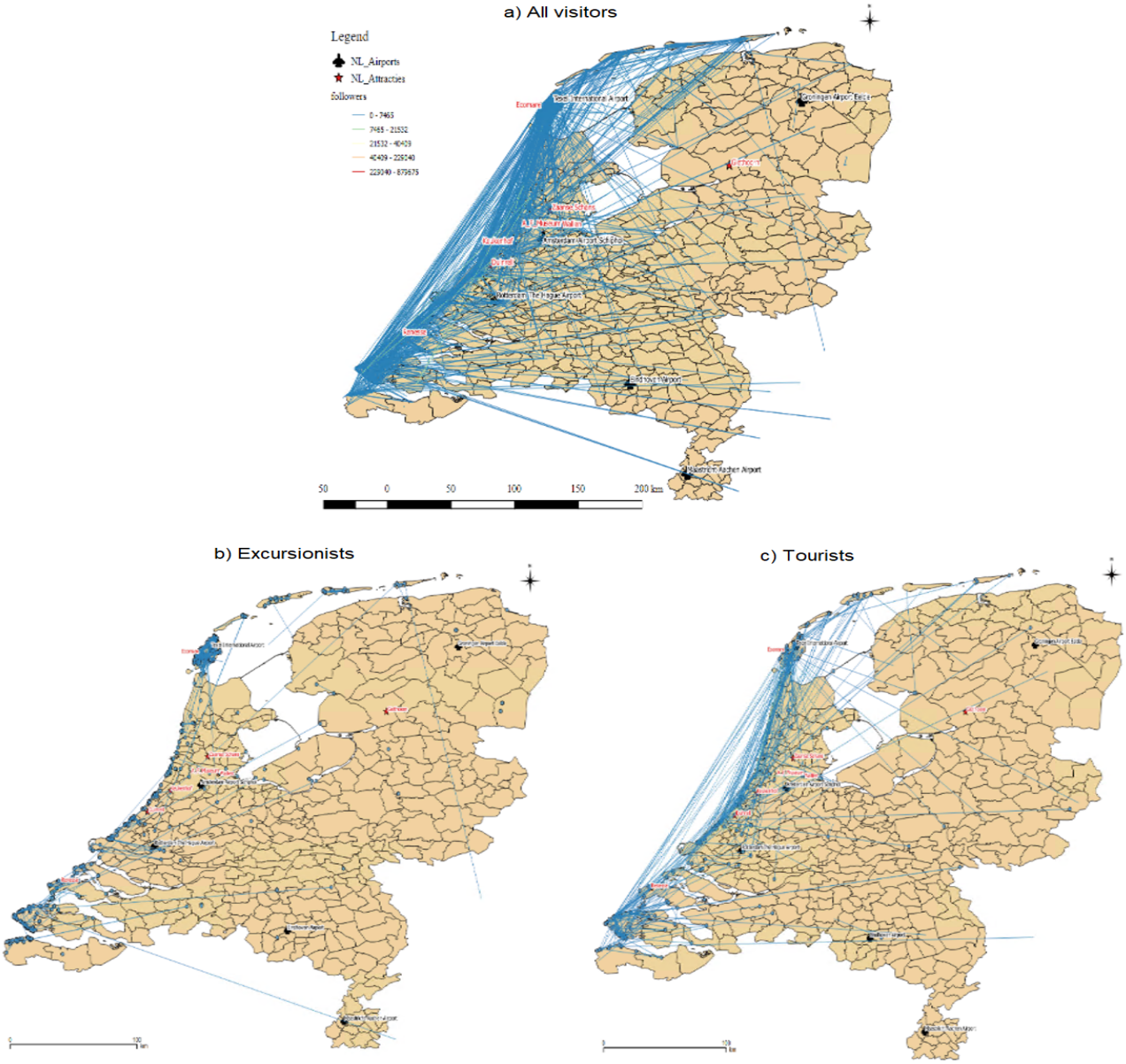
In this experiment all the obtained geo-locations from the social media messages of profiles in the data set of selection 2 (coastal area) were used. The analysis was carried out in QGIS.

Figure 5.13a shows the travel routes of visitors, so excursionists as well as tourists. It is interesting to see that a part of the visitors to the coastal area also travelled to other coastal areas. Furthermore, only a relatively small part of the visitors came from the East part of the Netherlands or from neighboring countries such as Germany and Belgium. That does not mean that only a few visitors from the East or from any other part of the Netherlands visit the coastal region. This result may have to do with the way the selection is made. One must also consider that not all visitors post messages.

After classifying the messages into profiles of excursionists and tourists also travel routes of these groups of visitors were derived and analyzed. Excursionist (Figure 15.3b) seem to travel mainly along the coast. A few of them post messages from the middle or eastern part of the Netherlands. Tourists (Figure 5.13c) on the other hand, seem to post messages along their whole travel route. These routes were also along the coast. Also, with these results, one has to consider the possible effects of the way the selection was made.

The last three experiments with social media data and then especially the translation to data on different groups of profiles, like visitors, excursionists and tourists, show that it is possible to follow the travel routes of these groups based on social media messages. These results can be visualized in a good way in a geographic information system or with R. This kind of information can also be combined with information on the supply side of tourism, like attractions, restaurants, accommodations etc. as well as geographic information on roads and cities. It is also clear that the methodological and empirical choices made are ultimately important for the end results. In addition, relatively little work was done to look more specifically at the quality of the cases: the profiles and the chronological order of the messages posted (timelines). For example, rather little time have been spent on outliers or missing data in the series of messages from a profile. These are all steps required to improve the quality of the data set and eventually the quality of the end results. Furthermore, it can be noticed that the number of cases can decrease quite rapidly when more detailing takes place. That can make the results dispersed. Nevertheless, social media data remain a good addition to, for example, mobile phone data, certainly when it would be possible to see what kind of activities are performed along the routes of visitors.

Figure 5.13 Routes travelled by a) all visitors, b) excursionists and c) tourists (selection 2)



Source: Statistics Netherlands

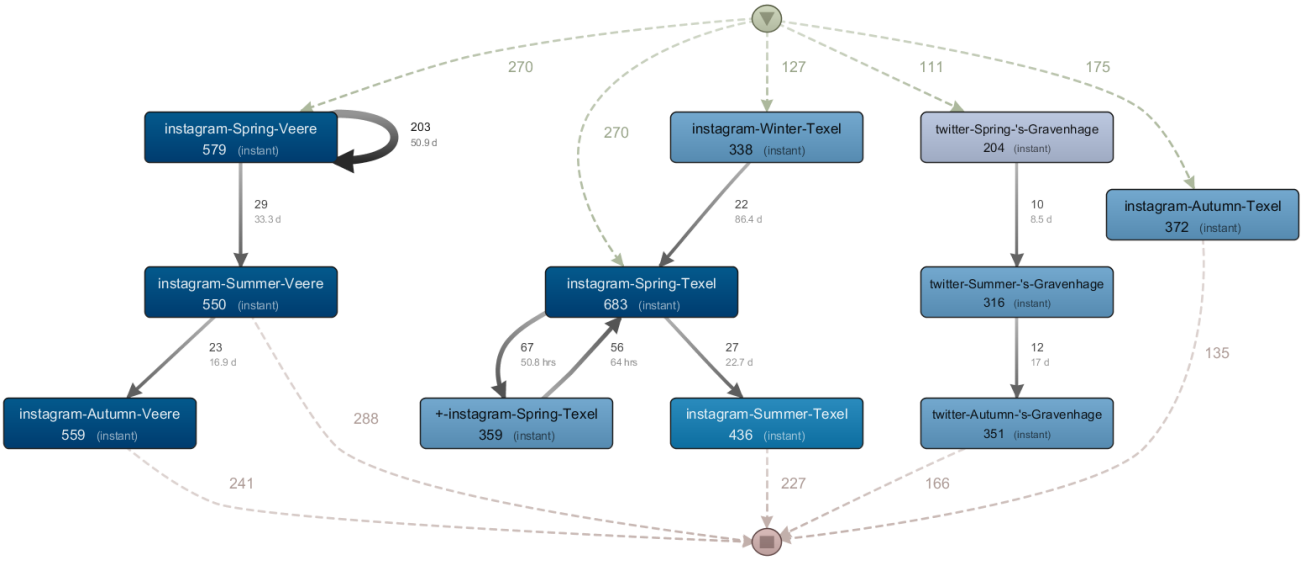
5.7 Travel routes: identifying routes per destination and per season

In this paragraph, results are presented of another kind of analysis of the travel routes of different groups of visitors. Variables, that were used to differentiate, include among others: types of social media platforms used, types of visitors (excursionists and tourists), places and seasons. For this analysis data from selection 2 were used: messages posted in the coastal area of the Netherlands. For this specific exercise, however, we zoomed in on three tourist destinations, i.e. Veere, Texel and The Hague ('s-Gravenhage).[17)]

17) These cities, that are situated along the coast of the Netherlands, have a tourist orientation. The city of Veere is known because of its water-side beach, but also because of the Delta Works or Windturbines on the Oosterscheldekering (major flood defense works) (Province of Zeeland). The island of Texel, to be reached by ferry, has a number of nature parks (Province of Noord-Holland) and The Hague ('s-Gravenhage) includes the beach of

In contrast to the previous paragraph, in this exercise, the analysis was carried out with the application software Disco.[18] Disco is first of all an application for process mining. But the package can also be used to analyze any data file with activities, timestamps and cases. These cases (e.g. trips) could be a series of messages posted per profile with a geo-location and a timestamp. Disco provides, for example, the possibility to decompose the cases and activities into different groups (attributes) and to present the relationship and chronological order between these decomposed groups. Besides the plots shown below, Disco also can present the data in real-time and an interactive way. So, it can present a dashboard with the paths which are followed in the plot per case (per profile). The boxes (vertices or nodes) and the related numbers in the plots below show the activities, in this case the messages posted. The lines (paths or edges) and the related numbers in the plots show the cases, i.e. the profiles (persons) in the data set. Furthermore, it is possible to play with the level of detail. One can leave out boxes and lines which are less important. Otherwise the plots often become too complicated. So, one should keep in mind that the presented plots in this paper also do not contain all boxes and lines. Finally, also the time between boxes (activities) are calculated in Disco. These are presented by the small numbers on the lines, indicated by, for example, hours, ("hrs"), days ("d") or weeks ("wks"). At the moment, these numbers are not very helpful. It would be more interesting to look at either the "mean" or the "median" of these paths.

Figure 5.14 Type of social media, season of the year and location of messages posted. Cases of Veere, Texel and The Hague (Disco)



Source: Statistics Netherlands

In paragraph 4.1, it was already indicated that social media platforms with the highest number of messages were Twitter, Facebook and Instagram. See Figure 4.2. Instagram messages contain often more geo-location data than Twitter messages. On the contrary, Facebook messages often lack geo-location data.

18) Scheveningen and is the administrative capital (government) of the Netherlands (Province Zuid-Holland)
See https://fluxicon.com/disco

Figure 5.14 analyzes mainly the season in which social media messages were posted and the type of platform used. So, the figure shows a decomposition of the social media messages, that were posted by the visitors. In this example social media data related to the city of Veere during the period that the data were collected from the Coosto-database, per season and per type of platform.

The plot of Figure 5.14 shows that the visitors to Veere posted messages by using predominantly the platform Instagram. This accounts for the spring and the summer as well as the autumn. In the winter season no messages were posted. Also, only a small portion of the profiles posted messages in the spring as well as in the summer (29 persons or profiles). This could indicate that most people do not go to Veere more than one time a year. The loop at the box "Instagram Spring Veere" means that people have posted more than one message during this period of time.

The visitors to Texel also used Instagram the most to post social media messages. These messages were predominantly sent in the winter, the spring and the summer. Figure 5.14 also shows that about 50 percent of the messages of the visitors to Texel were posted in the spring. These messages had a positive sentiment (359 out of 683). On the right side of Figure 5.14 there is a separate branch. This branch shows another cluster of messages send by visitors to Texel, but these messages were only sent in the autumn. This might indicate that the visitors in the autumn do not visit the island of Texel in other seasons. This does not hold for the messages of the branch on the left side of Figure 5.14. These messages could be posted by multiple visitors or owners of a second home in Texel. When looking at Instagram, the number of repeat visits within a year looks rather small.

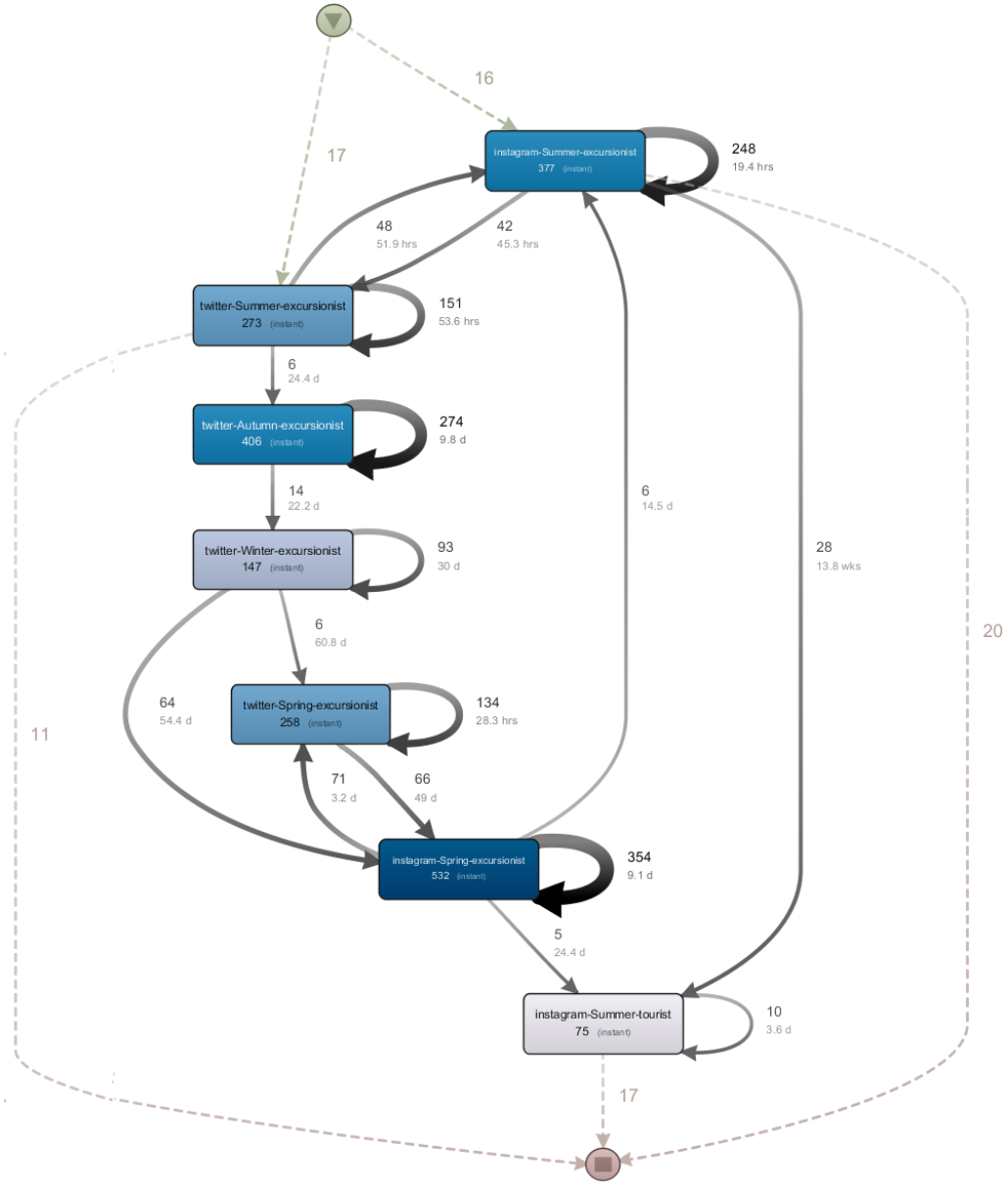
The visitors of the city of The Hague predominantly used Twitter in the period under research. The posts were mostly sent in the autumn, the summer and the spring.

Figure 5.15 zooms in on the visitors who went to Texel (see the middle part of Figure 5.14), by type of visitor, per season and per social platform. Tourism is a key economic activity for this island. It is estimated that 80 percent of the residents of Texel directly or indirectly depends on the expenditures of visitors.[19)] Figure 5.15 shows that Twitter seems to be used mostly by excursionists. The largest number of messages of excursionists were registered in the autumn (406 messages). The least were posted in the winter (147 messages). Spring and summer show about the same number of social media messages posted. Figure 5.15 also shows that Instagram was mostly used to post messages. This was predominantly the case in the summer and in the spring. This is consistent with Figure 5.14. Social media messages that were posted by tourists in the summer (75 messages, related to accommodations) used mainly Instagram. The number of messages from tourists were much lower than the number of messages that were posted by excursionists using either Twitter or Instagram. In particular, Instagram seems to be the social network for excursionists in the spring.

Another experiment, that was carried out with Disco, focused mainly on the different routes travelled by different types of visitors. In addition, it was examined whether something could be said about the duration of the visit.

19) Melchior, M., "Reportage Texel: De Europese massatoerismestad in het klein: Texels levensader slibt dicht", Volkskrant, August, 2018.

Figure 5.15 Type of social media messages, per social platform, per season and per type of visitor, case Texel (Disco)



Source: Statistics Netherlands

Figure 5.16 shows the results of the different routes that were followed by excursionists and tourists. In order to keep the figure readable, it should be noted once again that not all boxes and lines are shown. Two routes stand out, that is:

A) The route Sluis-Vlissingen-Veere-(Vlissingen)-Sluis (in the province of Zeeland) A share of the social media messages has a geo-location belonging to the city of Sluis, a small town on the border with Belgium. This set is classified as excursionists. A share of the visitors which have been classified as tourists heads to Vlissingen. The share of visitors that is classified as excursionist goes further to Veere and then goes back either straight to Sluis or via Vlissingen. Tourists that are going back to Vlissingen, probably do this to stay overnight. Another route to get to Veere is via the route Schouwen-Duiveland-Veere-Vlissingen. It comes from another cluster of excursionists and tourists. It can be a bit confusion to find tourists in the routes of excursionists. This could be because not all boxes and lines are shown in the plot of Figure 5.16. But it can also be that insufficient care is taken of profiles (people) with multiple trips in one year. If that is the case, then people can visit this region more often: one or more times as an excursionist and one or more times as a tourist.

B) The route Schiermonikoog-Terschelling and the route Texel-Rotterdam. The right side of Figure 5.16 shows two other examples of routes that were chosen by visitors. Excursionists that visited Schiermonikoog also liked to visit the island of Terschelling (one of the other Wadden islands) in that same year. Excursionists to Texel look to have a strong relationship with the city of Rotterdam. In a lesser degree that accounts also for the small coastal city of Bergen in Noord-Holland.

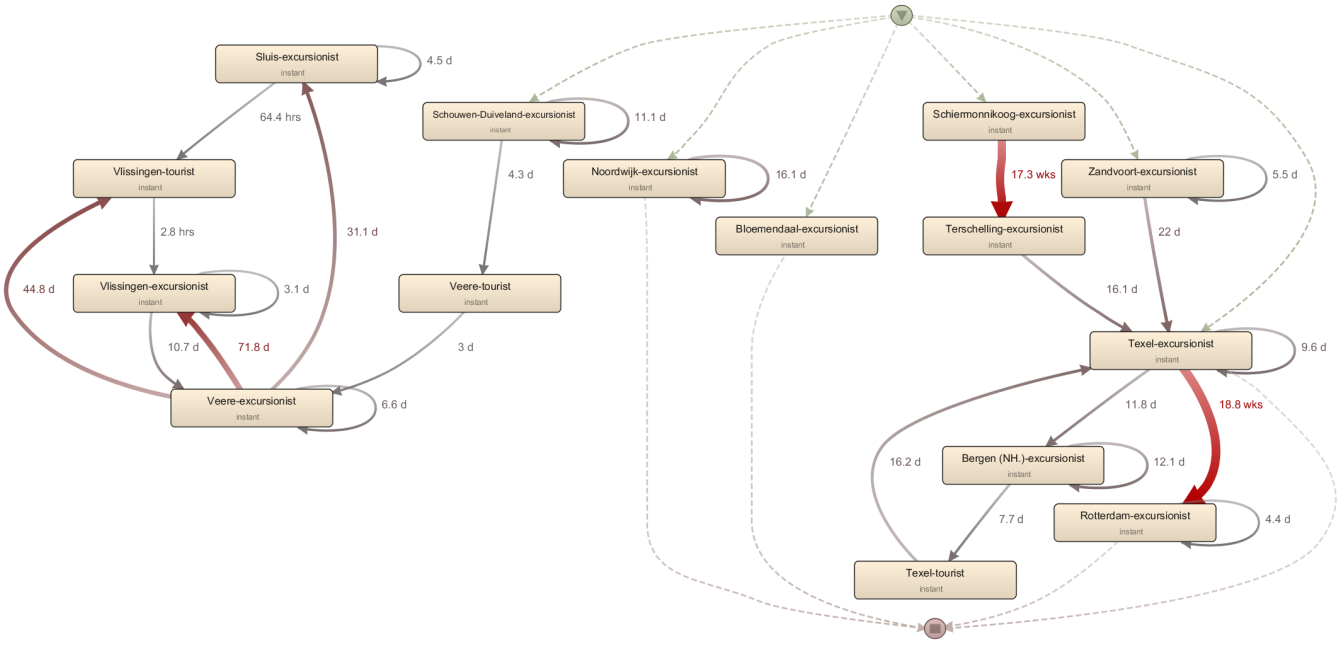
More experiments with Disco were carried out in this study. It goes a step too far to show them all. For example, we looked at visits to or from cities, like Rotterdam, Amsterdam, Noordwijk and Zandvoort.

Also, an experiment, was carried out, to look at the possibilities to say something about the duration of stay of tourists in the cities of Amsterdam, The Hague and Veere. It accounts for the median duration derived from the timestamps in the social media messages. The first experiments resulted in 4.8, 5.5 and 2.9 days, respectively for Amsterdam, The Hague and Veere. However, it should be noted, that these are estimated median durations which need to be further improved using a more refined classifier.

Again, the experiments described above were not intended to come up with exactly the right figures. The main goal was to see whether certain questions in the domain of tourism could be answered with social media data. And, what kind of methods and techniques are useful to explore further. In this paragraph the focus was on analyzing the different routes taken by different groups of visitors and to see which places they have visited. The package Disco looks to be a good option to do so. However, there are also other software applications that have the same features. A pre-requisite is that the results of these applications can immediately feed into the geographic information system used.

Besides this, there are still many areas where improvements can be made. This focuses primarily on improving the number and quality of the cases used in these applications. An example is dealing with multiple trips from one profile in a given period. In principle, every single trip should be classified as a separate case. Outliers should be detected and removed. There should also be an attempt to fill in missing data in a series of social

Figure 5.16 Flows of social messages along the coast, per type of visitor and based on the average duration (stay).



Source: Statistics Netherlands

messages posted by a profile (timeline) or, if necessary, they should be deleted from the data set. Maybe it is also better that different types of visitors, like excursionists, tourists but also foreign visitors, are analyzed independently as separate groups. In addition, one should always be aware of the fact that as more detailed figures are required, the number of cases can decrease rapidly. Much also depends on how the social media messages and associated profiles are selected and how big the data set is. It also depends on the type of social media platforms included. Some platforms lend themselves more to messages with geo-locations and timestamps then others.

Finally, Figure 5.17, as shown here, is an example of a previously published method (Wageningen University, Dat.mobility and Amsterdam Institute for Advanced Metropolitan Solutions) for the city of Amsterdam with information based on the platform Flickr. This relatively simple method plots messages with pictures on the basis of timestamps and geo-locations. It also distinguishes between tourists and local residents and figures per month. It is assumed that the term “tourists” here is equal to the right term “visitors”. The nice thing of the map on the left-side is that it is a 3-D image, while the resulting heat map on the right is probably more informative for policymakers. The disadvantage is that messages with no text cannot be interpreted. On the other hand, it would be nice if this kind of images could be supplemented with other information, for example on the basis of 100x100 meters if it concerns data on the supply-side of tourism or the background of people who are living there. This allows to look, for example, at over-tourism.

Figure 5.17 A plot based on Flickr messages with a geo-location and a timestamp (left) and the resulting heat map (right)(Amsterdam)

Source: Wageningen University, Dat.mobility and Amsterdam Institute for Advanced Metropolitan Solutions

5.8 Comparison with Tourism Accommodation Statistics (TAS)

In this last paragraph of this chapter, the results of a comparison made between the outcomes of the Dutch Tourism Accommodation Statistics (TAS) on the one side and estimated numbers of visitors and tourists based on social media data on the other side is presented. The TAS produces figures on the number of guests and the number of nights spent in accommodations establishment in the Netherlands, distinguished by country of residence.

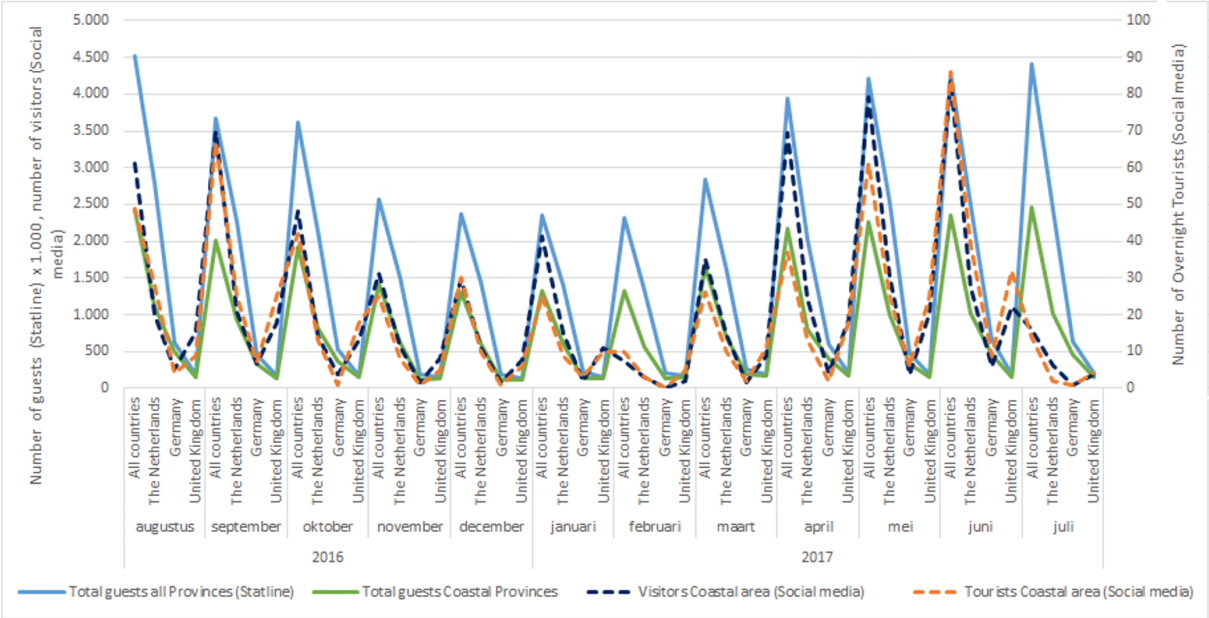
To validate the outcomes, first a language detection package in R ("textcat") was used to identify the share of messages in Dutch, English, German and other languages. The comparison is only based on the messages of profiles that are classified as tourists in selection 2, meaning: persons who spend at least 1 night in an accommodation and not more than 14 nights. Profiles with Dutch messages were then classified as domestic tourists and all profiles with non-Dutch messages were classified as inbound "foreign" tourists. Also here, it can be noticed that the number of usable cases decreased significantly.

The second step was then to make the comparison between the published numbers of tourists visiting all provinces of the Netherlands and the guests visiting only the coastal provinces[21)] on the one hand with the numbers of visitors and tourists derived from social media messages on the other. See Figure 5.18 for the results.

The results show that the peaks and trends, that were found, are quite consistent for the months of September 2016, January 2017 and April – June 2017. As a first result, this is not bad so far, given the restrictions and assumptions that were made. One of the restrictions was that it was presumed that all non-Dutch messages belonged to foreigner visitors or tourists. It has already been said that this is slightly more complicated. For

20) StatLine is the central web base of Statistics Netherlands, with which all data are made public.
21) These are Friesland, Noord-Holland, Zuid-Holland and Zeeland.

Figure 5.18 The number of guests in tourism accommodation establishments in the Netherlands and the coastal provinces (Statline [20]) versus number of messages of visitors and tourists based on social media data



Source: Statistics Netherlands

example, it is difficult to distinguish between messages from Flemings and messages posted by Dutch people. Although, an initial distinction was made between German, English, Italian, French, Polish and Spanish messages, at the end, it seemed more sensible just to aggregate all non-Dutch messages. Also, only German and English messages looked to be relevant. Assumptions made are, for example, that the social media data are representative for the population and that the classifiers are working well.

As said before, social media research cannot be used very well to look at numbers and volumes. This is reinforced in this research by the small number of cases available in the data set. That is why we only looked at the trends and peaks in the resulting figures. Therefore, it makes not too much sense to look carefully at the scales on the right and left side of the Figure 5.17. For information: the left-vertical axis stands for the Statline figures (with scale in thousands) and the visitors based on the social media profiles. While the right-vertical axis shows the number of tourists obtained by using social media data.

6 Conclusions and recommendations

The number of people nowadays who are using social media platforms worldwide is impressive and still growing.[22)] That makes it attractive to use publicly available social media messages as a new source of big data for research. This paper tries to provide some more insight into the possibilities in this field. Following on from an earlier study into social media data for businesses statistics, this experimental study examined the domain of tourism. At least before the COVID-19 crisis, the tourism industry was a substantial sector of the Dutch economy: 4,4 percent of GDP and almost 800 thousand jobs. Growth was expected to continue. This continuous growth of tourism also led more and more to negative developments, such as pressure on the environment and social pressure. This increased the need for more up to date and detailed figures. However, recently the COVID-19 crisis has hit tourism especially hard: most flights have been cancelled, museums and restaurants are closed, major events are not allowed and often there is an entry ban. This crisis did not change the need for more up to date and detailed figures, on the contrary, but it did change the nature of the questions. So, the need for more (big data) research still stands.

The main intention of this exploratory research on the use of data from social media for tourism statistics was not to produce exactly the right figures, but just to explore the possibilities in general by carrying out a series of experimental studies. For practical reasons, like processing time and amount of work, data sets from another study from 2017/2018 were used. Some of these data sets were already processed and ready to be analyzed. It made little difference to the end results and the conclusions.

Social media research can be used in the different stages of the customer journey of visitors. In the literature most attention is paid to marketing research. However, that was not the topic of this study. Social media research can also provide other information on, among others, the behavior and movements of tourists, on their activities and satisfaction. In fact, social media are unique as a resource because of the combination of content on the one side and information on geo-locations and time on the other. This study has explored a number these possibilities, but also looked at the limitations of social media research. The intention was not to be comprehensive. The study has led to the following conclusions.

First of all, it is important to realize that research into social media data and tourism statistics or any other statistic is not yet sufficiently based on a strong conceptual and methodological foundation. So, further research requires more input from methodologists when it comes to questions like representativeness and validity of the results. However, that makes the results no less valuable. Figures from social media research should, at the moment, be seen as beta-indicators with sometimes (many) noise in the end results. Researchers should be clear about this limitation and its influence on the end results. Producing volume figures should, therefore, not be the first goal. It is not easy, for example, to translate the number of profiles or the number of social media messages into number of visitors. It is better to look at developments, shares and benchmarking. Social media research can also be seen as an addition to other (big) data sources, especially like

22) Zie https://datareportal.com/reports/digital-2020-global-digital-overview

mobile phone data. Besides research on representativeness, methodology, that can bring social media research for tourism statistics a step further, are techniques like text mining, machine learning and deep learning, especially when it concerns the content of the messages. Also work has to be done on techniques to develop so called classifiers. In this study some examples are discussed. In order to obtain better data sets, it is essential to look more closely at the step of data cleaning. Finally, it should also be borne in mind that in time there can be substantial shifts between the popularity of social media platforms. This has a continuous effect on the results of social media research.

A second element of attention is that of privacy. Not discussed in this paper. Although this data source concerns publicly accessible data that can be scraped from the internet, it is an obligation of every researcher the protect all the negative aspects thereof.

A third element of social media research for tourism statistics is that always a distinction must be made between social media messages from visitors, like excursionists, tourists or any other valuable detailing for tourism, and social media messages from other groups, like residents and businesses. Although certainly not 100 percent correct, this study shows that, with some assumptions, there are certainly possibilities for classifiers here. However, concepts and methods in this field need to be refined, for example by incorporating new text mining techniques and machine learning.

A fourth element that plays a role is that for research into mobility one has to limit the data set to social media messages that have a timestamp and a geo-location. A disadvantage is that this can limit the number of available cases in the data set considerably, certainly if you want to further detail. What is also important here, is that per profile and per trip good matrices must be set up of every step in the trip (origins and destinations). More attention should be given to the 'cleaning' of the data in this respect, meaning: supplementing missing data, researching outliers, breaking up multiple trips into a set of single trips and removing incorrect data. It is also important to have a good look at the selections which are chosen and made, the groups of visitors that are analyzed in conjunction and the social media platforms that are used.

A fifth conclusion is that social media research often allows for more detailed research then meets the eye. This can lead to new and quicker indicators. On the other hand, as said, the limitation of social media research is that through further detailing the number of cases can decrease substantially. The problem is that the number of so-called contact points of social media (messages posted) is limited, certainly when compared to mobile phone data. Moreover, the flow of contact points is not nicely and evenly distributed over time. Visitors, who use social media, only post messages when they like to do so. This leads to gaps in the series of data.

A sixth conclusion of this study is that the use of graphical tools for visualizations of the results can be very helpful, especially as a part of tourism is all about mobility. The possibilities for research are further increased when these tools, like geographic information systems, already contain information about roads, areas, POI's, restaurants, accommodations etc. (the supply side of tourism). These tools can help to analyze different sets of data in conjunction. A good starting point would be to merge social media data in such a system with financial data and data from mobile phones.

Finally, one should continuously realize that by only looking at visitors, who use social media, one is always missing a large part of the visitors: visitors that do not use social

media, visitors that did not turn on their GPS and visitors that do not use certain keywords that are used in the selections.

In more detail, this study looked at, among other things:

- Possibilities to distinguish between different groups related to tourism, such as excursionists, tourists and residents. In particular work must be done on improving classifiers;
- Sentiment-analysis. This study shows that in addition to deriving the sentiment (or satisfaction), it is also worthwhile to look at the context of the sentiment by analyzing the content of the related messages. Therefore, techniques such as word clouds and machine learning, can be used. This also applies, for example, when one wants to look at activities and behavior of different groups of visitors;
- Linking social media messages to points of interest (POI's). The linkage of social media messages to points of interest (supply side of tourism) were carried out here by using buffer areas. The data shrinkage is however, a bit of a problem. This can be sorted out by using more data with geo-locations or by refining the discrimination of the social messages using emoticons;
- Visualizing flows or travel routes of visitors, that use social media. In this study origin and destination matrices were used as input for, among others, the data mining tool Disco. It seems possible to visualize the routes of visitors with social media, comparable with mobile phone data. In itself this can be done on a fairly detailed level. However, also here, the more detail, the less data is available. Combining social media data with mobile phone data is a promising option.

So, in general, the results of this study provide a positive picture of the possibilities of using social media messages for (new and quicker) tourism indicators. However, there are also clear limitations. The most important of these limitations is the lack of a solid conceptual and methodological foundation, of which representativeness is a major issue. So, it is recommended:

- to keep it simple for the time being;
- to present results as beta-indicators;
- to combine results as much as possible with other sources, like mobile phone data, financial data and above all data from the supply-side of tourism (geographic information systems);
- to look more closely at those social media platforms that contain messages with a geo-location and a timestamp;
- to look more closely at the keywords when selections are made;
- ensure a sufficient solid conceptual and methodological foundation. Important subjects for further research are, among others, representativeness, classifiers (distinction of messages from different groups of visitors), data cleaning (generating quality cases), origin and destination matrices (for routes and flows of visitors), effects of retweeting, visualization techniques and last but not least content analysis by using the latest techniques of text mining and machine learning;
- basic concepts, that apply for tourism research, should not be forgotten. An example is the "usual environment";
- Finally, in order to make substantial steps forward, it is important, beside content experts, to involve also methodologists, graphical experts and data-scientists. This multi-disciplinary aspect is often forgotten.

7 References

- Alaei, A.R, Becken, S. and Stantic,B., (2017), *Sentiment analysis in tourism: capitalizing om big data*.
- Agarwal, B. and Mittal, N., (2015), *Machine learning approach for Sentiment Analysis*.
- CBS, (2019), *ICT-use of persons in the Netherlands, 8 October 2019*. See also: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83429NED/table?ts=1588270850247
- Cloudfront.net, (2018), *GIS-layers displayed*, Last visited on 7 January 2020.
- Daas, P.J.H. and Puts, M.J.H., (2014), *Social Media Sentiment and Consumer Confidence*, Statistics Paper Series, No5. September 2014, page 31-32.
- Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A.Y.A., Gelbukh, A. and Zhou Q., (2016), *Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques*, 1 June 2016.
- David K. and Wickham H., (2013), *ggmap: Spatial Visualization with ggplot2*, in R Journal, Vol 5/1, June 2013.
- Habib, M. B. and Krol, N.C., (2017), *What Does Twitter Tell Us About Tourists Mobility Behaviour*, (2017), Pacific Asia Conference on Information systems, (PACIS) 2017 Proceedings, 19 July 2017.
- Hort, M., Zhang C., Shingjergji K., Ignéczi M. and Habib, M., (2018), *Tourists Mobility on Social Media*, 26 January 2018.
- Ortega, S. and Heerschap, N.M., (2019), *Verkennend onderzoek naar de mogelijkheden van het gebruik van sociale media voor statistiek* (in English: Exploratory research into the use of social media for business statistics), publication Statistics Netherlands.
- Spinder, S.T., (2019), *Estimation of the number of guests and overnight stays in platform-related accommodations*, Discussion Paper, Statistics Netherlands (CBS), November 2019.
- Taj, S., Shaikh, B.B. and Meghji, A.F., (2019), *Sentiment analysis of new articles: a lexicon based approach.*

Appendix

Matrix 5.1 Experimental lexicon-based approach

Taal_NL	Lang_ENG	Sprache_DEU	Landen	Countries	Steden	Cities	Sent_Pos_NL	Sent_Pos_ENG	Sent_Neg_NL	Sent_Neg_ENG
bezoek	visit	besuch	Belgie	Belgium	Amsterdam	Amsterdam	zon	sun	regen	rain
reis	trip	reise	Duitsland	Germany	Den Haag	The Hague	warm	warm	koud	cold
overnacht	sleep	nacht	Frankrijk	France	Rotterdam	Rotterdam	lekker	tasty	ijskoud	freezing
gast	guest	gast	Engeland	England	Middelburg	Middelburg	erg lekker	delicious	asociaal	asocial
dagje uit	day out	tagesausflug	Belg	Belg	Giethorn	Giethorn	heerlijk	yummy	vies	dirty
uitstapje	outing	ausflug	Duits	German	Den Helder	Den Helder	zalig	gorgeous	vuil	garbage
excursie	excursion	tagesausflug	Frans	French	Scheveningen	Scheveningen	smakelijk	heartily	vervuild	filthy
toer	tour	tour	Brits	English	Parijs	Paris	verrukkelijk	delicious	smerig	messy
trein	train	zug			Londen	London	mooi	lelijk	onaangenaam	unpleasant
Thalys	Thalys	Thalys			Brussels	Brussels	zoet	sweet	afval	waste
verbinding	connection	verbindung			Rome	Rome	fijn	goed	walgelijk	offensive
verblijf	stay over	bleib			Barcelona	Barcelona	vreugde	happyness	triest	sad
schema	itinerary	route			Munchen	Munich	geneugte	joy	boos	angry
reiziger	traveller	traveller			Frankfurt	Frankfurt	genot	delight	woede	rage
stad	city	stadt			Beieren	Bavaria	graag	pleasure	furie	anger
steden	cities	städte					zeer aangenaam	great	angst	fear
grens over	cross-border	grenzüberschreitend					schoon	clean	walging	disgust
bestemming	destination	bestemming					netjes	neat	rouw	grief
vertrek	departure	vertrek					opgeruimd	tidy	verdriet	distress
aankomst	arrival	ankunft					gezellig	cossy	blijdschap	cheersfully
seizoen	season	jahreszeit					genieten	enjoy	boef	thief
winter	winter	winter					blauwelucht	bluesky	helaas	unfortunatelly
zomer	summer	sommer					mooi	beautiful		
lente	spring	frühling								
herfst	autumn	herbst								
route	route	route								
airline	airline	fluggesellschaft								
luchtvaartma	airline	fluggesellschaft								
vlucht	flight	flug								
schema	schedule	zeitplan								
vrienden	friends	freunde								
familie	family	familie								
vakantie	holiday	urlaub								

Source: Statistics Netherlands

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information