1-2020

GI_Forum

Journal for Geographic Information Science

ÖAW
AUSTRIAN
ACADEMY OF
SCIENCES

GI_Forum 2020, Issue 1
Page: 3 - 14
Full Paper
Corresponding Author:
sabine.hennig@sbg.ac.at
DOI: 10.1553/giscience2020_01_s3

Implementing Geo Citizen Science Solutions: Experiences from the *citizenMorph* Project

Sabine Hennig[1], Lorena Abad[1], Daniel Hölbling[1] and Dirk Tiede[1]
[1]Salzburg University, Austria

Abstract

To exploit the potential of geo citizen science, technological solutions are needed that are tailored to the requirements of citizens and scientists. To create suitable solutions, participatory design is a valuable means. While information on techniques for requirement-gathering in cooperation with future solution users exists, less knowledge is available regarding tools for creating solutions together with future solution users. One tool used in a professional setting is ESRI's *Survey123 for ArcGIS*. The suitability of *Survey123 for ArcGIS* to implement geo citizen science solutions was evaluated within the *citizenMorph* project. The experiences showed that by using *Survey123 for ArcGIS* most requirements were met, but citizens faced a number of challenges using the *citizenMorph* solution developed.

Keywords:

contributory web maps, geomorphological phenomena, geo citizen science, landscape dynamics, participation, Survey123 for ArcGIS

1 Introduction and research questions

The rapid advance of Information and Communication Technologies (ICT) triggered a shift from traditional to online participation. The use of geospatial technologies – allowing, e.g., the public to contribute spatial information – has also received increasing attention. An example of making broad use of online participation, including geospatial technologies, is citizen science. Citizen science is the engagement of citizens in scientific processes with the aim of actively integrating them, their knowledge and commitment into scientific research and, thus, gaining new scientific knowledge. This can take different forms (Haklay 2013): (i) crowdsourcing projects (passive generation of data), (ii) contributory projects (citizens' active contributions of data based on their own observations), (iii) collaborative projects (actively contributing data and taking part in project design), and (iv) co-created projects (participating in project design and implementation). The integration of spatial data into citizen science is also called geo citizen science (Murray 2018). Contributory web maps can play an important role in geo citizen science because of their popularity among the public.

To fully exploit the existing potential of (geo) citizen science, technological solutions are needed that are tailored to the requirements of citizens and scientists and that take into account

experiences gained in the field of participation (Hennig et al. 2019). However, citizens' demands, particularly with regard to spatial data products, often differ substantially from the requirements of experts and are less well known (Tsou & Curran 2008).

To understand users, their resources and needs, and to consider these aspects when creating technological solutions, the active and direct participation of the future users in the development process (participatory design) is a valuable approach. It supports the development of solutions that deliver better user experience, increases the acceptance of the product in use, and ensures that the tool meets the requirements of the intended target group (Muller & Druin 2012; Steen et al. 2007). To successfully use the participatory design approach, the techniques and tools used in the development process must support the intended involvement of future users. While information regarding requirement-specification techniques is available (e.g., related to user-centered design), information regarding tools used for solution implementation (particularly geospatial technologies) is lacking.

One tool used to deliver off-the-shelf solutions for spatial information collection is ESRI's *Survey123 for ArcGIS* (hereinafter referred to as *Survey123*). It is used mainly to support research in ecology, biology and the social sciences (Ahmed II & Pradhan 2019), and less in areas such as geomorphology and landform dynamics. Although there are several benefits to using *Survey123* (e.g., easy and intuitive to learn and use, well documented, interoperability of various ESRI products), the question is how suitable *Survey123* is for the creation of geo citizen science solutions. How well can project-specific requirements (in terms of the research domain and the target group) be met by using *Survey123*? The *citizenMorph* project, which aims at developing technological solutions for citizens to contribute (spatial) information on landforms, addresses these questions.

2 The citizenMorph project

The *citizenMorph* project (Observation and Reporting of Landscape Dynamics by Citizens; http://citizenmorph.sbg.ac.at) is an expansion of the research project MORPH (Mapping, Monitoring and Modelling the Spatio-Temporal Dynamics of Land Surface Morphology, http://morph.zgis.at). The project is funded by the Austrian Science Fund (FWF) as part of the Top Citizen Science (TCS) funding initiative, which aims to include citizen science components in ongoing FWF projects. The main MORPH project focuses on the development of novel methods addressing the spatial-temporal dynamics of surface morphology by integrating various optical and radar remote-sensing data for a study area in Iceland. In connection with this goal of the main project, there is (still) a high demand for data (including images) gathered directly in the field (e.g., recording actual events, landform characteristics or landscape changes). The field data can be used for 3D reconstruction of the surface using Structure from Motion (SfM) and dense image matching (DIM) techniques, for enriching and validating remote-sensing based mapping results, as well as for increasing their detail and information content. The joint availability of field and remote-sensing data is of importance for comprehensive analysis and helps to broaden knowledge about geomorphological landscape dynamics and the prevalence of particular landforms.

Since field data cannot be delivered by scientists only (due, e.g., to time, budget and distance constraints), citizen science (i.e., citizens' data contributions) is beneficial in two ways: first, the contribution of field data and, second, citizens' input to the development of a technological solution that is tailored to their needs, which thus secures the contribution of extensive, high-quality field data. The *citizenMorph* project addresses these issues by developing a pilot solution, in cooperation with citizens, that allows them to contribute field data on landforms regarding mass movements (e.g., rockfall, debris flow), volcanism (e.g., lava flow, lahar), glacial features (e.g., moraine, drumlins), and coastal processes (e.g., cliff erosion).

Although the main project's study area is limited to Iceland, the *citizenMorph* project is aimed at collecting data anywhere in the world, with the collaboration of local citizens and scientists. Consequently, not only is testing the *citizenMorph* solution in the MORPH study area in Iceland key, but testing it in other regions, taking into account different types of landform and landscape dynamics, is also important.

3 Survey123 for ArcGIS

Survey123 for ArcGIS from ESRI was introduced in 2016. It is a simple, form-centric solution for creating, collecting, sharing and analysing so-called smart forms or surveys that allow collecting various types of information (including spatial data) using web or mobile devices (ESRI 2018). In general, a form is an online document that contains different types of questions, and text boxes in which to insert the required information. Multimedia (images, audio and video files) can usually be embedded to support the questions in various ways. Additionally, smart forms contain validation and logic, which means, for instance, that grouping of questions is possible and that people are only asked questions which apply to them (i.e., questions may appear or disappear depending on earlier responses).

The off-the-shelf smart forms created by *Survey123* – using either the online tool (*Survey123 web designer*) or the desktop application *(Survey123 Connect for ArcGIS)* – are in line with these characteristics. There is the possibility of using different question types (closed questions: single- and multiple-choice questions, single-choice grid questions, rating and Likert-scale questions; open questions: adding text, number, date and time, and contributing images). Crucial is the *GeoPoint* question, which allows citizens to report on a location using their mobile device's GPS sensor, or to choose the location themselves on an interactive map. Question logic and grouping of questions can be used (ESRI 2017). In addition, to provide project-relevant and further related information (e.g., to support and guide participants in how to complete the survey), single- and multi-line text boxes and notes can be added along with hints accompanying each question and form field. All these features can be leveraged using *Survey123 web designer*; even more options are available using *Survey123 Connect for ArcGIS*. For instance, repeats of questions to capture multiple versions of the same information can be implemented, and images and audio files can be added to the choices for single- and multiple-choice questions (ESRI 2016).

The comparatively simple structure of *Survey123* makes it an easy-to-use and intuitive tool to create off-the-shelf smart forms. Good support is available through the official ESRI websites, blogs and forums. In particular, creating and sharing surveys by using *Survey123 web designer* is

straightforward. It allows the design of surveys in a short time and without special ICT or GI/GIS expertise. Compared to *Survey123 web designer*, the use of *Survey123 Connect for ArcGIS* is technically more demanding.

Several options exist for distributing and using a *Survey123* smart form. The most common is the use of the *Survey123 field app*, available for Android and iOS, which allows participants to download surveys and start collecting data. Another possibility is to share surveys as a web link (URL, QR Code) that can be opened and filled in through a web browser. Both possibilities provide different capabilities, which depend on either online or offline usage (Table 1).

The data captured and submitted by collaborators can be immediately accessed via the *Survey123* website, which includes various reporting and mapping possibilities. Due to the interoperability of ESRI products, the data collected are available for visualization and analysis in other ESRI products (e.g., ArcGIS Online, ArcGIS Pro). However, the use of ESRI products (including *Survey123)* is not free of charge; it requires a licence. For participants, that does not matter: the *Survey123* smart forms can be completed without having an ESRI account.

Table 1: Selection of Survey123 smart form characteristics: Survey123 field app and browser-based option.

	***Survey123* field app**		**Browser-based option**
Accessibility	mobile application via Apple App Store or Google Play Store plus adding the survey		open URL/ scan QR Code; add to home screen
Update	needs to be loaded on the mobile device, but there is no option to notify the user of a survey update through the native application		no problem using the same URL
Usage	online	offline	online
Information provision	question hints, multi-/ single-line text (expand/ collapse)		
Access to URL, multimedia	accessible	not accessible	accessible
Style changes (font, colour, etc.)	limited – only small font size in info boxes		font size and colour can be adjusted
Use of audio files	limited control over audio playback (e.g., no pausing the audio or fast-forwarding)		control over audio playback possible
Use of images (single-/ multiple- choice questions)	no zooming/ flipping, no external links		provision via external links that allow zooming/ flipping
Basemaps	automatically provided (ESRI basemaps)	custom basemaps, complex to provide on mobile devices	automatically provided (ESRI basemaps)

Map symbols on *GeoPoint* question	no customized ones	
Ongoing participation	easy, asked at the end of the survey-filling process	reloading web page

4 Workflow, methods and tools

Various approaches and methods were used in the development of the *citizenMorph* solution (Figure 1). First, the participatory design approach provided the general idea behind the development process. It allowed the direct, active involvement of target-group representatives in activities such as the specification of requirements, and the design, implementation and testing of the solution; the representatives of the target group were also included in decision-making (see, e.g., Baek et al. 2007). In addition, the stages in the development of the *citizenMorph* system were based on a prototyping process model (Kumar 2003): requirements are specified, prototypes are implemented and discussed in an iterative manner, and finally the final product is designed and implemented (Figure 1).

The citizen representatives (25 high school students, 14 undergraduate students, and eight older adults enrolled in continuing education) contributed to the various tasks, delivered prototypes, and to varying degrees were involved in decision-making. Different methods suitable for involving citizens (ICT, GI/GIS laypeople) were applied (Figure 1). The implementation of prototypes and the final system, in cooperation with citizen representatives, took place using *Survey123 web design* (Version 3.7), *Survey123 Connect for ArcGIS* (Version 3.3.51), and the content management system (CMS) *WordPress* (Version 5.2.5).

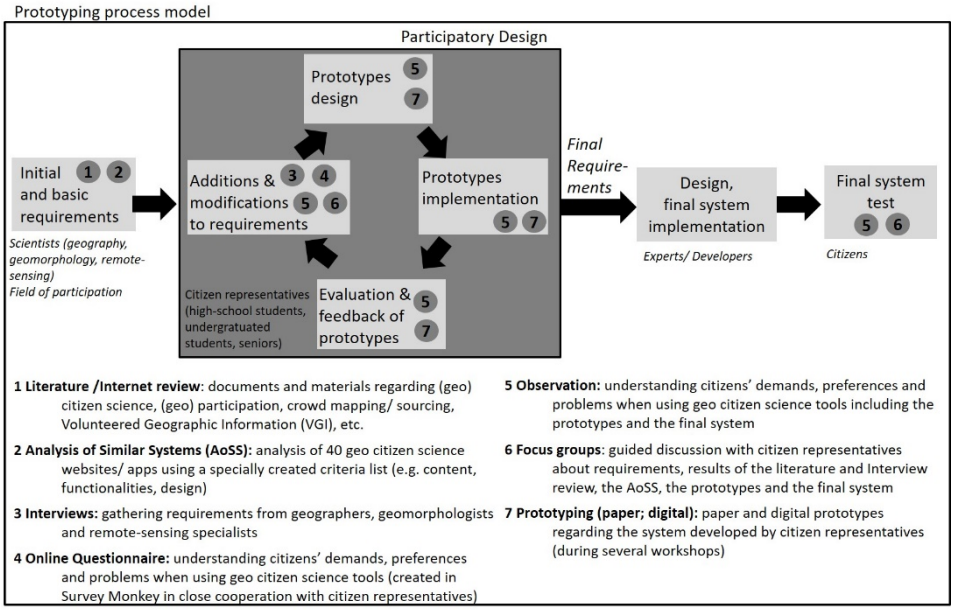

Figure 1: Workflow for the citizenMorph system development

The final *citizenMorph* system was tested and evaluated on four occasions by citizen representatives and experts:

- Excursion to the Berchtesgaden National Park, Germany, and the Weißbach Nature Park, Austria (14 participants: high-school students, seniors, experts; 11 July 2019)
- Excursion with workshop in Höfn, Iceland (15 participants: high-school students and experts; 5 September 2019)
- Workshop at Lomonosov State University Moscow, Russia (8 undergraduate students; 19 September 2019)
- Workshop (international GIS day) at Salzburg University (58 high-school and undergraduate students; 13 November 2019)

Observing those testing the solution (while they were using the *citizenMorph* system to complete the survey) and carrying out focus groups (after they had used the *citizenMorph* system to complete the survey) gave an insight into problems that citizens face using the *citizenMorph* solution. The findings from the observation and focus groups were coded and grouped under the categories 'survey distribution and installation', 'registration and login', 'design and usability', 'data contribution' and 'help and support'.

5 Requirements and system structure

The different stakeholders (experts: geomorphologists, geographers, remote sensing specialists; citizens) have different requirements of the *citizenMorph* technological solution. These include knowledge regarding participating in the research area, people's motivations, their digital skills, and the importance of building and maintaining a project community (Hennig et al. 2019; van Dijk 2012).

To meet the requirements specified (Table 2), the implementation of a system that only allows information to be contributed is not enough. As stressed by Hennig & Begiu (2011) and Murgante et al. (2011), additional components are required, notably information for volunteers about the project more generally, the data collection and reporting processes, security and safety issues, and feedback. The system must also allow social-networking possibilities (i.e. communication and interaction options) and facilitate building and maintaining a project community by participants. All these components are present in the *citizenMorph* system.

Table 2: Selected citizenMorph system requirements

	Requirements/ needs/ preferences
General usability/ design	• Easy to access and use; as self-explanatory as possible; attractive design • Well-written and understandable text; short, dense and well-structured content • Online and offline uses possible for a variety of mobile devices • Use of different media to provide information, support and help

	• Customizable text size (readability); possibility to enlarge and flip images
Help/ support/ guidance	• Information about related domains; project baseline information • Information about how to collect and report data, including support with spatial literacy skills • Information about how to take images (single image, image series) • Information about safety/ security issues (being on-site, intellectual property rights, personal data)
Data contribution	• Intuitive use; comfortable, quick and easy input (only relevant questions) • Support in identifying landforms: guiding users by questionnaire logic and information • Possibility to add a single image of the landform as well as a series of overlapping images (e.g. prerequisite for SfM-based 3D reconstruction) • Possibility to contribute data on-site and/ or at home using different devices • Possibility to edit data entries after submission
Community/ contact	• Directly addressing participants in the context of the project • Direct feedback to participants in the context of the project (optional) • Opportunities for contact and exchange with others (citizens, project team) • Opportunities to gain insight into the project community • Provision and collection of a very limited amount of personal data; not mandatory

6 Survey123 smart form implementation

Even though citizen representatives need to be able to use the survey both online and offline, of the two common ways to share a survey (field app; browser-based variant), the browser-based variant was chosen (https://arcg.is/15WPKv0). By using many of the capabilities provided by *Survey123* (online and desktop tool), it was possible to meet most of the requirements (Figure 2).

Different kinds of information are provided in the *Survey123* smart form: welcoming, project baseline information, support and guidance for collecting and reporting data, data protection information, and thanks and feedback to volunteers. Question hints and multi-line text boxes (possibility to expand/ collapse) allow the delivery of various levels of information to participants (for newcomers: more information; for experienced users: less/ no information). In the case of citizens' strong interest in selected topics, text boxes and question hints include links to additional information available on the *citizenMorph* website. Providing different levels of information to participants has various benefits: it allows the provision of survey-relevant descriptions and guidelines only when needed; it prevents the demotivation of volunteers while

they are familiarizing themselves with and completing the survey, since they are not overwhelmed by unwanted information.

Since users often refuse to read (lengthy) text (Hennig & Vogler 2016), audio files were created in addition to the written text. The audio files were produced by citizen representatives with the aims of attracting and motivating citizen participants to complete the survey and, at the same time, of giving the most relevant information (necessary survey-filling instructions). Having citizen representatives create the audio files, only terms that the citizens themselves are familiar with are used, and it is more likely that the content will be neither too long nor too technical.

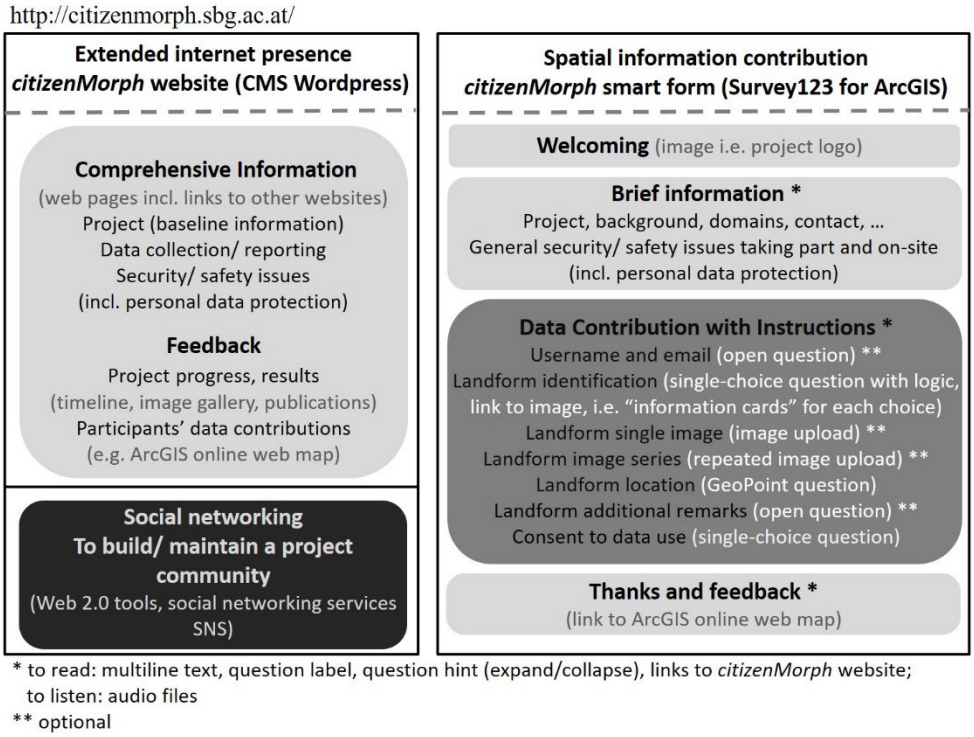

Figure 2: Structure and components of the citizenMorph system and its implementation using Survey123 web design and Survey123 Connect for ArcGIS

The question types used in the *citizenMorph* survey include: (i) open questions to be answered optionally (e.g. participant's username and email; additional information about the landform under investigation); (ii) single-choice questions (select the type of landform; participant's consent on using the data contributed); (iii) the possibility of adding images (landform single image; landform image series using the *repeat question* option); and (iv) a *GeoPoint* question to report the landform location.

To guide and support citizens in the identification of landforms, question logic is used. Thus, a subsequent question (e.g., about landform type/ category) will depend on the answer to an earlier question. For example, if a participant selects the landform category 'glacier', a

subsequent question will display related landforms as options for selection (moraine, rock glacier, etc.). In addition, *information cards* (jpg image presenting the name of the individual landform type/ category, a brief definition, explanatory images) were created to support participants in identifying landforms correctly. It is possible to add images to the answers to single- and multiple-choice questions, but the opened images cannot be zoomed or flipped. This is not in line with user needs for enlarging text and images. Thus, *information cards* were made accessible through links provided in the corresponding question hints. When opened in a browser, they can be zoomed and flipped.

7 Survey123 smart form evaluation

During the test events (focus groups), 80 statements relevant to *Survey123* usage were selected and summarized, quoting the most common statements per category (survey distribution & installation: 10%; registration & login: 5%; design & usability: 21%; data contribution: 35%; help & support: 29%). Together with the findings from observing the testers, these issues are described further in the following sections.

a) Survey distribution and installation

Even though the testers had no problem in making the browser-based survey available on their mobile device, they suggested making a 'typical app' available. From their perspective, the survey would be easier to install, and apps are the standard way of distributing (mobile) web applications today. Using the *Survey123 field app* does not meet testers' demands either, since it requires several steps: download the app from a mobile application store, install it, and add the survey of interest to the app. For use offline, an additional step is loading custom basemaps. For citizens, this is a complicated task that might discourage them from taking part in the survey. However, a major disadvantage of using the browser-based variant is that it is impossible to fill out the survey in remote areas where mobile network connectivity is poor or unavailable. In this case, participants have to wait for an internet connection to fill out the form with images taken earlier, and need to map the landform location from memory or by using recorded coordinates.

b) Registration and login

Even though the use of ESRI products usually requires an account, no registration and login are needed to complete a *Survey123* smart form. Because *Survey123* does not offer the option of participant registration, community building is not possible. The testers discussed the topic of registration and login. Filling out the survey without registration and the possibility of submitting data anonymously were considered positive, simplifying and accelerating the contribution process, and leading to fewer data privacy concerns. However, some testers mentioned that having a registration option would enhance community building and increase data quality. In this context, Jay et al. (2016) found that not having a registration step increases the number of contributors to citizen science projects by more than 60%, but offering the option to create an account, without making it a requirement, maximizes the contribution rates. To address this issue (i.e., support community building), the *citizenMorph* system includes social-networking activities (i.e., communication and interaction options). Further, asking participants for username and email (which are optional) makes it possible to contact those

who supply these details and give personal feedback. It also allows for the delivery of an ArcGIS Online web map showing the individual participant's data input.

c) Design and usability

Using *Survey123,* developers work on a template that is provided. This gives little leeway for the design and layout of surveys to be modified (e.g., using corporate design, customized map symbols). Thus, it is not surprising that a recurrent remark of the testers was that the design of the *citizenMorph* survey was not attractive enough. Several ideas for improving its appearance and design were suggested. Examples are a clear typeface, effective use of images, and the position of buttons, links and arrows.

d) Data contribution

In a broad sense, the testers found the survey intuitive and had no major difficulties answering the questions. They only faced problems when capturing image series and navigating through the *GeoPoint* question. To submit a single image, the process is straightforward (the user is prompted either to use the built-in camera on a mobile device or to access its gallery to find an image). However, the process of submitting image series is described by testers as tedious and annoying. There is no possibility in *Survey123* of taking several images or of selecting them from the gallery in a single step. Instead (using the *repeat question* option), participants need to add each image separately, which requires them to actively add a new field in the survey to upload each new image. In addition, how to map a feature (*GeoPoint* question) was not always clear to the testers. Panning and zooming in the map manually to localize oneself and/or the landform of interest were considered difficult and confusing. Using the get-my-location functionality (based on the mobile device's GPS sensor) was not always evident; feedback here was negative. For some testers, typing their approximate address was a workaround when trying to locate themselves.

e) Help and support

Even though citizens need at least some basic information to start contributing, the testing revealed that most volunteers did not read the information available to guide them through the form-filling process, although this is vitally important. Instead, they attempted to complete the survey directly. Reasons for this are people's general dislike of reading text online and the fact that reading onscreen in an outdoor setting is often difficult (e.g., because of light conditions). In this regard, the audio files proved to be more useful and were mostly positively remarked on by the testers.

8 Conclusion and outlook

Survey123 is a usable tool for cooperating with citizens in the context of participatory design. To meet all requirements of citizens, scientists and the participation domain in general, workarounds (e.g., use of images accompanying question choices) and compromises (e.g., a trade-off between online and offline use associated with different challenges) are required in some cases. In general, using *Survey123 web designer* and *Survey123 Connect for ArcGIS* to create off-the-shelf surveys is in line with off-the-shelf online (map-based) questionnaires: there are few possibilities regarding design and usability, and little focus on community building. Nevertheless, citizen science requires these features since they help encourage people to take

part in initiatives. This underlines the importance of the multi-component nature of the *citizenMorph* system.

Another possibility that allows custom applications to be built is ESRI's AppStudio (incl. QT Creator). Nevertheless, the use of this tool is more demanding. For the *citizenMorph* project, this was not an option (because of limited budget, and citizen involvement in the solution development process). However, ESRI is constantly enhancing its products, fixing bugs, and adding new functionalities with each release. The community of developers who use these tools are able to present their concerns and get support. It is therefore expected that some of the issues we have mentioned will be improved for future releases of *Survey123*.

Acknowledgments

The Top Citizen Science (TCS) project citizenMorph (FWF-TCS 47), and the main project MORPH (FWF-P29461-N29) are funded by Austrian Science Fund (FWF).

References

Ahmed II, J.B. & Pradhan, B. (2019). Spatial assessment of termites' interaction with groundwater potential conditioning parameters in Keffi, Nigeria. *Journal of Hydrology*, 578, 124012.

Baek, E.-O., Cagiltayik, K., Boling, E. & Frick, T. (2007). User-Centered Design and Development. In *Handbook of Research on Educational Communications and Technology,* ed. J. M. Spector, M. D. Merrill, J. J. van Merrienboer, & M. F. Driscoll, (pp. 659–670). New York: Routledge Chapman & Hall.

ESRI (2016). Media. Retrieved January 15, 2020, from https://doc.arcgis.com/de/survey123/desktop/create-surveys/xlsformmedia.htm.

ESRI (2017). XLSForm essentials. Retrieved January 15, 2020, from https://doc.arcgis.com/en/Survey123/desktop/create-surveys/xlsformessentials.htm.

ESRI (2018). Survey123 for ArcGIS starten. Retrieved January 15, 2020, from https://www.esri.com/en-us/arcgis/products/Survey123/overview.

Haklay, M. (2013). Citizen Science and Volunteered Geographic Information – overview and typology of participation. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice,* ed. D. Z. Sui, S. Elwood & M.F. Goodchild, (pp. 105–122). Berlin: Springer.

Hennig, S. & Belgiu, M. (2011). User-centric SDI: Addressing User Requirements in Third-Generation SDI - The Example of Nature-SDIplus. *Perspective,* (pp. 30-42).

Hennig, S.; Hölbling, D.; Ferber, N. & Tiede, D. (2019): Rahmenkonzept und Komponenten für Citizen Science Projekte. Das Projekt citizenMorph. *AGIT – Journal* (5).

Hennig, S. & Voger, R. (2016). User-centred map applications through participatory design: Experiences gained during the "YouthMap 5020" project. *The Cartographic Journal*, 53(3), (pp. 213-229).

Jay, C., Dunne, R., Gelsthorpe, D., & Vigo, M. (2016). To sign up, or not to sign up? Maximizing citizen science contribution rates through optional registration. Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI'16), (pp. 1827–1832).

Kumar S. (2003). What is Prototype model- advantages, disadvantages and when to use it? Retrieved January 15, 2020, from http://tryqa.com/what-is-prototype-model-advantages-disadvantages-and-when-to-use-it/.

Muller, M. J., & Druin, A. (2012). Participatory design. The third space in HCI. In *The human-computer interaction handbook*, ed. J. Jacko, (pp. 1051–68). Hillsdale Erlbaum.

Murray, T. (2018): Tracking ecosystem change through citizen science. Retrieved Dec 25, 2019, from https://www.biodiversityireland.ie/wordpress/wp-content/uploads/2018_10_18_TrackCitSci_TMurray.pdf.

Murgante, B., Tilio, L., Lanza, V. & Scorza, F. (2011). Using participative GIS and e-tools for involving citizens of Marmo Platano–Melandro area in European programming activities. *Journal of Balkan and Near Eastern Studies,* 13(1), (pp. 97–115).

Steen, M., L. Kuijt-Evers, & Klok. J. (2007). Early user involvement in research and design projects—A review of methods and practices. 23rd EGOS Colloquium, Vienna, Austria.

Tsou, M.-H. & Curran, J. M. (2008). 'User-centered design approaches for web mapping applications: a case study with USGS hydrological data in the United States', in *International Perspectives on Maps and the Internet. Lecture Notes in Geoinformation and Cartography*, ed. M.P. Peterson, (pp. 301–321), Berlin: Springer.

van Dijk, J. (2012). The Evolution of the Digital Divide - The Digital Divide Turns to Inequality of Skills and Usage. *Digital enlightenment yearbook 2012,* ed. J. Bus et al., (pp. 57-75), Amsterdam: IOS Press.

Urban Activity Detection Using Geo-located Twitter Data

GI_Forum 2020, Issue 1
Page: 15 - 31
Full Paper
Corresponding Author:
anna.kozlowska@ait.ac.at
DOI: 10.1553/giscience2020_01_s15

Anna Kozlowska[1] and Klaus Steinnocher[1]
[1]Austrian Institute of Technology GmbH, Austria

Abstract

More and more studies are based on freely available social media data. Using microblogs, a midpoint between instant messaging and content production, analyses of urban activities are possible. This paper focuses not only on mapping human activities but also on defining urban function in the city. Using geotagged Twitter data, the research carried out separate spatial and temporal analyses, in conjunction with combined spatio-temporal analyses. Tweets were categorised into six activity groups: *Working*, *Eating*, *Shopping*, *Leisure*, *Home* and *Education*, based on selected keywords. The results show stronger performance for the detection of *Leisure*, *Eating*, *Shopping* and *Education* activities and less successful performance for *Working* and *Home* activities. The first four cluster near the centre of the city, while the rest are scattered all over the city. Moreover, each activity shows its own temporal pattern. This study finds characteristic patterns for everyday activities and shows the possibility of using social media data to define urban function for places where land-use information is not available.

Keywords:

geotagged Twitter data, spatial analysis, temporal analysis, urban data, social media

1 Introduction

Leveraging location-based data offers new perspectives on, and better understanding of, events taking place in the world. Studying human behaviour and activity using social media was not possible a decade ago. That changed when social networks grew and the use of the Internet increased. The availability of data has made Twitter one of the most popular data sources for scientific research. With 320 million accounts creating over 500 million messages a day (*The Number of tweets per day in 2019*, 2019), Twitter is one of the largest social networks. It is also one of the preferred platforms for large-scale studies of human behaviour, thanks to its openness, global range, and the large number and variety of its users (Steinert-Threlkeld, 2018). Microblogs such as tweets help to validate socio-economic theories, predict social phenomena, or find spatial, temporal or thematic patterns in society. Moreover, according to Juhász & Hochmair (2019), among social-media microblogs, tweets relate the best to locations of daily activities.

While most of the literature focuses on the text of the tweets (Miller, 2011), few studies use the geographic information attached to the tweets (Hawelka et al., 2014; Leetaru et al., 2013). Geotagged text strings from Twitter are used mostly in research on social relationships and human dynamics. For example, they can be a support for analysing spatio-temporal patterns of happiness and public sentiment (Cao et al., 2018; Dodds et al., 2011; Nguyen et al., 2016), mobility (Hawelka et al., 2014; Kurkcu et al., 2016), or crime counts (Vomfell et al., 2018). Several studies use geo-located social-media data to track activities. Sakaki et al. (2010) used tweets to detect certain big events, like earthquakes or accidents, by searching for keywords related to the events. A different approach was presented by Martín et al. (2019), who used top tweeted words to obtain a clear idea about activity on a specific day. Zhang et al. (2018) used geo-tagged photos collected from social media to learn about principal tourist destinations.

Contrary to previous studies, the work presented here focuses more on detecting everyday activities than looking at extraordinary events. The objective is to learn about urban function in different parts of the city in order to determine urban land use. Land-use classification using social media data has already been carried out by Jiang et al. (2015), but their work was based on POI data rather than textual information.

The work closest to our approach was done by Andrienko et al. (2013). Although they used spatial and temporal clustering to analyse different activities, the choice of activities was based on the most frequently used keywords in their dataset. Also, unlike our study, it did not show hourly or daily spatial distributions. In pursuing the goal of this study, spatial, temporal and spatio-temporal descriptive analyses of geo-located data from the City of Manila, Philippines were carried out. The spatial analyses are based on Kernel density estimation. As shown in previous studies, this method is useful when analysing changes in density distribution of chosen events (Ma et al., 2009; Polonczyk & Lesniak, 2018; Zhang et al., 2009).

The paper is organised as follows: Section 2 outlines the data collection, pre-processing and activity classification; in Section 3, we present the analyses and results; Section 4 provides discussion and concludes the paper.

2 Methodology

2.1 Data collection

Twitter, as described by its owners, is 'what's happening in the world and what people are talking about right now' (*Twitter About Page*, 2019). It is an online social networking service where anyone can post short text messages ('tweets', max. 280 characters) and interact. Communication takes place in real time, by posting a message, commenting on a message, or redistributing another user's message (retweet). The message may also include a picture, a video or a link. Certain information can be marked with a hashtag '#', which facilitates the search for tweets within a chosen topic.

The study is based on geo-located Twitter data, which means that only tweets with an assigned location are used. On Twitter, the location can be set automatically by activating the precise

location option from the user account or the mobile device; alternatively, it can be set manually, each time a post is uploaded, by selecting a location from a predefined list. While the first option gives precise information on longitude and latitude, the degree of precision for the second ranges from the name of a city or neighbourhood to that of a specific public place (e.g. the name of a restaurant or other point of interest recognised by the Twitter service). Only tweets with original content can be georeferenced. Retweets, which are not classified by Twitter as original content, cannot be geotagged.

The Twitter data was acquired using the streaming API (Application Programming Interface) and the R Studio environment, with twitteR and streamR packages. The connection to the Twitter Search API was created through a Twitter account and Access-Token. Twitter provides data encoded in JavaScript Object Notation (JSON), which is based on key value pairs with named attributes and related values. All core attributes that accompany the tweet are encapsulated in that format. Each record stores the text of the tweet, the exact time of its publication and, in this case, information on geolocation. Whenever a tweet is georeferenced, a combination of the JSON keys 'geo', 'coordinates' and 'place' is filled with values. Specifically, each geo-tweet contains exact coordinates (longitude/latitude) in WGS84 as a single point.

The subject of this study was the City of Manila, Philippines. Data was collected for 9 months (20.06.2016 – 03.04.2017), with a total number of 608,667 tweets. The datasets were stored in CSV (Comma Separated Values) files.

Manila is a 'perfect' use case for any twitter data analysis. It is the world's most densely populated city, with an area of 42.88 km² and 1.78 million inhabitants (*Manila Population*, 2019). At the time of this study, the Republic of the Philippines was one of the most Twitter-active spots in the world (Figure 1), with approximately 200,000 tweets posted per day (*The one million tweet map*, 2017).

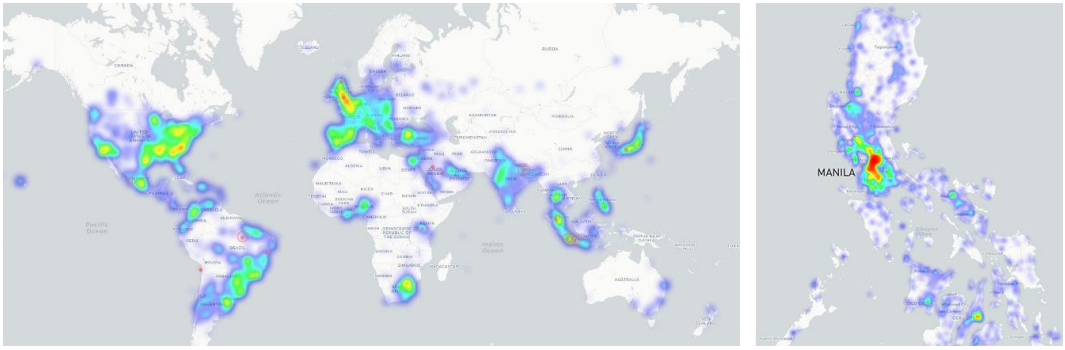

Figure 1: Twitter activity around the world (The one million tweet map, 2017).

2.2 Data processing

As argued by Symeonidis et al. (2018), pre-processing is a necessary and very important step in any analysis of text strings. Before addressing the actual content of the tweets, the geo-

location information was verified. All tweets lacking spatial reference and all location outliers (tweets lying outside the area of interest) were removed from the dataset.

In general, social media data are characterized by a large amount of noise. This noise includes all the special characters and punctuation embedded in a text string. In order to perform a successful classification and limit erroneous results, a clean text string is essential. Therefore, it is necessary to carry out several steps of text cleansing (Hangya & Farkas, 2013; Symeonidis et al., 2017). As most studies use tokenization – dividing the text string into separate words (Balazs & Velásquez, 2016) – this approach was also applied here. Furthermore, it was necessary to detect and delete duplicates and retweets. Later, the following techniques were used:

- Removing Unicode characters like comma (u002c) and unnecessary characters <, >, “, $
- Removing URLs which are part of most tweets but are not useful for the analysis and might also release sensitive information
- Unifying user tags (user account preceded by the ‘@’ symbol)
- Removing whitespaces
- Removing ‘#’ (commonly used on Twitter to categorize tweets).

After normalization, the next step was to identify and remove spam messages. A large group of tweets deemed not to be useful for this study were those generated automatically. Two types were detected and removed: tweets created by a bot (web robot), e.g. job offers and weather forecasts; tweets created automatically though external web sources like apps for music, running or games.

The final preparatory step was sorting date and time information. In this dataset, the time zone had to be corrected, from Greenwich Mean Time, by adding 8 hours (GMT+8). The time formats were normalized and additional information about the month (January–December), number of the week (1–52) and day of the week (Monday–Sunday) was assigned.

In the Philippines, there are around 200 unique languages and dialects, and two official languages: Filipino (Tagalog) and English. Therefore, before the text analysis the predominant language for all tweets was identified. If the tweet was too complex or there was no leading language in the text, the dominant language was not defined. The language detection showed that at least 2/3 of tweets in the dataset were written in English (Figure 2). The other 1/3 of the tweets may still contain English words. Hence, it was decided to proceed with text analysis for the entire dataset in English.

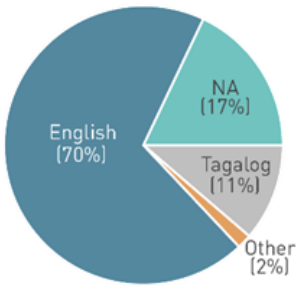

Figure 2: Results of language detection

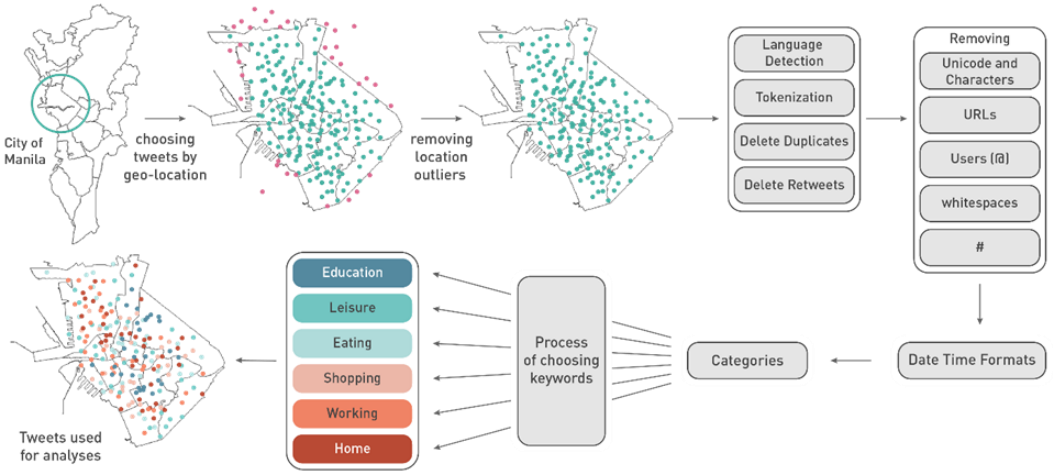

Figure 3: Cleansing data and activity classification process

2.3 Activity classification

Activity mapping and pattern analyses were based on tweets allocated to one of the six chosen activities: *Education*, *Eating*, *Leisure*, *Working*, *Home* and *Shopping*. *Education* covers studying at the locations of universities and schools; *Leisure* includes indoor and outdoor free-time activities, e.g. music events, cinema; *Eating* refers to eating and drinking, and places like restaurants, bars and cafés; *Shopping* refers to buying products in certain locations, e.g. malls; *Working* and *Home* refer to posts from work or home respectively, and do not match to any of the other groups. The allocation was done by choosing the relevant keywords for each activity. In this step, manual classification was chosen over automated text detection techniques. According to Hahmann (2014), classification of tweets by humans, although a subjective process, is more accurate.

The first simplified attempt to categorize tweets gave misleading results. Some of the tweets were wrongly assigned due to an ambiguity of individual word combinations. Expressions like 'after' or 'before' in tweets like 'Lunch before going to work' might suggest an action not necessarily happening at the place the tweet is posted. Phrases like 'going to', 'on my way', 'off to' express movement rather than an activity. Posts like 'Shopping for office supplies' are incorrectly classified to more than one group (*Shopping* and *Working*). Expressions such as 'feel like home' or 'second home' ought not to be classified as *Home*, just as 'working future' or 'work angels' do not concern actual *Working*. Moreover, activities related to *Eating*, *Shopping*, *Education* or *Leisure* are not considered if a location fitting another activity is included, e.g. 'Having lunch at University'. To correct these errors, a further group of keywords and phrases to be excluded from activity groups was created. Figure 4 shows a sample of keywords included and excluded from groups. As a result, only 14% of tweets (86,007 tweets) were successfully categorized (Figure 5).

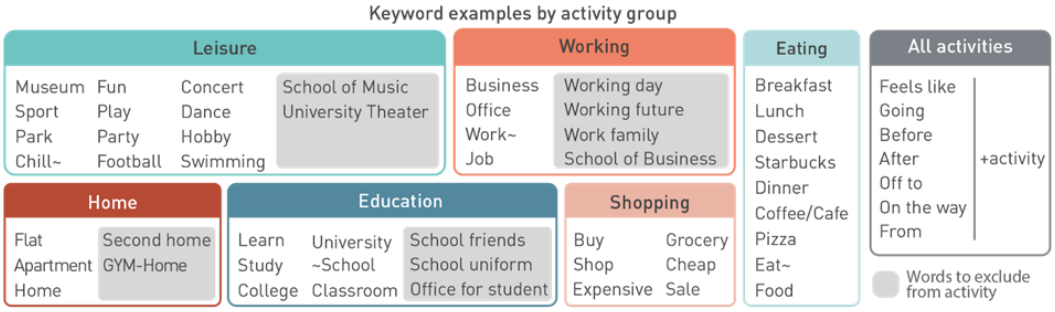

Figure 4: Selection of keywords for each activity group

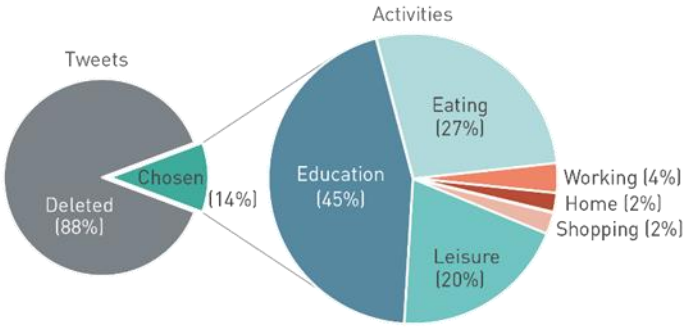

Figure 5: Final dataset used for analysis

3 Spatial and temporal analyses

Once the tweets have been correctly sorted into the six activity groups, the data is ready to be explored and analysed within the spatio-temporal context. The dataset stores information about date and time, as well as longitude and latitude. Thus, both temporal and spatial analyses can be performed. The results from the analyses presented below are therefore divided into three types: temporal analysis, spatial analysis and spatio-temporal analysis. By carrying out these analyses separately, we can gain an insight into when the activities are more present in the city, where people spend time depending on the activity, and how the patterns change throughout the day or week.

3.1 Temporal analysis

Looking at the total number of tweets and their temporal distribution, it is clear from the results that the numbers of posts vary throughout the period analysed (Figure 6). It can also be noted that for some dates data are missing (gaps in Figure 6). In general, most of the peaks are observed on Saturdays, when Leisure appears to be more present. The fluctuation in data can also be explained by the Education cycle of school holidays and the major exam period. Some changes might be related to national holidays or celebrations. It is interesting to note that Eating appears to have the least fluctuation throughout the period investigated.

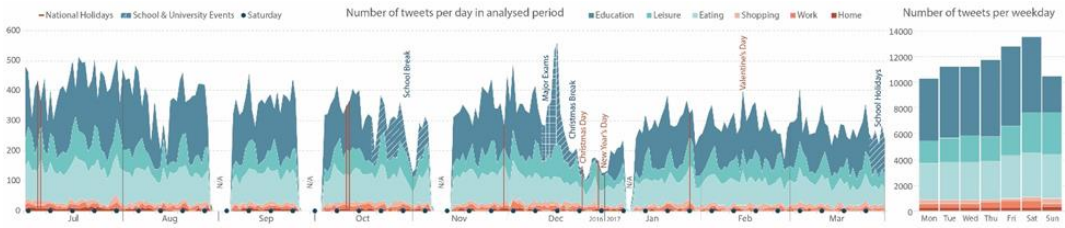

Figure 6: Number of tweets and major events taking place in Manila during the period analysed

The temporal patterns of each activity group were analysed according to daily, weekly and monthly distributions. It was expected that each activity group would have its own characteristic temporal pattern. Using the example of *Home* and *Working*, these activities were expected to have opposite patterns, because most people live and work in different places. Another expectation was that tweets related to *Education* or *Working* would be more intense during traditional working hours (Monday to Friday, 8am-6pm), while all activities analysed were expected to reduce to zero during night-time hours.

The temporal analyses of the general weekly trend (Figure 7a) show an expected pattern. Starting in the morning, the number of tweets rises throughout the day, reaching a peak in the evening hours, especially on Fridays and Saturdays, and finally declining during the night. The results for each activity group (Figure 7b) show more differentiated patterns.

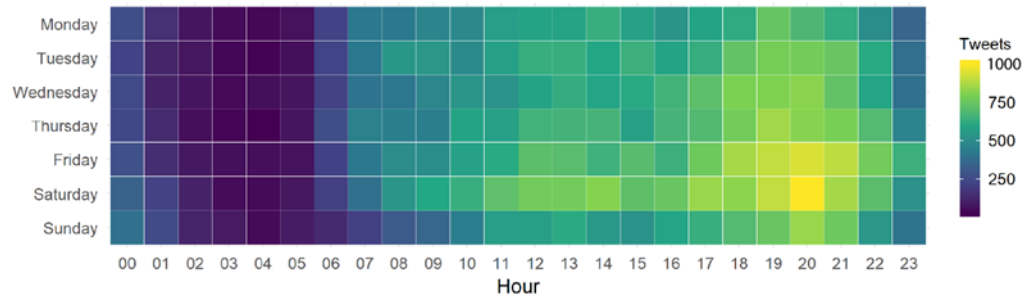

Figure 7a: Temporal heatmaps for each day of the week for all tweets

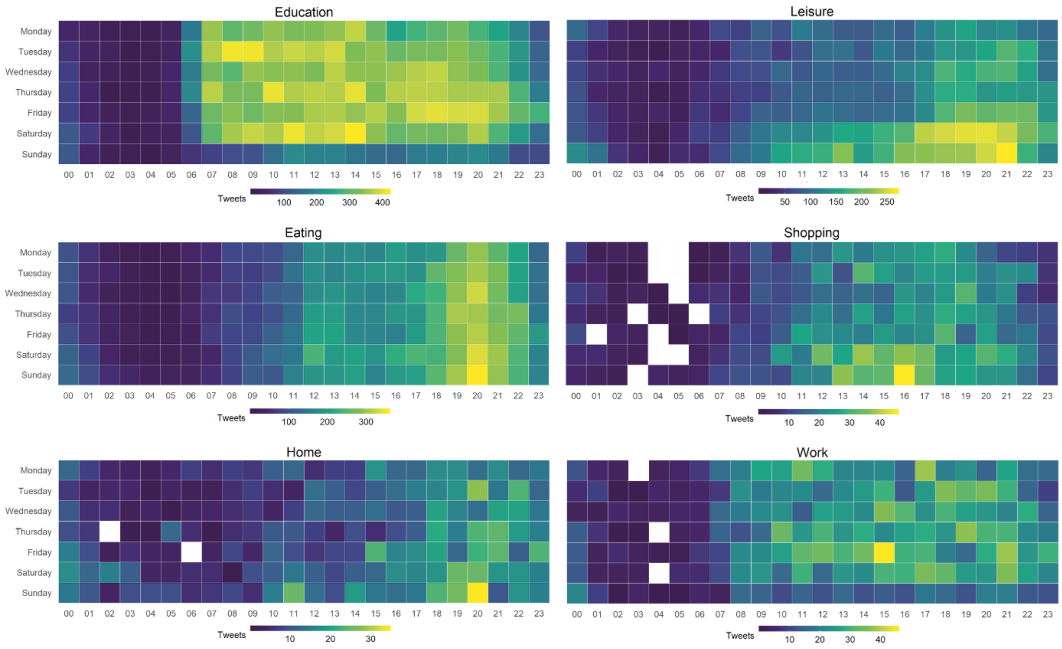

Figure 7b: Temporal heatmaps for each day of the week for each activity group

Tweeting activity related to *Education* is characterized by a regular pattern for the entire week excluding Sunday. Tweeting starts at 7am, with the number staying at a high level until the end of the day. There are no significant peaks or troughs. *Leisure* and *Eating activities* are also characterized by definable and stable, but contrasting, weekly-cycle patterns. For *Eating*, the first concentrated activity takes place between 12am and 2pm, and the main peak is between 7pm and 9pm regardless of the day of the week. By contrast, *Leisure* is characterized by significant peaks, mostly on weekend afternoons and in the evenings. *Shopping*, *Home* and *Working* show irregular temporal distribution. While *Working* differentiates between inactive Sundays and the rest of the week, the other two activities do not present a significant pattern.

3.2 Spatial analysis

To identify where activities take place across the city, a statistical analysis of spatial point patterns was carried out. The study focuses on visual analysis using 2D Gaussian Kernel density estimation, where a spatial relationship of tweets is visualized as a density surface using a graduated colour scheme. The result is a collection of density maps – heatmaps – with a spectrum of 'high' and 'low' point densities. The Kernel bandwidth in this case was based on a number of educated trial runs. The aim was to show the targeted distributions in a more interpretable way and to avoid over- or underfitting. First, the analysis was run for the entire dataset, then for each of the six activity groups separately.

It was expected that each activity would have its hotspots (clustering occurrences) in multiple locations throughout the entire city. This could reveal both overlaps and clear distinctions between activities. Social activities like *Leisure*, *Eating* and *Shopping* were predicted to take place

in close proximity to each other or even at the same location, and most likely in the centre of the city. *Education* activity was expected to show up around the main university campuses and schools, while it was anticipated that *Working* activity would be spread through the entire city, with a focus on the main business areas. *Home* activity was foreseen to be the most widespread activity, as citizens live in different parts of the city. To allow an informed analysis of the spatial distribution of tweets and their underlying localities, a land-use map of the City of Manila was used to compare the results of the analysis with the actual distribution of land use (Figure).

Figure 8: Land use map of City of Manila. Information source: City Planning and Development Office Manila (2017)

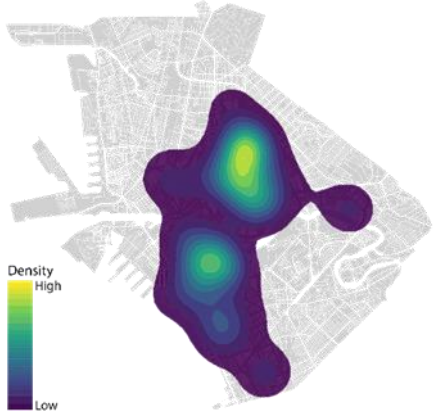

Figure 9: Heatmap for all activities combined

The density analysis shows that tweeting is not spread evenly across the city, with most tweets being located in the city centre (Figure). Analysis shows clear variations between categories (Figure). The most widespread activities are *Home* and *Working*. Tweeting from work occurs in most locations in the city. *Home* activity omits industrial and recreational areas along the riverbank and the northern industrial area. Moreover, *Home* hotspots do not overlap with *Working* activities. The third most widespread activity is *Eating*, covering almost half of the city, mostly where commercial, retail, recreational and institutional areas are located. Much more

concentrated and showing more details are *Shopping*, *Education* and *Leisure*. The hotspots for these activities were compared with corresponding places in the land-use map (Figure). *Education* tweets occur around the main universities and schools. *Shopping* coincides with areas where large shopping centres are located, but does not cover most of the commercial areas and local markets. *Leisure* is a very complex group as it comprises several kinds of activity. The analysis reveals that the main concentration is on the south side of the city, with a smaller spot in the north. These areas are where most of Manila's museums and parks are located. There is no spatial link detected for other entertainment and nightlife locations. The spatial analysis shows that there is a higher chance of finding a more precise location corresponding to *Education*, *Shopping*, *Leisure* or *Eating*, than for *Working* or *Home*.

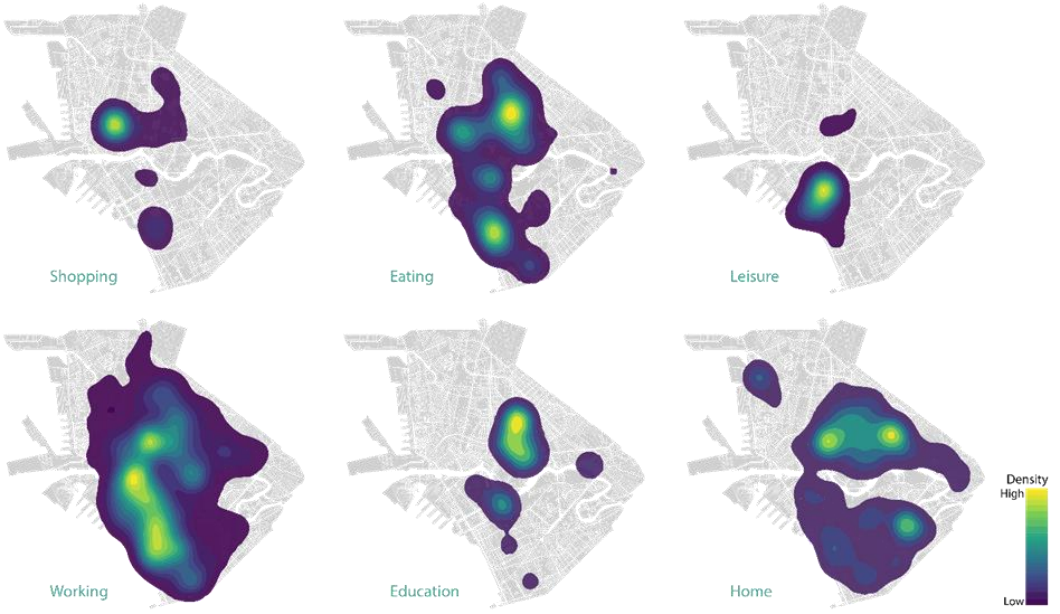

Figure 10: Heatmaps for each activity: Shopping, Eating, Leisure, Working, Education, Home

Figure 11: Heatmaps for selected activity groups checked against identified locations in the city of Manila. Left to right: Education, Shopping, Leisure

The contours representing the extent of the spatial distribution of six activities were overlaid, as shown in Figure, a simplified urban function map which helps to narrow down the areas where everyday activities take place in the city.

Figure 12: Urban function map based on Twitter analysis

3.3 Spatio-temporal analysis

Finally, the combination of both temporal and spatial aspects was analysed. Using time stamps for each point of data allows the mapping of tweets for selected time periods. As for the previous analyses, this analysis focused on different time scales, for various activity groups. This time, it was expected that the spatial distribution of the tweets would differ depending on the time frame. It seemed more likely that more significant results would be obtained from daily or weekly distributions than monthly ones.

Figure shows monthly and weekly distributions of all tweets, for which there are no significant variations. They vary only slightly from one month to another, and between days of the week, showing slight differences in intensity for some areas. As the interest lies in differences between activities, the next step was to explore hourly differences depending on the results from the temporal analysis. Due to the irregularity in the size of the activity groups, the analysis was done only for *Eating*, *Education* and *Leisure*.

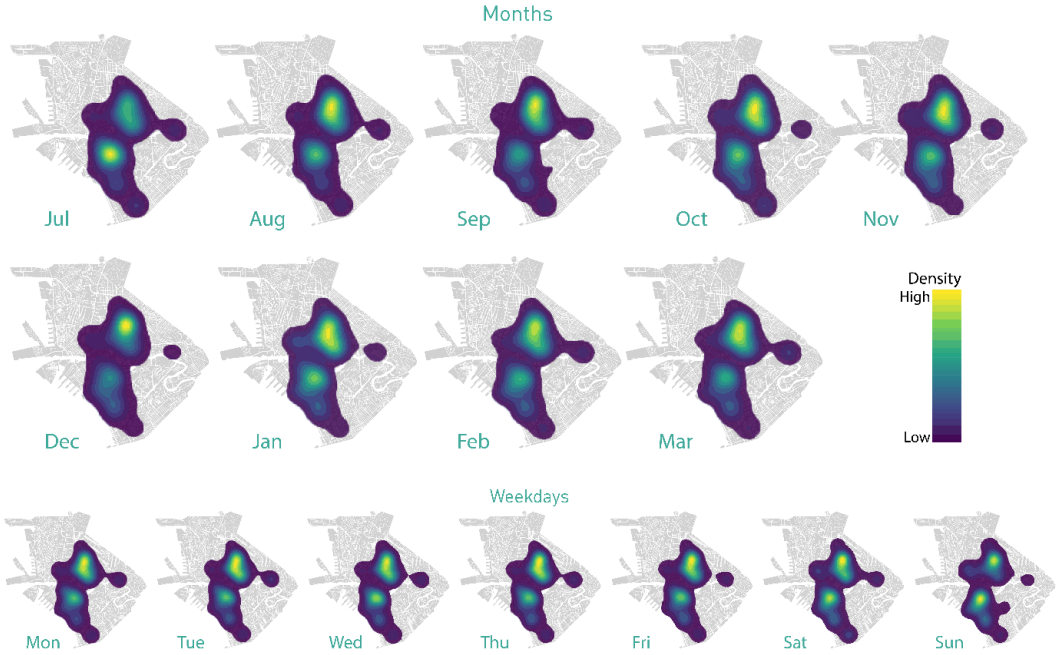

Figure 13: Heatmaps for all tweets, months (top) and weekdays (bottom)

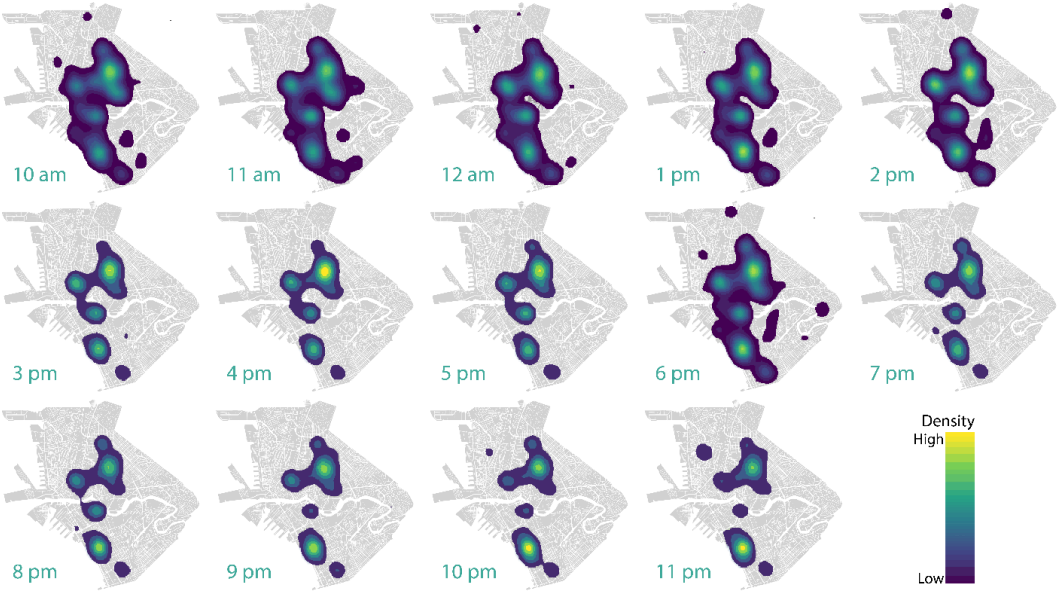

Figure 14: Heatmaps for Eating (10am to 11pm)

Eating and *Leisure* were examined for hourly patterns for a reference day derived as an average from the whole period under investigation (Figure 14, Figure 15). Results show that tweets

related to *Eating* are posted from a wider range of locations between 10am and 2pm and at 6pm. Their locations can be seen to intersect with the *Working* spatial distribution. Between 3pm and 5pm, and from 9pm to 11pm, the action is more focused on certain spots. Between 3pm and 5pm, *Eating* overlaps with *Education* and *Shopping* locations, with the strongest focus in the north where one of the universities is located. The second timeframe (9pm to 11pm) shows a pattern similar to *Leisure* and *Shopping*, with activity in areas where bars and restaurants are concentrated.

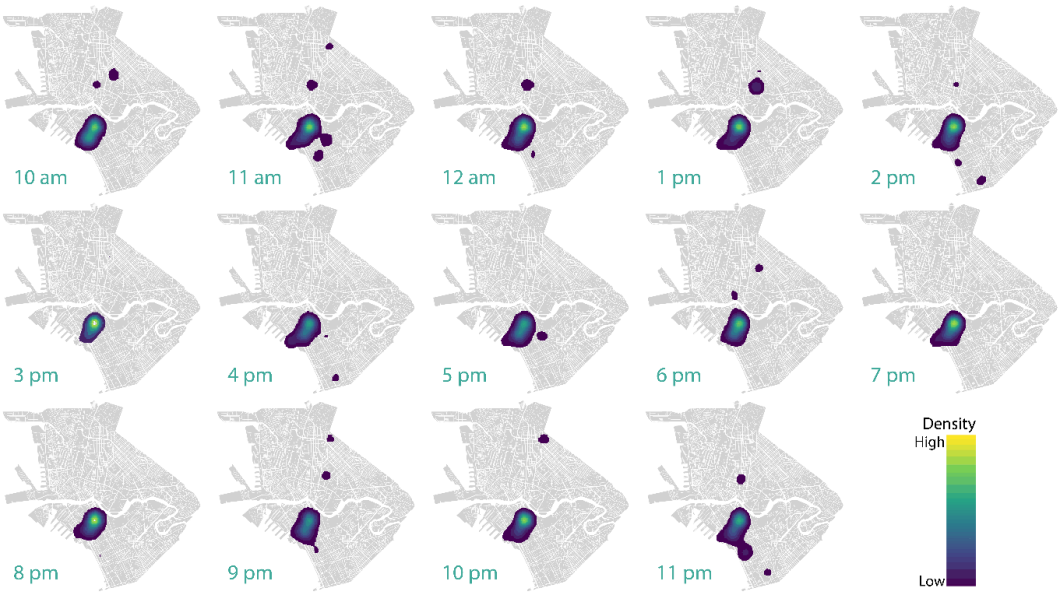

Figure 15: Heatmaps for Leisure (10am to 11pm)

The spatial daily distribution of *Leisure* is presented for a typical Saturday derived as an average over the whole period (Figure 15). The plot shows no significant hourly changes. The main hotspot identified earlier is still visible, with additional spots occurring with no apparent pattern. *Education* was divided into two temporal periods: 1) more activity from Monday to Saturday; 2) less activity on Sunday. *Leisure* shows more hotspots spread across the city, while *Education* has its focus mostly in the University area (Figure 16).

Figure 16: Heatmaps for Education

The results show the change in spatial distribution throughout the day or week but, in most cases, this distribution does not relate to the temporal patterns examined earlier. The intersecting spatial distribution of two or three activities suggests the hourly or daily changes of functionality of different parts of the city. *Working* or *Education* can be replaced by *Eating* during lunch breaks or by *Leisure* in the evening.

4 Discussion and Conclusion

The analyses presented here show the advantages of using data from social media for defining temporal and spatial patterns in the city. Results derived from analysing Twitter data can be used to determine urban function and list major activities taking place in certain parts of the city. The spatial aspect represents gatherings of people and main points of activity. The temporal aspect makes it possible to estimate an average time pattern for individual activities, which can be used to improve various aspects of the urban context, such as public transport or safety. The combination of time and location helps us to understand how the spatial distribution changes during the day. The study shows the effectiveness of using social media for detecting activities which are connected to social interactions or public spaces. It also shows that mapping *Education*, *Leisure* and *Eating* is more precise than mapping *Working* and *Home*; *Working* and *Home* had significantly fewer tweets than the rest of the activities. There are several reasons for this. Firstly, not everyone is willing to share their real location on social media, preferring instead to tag one of the locations from Twitter's predefined list of POIs. As Ludford et al. (2007) show, most people are likely to share about activities in public places, but they are not willing to share the exact location of their home or workplace. Moreover, most people are active on social media when they want to share exciting news, new locations or interesting events in which they are participating, and these are most likely to happen outside their work and home. Furthermore, tweeting about spending time at home or work is not particularly common among users of Twitter. The small sample size for *Working* and *Home* might have been a reason for the fluctuations observed in the temporal analyses.

The temporal and spatial patterns do not show a 1:1 relationship. It was expected that a temporal peak would be reflected in an even stronger peak in the spatial distribution. However, while there is an increase in tweets reflected in higher activity levels at individual hotspots, there is no significant increase in the number of activity hotspots.

Leisure differs from other groups because it combines more than one activity and can take place in different locations. It can be associated with morning or evening activities, or both. The activities can also be referred to as outdoor and indoor events. As a result, this category is very complex, which explains the lack of conclusive results in the *Leisure* analyses. This situation could be improved by predefining more activity groups related to leisure.

Although multiple studies show the strength of Twitter data for understanding urban processes, there are several drawbacks and limitations to using social media data in analyses. They depend on people being willing to share their opinions and feelings with the public. Some text messages comprise incomplete sentences or words reserved for certain social groups, which might be misinterpreted during the word classification. Moreover, the reliability of Twitter data analysis for cities like Manila is compromised by the high number of languages

used, some of which are not easily translated or widely understood. As this study was carried out entirely in English, the results might not reflect all activities typical for the region. According to Longley et al. (2015), analyses using microblogs do not represent the whole of society but only certain demographic groups. However, this should not have a severe impact on this study, as land-use classification does not necessarily depend on demographic or social groups, and thus does not require a complete representation of society. Another source of error are wrong locations assigned to tweets. Hecht et al. (2011) stated that for 34% of the tweets they analysed, the location was wrongly assigned, mainly due to the deliberate indication of a false location. Moreover, only 1% of tweets can be freely downloaded, while the full dataset is very expensive (Morstatter et al., 2013). Choosing only geo-located tweets reduces the sample size even more.

However, despite these limitations, the study brings a new perspective to using social-media data. Using Twitter data only, it is possible to learn about everyday activities taking place in a chosen area. Twitter data can be leveraged to provide information about hourly, daily or weekly patterns for common activities, especially those taking place in public spaces. More importantly, this study shows that by using Twitter data, it is possible to define urban function for places where land-use information is not available.

Acknowledgements

This study is a part of the project INTERSENSE funded by FFG, Vienna, in the frame of the Austrian Space Applications Programme, contract number 865977.

References

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., & Thom, D. (2013). Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics. *Computing in Science Engineering*, *15*(3), 72–82. https://doi.org/10.1109/MCSE.2013.70

Balazs, J. A., & Velásquez, J. D. (2016). Opinion Mining and Information Fusion: A survey. *Information Fusion*, *27*, 95–110. https://doi.org/10.1016/j.inffus.2015.06.002

Cao, X., MacNaughton, P., Deng, Z., Yin, J., Zhang, X., & Allen, J. G. (2018). Using Twitter to Better Understand the Spatiotemporal Patterns of Public Sentiment: A Case Study in Massachusetts, USA. *International Journal of Environmental Research and Public Health*, *15*(2). https://doi.org/10.3390/ijerph15020250

City Planning and Development Office Manila. (2017). *Existing Land Use Map 2017*. https://upload.wikimedia.org/wikipedia/en/8/82/Existing_Land_Use_Map_of_Manila_2017.jpg

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, *6*(12), e26752. https://doi.org/10.1371/journal.pone.0026752

Hahmann, S., Purves, R., & Burghardt, D. (2014). Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. *Journal of Spatial Information Science*, *9*, 1–36. https://doi.org/10.5311/JOSIS.2014.9.185

Hangya, V., & Farkas, R. (2013). Target-oriented opinion mining from tweets. *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*, 251–254.

https://doi.org/10.1109/CogInfoCom.2013.6719251
Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260–271. https://doi.org/10.1080/15230406.2014.890072
Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11*, 237.
https://doi.org/10.1145/1978942.1978976
Jiang, S., Alves, A., Rodrigues, F., Ferreira, J., & Pereira, F. C. (2015). Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, *53*, 36–46.
https://doi.org/10.1016/j.compenvurbsys.2014.12.001
Juhász, L., & Hochmair, H. (2019). Comparing the Spatial and Temporal Activity Patterns between Snapchat, Twitter and Flickr in Florida. *GI_Forum*, *1*, 134–147.
https://doi.org/10.1553/giscience2019_01_s134
Kurkcu, A., Ozbay, K., & Morgul, E. F. (2016). Evaluating the Usability of Geo-located Twitter as a Tool for Human Activity and Mobility Patterns: A Case Study for New York City.
Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, *18*(5).
https://doi.org/10.5210/fm.v18i5.4366
Longley, P. A., Adnan, M., & Lansley, G. (2015). The Geotemporal Demographics of Twitter Usage. *Environment and Planning A: Economy and Space*, *47*(2), 465–484.
https://doi.org/10.1068/a130122p
Ludford, P. J., Priedhorsky, R., Reily, K., & Terveen, L. (2007). Capturing, sharing, and using local place information. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1235–1244. https://doi.org/10.1145/1240624.1240811
Ma, J., Yang, S., You, J., & Zhang, M. (2009). Spatial Pattern Detection and BP Neural Network Analysis of Bank Mesh Point in Urban Area. *2009 Fifth International Conference on Natural Computation*, *3*, 639–643. https://doi.org/10.1109/ICNC.2009.473
Manila Population. (2019). http://worldpopulationreview.com/world-cities/manila-population/
Martín, A., Julián, A. B. A., & Cos-Gayón, F. (2019). Analysis of Twitter messages using big data tools to evaluate and locate the activity in the city of Valencia (Spain). *Cities*, *86*, 37–50.
https://doi.org/10.1016/j.cities.2018.12.014
Miller, G. (2011). Social Scientists Wade Into the Tweet Stream. *Science*, *333*(6051), 1814–1815. https://doi.org/10.1126/science.333.6051.1814
Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. 9.
Nguyen, Q. C., Kath, S., Meng, H.-W., Li, D., Smith, K. R., VanDerslice, J. A., Wen, M., & Li, F. (2016). Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity. *Applied Geography*, *73*, 77–88.
https://doi.org/10.1016/j.apgeog.2016.06.003
Polonczyk, A., & Lesniak, A. (2018). The Impact of Generalised Spatial Data on the Incidence Density of Selected Offences in Krakow. *2018 Baltic Geodetic Congress (BGC Geomatics)*, 328–334. https://doi.org/10.1109/BGC-Geomatics.2018.00068
Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. *Proceedings of the 19th International Conference on World Wide Web*, 851–860. https://doi.org/10.1145/1772690.1772777
Steinert-Threlkeld, Z. C. (2018). *Twitter as Data* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108529327

Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, *110*, 298–310. https://doi.org/10.1016/j.eswa.2018.06.022

Symeonidis, S., Effrosynidis, D., Kordonis, J., & Arampatzis, A. (2017). DUTH at SemEval-2017 Task 4: A Voting Classification Approach for Twitter Sentiment Analysis. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 704–708. https://doi.org/10.18653/v1/S17-2117

The Number of tweets per day in 2019. (2019). David Sayce. https://www.dsayce.com/social-media/tweets-day/

The one million tweet map. (2017). http://onemilliontweetmap.com

Twitter About Page. (2019). https://about.twitter.com/en_gb.html

Vomfell, L., Härdle, W. K., & Lessmann, S. (2018). Improving crime count forecasts using Twitter and taxi data. *Decision Support Systems*, *113*, 73–85. https://doi.org/10.1016/j.dss.2018.07.00

Zhang, W., Tan, G., Lei, M., Guo, X., & Sun, C. (2018). Detecting tourist attractions using geo-tagged photo clustering. *Chinese Sociological Dialogue*, *3*(1), 3–16. https://doi.org/10.1177/2397200917752649

Zhang, Z., Li, J., & Liu, Y. (2009). GIS-Based Spatial Distributions and Evolvement Analysis of Urban Affordable Housing: A Case Study. *2009 International Conference on Environmental Science and Information Application Technology*, *2*, 419–422. https://doi.org/10.1109/ESIAT.2009.130

GI_Forum 2020, Issue 1
Page: 32 - 46
Full Paper
Corresponding Author:
renato.souza@oeaw.ac.at
DOI: 10.1553/giscience2020_01_s32

Uncertain Spaces, Uncertain Places. Dealing with Geographic Information in Digital Humanities: The Example of a Language Legacy Dataset

Amelie Dorn[1], Renato Rocha Souza[1], Barbara Piringer[1] and Eveline Wandl-Vogt[1]
[1]Austrian Academy of Sciences (ÖAW), Austria

Abstract

In addition to their purely linguistic content, legacy language collections often contain other information, such as geographical and spatial details, e.g. locations, regions and municipalities. Such information may offer valuable insights into the linguistic landscape, but it may also pose challenges when some aspects remain ambiguous. This paper outlines and discusses various known and unknown uncertainties of spatial aspects contained in a non-standard German language legacy dataset (DBÖ) that has undergone several stages of data conversion since the early nineties. The authors introduce and discuss their taxonomy of uncertainties, exemplified by applying it to the spatial information contained in the DBÖ, the origins of which date back one hundred years. Finally, the authors discuss how the uncertainties found in the dataset affect Digital Humanities practice more widely.

Keywords:

Digital Humanities, spatial uncertainty, taxonomy, historic collections

1 Introduction

Uncertainty is an integral part of everyday life. However, it is only in recent times that it has received heightened attention in academic disciplines and beyond. As Jim Gray (quoted by Hey, Tansley, & Tolle, 2009) put it recently, we have seen a transformation in the whole research cycle, from data capture and data curation to data analysis and data visualization, but the intensive use of analytic frameworks does not necessarily contribute to better research data. Uncertainty, in the light of recent developments in the European policy landscape regarding science, research and innovation, has been taken up in scholarly and scientific discourses. Scientific research and innovation processes are inherently uncertain, the more so as they evolve towards ecosystem networks of actor groups with increased inclusion, collaboration and participation of different stakeholders, and the pressing necessity to meet human needs and face societal challenges. Uncertainty has, however, also been viewed as a chance for new opportunities and progress (see e.g. Nowotny, Scott, & Gibbons, 2013; Nowotny, 2015). Consequently, embracing uncertainty, creating a culture of learning from errors, and allowing

the creation of the conditions required for serendipitous discovery are essential and lie at the centre of the ongoing discussions (which extend well beyond the policy level) around scientific innovation and progress.

Digital humanists have been exhorted to embrace data-driven approaches to doing science, and have been inundated by the sheer amounts of data, from both legacy and modern systems and sources, in which uncertainty is inherent.

Various types of uncertainty have been described in the academic field, typically associated with unknown or lacking information, imprecise or incomplete knowledge, inaccurate measurements, and risk. They have also been addressed by different disciplines, including philosophy (Dow, 2012), psychology (Downey, Hellriegel, & Slocum, 1975), physics (Taylor, 1997), information science (Kuhlthau, 1993), economics (Shackle, 2010), law (Weiss, 2003), and statistics (Stigler, 1986) (see also Bammer & Smithson, 2008). While uncertainties in the natural sciences are mostly related to the limits in the possibilities of making measurements, uncertainties in the Humanities can involve subjective aspects related to perception, ambiguity, vagueness, incompleteness or credibility.

Here, we present a previously developed taxonomy of uncertainties for spatiotemporal and linguistic domains; an overview of the exploreAT! project and its associated data; and specific examples of uncertainty related to the geospatial domain, notably when we deal with data that was collected and transformed over long periods of time.

Across academia, researchers have attempted different ways of classifying uncertainties, resulting in a variety of taxonomies. The New World Encyclopedia (2016) entry on uncertainty presents a general taxonomy; Thomas (2013) introduces a fairly comprehensive one, adapted from Smithson's (1989) taxonomy of ignorance and uncertainty. In Thomas's (2013) taxonomy, uncertainty appears as a specific kind of incompleteness, but not as an error. Specific taxonomies of uncertainty can be found for various areas, including biology (Regan, Colyvan, & Burgman, 2002), health (Fox, 2000), and trading regulations (Hoffmann, Trautmann, & Schneider, 2008). Shattuck, Lewis Miller and Kemmerer (2009), on the other hand, make the distinction between the uncertainty produced by the flow of information and the uncertainty of individuals interpreting any given information. Lovell (1995), in an extended digression on the topic, presents a detailed compilation of uncertainties from many different sources. In this view, uncertainties can originate in the world itself, in the empirical evidence, and in the human subjects who interpret them. Vullings, de Vries and de Borman (2007), based on Fisher, Comber and Wadsworth (2005), devised a fairly complete model for dealing with spatial uncertainties. Temporal uncertainties are often associated with spatial data, as pointed out by Cressie and Wikle (2015). Aigner, Miksch, Müller, Schumann and Tominski (2007) distinguish time points and time intervals, and also draw attention to the kind of events that are being described when they involve other variables (such as space). Kissling et al. (2018) identify the differing lengths of time series and the precision of time in the collection process as sources of temporal uncertainty. Uncertainty in data pertaining to Geographic Information Systems (GIS) and spatial information in general is a frequently explored topic (see e.g. Couclelis, 2003; Fisher, 1999; Fusco et al., 2017; Züfle et al., 2017) and finds its own entry in

the GIS dictionary[1]. We aim to illustrate how these uncertainties can arise and affect a legacy language collection that contains other aspects of information, such as geographical and spatial details.

2 Taxonomies of uncertainty

In the scope of our research, we explore uncertainty in the Humanities, in particular within Digital Humanities (DH), where uncertainty has in recent years been under the spotlight (see Rocha Souza, Dorn, Piringer; Wandl-Vogt, 2019) and generating increased interest, particularly in relation to data and data treatment. Data includes imprecise or erroneous information and knowledge, incomplete information, spelling variations, abbreviations, ambiguous information, missing information, or uncertainties introduced by tools or human beings in the process of digital data transformation and standardization. In combination with such language phenomena, and linguistic changes, such as shifts in language borders/boundaries, uncertainties in the spatio-temporal aspects play an important role and also give insights into the history and workflow of data collections. In order to facilitate such insights, we based our analysis of uncertainties on existing categories of uncertainty, which we eventually modified to include novel aspects found in our data, developing our own taxonomy of uncertainties (Rocha Souza, Dorn, Piringer, & Wandl-Vogt, 2019) (see Figure 1).

Common to long data transformation and conversion processes, uncertainties have been both remedied and reintroduced over time – for example differences in database schemas due to assignment of fields without proper semantics during DB conversion; imperfect matches between the original terms/lexical concepts and DBpedia concepts in the enrichment process. While most of these uncertainties are common to a plethora of long-term, data-intensive projects, some are particular to this collection.

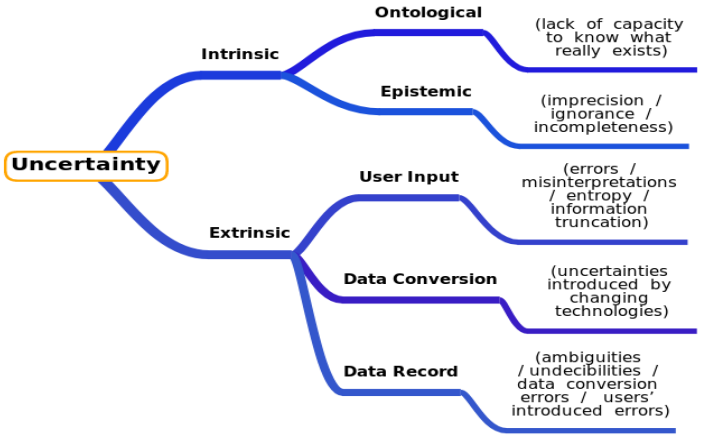

Figure 1: Uncertainty dimensions

[1] https://support.esri.com/en/other-resources/gis-dictionary/term/9ac5d78f-2a00-4c24-81ba-346ad51bf302

3 The exploreAT! project, the DBÖ collection and the PROVIDEDH project

This study was carried out in the context of the Digital Humanities project *exploreAT! – exploring Austria's culture through the language glass* (see Wandl-Vogt, Kieslinger, O'Connor, & Therón, 2015). exploreAT! was implemented in 2015 as a cross-disciplinary project at the Austrian Centre for Digital Humanities (ACDH-OeAW), the Austrian Academy of Sciences. It brings together expertise from different disciplines and partners in the fields of cultural lexicography and Open Innovation (OI) (ACDH-OeAW, Austria), semantic technologies (ADAPT Centre, DCU, Ireland), and human–machine interaction via visualization (VisUSAL, Universidad de Salamanca, Spain) (see Abgaz, Dorn, Piringer, Wandl-Vogt, & Way, 2018a, 2018b; Benito et al., 2016; Benito, Losada, Therón, Dorn, & Wandl-Vogt, 2018; Dorn, Wandl-Vogt, Abgaz, Benito Santos, & Therón, 2018).

The exploreAT! project has at its core a digitized non-standard language resource of the Bavarian Dialects in Austria (*Datenbank der bairischen Mundarten in Österreich [DBÖ])* and the related dbo@ema (*database of Bavarian dialects @ electronically mapped*) (Wandl-Vogt, 2008). Initially conceived as a dictionary project (*Wörterbuch der bairischen Mundarten in Österreich* [WBÖ, 1970–]; see *Arbeitsplan*, 1912), this heterogeneous collection not only captures the historical language in an area of the former Austro-Hungarian Empire, but also contains detailed cultural information of the former day-to-day life of the rural population, including their professions, customs, religious festivities, folk medicine, etc. In addition, the DBÖ collection contains digitized information extracted from excerpts of folk literature, vernacular dictionaries and historical documents. The data follows a lexicographical structure consisting of lemmas, definitions, sources and a variety of other fields. As well as this richly textured linguistic and societal content, the collection also makes available information on people (authors, collectors, editors) (Piringer, Wandl-Vogt, Abgaz, & Lejtovicz, 2017), and spatio-temporal information (places, regions, GIS locations, etc.) (Scholz, Hrastnig, & Wandl-Vogt, 2018).

The DBÖ collection has undergone various transformation processes since its beginning in 1911. The collection started by means of questionnaires, covering around 100 different topics pertaining to everyday life, which were distributed across the population. Together, the questionnaires totalled approximately 17,000 questions. Answers to these questions were first noted on individual paper slips, then the data passed through several stages of digitization and digital data conversion (Figure 2), until the collection reached its current state.

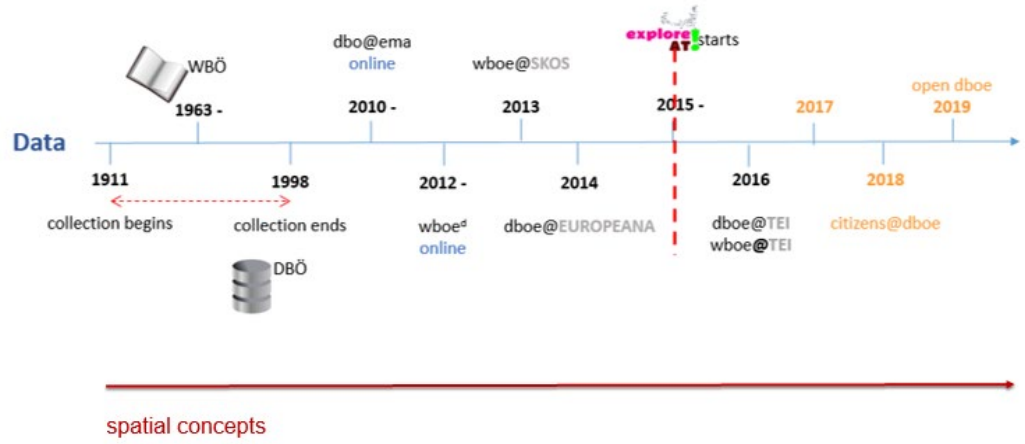

Figure 2: Timeline of the data-transformation process in relation to the beginning of the exploreAT! project. Image © Amelie Dorn, Eveline Wandl-Vogt 2018

In the first stage of digitization (1993–2011), all available information noted on the paper slips (including headword, meaning, pronunciation, location, date, collector's name) was manually entered into TUSTEP (*TÜbinger System von TExtverarbeitungs-Programmen / Tuebingen System of Text Processing tools)*[2], resulting in ~2.43 million entries (Bergmann, Glauninger, Wandl-Vogt & Winterstein, 2010). Towards the end of this first digitization process, parts of the TUSTEP data (auxiliary databases for biographies, bibliographies, plant names, locations) and the institute's library database (MS-Access) were transferred to a relational database (MySQL and PostgreSQL) cluster as part of the dbo@ema project (*Datenbank der Bairischen Mundarten in Österreich electronically mapped*) (Wandl-Vogt, 2012). For the first time, separate datasets were joined, and a geographic visualization interface (maps) and georeferencing of data (coordinates: latitude/longitude and altitude) were added, creating a real-world relationship. Further, visualization and analysis of the data via interactive web-based maps were enabled, re-using a system that was already in place for another dataset; data were made publicly accessible and visible on the internet via an interactive project website.[3] dbo@ema was in use for editing purposes by more than 20 people during 2010–2012, and for geo-spatial hierarchization.

From this point, the heterogeneity of the data increased again, with parts of the data being converted to an Entity-Relationship model in the MySQL database (Wandl-Vogt, 2010, 2012). In 2015, with the start of the exploreAT! project, data conversion into two formats evolved: 1) TEI/XML format (Schopper, Bowers & Wandl-Vogt, 2015), based on information from both the TUSTEP files and dbo@ema; 2) RDF (Resource Description Framework), linked to the LOD Cloud[4] (2017–) (Abgaz et al., 2018a, 2018b).

[2] https://www.tustep.uni-tuebingen.de/tustep_eng.html

[3] https://dboema.acdh.oeaw.ac.at/projekt/beschreibung/

[4] https://lod-cloud.net/

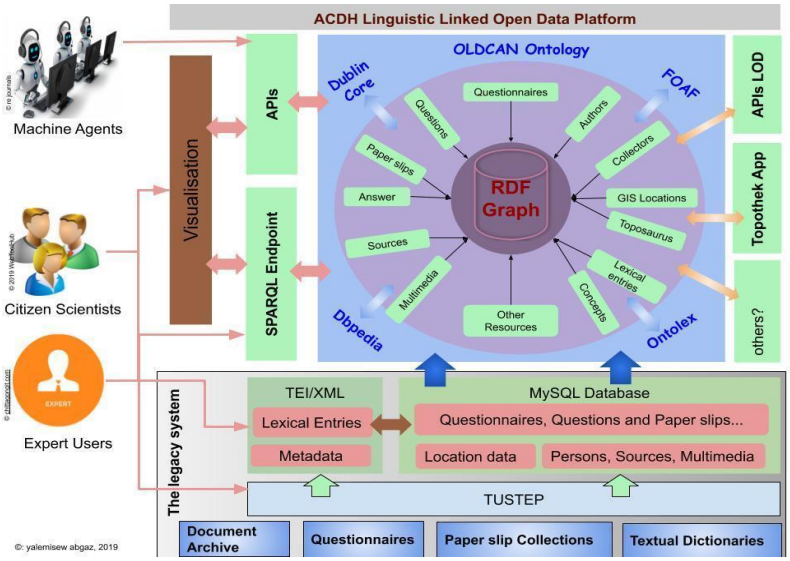

Figure 3: Overview of the data transformation process. Source: Yalemisew Abgaz

To give a concrete example, Figure 4 presents two stages in the conversion process, from a paper slip (a), to a TUSTEP entry (b), and an XML/TEI file excerpt (c).

Figure 4: Example of the data conversion process for the word 'Strützel'. (a): original TUSTEP entry; (b): screenshot of a TUSTEP entry; (c): XML data entry.

Tables 1, 2 and 3 present temporal and spatial information relating to the collection, in two of the main current digital sources (XML/TEI files and MySQL database).

Table 1 shows the time span for entries in each of the main sources.

Table 1: Numerical overview of temporal information for the entries. Source: the authors

	XML/TEI files		MySQL DB	
time span for entries	oldest	newest	oldest	newest
year	1010	2008	1196	2012

Table 2 presents the numbers of entries with and without spatial information.

Table 2: Numerical overview of entries with and without spatial information. Source: the authors

	XML/TEI files		MySQL DB	
number of entries	2,416,499		65,839	
	with location	without location	with location	without location
	1,712,705 (71%)	703,794 (29%)	7,333 (11%)	58,506 (89%)

For each of the main databases, Table 3 shows the number of entries with spatial information, with a breakdown by level of location.

Table 3: Numerical overview for spatial information per hierarchical, partly administrative spatial level. Source: the authors

	XML/TEI files		MySQL DB
Location level	number of distinct locations per level	number of entries with locations	number of entries with locations
• Bundesland	9	1,316,889 (55%)	-
• Großregion	32	1,296,722 (54%)	-
• Kleinregion	323	1,286,463 (53%)	415 (0,6%)
• Gemeinde	1,146	1,198,447 (50%)	3,058 (4,6%)
• Ort	1,145	1,198,447 (50%)	19,946 (30%)
• Ort (without associated Gemeinde)	24,788	395,186 (16%)	-

The specific spatial parameters are: Bundesland (county; e.g. Steiermark/St.), Großregion (big region; e.g. mittelbairische Obersteiermark/mbair.Obst.), Kleinregion (small region; e.g. Erzberger Gegend/Erzbg.Geg.), Gemeinde (municipality; e.g. Radmer), Ort (location; e.g. Radmer), and entries without a given location. The distinctions between the different types/sizes of regions were made according to the so-called 'Sigles' (a system of identifiers for regions), which consists of a combination of numbers and letters denoting a hierarchical structure, as we can see in Figure 5.

```
<listPlace xml:id="sigle:3.2q02">
  <place type="Bundesland">
    <placeName>St.</placeName>
    <idno>3</idno>
    <listPlace>
      <place type="Großregion">
        <placeName>mbair.ObSt.</placeName>
        <idno>3.2</idno>
        <listPlace>
          <place type="Kleinregion">
            <placeName>Erzbg.Geg.</placeName>
            <idno>3.2q</idno>
            <listPlace>
              <place type="Gemeinde">
                <placeName>Radmer</placeName>
                <idno/>
                <listPlace>
                  <place type="Ort">
                    <placeName>Radmer</placeName>
                    <idno>3.2q02</idno>
                  </place>
                </listPlace>
```

Figure 5: Example of the nested location codes in an entry from the XML files. Source: the authors

If we compare the total numbers of unique locations, we note considerably more entries in the XML dataset than in MySQL, but also striking structural differences between the two datasets. Whereas the majority of XML entries contain a hierarchical structure of location information (Bundesland > Großregion > Kleinregion > Gemeinde > Ort), some parameters (Bundesland, Großregion, Kleinregion) are not accessible in a structured way, but have been merged in a single column. A noticeable difference between the datasets emerges: the MySQL dataset contains a higher percentage of unique location entries. However, this can be explained by the huge difference in the number of records - the MySQL data is 2,7% of the size of the TEI-XML data.

Looking finally at entries that are, or are not, linked to location parameters, again an overall higher number can be observed for the XML dataset. In this dataset, compared to the MySQL dataset, a higher number of entries are linked to location information.

This numerical overview can only offer an impression of the type and quantity of data contained in the dataset; it does not cover the various levels at which uncertainties in this particular dataset can arise or the extent of heterogeneity. The records are not homogeneous, given differences in the details from the myriad of sources, and also because of differences in the transformation and conversion processes from the legacy sources to the current records.

4 Geospatial uncertainties in the DBÖ collection

Geospatial aspects and properties pertaining to the DBÖ collection and dbo@ema database have been dealt with in various ways over recent years (Wandl-Vogt et al., 2008; Scholz et al., 2008; Bartelme & Scholz, 2010; Benito et al., 2018; Scholz et al., 2018; Hrastnig, 2018).

As commonly occurs in long data transformation and conversion processes, uncertainties have been both remedied and introduced over time. It is also important to note that the administrative hierarchy may change over time: for example, an 'Ort' may now be in a different region from the one it was in at the time the record was created. Most of these uncertainties are common to a plethora of long-term, data-intensive projects. Table 5 presents the classes and sources of uncertainties regarding spatial dimensions in our collection.

Visualization and GI techniques were employed to mitigate these problems, as can be seen in earlier related work (Wandl-Vogt et al., 2008; Wandl-Vogt, 2010; Wandl-Vogt et al., 2015; Scholz, Lampoltshammer, Bartelme, & Wandl-Vogt, 2016; Benito et al., 2018; Scholz et al., 2018).

Table 4: Classes and sources of spatial uncertainties. Source: the authors

	Uncertainties				
	Intrinsic		**Extrinsic**		
	Ontological (lack of capacity to know what really exists)	**Epistemic** (imprecision / ignorance / incompleteness)	**User input** (errors / misinterpretations / entropy / information truncation)	**Data conversion** (uncertainties introduced by changing technologies)	**Data record** (ambiguities / Undecidable elements / data conversion errors / users' introduced errors)
Spatial uncertainties	- Places that ceased to exist	- Unknown places - Exact place vs. approximate/region	- Typos - Abbreviations - Changing transcription guidelines - Assumptions about certain spelling variations - Lack of precision in creating data records - Guessing - Prejudice and biases	- Language codification errors - Errors in the conversion of formats and databases - Heterogeneity of data sources	- Identical toponyms - Difference in details among records

5 Discussion

We have presented some of the aspects of uncertainty in the DBÖ collection as regards the spatial domain. Our research has offered insights into contributing factors, including the multiple sources, highlighting also the sheer extent of heterogeneity in this legacy dataset. To cope with the specificities of the collections, a handful of established taxonomies for classifying uncertainties were consulted, which led us to devise a specific one, suitable for our data. What has become apparent is that the continuous process of data transformation, aimed at promoting accessibility and enriching the collection informationally, also introduced new types of uncertainties, despite the availability and use of guidelines, standards and manual corrections. Where the spatial dimensions in particular are concerned, the constantly evolving nature of geopolitical entities in the real world (changes in borders, names of places, regions, territories and so on) have affected not only the historical but also the current datasets.

Nevertheless, many of the uncertainties have also been partially resolved in the course of data transformation processes, and new opportunities for exploration have been created. In this context, the dbo@ema project (Wandl-Vogt et al., 2008), for the first time, enabled the geo-referencing of all data and its immediate publication in a map, making available interlinked publications, and the interactive navigation and analysis of data in connection to a map. Thanks to the collaboration between teams from different disciplines, diverse views on the data and information were enabled, such as a the distribution of homonymous toponyms, mapping of places with collections on Google maps, or a web-browser-based query and headword presentation (Wandl-Vogt, 2010). In the context of the exploreAT! project, data beyond the map was explored further (Theron & Wandl-Vogt, 2014). Subsequently, a web-browser-based visual analysis of the TEI-encoded data, drawing on network visualizations of data chunks, was also enabled, in a prototype, for data with and without precise temporal or spatial information (Benito et al., 2016). In addition, an interactive web-based exploration of the DBÖ content was developed by Benito et al. (2018) by revisiting and building on previous work. In spite of the efforts to deal with these uncertainties, these uncertainties cannot be fixed or solved retroactively. This impossibility demands a pragmatic / probabilistic approach when dealing with the linguistic information in the DBÖ resource.

We understand that much of what we have illustrated in this paper regarding spatial uncertainties is common to many corpora formed through time, such as collections of heritage and historical documents. Although many processes of data gathering, input and conversion are inherently *ad hoc*, the possible extrapolations and generalizations may serve as a warning for the difficulties of maintaining huge textual, imagetic and multimedia collections which are so common nowadays. The majority of computer database collections were compiled in the last three decades, and collections formed over long periods (in this case, a whole century) are key to understanding the long-term consequences of each and every decision regarding data maintenance. Although uncertainty is impossible to avoid, keeping it at its lowest acceptable level is an essential goal of data humanists. At the same time, uncertainties may open up new possibilities for collaboration across disciplines, and potential for creating and exploring new insights – something which is particularly suited to the Digital Humanities field.

Acknowledgements

This research was partially supported by the Nationalstiftung of the Austrian Academy of Sciences Sciences under the funding scheme: Digitales kulturelles Erbe, grant number DH2014/22, as part of the exploreAT! project, carried out in collaboration with the VisUSAL Group, Universidad de Salamanca, Spain and the ADAPT Centre for Digital Content Technology at Dublin City University, Ireland, which is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. This research was also partially supported by the PROVIDEDH project, funded within the CHIST-ERA programme under the national grant agreement PCIN-2017-064 (MINECO, Spain), in the context of which the Austrian Centre for Digital Humanities as a project partner receives funding under the national grant agreement FWF (Project number I 3441-N33).

References

Abgaz, Y., Dorn, A., Piringer, B., Wandl-Vogt, E., & Way, A. (2018a). A Semantic Model for Traditional Data Collection Questionnaires Enabling Cultural Analysis. In McCrae, J.P., Chiarcos, C., Declerck, T., Gracia, J., & Klimek, B. (Eds.), Proceedings of the LREC 2018 Workshop '6th Workshop on Linked Data in Linguistics (LDL-2018)' (pp. 21–29). Miyazaki, Japan.

Abgaz, Y., Dorn, A., Piringer, B., Wandl-Vogt, E., & Way, A. (2018b). Semantic Modelling and Publishing of Traditional Data Collection Questionnaires and Answers. Information, 9(12), 297:1–297:24. https://doi.org/10.3390/info9120297

Aigner, W., Miksch, S., Müller, W., Schumann, H., & Tominski, C. (2007). Visualizing time-oriented data—A systematic view. Computers & Graphics, 31(3), 401–409. https://doi.org/10.1016/j.cag.2007.01.030

Arbeitsplan und Geschäftsordnung für das bayerisch-österreichische Wörterbuch. 16. Juli 1912. Karton 1. Arbeitsplan-a-h Bayerisch-Österreichisches Wörterbuch. Wien: Archive of the Austrian Academy of Sciences.

Bammer, G., & Smithson, M. (Eds.) (2008). Uncertainty and Risk. Multidisciplinary Perspectives. London, UK: Earthscan.

Bartelme, N., & Scholz, J. (2010). Geoinformationstechnologien zur Analyse des Raum- und Zeitbezugs bei Dialektwörtern. In Bergmann, H., Glauninger, M.M., Wandl-Vogt, E., Winterstein, S. (Eds.), Fokus Dialekt. Analysieren – Dokumentieren – Kommunizieren. Festschrift für Ingeborg Geyer zum 60. Geburtstag (= Germanistische Linguistik 199–201) (pp. 65-78). Hildesheim, Germany: Georg Olms Verlag.

Benito, A., Losada, A.G., Therón, R., Dorn, A., Seltmann, M., & Wandl-Vogt, E. (2016). A Spatio-temporal Visual Analysis Tool for Historical Dictionaries. In García-Peñalvo, F.J. (Ed.), Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality (pp. 985-990). New York, NY: ACM. https://doi.org./10.1145/3012430.3012636

Benito, A., Losada, A.G., Therón, R., Dorn, A., & Wandl-Vogt, E. (2018). Creating Meaningful Narratives in Collections of Historical Lexical Data. GI_Forum, 6(2), 50–57. https://doi.org/10.1553/giscience2018_02_s50

Bergmann, H., Glauninger, M., Wand-Vogt, E., & Winterstein, S. (Eds.) (2010). Fokus Dialekt. Analysieren – Dokumentieren – Kommunizieren. Festschrift für Ingeborg Geyer zum 60. Geburtstag (= Germanistische Linguistik 199-201). Hildesheim, Germany: Georg Olms Verlag.

Couclelis, H. (2003). The certainty of uncertainty: GIS and the limits of geographic knowledge. Transactions in GIS, 7(2), 165-175. https://doi.org/10.1111/1467-9671.00138

Cressie, N. & Wikle, C.K. (2015). Statistics for spatio-temporal data. Hoboken, NJ. Wiley & Sons.

Dorn, A., Wandl-Vogt, E., Abgaz, Y., Benito Santos, A., & Therón, R. (2018). Unlocking Cultural Conceptualisation in Indigenous Language Resources: Collaborative Computing Methodologies. In Soria, L., Besacier, L., & Pretorius, L. (Eds.), Proceedings of the LREC 2018 Workshop 'CCURL2018 – Sustainable Knowledge Diversity in the Digital Age' (pp. 19–22).

Dow, S.C. (2012). Uncertainty about uncertainty. In Dow, S.C., Foundations for New Economic Thinking (pp. 72–82). London, UK: Palgrave Macmillan.

Downey, H.K., Hellriegel, D., & Slocum, J.W. (1975). Environmental Uncertainty: The Construct and Its Application. Administrative Science Quarterly, 20(4), 613–629

Fisher, P. F. (1999). Models of uncertainty in spatial data. Geographical information systems, 1, 191–205.

Fisher, P., Comber, A., & Wadsworth, R. (2005). Approaches to Uncertainty in Spatial Data. In Devillers, R., & Jeansoulin, R. (Eds.), Qualité de l'information géographique (Traité IGAT), (pp. 9–64). Paris, France: Hermes/Lavoisier.

Fox, R. C. (2000). Medical uncertainty revisited. In Albrecht, G.L., Fitzpatrick, R., & Scrimshaw, S.C. (Eds.), Handbook of social studies in health and medicine (pp. 409-425). Thousand Oaks, CA: SAGE Publishing.

Fusco, G., Caglioni, M., Emsellem, K., Merad, M., Moreno, D., & Voiron-Canicio, C. (2017). Questions of uncertainty in geography. Environment and Planning A: Economy and Space, 49(10), 2261–2280. https://doi.org/10.1177/0308518X17718838

Hey, T., Tansley, S., & Tolle, K. (2009). Jim Gray on eScience: A Transformed Scientific Method. In Hey, T., Tansley, S., & Tolle, K. (Eds.), The Fourth Paradigm. Data-Intensive Scientific Discovery (pp. xvii-xxxi). Redmond, WA: Microsoft Research. Retrieved from https://www.microsoft.com/en-us/research/wp-ontent/uploads/2009/10/Fourth_Paradigm.pdf

Hoffmann, V.H., Trautmann, T., & Schneider, M. (2008). A taxonomy for regulatory uncertainty—application to the European Emission Trading Scheme. Environmental Science & Policy, 11(8), 712-722. https://doi.org/10.1016/j.envsci.2008.07.001

Hrastnig, E. (2018). A Linked Data approach for Digital Humanities (Master's Thesis). Technische Universität Graz, Graz, Austria. January 2018. Retrieved from https://diglib.tugraz.at/download.php?id=5b073aaa542f6&location=browse

Kissling, W.D., Ahumada, J.A., Bowser, A., Fernandez, M., Fernández, N., Alonso Garcia, E., … Hardisty, A.R. (2018). Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. Biological reviews, 93(1), 600-625. https://doi.org/10.1111/brv.12359

Kuhlthau, C.C. (1993). A principle of uncertainty for information seeking. Journal of Documentation, 49(4), 339-355. https://doi.org/10.1108/eb026918

Lovell, B.E. (1995). A Taxonomy of Types of Uncertainty (Doctoral dissertation). Portland State University, Portland, OR, USA. Dissertations and Theses. Paper 1396. https://doi.org/10.15760/etd.1395

Nowotny, H. (2015). The radical openness of science and innovation. Why uncertainty is inherent in the openness towards the future. EMBO Reports, 16(12), 1601-1604. https://doi.org/10.15252/embr.201541546

Nowotny, H., Scott, P.B., Gibbons, M.T. (2013). Re-thinking science: Knowledge and the public in an age of uncertainty. New York, NY: Wiley & Sons.

Österreichische Akademie der Wissenschaften (2018, January 15). Datenbank der bairischen Mundarten in Österreich [Database of the Bavarian Dialects in Austria] (DBÖ) [Data file].

Piringer, B., Wandl-Vogt, E., Abgaz, Y., & Lejtovicz, K. (2017). Exploring and exploiting biographical and prosopographical information as common access layer for heterogeneous data facilitating inclusive, gender- symmetric research. In Wandl-Vogt, E., & Lejtovicz, K. (Eds.), Biographical Data

in a Digital World 2017. A conference in the framework of the project APIS, 6–7 November 2017. Abstracts. https://doi.org/10.5281/zenodo.1041978

Regan, H.M., Colyvan M., & Burgman, M.A. (2002). A taxonomy and treatment of uncertainty for ecology and conservation biology. Ecological Applications, 12(2), 618-628. https://doi.org/10.1890/1051-0761(2002)012[0618:ATATOU]2.0.CO;2

Rocha Souza, R., Dorn, A., Piringer, B., & Wandl-Vogt, E. (2019, September). Towards a taxonomy of uncertainties: Analysing sources of spatio-temporal uncertainty on the example of non-standard German corpora. In Informatics (Vol. 6, No. 3, p. 34). Multidisciplinary Digital Publishing Institute. DOI:10.3390/informatics6030034.

Scholz, J., Bartelme N., Fliedl G., Hassler M., Mayr H.C, Nickel J., … Wandl-Vogt, E. (2008). Mapping Languages – Erfahrungen aus dem Projekt dbo@ema. In Angewandte Geoinformatik 2008 - Beiträge zum 20. AGIT-Symposium (pp. 822–827). Heidelberg, Germany: Wichmann.

Scholz, J., Hrastnig, E., & Wandl-Vogt, E. (2018). A Spatio-Temporal Linked Data Representation for Modeling Spatio-Temporal Dialect Data. In Fogliaroni, P., Ballatore, A., Clementini, E. (Eds.), Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017) (pp. 275–282). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-63946-8_44

Scholz, J., Lampoltshammer, T.J., Bartelme, N., & Wandl-Vogt, E. (2016). Spatial-temporal Modeling of Linguistic Regions and Processes with Combined Indeterminate and Crisp Boundaries. In Gartner, G., Jobst, M., & Huang, H. (Eds.), Progress in Cartography. Lecture Notes in Geoinformation and Cartography. Cham, Switzerland: Springer. pp. 133–151.

Schopper, D., Bowers, J., & Wandl-Vogt, E. (2015). dboe@TEI: remodelling a database of dialects into a rich LOD resource. In Text Encoding Initiative. Conference and members' meeting 2015. October 28-31, Lyon, France. Papers. Retrieved from http://tei2015.huma-num.fr/en/papers/#146

Shackle, G.L.S. (2010). Uncertainty in economics and other reflections. Cambridge, UK: Cambridge University Press.

Shattuck, L.G., Lewis Miller, N., & Kemmerer, K.E. (2009). Tactical Decision Making Under Conditions of Uncertainty: An Empirical Study. Proceedings of the Human Factors and Ergonomics Society. Annual Meeting, 53(4), 242–246. https://doi.org/10.1177%2F154193120905300417

Smithson, M. (1989). Ignorance and uncertainty: emerging paradigms. Berlin, Germany: Springer Science & Business Media.

Stigler, S. M. (1986). The history of statistics: The measurement of uncertainty before 1900. Cambridge, MA: Harvard University Press.

Taylor, J. (1997). Introduction to Error Analysis, the Study of Uncertainties in Physical Measurements. (2nd ed.). New York, NY: University Science Books.

Therón, R., Losada, A.G., Benito, A., & Santamaría, R. (2018). Toward supporting decision-making under uncertainty in digital humanities with progressive visualization. In García-Peñalvo, F.J. (Ed.), Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality (pp. 826–832). New York, NY: ACM. https://doi.org/10.1145/3284179.3284323

Therón, R. & Wandl-Vogt, E. (2014). The Fun of Exploration: How to Access a Non-Standard Language Corpus Visually. In Hautli-Janisz, A., Lyding, V., & Rohrdantz, C. (Eds.), Proceedings of the LREC 2014 Workshop 'VisLR – Visualization as added value in the development, use and evaluation of LR's' (pp. 9–12)

Thomas, R.C. (2013). The Rainforest of Ignorance and Uncertainty [Blog post]. Retrieved from https://exploringpossibilityspace.blogspot.com/2013/07/the-rainforest-of-ignorance-and.html

Uncertainty. (2016). In New World Encyclopedia. Retrieved from http://www.newworldencyclopedia.org/p/index.php?title=Uncertainty&oldid=993112

Vullings, W., de Vries, M., & de Borman, L. (2007). Dealing with uncertainty in spatial planning. In Wachowicz, M., & Bodum, L. (Eds.), Proceedings 2007. The 10th AGILE International Conference

on Geographic Information Science. Retrieved from https://agile-online.org/conference_paper/cds/agile_2007/proc/pdf/164_pdf.pdf

Wandl-Vogt, E. (2010). Multiple access routes. The dictionary of Bavarian dialects in Austria / Wörterbuch der bairischen Mundarten in Österreich (WBÖ). In Granger, S., Paquot, M. (Eds.), eLexicography in the 21st Century: New Challenges, New Applications. Proceedings of eLex 2009, Louvain-la-Neuve, 22–24 October 2009 (= Cahiers du Cental 7) (pp. 451–455). Louvain-la-Neuve, France: Presses Univ. de Louvain.

Wandl-Vogt, E. (2012). Datenbank der bairischen Mundarten in Österreich @ electronically mapped. Projektbeschreibung. Retrieved from https://dboema.acdh.oeaw.ac.at/projekt/beschreibung/

Wandl-Vogt, E. (2018, January 15). Datenbank der bairischen Mundarten in Österreich electronically mapped [Database of the Bavarian Dialects in Austria electronically mapped] (dbo@ema) [Data file].

[Wandl-Vogt, E., Kieslinger, B., O´Connor, A., & Theron, R.] (2015). exploreAT! Perspektiven einer Transformation am Beispiel eines lexikographischen Jahrhundertprojekts. In DHd2015. Von Daten zu Erkenntnissen. 23. Bis 27. Februar 2015, Graz. Book of Abstracts. Retrieved from http://gams.uni-graz.at/o:dhd2015.abstracts-gesamt

Weiss, C. (2003). Expressing scientific uncertainty. Law, Probability and Risk, 2(1), 25-46. https://doi.org/10.1093/lpr/2.1.25

Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Bayerisches Wörterbuch: I. Österreich (1970–). Ed. by Österreichische Akademie der Wissenschaften. Wien, Austria: Verlag der Österreichischen Akademie der Wissenschaften.

Züfle, A., Trajcevski, G., Pfoser, D., Renz, M., Rice, M.T., Leslie,T., Delamater, P., & Emrich, T. (2017). Handling Uncertainty in Geo-Spatial Data. In Proceedings. 2017 IEEE 33rd International Conference on Data Engineering – ICDE – 19–22 April 2017, San Diego, California, USA (pp. 1467–1470). Piscataway, NJ: IEEE. https://doi.org/10.1109/ICDE.2017.212

GI_Forum 2020, Issue 1
Page: 47 - 62
Full Paper
Corresponding Author:
pointner@bgeos.com
DOI: 10.1553/giscience2020_01_s47

Interannual Variability of Lake Ice Backscatter Anomalies on Lake Neyto, Yamal, Russia

Georg Pointner[1,2] and Annett Bartsch[1,2]

[1]b.geos, Korneuburg, Austria
[2]Austrian Polar Research Institute, Vienna, Austria

Abstract

Anomalous areas of varying shape and location characterized by low backscatter in Synthetic Aperture Radar (SAR) imagery of lake ice on lake Neyto on the Yamal Peninsula in Russia have been described qualitatively in the literature for many years. Possible suggested causes are the formation of eddies or the release of gas through the lake sediments, which could both lead to local thinning of the ice layer and alter radar backscatter. To date, the phenomenon, its cause, and its spatial and temporal dynamics are poorly understood, and studies from other geographic regions are completely absent. In order to perform first steps towards a better understanding of the phenomenon, we developed a workflow to quantitatively assess the spatial variability of the anomalies in the years 2015 to 2019 for lake Neyto. We introduce a binary image classification algorithm developed with state-of-the-art open-source image processing tools and employ metrics commonly used for describing spatial relationships of vector and raster data. This includes polygon distances, polygon intersections and cumulative pixel counts deduced from the classification results in order to quantify, for the very first time, the dynamics over a number of years. The geospatial analysis reveals large spatial variations, but also some overlap between different years. Locations of anomalies do not seem more similar between consecutive years than when they are compared over the longer period. Some of the spatial properties of the clusters of low backscatter may support the explanation of gas release as the primary cause of the observed patterns.

Keywords:

remote sensing, Arctic, lakes, image processing, synthetic aperture radar

1 Introduction

Today's pupils will live in cities that are organized in a fundamentally different way compared to present urban spaces. They will live in smart cities. Three developments have facilitated the spread of the smart city: more efficiently. Ideally, citizens have more influence through e-participation in governmental decisions (Mandl & Zimmermann-Janschitz, 2014, p. 616).

Arctic lakes are important features of the hydrosphere and the cryosphere. They occupy significant parts of the Arctic tundra and play an important role in the carbon cycle (e.g. Walter Anthony et al., 2012; Wik et al., 2016).

In winter, space-borne C-band Synthetic Aperture Radar (SAR) data can be useful for monitoring lake ice phenology (e.g. Surdu et al., 2015; Duguay & Pietroniro, 2005), and especially the grounding state of lake ice (e.g. Duguay & Lafleur, 2003; Surdu et al., 2014; Grunblatt & Atwood, 2014). Regions of floating lake ice appear bright (high backscatter) in SAR images due to the high reflection of the radar signal, which is caused by high dielectric contrast at the ice–water boundary (Duguay et al., 2002).

However, for a range of lakes on the Yamal Peninsula in northwestern Siberia, patterns of low backscatter in central parts of lakes where floating lake ice is assumed were identified by Bogoyavlensky et al. (2018). Low backscatter is usually observed from shallow shelves of lakes, where the ice is grounded (Duguay et al., 2002). The extent of these areas remains fairly constant throughout the winter, but the extent of the area with low backscatter patterns changes significantly throughout the winter. Most of the zones of low backscatter outside the shelves can be identified first in the SAR imagery in mid to late winter (usually March or April), when they start to appear mostly as circular or elongated objects. These subsequently widen until the onset of snowmelt. Possible explanations for these backscatter anomalies include the formation of eddies and the accumulation of methane released through pockmarks in the lake sediments migrating upwards in the water column under the ice layer. Both would lead to a local thinning of the ice layer and thus to lower backscatter due to increased specular reflection from the water surface.

Understanding the origin and dynamics of these patterns may be important for climate research and for understanding sub-lake permafrost dynamics in the case of methane emissions, or for hydrological research in the case of eddies.

To date, the literature has included only visual descriptions of anomalies, in only a few SAR images of lake Neyto and lakes in its vicinity. To our knowledge, there are no descriptions of similar backscatter anomalies for lakes in other geographic regions. Engram et al. (2013) demonstrated a positive statistical relationship between L-band backscatter and bubbles of methane trapped in lake ice for a range of lakes in Alaska, but they did not show consistent areas of anomalous backscatter and noted that such a relationship could not be deduced for C-band data.

In this study, we perform the first-ever quantitative analysis on these objects of varying location and shape on lake Neyto, which will contribute to understanding the phenomenon. Understanding the nature of the phenomenon may lead to new applications of Sentinel-1 data for the monitoring of gas emissions or eddies in remote Arctic locations.

An important part of the research to understand the phenomenon is the analysis of changes in the locations of anomalies in different years. This study aims to describe the variability of patterns from 2015 to 2019 based on Sentinel-1 Extra-Wide Swath (EW) data. First, a method needed to be identified which would allow the retrieval of the anomalies. In the second step, temporal patterns of object metrics were analysed. Objects of low backscatter intensity were

extracted, and distance and intersection metrics were used to describe the spatial variability between the years.

2 Data and Methods

2.1 Study area selection

The study site is one of the largest lakes on the Yamal Peninsula, lake Neyto, which is among the lakes with the largest clusters of pixels of low backscatter in central parts of the lake in late winter. The visual appearance of patterns in different years have already been described by Bogoyavlensky et al. (2018), but no quantitative analyses to characterize spatial and temporal properties had so far been carried out. Similarities to optical data were also identified (an example is shown in Figure 1). Due to these characteristics, we chose this lake as our primary study area. Because of its size and large clusters of low backscatter outside the shelf, it can be studied using relatively coarse Sentinel-1 Extra-Wide Swath (EW) data at 40-metre pixel spacing. This is crucial as data over central Yamal are mostly acquired in this mode by Sentinel-1.

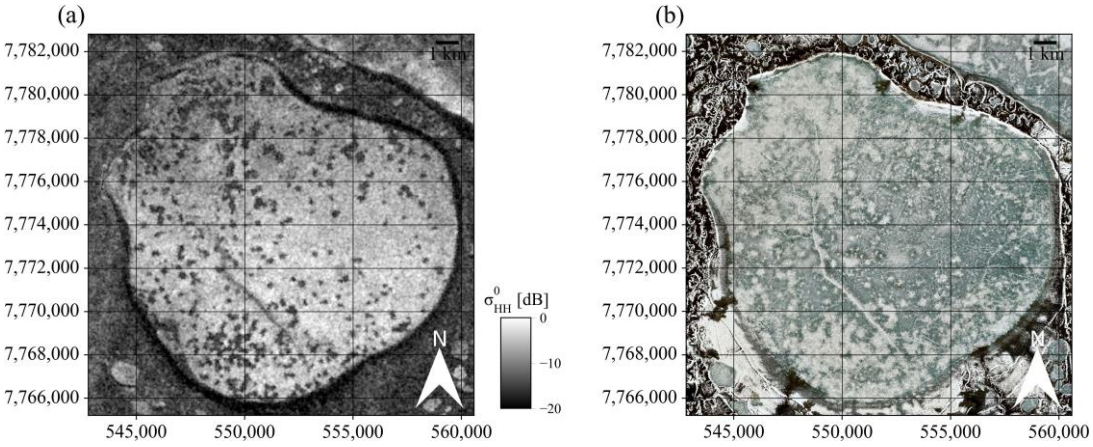

Figure 1: Example for clusters of low backscatter in SAR imagery and similarities to optical imagery: (a) Sentinel-1 EW HH-polarized acquisition from 27 May 2017; (b) Sentinel-2 true-colour composite from 8 June 2017.

2.2 Data

Sentinel-1 SAR Data

The primary data used in this study come from the two polar-orbiting Sentinel-1 satellites, which are part of the EU's Copernicus programme. Sentinel-1A and Sentinel-1B were launched in April 2014 and April 2016, respectively. The main scientific instrument on the two satellites is an identical SAR instrument, called "C-SAR", which can be operated using different spatial resolutions and swath widths (ESA, 2013). Additionally, data can be acquired using

various polarization modes: single co-polarized acquisitions, or dual co-polarized plus cross-polarized acquisitions are possible.

The default operating mode over land is the Interferometric Wide Swath mode (IW) with vertical-vertical (VV) and vertical-horizontal (VH) dual-polarization. However, data over lake Neyto are much more frequently acquired in Extra Wide Swath mode (EW) with horizontal-horizontal (HH) and horizontal-vertical (HV) dual polarization. The number of EW acquisitions over lake Neyto is more than seven times greater than the number of IW acquisitions, and no IW data are available for 2016. Therefore, we only consider EW data in this study. The main differences between the two modes are the larger swath width and coarser spatial resolution in EW mode compared to IW mode. Common pixel spacing used after pre-processing is 40 metres for EW data and 10 metres for IW data.

Sentinel-1 EW data with both HH and HV polarization channels were used to classify clusters of low backscatter in the central part of lake Neyto. The locations of mapped clusters were compared to each other between the years 2015 to 2019.

In order to assess changes in the locations between years, we used two Sentinel-1 EW acquisitions per year, hence ten images in total. For a time series analysis of the evolution of the area of backscatter anomalies, we used all available later-winter Sentinel-1 EW images, a total of 395 images (60 in 2015, 113 in 2016, 111 in 2017, 62 in 2018, 49 in 2019).

Sentinel-2 Optical Data

The two Sentinel-2 satellites are also part of the EU's Copernicus programme and were launched in June 2015 and March 2017. Sentinel-2A and Sentinel-2B each carry an identical multispectral imager, the 'MultiSpectral Instrument' (MSI), which acquires data in the optical and near-infrared regions of the electromagnetic spectrum in 12 spectral bands (Drusch et al., 2012). The spatial resolution depends on the band and ranges from 10 to 60 metres.

In this study, one Sentinel-2 image was used to visually highlight similarities of patterns between SAR and optical imagery for lake Neyto.

Global Historical Climatology Network (GHCN) – Daily Data

GHCN-Daily (Menne, Durre, Korzeniewski, et al., 2012) is a database that provides daily records of temperature, precipitation and snow over land areas worldwide (Menne, Durre, Vose, et al., 2012). In this study, we used daily air temperature records from the Seyaha station, the station closest to lake Neyto and located on the east coast of the Yamal Peninsula at a distance of approximately 80 kilometres, to assess the influence of weather conditions on ice properties in relation to observed backscatter.

2.3 Pre-processing of Sentinel-1 SAR images

The pre-processing of Sentinel-1 EW data was conducted with the Sentinel Application Platform (SNAP) toolbox provided by the European Space Agency (ESA). The main steps applied were sub-setting, radiometric calibration to backscatter coefficient σ^0, thermal noise removal, terrain correction, conversion to decibels (dB), and incidence angle normalization. All these steps were performed on both polarization channels (HH and HV).

2.4 Binary Classification of Sentinel-1 SAR images

The classification algorithm will be briefly outlined here. We describe the main steps and provide visualizations of classification outcomes in the results section of this paper.

The inputs for the binary classification algorithm are the pre-processed Sentinel-1 images in map geometry. All steps described below were applied in identical fashion to both polarization channels. The main tool used for the classification was the Python module 'scikit-image' (van der Walt et al., 2014).

First, areas outside the lake and its shelf area where ground-fast ice is present needed to be masked from the imagery. We deduced lake masks from late-autumn Sentinel-1 EW imagery and shelf masks from earlier-winter Sentinel-1 EW imagery through binary classification. The images were rescaled to fit the pixel values from -1 to 1 required for the image-processing algorithms. The steps for the image processing included bilateral filtering to remove noise from the images, auto-levelling to balance out the unevenly distributed backscatter level across the lake, and Yen-thresholding (Yen et al., 1995) to automatically classify the images. The outputs of these steps are two classified binary images, one for the HH-channel and one for the HV-channel.

To counter the problem of the lack of in-situ data for calibration, we chose a conservative approach: for the final classification map, we kept only pixels belonging to low backscatter patterns (positive class) in the binary classification outcome of both polarization channels; otherwise the pixels were assigned to the negative class (regular floating lake ice). This corresponds to a logical AND between the classification on the HH-channel and the classification on the HV-channel.

Because Yen-thresholding determines thresholds automatically, it is only applicable if clusters of low backscatter are actually present in the image. We therefore needed to apply a mechanism to detect whether these clusters were present. Our approach tests the similarity between binary classification outcomes of the two polarization channels using Cohen's Kappa score κ (Cohen, 1960). Only if κ was above 0.2, which corresponds to 'fair agreement' following Landis & Koch (1977), was the final classification map produced as defined above. If κ was below 0.2, all pixels in the image were assigned to the negative class.

2.5 Determination of variations of locations between years

Clusters of low backscatter emerge primarily from locations of only a few pixels. Over time, these clusters widen out significantly. New clusters can form later and merge with the widening clusters, but only very rarely do pixels of low backscatter clusters revert to high backscatter within a single year. A comparison of Figures 2 and 3 in the Section 3 (Results) provides some explanation for these observations.

The basis of our analysis are two comparisons of five single binary classification results from the years 2015, 2016, 2017, 2018 and 2019. The images and their respective acquisition dates for the comparisons were chosen according to two criteria based on the observations of pattern development over time as described above. The images selected for the first comparison are the ones where the classified pattern area first exceeded 20km^2. The images

for the second comparison were acquired on the last date the patterns were detectable. After snowmelt sets in, the clusters of low backscatter can no longer be observed, because very low backscatter is observed from the entire lake surface.

For the two comparisons themselves, we calculated the mean minimum distance and percentages of intersecting areas between polygonized classification outcomes of the positive class (backscatter anomalies), pairwise for all years. This is the first time that study of the phenomenon has focused on single objects and the spatial relationships among them. For the most part, the Python packages Shapely (Gillies, 2007) and Fiona (Gillies, 2011), which are essential tools for geospatial programming, were used for the calculations. Since the mean minimum distance calculated from one polygon set A to another polygon set B is not equal to the mean minimum distance of polygon set B to polygon set A, the result of our calculations is a square matrix of shape 5x5 (because of 5 distinct years). Similarly, the percentage of intersecting areas is also asymmetrical, as it is calculated as the area of intersection between images of two years divided by the classified area in one year. Hence, the result is also a 5x5 matrix. Additionally, we calculated cumulative counts of positively classified pixels for the two comparisons, ranging from 0 (no occurrence in any year) to 5 (the pixel was classified positively in all five years). Further, we considered a time series of classified pattern areas for our interpretations.

3 Results

3.1 Classification results

The classification results of the positive class for the first criterion, where the pattern area first exceeded 20 km² in 2015 to 2019, are shown in Figure 2 (a)–(e). Similarly, the classification results of the positive class for the second criterion, the date (2015 to 2019) of the last available useful acquisitions in the years concerned are shown in Figure 3 (a)–(e). These images may serve for a visual assessment of the binary classification outcome and to aid understanding of the other results. The expansion of clusters of low backscatter can seen in part by comparing Figure 2 (a)–(e) with Figure 3 (a)–(e).

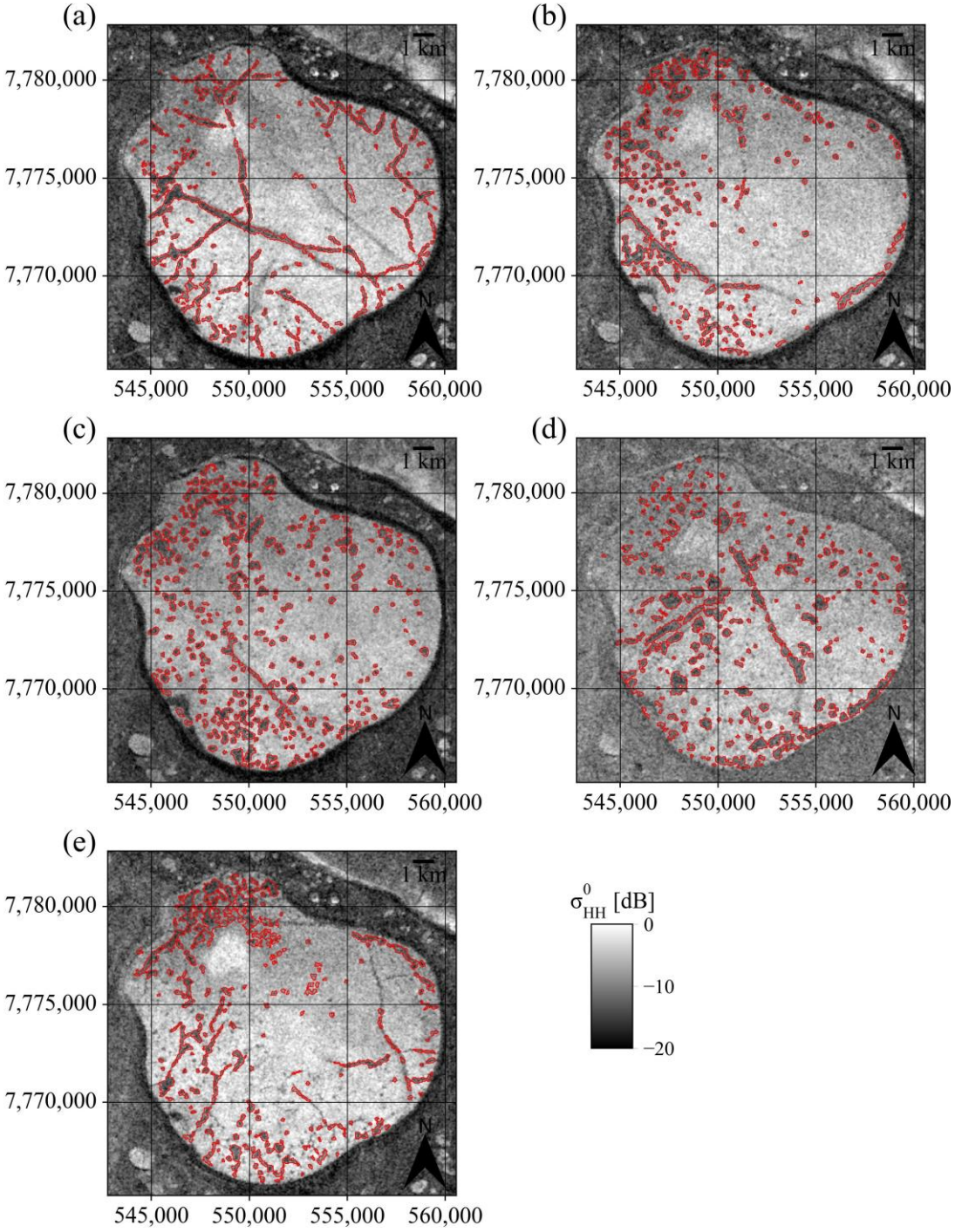

Figure 2: Sentinel-1 EW HH-polarized SAR images where the classified pattern area first exceeded 20 km2. Red outlines show polygonized results of the positive class from the automatic binary classification. (a) 2015, (b) 2016, (c) 2017, (d) 2018, (e) 2019.

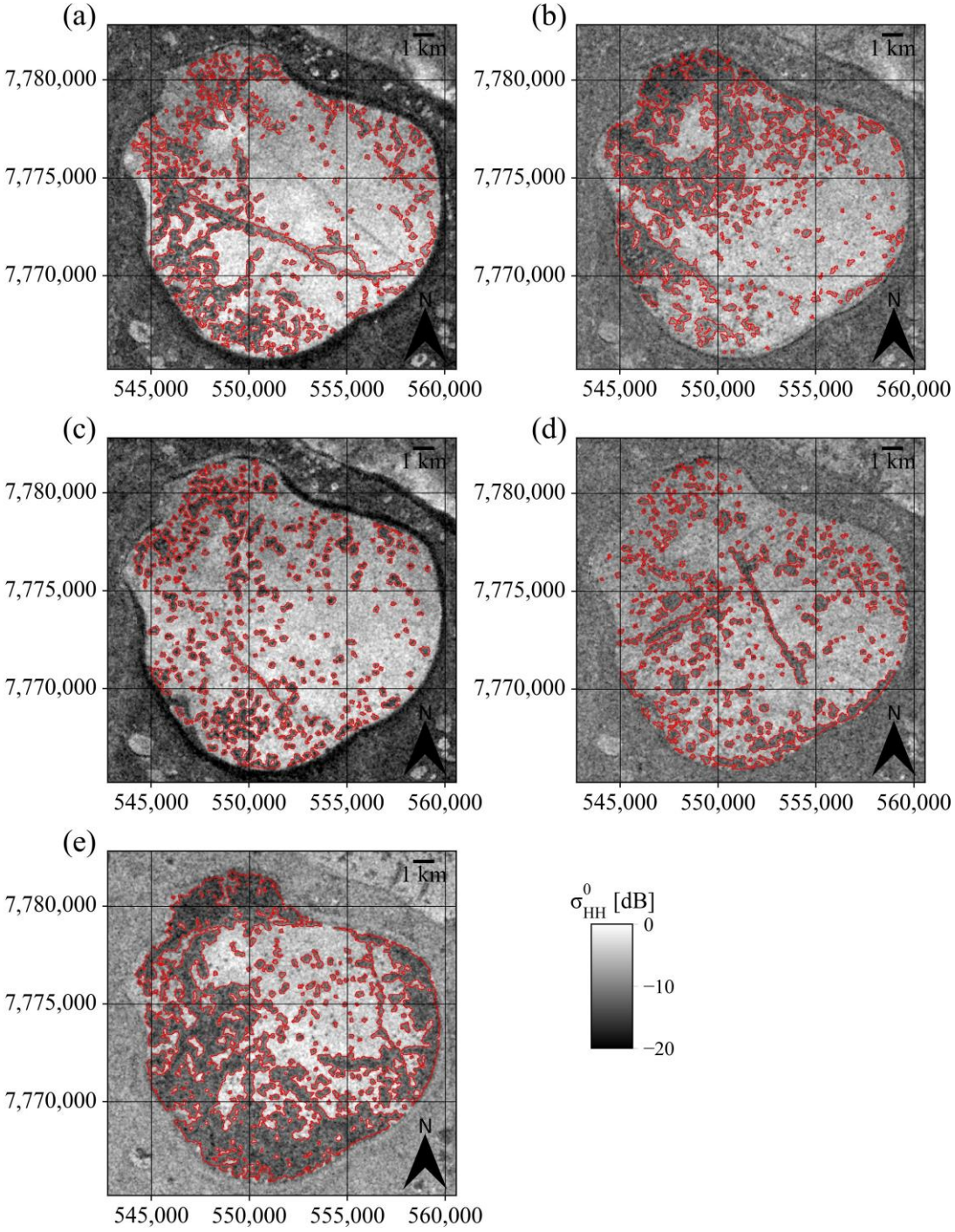

Figure 3: Last available Sentinel-1 EW HH-polarized SAR images before snowmelt onset in each year from 2015 to 2019. Red outlines show polygonized results of the positive class from the automatic binary classification. (a) 2015, (b) 2016, (c) 2017, (d) 2018, (e) 2019.

For an assessment of differences in ice conditions across the years, backscatter levels of regular floating lake ice (negative class) may be of interest. For the images in Figures 2 and 3, mean σ^0 of pixels in the negative class of each polarization channel is of similar magnitude across years (Table 1).

Table 1: Mean backscatter coefficient σ0 of pixels in the negative class (regular floating lake ice) for the years 2015 to 2019 and the two criteria used for image selection. Images for criterion 1 are those where the area of the anomalies first exceeded 20 km²; images for criterion 2 are the last useful acquisitions in the year. The HH-channel images for both criteria are shown in Figures 2 and 3, respectively.

Year	σ^0_{HH} criterion 1 (Figure 2)	σ^0_{HV} criterion 1	σ^0_{HH} criterion 2 (Figure 3)	σ^0_{HV} criterion 2
2015	-5.9 dB	-17.8 dB	-5.5 dB	-17.2 dB
2016	-5.8 dB	-18.3 dB	-6.5 dB	-18.5 dB
2017	-7.1 dB	-17.1 dB	-6.2 dB	-17.4 dB
2018	-6.8 dB	-19.7 dB	-7.6 dB	-20.3 dB
2019	-5.4 dB	-19.4 dB	-5.5 dB	-18.8 dB

In general, a steady increase of pattern area can be observed in late winter in every year from 2015 to 2019 (Figure 4 (a)–(e)). Alongside this general trend, minor fluctuations in the classified area of low backscatter are visible for 2016, 2017, 2018 and 2019. These fluctuations are particularly apparent in early 2018. They may be caused partly by noise in the images or by imperfections in the classification method. Rare small clusters of low backscatter that revert to high backscatter may also contribute. Since no reference data are available, it is impossible to state the cause of the fluctuations with confidence, but air temperatures close to or slightly above 0°C may play a role. However, for this study, the important aspects are the relatively steady increase of pattern area in late winter and the time of the start of this general uptrend. Variations in air temperature behave differently among the years, but are of similar magnitude, and air temperature rarely exceeds 0°C during the analysis periods of backscatter anomalies.

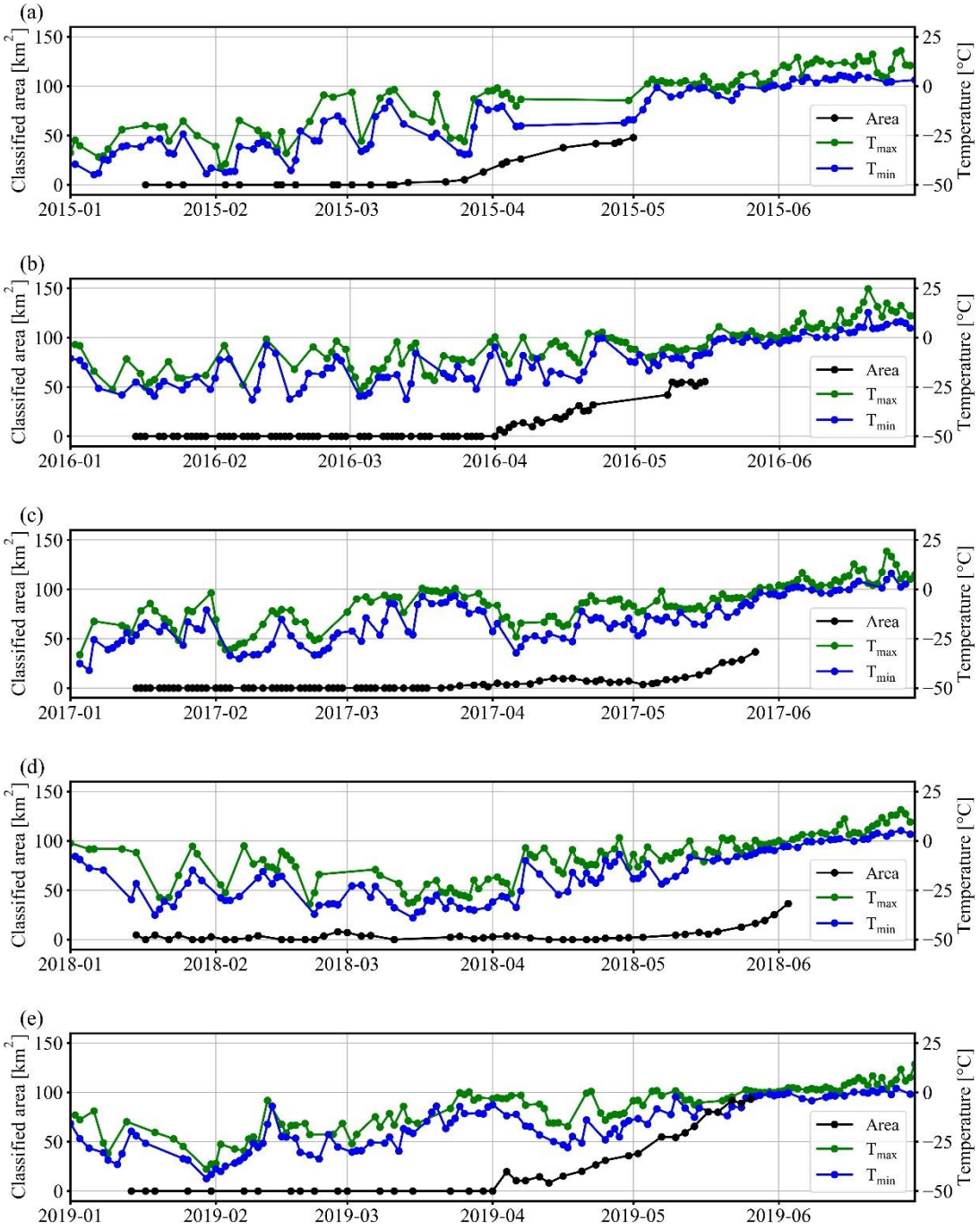

Figure 4: Temporal evolution, January to June, of classified area of clusters of low backscatter (black), and minimum (blue) and maximum (green) air temperatures recorded at the Seyaha weather station for (a) 2015, (b) 2016, (c) 2017, (d) 2018, (e) 2019.

3.2 Mean Minimum Distances

Table 2 shows the results of mean minimum distance calculations for the first criterion, where pattern area first exceeded 20 km². The mean minimum distances range from 65 metres to 227 metres, where most lie between 100 and 200 metres. Similarly, mean minimum distances for the second criterion, the last available useful acquisitions in the year, can be seen in Table 3. The mean minimum distances range from 52 to 227 metres. Of particular note is that mean minimum distances in the 2019-column in Table 3 are significantly smaller in comparison with Table 2. Smaller distances would be expected, due to the expansion of patterns as described above. Distances between consecutive years are of similar magnitude to those found over the longer period.

Table 2: Mean minimum distances between objects pairwise for different years, where pattern area first exceeded 20 km2. The figures are for the mean minimum distance between classified objects in the year in the row to all classified objects in the year in the column.

Year	2015	2016	2017	2018	2019
2015	-	175 m	69 m	118 m	137 m
2016	139 m	-	65 m	97 m	123 m
2017	137 m	210 m	-	117 m	178 m
2018	156 m	227 m	107 m	-	176 m
2019	173 m	170 m	88 m	91 m	-

Table 3: Mean minimum distances between objects pairwise for different years, for the last useful acquisitions in the year in question. The figures are for the mean minimum distance between classified objects in the year in the row to all classified objects in the year in the column.

Year	2015	2016	2017	2018	2019
2015	-	60 m	63 m	120 m	25 m
2016	132 m	-	94 m	140 m	18 m
2017	65 m	52 m	-	117 m	15 m
2018	82 m	77 m	77 m	-	18 m
2019	201 m	79 m	108 m	161 m	-

3.3 Intersections

The intersections deduced from the results for the first criterion, where pattern area first exceeded 20 km², are generally rather low, with 30% being the maximum and most others being between 10% and 20% (Table 4). In comparison, the results of the intersection calculations for the second criterion, the last available useful acquisitions in a particular year, are displayed in Table 5. As could be expected, intersections are generally larger, but they do not exceed 50%, except for the percentages of intersections deduced using the polygon set

from 2019, which are approximately two thirds (Table 5, 2019-column). As with the mean minimum distance calculations, no clear differences between the intersections for consecutive years and those between other years can be seen.

Table 4: Intersections of area of objects classified in one year (row) with area of objects classified in another year (column), where pattern area first exceeded 20 km2. The results correspond to the area of intersection between the two years, divided by the area in the row-year.

Year	2015	2016	2017	2018	2019
2015	-	16 %	18 %	17 %	14 %
2016	17 %	-	23%	25 %	30 %
2017	15 %	18 %	-	18 %	17 %
2018	14 %	21 %	18 %	-	15 %
2019	15 %	30 %	22 %	19 %	-

Table 5: Intersections of area of objects classified in one year (row) with area of objects classified in another year (column) for the last useful acquisitions in the years concerned. The results correspond to the area of intersection between the two years, divided by the area in the row-year.

Year	2015	2016	2017	2018	2019
2015	-	45 %	28 %	27 %	68 %
2016	38 %	-	29%	27 %	65 %
2017	37 %	45 %	-	27 %	66 %
2018	36 %	44 %	26 %	-	65 %
2019	35 %	39 %	26 %	25 %	-

3.4 Cumulative counts of occurrences

Cumulative pixel counts are visualized in Figure 5 (a) for the images where pattern area first exceeded 20km², and in Figure 5 (b) for the last useful acquisitions in the years concerned. The colourbar indicates how often a pixel was classified positively in the five images taken in the five different years (one image/year). Figure 5 (a) is clearly dominated by cumulative counts of 1 and 2, while in Figure 5 (b), where pattern area in the single images is generally larger, wider areas of counts higher than 2 can be observed, although counts of 4 and 5 are still relatively rare. Both sub-figures clearly show that the occurrence of clusters of low backscatter is a lot more frequent in the northern and western parts of the lake.

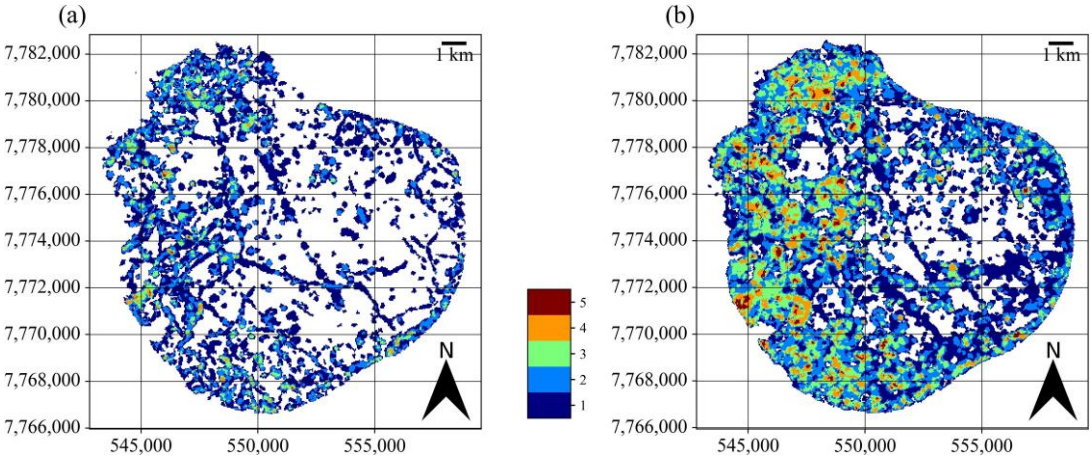

Figure 5: Cumulative pixel counts of backscatter anomalies identified from Sentinel-1 acquisitions from 2015 to 2019: (a) for images where classified area first exceeded 20 km2, (b) for last useful images in the years concerned.

4 Discussion

In this study, we base our analyses on the comparison of images from five different years selected by applying two criteria. Conventionally, time series analyses refer to day of year (DOY). Our results demonstrate (Figure 4) that the time at which the pattern area starts to increase steadily varies significantly between years. This fact supports our choice of an alternative approach – the spatial extent of the phenomenon. In general, it difficult to find a suitable criterion, since some clusters form earlier and some later in the year, and this timing may also vary spatially between years. With our approach, we tried to provide an overview of similarities and variations of locations from 2015 to 2019. The threshold of 20 km^2 was chosen for our analysis of the main locations from which large clusters expand. A lower threshold may not be useful due to limited spatial resolution (40-metre pixel spacing).

Binary classification results generally give a good visual impression. In situ data are, however, not available due to remoteness and safety reasons. It can be assumed that locations of anomalies are characterized by thin ice which cannot be traversed. This assumption is supported by reports from reindeer herders, who observed very thin ice on one large lake on Yamal, where ice thickness is usually more than one metre (Pointner et al., 2019). The lack of direct reference data collected on site also impedes an assessment of ice conditions in relation to the weather in different years. We could only compare our results with air temperature data recorded at a weather station located 80 km away on the coast. However, these temperatures were below 0°C throughout almost the whole of the analysis periods and variations were of similar magnitude across years, which suggests similar ice conditions in different years. This assumption is also supported by similar backscatter values reported for regular floating lake ice (negative class in our classification) across the years.

Some point-like and elongated objects can be seen, especially in Figures 2 (a), (b) and (c), which are characterized by medium contrast with the surrounding high backscatter of floating lake

ice and may belong to the class of low backscatter clusters. However, since no in-situ data are available, we used a conservative strategy and focused on mapping objects that are characterized by higher contrast.

Some objects, primarily visible in earlier acquisitions, may resemble linear fractures in the Sentinel-1 SAR images (especially in Figures 2 (a) and (e)). However, the characteristic expansion of the areas over time is the same for these objects (compare Figures 3 (a) and (e)) and fractures commonly exhibit high backscatter, which is often obscured by the high backscatter of floating lake ice in late winter (Duguay & Pietroniro, 2005).

Linear features in optical and SAR imagery of ice on the lakes on Yamal may also be attributed to leakage of methane and associated geological structures and faults (Bogoyavlensky et al., 2016, 2018). If the clusters of low backscatter for ice on lake Neyto were indeed caused by upwelling gas and accumulation under the ice layer, analysing any changes of locations could be interesting to determine the duration of seeping from a particular point source. However, the limited spatial resolution may be problematic for that purpose.

The calculated mean minimum distances may not at first seem large when compared to the 40-metre pixel spacing, but keep in mind that if two polygons overlap, the minimum distance is zero. So, we still consider the calculated distances as signs of large spatial variations of occurrences. Although we argue that this metric is the most difficult to interpret, we nevertheless think it is useful for highlighting that the variations between consecutive years are similar to those between other years.

Intersection metrics are easier to interpret. Intersections calculated for the images where the classified pattern area first exceeded 20 km² (Table 4) are generally rather low, often below 20%, which shows that clusters emerge mainly from different regions of the lake in different years, although there is always some overlap. Intersections calculated for the last useful acquisitions in any one year (Table 5) are significantly higher, which can be explained partly by the fact that the classified pattern area is generally larger for the images used here. Intersections with the polygon set deduced from the 2019 image range from 65% to 68% (last column on the right in Table 5), but the classified area for the 2019 image also covered nearly half of the total lake area (compare Figures 3 (e) and 4 (e)).

Strong spatial variations of cluster locations can also be seen in Figure 5. There is a strong spatial difference in the emergence of patterns between the five years (Figure 5 (a)), although higher cumulative pixel counts can be seen in Figure 5 (b), where the classified pattern areas in the single images were generally higher. Especially interesting is the more frequent occurrence in northern and western parts of the lake. Bogoyavlensky et al. (2018) discuss the occurrence of patterns in these same parts of lake Neyto in single images. They associate the patterns (but without quantitative analyses) with a nearby gas field that stretches under these areas of lake Neyto. Our results may support their assumption that the clusters of low backscatter are caused by gas emissions.

The cumulative pixel counts (Figure 5) provide some insight into the spatial variation of backscatter anomalies during the five years. However, the analysis of spatial relationships between individual objects (mean minimum distances and intersections), as presented here in relation to backscatter anomalies of lake ice for the first time, may reveal actual variations

between the years. Results suggest that spatial changes between consecutive years are similar to changes over the entire time period, which cannot be deduced from cumulative pixel counts.

5 Conclusions

The purpose of this study was to examine the interannual variability of clusters of low backscatter on Sentinel-1 SAR images of lake ice on lake Neyto in northwestern Siberia. Our results show that there are significant spatial variations in occurrences of clusters of low backscatter between the years 2015 to 2019, although there is also always some overlap. Geospatial analysis reveals that variations are of similar magnitude, whether we look at consecutive years or the longer period. Linear structures and the more frequent occurrence of backscatter anomalies in the northern and western parts of lake Neyto may point to gas release as the primary cause of the anomalies. Methods commonly used for assessing spatial relationships of vector data can provide valuable insight into the phenomenon.

Acknowledgements

This work was supported by the HORIZON2020 (BG-2017-1) project Nunataryuk and the doctoral college DK GIScience at the University of Salzburg.

We thank the reviewers and editors for their time and the constructive review process.

Contains modified Copernicus Sentinel data (2015, 2016, 2017, 2018, 2019).

References

Atwood, D. K., Gunn, G. E., Roussi, C., Wu, J., Duguay, C., & Sarabandi, K. (2015, Nov). Microwave Backscatter From Arctic Lake Ice and Polarimetric Implications. IEEE Transactions on Geoscience and Remote Sensing, 53(11), 5972-5982.

Bogoyavlensky, V. I., Sizov, O. S., Bogoyavlensky, I. V., & Nikonov, R. A. (2016). Remote Detection of Surface Gas Shows Zones and Gas Blowouts in the Arctic: The Yamal Peninsula. *Arctic: Ecology and Economy, 3*(23), 4–15. (in Russian)

Bogoyavlensky, V. I., Sizov, O. S., Bogoyavlensky, I. V., & Nikonov, R. A. (2018). Technologies for Remote Detection and Monitoring of the Earth Degassing in the Arctic: Yamal Peninsula, Neito Lake. *Arctic: Ecology and Economy, 2*(30), 83–93. (in Russian)

Burn, C. R. (2005). Lake-bottom thermal regimes, western Arctic coast, Canada. *Permafrost and Periglacial Processes, 16*(4), 355–367.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Drusch, M., Bello, U. D., Carlier, S., Colin, O., Fernandez, V., Gascon, F., ... Bargellini, P. (2012). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment, 120* (Supplement C), 25 -36.

Duguay, C. R., & Lafleur, P. M. (2003). Determining depth and ice thickness of shallow sub-Arctic lakes using space-borne optical and SAR data. *International Journal of Remote Sensing, 24*(3), 475-489.

Duguay, C. R., & Pietroniro, A. (2005). *Remote Sensing in Northern Hydrology: Measuring Environmental Change*. Washington DC American Geophysical Union Geophysical Monograph Series, 163.

Duguay, C. R., Pultz, T. J., Lafleur, P. M., & Drai, D. (2002). RADARSAT backscatter characteristics of ice growing on shallow sub-Arctic lakes, Churchill, Manitoba, Canada. *Hydrological Processes, 16*(8), 1631–1644.

Engram, M., Anthony, K. W., Meyer, F. J. & Grosse, G. (2013). Synthetic Aperture Radar (SAR) Backscatter Response from Methane Ebullition Bubbles Trapped by Thermokarst Lake Ice. *Canadian Journal of Remote Sensing, 38*(6), 667–682.

ESA. (2013). *Sentinel-1 User Handbook*. European Space Agency.

Gillies, S. (2007). Shapely: manipulation and analysis of geometric objects. Retrieved from https://github.com/Toblerity/Shapely

Gillies, S. (2011). Fiona is OGR's neat, nimble, no-nonsense API. Retrieved from https://github.com/Toblerity/Fiona

Grunblatt, J., & Atwood, D. (2014). Mapping lakes for winter liquid water availability using SAR on the North Slope of Alaska. *International Journal of Applied Earth Observation and Geoinformation, 27*, 63–69.

Jeffries, M. O., Morris, K., Weeks, W. F., & Wakabayashi, H. (1994). Structural and stratigraphic features and ERS 1 synthetic aperture radar backscatter characteristics of ice growing on shallow lakes in NW Alaska, winter 1991–1992. *Journal of Geophysical Research: Oceans, 99*(C11), 22459–22471.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159–174.

Menne, M. J., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., . . . and others (2012). *Global Historical Climatology Network-Daily (GHCN-Daily), version 3.26*. NOAA National Climatic Data Center. (accessed on 6 April 2020)

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., & Houston, T. G. (2012). An overview of the global historical climatology network-daily database. *Journal of atmospheric and oceanic technology, 29*(7), 897–910.

Pointner, G., Bartsch, A., Forbes, B. C., & Kumpula, T. (2019). The role of lake size and local phenomena for monitoring ground-fast lake ice. *International Journal of Remote Sensing, 40*(3), 832-858.

Surdu, C. M., Duguay, C. R., Brown, L. C., & Fernández Prieto, D. (2014). Response of ice cover on shallow lakes of the North Slope of Alaska to contemporary climate conditions (1950-2011): radar remote-sensing and numerical modeling data analysis. *The Cryosphere, 8*(1), 167–180.

Surdu, C. M., Duguay, C. R., Pour, H. K., & Brown, L. C. (2015). Ice Freeze-up and Break-up Detection of Shallow Lakes in Northern Alaska with Spaceborne SAR. *Remote Sensing, 7*(5), 6133–6159.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., ... and others (2014). scikit-image: image processing in Python. PeerJ, 2, e453. Retrieved from https://doi.org/10.7717/peerj.453

Wakabayashi, H., Weeks, W. F., & Jeffries, M. O. (1993). A C-band backscatter model for lake ice in Alaska. In *Proceedings of IGARSS '93 - IEEE International Geoscience and Remote Sensing Symposium* (Vol. 3, pp. 1264-1266).

Walter Anthony, K. M., Anthony, P., Grosse, G., & Chanton, J. (2012). Geologic methane seeps along boundaries of Arctic permafrost thaw and melting glaciers. *Nature Geoscience, 5*(6), 419–426.

Wik, M., Varner, R. K., Anthony, K. W., MacIntyre, S., & Bastviken, D. (2016). Climate-sensitive northern lakes and ponds are critical components of methane release. *Nature Geoscience, 9*(2), 99–105.

Yen, J.-C., Chang, F.-J., & Chang, S. (1995). A new criterion for automatic multilevel thresholding. *IEEE Transactions on Image Processing, 4*(3), 370–378.

GI_Forum 2020, Issue 1
Page: 63 - 72
Full Paper
Corresponding Author:
irada.ismayilova@geo.uni-augsburg.de
DOI: 10.1553/giscience2020_01_s63

Land Suitability Analysis of Alvar Grassland Vegetation in Estonia Using Random Forest

Irada Ismayilova[1], Evelyn Uuemaa[2], Aveliina Helm[2], Christian Röger[1] and Sabine Timpf[1]

[1]University of Augsburg, Germany
[2]University of Tartu, Estonia

Abstract

Calcareous alvar grasslands are one of the most species-rich habitats in Estonia. Land-use change and cessation of traditional agricultural practices have led to a decrease of the area of these valuable grasslands during the past century. Therefore, their conservation and restoration are becoming increasingly important. Efforts to restore these habitats have already been made in recent years. Land suitability analysis for potential restoration sites, using the machine learning technique Random Forest (RF), was performed for the first time in this study, which aimed to assess the use of RF for a suitability analysis of alvar grassland. RF predicted 610.91 km^2 of areas suitable for restoring alvar grasslands or for creating alvar-like habitats in Estonia. These areas include all existing alvar areas as well an additional 140.91 km^2 suitable for establishing new habitat similar to calcareous alvar grasslands. We discuss suitability analysis to help with restoration planning and find it to be a reasonable and efficient tool that has potential to provide relevant information. The quality of the prediction could be improved by including additional data relevant for alvar grasslands, such as soil depth, but such data was unfortunately unavailable.

Keywords:

alvar grasslands, restoration, land suitability, machine learning, Random Forest

1 Introduction

Alvars are calcareous grassland habitats with a limited distribution on Earth. They are found mostly in Estonia, Sweden and a few other smaller areas in the Northern hemisphere (Pärtel et al. 1999). These grasslands have high conservation value both in Estonia and in Europe as a whole due to their species richness, the variety of their important ecosystem services, and their high relevance in supporting natural and cultural heritage in European landscapes. Alvar grasslands are among Annex I of priority habitat types in the EU Habitats Directive (*6280 Nordic alvar and Precambrian calcareous flatrocks) (Rosén 1982). Over the millennia, Estonian alvar grasslands have been supported by moderate human influence, especially grazing. Land-use change, leading to the cessation of grazing, afforestation or direct destruction, have resulted in substantial decreases in area of alvars and the subsequent loss of

their species richness (Helm, Hanski & Pärtel 2006). The current distribution of alvar grasslands is shown in Figure 1.

Considering their high conservation value and position among priority habitat types in the Natura 2000 network, alvar grasslands as well as other types of dry calcareous grassland are in urgent need of restoration and more effective conservation (Helm, Urbas & Pärtel 2007). The most recent restoration work has focused on shrub clearance and the removal of other unwanted vegetation in long-abandoned or afforested alvar grasslands, with the aim of restoring original alvars (Holm 2019, Helm 2019). One managerial measure would be to look for suitable areas where new alvar-like habitats could be created.

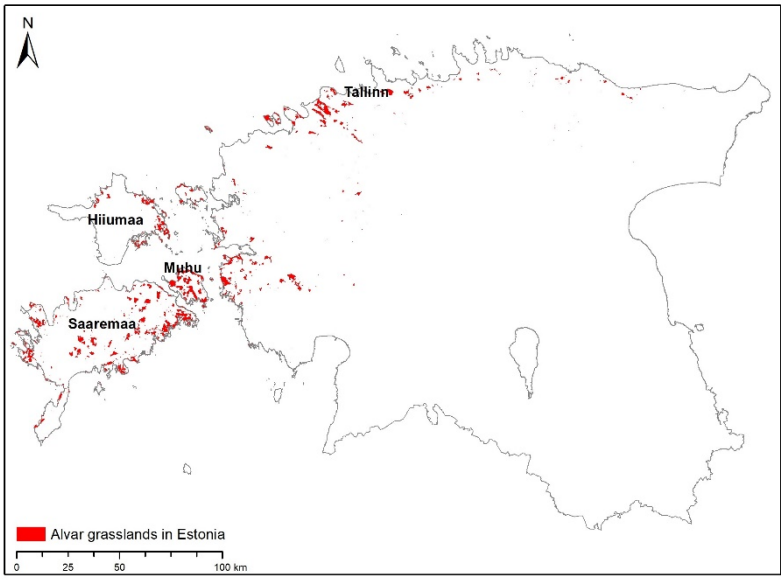

Figure 1: Distribution of alvar grasslands in Estonia

From 2014 to 2019, ca 2,500 ha of overgrown alvar grasslands were restored in western Estonia as part of the LIFE to Alvars project (Holm 2019). This as well as other grassland restoration projects have focused solely on historical grassland areas that have been degraded by becoming overgrown with trees and shrubs. However, for future restoration planning it would be helpful to identify all potential regions where grassland restoration or the creation of new grassland areas would be environmentally feasible.

Land suitability analysis is one of the most frequently used techniques in environmental management and the planning of habitat restoration. For example, Novak and Short (2000) performed a suitability analysis for eelgrass meadows on Plum Island. Hunter et al. (2016) carried out a restoration suitability assessment for swamps in order to safeguard and improve the provision of important ecosystem services. However, land suitability analysis for alvar grasslands in Estonia has not been performed so far. Therefore, this study aims to identify potentially suitable areas of alvar grassland for restoration or for the creation of alvar-like

habitats. We use a method from machine learning, called Random Forest (RF), because of the limitations of available datasets relating specifically to alvar grasslands and yet a large amount of data to process. The literature on RF confirms that it is timesaving for handling large amounts of data and capable of highly accurate predictions. Our analysis covers the whole of Estonia.

2 Random Forest method

Machine learning methods are becoming increasingly popular in land suitability analysis thanks to their ability to deal with complex relationships between predictor variables, robustness in managing big and noisy data, and being economical in terms of time required (Lahssini et al. 2015). RF, as proposed by Breiman (2001, p. 6), is "a classifier consisting of a collection of tree-structured classifiers {h(x,Θk), k=1, ...} where the {Θk} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x". This method is an extension of classification and regression trees and uses the Classification and Regression tree algorithms (CART). The classification algorithm is used when aiming to predict categorical values or labels; the regression algorithm is used when aiming for numerical predictions (Strecht et al. 2015). RF can be described as trees where branches are formed by answers to yes/no questions and are not pruned (but can be). Each tree in the forest is constructed using bootstrap samples from the original dataset. It uses a random selection of explanatory variables or factors to split the tree at nodes.

The goal of RF is to identify the best model to analyse the relationship between dependent and independent variables (Friedman and Meulmann, 2003). In order to evaluate how good the RF model is, the data needs to be split into two parts: training and testing data. This helps to evaluate the performance of the algorithm for the chosen problem by training one sample of data and validating it using the test sample. RF, in both classification and regression models, also provides a measure of the importance of a variable based on the contribution of the variable to the model at each node and each tree where it appears. Another estimated value that can be obtained from the model is the Out Of Bag (OOB) score, an average error of prediction of out of bag samples (samples that do not appear in bootstrap samples (Breiman 2001). RF has also proved to be a suitable method when there is a correlation between the variables involved in the analysis (Georgian et al. 2019).

Several machine learning techniques have already been incorporated into land suitability analysis. For instance, Wen et. al. (2009) used classification and regression trees to investigate hydrological requirements of the river Red, while Park et al. (2003) applied an artificial neural network to predict aquatic insect species. Landscape configuration and habitat suitability were analysed by Holzkaemper et al. (2006) using genetic and simulated annealing algorithms. However, many studies have shown that RF most often attains the best predictive performance (Garzon et al. 2006). Lahssini et al. (2015) and Vincenzi et al. (2011) used RF to detect cork oak suitability and *Ruditapes philippinarum*'s potential spatial distribution assessment respectively. The probability of correct predictions in both studies was more than 90%.

3 Using Random Forest to Predict Suitability for Alvar grasslands

We predicted the probability of the suitability of areas throughout Estonia for restoration of alvar grassland or for the creation of alvar-like habitats. In stage one, we focused on defining environmental variables and extracting predictor variables. In the second stage, we used the Scikit-Learn library in Python for the prediction of suitability and defined the most suitable probability threshold for the given question. Results were visualized, and were examined for reliability by experts in the department botany at the University of Tartu (UT).

3.1 Choosing environmental predictors

The main data sources used in this work were the Estonian Soil Database (ESD), a LiDAR-based Digital Elevation Model (DEM), and the Estonian Geological Database. From these datasets, we chose eight predictor variables for further use.

The occurrence of alvar grasslands is limited to three main bedrocks: Silurian, Ordovician, and to a lesser extent Cambrian (Pärtel et al. 1999). Bedrock is considered the most important predictor variable in identifying suitable areas for alvar grassland. Alvar-type vegetation occurs only on thin and calcareous soils. Usually the soil depth over the bedrock is less than 20 cm, and in some alvars it is even less than 5cm (Pärtel et al. 1999). Therefore, we extracted soil-type and soil-texture information from the ESD. Since soil silt, soil sand and soil rock content describe soil state and condition, they were also used as predictor variables. No concrete example of a correlation between the slope of terrain, Topographic Wetness Index (TWI) and alvar-like vegetation in Estonia has been established. However, prior statistical analysis of available datasets showed that slope and TWI values under alvar grasslands are always within a certain range. Therefore, we considered using slope and TWI as further predictor variables to ensure higher information gain for the RF models. A DEM with 1 m resolution was used to calculate slope and TWI. The datasets, their sources, and the predictor variables involved in the suitability analysis are shown in Table 1.

Table 1: Environmental datasets and variables used in land-use suitability predictions

Datasets	Source	Predictor variables (data type)
Soil Database	Estonian Land Board & Kmoch et. al., (2019)	Soil type (categorical)
		Soil texture (categorical)
		Soil silt content (numerical)
		Soil sand content (numerical)
		Soil rock content (numerical)
DEM	Estonian Land Board	Slope (numerical)
		TWI (numerical)
Geological Database	Estonian Land Board	Bedrocks (categorical)

Data for the location of Alvar grasslands in Estonia was provided by the Botany Department of UT, in the form of two datasets. One of the two contained the most recent alvar grassland distribution information available for Estonia. This dataset is a product of the survey of the Estonian Semi-Natural Community Conservation (2000–2010) and alvar distribution mapping based on the Estonian state-run database EELIS. The second dataset is a result of the Estonian vegetation mapping from 1930 to 1950 and was helpful to understand the historical distribution of alvars. We merged these datasets to create a single one in which we assigned "1" to all areas indicating presence of alvar grasslands. For absence data, we generated random points in the areas outside of alvar grasslands with suitable bedrocks. The absence of alvar grasslands was indicated by "0".

In order to assure identical extent, cell size and coordinate systems for the suitability analysis, pre-processing of layers was carried out in ArcGIS. This step resulted in one big joint database, with 41,657 objects indicating presence ("1") and absence ("0") of alvar grasslands and containing predictor variables.

3.2 Suitability modelling using Random Forest

In order to identify suitable areas using RF, a list of variables (predictors), summarized in Table 1, was used. As a first step, we did one-hot encoding for categorical variables. We then created five models with different combinations of predictor variables in order to find the most suitable combination. As part of the general procedure, we split the data into training and test sets. There is no information available on which proportions for splitting the data work best, but in many similar studies, datasets have been divided into 60/40 or 70/30 ratios. We therefore used both these ratios and chose the better option (Table 2). The target variable in the training phase was the alvar grassland presence or absence data. In order to assess the performance of the trained model (how well it can recognize alvars), test sets were used. Using the "RandomizedSearcCV" function from the scikit-learn library in Python, we aimed to define the best set of parameters (e.g. n_estimators, max_depth).

We estimated the accuracy of the models using the R-squared value and OOB Error estimate produced by k-fold cross-validation with 3-fold. Using the best model parameters from the hyperparameter tuning process and the dataset covering the whole of Estonia (alvar grassland vegetation presence or absence was excluded), we fitted the RF model and obtained continuous values between 0 and 1 representing the probability of suitability. In order to convert the values into binary maps, we had to identify the threshold above which the areas were most suitable, and below which areas had low suitability or were unsuitable. Because of the nature of the analysis, we set the suitability threshold after examining the results.

Table 2: RF Models with different combinations of predictor variables and split options

Base set of predictor variables used in training/test models	Split options
[Model 1] Soil type, Soil texture, Soil clay, Soil silt, Soil sand, Soil rock, Slope, TWI, Bedrock [Model 2] Soil type, Soil clay, Soil silt, Soil sand, Soil rock, Slope, TWI, Bedrock [Model 3] Soil type, Soil texture, Slope, TWI, Bedrock [Model 4] Soil texture, Soil clay, Soil silt, Soil sand, Soil rock, Slope, TWI, Bedrock [Model 5] Soil type, Soil texture, Soil clay, Soil silt, Soil sand, Soil rock, Slope, TWI	[Split option **1**] 60/40 & [Split option **2**] 70/30

4 Results of the suitability analysis

Once all the models had been run, their prediction statistics were checked and R^2 values and OOB scores were examined. The results are shown in Table 3.

Table 3: R-squared and OOB scores for individual RF models

Variant	Split option	R^2 score	OOB score
Model 1	1	0.7617	0.75438
Model 2	1	0.7457	0.74309
Model 3	1	0.7888	0.75424
Model 4	1	0.7783	0.77606
Model 5	1	0.7631	0.76548
Model 1	2	0.7756	0.77503
Model 2	2	0.7884	0.79678
Model 3	2	0.7919	0.79670
Model 4	2	0.7861	0.79668
Model 5	2	0.6692	0.68487

Table 3 shows R^2 and OOB scores for different model variations. There were almost no differences between these figures across all models. Only one model (Model 5) stood out slightly from the rest. It contained soil type, soil texture, slope, TWI and bedrock as predictor variables. Further, split option 2 (70/30) was chosen as suitable for alvar grassland

identification. The accuracy of the selected model was 0.79 and 0.8 for the R^2 and OOB scores respectively. This means that RF has a useful prediction capability. We also checked which predictor variables contribute the most to the information gain of the RF model, and to what extent the accuracy will decrease if a certain variable is removed from the model. We used a Mean Decrease in Accuracy metric which was calculated via the permutation feature importance algorithm (rfpimp; Breiman 2001). The results are shown in Figure 2. In the chosen model, bedrock was the most important predictor while TWI was the least important.

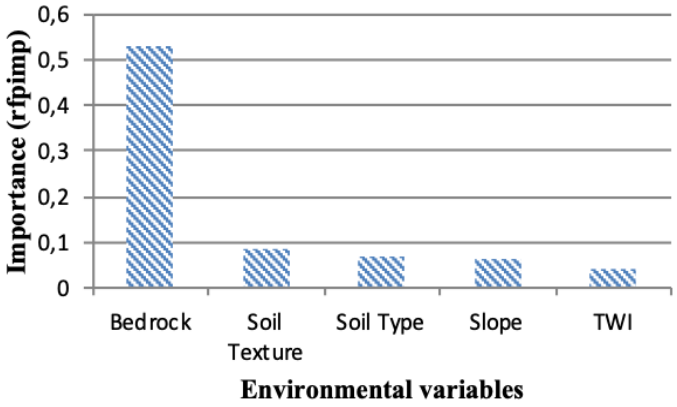

Figure 2: Importance of each variable in the final RF model

The predicted occurrence ranges from 0.04 to 0.884%. Therefore, probability of more than 80% was considered highly suitable; probability of less than 80% was considered unsuitable or of low suitability for alvar grasslands. The aim was to find suitable areas with very high probability, and therefore a high threshold was selected.

Of all the suitable areas for alvar grasslands, 45% are currently forests, 34% are croplands, and 11% are grasslands, as shown in Table 4.

Table 4: Actual land use of the areas predicted to be suitable for alvar grassland restoration using RF

Land use	Areal percentage (%)
Forest	44.94
Cropland	33.49
Grassland	11.02
Other	6.66
Shrubland	2.32
Urban	1.14
Wetland	0.34
Water	0.09

The final result of our analysis is illustrated in Figure 3. RF predicts a total of 610.91 km² where currently no alvar grasslands exist. Out of those, 470 km² were once alvar grasslands. The most suitable areas for alvar grassland restoration in Estonia are in the western islands (Saaremaa, Muhu, Hiiumaa), and north-western and northern inland areas. Low suitability or unsuitable areas fall in southern Estonia.

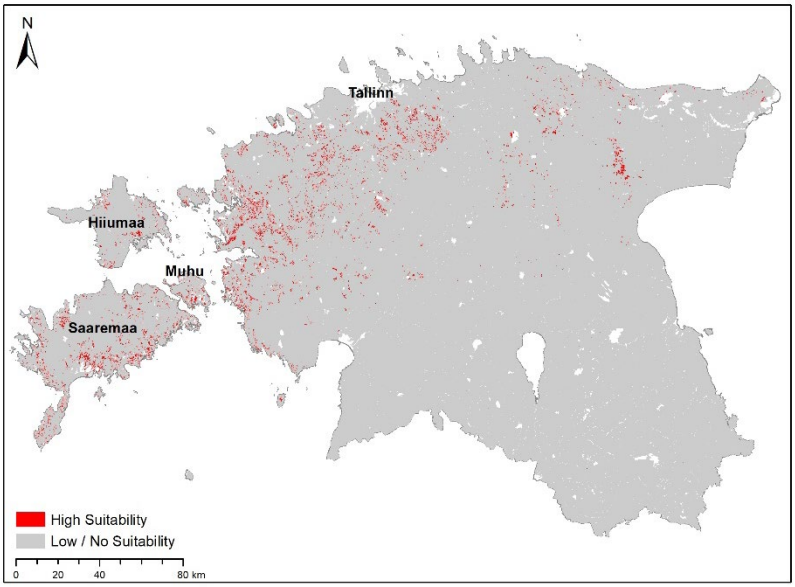

Figure 3: Results of the RF model for highly suitable and less suitable areas for alvar grassland restoration or creation of alvar-like habitats

5 Discussion and Conclusions

In this study, we predicted the potential occurrence of alvar grassland vegetation in Estonia. Limited distribution of these grasslands in Estonia and the small number of environmental variables that characterize alvar grasslands in Estonia makes the suitability analysis very challenging. Our case study shows that RF is a suitable method for finding these areas. It has the ability to learn by itself from the available data (in this case, current locations of alvar grasslands and their properties) and make predictions based solely on data – i.e. without human constraints such as the need for expert knowledge. Compared to other well-known methods such as Multi Criteria Decision Making (MCDM), this reduces massively the time spent preparing the data prior to performing land suitability analysis. Thus, RF is used in many studies dealing with land suitability analysis. Our model performs well, with an accuracy of 80%. Experts in the Botany Department at the University of Tartu, Estonia, confirmed our results. Similar studies applied to other areas and land suitability types such as Garzon et al. (2006) reach an accuracy of 90% or even higher. We believe that the difference between our accuracy and that reached by similar studies is due to limitations caused mainly by our choice of datasets for the analysis.

We expected bedrock, soil texture and soil type to provide the highest contributions to the final model. The data we use does not contain soil depth information, as this is not (yet) available for the whole of Estonia. Albert (1998) states that alvar grasslands occur on thin soils of no more than 20 centimetres. In our dataset, some areas have an actual soil depth record while others show the depth of a soil profile up to one meter. Additionally, soil pH could not be included in our suitability analysis. Including this information should increase the prediction accuracy.

We conclude that RF is a reliable method for performing land suitability analysis for alvar grasslands. An accuracy of 80% is acceptable considering the limitations of our data. Although other information such as soil depth and soil pH is missing, the datasets used perform well with RF.

Future research will focus on applying other land suitability analysis methods such as MCDM in order to allow a comparison of the results, and on optimizing the choice of datasets. Taking into account factors like soil depth and soil pH will most probably enhance our results. Our goal is to reach an accuracy of 90% or higher.

References

Albert, D. A., & Kost, M. A. (1998). Natural community abstract for lakeplain wet prairie. *Michigan Natural Features Inventory, Lansing, MI.*

Breiman, L. (2001). Random forests. *Machine Learning* 45: 5–32

Friedman, J.H., Meulman, J.J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in medicine.* 22 (9): 1365–1381

Garzon, M. B., Blazek, R., Neteler, M., De Dios, R. S., Ollero, H. S., & Furlanello, C. (2006). Predicting habitat suitability with machine learning models: the potential area of Pinus sylvestris L. in the Iberian Peninsula. *Ecological modelling*, 197(3-4), 383-393.

Georgian, S., Anderson, O., Rowden, A. (2019). Ensemble habitat suitability modelling of vulnerable marine ecosystem indicator taxa to inform deep-sea fisheries management in the South Pacific Ocean. *Fisheries Research* 211:256–274

Helm, A. (2019). Large-scale restoration of Estonian alvar grasslands: im-pact on biodiversity and ecosystem services: Final Report of Action D.1

Helm, A., Hanski, I., & Pärtel, M. (2006). Slow response of plant species richness to habitat loss and fragmentation. *Ecology Letters* 9: 72–77

Helm, A., Urbas, P., & Pärtel, M. (2007). Plant diversity and species characteristics of alvar grasslands in Estonia and Sweden. *Acta Pythogeographica Suecica.* 88: 33-42

Holzkaemper, A., Lausch, A. and Seppelt, R. (2006). Optimizing Landscape Configuration to Enhance Habitat Suitability for Species with Contrasting Habitat Requirements. *Ecological Modelling* 198: 277-292

Holm, A. (2019). Life to alvars project: Report of Action C.1

Hunter, R., Day, J., Shaffer, G., Lane, R.,Englande, A., Reimers, R., Kandalepas, D., Wood, William B., Day, J., Hillmann, E., Bank, E. (2016). Restoration and Management of a Degraded Baldcypress Swamp and Freshwater Marsh in Coastal Louisiana. *Water* 8: 79-101

Lahssini, S., Lahlaoi, H., Mharzi, H., Bagaram, M., Ponette, Q. (2015). Predicting Cork Oak Suitability in Maˆamora Forest Using Random Forest Algorithm. *Journal of Geographic Information Systems* 7: 202-210

Novak, B., Short, T. (2000). Creating the Basis for Successful Restoration: An Eelgrass Habitat. *Ecological Engineering* 15: 239-252

Park, S., Céréghino, R., Compin, A. and Lek, S. (2003). Applications of Artificial Neural Networks for Patterning and Predicting Aquatic Insect Species Richness in Running Waters. *Ecological Modelling* 160: 265-280

Pärtel M.,Mändla R. & Zobel M. (1999). Landscape history of a calcareous (alvar) grassland in Hanila, western Estonia, during the last three hundred years. *Landscape Ecology* 14: 187-196

Rosén, E. (1982). Vegetation development and sheep grazing in limestone grasslands of south Öland, Sweden. *Acta Phytogeographica Suecica*. 72: 1-104

Strecht, P., Cruz, L., Soares, C., Moreira, J., Abreu, R. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. *International Educational Data Mining Society*.

Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranovi, F., De Leo, G.A. and Torricelli, P. (2011). Application of a Random Forest algorithm to Predict Spatial Distribution of the Potential Yield of Ruditapes philippinarum in the Venice Lagoon, Italy. *Ecological Modelling* 222: 1471-1478

Wen, L., Ling, J., Saintilan, N. and Rogers, K. (2009). An Investigation of the Hydrological Requirements of River Red Gum (Eucalyptus camaldulensis) Forest, Using Classification and Regression Tree Modelling. *Ecohydrology* 2: 143-155

GI_Forum 2020, Issue 1
Page: 73 - 88
Full Paper
Corresponding Author:
t.schneider@fh-kaernten.at
DOI: 10.1553/giscience2020_01_s73

Towards Predicting Vine Yield: Conceptualization of 3D Grape Models and Derivation of Reliable Physical and Morphological Parameters

Thomas Schneider[1], Gernot Paulus[1] and Karl-Heinrich Anders[1]
[1]Carinthia University of Applied Sciences (FH Kärnten), Austria

Abstract

In viticulture, yield prediction plays an important role, helping winegrowers to predict the start of the next growth stage of vines and to improve vineyard management decision-making. To predict a vineyard's yield, it is necessary to gather accurate local information about the vine's phenology and morphology, such as the volume of individual grapes. Traditional collection of these data and yield prediction rely on resource- and time-intensive direct visual and manual in-field work by viticulturists. Thus, only limited sampling in the vineyards is possible, carried out by humans. Automated procedures utilizing sensor-based systems reduce the data acquisition time and enable the collection of high-resolution data from the entire vineyard. Large-scale 3D models of vineyards can be generated from these data and used to analyse, for example, the vineyard's yield or the vegetative stage of individual vines.

We propose a concept for a 3D model that uses close-range photogrammetry. In a laboratory experiment, we tested the acquisition of multi-view image datasets from grapes using close-range photogrammetry and derived physical and morphological parameters from 3D grape models. The results could contribute to the design and implementation of a large-scale in-field experiment.

Keywords:

precision viticulture, vine, 3D grape model, physical parameters, morphological parameters

1 Introduction

Vines play an important role in the lives of many, and their connection to mankind can be traced back to ancient times (Poupin et al., 2011). About 28.4 million litres of wine were produced in 2015, an increase of 3.5% since 2014 (Wine Institute, 2019). Due to the constantly increasing consumption and demand for high-quality wines, winegrowers need to find efficient approaches to guarantee effective vine-growing and high-quality grapes. To facilitate efficient and high-quality viticulture, vineyard managers record phenological data of their vines on a weekly basis (Westover, 2018). The knowledge gained through analysis of phenological data

helps them to execute vineyard management practices at the right time and to determine the beginning of the next growth phase of the plants (Westover, 2018). Furthermore, estimated crop yields derived from phenological information is also helpful for wineries so that they can predict better the size of their crops and how much wine they will produce (Moyer and Komm, 2015). However, phenological data collection is traditionally performed manually (Moyer and Komm, 2015), is time-consuming (Nuske et al., 2014), and can be prone to human error (Rose et al., 2016). The application of sensor-based systems to collect and analyse phenological data can help to avoid gaps in phenological datasets (Westover, 2018), reduce human error and data collection time, make yield prediction more reliable, and reduce labour costs (Rose et al., 2016).

This preliminary study deals with the conceptualization of 3D models of harvest-ripe grapes using close-range photogrammetry and the derivation of physical and morphological parameters from the models. It was conducted in the form of a laboratory experiment.

The knowledge from this study could be utilized for the design and implementation of in-field experiments at vineyard scale. Furthermore, the information obtained helps to evaluate how well empirical physical and morphological measurements of the phenological stages of vines (in this case from harvest-ripe grapes) compare to model-based measurements. Our study also provides data on the image resolution required to create 3D high-density point clouds of individual grapes from images taken by unmanned aircraft system (UAS) flight missions.

2 Precision Viticulture, Phenological Stages of Vines and Yield Prediction

2.1 Precision Viticulture

Scientists and viticulturists strive constantly to find new approaches to improve the efficiency of vineyard management and to increase the quality of grapevines. Rapid advances in the 20th and 21st centuries have provided new digital and spatial technologies and enabled new sciences with a focus on viticulture, including precision viticulture. Nowadays, sensor-based systems are often used in viticulture, for example to derive phenological data from vineyards (Rose et al., 2016) or to predict a vineyard's yield (Nuske et al., 2014).

Precision viticulture is defined as the application of geospatial technologies, devices and tools which exploit spatial location to collect, store, manipulate, analyse and visualize environmental data from vineyards. Technologies and systems developed through precision viticulture give winegrowers and viticulturists the opportunity to collect data from their vineyards in real-time with accurate positional information (Goldammer, 2015).

Thus, the application of precision viticulture technologies has a wide range of advantages for viticulture experts. The foremost advantage is the improvement of viticulture management, including not only the enhancement of management techniques but also the ability to improve decisions on when to apply particular techniques. Deploying vineyard-related practices in timely fashion, such as the adjustment of canopy and nutrients, or the detection and elimination of diseases and insects, leads to an increased yield, yield quality and yield security.

Furthermore, it helps to reduce operating costs and the use of chemicals, and to improve the quality of soil and groundwater around a vine (Maniak, 2004).

2.2 Phenological Stages of Vines

In the context of viticulture management and yield prediction, the topic of "phenology" plays an important role. Phenology is the science which focuses on the natural changes and development of organisms and their relationship with seasonal variations in climatic parameters (Centinari, 2018; Westover, 2018; Hellman, 2003). It gives winegrowers crucial information about the health and condition of their plants. An important phenological characteristic of vines are their vegetative changes and developments over the course of a season. The weekly collection of phenological data such as plant growth information is recommended in order to avoid gaps in the datasets (Westover, 2018). One difficulty in trying to measure a vine's phenology is that plant growth is an ongoing process, and individual vines have their own natural growth patterns which are influenced by many different factors. Thus it is difficult to divide the seasonal growth of vines into exact, discrete, phases.

Various scales and classification systems support winegrowers and scientists by dividing the growth process into several phenological stages. A frequently used phenological classification for vines is the Modified Eichhorn-Lorenz (E-L) system. This system was originally designed by Dr. K. W. Eichhorn and Dr. D. H. Lorenz in 1977, revised in 1995 by B. G. Coombe, and modified further in 2004 by Coombe and P. Dry, becoming known as the Modified Eichhorn-Lorenz system (Westover, 2018). The modifications of the original E-L system concern changes in the classification of bud growth, since the visual characteristics in the early stages of bud growth vary among grapevine varieties (Centinari, 2018).

The Modified E-L system classifies the annual growth of vines into 47 vegetative stages, to each of which is assigned a specific E-L number. The system highlights eight of these as major stages: Bud Burst (E-L 4), Shoots 10 cm (E-L 12), Flowering begins (E-L 19), Flowering (E-L 23), Setting (E-L 27), Berries pea-size (E-L 31), Véraison (E-L 35) and Harvest (E-L 38). In the context of precision viticulture, these key phenological stages and their corresponding characteristics are used as benchmarks by winegrowers and scientists to check the condition of vines, carry out necessary practices, monitor nutrients, and reduce pests and diseases. Furthermore, tracking and collecting phenotypical data of vines during the major stages is useful to estimate a vineyard's yield (Goldammer, 2015).

This research project focuses on the last key phenological stage, "Harvest" (E-L 38), which is characterized by ripened grapes having reached their potential berry size and adopted their variety's typical berry colour (Westover, 2018). Although the grapes are already ripe, the precise date of the harvest depends on their intended use and the vine-grower's desired ripeness parameters (Hellman, 2003). The aim of the paper is to model harvest-ripe grapes in 3D and to derive physical and morphological parameters, such as volume (which correlates to berry size), from these models.

2.3 Yield Prediction

Yield is another vital topic in viticulture. It is measured in tons of grapes over a vine block or other spatial unit. In specific applications, yield is also referred to as the amount of fruit on a single vine (Moyer and Komm, 2015).

Information on a vineyard's current anticipated yield enables winegrowers to check whether the yearly yield and quality goals are likely to be achieved. Yield prediction is necessary to assess when to utilize specific vineyard practices or to determine when grapes will be ready for harvest (Nuske et al., 2011). Data about physical and morphological features of grapes, such as their weight or volume, are valuable for vine experts as they help to predict yield (Hemming, 2016). Yield-prediction approaches include the in-season cluster counting method and dormant winter bud dissection (Moyer and Komm, 2015). To obtain reliable yield prediction results, yield prediction must be carried out throughout several blocks of a vineyard and throughout the seasons (Dunn, 2010). Traditionally, samples used for yield prediction are collected manually by winegrowers in the field (Nuske et al., 2014). A feature that is common to these traditional non-sensor-based approaches is their reliance on direct visual and manual in-field measurements. These traditional methods are thus very labour-intensive and allow the collection of only a limited number of samples.

The application of sensor-based technologies, such as photogrammetric approaches in combination with UAS, overcome some of these challenges as they provide a high adaptability to geographical extent and can produce highly detailed point clouds covering an entire vineyard (Rose et al., 2016). This preliminary photogrammetric study focuses on digital data and image collection associated with yield prediction. It looks at the possibility of reducing time-intensive and expensive labour, decreasing human error, and preventing destructive sampling. It takes the form of a laboratory experiment for the modelling of harvest-ripe grapes as digital 3D models based on close-range photogrammetry. The approach also allows the derivation of reliable physical and morphological parameters such as grape weight and volume.

3 Data Acquisition and Data-Modelling Methodology

The workflow of the laboratory experiment for the derivation of morphological parameters of harvest-ripe grapes is illustrated in Figure 1. The first step is to assemble the necessary equipment for data-capture. The subsequent data-capture procedure involves close-range photogrammetry to record multi-view image datasets of grapes and the manual measurement of grape features such as length, width and weight of individual grapes, which are utilized for data validation. The image datasets acquired are utilized in a multi-view 3D photogrammetric analysis to generate high-density 3D grape models. Next, the RANSAC shape-detection tool is used to derive the structure of individual grapes and parameters, such as berry diameter, from the 3D grape models. Characteristics such as the length, width, volume and weight of the digital grape clusters and of individual digital grapes are determined based on the parameters derived from the 3D model. The final step is a validation process, which involves comparing the calculated morphological parameters of the digital grapes and the manually measured parameters.

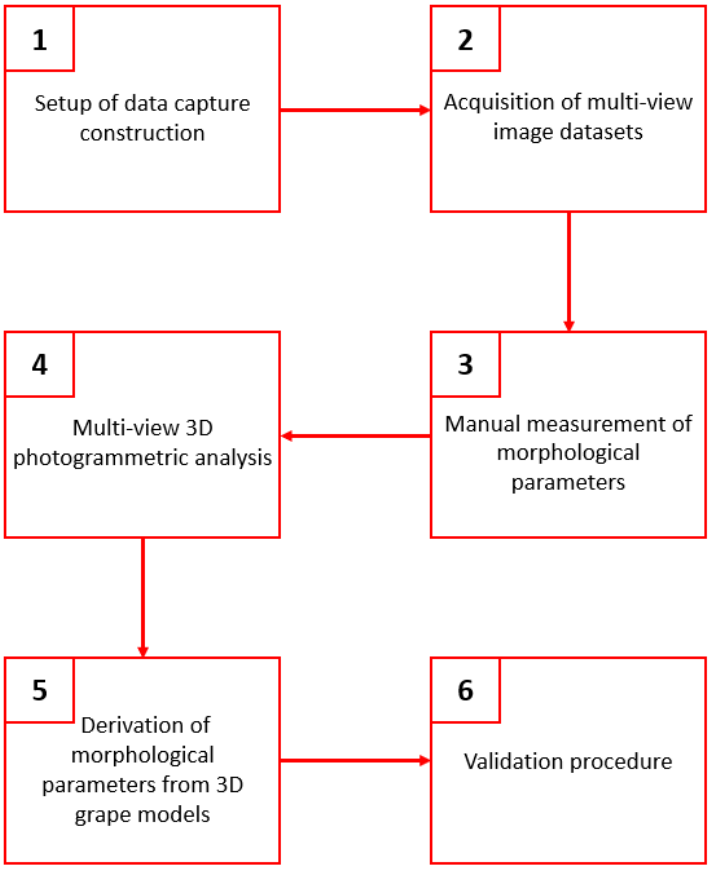

Figure 1: The laboratory experiment workflow

3.1 Laboratory Experiment Setup and Data Acquisition Process

Setting up the laboratory experiment requires gathering and installing the necessary equipment for the data capture. The equipment and objects used in the data acquisition process are:

1. Camera
    - A Sony Alpha 7r3 (Sony, 2018) was used. Relevant technical details of the camera are:
        i. 42.4megapixels
        ii. Sensor size of 35.9 times 24.0mm
        iii. Pixel dimension of 7953 times 5304pixel
2. Tripod
3. Flashers
4. Metal bars
5. Strings

6. Metal plate with holes
7. Weighing scale
8. Calliper ruler
9. Table to record measurements
10. Commercially available green and red grapes (which represent grapes from vines at the harvest stage).

The camera was used with a wide-angle lens and mounted on a tripod. The focal length of the camera was set to 90mm and the aperture to 22 for image capturing. These settings proved to provide the best results. No camera calibration was applied as the calibration is automatically estimated in the 3D modelling software.

The data-capturing construction is a framework consisting of three metal bars (see Figure 2, which illustrates the final version of the construction). The grapes are attached with strings, tape and a metal plate to the middle of the horizontal bar in such a way that they can be rotated while the tripod with the camera remains fixed. The tripod and camera are placed 1 metre away, in front of the grapes. No position marks are used in this setup, as only the relative position of the camera to the grapes is of interest, and there is a very high overlap between the images.

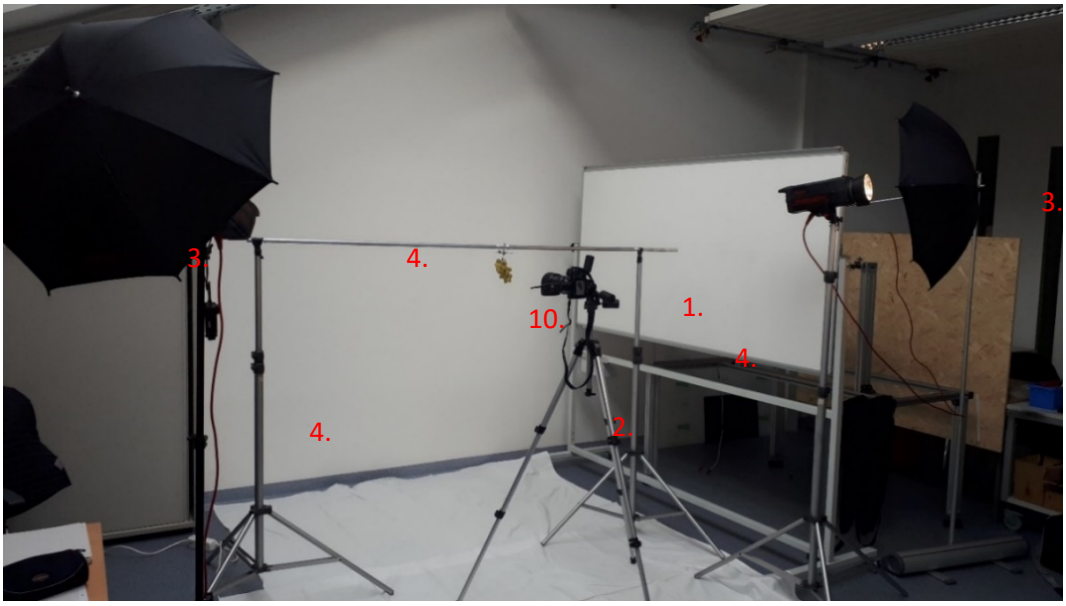

Figure 2: Data capture construction: 1. Camera, 2. Tripod, 3. Flashers, 4. Bars, 10. Grapes

The processing steps of the data-capturing procedure are shown in Figure 3. The procedure consists of two principal parts:

1. The image-capturing process applies close-range photogrammetry to generate high-resolution images of the grapes from multiple points of view. During this process, after an image is created, the grapes are rotated by a few degrees, which changes the

relative viewing angle of the camera on the grapes without moving the tripod. This ensures that images of the grapes from different points of view are recorded. The multi-view image dataset thus generated is used in the data-modelling process to create the 3D grape models.

2. The manual measurement is carried out using a calliper ruler, a weighing scale, and a table to record the measurements. First, the total weight of the grape cluster is measured and noted. Then, each grape is carefully removed from the cluster, and its length, width and weight are recorded. Figure 4 shows how the length and width of the grapes are measured. Finally, the grapes are counted and the number of grapes from the individual grape clusters are recorded in a table. Figure 5 shows the various elements required for the manual measurement of a grape cluster.

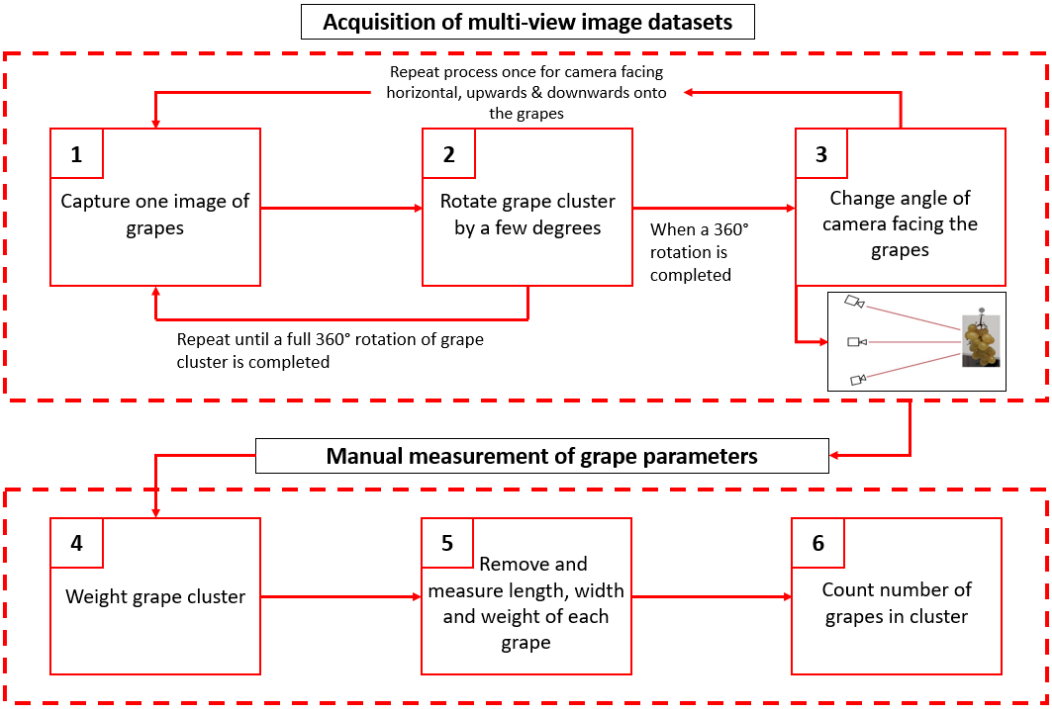

Figure 3: Workflow of the data capture process

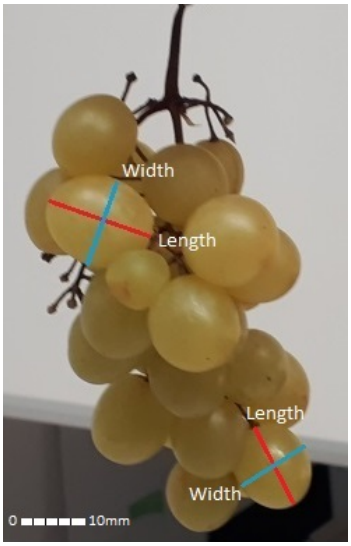

Figure 4: Measuring length and width of grapes

Figure 5: Material for manual measurement of grapes

3.2 Multi-View 3D Photogrammetric Analysis

The next phase of the study deals with a multi-view 3D photogrammetric analysis of the captured grape images. This multi-step procedure produces high-density photogrammetric 3D point clouds that are used to generate 3D grape models. The program Agisoft Photocan Professional© 1.4.3 (build 6529) is used (Agisoft, 2019). Figure 6 illustrates the workflow for the multi-view 3D photogrammetric analysis.

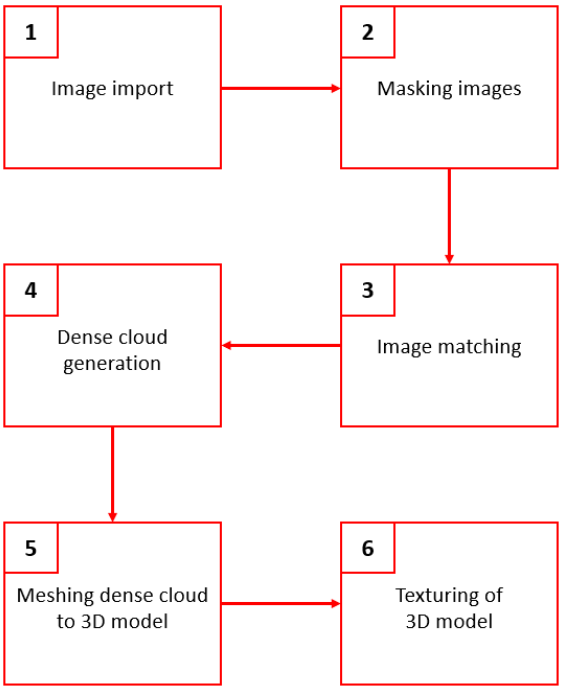

Figure 6: Workflow of the multi-view 3D photogrammetric analysis

First, the high-resolution grape images are loaded into the program. Then, using the "magic wand" tool within Agisoft Photoscan, masks are created for each image to hide such things as the bars, unnatural distortions, or white pixels representing the wall. The masking results in the 3D grape models being more precise because the modelling will only use unmasked parts of images for the calculation of the point clouds and the 3D models. The data-modelling process was tested with unmasked images, and although the resulting 3D models showed generally good results, some grapes showed 3D distortions and unnatural white strips.

Next, the images are aligned in an image-matching process. Image-matching is a prerequisite for 3D modelling of structures as it extracts 3D information from multiple overlapping images. The 3D information thus acquired can be applied to construct 3D models of the surveyed object or scene (Zhang, Xiong and Hao, 2011). The image-matching procedure uses extracted 3D information to align the images and to compute the camera positions, image orientation, and a sparse point cloud. Next, a dense point cloud is generated based on the 3D information contained in the sparse point cloud, the matched images, and the camera parameters that have been calculated. By executing a meshing algorithm, the dense clouds are meshed to 3D grape models.

Finally, a texturing process adds textures from the photographs to the corresponding parts in the previously generated 3D grape models, resulting in the final 3D models of the grapes (Figure 7, green grapes; Figure 9, red grapes). To illustrate the quality of the modelled grapes, photos of the actual grapes are also shown in Figures 8 and 10.

Figure 7: 3D model of the green grapes

Figure 8: Photo of the actual green grapes

Figure 9: 3D model of the red grapes

Figure 10: Photo of the actual red grapes

3.3 Derivation of Morphological Parameters from 3D Grape Models

The next step of the methodology is to identify the shapes of the individual grapes in the 3D models and to derive the physical and morphological parameters of single grapes and of grape clusters. Data such as their weight and volume are important for winegrowers and wineries as they can be utilized to predict yield (Hemming, 2016).

In order to identify the shapes of individual grapes, the RANSAC shape detection plugin (Schnabel, Wahl and Klein, 2007) of the program CloudCompare (CC) version v2.10-alpha [64-bit] (CloudCompare, 2019) is deployed to fit spherical shapes into detected grapes in the 3D models. If a sphere-like shape is found in a 3D model, the corresponding part of the model is separated, colour-coded, and a sphere is placed at that location. The grapes detected in the 3D models are shown in Figures 11 (green grape model) and 12 (red grape model). The grapes in these figures are shown in different hues to allow easier visual distinction between single grapes.

The CC plugin also calculates parameters of these shapes, such as the radius. The radius (r) of an identified digital grape is, by definition, half the grape's width. Thus the width (b) of a 3D grape is derived from its radius. Next, the length (a) of each modelled grape is determined using its width and the length–width ratio (c) of the actual grape. The volume of the digital grapes is calculated using the volume formula for oblate ellipsoids (Eq. 1):

Eq. 1: Ellipsoid volume formula

$$V = \frac{4}{3} * \pi * \left(\frac{b}{2}\right)^2 * \left(\frac{a}{2}\right) = \frac{4}{3} * \pi * r^2 * (r * c)$$

Variable "V" is the volume, and the variables "a" and "b" represent the length and width, respectively. We assume that $\frac{b}{2} = r$ and $\frac{a}{2} = r * c$ where r is the radius calculated by the tool, b is the grape width (i.e. the small semi-axis of the ellipsoid), a is the large semi-axis or the grape length, and c is the mean ratio of length and width measured manually on actual grapes.

In the next step, the mean length and width, length variance and width variance, mean volume and volume variance of the grapes, and total volume of the 3D grape clusters are determined. To validate these derived morphological parameters of the digital grapes, the same parameters are calculated for the actual grapes using the manual measurements. Since the weight of the actual grapes was also measured in the data acquisition process, the density of these grapes can be calculated using the formula (Eq. 2):

Eq. 2: Density formula

$$\varrho = \frac{W}{V}$$

where "ϱ" (Rho) is the density, "W" is the weight, and "V" is the volume. The mean density of the green and red grapes is determined. Now, the weight, mean weight and weight variance of the modelled grapes are calculated using the mean density values of the actual grapes. The results of these calculations are given in Table 1. The values in brackets in Table 1 are the total volume and weight of the digital red grapes, as if all 62 grapes had been detected. For the

missing grapes, the mean volume and weight values of the other 3D modelled red grapes were used as proxies.

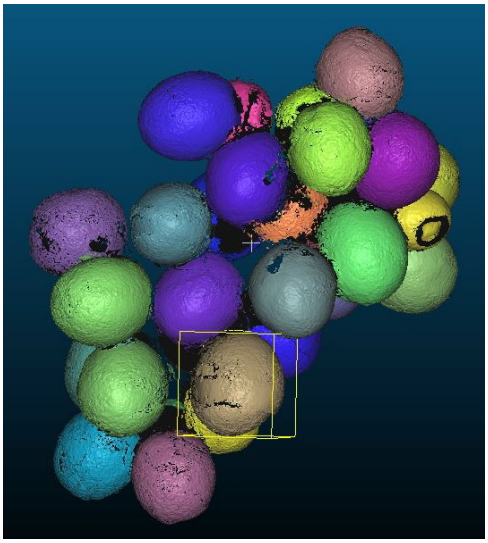

Figure 11: Shapes of the detected green grapes

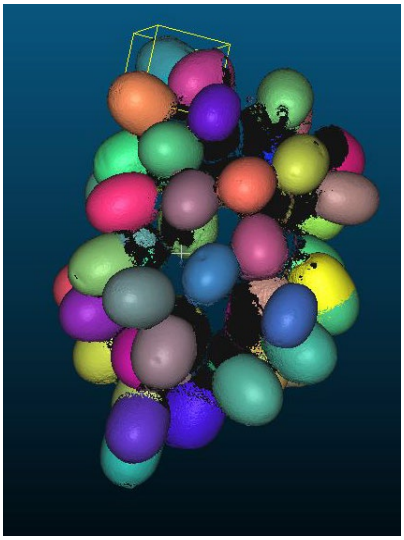

Figure 12: Shapes of the detected red grapes

Table 1: Calculated morphological parameters of actual and digitally derived green and red grapes. * Values in brackets are the total values, as if all actual grapes had been detected

Parameter [unit]	Green grapes		Red grapes	
	Actual grapes	Digital grapes	Actual grapes	Digital grapes
Number of grapes [-]	29	29	62	55
Mean Length [mm]	21.18	21.18	24.11	24.11
Length Variance [mm]	1.46	1.07	3.81	1.22
Mean Width [mm]	18.75	18.75	18.41	18.41
Width Variance [mm]	1.04	0.94	2.06	0.93
Total Volume [mm³]	115,444.02	115,005.25	269,494.7	239,288.22 (269,743.08)*
Mean Volume [mm³]	3,980.83	3,981.05	4,346.69	4,350.69
Volume Variance [mm³]	639.16	556.33	1,565.36	684.93
Total Weight [g]	138.00	137.48	346.00	307.22 (346.32)*
Mean Weight [g]	4.76	4.74	5.58	5.59
Weight Variance [g]	0.76	0.67	2.01	0.88
Mean Density [g/mm³]	0.001195	-	0.001284	-

4 Results and Discussion

The laboratory experiment was performed three times. The setup and methodology described in this paper were followed in experiments 2 and 3 (the setup for experiment 1 was smaller), and the quantification was carried out only in the third experiment. In total, 6 grape clusters, one green and one red grape cluster per experiment, were examined. The final 3D grape models are illustrated in Figure 7 and Figure 9, and the quantification results are shown in Table 1.

The results of the approach demonstrate the reconstruction of 3D grape models under laboratory conditions. The visual comparison of the models shown in Figure 7 and Figure 9 with the actual grapes in Figure 8 and Figure 10 shows that the modelled digital grapes are very similar to their actual counterparts. However, the comparison also reveals that some digital grapes were not correctly reconstructed, which can be seen in certain deformations in the digital grapes. It can be concluded that the process used in this laboratory experiment worked well for detecting grapes: the results of the shape-detection illustrated in Figure 11, Figure 12, and row 2 "Number of grapes [-]" in Table 1 show that all 29 green grapes were detected. In addition, 55 of the 62 red grapes were correctly identified, an accuracy of 88%. However, the shape-detection procedure shows difficulties in identifying strongly elliptical grapes, such as the 7 undetected red grapes. Furthermore, the methodology is reliable for the derivation of physical and morphological parameters from 3D grape models. The calculated total weight, displayed in row "Total Weight [g]" in Table 1, of the digital green grape cluster (137.48g) differs by just 0.52g (0.4%) from the weight of the actual green grape cluster (138.0g). The estimated total weight of the digital red grape cluster (346.32g) differs by just 0.32g from its actual counterpart (346.0g).

The results of this work can be compared with the measurements of Coetzee and Lombard (2013), who examined 300 berries from 37 grape clusters to determine specific grape parameters. They calculated an average grape cluster weight of 101.3g, which differs significantly from the estimated total grape cluster weights presented in this paper (138.0g for the digital green grape cluster; 346.0g for the digital red grape cluster). The average grape densities (1,130kg/m³ in Coetzee and Lombard (2013); 1,195kg/m³ (green grapes) and 1284kg/m³ (red grapes) in this study) are also significantly different. These disparities can be attributed to the use of differently sized grapes in the two studies. Coetzee and Lombard (2013) calculated an average grape length of 14.6mm and grape width of 12.5mm, whereas the mean grape lengths in this work were 21.18mm (green grapes) and 24.11mm (red grapes), and the mean grape widths were 18.75mm (green grapes) and 18.41mm (red grapes).

Some of our results were inconclusive. The calculated variance values differ greatly between the various parameters measured for the digital and the actual grapes. An example is the large difference in length variance, which is 0.39mm for the digital green grapes and the actual grapes, and 2.59mm for the red grapes. The large variance between the digital and the actual grapes is thought to be caused by the shape detection method we applied, which utilized spheres to approximate the grape shapes. We also experimented with using cylinders to detect grapes, but this proved inadequate as it resulted in a model of just 10 grapes.

Several problems and limitations were identified. The approach showed difficulties with the accurate reconstruction of misshapen or strongly elliptical grapes. Examples are highlighted by red circles in Figure 13 and Figure 14. These grapes could be correctly detected with the use of LiDAR-based systems. Another technology which can be applied to analyse the shape of any grape is described in American Society for Horticultural Science (2009). An advantage of the SigmaScan®-based approach described there over the method presented in this paper is the grapes' independency from geometry and spatial position through the measurement of grapes based on difference in colour between the surveyed grape and the background. Additional errors such as holes in parts of stems were found because adequate point density cannot always be achieved for the modelling of thin structures. Another limitation is the immobility of the data-capture setup. It took several attempts to find an appropriate procedure and settings. Furthermore, a new construction would need to be built if an entire row of vines was to be simulated and modelled.

The results of our preliminary laboratory experiment indicate that close-range photogrammetry can be applied to generate 3D grape models and that parameters such as the volume of the grape can be derived from these digital models.

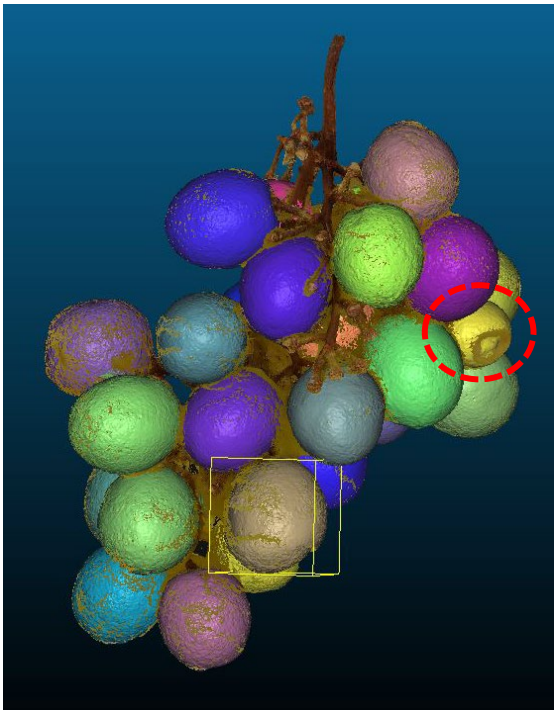

Figure 13: Difficulties detecting misshapen grapes

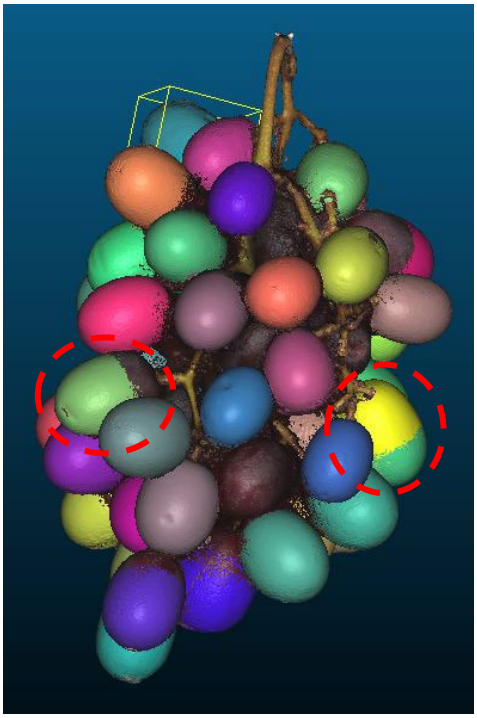

Figure 14: Incorrectly identified elliptical grapes

5 Conclusions and Future Work

In this paper we conceptualized and modelled 3D grapes and used these models to derive physical and morphological parameters. In the laboratory experiment, a special data-capture setup was constructed to record multi-view high-resolution image datasets of commercially available green and red grapes, representing grapes at harvest stage (E-L 38). The data-modelling process focused on the generation of high-resolution 3D grape models, the derivation of parameters from them, and the comparison of the parameters for 3D grape models and actual grapes. The approach shows that close-range photogrammetry can be applied in a lab to create high-resolution 3D grape models and to derive reliable physical and morphological parameters.

In a future project, a similar approach could be applied in-field by, for example, creating a system of multiple cameras mounted on an unmanned autonomously moving vehicle, such as a UAS, which generates close-range multi-view image datasets. These datasets could then be utilized to generate 3D models of grapes on the vines, which in turn could be used to derive parameters such as grape volume to estimate the yield of the grapes.

Acknowledgments

The authors would like to thank the three anonymous reviewers and editor Mary Rigby for their valuable input into the content of this paper, which certainly improved the quality of it.

References

Agisoft. (2019). Professional Edition. Retrieved from https://www.agisoft.com/features/professional-edition/

American Society for Horticultural Science. (2009). Grape shapes. Retrieved from https://www.eurekalert.org/pub_releases/2009-02/asfh-gs021709.php

Centinari, M. (2018). Grapevine Bud Break 101 [Blog post]. Retrieved from https://psuwineandgrapes.wordpress.com/2018/05/14/grapevine-bud-break-101/

CloudCompare (version 2.10-alpha) [GPL software]. (2019). CloudCompare - Open Source project. Retrieved from http://www.cloudcompare.org/

Coetzee, C. and Lombard, S. (2013). The destemming of grapes: Experiments and discrete element modelling. *Biosystems Engineering.* 114(3), 232-248.

Dunn, G. M. (2010). *Yield Forecasting* [Fact sheet]. Retrieved from https://www.wineaustralia.com/getmedia/5304c16d-23b3-4a6f-ad53-b3d4419cc979/201006_Yield-Forecasting.pdf

Goldammer, T. (2015). *The grape grower's handbook*. 2nd edition. USA: Apex Publishers.

Hellman, E. W. (2003). Grapevine Structure and Function. Retrieved from https://www.growables.org/information/LowChillFruit/documents/GrapeExtOrg.pdf

Hemming, R. (2016). Wine by numbers: viticulture, part one. Retrieved from https://www.jancisrobinson.com/articles/wine-by-numbers-part-one

Maniak, S. (2004). *Datenaustausch in geographischen Informationssystemen.* [Data exchange in geographic information systems]. Düren, Germany: Shaker, 5-14.

Moyer, M. M. & Komm, B. (2015). *Vineyard Yield Estimation.* Washington State University Extension. Retrieved from https://www.vineyardteam.org/files/resources/Vineyard%20Yield%20Estimation-%20WSU.pdf

Nuske, S., Achar, S., Bates, T., Narasimhan, S. G. & Singh, S. (2011). Yield estimation in vineyards by visual grape detection. *IEEE International Conference on Intelligent Robots and Systems.* 2352-2358. doi:10.1109/IROS.2011.6095069.

Nuske, S., Wilshusen, K., Achar, S., Yoder, L., Narasimhan, S. & Singh, S. (2014). Automated Visual Yield Estimation in Vineyards. *Journal of Field Robotics, 31*(5), 837-860.

Poupin, M. J., Matus, J. T., Leiva-Ampuero, A. & Arce-Johnson, P. (2011). *The Flowering Process and its Control in Plants: Gene Expression and Hormone Interaction.* 1st edition. Kerala, India: Research Signpost, 173-197.

Rose, J., Kicherer, A., Wieland, M., Klingbeil, L., Töpfer, R. & Kuhlmann, H. (2016). Towards Automated Large-Scale 3D Phenotyping of Vineyards under Field Conditions. *Sensors, 16*(12), 1-7.

Schnabel, R., Wahl, R. and Klein, R. (2007). Efficient RANSAC for Point-Cloud Shape Detection. *Computer Graphics Forum*, *26*(2), 214-226.

Sony (2018). Sony α7R III 35 mm full-frame camera with autofocus. Retrieved from https://www.sony.co.in/electronics/interchangeable-lens-cameras/ilce-7rm3/specifications

Westover, F. (2018). Grapevine Phenology Revisited. Retrieved from https://www.winesandvines.com/features/article/196082/Grapevine-Phenology-Revisited

Wine Institute (2019). World Wine Production by Country. Retrieved from https://personalpages.manchester.ac.uk/staff/fumie.costen/pastwork/grapes/XiaoWenyu.pdf

Zhang, Y., Xiong, J., & Hao, L. (2011). Photogrammetric processing of low-altitude images acquired by unpiloted aerial vehicles. *Photogrammetric Record. 26*(134), 190-211.

GI_Forum 2020, Issue 1
Page: 89 - 98
Short Paper
Corresponding Author:
saskia.wolff@geo.hu-berlin.de
DOI: 10.1553/giscience2020_01_s89

Characterising Agricultural Landscapes using Landscape Metrics and Cluster Analysis in Brandenburg, Germany

Saskia Wolff and Tobia Lakes
Humboldt-Universität zu Berlin, Germany

Abstract

An increasing demand for agricultural products within the past years has led to increasing agricultural intensification. Various agricultural compositions and landscape configurations can have different impacts on the provision of ecosystem services. The EU follows the aim of supporting and developing sustainable food production systems. We use the plot-based data provided by the Integrated Administration and Control System (IACS) to identify different types of agricultural landscapes and their spatial distribution in Brandenburg, Germany. By calculating a set of landscape metrics to characterise agricultural land use, we were able to identify six types of agricultural landscapes by a Two-Step cluster analysis for a hexagonal grid. Thereby, the majority of Brandenburg is covered by agriculture characterised by high share of cropland but different degrees of fragmentation. By providing a framework using landscape metrics derived from IACS data, the approach of clustering to identify typologies is highly transferable to other regions within the EU and may provide an important asset for offering new units of analysis for a better tailored environmental and agricultural planning depending on the local to regional characteristics.

Keywords:

agricultural land use pattern, agricultural intensification, landscape metrics, cluster analysis

1 Introduction

European agricultural landscapes have featured considerable changes towards intensification and marginalization of areas, and these major trends are expected to continue in the future (Lüker-Jans, Simmering, & Otte, 2016; Rounsevell, Annetts, Audsley, Mayr, & Reginster, 2003). We define agricultural landscapes as the result of land uses and management in an area following the definition of Kizos and Koulouri (2006). These landscapes provide ecological functions, e.g. habitat provision; economic functions, e.g. income generation; and cultural functions, e.g. landscape aesthetics. According to Lüker-Jans et al. (2016), marginal agricultural landscapes are characterised by unfavourable biophysical conditions, such as steep slopes, shallow and/or poor soils, and inferior accessibility. They often show increased biodiversity and habitat richness due to low intensities of cultivation, crop and grassland rotation and small-parcelled mosaics. Conversely, intensive agriculture often goes along with

larger field sizes, lower heterogeneity in habitat structure, and more monoculture (Ruiz-Martinez, Marraccini, Debolini, & Bonari, 2016). Thus, intensification is frequently associated with a decrease in biodiversity and negative effects on the environment, i.e. soils and water quality (Thomson et al., 2019). A sustainable pathway is needed for maximising agricultural production and particularly achieving future food security while at the same time reducing the negative environmental effects of agricultural land use. In recent years, the provision of ecosystem services from agricultural land has been increasingly highlighted by science and enacted in policy changes (Schaller et al., 2018). The European Common Agricultural Policy (CAP), the major policy instrument driving agricultural land use in Europe, aims to support the sustainable management of natural resources such as water, soil and air and to contribute to the protection of biodiversity, enhance ecosystem services and preserve habitats and landscapes (European Union, 2019).

In the past decades, landscape metrics have been successfully applied to characterise and compare (agricultural) landscapes across space and time in a quantitative manner (Uuemaa, Mander, & Marja, 2013). Typically, number, size, shape and arrangement of patches of different land-use/land cover types are used quantify landscape structure, composition and dynamics. Lately, metrics have also been used as proxies for characterising agricultural land use intensity, e.g. area under cultivation, mean patch size and Shannon's Diversity Index (Schlesinger and Drescher 2018). In contrast, others have analysed inputs, such as labour, capital or management practices, and outputs, such as yields (Shriar, 2000) or the dependence on industrial goods, e.g. machinery and fertilizer (Temme & Verburg, 2011; Zasada et al., 2013) to characterise agricultural land use intensity. However, these studies face the problem of data availability and are therefore often restricted to small areas and selected farms. A promising dataset to achieve area-wide characterization by different types of agricultural landscapes comes from the Integrated Administration and Control System (IACS, in German: Invekos). In recent years, initial studies successfully used this dataset that is derived from the subsidy-payments to the farmers to analyse agricultural land use change (Lüker-Jans et al., 2016; Tomlinson, Dragosits, Levy, Thomson, & Moxley, 2018) and farm-level agriculture characterization (Lomba et al., 2017; Uthes, Kelly, & König, 2020).

The aim of this paper is to identify and characterise different types of agricultural landscapes and to depict their spatial patterns using landscape metrics and a cluster analysis for the case study of Brandenburg, Germany. While landscape metrics are most frequently applied to grids and administrative areas, we use hexagons. They have shown to better capture spatially continuous phenomena such as agricultural landscapes because of their spatial smoothing effect towards the edges of the hexagons (Birch, Oom, & Beecham, 2007; Schindler, Poirazidis, & Wrbka, 2008). The outcomes of this study may provide an important asset for providing new units of analysis for better-tailored environmental and agricultural policies depending on the local to regional characteristics.

2 Material and Methods

2.1 Study Area

We focus on the state of Brandenburg, which is located in the northeast of Germany covering 29.640 km² of which 45% are used for agriculture (Figure 1). Ongoing pressure on agricultural land to convert into residential land is observed in the suburban areas of Berlin, while an increasing demand for regional food production can also be observed. At the same time, Gutzler et al. (2015) anticipate an increased use of cropland for renewable energy production. Farms in Brandenburg are comparatively large, around 240 ha, four times the German average (Gutzler et al., 2015). In addition, general low soil quality with almost two-thirds being sandy and sandy-loamy soils, low rainfall at only 591 mm/year and a high technological level characterise the agricultural land use. Compared to other German states, Brandenburg shows a relatively high share of organic agriculture (12 % of agricultural area) that is further increasing in recent years (Ministerium für Landwirtschaft, Umwelt und Klimaschutz [MLUK], 2019).

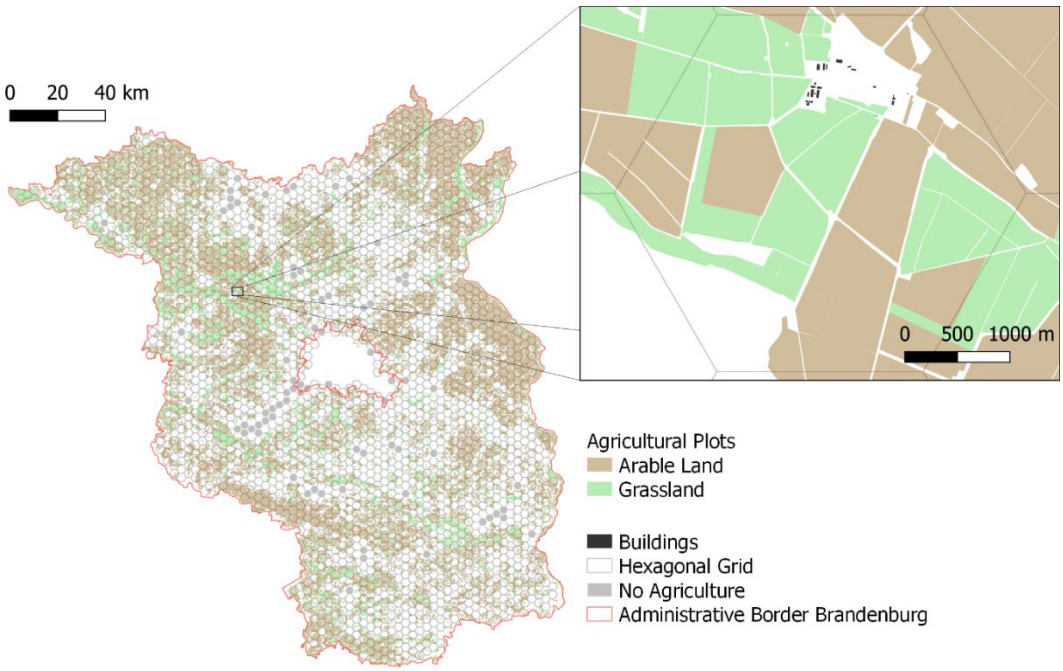

Figure 1: Agricultural land use and hexagons grid outline of the state of Brandenburg, Germany.

2.2 Data

We used plot-based information on cultivation for Brandenburg agriculture in 2018 (reported for 31.5.2018) provided by the Integrated Administration and Control System (IACS). We selected and reclassified the data into the categories: grassland, cropland, and maize as a single crop. We also derive the plot sizes and edges and if a plot is managed conventionally or

organically. In addition, we use Open Street Map (OSM) data on buildings and soil quality data that captures the yield potential (Bundesanstalt für Geowissenschaften und Rohstoffe, 2014).

2.3 Methods

We created a hexagonal grid with a cell size of 10 km² (N = 2836; 178 were deleted because of missing data). The size of the cells captured the landscape level and the spatial configuration of plots within. We selected the following indicators to assess different types of agricultural landscapes based on a literature review: soil quality (values from 0-100), number of buildings (N), edge density (calculated as share of total hexagon area, in km/10km²), median plot size (ha), organic share of total agricultural area (%), maize share of total agricultural (%), cropland share (%), Shannon Diversity Index, share agriculture of total area (%) and mean distance to settlements (km). We measured cropland intensity by the share of maize that is likely to be used for biogas and cultivated as a long-term self-following crop (i.e. without crop rotation; (Gutzler et al., 2015; Lüker-Jans et al., 2016). We included both maize types (i.e. silage maize and corn maize) in our analysis. According to the Fachagentur Nachwachsende Rohstoffe e. V. (2013), the expansion of maize is expected to be on par with intensification of crop production. We calculated the respective indicator values for the year of 2018 for the hexagons. To reduce redundancies in the datasets we calculated Spearman's correlation coefficients (Lausch & Herzog, 2002) and dropped those indicators with a correlation of 0,4 or more, i.e. share of agriculture, Shannon's Diversity Index, distance to settlements. We then applied a cluster analysis on the remaining 7 indicators: number of buildings, soil quality, median plot size, edge density and share of cropland, maize and organic agriculture. The Two-Step clustering offers the advantage automatic determination of the optimum number of clusters and was originally developed for large datasets by Chiu, Fang, Chen, Wang, and Jeris (2001). For validation of the cluster number, the model fit was evaluated by the silhouette coefficient, which is a measure of cohesion and separation of clusters. A value above 0,2 thereby indicated a fair quality of clusters (Tkaczynski, 2017).

To measure spatial autocorrelation for the categorical cluster values, we calculated the join count (Plant, 2012). This determines the degree of clustering or dispersion among a set of spatially adjacent polygons. To calculate the join count for each cluster value, we set the reference cluster value to 1 and all other cluster values to 0, and we calculated the join count separately for each cluster.

3 Results

We identified 6 different types of agricultural landscapes in Brandenburg: 1 Peri-urban, 2 High Fragmentation, 3 Low Fragmentation, 4 High Intensity, 5 Low Intensity (marginal grasslands), 6 Organic Production (see Table 1). The Two-Step clustering for these 6 clusters returned the best results with relatively low Bayesian Information Criterion (BIC) values (7894,076) and distance measure is the highest (1,546). A The silhouette measure of cluster cohesion and separation indicates a fair quality (0,3) for these 6 clusters.

Table 1: Centroid of clusters with indication of lowest (green) and highest (red) values

	Centroid						
Cluster	Soil Quality	Number of Buildings	Edge Density (km/10km²)	Median Plot Size (ha)	Organic Share (%)	Maize Share (%)	Cropland Share (%)
1: Peri-urban	49,4	3206,2	5,0	3,0	7,6	10,1	68,9
2: High Fragmentation	49,4	194,7	10,4	4,4	5,1	18,4	83,7
3: Low Fragmentation	51,3	197,4	4,1	3,5	5,3	19,3	86,7
4: High Intensity	62,8	173,9	7,9	11,2	3,2	20,5	93,7
5: Low Intensity	47,2	207,8	8,3	4,5	12,9	7,2	35,7
6: Organic Production	50,4	244,6	6,3	3,2	68,9	4,8	72,1

More specifically, the identified types of agricultural landscapes can be characterised as the following:

Cluster 1 (**Peri-Urban**: 5,8 % of all clusters, N = 149) can be described as the peri-urban agriculture cluster mainly characterised by very high mean numbers of 3206 buildings (Table 1). Hence, mean share of agricultural area is lowest amongst the clusters with a calculated average of 24,5 %. Consequently, edge density is also relatively low (mean 5,0 km/10 km²). With the lowest average median plot size (3,0 ha) plots in this cluster tend to be smaller than plots in other clusters. Share of maize and cropland in general tend to be lower than in the other clusters. Additionally, the areas of this cluster are characterised by lower soil quality (49,4) in terms of yield potential.

Cluster 2 (**High Fragmentation**: 36,1 % of all clusters, N = 933) characteristics are that of high fragmentation and high mean of agriculture share (66,0 %). Cropland share in general and particularly share of maize is relatively high.

On the contrary, **Cluster 3** (**Low Fragmentation**: 22,4 % of all clusters, N = 579) is characterized by low fragmentation of the agricultural landscape explained by a low mean agriculture share of 25,5 %. Furthermore, it shows a high share of cropland, relatively high soil quality and low edge density. The landscape is generally not characterised by agriculture, but other land covers such as water or forest.

Cluster 4 (**High Intensity**: 8,9 % of all clusters, N = 229) shows the highest mean agriculture share (66,3%) as well as high quality soil (62,8). It is characterised by large plot sizes (11,2 ha) with large share of cropland (93,7 %) and maize (20,5 %).

Cluster 5 (**Low Intensity**: 15,6 % of all clusters, N = 404) mainly represents marginal grasslands with a mean agriculture share of 44,5 %. The low soil quality (47,2) leads to plots

mainly used for grassland (low share of cropland = 35,7). Compared to other clusters (except 6) grassland is thereby often managed organically (mean organic share = 12,9 %).

Cluster 6 (**Organic Production**: 11,2 % of all clusters, N = 289) represents organic farming. It is characterised by a low share of cropland and maize, smaller median plot sizes (3,2 ha), and a mean agricultural share of 32,5 %.

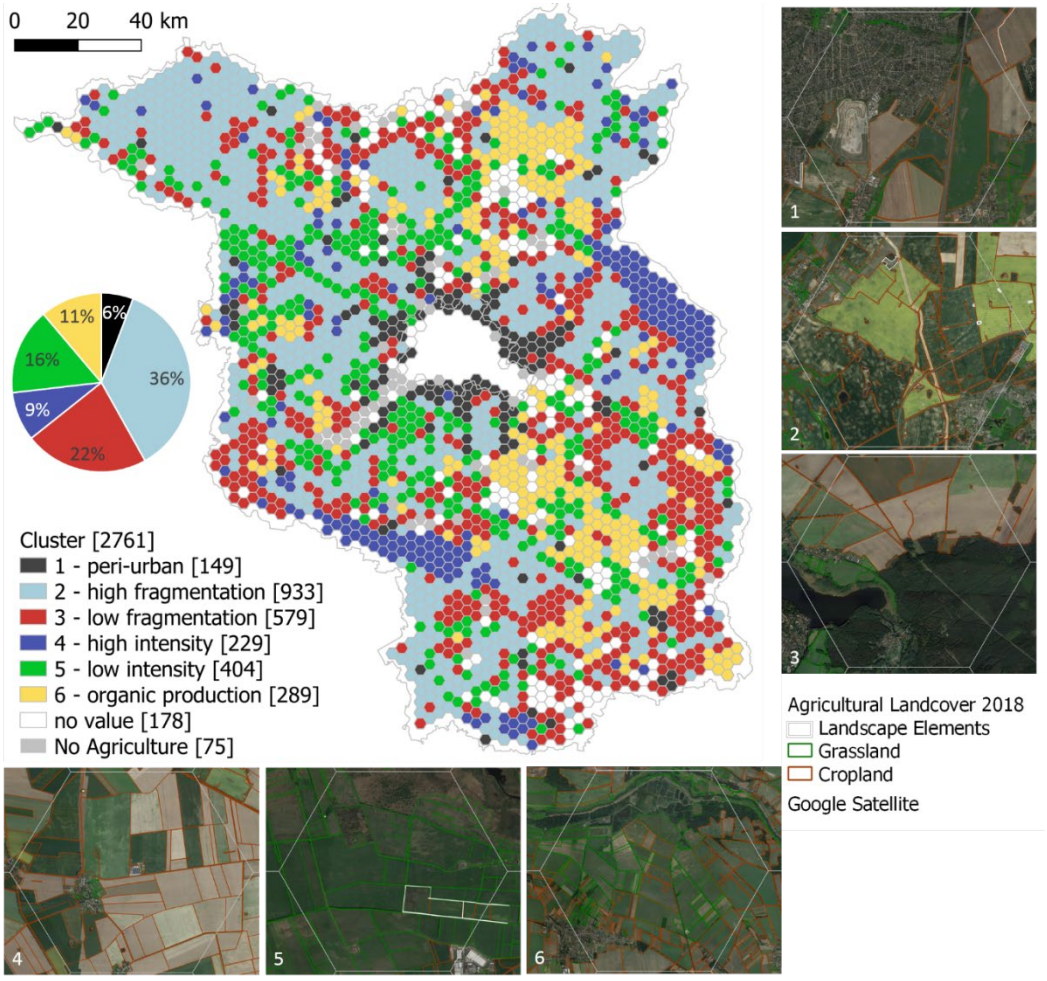

Figure 2: Map and exemplary satellite imagery (Google) of Agricultural Land Use clusters in Brandenburg 2018

We identified a high positive spatial autocorrelation for the 'high intensity' (N = 98) and 'organic production' (N = 95) clusters. This means that one agricultural landscape type is located next to another agricultural landscape of the same type. The spatial clustering of 'high intensity' agriculture that we find in our results may be attributed to the underlying spatial clustering of high soil quality. One reason for spatial clustering of 'organic production' might

be that it occurs often in nature preserves under stringent conditions (Venghaus & Acosta, 2018). In contrast to other studies and literature, we could not find significantly higher soil qualities in areas under organic production. Other influencing factors could be operational determinants, for example the share of grassland which is higher in our organic production type than in other clusters (Bichler & Häring, 2003). Another reason the potential agglomeration effect of organic agriculture (Schmidtner et al., 2012). In contrast, the 'low fragmentation' (N = 34) and 'low intensity' (N = 43) clusters do not show a high degree of spatial autocorrelation and are distributed across the state. The 'peri-urban' (N = 54) and 'high fragmentation' (N = 71) clusters show medium spatial autocorrelation and are mostly randomly spatially distributed whereby the peri-urban cells are concentrated around Berlin.

4 Discussion

Our results complement information on agricultural landscapes, such as the agro-ecological zones of Brandenburg (*Landbaugebiete*), that have been given a suitability rating for crop production (*Ackerzahl*; Landesamt für Ländliche Entwicklung, Landwirtschaft und Flurneuordnung, 2016) and the maps available in the Thünen Atlas, including the distribution of crop types or grassland on a municipal scale (Thünen Institut, 2014). Our types thereby also include information on composition, diversity and intensity based on a plot-based analysis instead of representing a single indicator (e.g. soil quality). They can help to understand the agricultural landscape structure in Brandenburg and identify regions where monitoring and specified support measures are necessary.

Typologies of Brandenburg's agriculture have been created mainly through farmer decisions with reference to renewable energy production (Venghaus & Acosta, 2018). Thereby the farmer is the decision-making "designer" of agricultural landscapes whereby we used landscape metrics as input for typologising agriculture. Consistent with Lüker-Jans et al. (2016) using k-means clustering, we identified similar agricultural types focused on cropland share with maize as a particular crop. In contrast to our hexagons providing a smooth surface allowing the unambiguous definition of neighbourhoods for the study area, they analyze metrics on a municipal level which provides higher variance in shape and size than grid-based analysis. In general, landscape metrics prove to be an adequate tool for analysing configuration and composition of landscapes. Similar to Lomba et al. (2017), Uthes et al. (2020) and Lüker-Jans et al. (2016), we were able to show the potential of IACS data for analysing agricultural land use. Other studies have used remote sensing, e.g. to identify patchiness of the agricultural landscape (Weissteiner, García-Feced, & Paracchini, 2016). The analysis on a finer spatial scale could enable the possibility of investigating finer landscape structures and, additionally, changes in e.g. agricultural composition. A common problem in ecological analysis of spatial indicators is scale. Scale dependence can be addressed by sensitivity analysis via up- and downscaling the grid cell size and can be applied in further studies. Oberlack et al. (2019) emphasised that archetypes can help tailor intensification strategies to particular contexts. Additionally, to increase the quality of the "archetypes", Eisenack et al. (2019) proposed a framework to merge quantitative and qualitative approaches. However, this paper focuses on the methodological suitability of landscape metrics as an input for cluster analysis within a

hexagonal grid. One of the advantages of using IACS data is thereby the high possibility of transferability to other study regions.

Conclusions

Our findings reveal six different types of agricultural landscapes and their respective spatial patterns. We conclude that Brandenburg is characterised by highly fragmented agriculture and high spatial clustering of high intensity agriculture and organic production.

The chosen landscape metrics derived from IACS data have proven to be adequate for improving the understanding of agricultural landscapes, and they are suitable for measuring agricultural intensity and diversity in terms of plot composition and configuration at the EU level since IACS data are available across the EU. Our paper proposes an approach at the landscape level which is, according to Thomson et al. (2019), a fundamental connection between the diverse array of relevant disciplines at the plant to field level and can inform national and global decision making. Future work will focus on relations of these different types with land price development, ownership patterns and trade-offs for example between food and energy production.

Acknowledgement

This work was supported by the Deutsche Forschungsgemeinschaft in the Research Unit 2569 "Agricultural Land Markets—Efficiency and Regulation".

References

Bichler, B., & Häring, A. M. (2003). Die räumliche Verteilung des ökologischen Landbaus in Deutschland und ihre Bestimmungsgründe. Retrieved from https://orgprints.org/5046/1/5046-02OE469-uni-hohenheim-2003-raeuml-verteilg.pdf

Birch, C. P.D., Oom, S. P., & Beecham, J. A. (2007). Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological Modelling*, *206*(3-4), 347–359. https://doi.org/10.1016/j.ecolmodel.2007.03.041

Bundesanstalt für Geowissenschaften und Rohstoffe (2014). Ackerbauliches Ertragspotential der Böden in Deutschland. Retrieved from https://www.bgr.bund.de/DE/Themen/Boden/Ressourcenbewertung/Ertragspotential/Ertragspotential_node.html

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). *A robust and scalable clustering algorithm for mixed type attributes in large database environment*. Retrieved from KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining website: http://dl.acm.org/citation.cfm?id=502549

Eisenack, K., Villamayor-Tomas, S., Epstein, G., Kimmich, C., Magliocca, N., Manuel-Navarrete, D., . . . Sietz, D. (2019). Design and quality criteria for archetype analysis. *Ecology and Society*, *24*(3). https://doi.org/10.5751/ES-10855-240306

European Union (2019). THE POST-2020 COMMON AGRICULTURAL POLICY: ENVIRONMENTAL BENEFITS AND SIMPLIFICATION. Retrieved from https://ec.europa.eu/info/sites/info/files/food-farming-fisheries/key_policies/documents/cap-post-2020-environ-benefits-simplification_en.pdf

Fachagentur Nachwachsende Rohstoffe e. V. (2013). *Biogas an introduction*. Retrieved from https://mediathek.fnr.de/media/downloadable/files/samples/b/r/brosch.biogas-2013-en-web-pdf.pdf

Gutzler, C., Helming, K., Balla, D., Dannowski, R., Deumlich, D., Glemnitz, M., . . . Zander, P. (2015). Agricultural land use changes – a scenario-based sustainability impact assessment for Brandenburg, Germany. *Ecological Indicators*, *48*, 505–517. https://doi.org/10.1016/j.ecolind.2014.09.004

Kizos, T., & Koulouri, M. (2006). Agricultural landscape dynamics in the Mediterranean: Lesvos (Greece) case study using evidence from the last three centuries. *Environmental Science & Policy*, *9*(4), 330–342. https://doi.org/10.1016/j.envsci.2006.02.002

Landesamt für Ländliche Entwicklung, Landwirtschaft und Flurneuordnung (2016). Datensammlung für die betriebswirtschaftliche Bewertung landwirtschaftlicher Produktionsverfahren im Land Brandenburg. Retrieved from https://lelf.brandenburg.de/media_fast/4055/Datensammlung%202016_web.pdf

Lausch, A., & Herzog, F. (2002). Applicability of landscape metrics for the monitoring of landscape change: issues of scale, resolution and interpretability. *Ecological Indicators*, *2*(1), 3–15. https://doi.org/10.1016/S1470-160X(02)00053-5

Lomba, A., Strohbach, M., Jerrentrup, J. S., Dauber, J., Klimek, S., & McCracken, D. I. (2017). Making the best of both worlds: Can high-resolution agricultural administrative data support the assessment of High Nature Value farmlands across Europe? *Ecological Indicators*, *72*, 118–130. https://doi.org/10.1016/j.ecolind.2016.08.008

Lüker-Jans, N., Simmering, D., & Otte, A. (2016). Analysing Data of the Integrated Administration and Control System (IACS) to Detect Patterns of Agricultural Land-Use Change at Municipality Level. *Landscape Online*, *48*, 1–24. https://doi.org/10.3097/LO.201648

Ministerium für Landwirtschaft, Umwelt und Klimaschutz (2019). Massnahmeprogramm Oekologische Produktion. Retrieved from https://mluk.brandenburg.de/sixcms/media.php/9/Massnahmeprogramm_Oekologische_Produktion.pdf

Oberlack, C., Sietz, D., Bürgi Bonanomi, E., Bremond, A. de, Dell'Angelo, J., Eisenack, K., . . . Villamayor-Tomas, S. (2019). Archetype analysis in sustainability research: meanings, motivations, and evidence-based policy making. *Ecology and Society*, *24*(2). https://doi.org/10.5751/ES-10747-240226

Rounsevell, M.D.A., Annetts, J.E., Audsley, E., Mayr, T., & Reginster, I. (2003). Modelling the spatial distribution of agricultural land use at the regional scale. *Agriculture, Ecosystems & Environment*, *95*(2-3), 465–479. https://doi.org/10.1016/S0167-8809(02)00217-7

Ruiz-Martinez, I., Marraccini, E., Debolini, M., & Bonari, E. (2016). *Indicators of agricultural intensity and intensification: A review of the literature*. Retrieved from https://hal.archives-ouvertes.fr/hal-01277628/document

Schaller, L., Targetti, S., Villanueva, A. J., Zasada, I., Kantelhardt, J., Arriaza, M., . . . Viaggi, D. (2018). Agricultural landscapes, ecosystem services and regional competitiveness—Assessing drivers and mechanisms in nine European case study areas. *Land Use Policy*, *76*, 735–745. https://doi.org/10.1016/j.landusepol.2018.03.001

Schindler, S., Poirazidis, K., & Wrbka, T. (2008). Towards a core set of landscape metrics for biodiversity assessments: A case study from Dadia National Park, Greece. *Ecological Indicators*, *8*(5), 502–514. https://doi.org/10.1016/j.ecolind.2007.06.001

Schmidtner, E., Lippert, C., Engler, B., Häring, A. M., Aurbacher, J., & Dabbert, S. (2012). Spatial distribution of organic farming in Germany: does neighbourhood matter? *European Review of Agricultural Economics*, *39*(4), 661–683. https://doi.org/10.1093/erae/jbr047

Shriar, A. J. (2000). Agricultural intensity and its measurement in frontier regions. *Agroforestry Systems*, *49*(3), 301–318. https://doi.org/10.1023/A:1006316131781

Temme, A.J.A.M., & Verburg, P. H. [P. H.] (2011). Mapping and modelling of changes in agricultural intensity in Europe. *Agriculture, Ecosystems & Environment*, *140*(1), 46–56. https://doi.org/10.1016/j.agee.2010.11.010

Thomson, A. M. [Allison M.], Ellis, E. C., Grau, H. R., Kuemmerle, T., Meyfroidt, P., Ramankutty, N., & Zeleke, G. (2019). Sustainable intensification in land systems: trade-offs, scales, and contexts. *Current Opinion in Environmental Sustainability*, *38*, 37–43. https://doi.org/10.1016/j.cosust.2019.04.011

Thünen Institut (2014). Der Thünen Agraratlas. Retrieved from https://www.thuenen.de/de/infrastruktur/thuenen-atlas-und-geoinformation/thuenen-atlas/

Tkaczynski, A. (2017). Segmentation Using Two-Step Cluster Analysis. In T. Dietrich, S. Rundle-Thiele, & K. Kubacki (Eds.), *Segmentation in Social Marketing: Process, Methods and Application* (pp. 109–125). Singapore: Springer Singapore; Imprint: Springer. https://doi.org/10.1007/978-981-10-1835-0_8

Tomlinson, S. J., Dragosits, U., Levy, P. E., Thomson, A. M. [Amanda M.], & Moxley, J. (2018). Quantifying gross vs. Net agricultural land use change in Great Britain using the Integrated Administration and Control System. *The Science of the Total Environment*, *628-629*, 1234–1248. https://doi.org/10.1016/j.scitotenv.2018.02.067

Uthes, S., Kelly, E., & König, H. J. (2020). Farm-level indicators for crop and landscape diversity derived from agricultural beneficiaries data. *Ecological Indicators*, *108*, 105725. https://doi.org/10.1016/j.ecolind.2019.105725

Uuemaa, E., Mander, Ü., & Marja, R. (2013). Trends in the use of landscape spatial metrics as landscape indicators: A review. *Ecological Indicators*, *28*, 100–106. https://doi.org/10.1016/j.ecolind.2012.07.018

Venghaus, S., & Acosta, L. (2018). To produce or not to produce: an analysis of bioenergy and crop production decisions based on farmer typologies in Brandenburg, Germany. *Regional Environmental Change*, *18*(2), 521–532. https://doi.org/10.1007/s10113-017-1226-1

Weissteiner, C. J., García-Feced, C., & Paracchini, M. L. (2016). A new view on EU agricultural landscapes: Quantifying patchiness to assess farmland heterogeneity. *Ecological Indicators*, *61*, 317–327. https://doi.org/10.1016/j.ecolind.2015.09.032

Zasada, I., Loibl, W., Berges, R., Steinnocher, K., Köstl, M., Piorr, A., & Werner, A. (2013). Rural–urban Regions: A Spatial Approach to Define Urban–Rural Relationships in Europe. In K. Nilsson (Ed.), *Peri-urban futures: Scenarios and models for land use change in Europe* (pp. 45–68). Heidelberg: Springer. https://doi.org/10.1007/978-3-642-30529-0_3

GI_Forum 2020, Issue 1
Page: 99 - 106
Short Paper
Corresponding Author:
juergen.knies@hs-bremen.de
DOI: 10.1553/giscience2020_01_s99

A Blueprint to Integrate and Exploit the Benefits of E-Government and BIM in the Urban Heat Planning Process

Jürgen Knies
Hochschule Bremen, Germany

Abstract

Against the background of the climate protection plan, the transition of cities towards a renewable heat supply is a particular challenge, involving planning and participation at various levels. In addition simply to the implementation of measures to reduce the heat demand of a building, the question arises as to the specific role that an individual building can play in strategic energy planning. Strategic energy planning enables a spatial framework for interaction. Communications between the building and the planning process have to be established. This paper outlines processes and communications between building information modelling (BIM) and e-government standards in Germany (XBau, XFall, XPlanung), and the need for further research and development. The combination and modification of existing standards and procedures can create something new in the long term: an indispensable data basis for municipal heat planning.

Keywords:

BIM, E-Government, XPlanung, heat planning, energy transition

1 Motivation

Meeting the targets of the climate protection plan set by the German government is an urgent and pressing challenge. Approximately one third of the energy consumed in Germany is used for room heating and hot water supply (BMWi, 2018). The reduction targets set out in the climate protection plan (BMUB, 2016) differentiate between the various sectors. In the building sector, the aim is to achieve a more or less climate-neutral building stock. This is outlined in the Energy Efficiency Strategy for Buildings issued by the German federal government (Thamling et al., 2015). In the strategy, the reduction of the overall heat requirement and the increased use of renewable heating sources are seen as complementary approaches, with different proportional weighting depending on the scenario in question. The strategy envisages a reduction of the energy demands for water and room heating of between 40% and 60% compared to 2008 for buildings used for accommodation, public services and the service sector. For the industrial sector, this figure is set at 20%. The strategy also takes into account building and planning regulations, the difficulties involved in retrofitting, demographic shifts, and the reduction in heat demand due to climate change. Heating

networks have the potential to integrate renewable energies, into the heating supply system in particular. However, on the basis of the measures that have been implemented to date, it is predicted that the reduction targets cannot be met (Graichen et al., 2017).

Recent developments in the digitalization of the construction industry and of administrative processes, known respectively as building information modelling (BIM) and e-government, as well as the ongoing development of standards and processes relating to the use of geo-data, all offer an unprecedented opportunity to cluster and synergize their respective potentials and take on the challenges of energy transition in urban settings.

This central idea is being heard in the current debate around the concept of the smart city. Given the numerous definitions of, and variety of opinions on, what the 'smart city' actually means (Albino et al., 2015), no attempt will be made here to clarify the issue further.

2 Energy planning

In the public discussion, energy transition has largely focused on electrical power supply and taken place outside cities. This is set to change in the near future, with attention turning to urban heat supply – the so-called heat transition. The transformation of the energy supply to entire cities will introduce a new dimension to the planning and participation processes as well as new technological requirements.

A municipal energy plan can indicate the nature of future developments, which are not simply derived from existing funding budgets but also take into account the potential of local sources of renewable energy as well as urban planning data relating to, for example, areas of future or current urban development, demography and mobility. Habermann-Nieße et al. (2012) propose a combination of urban development funding and the funding of energy measures. Such funding could be spatially differentiated by priority areas of energy supply options, which have to be defined from an energy planning perspective.

Municipal heating plans will play a significant role in the future (Schubert, 2015). However, the funding programmes and planning instruments that are currently available are not considered adequate to the task of transforming the energy concept of entire cities: 'The long-term vision to transform urban energy systems is often lacking at the communal planning level' (Riechel, Koritkowski, Libbe, & Koziol, 2016), and appropriate planning frameworks and standards are not in place. There is more to integrated heating planning than just planning a district heating network. Local conditions and heat supply options of all kinds must be taken into account – for example, the use of heat pumps for individual buildings, LowEx heat networks etc. (Knies, 2018).

Incentives, such as the KfW (German government-sponsored bank dedicated to regeneration projects) programme No. 432, are available. This programme provides consultancy and full planning support for areas which are to be redeveloped and can be combined with the tax incentives granted under German construction law (§ 136 BauGB) for the implementation of energy efficiency measures (Langenbrinck et al., 2017). Furthermore, regulations and laws can be invoked to enforce connection to and supply from district heating systems. It is very rare,

however, that such measures are implemented in the context of energetic urban redevelopment and their use remains very controversial (Langenbrinck et al., 2017, p. 84).

Currently, it is difficult to predict the impact of individual, funding-led planning decisions. This is because the funding and its impact on the energetic performance of a building are not recognized as inherent attributes of the building.

3 Relevant standards

A variety of standards are in use at the building-planning, urban-planning and implementation levels.

Building Information Modelling (BIM) is seen as a data and process management system, but it is also regarded as the foundation of a new culture of transparency in the construction industry. The individual components of a construction (shell, supply and disposal systems etc.) are all classified according to the Industry Foundation Classes (IFC). Very detailed information can be attached to each element, and this can be of great significance in the overall energetic assessment of a building (e.g. U-value for walls, windows and doors; for more details see http://www.buildingsmart-tech.org/specifications/ifc-releases/ifc4-add2). A comprehensive guide can be found in the German-language handbook Anwenderbuch Datenaustausch BIM/IFC (Liebich & Hoffeller, 2006). Certain parameters are important input variables for the purposes of GIS visualizations. However, to date, BIM/GIS interaction has mainly consisted in data exchange for structural engineering projects and exploring 3D-GIS visualization possibilities (Barbato et al., 2018), which has led the way to some very complex approaches towards city information modelling (Xu et al., 2014). Agugiaro et al. (Agugiaro, Benner, Cipriano, & Nouvel, 2018) examine the spatial commonalities between the levels covered by BIM, CityGML and INSPIRE, and demonstrate that the greatest commonalities in coverage are between BIM and CityGML. The authors also describe (ibid.) the Energy Application Domain Extension for CityGML (Energy ADE), which aims to include relevant energy data across an urban area and make it available for simulation purposes.

Work is currently in progress to incorporate an IT-driven process view in urban planning. The main standards under consideration here are XBau and XPlanung (IT-Planungsrat, 2017; Krause & Munske, 2016). Building application processes and notifications which have been standardized by XBau and are based on the XÖV process standards can be used as BIM data, thus enabling a seamless digital process (Krause, 2018).

To facilitate applications for the energetic funding mentioned above, the continued development of XFall for building applications is strongly recommended (http://xfall.eu/). XFall is a universally interoperable standard for application data and can be used for centralized application platforms that meet the requirements of the EU Services in the Internal Market Directive (2006/123/EG). XFall is used to transmit application documents and attachments, signatures, status information, information updates and interim reports etc. It could be used to process applications not only for KfW-funded building measures but also for grants available under German renewable energy legislation (Erneuerbaren-Energien-Gesetz,

EEG) and the market incentive programme (Marktanreizprogramm, MAP) targeted at the heating sector.

XPlanung, on the other hand, allows both the loss-free exchange of spatial planning data between different systems and, in conjunction with XBau, the integration of the plane geometry of buildings in the application process (Krause & Munske, 2016). Since October 2017, the use of XFall, XBau and XPlanung has been obligatory for all planning authorities in Germany.

4 Combining standards and processes

The processing of building applications is now digitally continuous thanks to the use of BIM, XBau and XPlanung, and this can be taken as a model for similarly continuous urban energy planning.

Energetic redevelopment areas or suitability areas for heating options should be integrated as objects in XPlanung, and their objectives should be differentiated and formalized. They are taken into account as an area setting in the further process. Even if an individual building is located outside these areas, the process would proceed.

In addition, rudimentary BIM models of the existing building stock should be available. Due to the extremely varied nature of buildings, it will not be possible to include detailed differentiations, but the aim is to cover all the existing buildings within the city. One possibility is to collect information from heating energy requirement data sources and other available city models. Energy-ADE, mentioned above, would be suitable for this purpose. As soon as the KfW receives an application for redevelopment funding, for example, the dataset can be updated to include this specific information. Depending on the proposed measures, the data on the building elements in question (windows, exterior walls, roof etc.) or building utilities (heating supply system, photo-voltaic units etc.) can be adjusted and revised. In the course of time, the overall density of data for a given city will increase.

Redevelopment measures change not only the energetic performance of the building in question. Given sufficient spatial density, such alterations can also alter the nature of the area setting by, for example, reducing the heat demand so that a sufficient number of buildings become suitable and available for connection to a low-temperature heat network. Energy suppliers can use this information as a basis for their offers and proposals. In addition, municipalities can promote local developments using targeted public information campaigns.

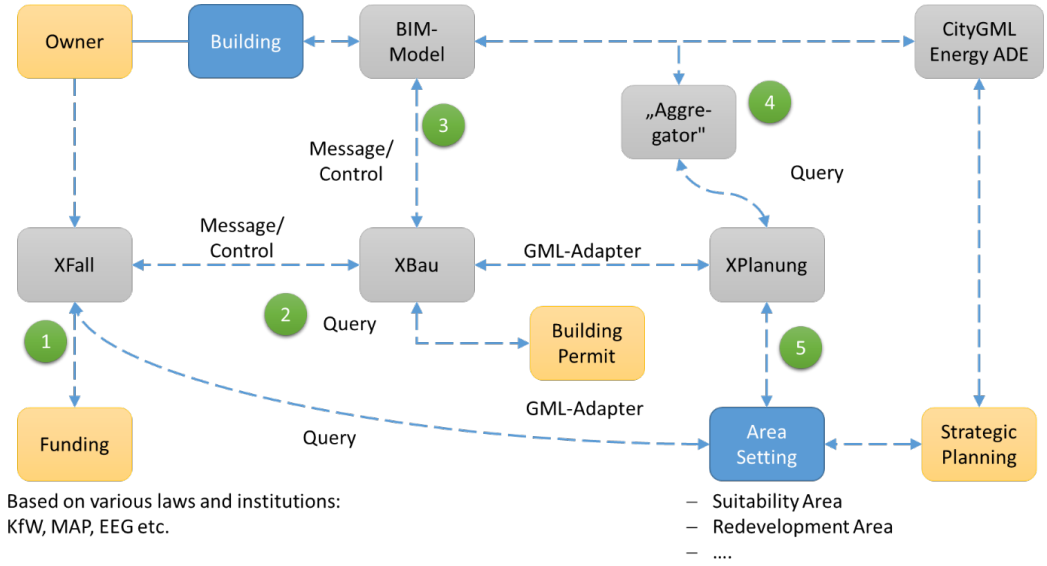

Figure 1: Proposed process chain – Yellow: actor-related; grey: standards and technical-related; blue: physical world-related

The following additional developments would also be beneficial in the context of municipal heat planning. (The numbering in the list corresponds to that in Figure 1.) The points listed represent a proposed process scheduling, starting with the initial funding application:

1. Building owners use XFall to apply for funding for energetic redevelopment measures (building components, EEG, MAP). The area setting of the building is automatically taken into account (location in a redevelopment area, in an area of suitability for heat supply options etc.).
2. XFall and XBau then communicate and exchange data, thus integrating energetic measures in the building application documents, as well as requesting information from XBau on any existing building permissions. Energy-related information, e.g. on thermal insulation of the shell and windows at the time of approval, which can be compared with the renovation measures applied for, is particularly relevant here.
3. XBau then communicates with the (rudimentary) BIM model of the building so that changes to the building and energetic data can be recorded upon completion of the redevelopment measures. This assumes, of course, that for all existing buildings a (rudimentary) BIM model exists (and is frequently updated) based on 3D city models (CityGML).
4. Aggregated energetic values are derived from the building model ('Aggregator'), which allow an appropriate heat supply option to be allocated to the building. So, for example, a building which is to be redeveloped and is currently connected to the natural gas supply could, if some further minor alterations were to be carried out to the domestic water system, be suitable for connection to a low-temperature network. Given sufficient spatial density of buildings with similar potential, the borders of the

redevelopment area can be redrawn to fit and simulations carried out for the whole area, using, for example, Energy-ADE (Agugiaro et al., 2018).

5. Based on the results of the simulation, energetic targets are formalized for the redevelopment areas and areas of suitability in question. These are communicated using XPlanung, and the information is then available for, and can be taken into account in, the funding application process.

The so-called aggregator in Figure 1 has a crucial role in the process. As no further detailed information about a building is required for subsequent strategic planning steps and the use of such data would, in any case, raise data protection issues, the detailed data from the BIM model is aggregated for the purposes of strategic planning. This means that, for any individual building, only a handful of energetic values are actually taken into account. Currently, the aggregator could best be defined using MVD (Model View Definition), which extracts the necessary information in aggregated form and transfers it to, for example, Energy-ADE.

This would allow both the interaction between, and the delineation of, BIM and CityGML to be clearly defined. It would thus enable seamless digital communication throughout the process from the funding application to the building level (BIM), and then to the strategic planning level.

In a subsequent step, detailed technical planning would start, using detailed information from the BIM models. The clear and secure delineation between the data required for strategic planning and those required for detailed technical planning is a delicate and contentious area: more work is required on this issue.

5 The way ahead

The approach outlined here is a rough draft and, as such, intended merely as the basis for further discussion on how to enhance existing standards and make use of them in municipal energy planning.

Nearly every necessary component is already there to realize an integrated process chain and to join the strengths of BIM, GIS and e-governmental standards. The lack of interoperability is not really a technical issue, but rather a question of institutional responsibility and competence. Thus the model represents a process chain but deliberately does not allocate responsibilities. This crucial point can best be addressed by the authorities responsible for building permits, funding etc., because it requires an overarching mandate.

The lack of data is a fundamental problem and a real obstacle to strategic energy planning which can only be overcome by continuously adding to the relevant databases. This can only be achieved by integrating and interlocking various standards and data streams with an explicitly spatial component.

The interaction between BIM and CityGML needs further clarification. It must be emphasized that this paper starts from the perspective of the funding of retrofitting measures and, therefore, focusses on the BIM level. Valuable datasets are being built up over time, which can then be incorporated into a CityGML and used to create city-wide perspectives and

simulations. Implementation takes place at the level of the local area or individual building, and the data from the detailed planning is processed at BIM level, continuing the information loop. In order to eliminate redundant information and inconsistencies, we need to ask at what point in the loop any updates should be carried out. This is not just a technical interface, but a question of responsibilities.

The building stock will, naturally, change over time as rebuilding measures are carried out and recorded in BIM models. It is crucial that the changing potential of an area (neighbourhood, district, suitability area etc.) to be converted to renewable heating be recognized in time. This is possible by continuously updating the energetic value data and, thus, the process of planning appropriate measures can begin in good time. Depending on the spatial distribution of the individual buildings, the extent of areas for redevelopment can be (re-)defined in an energetically rational way.

Probably the most important question is to what extent the funding mechanisms can adapt to, and take into account, the specific surroundings of any given building. A draft bill to the German parliament in 2018 (§107 GEG Entwurf) addressed this issue. However, this would require current funding policy to be changed in order to accommodate rational spatial differentiation and allow the strengths of various technologies to be combined and matched in spatial clusters.

Currently, planning law does not define any framework for municipal heat planning, so the limits of redevelopment areas and areas of suitability cannot be defined except under urban redevelopment regulations (§ 136 BauGB – German federal building regulations). There is a great need for further research into these legal aspects and how planning law would have to be revised. (On planning obstacles, see Riechel et al. (2016).) Any such developments would necessarily impact on the future specifications in XPlanung.

By combining and modifying current standards and processes, something novel can be created: continuous communication from the level of funding to the building itself and, further, to the level of strategic energy planning and the creation of a data resource available as an indispensable tool for decision-makers involved in heat transition at municipal level.

References

Agugiaro, G., Benner, J., Cipriano, P., & Nouvel, R. (2018). The Energy Application Domain Extension for CityGML: enhancing interoperability for urban energy simulations. Open Geospatial Data, Software and Standards, 3(1), 30. https://doi.org/10.1186/s40965-018-0042-y

Albino, V., Berardi, U., & Dangelico, R. M. (2015). Smart Cities : Definitions , Dimensions , Performance , and Initiatives. Journal of Urban Technology, 22(1), 3–21.

H ttps://doi.org/http://dx.doi.org/10.1080/10630732.2014.942092

Barbato, D., Pristeri, G., & Marchi, M. De. (2018). GIS-BIM Interoperability for Regeneration of Transurban Areas. In M. SCHRENK, V. V. POPOVICH, P. ZEILE, P. ELISEI, C. BEYER, G. NAVRATIL, & 243 (Eds.), REAL CORP 2018 – EXPANDING CITIES – DIMINISHING SPACE (pp. 243–250). Wien.

BMUB. (2016). Klimaschutzplan 2050 - Kabinettbeschluss vom 14. November 2016 (p. 91). p. 91. Retrieved from

http://www.bmub.bund.de/fileadmin/Daten_BMU/Download_PDF/Klimaschutz/klimaschutzplan_2050_bf.pdf

BMWi. (2018). Verteilung des Energieverbrauchs nach Anwendungsbereich in Deutschland im Jahresvergleich 2008 und 2016. Retrieved from https://de.statista.com/statistik/daten/studie/253748/umfrage/anteil-der-anwendungsbereiche-am-gesamtenergieverbrauch-in-deutschland/

Graichen, P., Peter, F., & Litz, P. (2017). Das Klimaschutzziel von -40 Prozent bis 2020: Wo landen wir ohne weitere Maßnahmen? (p. 10). p. 10. Retrieved from https://www.agora-energiewende.de/fileadmin/Projekte/2015/Kohlekonsens/Agora_Analyse_Klimaschutzziel_2020_07092016.pdf

Habermann-Nieße, K., Jütting, L., Klehn, K., & Schlomka, B. (2012). Strategien zur Modernisierung: Mit EKO-Quartieren zu mehr Energieeffizienz; eine Studie. Band 24 Der Schriftenreihe Ökologie, p. 86. Retrieved from: https://www.boell.de/sites/default/files/Endf_Strategien_zur_Moderisierung_2_kommentierbar.pdf

IT-Planungsrat. (2017). Betriebskonzept XBau / XPlanung, 28.04.2017 / Version 1.0 final (p. 20). p. 20. Retrieved from https://www.it-planungsrat.de/SharedDocs/Downloads/DE/Entscheidungen/23_Sitzung/StandardisierungsagendaAnlage2.pdf?__blob=publicationFile&v=2

Knies, J. (2018). A spatial approach for future-oriented heat planning in urban areas. International Journal of Sustainable Energy Planning and Management, 16, 3–30. https://doi.org/10.5278/ijsepm.2018.16.2

Krause, K.-U. (2018, February). Xplanung / Xbau Standards des IT-Planungsrates für den Bau- und Planungsbereich: BIM ready. Retrieved from https://www.innovationsforen-bauen40.de/wp-content/uploads/2018/03/XPlanung-XBau-BIM-Ready.pdf

Krause, K. U., & Munske, M. (2016). Geostandards XPlanung und XBau. ZFV - Zeitschrift Fur Geodasie, Geoinformation Und Landmanagement, 141(5), 336–342. https://doi.org/10.12902/zfv-0137-2016

Langenbrinck, G., Rensing, L., Wüllner, L., Klaus Habermann-Nieße, K. K., & Rosenau, L. (2017). KfW-Programm 432 „Energetische Stadtsanierung – Zuschüsse für integrierte Quartierskonzepte und Sanierungsmanager“ Ergebnisse der Begleitforschung (p. 98). p. 98. Retrieved from http://www.bbsr.bund.de/BBSR/DE/Veroeffentlichungen/BBSROnline/2017/bbsr-online-25-2017-dl.pdf

Liebich, T., & Hoffeller, T. (2006). Anwenderhandbuch Datenaustausch BIM / IFC (1.1; IAI - Industrieallianz für Interoperabilität e.V., Ed.). Retrieved from http://www.dds-cad.de/fileadmin/redaktion/PDF-Dateien/buildingSMART-IFC_Anwenderhandbuch_Version1.0_4MB.pdf

Riechel, R., Koritkowski, S., Libbe, J., & Koziol, M. (2016). Wärmewende im Quartier - Hemmnisse bei der Umsetzung am Beispiel energetischer Quartierskonzepte (p. 28). p. 28. Retrieved from http://edoc.difu.de/edoc.php?id=FZRP4QJM

Schubert, S. (2015). Die Rolle räumlicher Planung zur Förderung klimaschonender Wärme- und Kälteversorgung in Deutschland und der Schweiz (Dissertation). Verlag Dorothea Rohn, Lemgo.

Thamling, N., Pehnt, M., & Kirchner, J. (2015). Hintergrundpapier zur Energieeffizienzstrategie Gebäude (p. 131). p. 131. Retrieved from https://www.bmwi.de/BMWi/Redaktion/PDF/E/energieeffizienzstrategie-hintergrundinformation-gebauude.pdf

Xu, X., Ding, L., Luo, H., & Ma, L. (2014). From Building Information Modeling to City Information Modeling. Journal of Information Technology in Construction (ITcon), 19 (December 2013), 292–307.

GI_Forum 2020, Issue 1
Page: 107 - 118
Short Paper
Corresponding Author:
lukas.goetzlich@researchstudio.at
DOI: 10.1553/giscience2020_01_s107

GIS-based Heat Demand Modelling for Tourist Accommodation. A Case Study in the State of Salzburg

Lukas Götzlich[1], Martin Santa Maria[1], Markus Biberacher[1] and Ingrid Schardinger[1]
[1]Research Studios Austria Forschungsgesellschaft, Salzburg, Austria

Abstract

Spatial energy planning plays a key role in energy transition. Geo-information systems (GIS) make an important contribution in this context: spatially differentiated modelling, representation, and analysis of energy demands in the building sector are the basis for well-founded strategic energy planning. This paper presents a GIS-based method to model the heat demand for tourist accommodation in the federal state of Salzburg. The paper includes the development, description, implementation and validation of the heat demand modelling based on a bottom-up approach. The discussion reflects on possible improvements to the approach.

Keywords:

GIS model, tourist accommodation, heat demand

1 Introduction

Energy planning is indispensable in ensuring environmental sustainability and reducing the risk associated with climate change. The main task of energy planning is to quantify local heat demands and energy potentials, and to provide the data in a spatially differentiated form. Energy planners depend on reliable and transparent models that support them in their decision-making processes. The required depth of the models' data (spatial and temporal resolution) varies depending on the application, and energy planning normally includes additional spatially located information (e.g. building stock, supply infrastructure, energy potentials) (Mauthner, 2018). A technical and methodological challenge in creating models is to harmonize and standardize the available data to generate the greatest possible benefit.

In general, Geographic Information Systems (GIS) are suitable for processing, standardizing and visualizing several layers of information, which is why they have frequently been used in the field of energy planning for many years (Mardani et al., 2017). Improvements in data availability, data quality and computer performance enable us to conduct more context-specific analyses (including when using GIS), which often require a revision or extension of existing models or the creation of new models.

In an existing approach to heat demand modelling by Schardinger & Biberacher (2017), a concrete need for research into modelling the heat demand for tourist accommodation was exposed. In their project, Schardinger & Biberacher evaluated a heat demand model at building level using the heat consumption values of regional district heating providers. In the process, significant inaccuracies were found in areas with a high share of tourist accommodation. Their additional facilities and services (e.g. swimming pools, sports facilities, saunas and laundries) produce heat demands in addition to those of room heating.

The literature offers a variety of different methods to model the heat demand. Li et al. (2017) divided the models into top-down and bottom-up ones. Bottom-up models are characterized by a higher degree of spatial and temporal detail than top-down ones. They are usually based on empirical real data ('real example building') or representative building characteristics ('real average building' or 'synthetic average building'). The necessary information for all individual buildings is often unavailable, which is why the building stock investigated is classified into building types (Ballarini et al., 2014; Loga et al., 2016), and the building models are parameterized using the characteristics of these types (Nageler et al., 2017; Schiefelbein et al., 2019; Streicher et al., 2019). The disadvantages of such bottom-up models are the high data intensity and uncertainties due to the typology (Brøgger & Wittchen, 2018).

This paper uses an innovative bottom-up approach to model heat demand, focusing exclusively on tourist accommodation in the federal state of Salzburg. The buildings' gross floor areas[1] (GFA) and the energy consumption indicator (ECI) serve as a basis for the model. The paper validates the model by comparing modelled values with reported data.

It continues the research into heat demand models for tourist accommodation and contributes to two sub-fields of energy planning: (1) mapping of buildings, and (2) heat demand modelling.

The paper has three main objectives: (1) the development, description and application of a bottom-up modelling approach for the mapping and heat demand modelling of tourist accommodation in the federal state of Salzburg; (2) the partial validation of the modelling approach developed here, using reported data; (3) the development of a data concept for the localization and characterization of the tourist accommodation stock and, based on this, a building typology for heat demand modelling. The data concept is not limited to the building type 'tourist accommodation': it is transferable to other building types because the concept is reduced to a minimum set of input parameters and relies mainly on data that is available nationwide. ArcGIS Pro v. 2.2.4 and QGIS v. 2.18 and v. 3.6 served as GIS.

The paper has the following structure. Section two provides the methodological background. First, it lists the data sources and categories used (§2.1), before describing the workflow (§2.2). Section 3 provides the results, and the paper concludes with a discussion in Section 4.

[1] The gross floor area is defined as the sum of the aboveground and underground floor areas of a building that have to be heated or cooled during use (Amstutz & Schegg, 2003).

2 Method

2.1 Data sources and categorization

Table 1 provides an overview and a short description of the data sources used. The main data source for the model is the federal state of Salzburg. Additionally, the model integrates further building information (address, type, company, number of stars for hotels) from Herold, Bundesamt für Eich- und Vermessungswesen (BEV) and Wirtschaftskammer Salzburg (WKS).

Table 1: Data sources

Data Source	Name	Type	Date	Description
Salzburg State	Cadastre	shape	2016	All buildings in the federal state of Salzburg as polygons
Salzburg State	Communities	shape	2019	All communities of the federal state of Salzburg
Salzburg State	DEM[2]	raster	2016	DEM of the federal state of Salzburg, resolution 1m
Salzburg State	DSM[3]	raster	2016	DSM of the federal state of Salzburg, resolution 1m
Salzburg State	HDD[4]	csv	2016	HDD for every community in the federal state of Salzburg, years
BEV	Addresses	csv	2018	Tables with data about addresses, buildings and types in the federal state of Salzburg
WKS	Tourist accommodation	csv	2017 2018	Table with information about companies, their name, address, type, and (for hotels) information about stars, in the federal state of Salzburg
Herold	Tourist accommodation	shape	2016	Information about addresses, building types and company names in the federal state of Salzburg

The study uses the following categories for tourist accommodation (Hotel Energy Solutions, 2011): hotels, apartments, campsites, holiday homes, inns, guesthouses, private rooms, mountain huts and youth hostels. The category 'hotel' has three sub-categories, for 2/3, 4 and

[2] Digital Elevation Model

[3] Digital Surface Model

[4] Heating Degree Days; these represent a relationship between the room temperature and the outside temperature during the heating period and are used to find out about the heat demand. In Austria, a room temperature of 20°C and a temperature of 12°C are applied for the calculation of heat demand. This means that if the outside temperature is below 12°C, a room has to be heated to maintain a temperature of 20°C.

5 stars (see Table 2). The paper draws on several data sources, which differ in up-to-dateness and data collection interval. In order to avoid errors, it uses the data for 2016.

2.2 Workflow

The starting point for building mapping and heat demand modelling is to locate all relevant buildings of the type 'tourist accommodation', as well as to allocate values for building characteristics (especially GFA). A climate-adjusted heat demand modelling at the building level is then performed, based on this information.

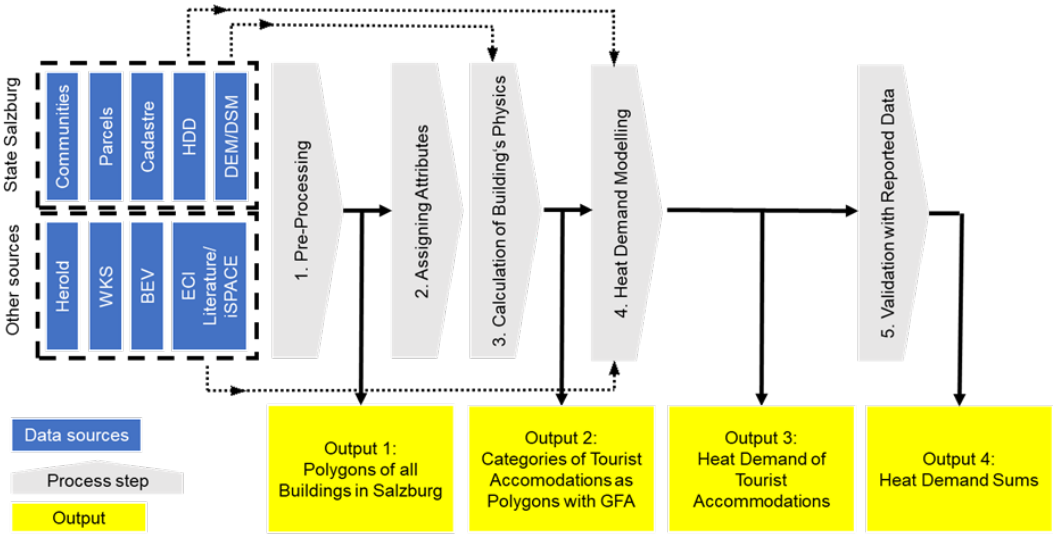

Figure 1: Workflow

Figure 1 shows the workflow used in the study, which consisted of five steps:

1. Pre-processing: In the first step, all data from the various data sources go through pre-processing to generate polygons of all buildings in the state of Salzburg as an intermediate output.

2. Assigning attributes: The polygons from step 1 are matched with all the attributes (sub-categories, addresses) relevant for the model.

3. Calculation of buildings' physical characteristics: The DEM and DSM provide the height and the GFA for all tourist accommodation.

Steps 2 and 3 both pay particular attention to the buildings on the same parcel and to buildings that touch tourist accommodation, because tourist facilities tend to extend over several parcels. To prevent having to manipulate the model further, these buildings are given polygon feature classes, before a later step dissolves and transforms all feature classes of one address to a single feature class.

4. Heat demand modelling: Step 4 comprises the heat demand modelling using ECI. The first part of the heat demand modelling is to select suitable ECIs.

The literature provides a variety of ECIs for tourist accommodation, obtained using various methods (Amstutz & Schegg, 2003; Bayer et al., 2011; Benke et al., 2012; Perincoli et al., 2010). The ECIs differ in their categorization of tourist accommodation or their regional origin. Our paper uses a combination of ECIs from Amstutz & Schegg (2003) and Benke et al. (2012), because of their regional proximity to the federal state of Salzburg.

Table 2 shows the final ECIs used for the categories of tourist accommodation, adapted to a reference climate[5] (for the town of Bischofshofen).

Table 2: Selected ECIs for the tourist accommodation types investigated

Category		Energy Consumption Indicator (ECI) in kWh/m²/a
Hotels	5 stars	156
	4 stars	137
	3 and 2 stars	118
	No information	125
Apartments		83
Guesthouses		100
Holiday Homes		83
Inns		131
Private Rooms		83
Youth Hostels		134

After the selection of the ECIs, the multiplication of the GFA value by the corresponding ECI delivers the heat demand of the building. For mountain huts and campsites, no ECIs could be found, so they could not be regarded in the heat demand model.

5. Validation with reported data: In the final step, the method for heat demand modelling is compared with reported data. The validation process allows us to identify the weaknesses and strengths of the model and to avoid possible statistical errors.

The federal state of Salzburg provided heat demand data for 52 of its 119 communities, with the number of reported values varying for each community (minimum 1, with up to 344 heat demand values for any one community). These data are independent of the data used in the

[5] The following formula was used to calculate the appropriate ECIs:

$$\left(\frac{ECI_{Lit}}{HDD_{Lit}}\right) * HDD_{B'hofen} = ECI_{B'hofen}$$

heat demand modelling; hence, the comparison of these two independent datasets validates the model.

For the validation, the sums of the heat demand for the individual categories of the two datasets are compared. The heat demand sum is the sum of the heat demand of all buildings for one category and is a hypothetical value. The deviation between model and data is the parameter to evaluate the degree of agreement.

Table 3 provides an overview of the output data.

Table 3: Characteristics of the output datasets

Name	Accommodation	Extent	Buildings	Geometry	Attributes
Output 1	All	Salzburg	Address, Parcel based and touching buildings	Polygon	Address, Company, Type
Output 2	Tourist accommodation	Salzburg	Address, Parcel based and touching buildings	Polygon	Address, Company, Type, Category, Stars, Height, GFA
Output 3	Tourist accommodation	Salzburg	Address, Parcel based and touching buildings	Polygon	Address, Company, Type, Category, Stars, Heat Demand
Output 4	Tourist accommodation	Salzburg	Address	None, table data	Category, Heat Demand Sums

3 Results

The results of the building mapping and heat demand modelling are a precise characterization of the building stock and its estimated heat demand. This section presents the results from outputs 2, 3 and 4.

3.1 Output 2: Calculation of buildings' physical characteristics

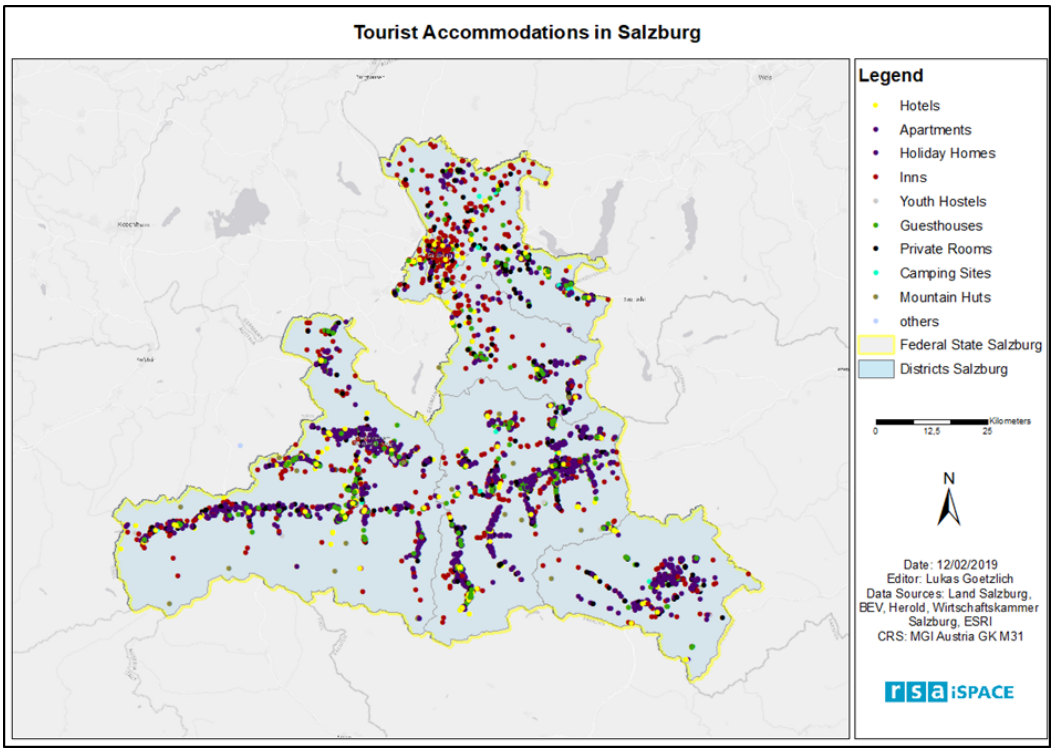

Figure 2: Tourist accommodation in Salzburg

Figure 2 shows the tourist accommodation in the state of Salzburg covered in the model. In total, 5,615 addresses and 11,125 buildings were categorized as tourist accommodation. At the district level, most of the accommodation is located in the city of Salzburg and its surroundings (1,100 addresses and 2,475 buildings), followed by the Saalbach Hinterglemm (585 buildings) and Flachau (470 buildings) districts. In terms of categories, holiday homes (1,576 addresses and 3,196 buildings) have the largest share, followed by inns (1,076 addresses and 2,639 buildings). Hotels are assigned to 802 addresses with 1,121 buildings. For about 60% of the hotels, there is no information available for the sub-categorization (number of stars). For the hotels with information, the subcategory '4-star hotel' has the largest share (170 addresses and 245 buildings).

3.2 Output 3: Heat demand modelling

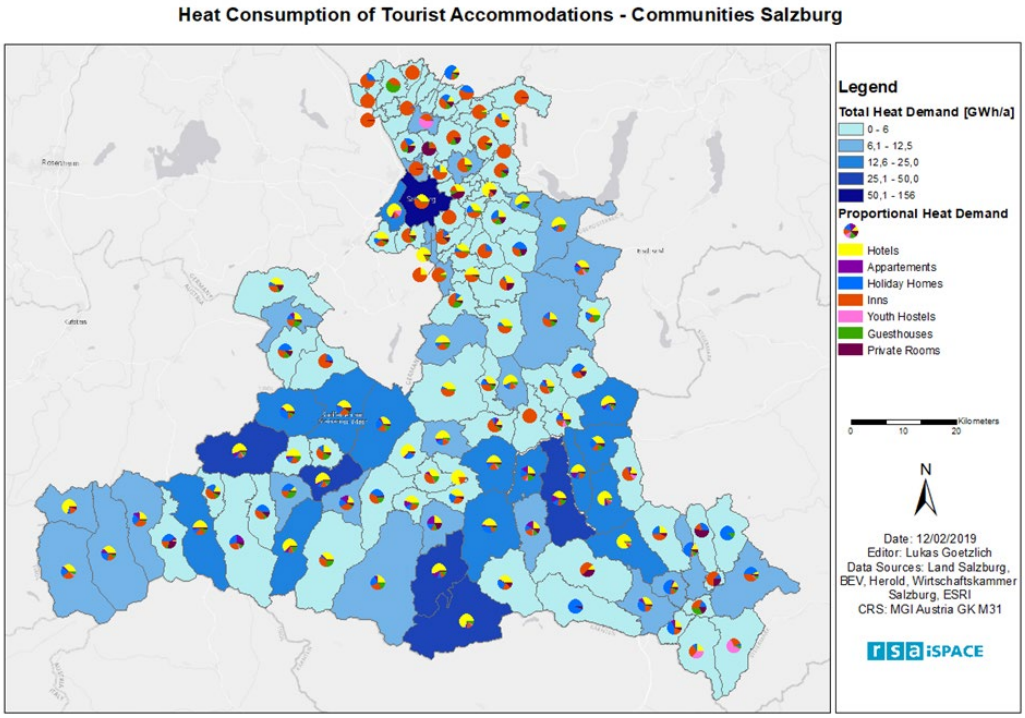

Figure 3: Heat consumption of tourist accommodation

3.3 Output 4: Validation using reported data

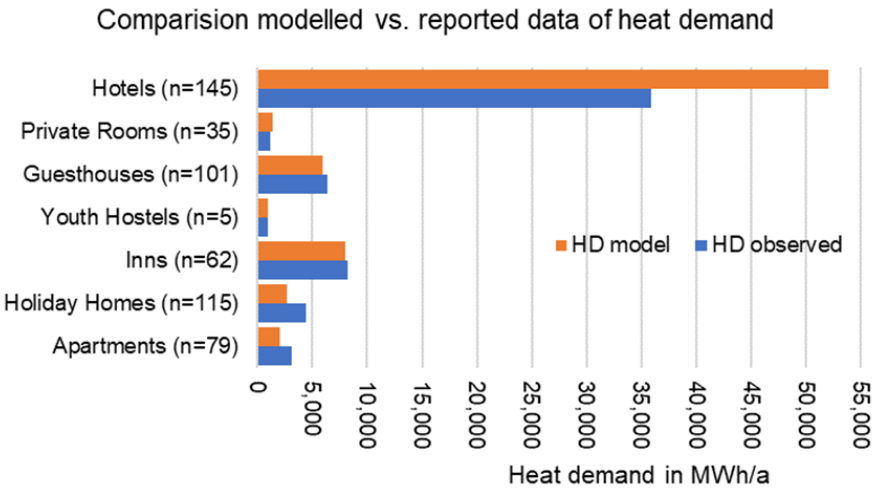

Figure 4: Comparison of modelled data vs. reported data of heat demand

Figure 4 compares the heat demand sums of the reported and modelled data of each category. The model tends to underestimate the demand, except for the categories 'hotels' and 'private rooms'. Out of the seven categories, six show a deviation of less than 40%. The largest variation is for the hotel category, although for this category the model has the highest level of detail due to the subcategories. As a result, a second validation, for the category 'hotel' only, was performed.

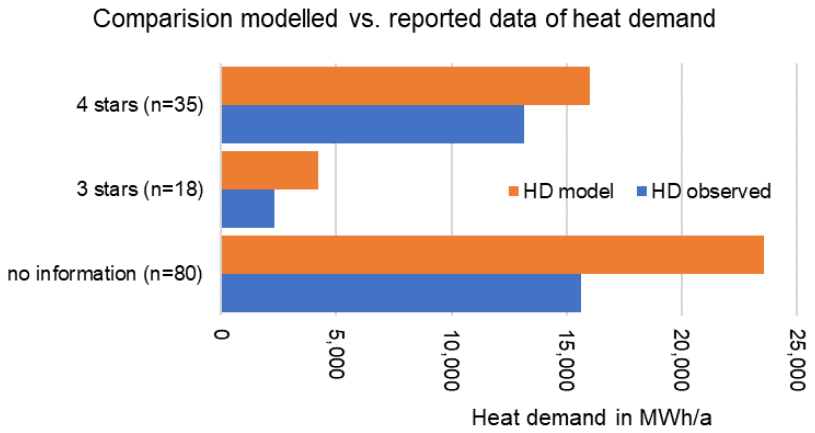

Figure 5: Comparison of modelled vs. reported data for heat demand (category 'hotel')

Figure 5 illustrates the comparison between the heat demand sums of the reported and modelled data for each sub-category of hotel. No reported values were available for the sub-category '2 stars'; therefore, Figure 5 does not list it. The Figure demonstrates the result from Figure 4 even more clearly, namely that the model generates excessively high values for the category 'hotel'. A further division into the subcategories '3-star' and '4-star' did not produce any additional findings.

The sample available for the validation is rather small (542 addresses with measured values from a total of 5,612 addresses in the study area). The quantitative comparison with real data has shown that the modelled heat requirements tend to underestimate the actual consumption. This finding contradicts several other studies (Bauer & Weiler, 2013; Rehbogen et al., 2017) and may be explained by differences in user behaviour.

The results for the 'hotel' category may be systematic misinterpretations of the modelled demand vs the actual demand. Further calibration of the indicators based on real data (e.g. measurement campaigns) should be the next step for further model validation and improvements.

4 Discussion and outlook

All assumptions and specifications made in the model offer the potential for improvements.

A critical point in the workflow is the assignment of attributes to buildings. All buildings on a parcel with tourist accommodation are typified as tourist accommodation, which is not necessarily true. This step is advisable because for many buildings no information is available. In the future, the quality of this information should be improved by a (preferably automated) comparison with other data sources (Google Maps, OSM, etc.). Clearly defined rules for data migration are also important to enable importing from different sources to the new schema.

The definition of the building types is an essential step in the workflow and all assignments for the building-specific heat demand modelling are based on these definitions. Established typologies from Hotel Energy Solutions (2011) were used to define the types.

Another sensitive point is the calculation of the height of the building using DEM and DSM. The study used the median value of all pixels located in the base area of a building to determine the height. One alternative parameter for the height is the mean value of all pixels.

The availability of more representative indicators for the different types of tourist accommodation can be further optimized. In particular, a calibration of the key figures based on real data (e.g. measurement campaigns) is necessary for further validations and improvements.

An alternative approach to the one presented here which uses ECI and GFA is to model the heat demand using the gross volume of the building (Kalasek & Brus, 2018).

An essential result of our study is the description and implementation of a methodology for building mapping and heat demand modelling. The methodology includes a data concept and a definition of building typologies to model the demand using energy consumption indicators. The method was applied for the federal state of Salzburg and validated using reported data. The approach we have presented is reduced to a few input parameters, is based mainly on data available nationwide, and is therefore transferable to other building types and areas.

The validation of the model results shows that the heat demand model tends to underestimate the actual demand and that a further calibration of the model parameters is necessary. Additionally, the amount of available reported data is small.

The present paper provides a method for building mapping and heat demand modelling that is applicable more widely. The method provides an important contribution to spatial energy planning in regions with a high share of tourist accommodation (like the state of Salzburg). The findings of the paper on building mapping and heat demand modelling of tourist accommodation will be further addressed in the ongoing project S/E/P - Spatial Energy Planning for Heat Transition.[6]

[6] Website of the project: http://www.waermeplanung.at/

References

Amstutz, M., & Schegg, R. (2003). Energieeffizienz und CO2-Emissionen der Schweizer Hotellerie. *Bundesamt Für Energie BFE*. https://www.hslu.ch/en/lucerne-university-of-applied-sciences-and-arts/research/projects/detail/?pid=259

Ballarini, I., Corgnati, S. P., & Corrado, V. (2014). Use of reference buildings to assess the energy saving potentials of the residential building stock: The experience of TABULA project. *Energy Policy*, *68*, 273–284. https://doi.org/10.1016/j.enpol.2014.01.027

Bauer, E., & Weiler, T. (2013). Investitions- und Nutzungskosten in Wohngebäuden gemeinnütziger Bauvereinigungen unter besonderer Berücksichtigung energetischer Aspekte. *Österreichischer Verband Gemeinnütziger Bauvereinigungen (Gvb)*, 56.

Bayer, D. G., Sturm, T., & Hinterseer, S. (2011). Bericht über Kennzahlen zum Energieverbrauch in Dienstleistungsgebäuden. *Klima- und Energiefonds im Rahmen des Programms „Neue Energien 2020"*, 27.

Benke, G., Leutgöb, K., Jandrokovic, M., Bayer, G., Baumgartner, D., Auer, M., & Mayer, B. (2012). Energieverbrauch im Dienstleistungssektor. *E7 Energie Markt Analyse GmbH, Österreichsicher Klima- und Energiefonds*.
https://www.e-sieben.at/publikationen/0910_Endbericht_EV_DLG/Endbericht-EV-DLG.pdf?m=1570012985&

Brøgger, M., & Wittchen, K. B. (2018). Estimating the energy-saving potential in national building stocks – A methodology review. *Renewable and Sustainable Energy Reviews*, *82*, 1489–1496. https://doi.org/10.1016/j.rser.2017.05.239

Ferrari, S., Zagarella, F., Caputo, P., & D'Amico, A. (2019). Results of a literature review on methods for estimating buildings energy demand at district level. *Energy*, *175*, 1130–1137. https://doi.org/10.1016/j.energy.2019.03.172

Hotel Energy Solutions. (2011). Analysis on energy use by European hotels: Online survey and desk research. *Hotel Energy Solutions Project Publications*.
http://www.nezeh.eu/assets/media/fckuploads/file/Reports/10.HESreserch.pdf

Kalasek, R., & Brus, T. (2018). Berechnungsgrundlagen Heizwärmebedarf HWBsk für Gebäude (unveröffentlicht). *Magistrat Der Stadt Wien MA20 - Energieplanung / TU Wien*.

Li, W., Zhou, Y., Cetin, K., Eom, J., Wang, Y., Chen, G., & Zhang, X. (2017). Modeling urban building energy use: A review of modeling approaches and procedures. *Energy*, *141*, 2445–2457. https://doi.org/10.1016/j.energy.2017.11.071

Loga, T., Stein, B., & Diefenbach, N. (2016). TABULA building typologies in 20 European countries—Making energy-related features of residential building stocks comparable. *Energy and Buildings*, *132*, 4–12. https://doi.org/10.1016/j.enbuild.2016.06.094

Mardani, A., Zavadskas, E. K., Khalifah, Z., Zakuan, N., Jusoh, A., Nor, K. M., & Khoshnoudi, M. (2017). A review of multi-criteria decision-making applications to solve energy management problems: Two decades from 1995 to 2015. *Renewable and Sustainable Energy Reviews*, *71*, 216–256. https://doi.org/10.1016/j.rser.2016.12.053

Mauthner, F. (2018). *Vergleich von GIS-basierten Methoden zur Kartierung von Wärmebedarfen—Grundlagen räumlicher Energieplanung am Beispiel der Stadtgemeinde Gleisdorf*.
https://doi.org/10.13140/RG.2.2.25449.0368

Nageler, P., Zahrer, G., Heimrath, R., Mach, T., Mauthner, F., Leusbrock, I., Schranzhofer, H., & Hochenauer, C. (2017). Novel validated method for GIS based automated dynamic urban building energy simulations. *Energy*, *139*, 142–154. https://doi.org/10.1016/j.energy.2017.07.151

Perincoli, L., Hotelleriesuisse, 3003 Bern, Bundesamt für Energie BFE, 3003 Bern, & Energie-Agentur der Wirtschaft, 8032 Zürich. (2010). Energiemanagement in der Hotellerie. *Leitfaden Energiemanagement in Der Hotellerie*, *3. Auflage*(Ing. Büro Energie & Umwelt). http://www.hotelpower.ch/sites/default/files/eidh_d_wkom_link_0.pdf

Rehbogen, A., Strasser, H., Koblmüller, M., Mostegl, N., Schardinger, I., & Biberacher, M. (2017). *Integrierter Wärmeplan Zentralraum Salzburg—Umsetzungsplanung für die Wärmewende der Energie-Vorzeigeregion Salzburg (heatswap_Salzburg)*. https://www.vorzeigeregion-energie.at/projekt/heatswap_salzburg/

Reinhart, C. F., & Cerezo Davila, C. (2016). Urban building energy modeling – A review of a nascent field. *Building and Environment*, *97*, 196–202. https://doi.org/10.1016/j.buildenv.2015.12.001

Schardinger, I., & Biberacher, M. (2017). Fernwärmepotenzial im Bundesland Salzburg. *Interner Projektbericht. Im Auftrag Des Amtes Der Salzurger Landesregierung. - Salzburg*.

Schiefelbein, J., Rudnick, J., Scholl, A., Remmen, P., Fuchs, M., & Müller, D. (2019). Automated urban energy system modeling and thermal building simulation based on OpenStreetMap data sets. *Building and Environment*, *149*, 630–639. https://doi.org/10.1016/j.buildenv.2018.12.025

Streicher, K. N., Padey, P., Parra, D., Bürer, M., Schneider, S., & Patel, M. (2019). Analysis of space heating demand in the Swiss residential building stock: Element-based bottom-up model of archetype buildings. *Energy and Buildings*, *184*, 300–322. https://doi.org/10.1016/j.enbuild.2018.12.011

GI_Forum 2020, Issue 1
Page: 119 - 136
Full Paper
Corresponding Author:
ljuhasz@fiu.edu
DOI: 10.1553/giscience2020_01_s119

Studying Spatial and Temporal Visitation Patterns of Points of Interest Using SafeGraph Data in Florida

Levente Juhász[1] and Hartwig Hochmair[2]

[1]Florida International University, Miami, FL, USA
[2]University of Florida, Ft. Lauderdale, FL, USA

Abstract

SafeGraph is a commercial provider of massive Point of Interest (POI) data, including visitation patterns in North America. Although the data source does not share specific travel trajectories, the data available includes daily and monthly POI visitation numbers for over 160 categories, as well as information about where visitors come from and which other POI categories they visit. This allows analysts to gain insight into travel behavior in a geographic region over time. This study analyzes various aspects of visitation patterns that can be derived from the SafeGraph dataset for Florida. Using three major Florida cities, namely Miami, Orlando and Jacksonville, temporal patterns of daily and monthly visit numbers are correlated between various POI categories, and the effect of a short event (Hurricane Irma) on daily visitation numbers around the event is explored. In addition, travel distances from home to POIs are compared between different POI categories, and Ordinary Least Squares (OLS) regression models are used to identify factors associated with increased or decreased distance between home and a specific POI category. The study concludes that the aggregated data provided on the SafeGraph platform helps the GIScience community to learn more about travel patterns in both the spatial and the temporal domains.

Keywords:

Travel behavior, visiting patterns, Point of Interest, Florida, hurricane, urban environment

1 Introduction

Society increasingly utilizes location-based services (LBS) that cover a wide range of functionalities including navigation, social networking, assistive healthcare, customized advertising, event recommendation and participatory decision making. LBS often use spatial information derived from Point of Interest (POI) information, for example when recommending overnight accommodations based on user reviews. The tech industry utilizes POIs in geo-gaming and mapping applications (Juhász & Hochmair, 2017; Juhász, Novack, Hochmair, & Qiao, 2020), and to derive detailed land use/land cover information (Spyratos, Stathakis, Lutz, & Tsinaraki, 2017). Apart from being a static collection of places, POI data combined with visitor patterns can be used to study urban dynamics and user markets. This

added information is commonly applied for location-based advertising or consumer analytics (Baik, Lee, Lee, Kim, & Choi, 2016). For example, Foursquare check-in information can help to predict the types of places a user will visit in the future (Zhuang et al., 2017), or one can provide POI visiting recommendations based on the analysis of visit trajectories (Massimo & Ricci, 2019).

Since POIs can serve as the data foundation for a variety of industry applications and solutions for answering societal questions, there is no single best source of POI in general. That is, different datasets can vary in content, completeness or quality depending on their purpose. For example, POIs compiled from social media services tend to be more abundant than POIs from business- and mapping-oriented sources, but at the same time they tend to have higher positional errors (Hochmair, Juhász, & Cvetojevic, 2018).

SafeGraph is a commercial provider of POI data that compiles its dataset from several sources, such as mobile phone GPS data and governmental open data, to build a comprehensive business listing in the United States and Canada. In addition to the POI data itself, the company derives visitation pattern information and aggregates it to POIs, which enables access to visitor and visit counts and certain aspects of demographics. SafeGraph POI visitation data, which is made available for academic research free of charge in aggregated form, can be a useful source of information for studying certain aspects of urban dynamics and travel behavior. For example, one study used SafeGraph data to reconstruct origin–destination pairs in Milwaukee, Wisconsin, in order to explore the spatial isolation of neighborhoods (Prestby, App, Kang, & Gao, 2019). Different machine learning models applied to SafeGraph data have also been used to predict parking violations (Gao et al., 2019) which depended on certain POI categories, such as retail stores or restaurants. The dataset can also be used to assess the effect of certain events or policy changes on visitation patterns. This has been illustrated in a study that analyzes the change in visitation patterns after Starbucks implemented an open bathroom policy allowing anyone, even without a purchase, to use their bathroom facilities. Results revealed a 6.8 % decline in store visits compared to other nearby restaurants and cafes (Gurun, Nickerson, & Solomon, 2020). More recently, SafeGraph data has been used to assess compliance with guidelines on social distancing in response to COVID-19 (Andersen, 2020) and for building a POI database in conjunction with other data sources for informed decision making (Killeen et al., 2020).

This research contributes to the growing body of literature using SafeGraph data and conducts exploratory analyses in the spatial and temporal domains using three major cities in Florida (Miami, Orlando and Jacksonville) as study areas. More specifically, the study has the following objectives:

1. To compare the temporal characteristics of visitation patterns between different POI categories using monthly and daily correlation analysis
2. To study the effect of short-term events on POI visitation patterns, and
3. To analyze distances between visitors' home locations and POIs visited.

In pursuing these objectives, the study showcases novel analysis approaches applied to the SafeGraph dataset. These include, for example, analyzing the effects of hurricanes on visitation frequency to certain POI types before, during and after the hurricane, reflecting the type of preparation that the population are making for such an event. It also identifies the localness of POI types, i.e. the role of a specific POI type for the local population versus its role for visitors from further away.

2 Study Setup

2.1 POI data collection and pre-processing

SafeGraph's main product is SafeGraph Places, which consists of three datasets, namely Core POI, Geometry and Patterns. The *Core POI* dataset contains basic information of about 6.1 million POIs in the US and Canada, such as the name, brand association (i.e. if the POI is part of a chain), address, category and opening hours, along with an internal place ID. As well as the point geometry of POIs, the *Geometry* dataset also contains their polygon representation, for example the outline of the buildings that POIs are housed in. The hierarchy of POI locations is also included in this dataset. That is, POIs can be nested within each other, which is often the case when a larger entity, such as a shopping mall, contains multiple individual stores. The *Patterns* dataset describes visitation patterns to over 3.6 million unique POIs. These patterns include monthly aggregated visitor and visitation numbers, daily visits and dwell times. The dataset also includes the number of home and work locations of visitors as well as number of smartphone devices observed, which are aggregated at the level of the US census block group. SafeGraph uses accurate smartphone GPS locations and machine learning to attribute visits to POIs.

SafeGraph provided us with their Florida dataset, which consists of 302,201 POIs. Among these, 258,658 POIs also contain Patterns data for at least one month between January 2017 and August 2019. The dataset is available as a collection of plain text flat files. These files were parsed and inserted into a spatially enabled PostgreSQL database using the Places schema (SafeGraph, 2020) for further processing. Standalone tables can be joined by the common internal place ID that is attached to all places, geometries and patterns. Even though it is stored in a relational database, the schema contains several fields represented as JSON documents. Custom SQL queries were designed to extract information (e.g. from JSON) to complete each analysis step (described in Section 2.2). The study is geographically limited to three major cities in Florida by filtering POI addresses. Further, the analysis was also limited to 17 POI categories. The POI numbers included in this study for the different POI categories and cities are summarized in Table 1.

Table 1: Summary of POI categories and the number of POIs per city included in the study

		# of POIs included		
Code	**Category name**	**Miami**	**Orlando**	**Jacksonville**
ACC	Accounting, Tax Preparation, Bookkeeping, and Payroll Services	173	150	141
AMU	Amusement Parks and Arcades	19	54	15
AUTO	Automobile Dealers	278	249	163
ALC	Beer, Wine, and Liquor Stores	113	74	71
UNI	Colleges, Universities, and Professional Schools	97	71	77
ELEC	Electronics and Appliance Stores	89	60	61
SCH	Elementary and Secondary Schools	743	445	502
GAS	Gasoline Stations	377	253	356
GROC	Grocery Stores	449	403	427
HOME	Home Health Care Services	102	45	58
LUG	Jewelry, Luggage, and Leather Goods Stores	311	158	101
MUS	Museums, Historical Sites, and Similar Institutions	182	132	146
DOC	Offices of Physicians	1,648	944	859
OTH_AM	Other Amusement and Recreation Industries	579	378	311
POST	Postal Service	37	23	25
REST	Restaurants and Other Eating Places	2,975	2,697	1,994
TRAV	Traveler Accommodation	184	335	146
	Sum	**8,356**	**6,471**	**5,403**
	Total POI included	**20,230**		

2.2 Analysis Methods

Temporal visiting patterns

Multiple approaches were used to assess the usability of the SafeGraph dataset for analyzing temporal visiting patterns at different time scales. This was achieved by pursuing three tasks: 1) comparison of monthly aggregated visitation patterns between different POI categories; 2) comparison of daily visitation patterns across different POI categories, and 3) exploration of the effects of a short-term event on visitation patterns.

For the first task, seasonal visitation patterns over a year were analyzed. Monthly aggregated visits and monthly unique visitor numbers from the Patterns dataset for each POI were used and correlated between different POI categories. POIs were filtered by category (see Table 1) and city (Miami, Orlando and Jacksonville). To smoothen monthly count data, the average of

the monthly counts from different years was used where available in our study time frame. To compare visitation patterns between POI categories, the average monthly visit and visitor count values were normalized to a range of 0 to 1 for each city and for each category, where the month with the lowest average visit or visitor number was attributed a value of 0, and 1 was assigned to the month with most visits or visitors. In the final step, Pearson correlation coefficients were calculated for each POI category pair. These correlation matrices were created for each city separately. This method of identifying temporal similarity of activity patterns through correlation has been used before, e.g. for comparing weekday and weekend cell phone communication counts (Sagl, Delmelle, & Delmelle, 2014).

Daily visit numbers from SafeGraph Patterns data were used to study visitation patterns at a more refined temporal scale. Since seasonal patterns greatly affect visitation numbers, only one month of data (May 2018) was used to analyze daily visits. This month was chosen because it does not include any national US holiday and summer break for schools has not yet started, which means that a typical visitation behavior can be expected. As for task 1, daily visit numbers were extracted and normalized for each city and POI category. Correlation matrices were computed as above, which allows the assessment of the similarity of daily visitation patterns between different types of POIs. The effects of a short-term event on POI visitation rates (task 2) were also explored using daily pattern data. As a showcase, we used Hurricane Irma, which affected the Miami metropolitan area in September 2017 through heavy wind gusts and flooding. The average daily visit per POI category between 25 August and 30 September 2017 was calculated and compared to a reference dataset for the same time period in 2018.

Distance from home

As one of its attributes, the SafeGraph monthly Patterns data includes the median distance between visitors' home locations and a POI. Median distances from June 2018 for the three cities were analyzed with respect to the 17 selected POI categories (see Table 1) in order to identify which types of POIs are associated with shorter or longer travel distances. Comparison of median distances for POIs within a city was conducted through a series of unpaired two-sample Wilcoxon rank sum tests. This type of test is a nonparametric test of the null hypothesis that the medians of two populations are equal. For this analysis, we expected that POIs from categories that provide local services for everyday activities (e.g. grocery stores, gas stations, post offices, or schools) would be closer to home locations than POIs used for recreational or travel-related activities, such as amusement parks, hotels or restaurants, which provide distinct services at specific locations and thus justify longer trips.

The same set of observations was used in a series of multiple linear regression models. Through the use of several explanatory variables, these models predict the median travel distance for a given POI category and city. The predictor variables can be subdivided into those describing spatial characteristics of other POIs in the same category, sociodemographic factors at the US census block group level, and location of a POI in the study area relative to city centers, highway access points, and major airports (Table 2). For each city–category combination, Spearman's rho correlation coefficient was computed between all candidate explanatory variables. To mitigate multicollinearity, predictor combinations with a high

correlation of $|r| > 0.7$ were avoided during the model-building process. All other predictors, even if not significant, were retained in the final models presented here.

Table 2: Candidate explanatory variables for the regression analysis

Variable	Operationalization	Data Source
POI		
Nearest neighbor (NN) distance	Distance to nearest POI in same category (in m)	SafeGraph
POI count	Number of POIs in same category within a 5-km buffer	
Sociodemographic		
Job density	Number of jobs per km² in census block group	US Census Bureau – LEHD
Population density	Population per km² in census block group	US Census Bureau
Location		
Distance to CBD	Direct distance between POI and Central Business District (in m)	http://www.city-data.com
Distance to highway	Direct distance to nearest highway access point	HERE NAVSTREETS
Distance to nearest major airport	Direct distance (in m) to Miami/Orlando/Jacksonville/Fort Lauderdale-Hollywood International Airport, whichever is closest	Natural Earth Airports

We hypothesized that, in general, a higher density of POIs in the vicinity of an analyzed POI (as operationalized in the first two variables) would facilitate short trips to a POI of that type. Census block groups with higher job densities mark areas to which people commute and hence where they perform activities away from home. Low density population areas mean less access to activity opportunities and are therefore expected to be associated with longer travel to POIs. Central Business Districts (CBDs) offer a wide range of activities for visitors from outside (tourists, business travelers), so that CBDs are expected to be associated with longer POI–home distances. Access to highways (i.e. shorter distance to highway access points) facilitates larger activity spaces for a given time budget (Parthasarathi, Hochmair, & Levinson, 2015), because of higher speed limits and the absence of intersections on highways. Proximity to highways can therefore be expected to lead to longer distances between home and POI. At least for some POI categories, e.g. gas stations or hotels, POIs near airports are more likely to be visited by air-travelers (many of whom live outside the local area) than other parts of a city, which means that POIs within short distances of airports can be expected to be associated with greater home-to-POI distances.

3 Analysis Results

3.1 Temporal Patterns

Comparison of POI categories

It was expected that different POI categories would show different visitation patterns due to the different nature of their target audiences. To analyze this, correlation matrices of monthly visit and visitor numbers for each city were computed between 17 POI categories based on the normalized values. Figure 1 plots the correlation of monthly visits between POI categories in Orlando. The main diagonal of the plot shows normalized visits between January and December as first and last data points. The lower triangular matrix plots the two normalized visit curves against each other for category pairs, while the upper triangular matrix reports the Pearson correlation coefficients along with their levels of significance. The matrix reveals that in Orlando monthly visitation to POIs generally follows the same pattern. This is indicated by most coefficients between category pairs being positive. Categories that do not follow the usual pattern are highlighted by negative correlation coefficients, in blue. An example of this is Universities, Colleges and Professional Schools, which are negatively correlated with all other categories except elementary schools, electronic stores and accounting services. This can be partially explained by the presence of the University of Central Florida (UCF) in Orlando. This institution has the largest university campus in the US in terms of enrollment, hosting more than 66,000 undergraduate and graduate students. The lower number of students present on campus during summer and winter breaks results in a deviation from the typical POI visit pattern, as indicated by the negative correlation coefficients (which, however, are mostly not significantly different from zero). Miami and Jacksonville show similar monthly visitation patterns with fewer cases of negative correlations between POI categories and some local differences. Two other statistically significant negative correlations are found in Miami between Universities and Amusement Parks, and between Schools and Amusement Parks.

This analysis can be complemented by calculating the same matrix for the number of monthly unique visitors (not shown in the Figures). This would allow the joint interpretation of visit and visitor matrices in order to gain more insights into urban dynamics.

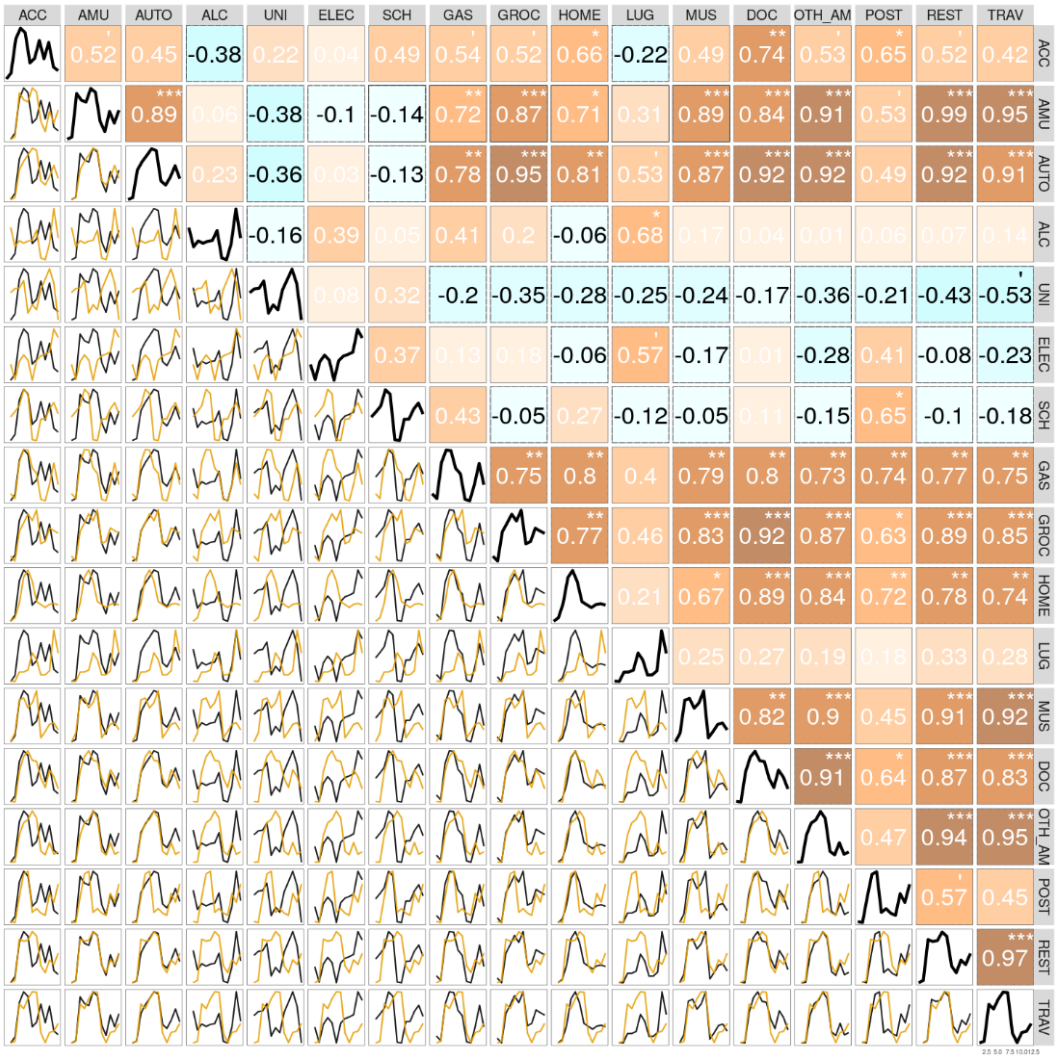

Figure 1: Correlation of monthly visits between selected POI categories in Orlando. (Significance codes: 0 ***, <0.001 **, <0.01 *, <0.05 '). Abbreviations of category labels are explained in Table 1.

Figure 2 shows the correlation plot of daily visits between POI categories in May 2019 for Orlando. Not surprisingly, among other things the plot reveals the negative relationship between visits to universities and amusement parks and hotels (Traveler accommodation). This is because universities are typically attended during weekdays, whereas entertainment facilities tend to be more visited on weekends. Outliers caused by local short-term events (e.g. storm, spike in gas price or athletic games) can influence the results of similar analyses by obfuscating real relationships.

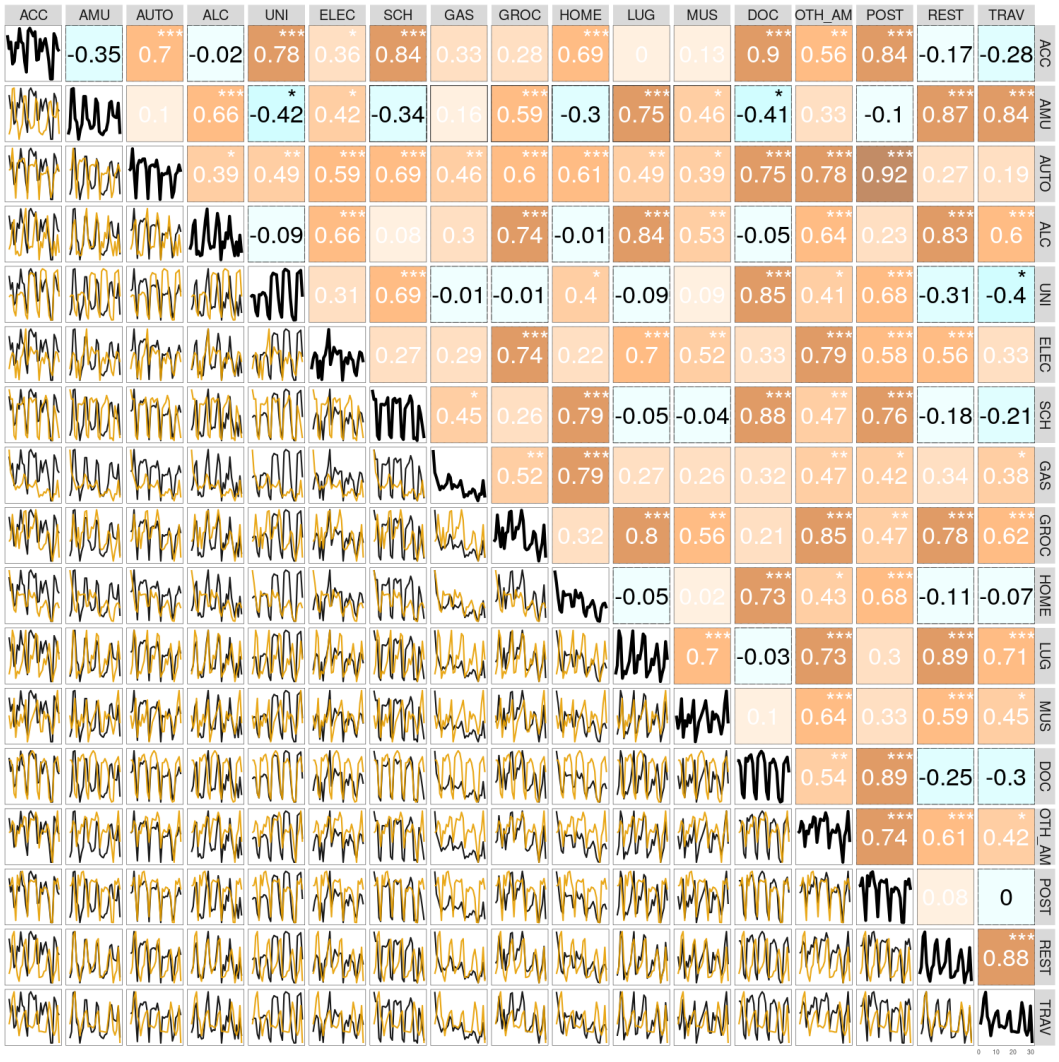

Figure 2: Correlation of daily unique visitors between selected POI categories in Orlando. (Significance codes: 0 ***, <0.001 **, <0.01 *, <0.05 '). Abbreviations of category labels are explained in Table 1.

Event analysis

The effect of short-term events on visitation patterns was explored in more detail by analyzing the effect of Hurricane Irma on the Miami metropolitan area. Hurricane Irma was the most intense hurricane to make landfall in contiguous United States since Hurricane Katrina in 2005. The hurricane had the potential to hit the Miami metropolitan area directly. However, it eventually struck the Florida Keys on September 10, 2017, less than 200 km from Miami. The effects of this hurricane are clearly visible in Figure 3a, which plots the average number of daily visits per POI in a given category between August 25 and September 30, 2017. In the shaded areas, Figure 3a also shows the 95% confidence interval. Yellow vertical bars denote

weekends; the time between September 9 and 11 when the hurricane was closest to Miami is highlighted with an orange vertical bar. On September 7, Miami-Dade County expanded the mandatory evacuation order for residents in all evacuation zones, which affected more than 650,000 people (red vertical line in Figure 3a). The figure shows increased visits to gas stations and grocery stores before the event, which is the usual behavior when a storm is projected to hit an area as residents stock up with non-perishable food and fill up their vehicles with gas. The data also reflects the response of higher educational institutes in the area, as Florida International University, the University of Miami and Miami-Dade College all canceled classes on September 6, with Miami-Dade College remaining closed for the remainder of the week. After the evacuation order, visits to grocery stores and gas stations decreased rapidly. All categories reached their minimum visitation rates on the day of the landfall (September 10). After the storm, visits to grocery stores and gas stations increased more rapidly than visits to universities and colleges. The data shows that it took 4–5 days to reach pre-storm visitation levels in gas stations and grocery stores, while universities and colleges did not resume normal operations until the following week.

In order to compare this pattern to an unaffected time period, average visits per POI in Miami were calculated for the same period in 2018. Slightly different dates (24 August – 29 September, 2017) were used for a pairwise comparison in order to match the days of the week between the two periods (Figure 3b). The difference between the average number of visits per POI in 2018 and 2017 (i.e., year 2018 visits minus year 2017 visits) is shown in Figure 3c. Positive values show an increase in the average number of visits in 2018 compared to the previous year, while negative values mean the opposite. Since the hurricane peaked on a weekend, the observed decrease in visit numbers during the hurricane is smallest for universities (Figure 3c), since these institutions do not generally host many activities on weekends.

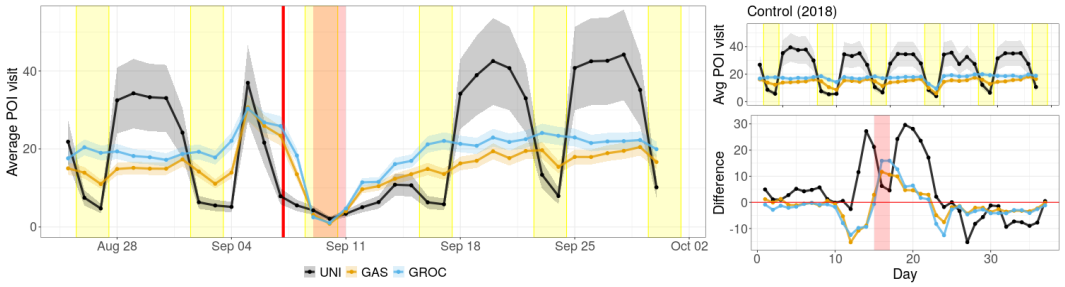

Figure 3: The effect of Hurricane Irma on visitation patterns in three POI categories in Miami, illustrated by (a) the average daily visits per POI in 2017; (b) the average daily visits per POI during the control time period in 2018; (c) the difference in the average daily visits between the control and hurricane periods

3.2 Distance from Home

Spatial and temporal patterns

Figure 4 maps the median distance from home to grocery store (MD = 7.1 km) and restaurant (MD = 9.7 km) POIs in the Jacksonville study area for June 2019. Yellow circles (restaurants)

tend on average to be larger than red ones (grocery stores), suggesting that grocery shopping is a more localized activity than visiting restaurants.

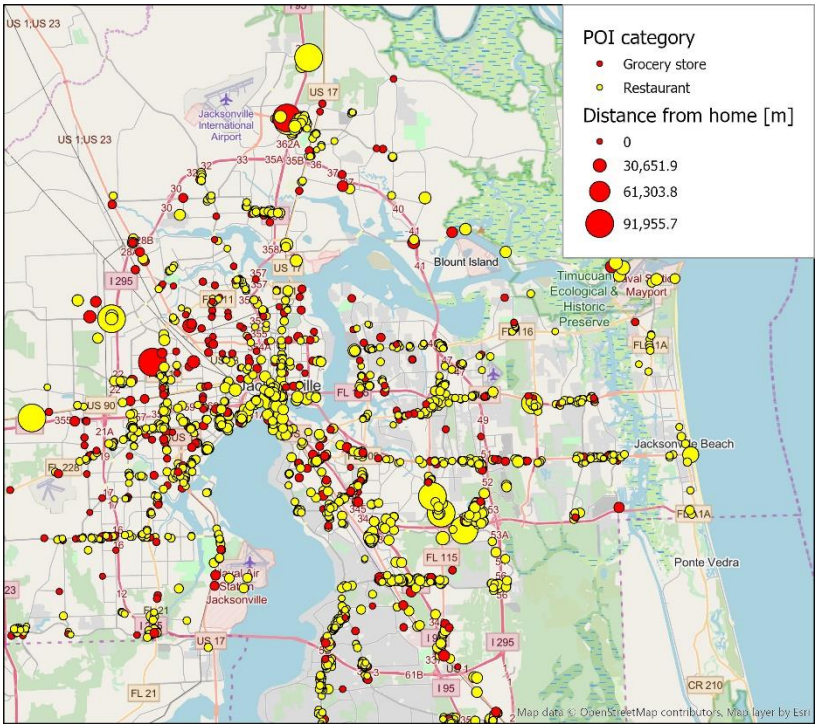

Figure 4: Median distance from home to grocery stores and restaurants for June 2019 POI visits in the Jacksonville study area

Flow maps (Figure 5) were generated by counting the number of residents in US census block groups who visited three Orlando theme parks (Animal Kingdom, Universal Studios, and SeaWorld) during August 2018 and December 2018, based on SafeGraph aggregated Pattern counts. Map comparison reveals a change in visitor patterns towards more people from the south-east and mid-west traveling to Florida in December, possibly to enjoy milder temperatures during the winter season.

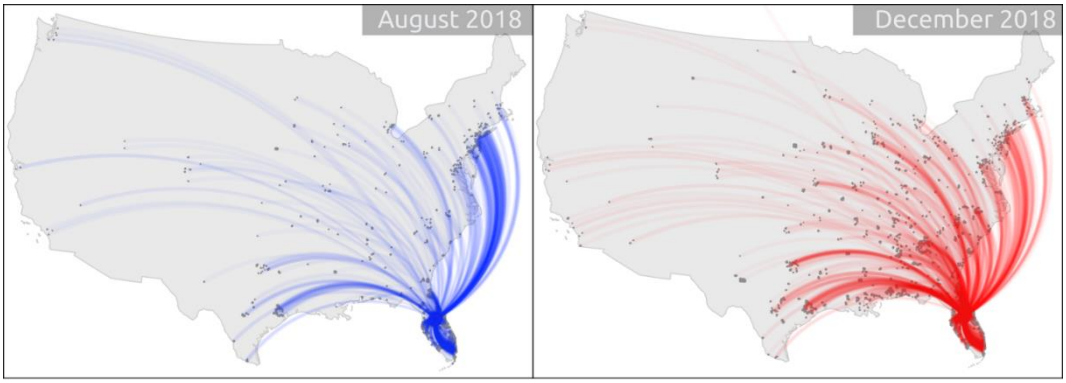

Figure 5: Home locations of Orlando theme park visitors in August and December 2018 at census block group level.

Distances between home and POI vary significantly over the year for some POI categories. As an example, monthly median distances between home and POI visits to Orlando amusement parks and traveler accommodations (hotels) between July 2018 and June 2019 are plotted in Figure 6. Distances to amusement parks (Figure 6a) peak in March, a month which includes the spring break, during which Florida is a traditional travel destination. Smaller peaks can be found in June and July (US holidays) and November (which includes Thanksgiving), all of which appear to contribute to traveling to Orlando theme parks from more distant locations than during the off-season. Distances from home to travel accommodation (Figure 6b) also peak in March, but are generally higher during the winter season (December through April), suggesting more visits from residents outside the Florida region during that time. Longer distances are also observed in June and July, when school vacation allows for extended family vacations and longer-distance travel.

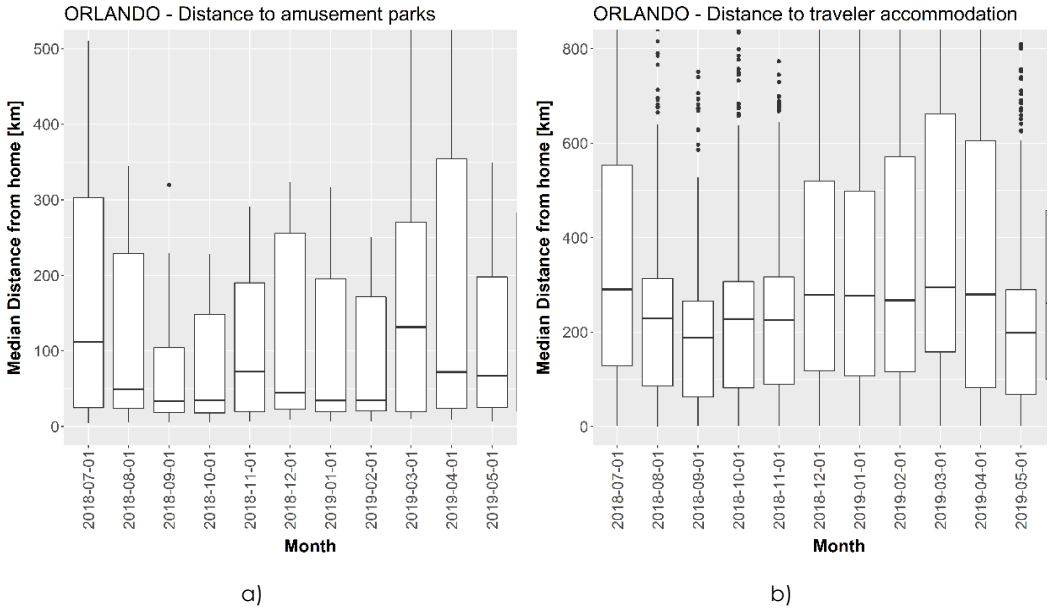

Figure 6: Box plots of median distances, over 12 months, from home to (a) Orlando amusement parks, GI2020GI

Variation across POI categories

Figure 7 shows box plots of distances from home to POIs for the 17 top POI categories analyzed for (a) Miami, (b) Orlando, and (c) Jacksonville for June 2018. In all three cities, travel accommodation (hotels) attracts those whose homes are farthest away, which can be expected since locals use hotels less frequently than people from outside the region. Another noticeable pattern is greater travel distances for Orlando amusement parks compared to those of Miami and Jacksonville. This can be attributed to the national and worldwide popularity of Orlando's theme parks, whereas Miami is known, rather, for its beaches and nightlife, and Jacksonville is a center for healthcare, retail, marine transportation, and finance. All other distance medians associated with different POI categories vary within only a small range – of about 5 to 15 km across each city.

Wilcoxon rank sum tests, with the Bonferroni correction used for multiple testing, were applied to identify which POI category / distance from home to POI pairs differ at a 5% level of significance. In all three cities, distance to travel accommodation is significantly greater than for all other categories. In Miami, median distances were lowest for postal service (4.7 km), grocery (4.9 km), liquor store (5.4 km), school (5.5 km), and gas station (5.6 km). There was no significant difference between these categories, but the distances were significantly shorter than those of all or most other categories (e.g. physician or museum). This shows that local services allow shorter trips. Orlando reveals a similar pattern, where schools (MD = 7.2 km) have the shortest distance, probably because this type of POI is not affected by tourists who travel from further away. In Jacksonville, distances for visits to grocery stores are shortest (MD = 7.1 km), followed by liquor stores (MD = 7.8 km) and gas stations (MD = 8.4 km); grocery store distances are significantly shorter than distances to all other POI categories.

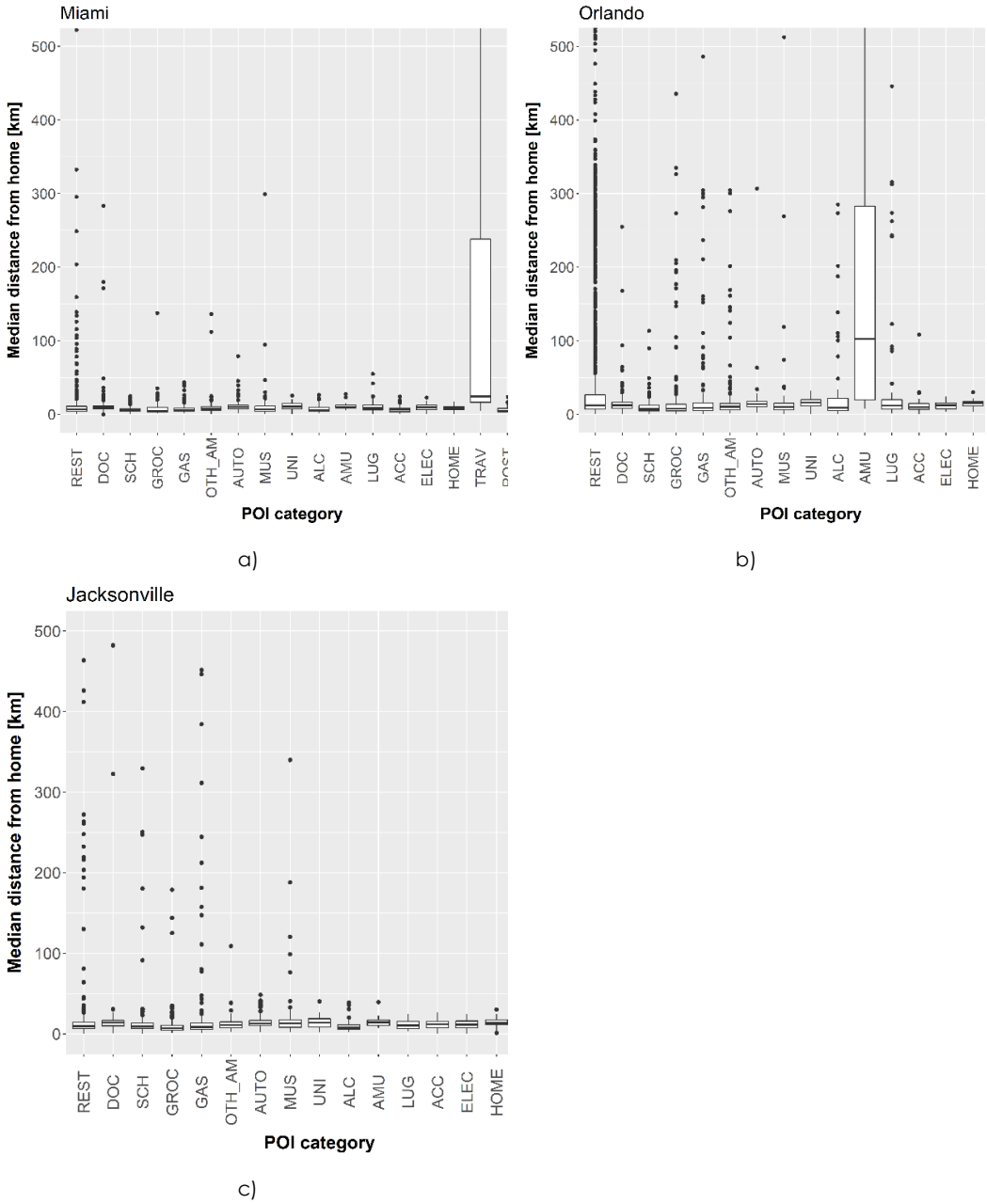

Figure 7: Scatterplots of distance from home for different POI categories in (a) Miami, (b) Orlando, and (c) Jacksonville

Regression analysis

Using the set of seven candidate explanatory variables listed in Table 2, linear regression models that predict the distance between POI and home census block group were constructed for each POI category in each of the three cities. Not every model resulted in significant coefficients. Some of the results appear to be affected by local characteristics of the layout of a city, whereas other findings hold across all three cities analyzed. **Fehler! Verweisquelle konnte nicht gefunden werden.** shows the model results for three different POI categories in each of the three cities. Models shown in the table were selected based on the explanatory power (adjusted R-square) and interpretability and usefulness of regression coefficients. Each model description comes with the number of observations (N) and the adjusted R-square value. Blank cells indicate predictors removed due to collinearity. POIs for everyday services (gas stations, grocery stores) showed the expected effect of POI abundance within a 5-km radius, namely a decrease in distance between home and POI. This suggests that customers of these POIs tend to choose a facility near their home, and that longer trips to another POI which provides similar services would not be justified. For POIs associated with more internal variability (restaurants, museums / historical sites / nature parks), the opposite is the case. This suggests that visitors from further away, such as tourists, tend to visit areas with a higher abundance and more clustering of such facilities. The positive coefficient of POI NN distance for Miami museums and historical sites can be explained by the long driving distances required to get to state and national parks (e.g. Shark Valley Visitor Center) located outside the city boundaries. Arithmetic signs of coefficients for job density (positive) and population density (negative) are as expected across the three cities where significant. Proximity to highway access points was associated with greater travel distances, as expected, for grocery and jewelry stores. However, gas stations further away from highways had greater travel distances, possibly due to generally lower gas prices further away from highways. Longer trips to museums/historical sites away from highways, as shown for Miami, are likely explained by the remoteness of parks from highways (mentioned above). While, as expected, POI proximity to airports (i.e. shorter distance) is associated with longer distances from home (due to visitors coming from out of the state), this is not the case for Orlando restaurants. One reason could be that many restaurants are clustered around the Orlando theme parks, which are themselves located 20 or more km away from the airport. These restaurants, often visited by tourists from far away, could explain the positive regression coefficient for airport distance. The same explanation could hold for the positive signs associated with CBD distance for Orlando restaurants and hotels, since theme parks are between 15 and 30 km from the CBD.

Table 3: Coefficient estimation results for selected OLS models (Significance codes: *** p<0.001, ** p<0.01, * p<0.05)

Miami	*Gas station*	*Jewelry store*	*Museum/historical site*
POI NN distance	-0.589	-1.15	**6.52E+01*** **
# POI within 5 km	**-9.15E+01*** **	**8.46E01*** **	**3.44E+03*** **
Job density	-5.73E-03	-1.68E-03	1.19
Population density	**-0.254** **	**-0.597*** **	-2.38
Highway distance	**0.529** **	-0.392	**4.64E+01*** **
CBD distance			
Airport distance	**-0.338*** **	5.85E-02	-1.33
Constant	**1.45E+04*** **	**7.48E0*** **	**-2.29E+05*** **
N (Adj. R^2)	347 (0.132)	167 (0.313)	151 (0.650)

Orlando	*Grocery*	*Restaurant*	*Travel accommodation*
POI NN distance	1.24E01	**-2.47E+01** **	**-4.34E+01*** **
# POI within 5 km	**-8.00E+02** **	**4.43E+02*** **	4.09E+02
Job density	0.179	-2.11E-03	1.948
Population density	-2.49	-0.429	-2.79
Highway distance	**-1.32E+01*** **	0.780	3.62E+01
CBD distance		**1.66E+01*** **	**2.35E+02*** **
Airport distance	1.28	**8.64*** **	-2.16
Constant	**6.28E+05** **	**-3.56E+05**	-2.76E+05
N (Adj. R^2)	353 (0.048)	2302 (0.304)	299 (0.227)

Jacksonville	*Gas station*	*Jewelry store*	*Car dealer*
POI NN distance	-2.42	**-1.49***	0.270
# POI within 5 km	**-3.48E+02** **	**2.07E+02***	-1.45E+01
Job density	**0.359** **	3.43E-02	0.782
Population density	**-3.81***	7.16E-03	**-1.15***
Highway distance	-1.06	**-1.21** **	-0.136
CBD distance		-3.76E-02	**0.430*** **
Airport distance	**-0.563** **	5.43E-02	**-0.219** **
Constant	**4.15E+05*** **	**1.06E+04*** **	**1.52E+04*** **
N (Adj. R^2)	320 (0.083)	67 (0.322)	152 (0.134)

4 Conclusions

This study analyzed spatial and temporal visitation patterns of POIs in Miami, Orlando and Jacksonville, using SafeGraph POI and Patterns data. It extends the growing body of literature (Andersen, 2020; Gao et al., 2019; Killeen et al., 2020; Prestby et al., 2019) using this data source and provides insights into potential use cases of the data.

Correlation matrices based on aggregated monthly and daily visitation counts identified POI category pairs of similar or dissimilar temporal activity patterns. This type of analysis can be used to reveal whether visitation patterns to a certain POI category deviate from the rest. As an example, it was shown that the visitation pattern to universities in Orlando was different from other categories in the city due to the absence of students during breaks. The ability to analyze the effects of short-term events on visitation patterns was demonstrated by using Hurricane Irma as a case study and by comparing the average number of visits to selected POI categories to a control period in Miami. In the future, this analysis could be extended to measure changes in visitation patterns caused by policy changes or other types of events, such as sporting events or disease outbreaks.

The research demonstrated that the dataset analyzed is a viable source of information for several analysis tasks, although only check-in data but not travel trajectory data is provided. The latter is a limitation compared to other freely available sources, such as tweets, which can be used to derive travel trajectories (Han, Tsou, Knaap, Rey, & Cao, 2019). The methodology presented here, such as co-interpreting visit and visitor correlation matrices, or determining travel distances to POIs for different categories, has the potential to aid urban planners and city managers to better understand the dynamics of a city and to complement data from local visitor surveys on travel and visitation behavior.

Acknowledgements

The authors thank SafeGraph Inc. (www.safegraph.com) for providing free access to Florida POI data, including Core Places, Geometry, and Patterns tables, for this research.

References

Andersen, M. (2020). Early Evidence on Social Distancing in Response to COVID-19 in the United States. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3569368

Baik, J., Lee, K., Lee, S., Kim, Y., & Choi, J. (2016). Predicting personality traits related to consumer behavior using SNS analysis. *New Review of Hypermedia and Multimedia, 22*(3), 189-206.

Gao, S., Li, M., Liang, Y., Marks, J., Kang, Y., & Li, M. (2019). Predicting the spatiotemporal legality of on-street parking using open data and machine learning. *Annals of GIS, 25*(4), 299–312.

Gurun, U. G., Nickerson, J., & Solomon, D. H. (2020). The Perils of Private Provision of Public Goods. Retrieved from http://www.umitgurun.com/wp-content/uploads/2019/12/PublicProvision_GNS.pdf

Han, S. Y., Tsou, M.-H., Knaap, E., Rey, S., & Cao, G. (2019). How Do Cities Flow in an Emergency? Tracing Human Mobility Patterns during a Natural Disaster with Big Data and Geospatial Data Science. *Urban Science, 3*(2), 3020051. doi:10.3390/urbansci3020051

Hochmair, H. H., Juhász, L., & Cvetojevic, S. (2018). Data Quality of Points of Interest in Selected Mapping and Social Media Platforms. In P. Kiefer, H. Huang, N. Van de Weghe, & M. Raubal (Eds.), *Progress in Location Based Services 2018* (Vol. Lecture Notes in Geoinformation and Cartography, pp. 293-313). Springer: Berlin.

Juhász, L., & Hochmair, H. H. (2017). Where to catch 'em all? – A geographic analysis of Pokémon Go locations. *Geo-spatial Information Science, 20*(3), 241-251.

Juhász, L., Novack, T., Hochmair, H. H., & Qiao, S. (2020). Cartographic Vandalism in the Era of Geo-Gaming -The Case of OpenStreetMap and Pokémon GO. *ISPRS International Journal of Geo-Information, 9*(4), 197.

Killeen, B. D., Wu, J. Y., Shah, K., Zapaishchykova, A., Nikutta, P., Tamhane, A., . . . Unberath, M. (2020). A County-level Dataset for Informing the United States' Response to COVID-19. Retrieved from https://arxiv.org/abs/2004.00756

Massimo, D., & Ricci, F. (2019). Clustering Users' POIs Visit Trajectories for Next-POI Recommendation. In J. Pesonen & J. Neidhardt (Eds.), *Information and Communication Technologies in Tourism 2019* (pp. 3-14): Springer.

Parthasarathi, P., Hochmair, H. H., & Levinson, D. M. (2015). Street Network Structure and Household Activity Spaces. *Urban Studies, 52*(6), 1090-1112.

Prestby, T., App, J., Kang, Y., & Gao, S. (2019). Understanding neighborhood isolation through spatial interaction network analysis using location big data *Environment and Planning A*.

SafeGraph. (2020). Places Schema. Retrieved from https://docs.safegraph.com/docs/places-schema

Sagl, G., Delmelle, E., & Delmelle, E. (2014). Mapping collective human activity in an urban environment based on mobile phone data. *Cartography and Geographic Information Science, 41*(3), 272-285. doi:10.1080/15230406.2014.888958

Spyratos, S., Stathakis, D., Lutz, M., & Tsinaraki, C. (2017). Using Foursquare place data for estimating building block use. *Environment and Planning B, Planning and Design, 44*(4), 693–717.

Zhuang, Y., Fong, S., Yuan, M., Sung, Y., Cho, K., & Wong, R. K. (2017). Location-based big data analytics for guessing the next Foursquare check-ins. *The Journal of Supercomputing, 73*, 3112-3127.

GI_Forum 2020, Issue 1
Page: 137 - 152
Full Paper
Corresponding Author:
andreas.petutschnig@sbg.ac.at
DOI: 10.1553/giscience2020_01_s137

Exploratory Spatiotemporal Language Analysis of Geo-Social Network Data for Identifying Movements of Refugees

Andreas Petutschnig[1], Clemens Rudolf Havas[1], Bernd Resch[1,3], Veronika Krieger[1] and Cornelia Ferner[2]

[1]University of Salzburg, Austria
[2]Salzburg University of Applied Sciences, Austria
[3]Harvard University, USA

Abstract

Refugee movements in recent years have caused enormous challenges for relief organizations and public authorities, but especially for refugees themselves. Organizations which have to allocate their resources to regions where large groups of arrivals are expected struggle to prepare the refugees' admission, transfer, care and accommodation in time. Events like the refugee movement of 2015/16 in Austria and Germany in the wake of the Syrian civil war have shown that many of these issues are caused by a lack of up-to-date information about logistical requirements. We evaluate various methods to acquire this information that utilize semantic, spatial and temporal features to analyse geo-social network data. A multimodal analysis of these features leads to information about refugee movements across borders and regions. Approaches based on user trajectories and attempts to identify refugees by the language they used showed little promise, whereas using spatiotemporal aggregation and hotspot analysis of keyword-based filtered data allowed us to retrace refugees' collective movement patterns. Using temporal bins, we were able to detect changes in these patterns caused by external factors such as border closures.

Keywords:

language, refugees, social media, GSND, ESDA

1 Introduction

The phenomenon of refugee movement is inherently geographical (Lewis, 1982), but it has also been studied from a variety of other viewpoints, focusing on the causes of flight (Warner, 2009; Black et al., 2011; Mueller et al., 2014), effects (Jacobsen, 1997; Garfi et al., 2009; Biswas & Tortajada-Quiroz, 1996), and demographic aspects (Randall, 2005; Greenwood, 1997). Because the phenomenon is so multifaceted, it warrants an analysis which incorporates multiple modalities.

The event under investigation is the refugee movement in 2015/16 during which over 2.5 million people[1] fled to Europe from war-torn countries, mostly in northern Africa and western Asia. During that time, it was largely unknown when, where and how many refugees would appear at European borders, which made planning the necessary distribution of goods and personnel challenging for relief organizations and public authorities. This led to authorities at many border regions being overwhelmed by the necessary logistics at such short notice, which in turn led to humanitarian and societal problems (Razum & Bozorgmehr, 2015; Breen, 2016).

One way to mitigate these problems is to provide the information needed for resource planning based on an up-to-date picture of the situation, which in turn necessitates comprehensive, near real-time information. Information about refugee movements, especially in regions around borders, derived from new data sources like social networks, news outlets or crowdsourcing platforms, fits these requirements because of the potentially high spatial and temporal resolution of the data.

The methods and findings presented in this paper are a contribution towards building a system that consolidates multiple data sources, processes the data, and provides relevant up-to-date information to relief organizations. For this approach to be useful requires the ability to extract information about refugee movements from a given set of geo-social network data (GSND). We outline methods to explore what information we can derive from such data, and present a multimodal analysis approach in which we collected georeferenced posts from the social media platform Twitter and analysed their spatial, temporal and semantic characteristics.

We begin by defining and checking the assumptions based on which we perform the analysis. For example, we need to check whether there are enough data available to make reasonable predictions at all. We further identify the languages used in the text data to examine whether Twitter users in Arabic-speaking countries actually use Arabic as their language of communication. This is critical because we use Arabic-language data as a proxy for populations originating from Arabic-speaking countries, who are the focus of this study. As the text corpus also contains other languages, mostly English, we include a selection of words from other languages in the keyword-based analysis as well. Using this setup, we aim to understand the potential for, and limitations of, detecting collective refugee movement patterns in a multi-disciplinary approach, extracting and combining information from geographic, temporal and semantic space. We also describe the measures we employed to preserve the privacy of individuals represented in the data.

2 Related Work

Besides food and shelter, smartphones are one of the essentials for refugees (Matthew, 2015) on their way to their destinations. Smartphones allow refugees to connect to social media networks where they can gather information and share their own experiences within their network. Refugees connect to various social media networks such as Instagram or the now

1 https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tps00191&plugin=1

defunct Google+, but Facebook and Twitter are the most popular networks in this user group. The information obtained from social media networks is used by refugees for decision-making on their way to their destinations or for planning where to settle (Dekker et al., 2018). Further, they also use smartphones to plan and document their journeys, and to contact friends, family and people who will help them to get to their desired destinations. Although they are useful, smartphones also hold many dangers for refugees, such as surveillance by other actors. Therefore, refugees frequently use encrypted or closed communication channels, like closed Facebook groups or encrypted WhatsApp messages (Gillespie et al., 2018). Although refugees must use social media networks carefully, they are the main resource for communication, which makes them a generally suitable data source for data analysis in the context of refugee movements. This allows us to lean on the principles of collective sensing as described by Resch (2013) for data collection and analysis.

Many challenges associated with the recent refugee movements in Europe have been pointed out, including humanitarian protection (Ostrand, 2015), policy making (Guild et al., 2015), public health systems (Catchpole & Coulombier, 2015) or news coverage (Chouliaraki & Zaborowski, 2017). One of the factors leading to such problems is the lack of real-time information about refugee movements. Curry et al. (2019) discuss the potential and challenges of alternative data sources such as Volunteered Geographic Information or social media data. They present multiple results from analysing data from social media networks such as Flickr or Instagram, which they used to detect refugee-related activities in Europe. They conclude that social media are a crucial data source for information associated with mass migration and can help fill the information gap between authoritative information products and the actual situation. Hübl et al. (2017) performed an exploratory spatial data analysis (ESDA) on a Twitter dataset to demonstrate the value of GSND analysis for refugee movements. In the context of the refugee crisis in 2015, they detected and visualized generalized trajectories from the Middle East and northern Africa to Europe. In order to filter their dataset, they used language-dependent keywords for German, Italian and Greek. They were able to identify only a few potential refugees as Twitter users, and consequently only a limited number of trajectories.

In many cases, the triggers for the refugee movements we are observing here were terror and war. Consequently, ethical and responsible handling of data, and the presentation of results in such a way as to protect the research subjects are paramount, as most refugees fear surveillance and surveillance by others (Gillespie et al., 2016). Kounadi & Resch (2018) give an overview of potential privacy threats that come with GSND analysis, while Kounadi et al. (2018) have drawn up a set of practical guidelines and design principles for researchers who work with GSND to mitigate these threats.

3 Methods

The original dataset consists of 354,116,330 georeferenced tweets in the area between 8.0°E - 43.2°E and 28.2°N - 50.0°N, covering the period January 2014 to January 2020. Spatially and temporally the data cover the 'Balkan route', an informal land route connecting Greece and central Europe, which many refugees took during the refugee movement in 2015/16. The time frame of the dataset exceeds the actual event in order to provide a baseline before and after

the event as context for interpretation. All our Twitter data are georeferenced (i.e., each tweet is linked to a geographic coordinate). This geolocation is only possible if the user chose to make their location public before tweeting. No additional georeferencing was carried out, because this would have introduced a bias. With this reliable reference in place, we can observe the number of tweets, the proportions of tweets using the different languages for each country, as well as the shift of languages or of relevant tweets in time. Besides coordinates, the data we used include a numeric user ID generated by Twitter, a timestamp, and the text of the message.

To be able to draw meaningful conclusions from the data, we used different methods for our analysis. We aggregated the Twitter data such that the data were grouped semantically by language, spatially by hexagonal bin or by the country from which the tweet was sent, and temporally using weekly bins for the timestamps. The aggregated data then served as input for further analysis. Language detection was carried out by applying the language detection module of the Python library Polyglot[2] on the tweets' texts. Polyglot can detect up to 196 languages, including Arabic, modern Greek, Turkish, German and English, which are the predominant languages used in tweets from the area between the Near East and Central Europe. The result of the language detection for each tweet is a probability distribution of the most probable languages used in the text. Due to a tweet's short length, there is usually just one language that exhibits dominant probability, and this is the language which we assigned to the tweet.

To learn about the spatial and temporal distribution of the data, we counted the yearly number of tweets within each country. This step was necessary to determine which assumptions about the data we could make in our analysis. Similarly, we counted the number of tweets grouped by language in each country. This way we could determine the composition of languages used within each country and whether this composition changes over time, which may in turn reflect changes in the population. As most of the refugees originated from predominantly Arabic-speaking countries, we designed part of our study to focus on the movement patterns of Arabic-speaking Twitter users. Our approach was similar to the one used by Bulbul et al. (2018) to identify Arabic-speaking refugees in Turkey. For this, we counted the number of Arabic tweets grouped by country and aggregated them in weekly bins. The resulting time series show signatures of changing use of the Arabic language, and therefore, by proxy, changes in the dynamics of the local Arabic-speaking population. For the keyword-based part of the study, we used refugee-related keywords. The study area covers a large region, which includes a number of language borders. In order to take into account the linguistic diversity in the Twitter data, we defined keywords in a set of locally used languages, which were defined and translated by language experts or native speakers who manually examined the text contents. The keywords are listed in Table 1. We matched tweets and keywords based on case-insensitive approximate string matching in the tweets' texts. We prepared the keyword-matched tweets for analysis by grouping them in hexagonal weekly bins and generating a series of maps for visual interpretation. The proportion of keyword-relevant tweets was 0.47% for German, 0.14% for English, 0.14% for Hungarian, 0.33% for Serbian or Croatian, 0.25% for Greek, and 0.58% for Arabic.

2 http://polyglot.readthedocs.org

Table 1: Keyword overview

German	English	Hungarian	Serbian or Croatian	Greek	Arabic
Flüchtling	refugee	migráns	Sirija	πρόσφυγες	لاجئ
Fluechtling	migra	migrans	Сирија	προσφύγων	لاجئين
Flucht	syria	Szíria	izbeglice	πρόσφυγας	لاجئون
Migrant	border*cross	sziria	izbjeglice	μετανάστης	مهاجر
Syrien	cross*border	Soros	избеглице	μεταναστες	مهجر
Grenz		StopSoros	migranti	μετανάστες	مهجرون
Asyl		brüsszel	мигранти	Συρία	مهجرين
unbegleitete Minderjährige		bruesszel	migracije	Βρυξέλλες	لاجئة
unbegleitete Minderjaehrige		Brusszel	миграције	Prosfuga	لاجئات
Schlepper		O1G	азил	Prosfiga	مهاجرة
Willkommenskultur		kvóta	granica	Prosfyga	مهجرة
Erstaufnahmezentrum		kvota	граница	Prosfuges	مهجرات
Balkanroute		fidesz	Brisel	Prosfiges	سوريا
		bevandorlas	Брисел	Prosfyges	لاجئ
		bevándorlás	Батровци	Prosfugwn	لاجئين
		keritesepites	Batrovci	Prosfigwn	لاجئون
		kerítésépítés	Хоргош	Prosfygwn	ملجأ
		ahataron	Horgoš	Metanastis	مأوى
		ahatáron	azilanti	Metanastths	مهاجر
		migráncs		Metanastes	مهجر
		bevandored		Metanastwn	مهجرون
		menekült		Brixelles	مهجرين
		dzsihadista		Bruxelles	حدود
		dzsihadistak		Synora	حدودية طة
		terrorista		Sunora	حدودي
		határainkat		Πρόσφυγ	معبر
		hatarainkat		Προσφύγ	لجوء
		védyük meg		Μετανάστ	أزمة
		muszlim		Μεταναστ	السورية زمة
		muzulman		Prosfug	السورية حرب
				Prosfig	البلقان ق
				Prosfyg	مخيم
				Metanast	الركبان
					الزعتري

4 Results

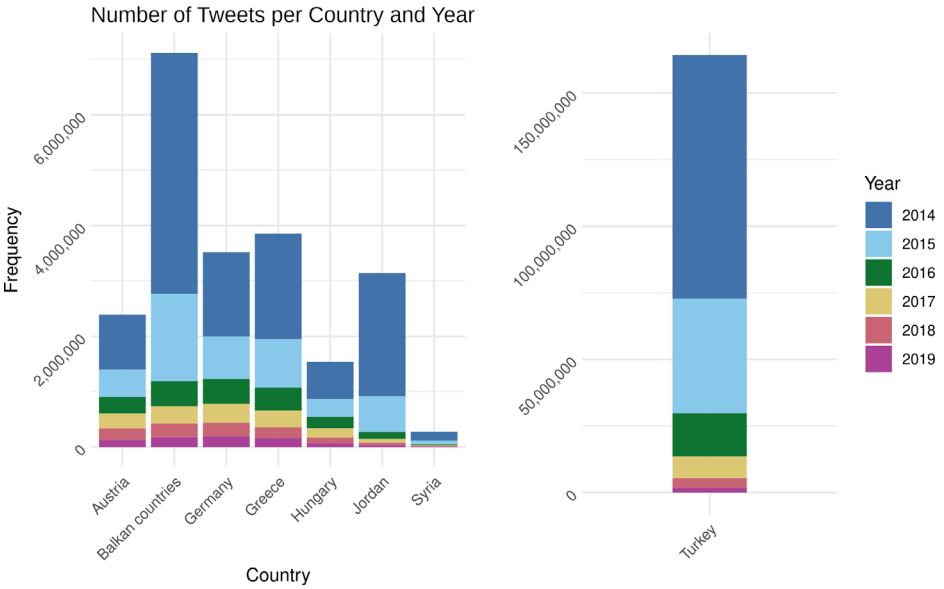

Figure 1: Number of tweets per country and year

The focus in this analysis is on the countries through which people passed on their way from Syria to Germany. Because the Balkan countries of Albania, Bosnia and Herzegovina, Croatia, Kosovo, North Macedonia, Montenegro, Serbia and Slovenia individually gave very little data, we aggregated them in some results under the term 'Balkan countries', for readability. The results show that in general the number of tweets collected is higher in 2014/15 than in later years. This is not necessarily attributable to changes in user behaviour: it may be affected by Twitter's data-sharing policy and the data collection process. However, one can see a change in the pattern for the different countries in Figure 1. For better readability, the chart for Turkey is presented separately, as significantly more tweets were collected in Turkey than in the other countries.

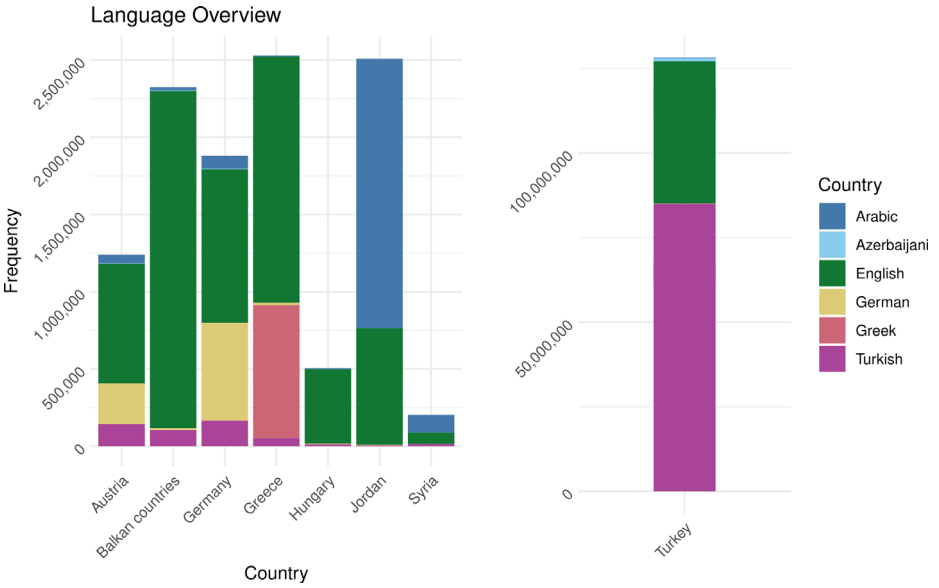

Figure 2: Number of tweets per language and country for 2014–2020

Figure 2 shows the number of tweets in Arabic, Azerbaijani, English, German, Greek and Turkish, which are overall the most dominant languages across the listed countries. All countries contain Arabic tweets, but in Turkey, Greece and Hungary their numbers are marginal.

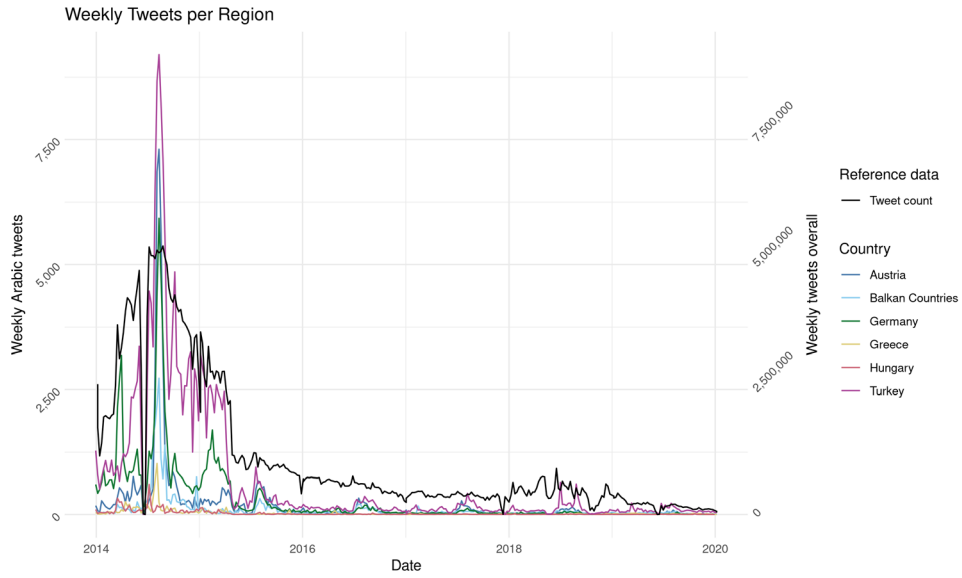

Figure 3: Time series of Arabic language tweets and overall tweet counts from 2014 to 2020 per country (note the secondary y-axis for total tweet counts)

Figure 3 shows a development of the number of Arabic tweets between 2014 and 2016. The peaks in the data mostly appear simultaneously in a seemingly seasonal interval throughout the countries with few exceptions. The grey line shows the total number of weekly tweets, providing an indication for how strongly the peaks deviate from the baseline. Within the group of the Balkan countries, the temporal signatures appear to be similar, but substantiated conclusions are difficult due to the relatively sparse data.

The next method we used aimed at the identification of trajectories to derive routes that are shared by larger numbers of refugees. A trajectory consists of a minimum of two tweets from the same user ID. To identify trajectories that meet our criteria for country of origin and time frame, we filtered our dataset using spatial and temporal constraints. We removed bots from the dataset by restricting the average number of tweets per day to 15, which we defined based on empirical evidence. We defined the start of the trajectory to be east of Greece, and the average distance between two tweets was chosen with the aim of eliminating 'extreme' values – i.e. to filter out small-scale as well as unrealistically long distances, such as tourists using planes to reach their destinations. The remaining data were assessed manually. The resulting trajectories during the major refugee movement period are shown in Figure 4. Visual analysis of the results did not yield any usable results because of the large number of trajectories with very low sampling rate. In addition, upon visual inspection of the text contents, we found that none of the trajectories were likely to belong to a refugee, but rather to either previously undetected bots or users such as tourists, business people, or journalists, who were not the focus of this study.

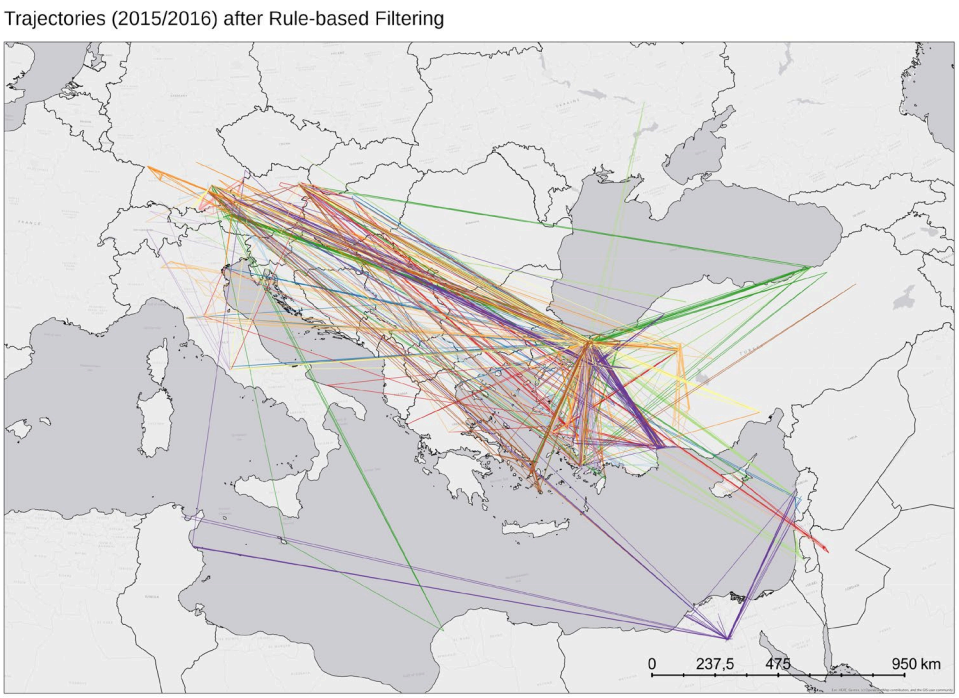

Figure 4: Coloured trajectories differentiating between individual users

Another possibility to detect refugee movements is to visualize the number of refugee-related tweets in an area of interest. The hypothesis is that if an unusually high number of refugees are passing through an area, people will refer to this on social media networks. We used the refugee-related keywords from Table 1 to identify relevant tweets and aggregated them spatially in hexagonal cells in the area of interest. We also binned the data temporally by week to detect temporal changes.

We observed that at the beginning of 2015 the refugee-related tweets were mostly sent from near the Greek coast. Later, more activity was observed at the border between Turkey and Greece, as shown in Figure 5. Figure 6 shows how, over time, the tweet frequency increased at the Greek and Bulgarian borders with Turkey. From August onwards, the number of refugee-related tweets rose in Serbia and Hungary. Figure 7 shows that between Belgrade through Hungary and Austria to Munich, the number of refugee-related tweets was high, which is in line with the actual path the refugees used as the primary route to central Europe during this specific time frame.[3] Furthermore, the tweet activity along the coast of Turkey and at the border with Greece is still high, comparable to the preceding months. This observation concords with articles that report consistent numbers of refugees arriving in Greece in 2015.[4]

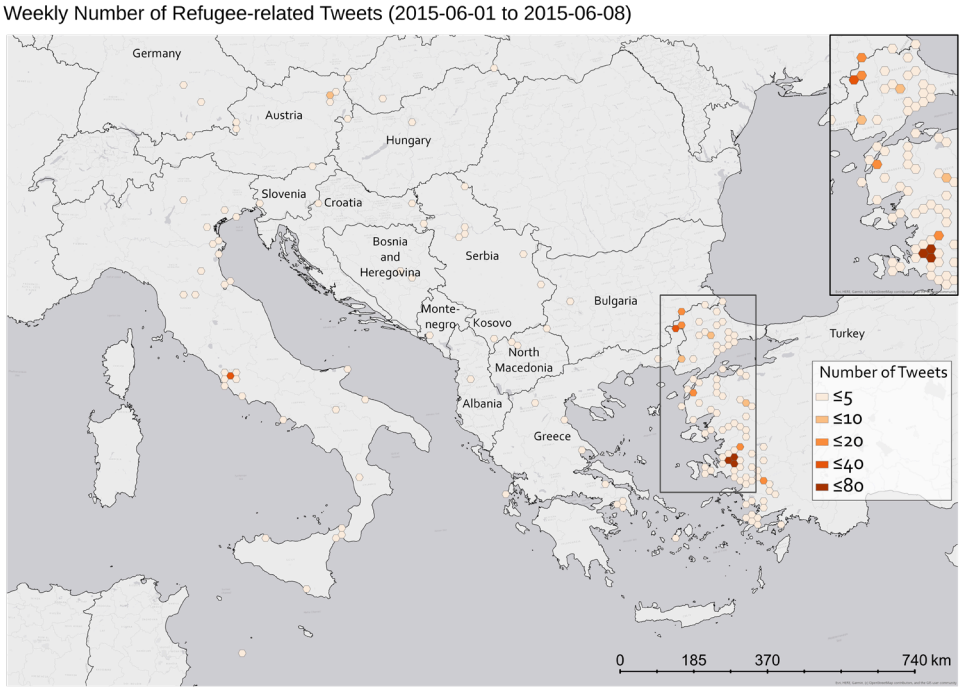

Figure 5: Weekly hexagonal aggregation of relevant tweets from 2015-06-01 to 2015-06-08

3 https://www.unhcr.org/news/latest/2015/9/55f0230e6/frustrated-refugees-migrants-serbia-hungary-border-seek-escape-poor-reception.html
4 https://www.unhcr.org/news/latest/2015/10/560e63626/refugee-sea-arrivals-greece-year-approach-400000.html

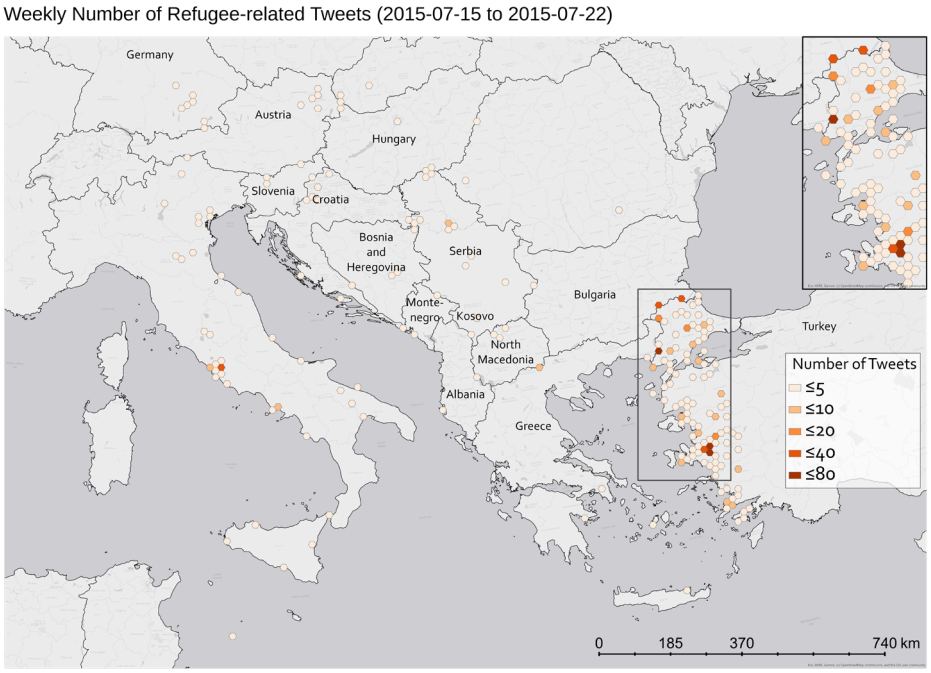

Figure 6: Weekly hexagonal aggregation of relevant tweets from 2015-07-15 to 2015-07-22

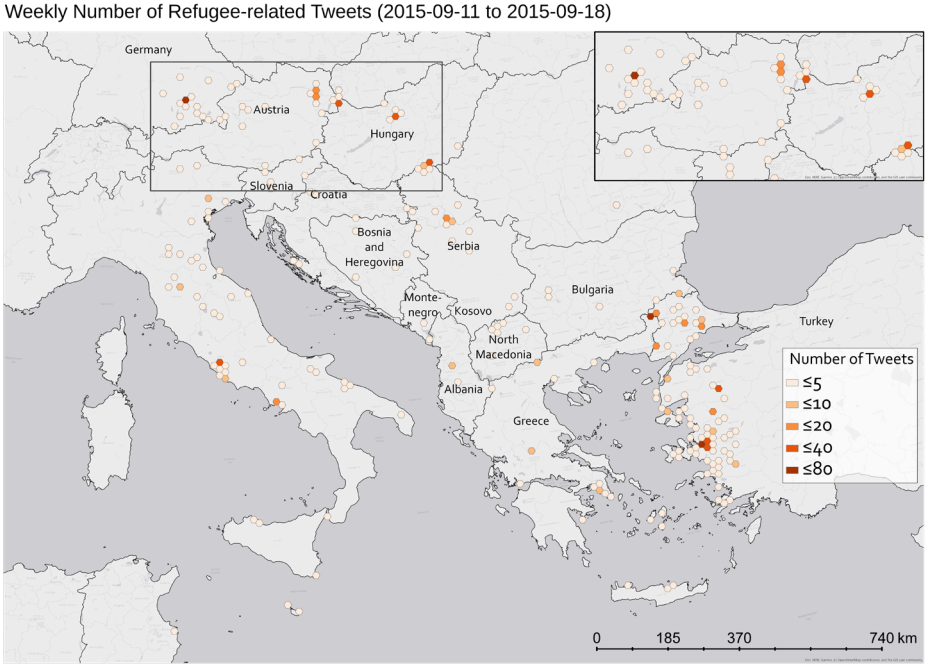

Figure 7: Weekly hexagonal aggregation of relevant tweets from 2015-09-11 to 2015-09-18

Spatiotemporal binning of the data provides a reasonable basis for drawing some conclusions from the data. However, there are several limitations to this approach. Firstly, it does not consider the cell's spatial neighbourhood, which results in outliers being visually very prominent on the map. Secondly, because social media activity in urban areas is higher than in rural areas, the raw counts for the two types of area cannot be compared meaningfully.

The spatial neighbourhood can be included by applying a hotspot analysis to the dataset. A hotspot analysis based on Getis-Ord $G_i^\star$ (Ord & Getis, 1995) identifies statistically significant cold- and hotspots in a dataset by including the values of the spatial neighbourhood of each cell. For the hotspot analysis, we defined the ratio between the refugee-related and the unfiltered social media as values in a fishnet grid. We chose the cell size based on the surface of the study area A and number of Tweets n (Wong & Lee, 2005) and adapted it based on the visual interpretation of results, resulting in cells with a side length of $l = \frac{1}{12} \cdot \sqrt{2\frac{A}{n}}$. We selected weekly bins experimentally because they are short enough to capture the large-scale refugee movement events under investigation and long enough to be reasonably robust against outliers. Figures 8 and 9 show two hotspot maps derived from the Twitter data which capture the situation before and after the Hungarian government closed their border with Serbia,[5] which in turn led to refugees using alternative routes via Croatia and Slovenia.[6] This results in a shift of hotspots from Hungary to the neighbouring countries, as seen on the maps.

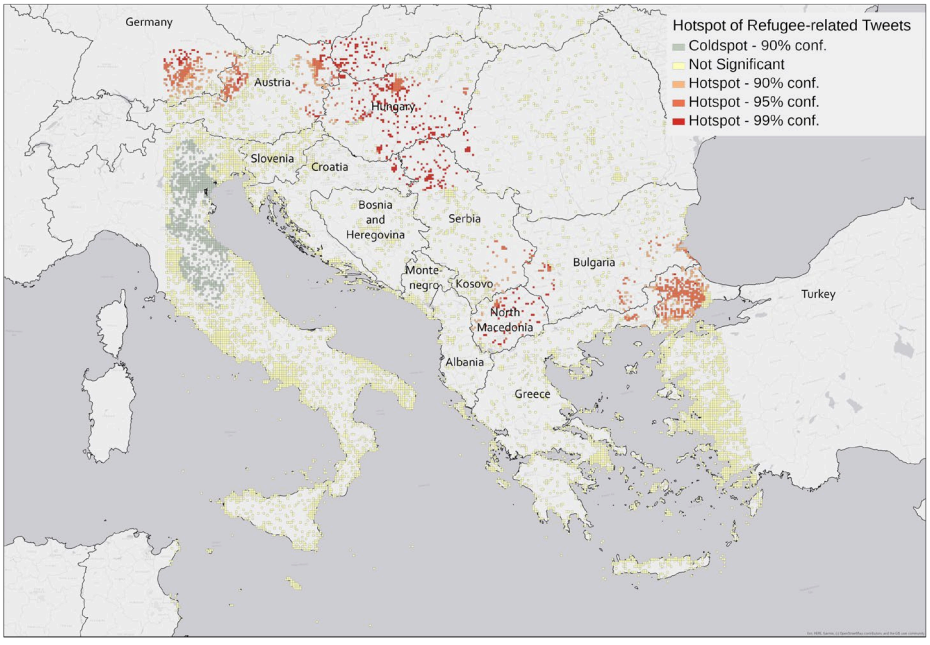

Figure 8: Weekly hotspots of aggregated relevant tweets from 2015-09-04 to 2015-09-11

5 https://www.bbc.com/news/world-europe-34252812
6 https://www.theguardian.com/world/2015/sep/19/young-migrants-trailblazers-hungary-croatia-serbia

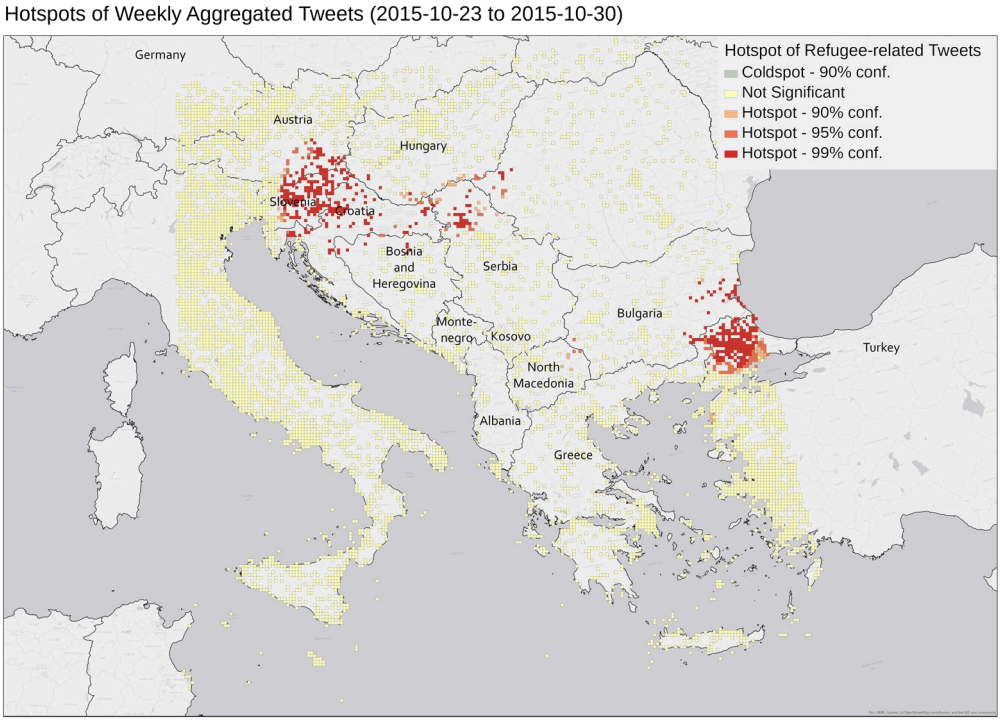

Figure 9: Weekly hotspots of aggregated relevant tweets from 2015-10-23 to 2015-10-30

5 Discussion and Conclusion

We conclude that Twitter data can be used as an indicator for refugee movements in the above scenario. Our findings are based on the large amount of GSND available in our area of interest and on the assumption that many users from Arabic-speaking countries use the Arabic language on social media, which we were able to confirm. We also found that the results vary strongly depending on the methods used. While reliable refugee movement trajectories would provide a good basis for the identification of collective movement patterns in given areas of interest, we were not able to derive the patterns from our data. Upon checking the contents of the tweets from which we constructed the trajectories, we found the authors to be business people, tourists, reporters or bots, instead of the expected refugees. One explanation for the lack of refugee trajectories in the results is the fact that many refugees avoid public communication channels in order to remain inconspicuous for safety reasons. They avoid sharing their location or do not use public social networks at all, opting instead for private communication platforms like WhatsApp or Facebook's Messenger. Findings such as this underline the importance of not relying solely on initial assumptions about the data, and the need to carefully check what new aspects of the data are revealed during the ESDA. The fact that our manual checking revealed some bot-generated messages mixed in with our results tells us that our threshold-based bot-filter approach would benefit from being combined with other more stringent bot-detection methods.

The time series analysis of Arabic-language tweets showed that even though we were able to identify large numbers of Arabic-language tweets in the observed countries, we were not able to clearly identify refugee traces from them. We noticed a series of peaks that appeared almost simultaneously across countries. Overlaying the overall tweet counts of the time period as a baseline, we ruled out the possibility that the peaks were a result of sampling irregularities. Upon manual inspection of the Arabic tweets in question, we found that most of them were posted by tourists who were visiting the regions concerned. This observation is also in line with the time of year, and the interval of approximately one year between peaks. Possible explanations for the lack of refugee-related data are, as already stated above, that refugees do not use public communication channels much, or at least refrain from using the Arabic language on these media.

Some of the most promising results were achieved by grouping the keyword-relevant tweets in weekly hexagonal bins and creating a series of maps. These offer the advantage of high spatial and temporal resolution, which makes them valuable for relief organizations and public authorities who might benefit from this study. The use of keywords for the identification of refugee-relevant tweets also means that we are not restricting our search to tweets created by refugees themselves, but are analysing tweets by a broader range of users engaging in the topic.

The hotspot maps based on the keyword-related tweet counts showed promising results as well. Because they not only report the absolute difference in numbers between rural and urban areas but also consider a cell's neighbourhood, they made some less frequent but locally strong hotspots stand out visually. This allows us to observe the small changes in keyword-related tweet counts that we are interested in. The maps in Figures 8 and 9 show good examples of this effect in North Macedonia, where only a few relevant tweets were located, but because of their homogeneous neighbourhoods the cells appear as hotspots. The opposite effect is visible in parts of Turkey, where the overall numbers are consistently high, but despite that, large areas do not contain significant hotspots. Such differences can also be observed on the temporal scale. The maps show the shift of refugee indicators around Hungary following the closure of the border between Serbia and Hungary.

From this we were able to retrace movements with a very good fit to what we were expecting based on reports from news sources or NGO reports about refugee arrivals, including specific destinations like cities in which many refugees arrived. Our approach also allowed us to partly retrace the paths that many people used in relatively remote regions. However, because the spatial and temporal resolution of our results is high, we have no reference data available for direct comparison, as statistics agencies and NGOs provide their data at a lower resolution. We therefore had to rely on anecdotal evidence in the form of news articles to confirm our findings.

For the end-users of information about refugee movements, it is important to have information available as early as possible and in a format that is easy to understand. Because the information is derived from data that are available as a constant stream, we can meet the time requirement by automating the data processing steps and operating a constantly running service. Based on our results, we believe that providing an interactive web map interface that displays spatially and temporally aggregated keyword-based tweet counts and hotspot maps

along with additional relevant context information as map layers would be a good way to communicate an overview of the situation in an easily accessible and interpretable manner.

It is essential that individuals must not be identifiable from the results. Therefore personally identifying characteristics were not stored in the data, following the principle of data economics and the guidelines developed in Kounadi & Resch (2018). Additional privacy protection measures are the spatial and temporal aggregation of results, with the intention of obscuring individual movement traces.

As the results are largely exploratory, the next steps for our work will focus on the development of a more sophisticated methodology for semantic information extraction to identify refugees in social media and messages that contain information about refugee movements. Potential approaches include supervised learning methods like convolutional (Dos Santos & Gatti, 2014) or recurrent neural networks (Lee & Dernoncourt, 2016), or unsupervised techniques like Latent Dirichlet Allocation (Qiang et al., 2016), as used in Resch et al. (2018), or dependency parsing (Di Caro & Grella, 2013). Furthermore, semantic analysis needs to be more dynamically integrated with spatial and temporal analysis methods such as spatiotemporal autocorrelation, and techniques for spatiotemporal analysis of semantically homogeneous user networks, which may be an indicator for collective refugee movements.

When using GSND as a proxy for an underlying population, because of the heterogeneous usage of these platforms we cannot assume that this population is represented uniformly (Duggan & Maeve Brenner, 2013; Tufekci, 2014). This affects the interpretation of our results insofar as we have to be aware that they only apply to a specific subset of users. Despite this limitation, GSND is still a valuable data source for many spatiotemporal data analysis applications, many of which were catalogued by Steiger et al. (2015). There are numerous possibilities to extend our research. Comparing our results with equally highly sampled official migration time series data from the Balkan countries, which are not available at this point, would be an important step towards statistical validation. The language and keywords a person uses are presumably not the only identifiers by which we can determine whether a person might be a refugee. Spatiotemporal patterns or specific regions that a person visits may also contain such information. Identifying, extracting and integrating such information in our models might significantly improve the validity of this research. Further, the language-based approach presented here could potentially be improved by including more languages that are being used in the area of interest. A related concern is the fact that Arabic can be written in both the Arabic and Latin alphabets. Our current method of language detection only recognizes texts written using Arabic letters, which potentially eliminates useful data. Lastly, applying our methods to similar situations in other parts of the world would give us insights into the spatial transferability of our approach.

Acknowledgements

This study was carried out in the HUMAN+ project, funded by the Austrian security research programme KIRAS of the Federal Ministry of Agriculture, Regions and Tourism (BMLRT), project number 865697. We would like to thank Harvard University's Center for Geographic Analysis for their support in providing us with the Twitter data for our study.

Preferences

Biswas, A. K., & Tortajada-Quiroz, H. C. (1996). Environmental impacts of the Rwandan refugees on Zaire. *Ambio*, *25*(6), 403–408. https://doi.org/10.2307/4314504

Black, R., Adger, W. N., Arnell, N. W., Dercon, S., Geddes, A., & Thomas, D. (2011). The effect of environmental change on human migration. *Global Environmental Change*, *21*(SUPPL. 1), S3–S11. https://doi.org/10.1016/j.gloenvcha.2011.10.001

Breen, D. (2016). Abuses at Europe's Borders. *Forced Migration Review*, *51*(January), 21–23. https://www.fmreview.org/destination-europe/breen

Bulbul, A., Kaplan, C., & Ismail, S. H. (2018). Social media based analysis of refugees in Turkey. *CEUR Workshop Proceedings*, *2078*, 35–40.

Catchpole, M., & Coulombier, D. (2015). Refugee crisis demands European Union-wide surveillance! *Eurosurveillance*, *20*(45), 30063. https://doi.org/10.1086/520454

Chouliaraki, L., & Zaborowski, R. (2017). Voice and community in the 2015 refugee crisis: A content analysis of news coverage in eight European countries. *International Communication Gazette*, *79*(6-7), 613–635. https://doi.org/10.1177/1748048517727173

Curry, T., Croitoru, A., Crooks, A., & Stefanidis, A. (2019). Exodus 2.0: crowdsourcing geographical and social trails of mass migration. *Journal of Geographical Systems*, *21*(1), 161–187. https://doi.org/10.1007/s10109-018-0278-1

Dekker, R., Engbersen, G., Klaver, J., & Vonk, H. (2018). Smart Refugees: How Syrian Asylum Migrants Use Social Media Information in Migration Decision-Making. *Social Media and Society*, *4*(1). https://doi.org/10.1177/2056305118764439

Di Caro, L., & Grella, M. (2013). Sentiment analysis via dependency parsing. *Computer Standards and Interfaces*, *35*(5), 442–453. https://doi.org/10.1016/j.csi.2012.10.005

Dos Santos, C. N., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. *COLING 2014 - 25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers*, 69–78.

Duggan, Maeve Brenner, J. (2013). *The Demographics of Social Media Users —2012* (Vol. 14). Pew Research Center's Internet & American Life Project. https://doi.org/10.1002/cd.23219957004

Garfi, M., Tondelli, S., & Bonoli, A. (2009). Multi-criteria decision analysis for waste management in Saharawi refugee camps. *Waste Management*, *29*(10), 2729–2739. https://doi.org/10.1016/j.wasman.2009.05.019

Gillespie, A. M., Ampofo, L., Cheesman, M., Faith, B., Iliadou, E., Issa, A., Osseiran, S., & Skleparis, D. (2016). *Mapping Refugee Media Journeys: Smartphones and Social Media Networks* (May).

Gillespie, M., Osseiran, S., & Cheesman, M. (2018). Syrian Refugees and the Digital Passage to Europe: Smartphone Infrastructures and Affordances. *Social Media and Society*, *4*(1). https://doi.org/10.1177/2056305118764440

Greenwood, M. J. (1997). *Chapter 12 Internal migration in developed countries* (pp. 647–720). https://doi.org/10.1016/S1574-003X(97)80004-9

Guild, E., Costelle, C., Garlick, M., & Moreno-Lax, V. (2015). The 2015 Refugee Crisis in the European Union. *CEPS Policy Brief*, *332*, 1–6. https://www.ceps.eu/system/files/CEPS PB332 Refugee Crisis in EU{_}0.pdf

Hübl, F., Cvetojevic, S., Hochmair, H., & Paulus, G. (2017). Analyzing refugee migration patterns using geo-tagged tweets. *ISPRS International Journal of Geo-Information*, *6*(10). https://doi.org/10.3390/ijgi6100302

Jacobsen, K. (1997). Refugees' environmental impact: The effect of patterns of settlement. *Journal of Refugee Studies*, *10*(1), 19–36. https://doi.org/10.1093/jrs/10.1.19

Kounadi, O., & Resch, B. (2018). A Geoprivacy by Design Guideline for Research Campaigns That Use Participatory Sensing Data. *Journal of Empirical Research on Human Research Ethics*, *13*(3), 203–222. https://doi.org/10.1177/1556264618759877

Resch, B. (2013). People as Sensors and Collective Sensing-Contextual Observations ComplementingGeo-Sensor Network Measurements. *Progress in Location-Based Services*, 373–388. https://doi.org/10.1007/978-3-642-34203-5

Resch, B., Usländer, F., & Havas, C. (2018). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, *45*(4), 362–376 https://doi.org/10.1080/15230406.2017.1356242

Steiger, E., Albuquerque, J. P. de, & Zipf, A. (2015). An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. *Transactions in GIS*, *19*(6), 809–834. https://doi.org/10.1111/tgis.12132

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *Eighth International Aaai Conference on Weblogs and Social Media*.

Warner, K. (2009). *In Search of Shelter Mapping the Effects of Climate Change on Human Migration and Displacement Koko Warner , Charles Ehrhart , Alex de Sherbinin, Susana Adamo , and Tricia Chai-Onn*. *January*.

Wong, D. W. S., & Lee, J. (2005). *Statistical analysis of geographic information with ArcView GIS and ArcGIS*. Hoboken, NJ: Wiley.

GI_Forum 2020, Issue 1
Page: 153 - 163
Short Paper
Corresponding Author:
anita.graser@ait.ac.at
DOI: 10.1553/giscience2020_01_s153

Extracting Patterns from Large Movement Datasets

Anita Graser[1,2], Peter Widhalm[1] and Melitta Dragaschnig[1]

[1]AIT Austrian Institute of Technology, Vienna, Austria
[2]University of Salzburg, Salzburg, Austria

Abstract

Extracting useful information from large spatiotemporal datasets is a challenging task that requires suitable visual data representations. Big movement data are particularly hard to visualize since they are prone to visual clutter caused by overlapping and crisscrossing trajectories. Different data aggregation approaches have been developed to address this challenge and to provide analysts with better visualizations for data exploration and data-driven hypothesis generation. However, most approaches for extracting patterns, such as mobility graphs or generalized flow maps, cannot handle large input datasets. This paper presents a flow extraction algorithm that can be used in distributed computing environments and thus make it possible to explore movement patterns in large datasets. We demonstrate its usefulness in a use case exploring maritime vessel movements

Keywords:

trajectories, spatiotemporal analysis, movement data analysis

1 Introduction

Large movement datasets that are collected by systems tracking vehicles, people or goods have the potential to improve our understanding of mobility and transport systems, in order to, for example, monitor vehicle emissions or tackle the issue of rising road traffic fatalities (WHO, 2018). Data exploration and data-driven hypothesis generation are important steps in the process of building data-driven models since they enable knowledge to be gained, and spatial modelling (Miller & Goodchild, 2015). However, we humans are not well equipped to understand large amounts of raw numerical data. Instead, we need to visually represent the data to extract useful information. The development of visualizations of big spatial data, however, is challenging (Robinson et al., 2017). Movement data in particular are hard to visualize due to visual clutter caused by intersecting and overlapping trajectories. Therefore, it is 'necessary to use appropriate data abstraction methods' (Andrienko & Andrienko, 2011).

Data aggregation is a common technique for dealing with large amounts of data (Andrienko et al., 2017a). Concerning movement data, density surfaces are probably the most commonly used aggregation technique, as can be inferred from their prevalence in review (Chen et al.,

2015; Andrienko et al., 2017b; He et al., 2019) and application papers (Willems et al., 2009; Aronsen & Landmark, 2016). The temporal dimension has been integrated into density concepts in Demšar & Virrantaus' (2010) space-time density volumes of trajectories. However, density approaches provide only limited data exploration capabilities.

More advanced aggregation techniques aim to extract mobility graphs or generalized flow maps from movement records. For example, Andrienko & Andrienko (2011) extract and cluster characteristic waypoints from trajectories to generate aggregated flow maps, as illustrated in Figure 1. However, their approach cannot deal with large datasets. As a work-around, they therefore suggest extracting and clustering characteristic points from a subset of the full trajectory dataset. 'Assuming that the sampling is done sufficiently well, i.e., the statistical and spatial distribution properties of the whole data set are preserved in a sample, we can use the territory division so obtained to summarize […] the whole database' (p. 216).

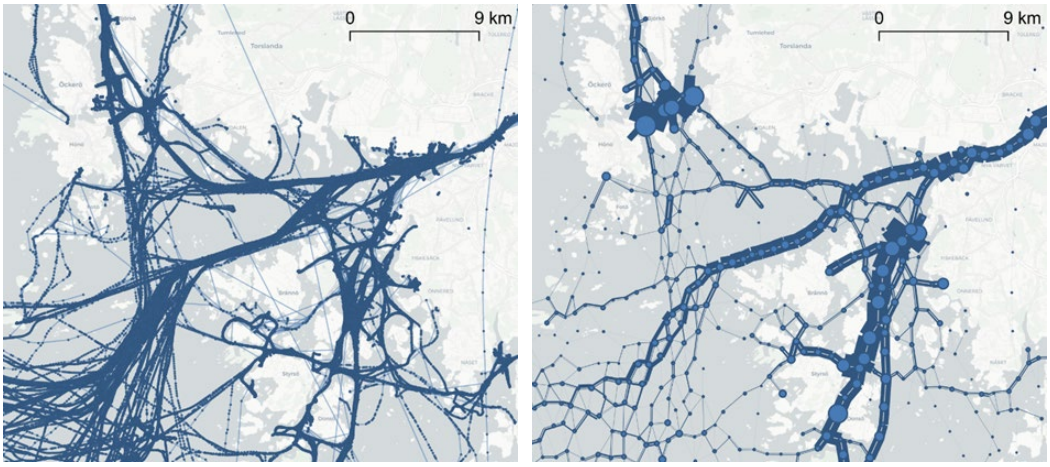

Figure 1: Example of raw movement data (left) and extracted flow map (right) using the algorithm by Andrienko & Andrienko (2011), applied to one day of vessel movement data in the area surrounding Gothenburg, Sweden. The flow map clearly communicates the relative popularity of the different route options in this area. (Background map: Positron © OpenStreetMap contributors and CARTO)

Extracting suitable samples from large movement datasets is no simple task. To avoid sampling, Pallotta et al. (2013) instead use incremental DBSCAN to identify waypoints. However, tuning DBSCAN parameters 'for good waypoint identification is not possible when dealing with areas with varying density' (Dobrkovic et al., 2018, p. 25). Approaches addressing the issue of varying density include lattice or grid-based DBSCAN (Xiao et al., 2017) as well as other clustering algorithms, such as OPTICS (Rinzivillo et al., 2008) or genetic algorithms (Dobrkovic et al., 2018). In Graser et al. (2020), we present M³ – a movement data exploration model that uses an incremental grid-based clustering algorithm. M³ runs in distributed computing environments and is therefore scaleable to large datasets that exceed the processing capacity of individual machines. Besides location information, M³ also takes other movement characteristics, such as speed and direction, into account. However, Graser et al. (2020) do not

cover the follow-up step of computing flows. To address this gap, this paper proposes an algorithm for computing flows from massive movement datasets.

The remainder of this paper is structured as follows: Section 2 presents our incremental flow computation algorithm; Section 3 presents a case study with massive vessel movement data; finally, Section 4 draws conclusions and provides an outlook for future work.

2 Methodology

Conceptually similar to Andrienko & Andrienko (2011), our proposed flow extraction method is based on a two-step process. First, we extract prototypes from the movement data. These prototypes describe movement characteristics in a certain geographic area and contain the following information:

- Number of input location records (similar to density surfaces but with support for multiple prototypes per grid cell)
- Geographical distribution (mean coordinates and variance) of location records
- Distributions (mean and variance) of direction, speed and other characteristics available in the location records (including temporal or seasonal information).

In the second step, we determine flows between prototypes, including information about:

- Distribution of travel speeds
- Number of observed transitions.

The details of both steps are described in the following subsections.

2.1 Extracting prototypes

This step is based on the M³ model introduced in Graser et al. (2020). In short, movement data records are clustered into prototypes using an incremental algorithm based on Vector Quantization. In Vector Quantization, probability density functions are modeled by the distribution of so-called prototype vectors. In our approach, these prototypes describe movement properties using Gaussian Mixture Models (GMMs). Each GMM consists of a set of components C. Each component c has a set of parameters $\theta_c = \{\boldsymbol{\mu_c}, \boldsymbol{S_c}\}$, where $\boldsymbol{\mu_c}$ is the mean value vector and $\boldsymbol{S_c}$ is the covariance matrix of the multivariate Gaussian. The Leader-Follower clustering (Duda et al., 2001) approach employed adds new data points to the closest existing cluster or creates a new cluster if a specified distance threshold $\mathrm{d_{max}}$ between the data point and the closest cluster is exceeded. To allow for distributed processing, movement data is split using a spatiotemporal grid. The contents of each grid cell are then processed independently.

2.2 Computing flows

After the prototypes have been computed, our new flow algorithm computes transitions between pairs of prototypes. Like the prototypes, our flows are also modeled using GMMs.

The information modeled in each flow includes but is not limited to the number of transitions and the speed distribution. An object moving from prototype A to prototype B triggers an update of the corresponding flow. To allow for distributed processing, each node in the distributed computing environment needs a copy of the previously computed prototypes. Before flows can be computed, movement records are grouped by moving-object ID, sorted chronologically, and then split into trajectories. Each moving object is processed independently. The complete flow algorithm (as illustrated in Figure 2) can be summarized as follows:

1. Create trajectories (i.e. sequences of chronologically ordered records) for individual moving objects:
    a. Split continuous movement tracks at stops and observation gaps and remove outliers
    b. Optionally: generalize the trajectories to reduce data size.
2. For each trajectory:
    a. Let x be the next record in the trajectory
    b. Find the most similar prototype μ^*
    c. Let $\mathrm{d_{max}}$ be the distance threshold
    d. If $|\mathrm{x} - \mu^*| \leq \mathrm{d_{max}}$ and μ^* is different from the previous prototype:
        i. Let γ be the flow between μ^* and the previous prototype
        ii. Let α_x, α_γ, α_{μ^*} be the directions of x, γ, and μ^* respectively
        iii. Let $\sigma_{\alpha_{\mu^*}}$ be the standard deviation of the direction of μ^*
        iv. Let φ_{max} be the direction difference threshold
        v. If $|\alpha_x - \alpha_{\mu^*}| \leq \varphi_{max}$ and $|\alpha_x - \alpha_\gamma| \leq \varphi_{max}$ and $|\alpha_x - \alpha_{\mu^*}| \leq 2\sigma_{\alpha_{\mu^*}}$:
            1. Update the flow properties: travel speed and number of transitions
            2. Update the previous prototype reference.

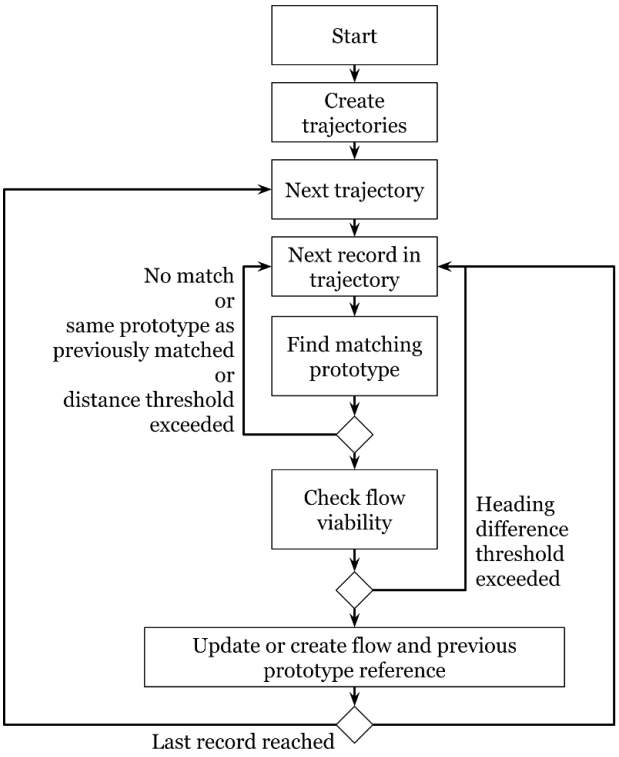

Figure 2: Flow diagram for the flow algorithm

Both algorithms for extracting prototypes and computing flows were implemented in Apache Spark. Spark (Zaharia et al., 2010) is a general-purpose cluster-computing framework that supports distributed computing on large datasets which do not fit into the available memory. This is important for processing large movement datasets. For the prototype extraction, only the (intermediate) prototypes and the particular movement record currently being worked on have to be kept in the memory. Similarly, for the flow computations, only the prototypes, the (intermediate) flow results, and the trajectory currently being worked on have to be kept in the memory for each iteration.

3 Case study

This case study aims to extract movement patterns from massive maritime vessel movement data. Vessel movements are tracked by the Automatic Identification System (AIS), which requires that vessels above a certain size broadcast their position and status.

3.1 Input data and cluster setup

The data used in this study were published by the Danish Maritime Authority; our dataset contains 350 million records covering July 2017. To store this data for distributed processing,

we use GeoMesa Accumulo. GeoMesa provides fast spatiotemporal indexing (Hughes et al. 2015) to help store and access spatiotemporal data. GeoMesa also provides spatial analysis functions that can be called by Spark.

The computer cluster used in this case study comprises eight data nodes: three nodes with two Intel Xeon E5-2430L CPUs and 32G RAM each, three nodes with two Intel Xeon E5-2660 v3 and 64G RAM each, and two nodes with two Intel Xeon Gold 6136 each. The operating system and HDFS file system reside on SSDs. The setup is based on Apache Hadoop 2.7 and managed using Ambari 2.6.

3.2 Results

Figure 3 and Figure 4 present the resulting prototypes and flows for two different vessel types – passenger and tanker vessels – in the area surrounding Gothenburg, Sweden. Wide flow lines in Figure 3 highlight frequent ferry connections in this area. These connections include local ferries that travel back and forth between the mainland and various islands along the coast, as well as long-distance ferry connections heading towards Denmark and Germany in the south-west.

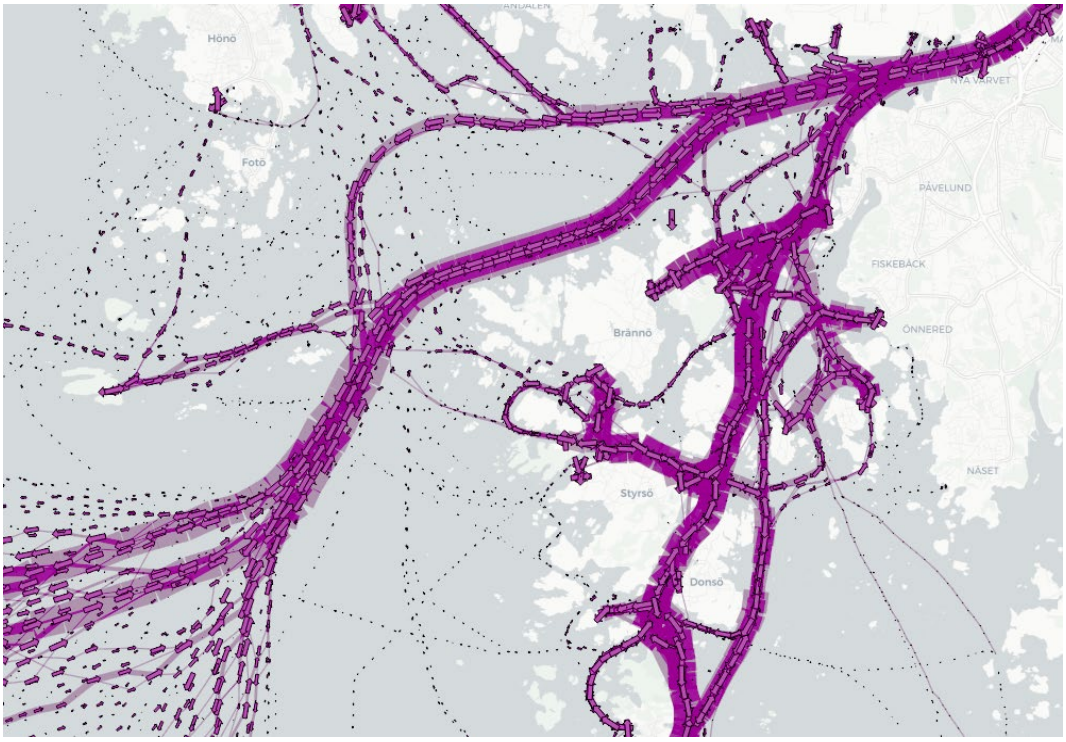

Figure 3: Passenger vessel prototypes (arrows) and flows (connections between arrows)

The tanker flows presented in Figure 4 are mostly focused on the main corridor entering the port of Gothenburg. Smaller flows indicate tankers providing services to islands. Tangles of flow lines and prototypes pointing in various directions in the highlighted region indicate an anchorage area. Indeed, comparisons of these movement patterns and mapped maritime information (Sjöfartsverket, 2016) confirm that this region is a dedicated anchorage area.

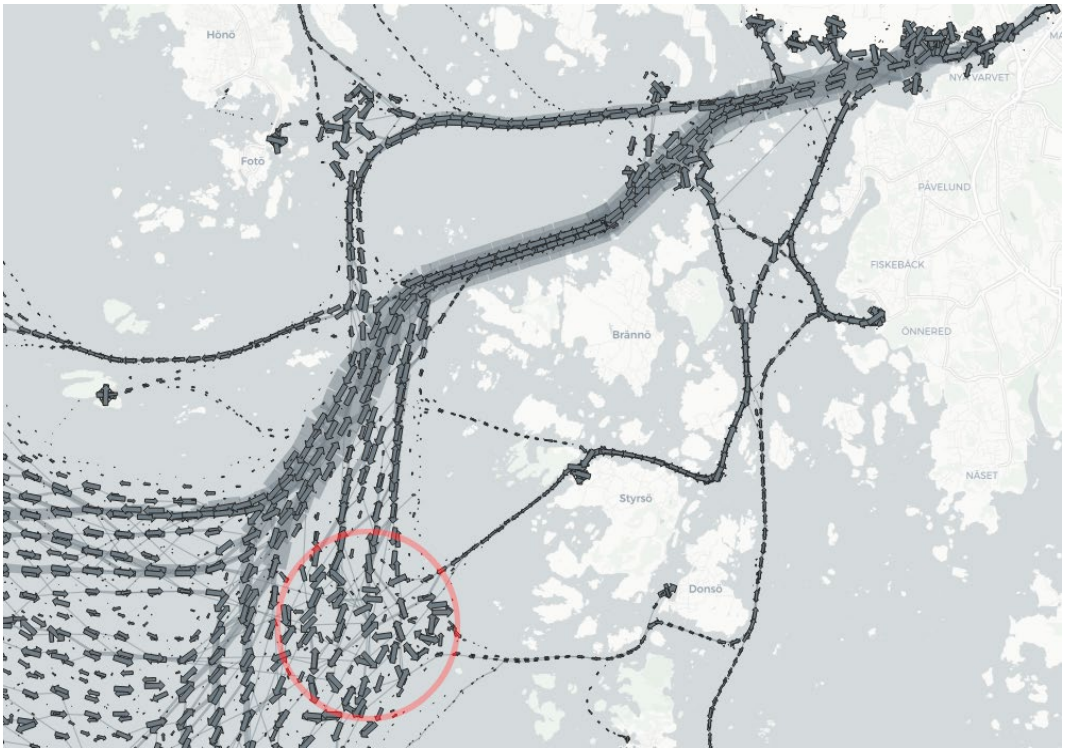

Figure 4: Tanker vessel prototypes (arrows) and flows (connections between arrows). The red circle marks an anchorage area containing tangles of flow lines and prototypes pointing in various directions

Since our flows also include information about mean movement speeds and speed distributions, they also support a more in-depth exploration of speed patterns than regular flow maps (Andrienko & Andrienko, 2011), which model flow strength but not flow speed distribution. For example, Figure 5 shows the speed patterns of passenger vessels. Wide lines indicate a high variation in speed values along a flow. The northern route into and out of the harbor of Gothenburg is particularly noteworthy for its high variations of speed. Information about regions with high speed variation is particularly relevant since these areas need to be watched more closely because accidents are more likely to happen where lots of vessels are moving at different speeds. At its western end, this route splits into darker (higher speed) and lighter (lower speed) routes with lower speed variation. This indicates that vessels with different speed characteristics follow different routes from thereon.

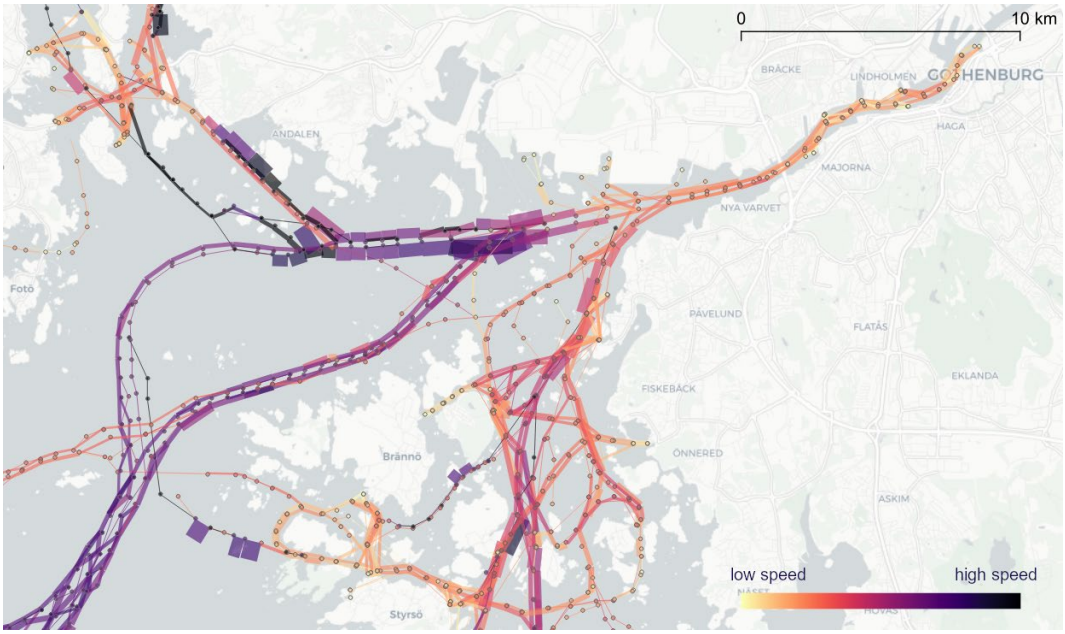

Figure 5: Passenger speed patterns: mean flow speeds (line colour: darker colours equal higher speeds) and speed variation (line width)

As this case study shows, our flows enable the exploration of movement patterns regarding the spatial distribution and density of trajectories, as well as the flow-specific distribution of movement speeds. We discovered, for example, anchorage areas that could be confirmed by official port maps, as well as routes with very high variations in vessel speed which could be important for safety considerations.

4 Discussion

The most critical step in the flow computation is the creation of trajectories. While Spark provides high-level functions for grouping and aggregating records, these are mostly geared towards dealing with unsorted data. If high-level Spark core functionality is used incorrectly, an aggregator needs to collect and sort the entire trajectory in the main memory of a single processing node. Consequently, analyses frequently run into out-of-memory errors when dealing with large datasets. Third-party libraries such as Spark-sorted (Tresata, 2020) provide groupSort, a functionality required to group, sort and iteratively process massive datasets. It never materializes the group for a given key in memory, but instead offers iterator-based streaming of the sorted data. This functionality helps to efficiently build the trajectories which are necessary for computing flows.

The runtime of the computations depends on a variety of factors, including the size of the input dataset, the characteristics of the compute cluster setup (such as the number of Spark executors and their assigned memory), and the spatial resolution of the model (d_{max} and n_{max}

in the prototype extraction step). Models with fewer prototypes and larger cells can be computed faster but provide a less detailed representation of the original observations. A detailed runtime evaluation and sensitivity analysis of the prototype algorithm is provided in Graser et al. (2020). The runtime of the trajectory creation step depends on the efficiency of the groupSort implementation. The runtime of the flow computation step depends on the efficiency of the implementation for finding the matching prototype and thus on the spatial indexing method used.

While the prototype algorithm allows for continuous updates and can therefore handle continuous streams of input data, the flow algorithm does not allow for continuous updates. Flows would have to be recomputed (at least locally) whenever prototypes changed. Therefore, the algorithm does not support exploration of continuous data streams. However, it can be used to explore large historical datasets. To support incremental updates of the flow model, it needs to be integrated into the prototype computation steps. An incremental flow model must keep track of the last observed positions of all moving objects within the system. This introduces considerable memory requirements, since every computational node needs access to this information.

The quality of the flows presented in the case study was assessed using visual plausibility checks. Both the form of the flows (geometries) as well as their strength and speed show expected patterns and are therefore deemed suitable for the exploration of this movement dataset. A quantitative evaluation requires a measure for how well the computed flows represent the original trajectories. To the best of our knowledge, there is no established method that addresses this specific issue. However, measures used to evaluate trajectory generalization algorithms may be adaptable to this issue.

5 Conclusions and future work

We have presented a novel algorithm for extracting flow patterns from large movement datasets. Our new flow algorithm builds on the distributed movement data exploration model M³ and enables the distributed computation of flows between prototypes. We have demonstrated the usefulness of this approach in a case study involving a large dataset of maritime vessel movements.

While the visualizations in this case study enable a qualitative evaluation of the resulting flows, questions remain pertaining to the quantitative evaluation of the flows. Future work should therefore include the development of quantitative measures that can be used to assess the quality of aggregated flow information.

Potential uses cases for flow data are not limited to data exploration. In the future, we plan to use movement patterns extracted from historical data in predictive analytics, for example, to provide location predictions as well as to estimate time of arrival.

Acknowledgements

This work was supported by the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT) within the 'IKT der Zukunft' programme under Grant 861258 (project MARNG).

References

Andrienko, N. & Andrienko, G. (2011). Spatial generalization and aggregation of massive movement data. *IEEE Transactions on visualization and computer graphics*, *17*(2), 205-219.

Andrienko, G., Andrienko, N., Chen, W., Maciejewski, R., & Zhao, Y. (2017a). Visual analytics of mobility and transportation: State of the art and further research directions. *IEEE Transactions on Intelligent Transportation Systems*, *18*(8), 2232-2249.

Andrienko, G., Andrienko, N., Fuchs, G., & Wood, J. (2017b). Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin-destination movement data. *IEEE transactions on visualization and computer graphics*, *23*(9), 2120-2136. 4

Aronsen, M., & Landmark, K. (2016). Density mapping of ship traffic. FFI-RAPPORT 16/02061. Norwegian Defence Research Establishment (FFI).

Chen, W., Guo, F., & Wang, F. Y. (2015). A survey of traffic data visualization. *IEEE Transactions on Intelligent Transportation Systems*, *16*(6), 2970-2984.

Demšar, U., & Virrantaus, K. (2010). Space–time density of trajectories: exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science*, *24*(10), 1527-1542.

Dobrkovic, A., Iacob, M. E., & van Hillegersberg, J. (2018). Maritime pattern extraction and route reconstruction from incomplete AIS data. *International journal of Data science and Analytics*, *5*(2-3), 111-136.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. John Wiley & Sons.

Graser, A., Widhalm, P., & Dragaschnig, M. (2020, in print). The M³ massive movement model: a distributed incrementally updatable solution for big movement data exploration. *International Journal of Geographical Information Science*.

He, J., Chen, H., Chen, Y., Tang, X., & Zou, Y. (2019). Diverse visualization techniques and methods of moving-object-trajectory data: a review. *ISPRS International Journal of Geo-Information*, *8*(2), 63.

Hughes, J. N., Annex, A., Eichelberger, C. N., Fox, A., Hulbert, A., & Ronquest, M. (2015). Geomesa: a distributed architecture for spatio-temporal fusion. In *Geospatial Informatics, Fusion, and Motion Video Analytics V* (Vol. 9473, p. 94730F). International Society for Optics and Photonics.

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

Pallotta, G., Vespe, M., & Bryan, K. (2013). Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction. *Entropy*, *15*(6), 2218-2245.

Rinzivillo, S., Pedreschi, D., Nanni, M., Giannotti, F., Andrienko, N., & Andrienko, G. (2008). Visually driven analysis of movement data by progressive clustering. *Information Visualization*, *7*(3-4), 225-239.

Robinson, A.C., Demšar, U., Moore, A.B., Buckley, A., Jiang, B., Field, K., Kraak, M.J., Camboim, S.P. & Sluter, C.R. (2017). Geospatial big data and cartography: research challenges and opportunities for making maps that matter. *International Journal of Cartography, 3*(1), 32–60.

Sjöfartsverket (2016). Passageplan Göteborg, Retrieved from https://www.sjofartsverket.se/pages/29206/Passageplan%20folder%20Göteborg%202016.pdf

Tresata (2020) Secondary sort and streaming reduce for Apache Spark: tresata/spark-sorted. Retrieved from https://github.com/tresata/spark-sorted

WHO (2018) Global status report on road safety 2018. Technical report, World Health Organization, Geneva. Retrieved from

https://apps.who.int/iris/bitstream/handle/10665/276462/9789241565684-eng.pdf

Willems, N., Van De Wetering, H., & Van Wijk, J. J. (2009). Visualization of vessel movements. In *Computer Graphics Forum*, 28(3), pp. 959-966. Oxford, UK: Blackwell.

Xiao, Z., Ponnambalam, L., Fu, X., & Zhang, W. (2017). Maritime traffic probabilistic forecasting based on vessels' waterway patterns and motion behaviors. *IEEE Transactions on Intelligent Transportation Systems*, *18*(11), 3122-3134.

Zaharia, M., Chowdhury, M., Franklin, M.J. et al. (2010). Spark: Cluster computing with working sets. In *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing,* p. 10.