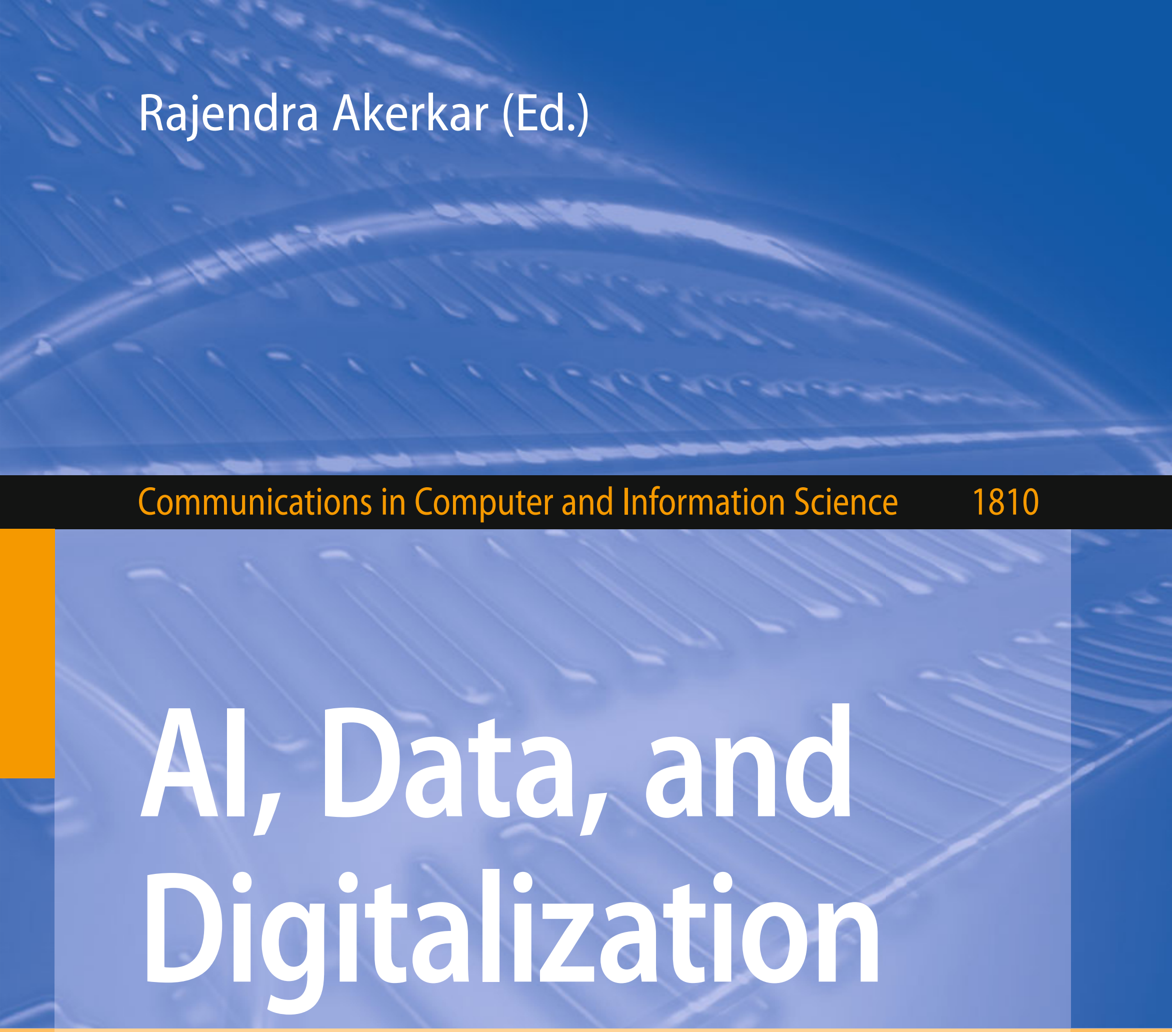


Rajendra Akerkar (Ed.)

Communications in Computer and Information Science 1810

AI, Data, and Digitalization

First International Symposium, SAIDD 2023
Sogndal, Norway, May 9–10, 2023
Revised Selected Papers

Springer

Communications in Computer and Information Science 1810

Rationale

The CCIS series is devoted to the publication of proceedings of computer science conferences. Its aim is to efficiently disseminate original research results in informatics in printed and electronic form. While the focus is on publication of peer-reviewed full papers presenting mature work, inclusion of reviewed short papers reporting on work in progress is welcome, too. Besides globally relevant meetings with internationally representative program committees guaranteeing a strict peer-reviewing and paper selection process, conferences run by societies or of high regional or national relevance are also considered for publication.

Topics

The topical scope of CCIS spans the entire spectrum of informatics ranging from foundational topics in the theory of computing to information and communications science and technology and a broad variety of interdisciplinary application fields.

Information for Volume Editors and Authors

Publication in CCIS is free of charge. No royalties are paid, however, we offer registered conference participants temporary free access to the online version of the conference proceedings on SpringerLink (http://link.springer.com) by means of an http referrer from the conference website and/or a number of complimentary printed copies, as specified in the official acceptance email of the event.

CCIS proceedings can be published in time for distribution at conferences or as post-proceedings, and delivered in the form of printed books and/or electronically as USBs and/or e-content licenses for accessing proceedings at SpringerLink. Furthermore, CCIS proceedings are included in the CCIS electronic book series hosted in the SpringerLink digital library at http://link.springer.com/bookseries/7899. Conferences publishing in CCIS are allowed to use Online Conference Service (OCS) for managing the whole proceedings lifecycle (from submission and reviewing to preparing for publication) free of charge.

Publication process

The language of publication is exclusively English. Authors publishing in CCIS have to sign the Springer CCIS copyright transfer form, however, they are free to use their material published in CCIS for substantially changed, more elaborate subsequent publications elsewhere. For the preparation of the camera-ready papers/files, authors have to strictly adhere to the Springer CCIS Authors' Instructions and are strongly encouraged to use the CCIS LaTeX style files or templates.

Abstracting/Indexing

CCIS is abstracted/indexed in DBLP, Google Scholar, EI-Compendex, Mathematical Reviews, SCImago, Scopus. CCIS volumes are also submitted for the inclusion in ISI Proceedings.

How to start

To start the evaluation of your proposal for inclusion in the CCIS series, please send an e-mail to ccis@springer.com.

Rajendra Akerkar
Editor

AI, Data, and Digitalization

First International Symposium, SAIDD 2023
Sogndal, Norway, May 9–10, 2023
Revised Selected Papers

Springer

Editor
Rajendra Akerkar
Western Norway Research Institute
Sogndal, Norway

ISSN 1865-0929　　ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-031-53769-1　　ISBN 978-3-031-53770-7 (eBook)
https://doi.org/10.1007/978-3-031-53770-7

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Preface

These post-proceedings comprise 13 refereed research papers that were presented at the Symposium on AI, Data and Digitalization (SAIDD 2023) during May 9–10, 2023, at Western Norway Research Institute, Sogndal, Norway. The symposium was organized by *Transnational Partnership for Excellent Research and Education in Disruptive Technologies for a Resilient Future* (INTPART DTRF) and *Big Data & Emerging Technologies Research Group* at the Western Norway Research Institute in Sogndal, Norway.

Artificial Intelligence, Big Data and Digitalization allow us more than ever before to make use of the data our society and public and private sectors generate every day. Institutions around the world are increasingly turning to such methods and technologies to help them solve complex problems, promote efficiency and improve performance and decision-making. In order to realise the full benefits of Big Data and AI technologies, we will need to act to support the growth of our research and innovation capabilities, building on strong foundations that already exist. Furthermore, new challenges and prospects require new theory, methodology, best practice and systems, and this should be developed, shared and discussed by a wide range of stakeholders. Therefore, the objective of SAIDD 2023 was to bring prominent researchers, policy experts and practitioners together in order to foster a deeper understanding of how data and AI are setting the stage for the digital revolution and contributing to solving societal challenges.

We received 42 papers submitted by authors coming from 12 countries around the world. Each paper was reviewed by at least three members of the international Scientific Committee (SC); reviews were single blind. Of these submissions, 12 position papers and 18 lightning talks were accepted for presentation at the symposium. Extended versions of all presentations were invited for a second review process. Finally, we selected 13 papers for publication in this volume of the Communications in Computer and Information Science series.

Many people contributed toward the success of the symposium. First, we would like to recognize the work of the scientific committee members for their cooperation in the reviewing process, an essential stage in ensuring the high quality of the accepted papers. We would like to thank the SC members for performing their reviewing work with diligence. We thank the Local Organizing Committee and Technical Support team for their terrific work before and during the symposium. Finally, we cordially thank all the authors, presenters and delegates for their valuable contribution to this successful event. The symposium would not have been possible without their support. Our special thanks are also due to the Research Council of Norway for their kind support and to Springer for publishing the post-proceedings. We also thank the EasyChair conference system, which made our job so much simpler. Finally, thanks go to the INTPART DTRF project team and the “Technology and Society” group at Western Norway Research Institute for supporting the symposium.

November 2023 Rajendra Akerkar

Organization

Scientific Committee Chair

Rajendra Akerkar	Western Norway Research Institute, Norway

Scientific Committee

Nick Bassiliades	Aristotle University of Thessaloniki, Greece
David Camacho	Universidad Politécnica de Madrid, Spain
Jasmien César	Mastercard, Belgium
Vadim Ermolayev	Ukrainian Catholic University, Ukraine
Anna Fensel	Wageningen University & Research, The Netherlands
Song Guo	Hong Kong Polytechnic University, China
Jason J. Jung	Chung-Ang University, Republic of Korea
Yusheng Ji	National Institute of Informatics, Japan
Pawan Lingras	Saint Mary's University, Canada
Pierre Maret	Université Jean Monnet, France
Andreas L. Opdahl	University of Bergen, Norway
Hemant Purohit	George Mason University, USA
Charles Robinson	Thales, France
M. Sasikumar	Centre for Development of Advanced Computing, India
André Skupin	San Diego State University, USA
Xuan Song	Southern University of Science and Technology, China
Abhishek Srivastava	Indian Institute of Technology, Indore, India
Fan Zipei	University of Tokyo, Japan

Organizing Committee

Reza Arghandeh	Western Norway University of Applied Sciences, Norway
Simen Eide	Schibsted, Norway
John Krogstie	NTNU, Norway
Kenth Engø-Monsen	Smart Innovation Norway, Norway

Sachin Gaur	NORA, Norway
Hoang Long Nguyen	Western Norway Research Institute, Norway
Vimala Nunavath	University of South-Eastern Norway, Norway
Endre Sundsdal	Norwegian Mapping Authority, Norway
Malin Waage	Western Norway Research Institute, Norway
Bjørn Christian Weinbach	Western Norway Research Institute, Norway
Sule Yildirim Yayilgan	NTNU, Norway
Svein Ølnes	Western Norway Research Institute, Norway

Reviewers

Abhishek Srivastava
Andreas L. Opdahl
Ankit Jain
Anna Fensel
David Camacho
Fan Zipei
Fedor Vitiugin
Hemant Purohit
Hoang Long Nguyen
Jason J. Jung
M. Sasikumar
Nick Bassiliades
Pawan Lingras
Pierre Maret
Reza Arghandeh
Rui Zhang
Rupendra Pratap Singh Hada
Song Guo
Sule Yildirim Yayilgan
Vadim Ermolayev
Vimala Nunavath
Xuan Song
Yasas Senarath
Yusheng Ji

Contents

Geolocation Data as a Research Tool for the Organization of the Settlement System and Mobility Mapping – Case Study of the Spatial Mobility Model in Czechia

Václav Jaroš(✉)

Department of Social Geography and Regional Development, Faculty of Science, Ministry of the Interior Charles University, Prague, Czech Republic
vaclav.jaros@mvcr.cz, vaclav.jaros@natur.cuni.cz

Abstract. Geolocation data is a widely used source of the spatial information about the population. Their great potential might be also used for population mobility research to identify spatial interactions forming the hierarchical structure of the settlement system. For this purpose, a model of data acquisition and their preliminary analysis was developed. This model represents an effective tool for mapping the mobility behavior of the population. Using the example of Czechia, primary commuting links are identified, which are subsequently analyzed in detail using GIS tools in both desktop and online environments. Therefore, important commuting centers of different hierarchical levels are defined by the volume and nature of spatial interactions. This approach is used as a source of important expertise for the proposals on subsequent administration reform in Czechia. Nevertheless, the entire model is generally transferable, and the entire method of using the geolocation data for mapping the hierarchy within the settlement system can be replicated in other countries as well.

Keywords: Geolocation data · Mobile phone data · Data acquisition model · Spatial interaction · GIS · Settlement system · Mobility mapping · Travel behavior patterns

1 Research on the Spatial Organization of the Settlement System and Useable Data Sources

This study is based on assumption that the settlement system represents a large set of complex processes between particular components of the society and the landscape variable in time and space. This process results in socio-spatial differentiation, which manifests itself the most as a spatial concentration of activities within society.

Implemented as part of the project "Improvement in Conditions for Decentralization and Accessibility of State Administration in the Territory" financed from EEA and Norway Funds 2014 – 2021.

R. Akerkar (Ed.): SAIDD 2023, CCIS 1810, pp. 1–15, 2024.
https://doi.org/10.1007/978-3-031-53770-7_1

The spatial concentration of activities is a natural process of development of social systems. A certain form of concentration is necessary, as it is not possible to ensure the availability of all activities, which have different degrees of rarity, in all locations equally. This is the very essence of the formation of settlement systems, which the concentration of activities allows to arise.

The naturality of these differentiation processes and a certain tendency of social processes to create spatial differences have historically led to significant interest in the study of these phenomena. Some of such works have become key studies establishing individual paradigmatic schools of thought not only of the entire social and regional geography but also e.g. of regional economics and other social sciences.

The fundamental studies of the regularities of the spatial arrangement of social activities and their reflection in the settlement structure are based on principles described in localization theories [26–29] and much later in theories of the new economic geography (NEG) which brings a more realistic view of the conditionality of population distribution and economic activities in space due to taking into account a number of additional influencing factors (see e.g. [30, 31], or [32]). Spatial differentiation is also the essence of polarization theories [33, 34], which, like NEG or structuralist, institutional and other theories of regional development generally defined at the global level, are also applicable to socio-spatial processes at various size levels, including microregional (more detailed information in e.g. [35, 36]).

However, with the increasing concentration/differentiation, there is also a growth in the integration of spatial units which creates complex systems (regions) including the core and the periphery (see e.g. [33] or [34]), These processes are applied at hierarchically different size levels (see e.g. [5–7], or [15]). Hence, the settlement system and its complexity have both horizontal and hierarchical (vertical) dimensions. Therefore, it's a complex socio-spatial process, including a whole scope of interactions, the result of which is a complex and hierarchically differentiated system of mutually overlapping ties holistically covering the entire spectrum of human activities in space, acting differently at various hierarchical size levels.

The relationships between the individual elements of the settlement system have a hierarchical character with diffusive processes occurring between them, that ensure the spread of development potential, trends, and innovations (see e.g. [37]). Spatial diffusion is one of the main differentiating processes forming the settlement system, although the nature of its action varies depending on the hierarchical size level and depending on the essence of the given phenomena. As a result of the hierarchization process, in the settlement (or in the economic) system, new bearers of differentiation appear, but in the case of developmentally lower phenomena, due to diffusion processes, there is a nivelisation in interregional differences [6, 38].

In perspective of the organization of the settlement system, these processes create specific regional structures (regions) at each hierarchical level (macroregional, mezoregional, microregional or sub-microregional). These regions are internally strongly integrated by processes of a certain type (depending on the hierarchical size level), and at the same time, they are relatively relationally closed to the outside.

The study of spatial interactions of the settlement system in order to understand sociospatial region-forming processes is based on the principles of gravity theories generally assuming a decreasing influence of the center on its surroundings with increasing distance, while also depending on the size/importance of the given center and the activities concentrated in it. Every element of the settlement system (or place of economic activity) interacts with its surroundings and represents both a generating role (supply) and an attracting role (demand). Gravitational models of spatial interactions built on these models differ mainly in the variety of parameters. Which take into account (see e.g. [17, 39–44]).

The interconnectedness of individual processes within the settlement system is so complex that it cannot be easily identified or even measured in any way. However, the external manifestations of these processes are measurable. These take the form of spatial interactions and manifest themselves as commuting relationships different at various hierarchical levels (see e.g. [5–12, 14, 16, 45, 46], or [37]). They are thus realized through transport links which have been measured for long period by transport geographers (see e.g. [1, 3, 4, 11, 13, 15, 47], or [22]). These interactions are not only quantitatively different but at the same time qualitatively different at various hierarchical levels of size-order. In general, the described spatial interactions can be called population mobility. It contains not only the actual journeys but also a reflection of the overall spatial pattern of each individual's behavior. Hence, the mobility/commuting behavior of an individual takes into account the intensity, frequency and repeatability of certain elements of spatial behavior [48, 49], This also determines the hierarchical position of the commuting destination and its relationship to the place of origin. In conclusion, the spatial behavior of the population completely reflects the relationships and processes within the settlement system and therefore, it is a suitable object of measurement for their explication.

A wide range of tools can be used in both local and large-scale statistical surveys focused on the traffic behavior/mobility of residents, such as questionnaire surveys, traffic diaries, GPS loggers, measuring passengers transported by individual modes of transport, or measuring traffic intensity (see e.g. [4, 11, 15, 17, 22], or [23]). In the Czechia, queries about commuting to work and schools are even part of the census, however, these available statistics have a low return in recent censuses, and it is assumed that up to 40% of commuting flows are missing from the census statistics. With this in mind, a significant potential for mobility measurement can be seen in the use of the geolocation data of mobile operators. Due to the high penetration of the population by mobile devices, and the possibility of tracking movement in unlimited random periods, this approach combines both the advantages of population-wide data collection and detailed (movement tracking) studies as well (see e.g. [21, 50, 51] or [47]).

The essence of the method are the records in the geolocation network, which are created every few minutes by every device joined to the GSM network via SIM cards. Determining the location is approximate by this technique, as only the transmitter (BTS) that registered the recording is precisely located. From the signal coverage map of individual transmitters, the approximate location of the SIM card can be deduced with an accuracy of hundreds of meters in urbanized areas and up to a few kilometers in rural areas.

In order to obtain this type of data, it is necessary to set up a complex mechanism of tools analyzing more than 10 million SIM cards (the case of the Czechia), each of which produces thousands of records within the measured periods. In addition to the technical solution and considerable computing capacity required for Big Data processing itself. Besides, it is also essential to consistently establish methodological procedures for the preliminary processing of primary records for the creation of databases of citizens' mobility/travel behavior.

In the past, it was the method of data claiming that was the main obstacle in the use of geolocation data concerning their low validity (me more detail e.g. [19, 20], or [21]). Research carried out in the past in the field of data analysis of mobile operators had to solve problems of representativeness of data and their evidential value when generalizing to the population (see e.g., [23], or [2]). Although this shortcoming is not an obstacle for use in research from the technical fields aimed at measuring the volume of journeys made or data transmitted, in the field of social geography the question of the generalizability of data to the population and the projection of spatial patterns of behaviors onto entire society in space and time is absolutely essential. For this purpose, a unique model was created, including a complete range of interconnected processes, which captures the mobility of the population and projects it on the social and settlement networks.

2 Model of Data Acquisition Process

The whole model (see Fig. 1) is based on the presumption that mobile phones move together with their users for most of the day (mentioned in e.g., [20], or [21]). Based on this assumption, the model eliminates records created by other devices than mobile phones, thereby largely eliminating the problem of duplicate records of a single user of multiple devices. Similarly, rarely used SIM cards that do not make enough records in the network are neglected. Furthermore, the assumption of high penetration of the population by mobile phones is also crucial. In general, it can be concluded that in contemporary societies of developed countries, both assumptions are fulfilled. The **reduction of the base dataset** of records in the network ensured that approximately 10.3 million SIM cards were included in the overall analysis in each of the 4 realized measurements. We distinguish two elemental states that SIM cards can acquire: stay or movement. As part of the records in the geolocation network, we only have information about the "stay" and the "movement" is detected as inferred, based on a change of stay. To demonstrate, if the SIM card (as part of a periodic update) logs in twice consecutively to different BTS transmitters, the location of stay changes, and it can be inferred that the SIM card has moved in the meantime.

When detecting the movement or stay of the SIM card, the proposed model must solve the problem of the inconsistency of the administrative boundaries of the municipalities with the boundaries of the signal transmitter´s service area. In reality, it is common for one BTS transmitter to serve several municipalities or their parts at the same time. In addition, overlapping of service areas of different transmitters is common as well. Based on the signal coverage maps of individual BTS transmitters, a "cell network" is created. The network of cells is continuous, without residues and overlaps covering the whole territory of Czechia. A cell represents an area in which a located SIM card will be served (with the highest probability) by one particular transmitter. There are usually 3 antennas on each BTS transmitter oriented in different directions (at 120° angles), and each of them has its own cell. The territorial detail of this network is therefore very high.

The **detection of stay/movement** itself and its assignment to particular territorial units (municipalities) occurs via the "cell-mapping process". In fact, it is an advanced algorithm designed just for the needs of this purpose. This tool distributes the measured records between specific settlements (municipalities) according to the amount of intravillan (build-up area) of each settlement extending into a specific cell (service area of the signal transmitter) and also reflecting the population density of each settlement. It is a complex mathematical computation that compares all defined cells with the land-use coverage map, especially the built-up areas of municipalities, of which there are more than 6 thousand in Czechia. In this case, any building or another way urbanized area is considered to be a built-up area, except traffic roads and railways. If several built-up areas extend into one cell, the mathematical algorithm evaluates the size of the built-up area of each municipality extending into the given cell. In this approach, the population density of each municipality is also included in the computation (obtained from official state statistics). Taking population density into account is an essential element for making the modeling more accurate, as a housing estate can interfere with a cell within one municipality and an industrial complex from another municipality, which, although they may occupy the same area, can be expected to have different occupancy by residents. Individual SIM cards with defined "stay" in a given cell are subsequently distributed among individual municipalities completely automatically based on this computation.

The accuracy of the distribution of SIM cards to the territory (cell-mapping. Process) mainly depends on the quality of the map showing the signal coverage network, i.e. service cells. In the case of our model, the situation is even more complicated, as the data is obtained from all 3 mobile network providers in Czechia, while each uses its own network of BTS transmitters, and therefore has its own signal coverage maps. In addition, there are several layers of these cell networks for each provider depending on the type of used technology (3G, 4G, 5G). The inputs to this calculation are varied, and the computing algorithm of the "cell-mapping process" repeatedly recalculates the entire task in case of any change in the inputs, e.g. due to technical repairs to part of the BTS network. In our specific case, the inputs changed not only between individual measured periods, but also between individual days of each measurement.

In conjunction with the clustering algorithm, cell mapping process can eliminate the unwanted effects of so-called "cell jitter" This represents random switching between neighboring transmitters serving the same location. It occurs especially when one transmitter is overloaded, or i.e., if weather conditions cause a change in the signal strength of individual neighboring transmitters in a given location (mostly locations near cell borders). This detects a change of SIM cards position and therefore "movement" in the resulting dataset, while the SIM card does not actually move at all. (for more details see [19], or [20]).

The clustering algorithm is an extension to the cell-mapping process. Monitors and recalculates individual short-term switching between neighboring cells (transmitters). Only those cases where the SIM card is logged in repeatedly in another cell, and the stay lasts at least 30 min are considered a change of stay (e.i. movement). It considers all other logins to other transmitters within its cluster of neighboring cells as a continuation of the current stay.

Databases obtained through this model also removes other undesirable elements that worsen the evidential value of the data, such as the **share of virtual operators** using the network, ownership of **multiple SIM cards** by one user, or on the contrary, not owning a mobile phone and thus **no SIM cards usage**. These effects cannot be technically (physically) eliminated or excluded from the dataset in any way, as they are a natural part (feature) of these data. For that reason, mathematical adjustments were made to compensate for these negative effects. As a result of the application of these compensations (application of coefficients unique to each municipality), a complex model is created, which no longer represents the number of measured SIM cards but the expected number of moving people. These compensations were based on a questionnaire survey on ownership/non-ownership of SIM cards (more than 8,000 respondents) and at the same time, on the measurement of territorial differences in shares of virtual operators in the signaling network. All described aspects are taken into account by the model of data acquisition, the process of which is shown in Fig. 1.

In addition, the model also contains **relocation mechanisms** that are capable to correct retroactively model errors in assigning the records to individual territorial units (e.i. cell-mapping process). The relocation process consists of several steps (automated and manual). Municipalities, where the cell mapping process may have failed during the detection of stay, are identified in this step. Thanks to the relocation mechanism, which contains elements of machine learning, the entire data acquisition model can adjust its settings even between different time measurements.

The relocation mechanism responds to the fact that the main problem with the entire data acquisition model is the impossibility of reliably verifying its validity, as there is no reference data to compare. With this in mind, there is the possibility of comparing only one parameter, namely the expected number of residents (defined in more detail later in the text) and the number of municipality inhabitants according to official state statistics. Although each of these data measures a different indicator, they both monitor the same phenomenon, i.e. the number of people living in the locality. Both data should not be completely identical, but at the same time, they cannot be diametrically different. Altogether, in more than 90–95% of all municipalities, the differences compared to official statistics on the number of inhabitants were minimal. In cases where the differences are significant (greater than ± 25%), a relocation mechanism was applied. In its automated part, it is searching for cases where SIM cards may have been incorrectly distributed (from cells) to the territories of municipalities. In practice, these errors were manifested mainly by neighboring extremely undervalued and extremely overvalued municipalities. In such cases, the mechanism proposed to move individual measured SIM cards. However, after the application of these transfers, the affiliation of some SIM cards to municipalities will change and, for that reason, different coefficients should subsequently be implemented on them within the data acquisition model (described in Fig. 1 – left part), therefore the individual steps are repeated again (from the cell-mapping process). Automatic relocation can ensure that municipalities that show extreme differences in the number of residents and the official population will be left with 1–2%. In these cases, it is subsequently necessary to individually assess whether these are local anomalies caused by the model setting or a condition that is justified in geographical reality. For example, smaller municipalities, which are, however, important tourist centers or spas, showed in individual measurements up to several times higher number of residents than their official population. If these data are justified in this way, these anomalies represent important information that, on the contrary, must be preserved in the data.

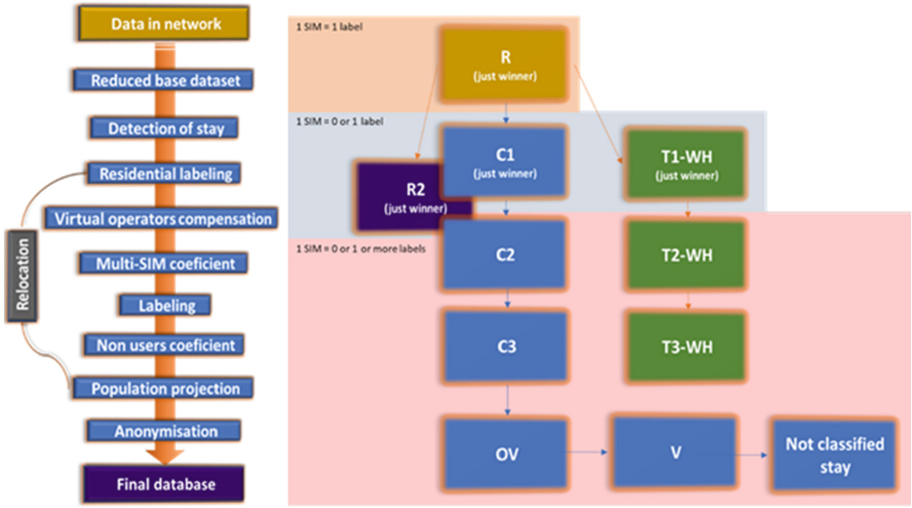


Fig. 1. Model of the data acquisition process and the mechanism of the labeling process

In addition to the mentioned data operations, in the final phase, a **projection on the population** is made, which ensures the representativeness of the data model according to the population. Hence the entire model is calibrated to the number of the official population of the state. In addition, the calibration was carried out to an absolutely minimal extent. The data measured by the model differed from the offical population of the state by only a few percentage units.

At the same time, sufficient **anonymization** of the final data is ensured. Consistent multi-phase anonymization ensures the impossibility of identifying specific persons with the specific data in the datasets. As a result of this process, some data is also artificially slightly deliberately altered for the purpose of ensuring the impossibility of potentially associating with specific persons. In fact, it potentially only applies in the case of very small municipalities with a small number of residents and a small volume of interactions.

This model is flexible in terms of the output databases produced. According to the primary setting, it produces a total of 15 attributes on the territorial detail of individual municipalities, structured into 3 basic interconnected datasets: a) statistical data for individual municipalities and characteristics of their residents, b) OD matrix showing commuting directions taking into account a total of 6 types of commuting intensity/types, and c) the average number of currently present population in every hour of the week (24/7) in each municipality with a breakdown by particular attributes.

The method of assigning attributes/labels to individual users in the network is also unique.it occurs within the **labeling process** as one of the steps of the model of the data acquisition (Fig. 1) Basically, the method does not monitor the actual volumes of the trips made but analyses the commuting rhythms and the overall spatial commuting behavior of each SIM card user during each of a total of 4 measured period (each 28 days). The list of individual labels and the process of their assignment is indicated in Fig. 1 (right part). Specifically, these are resident attribute (R1) i.e., the place where people most often spend the night, further there are 3 types of commuting - daily, weekly, occasional (C1-C3), overnight visitor (OV), one-time visitor (V), second residence (R2), and not classified stay. During the monitored period (28 days), "labels" of attributes are assigned to each SIM card according to its unique pattern of spatial behavior. The assignment of labels is carried out based on tracking the visited locations of individuals and analyzing the frequency of visits, their number, the total time spent at the destination, the periodicity/repeatability of these movements, and even the time of day when the movement/stay was realized. Each individual (SIM card user) can only have one label for each municipality, but he can have several labels for several municipalities - can be a resident in one municipality, commute to another for work or school, commute for services, or be an occasional commuter to another etc. The entire model of the data acquisition and label assignment is underway on two dimensions. One is general, and the other is limited to working days during working hours (Mon-Fri, 6 am–6 pm). These labels are assigned only to those users who fulfill specific characteristics of the given type of relationship in the specified time period (WH - working hours). Thus, two sets of labels are assigned independently, and the entire assignment algorithm s underway twice.

As a result of the labeling process the output databases themselves do not indicate specific measured values for a certain day or period, but each attribute represents basically the number of people who reports a given type of behavior. This is no longer the geolocation data itself, but a summary of time-spatially aggregated statistics about geolocation data.

3 Usability of the Resulting Databases

The resulting databases are a unique source of information on the residence and movement of persons in the long term. The data produced by this model are currently available from a total of 4 measurements from 2021 - 2023 and cover the specifics of mobility patterns in all individual seasons. It thus represents a fully comprehensive tool for creating a mobility model of the inhabitants and evaluating their travel behavior patterns.

It turns out that the stays and movements detected by this method are comparable in their accuracy to the reference data of national statistics. In addition, thanks to its volume, detail, and permanent monitoring, it can also identify the nature of relationships that cannot be identified by conventional approaches.

The identified territorial relations very faithfully describe the specifics of the spatial organization of the settlement system, its hierarchical levels, and mutual interactions. It thus expresses the territorial structure of social differentiation.

One of the most significant manifestations of the territorial dimension of the residential structure is the daily commute to work, to schools, or other activities that take up the main part of the day. These processes can be specifically captured in the data in two ways. 1) using data from the OD matrix showing links of inter-municipal commuting. These links can be divided according to the types of commuting relationships, which depend on the definition of the individual distributed labels. These data are the foundation for the creation of functional microregions based on natural interactions. An example of such a mapping of commuting relations can be seen in Fig. 4, which will be explained further in the text. 2) Use of the data from the dataset on the number of people present in the municipality in individual hours. From these data, a municipal occupancy model can be created, which is presented in Fig. 2 on the example of the Czech capital Prague and its wider hinterland (Central Bohemia region).

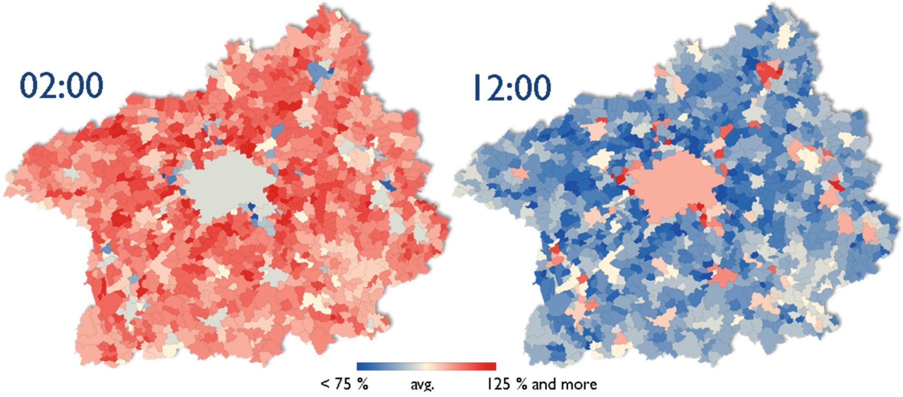


Fig. 2. Mobility model for the Prague and the Central Bohemia region. Current attendance of individuals in the municipalities during day according to the average occupancy of each municipality.

Mobility between the hinterland and a strong core such as Prague is very evident in the daily commute. For each municipality, the model points to its basic commuting parameter, namely the prevailing incoming ratio or outgoing ratio. The mobility model here shows the relative values of the incoming/outgoing commuting ratio during the day. For that reason, changes in the number of people present in the village during the day are more pronounced in small villages, where even a small number of moving people causes a significant change. On the other hand, in large cities/towns, especially Prague, even the enormous increase in the number of individuals during the day (up to 200,000 commuters) means relatively little change. In addition to Prague, other large cities in the region with a population of around 20–60 thousand also appear as important commuting centers. Incoming commuting settlements are also concentrated in the immediate vicinity of Prague, where important employers are located. Smaller municipalities at a greater distance then mainly fulfill the function of residential settlements with prevailing outgoing over incoming commuting. This regional mobility model shows the basic elements of the spatial interactions of the core and peripheral areas of the micro-regional level of the settlement system, which are precisely defined by the daily commute to work and school.

In more detail, we can follow the daily rhythm of each of the 6,254 municipalities in the Czech Republic. In addition to the number of people present at every hour of the day on all days of the week (24/7), we can also get a deeper breakdown into particular types of individuals from the model. Certain types of individuals are also dependent on the assigned attributes and therefore represent the relationship of each present person to the given place. The different daily or weekly rhythms of particular municipalities can be seen in Fig. 3, where an example of 4 municipalities of different sizes and different regional specifics is given.

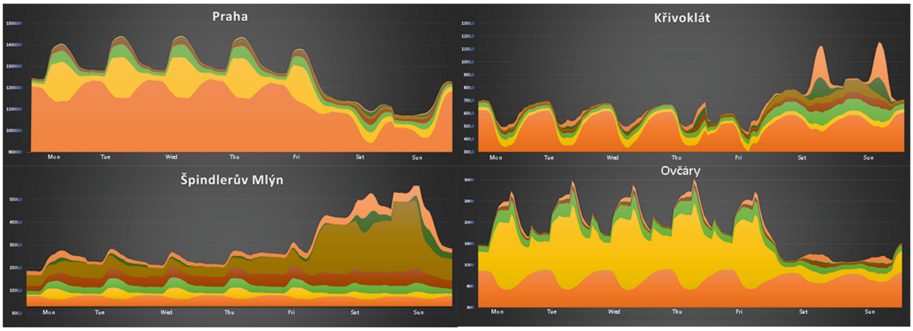


Fig. 3. Currently present individuals in the municipality (every hour of the week – 24/7). Divided into categories (sorted from the bottom): resident (R1), commuters I., II., and III. Type (C1-C3), overnight visitor (OV), visitor (V), transiting and other (not classified stay). The y-axis scale is adjusted individually for each case to provide a relative comparison.

The capital city of Prague has a completely different and relatively complementary weekly rhythm from Špindlerův Mlýn, which is a well-known ski resort. On the other hand, the daily rhythm on weekdays shows the complementarity of Prague and the small village of Křivoklát (Central Bohemia region), which is known as a one-day trip tourist center (medieval castle on its territory). A very specific case is the very small village of Ovčáry (Central Bohemia region), on whose territory a huge automotive factory is located. Its daily rhythm shows very clearly the specific hours when work shifts change in this factory as an extreme increase in the number of people present with regular periodicity.

From another point of view, the structure of the present population shows that, in the case of Prague, residents dominate and the daily rhythm is defined by 1st type of commuters (daily commute). Conversely, in Špindlerův Mlýn, local residents are the minority in the village throughout the week, and oversleep visitors dominate. In Křivoklát, during working days the regime of the village depends on the behavior of residents, however, on the weekend the mode is influenced by 2nd type commuters, one-time visitors, and people without an assigned label (random visitors or transiting persons). In the case of Ovčáry, the main component of the present population is daily commuters to the industrial area.

Such data are very valuable, especially in the agenda of spatial planning, development of the technical infrastructure, or in crisis management. They provide a piece of real, accurate, and very detailed information about the mobility needs of the residents of all municipalities of the state.

Processed databases enable visualization using GIS web tools, and make the data freely accessible. These advanced cartographic tools visualize data on the direction and strength of individual commuting relationships and supplement them with other interactive graphic elements (charts, tables, etc.).

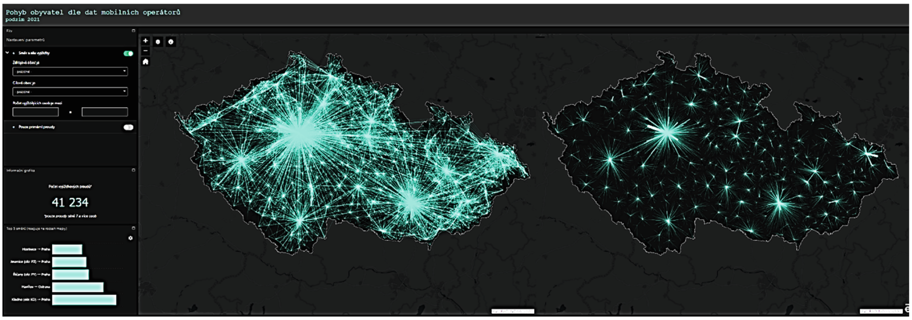


Fig. 4. A web map application showing the daily inter-municipal commuting links. Based on the OD matrix of the number of people commuting between municipalities, with the possibility of breakdown to particular types of commuting according to the assigned labels.

In the given example of the output from this web map application (see Fig. 4), 2 cases are displayed. 1) Case of all recorded commuting links (inter-municipal), of which there are more than 41 thousand in the database. And 2) case of primary commuting links for each municipality. Especially from the example of the map of primary commuting links (on the right), it is clear high internal integrity within the regions (clusters of individual links around important centers) and external relative closeness (disconnection between clusters by primary links). This map application will be available from this link after the approval of all project outputs.

These data enable the creation of further, advanced spatial analyses, for example, to define regions of territorial concentration and, based on them, to create the overall socio-economic regionalization of the state (see e.g., [2, 18, 21], or [25]). The freely accessible web application allows the general public to display this data, use, analyze, or download it and also allows various forms of sharing it. It is also possible that new web-based map applications using the primary dataset as a web map service can be created based on this data.

4 Resume

This article shows the usability of geolocation data of mobile operators and a new model for their acquisition for a purpose of population mapping and identification of patterns of their traffic behavior. It turns out that the output databases created by applying the described data acquisition model enable subsequent applications of geographic analyses identifying functionally integrated regions and their central areas at different hierarchical levels. Based on the principle of commuting to certain centers, the intensity and volume of these interactions, relatively closed (in terms of functional closeness of the interactions) and internally integrated regions might be recognized.

Primarily, the method is set to identify functional micro-regional commuting links. Microregions are territories in which a resident should be able to secure all his daily activities necessary and important for his everyday life. Their centers are primary commuting destinations for their surroundings and provide a sufficient range of job opportunities, primary and secondary education, health services, shops, etc. Visiting centers of a higher hierarchical levels providing services of a higher grade, however, is not needed daily. Nevertheless, thanks to the robustness of the analyzed data, it is also possible to define centers of higher (mezoregional) or lower (submicroregional) levels. Therefore, this method enables the implementation of a complete socio-economic regionalization of the state at individual hierarchical size-levels, including their hierarchical relationships. Socio-economic regionalization using data from this model represents natural commuting regions based on the assessment of the natural concentration of people in the territory based on their long-term monitoring. The concentration processes identified in the data reflect both commuting flows of varying intensity, i.e., the movement of individuals, as well as the concentration of people in a specific place during individual hours of each day of the week, i.e., the stay of individuals.

This approach is used not only for regionalization itself in the sense of academic research on the development of the settlement structure of the state, and its hierarchical organization but also for practical use, especially in the field of public and private service delivery. For individual municipalities, information on the stay and movement of people in the territory appears to be essential for the implementation of policies of regional development, urban planning, or crisis management in cases of solving risk treatments. Localization of public administration services should be implemented as part of the implementation of this project. However, the great potential of using this data is also for the localization of private services such as local shops, pharmacies, delivery service boxes, ATMs, and possibly even educational and medical facilities. All public infrastructure as well as the technical infrastructure of each municipality should thus respect the natural processes of concentration and rhythms of commuting behavior of citizens of surrounding areas. As a result, these steps indicate a significant increase in the standard of living of citizens and an improvement in public and private service delivery.

This approach was used in the Czech Republic for a comprehensive revision of the spatial units of the public administration structure. The purpose of this activity was to harmonize the administrative units with natural commuting regions. Particularly, the aim was to ensure that public administration offices were located where people naturally concentrated. This leads to streamlining and deconcentrating of the public administration and its adaptation to the needs of citizens. Based on this application example, it is also possible to conclude about the transferability of this approach and its applicability both in other territories (states) or in other scientific fields.

References

1. El-Geneidy, A., Levinson, M.: Access to Destinations: Development of accessibility Measures. Access to destinations study series, Report No. 1. Minnesota Department of Transportation (2006)
2. Halás, M., Blažek, V., Klapka, P., Kraft, S.: Population movements based on mobile phone location data: the Czech Republic. J. Maps **17**(1), 116–122 (2021). https://doi.org/10.1080/17445647.2021.1937730
3. Halás, M., Kladivo, P., Šimáček, P., Mintálová, T.: Delimitation of micro-regions in the Czech Republic by nodal relations. Moravian Geograph. Reports **18**(2), 16–22 (2010)
4. Halás, M., Klapka, P., Kladivo, P.: Distance-decay functions for daily travel-to-work flows. J. Transp. Geogr. **35**, 107–119 (2014). https://doi.org/10.1016/j.jtrangeo.2014.02.001
5. Hampl, M., et al.: Geography of Societal Transformation in the Czech Republic. Prague, Charles University in Prague, Faculty of Science, p. 242. (1999)
6. Hampl, M.: Geografická organizace společnosti v České republice: transformační procesy a jejich obecný kontext. Univerzita Karlova, Praha, 147 (2005)
7. Hampl, M., Marada, M.: Sociogeografická regionalizace Česka. Geografie **120**(5), 397–421 (2015). https://doi.org/10.37040/geografie2015120030397
8. Horňák, M., Kraft, S.: Functional transport regions in slovakia defined by passenger-car traffic flows. Mitteilungen der Osterreichischen Geographischen Gesellschaft **157**(1), 109–128 (2015)
9. Jaroš, V.: Social and transport exclusion. Geogr. Pol. **90**(3), 247–263 (2017). https://doi.org/10.7163/GPol.0099
10. Klapka, P., Halás, M., Netrdová, P., Nosek, V.: The efficiency of areal units in spatial analysis: Assessing the performance of functional and administrative regions. Moravian Geograph. Reports. **24**(2), 47–59 (2016). https://doi.org/10.1515/mgr-2016-0010
11. Kraft, S., Blažek, V., Marada, M.: Exploring the daily mobility rhythms in an urban environment: Using the data from intelligent transport systems. Geografie **127**(2), 127–144 (2022). https://doi.org/10.37040/geografie.2022.004
12. Kraft, S., Halás, M., Vančura, M.: The delimitation of urban hinterlands based on transport flows: a case study of regional capitals in the Czech Republic. Moravian Geograph. Reports **22**(1), 24–32 (2014). https://doi.org/10.2478/mgr-2014-0003
13. Kraft, S., Květoň, T., Blažek, V., Pojsl, L., Rypl, J.: Travel diaries, GPS loggers and Smartphone applications in mapping the daily mobility patterns of students in an urban environment. Moravian Geograph. Reports **28**(4), 259–268 (2020). https://doi.org/10.2478/mgr-2020-0019
14. Kraft, S., Marada, M., Popjaková, D.: Delimitation of nodal regions based on transport flows: case study of the Czech Republic. Quaestiones Geographicae **33**(2), 139–150 (2014). https://doi.org/10.2478/quageo-2014-0022
15. Marada, M.: Transport and geographical organization of society: case study of Czechia. Geografie **113**(2), 285–301 (2008)
16. Marada, M., Komárek, M., Šimbera, J.: Metropolitan polynodal cores as the basis of the new regional organization of Czechia. Geografie **128**(1), 49–74 (2023). https://doi.org/10.37040/geografie.2023.004
17. Marada, M., et al.: Fast connections among metropolitan areas: impact of (new) accessibility on labour market. Project TAČR TB0500MD005 (2016)
18. Marada, M., Zévl, J., Petříček, J., Blažek, V.: Interurban mobility: Eurythmic relations among metropolitan cities monitored by mobile phone data. Appl. Geograph. 156 (2023)https://doi.org/10.1016/j.apgeog.2023.102998
19. Mazouch, P. a kol.: Limits of data from mobile sites in statistical surveys of Czech statistical office. Project TAČR TD03000452 (2017)

20. MV ČR: Improvement of preconditions for decentralisation and availability of public administration in the territory. EAA grants, GG-PDP1–001. (2020)
21. Novák, J., Ahas, R., Aasa, A., Silm, S.: Application of mobile phone location data in mapping of commuting patterns and functional regionalization: a pilot study of Estonia. J. Maps **9**(1), 10–15 (2013)
22. Nutley, S.: Monitoring rural travel behaviour: a longitudinal study in Northern Ireland 1979–2001. J. Transp. Geogr. **13**(3), 247–263 (2005). https://doi.org/10.1016/j.jtrangeo.2004.07.002
23. Štraub, D., Jaroš, V.: Free fare policy as a tool for sustainable development of public transport services. Human Geograph. – J. Stud. Res. Human Geograph. **13**(1), 45–59 (2019). https://doi.org/10.5719/hgeo.2019.131.3
24. Šveda, M. - Barlík, P.: Daily commuting in the Bratislava metropolitan area: case study with mobile positioning data. Appl. Geograph. **4**(4), 409–423 (2018)
25. Šveda, M., Madajová, M.S.: Estimating distance decay of intra-urban trips using mobile phone data: the case of Bratislava, Slovakia. J. Transp. Geograph. **107**, 103552 (2023). https://doi.org/10.1016/j.jtrangeo.2023.103552
26. Christaller, W.: Die zentralen Orte in Suddeutschland: Eine okonomisch-geographische Untersuchung uber die Gesetzmassigkeit der Verbreitung und Entwicklung der Siedlungen mit stadtischen Funktionen. Jena (1933)
27. Lösch, A.: The Economics of Location. Yale University Press, Translated by William H. Woglom (1940)
28. Zipf, G.K.: Human Behaviour and the Principle of Least Effort. Addison - Wesley Press, Cambridge (1949)
29. Isard, W.: Location and Space-economy; a General Theory Relating to Industrial Location, Market Areas, Land Use, Trade, and Urban Structure. Published jointly by the Technology Press of Massachusetts Institute of Technology and Wiley, Cambridge (1956)
30. Krugman, P.: Geography and Trade. MIT Press, Cambridge, MA (1991)
31. Krugman, P.: Increasing returns and economic geography. J. Polit. Econ. **99**, 483–499 (1991). https://doi.org/10.1086/261763
32. Fujita, M., Krugman, P.: The new economic geography: past, present and the future*. Papers Regional Sci. **83**(1), 139–164 (2004). https://doi.org/10.1007/s10110-003-0180-0
33. Friedmann, J.: A general theory of polarized development. In: Hansen, N.M. (ed.) Growth Centres in Regional Economic Development, pp. 82–107. Free Press, New York (1966)
34. Perroux, F.: Economic space: theory and applications. Q. J. Econ. **64**(2), 89–104 (1950)
35. Higgins, B., Savoie, D.J.: Regional development theories and their application transaction publishers. New Brunswick (1995). https://doi.org/10.4324/9781315128269
36. Blažek, J., Uhlíř, D.: Teorie regionálního rozvoje: nástin, kritika, klasifikace. Praha: Karolinum (2002)
37. Hägerstrand, T.: Innovation diffusion as a spatial process. Translated by A. Pred. Chicago: University of Chicago Press (1967). https://doi.org/10.1111/j.1538-4632.1969.tb00626.
38. Derudder, B., Taylor, P.J., Witlox, F., Catalano, G.: Hierarchical tendencies and regional patterns in the world city network: a global urban analysis of 234 cities. Reg. Stud. **37**(9), 875–886 (2003). https://doi.org/10.1080/0034340032000143887
39. Reilly, W.J.: Methods for the study of retail relationships. University of Texas Bulletin, Monograph **4**, 2944 (1929)
40. Tobler, W.: Spatial Interaction Patterns. IIASA Research Report. IIASA, Laxenburg, Austria: RR-75–019 (1975)
41. Fotheringham, A.S., O'Kelly, M.E.: Spatial Interaction Models: Formulations and Applications. Kluwer Academic Publishers, Dordrecht (1989)
42. Haynes, K.E., Fotheringham, A.S.: Gravity and Spatial Interaction Models. Reprint. Edited by Grant Ian Thrall. WVU Research Repository (1985)

43. Daly, A.: Estimating choice models containing attraction variables. Transp. Res. Part B: Methodol. **16**, 5–15 (1982). https://doi.org/10.17226/14133
44. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. nature **453**(7196), 779–782 (2008). https://doi.org/10.1038/nature06958
45. Berry, B.J.L., Garrison, W.L.: Functional bases of the central place hierarchy. Econ. Geogr. **34**, 145–154 (1958)
46. Berry, B.J.L.: City size distribution and economic development. Economic Development and Cultural Change 6, str. 573ñ576 (1961)
47. Scherrer, L., Tomko, M., Ranacher, P., Weibel, R.: Travelers or locals? identifying meaningful sub-populations from human movement data in the absence of ground truth. EPJ Data Sci. **7**(1), 1–21 (2018). https://doi.org/10.1140/epjds/s13688-018-0147-7
48. Hannam, K., Sheller, M., Urry, J.: Editorial: mobilities, immobilities and moorings. Mobilities **1**, 1–22 (2006). https://doi.org/10.1080/17450100500489189
49. Sheller, M., Urry, J.: The new mobilities paradigm. Environ. Plann. A: Econom. Space **38**(2), 207–226 (2006). https://doi.org/10.1068/a37268
50. Ahas, R., Silm, S., Järv, O., Saluveer, E., Tiru, M.: Using mobile positioning data to model locations meaningful to users of mobile phones. J. Urban Technol. **17**(1), 3–27 (2010). https://doi.org/10.1080/10630731003597306
51. Steenbruggen, J., Borzacchiello, M.T., Nijkamp, P., Scholten, H.: Mobile phone datafrom GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. GeoJournal **78**(2), 1–21 (2013). https://doi.org/10.1007/s10708-011-9413-y

Returning Home Strategy Analysis Using Mobile Sensing Data in Tohoku Earthquake

Zhiwen Zhang[1], Hongjun Wang[2], Zipei Fan[1,2](✉), and Xuan Song[1,2]

[1] The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba, Japan
{zhangzhiwen,fanzipei,songxuan}@csis.u-tokyo.ac.jp
[2] Southern University of Science and Technology, Shenzhen, China
wanghj2020@mail.sustech.edu.cn

Abstract. In recent decades, there has been a significant increase in the frequency and intensity of natural disasters. Such catastrophic events often result in large-scale population movements and evacuations. Analyzing these human activities is crucial for effective planning of disaster control, and ensuring long-term social stability. While some research has been conducted on post-disaster analysis, particularly focusing on big earthquakes [15,22], very few studies have taken into account the influence of personal factors on decision-making. Understanding the key factors that drive individuals to choose a strategy, such as returning home, after a big earthquake is essential for comprehending human decision-making in such situations. Additionally, a considerable number of people may remain in companies or shelters due to the disruption of transportation networks. However, conducting such research is challenging due to the lack of big human mobility data. Furthermore, identifying the key factors that individuals consider when making decisions to return home after a big disaster is critical. To address these challenges, this study utilizes smartphone location data to track people's movements. A large and diverse dataset was collected during the Tohoku earthquake in Japan in 2011, allowing for the discovery of grid-based regions with different functions based on POI distributions in a region. The analysis conducted in this study aims to explore the fundamental laws governing human mobility following disasters. This paper is an extended version of our previous lightning talks [24].

Keywords: Earthquake · Big mobility data · Explainable knowledge · Decision-making strategy analysis

1 Introduction

On March 11, 2011, a massive "9.0" earthquake struck the Pacific Ocean, approximately 130 km off Sendai City, Japan. This event serves as a case study for our research [8,12,13]. We aim to jointly model the historical visitation patterns of functional areas and decision-making behavior following such a significant earthquake. Numerous studies have explored mobility patterns using historical

R. Akerkar (Ed.): SAIDD 2023, CCIS 1810, pp. 16–24, 2024.
https://doi.org/10.1007/978-3-031-53770-7_2

visit data collected through smartphones. These studies have been instrumental in analyzing people's behavior during disasters [2,3,17,18]. For instance, [19] developed a probabilistic model to simulate population evacuation across complex geographic features in Japan in response to future disasters. In addition, [3] demonstrated that location data can be utilized to track the distribution of earthquake risk areas, people's emergency responses, and overall behavioral patterns during disasters. However, previous works often overlooked the significance of functional regions [23] in understanding human mobility during disasters. These functional regions contain crucial information that can enhance our comprehension of human behavior in such scenarios. In conclusion, our study seeks to bridge this gap and includes an empirical prediction method to estimate how individuals chose to return home following the Tohoku earthquake. By jointly considering historical visitation data of functional areas and decision-making behavior, we aim to gain deeper insights into human mobility patterns after a major earthquake event.

More than a decade ago, seldom work focused on analyzing the human mobility pattern in natural disasters. Fortunately, mobile sensing technologies have been widely applied in various fields. This resulting multiple works [4,21] have been proposed to predict human mobility in large-scale disasters (such as earthquakes, tsunamis, and hurricanes). However, they may lack some necessary to analyze a human mobility strategy. Since large-scale disaster is rare, it's a challenge to understand human mobility from such data. Although there are a lot of related works for analyzing [15,22] after a big earthquake, few pieces of research consider the influence of personal factors on decision-making. In addition, the knowledge of what key factors impel a person to choose a returning home strategy is important for analyzing the human decision after a big earthquake.

Here, we present an overview of our framework. The system comprises two main steps: data mining and explainable knowledge.

In the data mining step, we begin by constructing a staying cuboid that collects the number of staying points generated during returning home trajectories for each region. And we could obtain different region topics of those historical personal visits by analyzing surrounding POI distributions. Regarding the transition mode, we employ a map-matching algorithm [16] for each trajectory to map the GPS points onto road segments. This allows us to determine the transportation modes based on features within each segment, such as railways, expressways, and footpaths.

After gathering the necessary data, we define the return home mode and departure time from the workplace as a multitask objective function. This approach helps us generate an explainable result to support our analysis effectively. Finally, we present the feature importance histogram to demonstrate the effectiveness of our proposed framework.

2 Related Work

The topics of human behavior prediction during disaster (e.g., crowd panics [14], fires [6,9,10], floods [5]) have received numinous attention in recently with a focus on small-scale or short-term emergencies. However, research on the dynamics of population movements across the country in large-scale disasters (such as earthquakes, tsunamis, and hurricanes) is minimal [11]. Meanwhile, the extensive volume of work has brought about a significant shift in the mobility pattern [2,3,17,18] derived from historical data collected by smartphones. This wealth of data is now being harnessed to analyze human behavior during stages of disasters. One noteworthy application is demonstrated by [19], who devised a comprehensive probabilistic model to simulate population evacuation across intricate geographic terrains in Japan, foreseeing potential future calamities. Moreover, by leveraging location data, [3] discovered that it is feasible to monitor the distribution of earthquake risk areas, assess people's emergency responses, and track their behavior during such critical events. Reference [20] points out that although people now will go further and faster than before, most of their time in people's lives will still locate in important places, such as home, workplace. However, those work ignore the functional region [23] in practice, which may lack some necessary information for understanding human mobility. For example, people's emergency measures are closely related to their located region, and historical visiting.

3 Data, Tasks and Feature Importance Analysis

3.1 Data Descriptions

POI Data. In our study, we gathered the Telepoint Pack DB of POI data in February 2011, which was provided by ZENRIN DataCom Co., Ltd [1]. The original database consists of records containing registered landline telephone numbers along with their associated coordinates (latitude, longitude) and industry category information. For the purpose of our research, we considered each "telepoint" as a specific point of interest (POI). To facilitate our analysis, we categorized the POIs into five broad types (i.e., commuting, public place, shopping, restaurant, and entertainment).

Human Mobility Data. In our study, we gathered a large GPS log dataset anonymously from approximately 1.6 million real mobile phone users in Japan. The data was collected over a 12-day period, specifically from March 1st to March 12th, 2011. The data collection was conducted by two entities: the mobile operator NTT DoCoMo, Inc. and the private workplace ZENRIN DataCom Co., Ltd. The mobile phone users provided their consent for the data collection process. To ensure privacy protection, the collected data were processed collectively and statistically. This procedure ensured that sensitive information such as gender or age was concealed and not accessible for analysis. The positioning function on the users' mobile phones was activated every 5 min by default. However, data acquisition could be affected by factors like signal loss or low battery power. When a mobile phone user remained stationary at a location, the positioning function was automatically turned off to conserve power. It's important to note that the dataset's age distribution slightly favors young users, as they tend to prefer mobile phones with positioning functionality compared to other age groups (e.g., the elderly). The representativeness of our dataset was verified through previous work, where its quality was evaluated [7]. For our specific analysis of human mobility after the Tohoku Earthquake, we focused on the period from March 1st to March 11th, 2011. This period allowed us to capture and study the mobility patterns and behaviors of individuals following the earthquake event.

3.2 Returning Home Strategy-Based Tasks

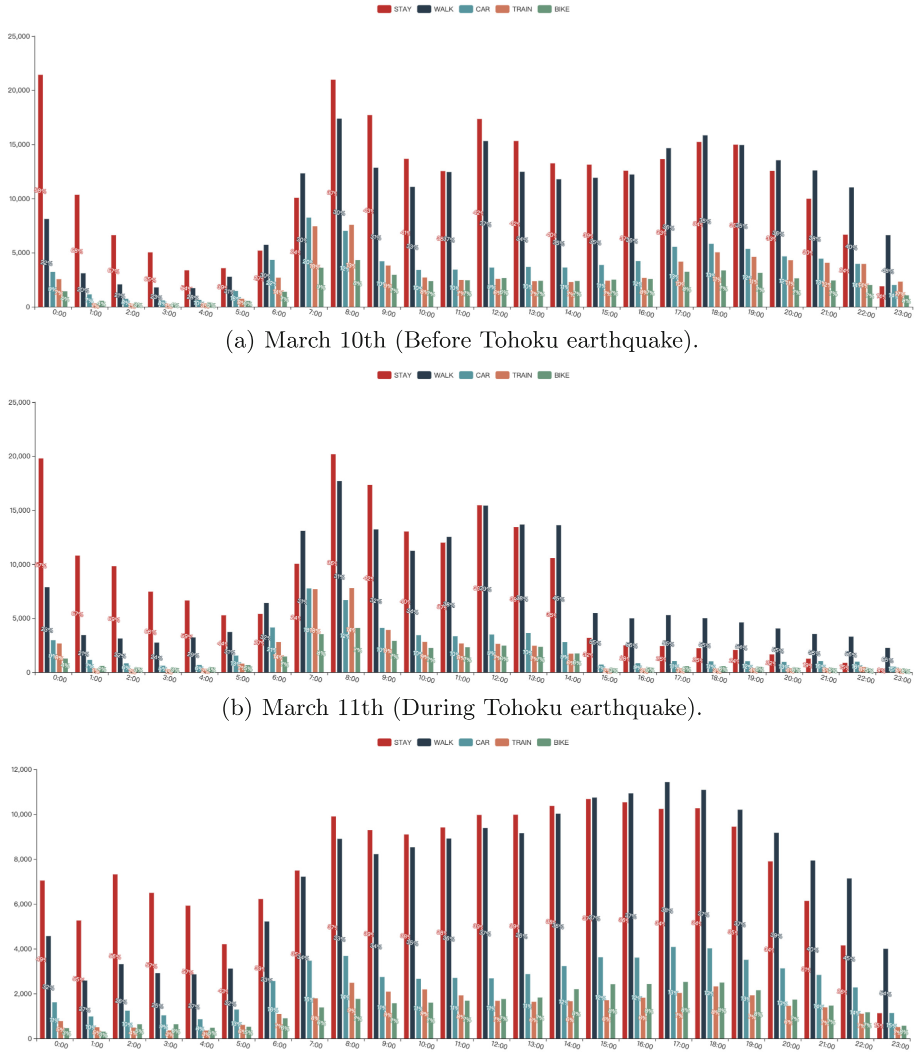
In this study, for those mobile phone users who returned home from their companies after the Tohoku earthquake, we divide proposed decision-making strategies into two main tasks. The first is the travel mode choice prediction, and then is the schedule prediction of departure time from the workplace. The travel mode and departure time from the workplace comprise the returning home strategy.

First, we conduct two sub-tasks regarding the travel mode choice prediction. As is shown as Fig. 1, most subway lines were halted due to the damage to the big earthquake. This caused the train-based travel mode to decline sharply. The primary railway transportation network would recover until 11:00 PM on March 12th. This raises a question: when it comes to the big earthquake if one mobile phone user often came back home from the workplace by railway before, how did him/her make choices for returning home on March 11th? The possible choice for him/her would be Choice 1 - Walk directly; Choice 2 - Wait for the recovery of the train and continue to choose the railway and Choice 3 - Not go home to a hotel/refuge. As a result, sub-task 1 is a three-class classification problem. Except for the decision-making strategy prediction for those who often come back home from their companies by railway, we also conduct a three-class travel mode prediction for all mobile phone users directly, i.e., train-based, walk-based (including Walk and Bike), car-based travel mode.

Secondly, estimating departure times from workplaces is a crucial aspect of comprehending decision-making behaviors during significant disasters. For instance, during the Tohoku earthquake that struck at 14:46 (Japanese standard time), numerous companies allowed their employees to leave work immediately. Nevertheless, with the transportation network severely disrupted, a considerable number of individuals opted to remain at their workplaces, anticipating the restoration of public transportation. As a result, we undertake the task of estimating departure times from workplaces for these mobile phone users.

3.3 Feature Importance Analysis

Although we have already known the feature importance of "HV" from the inference results, we also provide some interesting and reasonable discoveries for decision-making strategies based on feature importance analysis. In Fig. 2 and Fig. 3, we show the all considered factors/features importance values for two inference tasks, respectively. These values were captured by the Shapash library when using all features in LightGBM model. The higher values of the



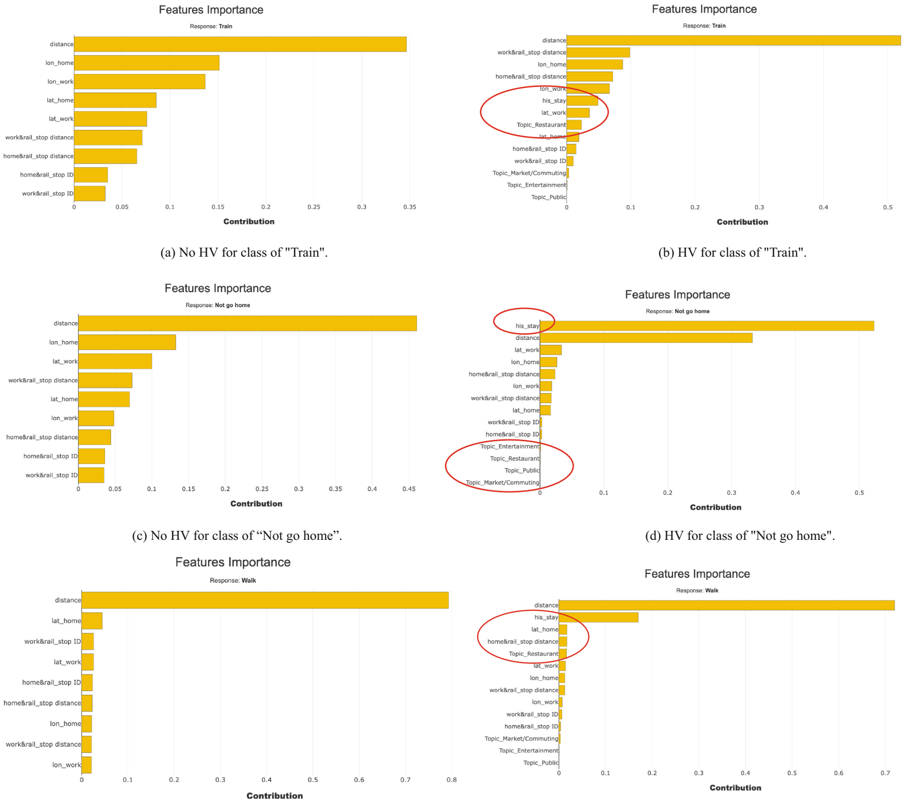
(a) March 10th (Before Tohoku earthquake).

(b) March 11th (During Tohoku earthquake).

(c) March 12th (After Tohoku earthquake)

Fig. 1. Traffic volume and ratio changes of each travel mode choice ("Stay", "Walk", "Bike", "Train" and "Car") before/during/after Tohoku Earthquake (March 10th/11th/12th) in Great Tokyo area. The earthquake happened at 14:46 PM on March 11th, and the horizontal axis is the 24 h of one day.

figures mean the big feature importance. At first, for the travel mode choice inference (sub-task 1), users' locational information (especially "distance") has relatively higher important values than users' historical visit information. But if one user intended to choose to wait for the recovery of the train, the feature "Topic_restaurant" has highly important values. It's reasonable because phone users who often visited the restaurant before the earthquake were more likely



(a) No HV for class of "Train".

(b) HV for class of "Train".

(c) No HV for class of “Not go home”.

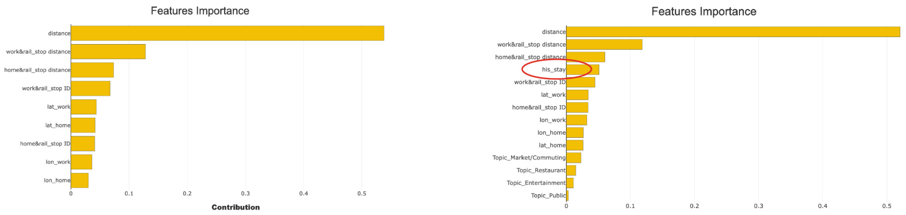
(d) HV for class of "Not go home".

(e) No HV for class of "Walk".

(f) HV for class of "Walk".

Fig. 2. Feature importance analysis for travel mode choice inference. Here, we implement the feature importance analysis of sub-task 1.

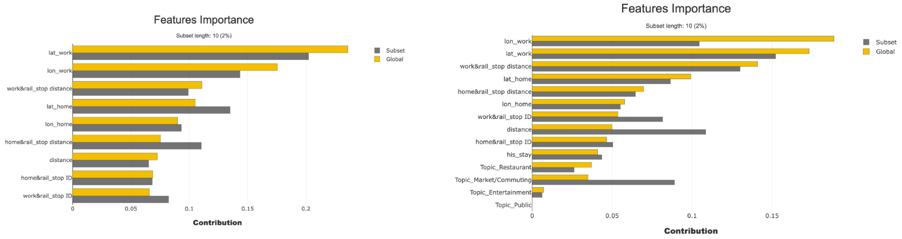
to choose to wait for the recovery of the train in nearby bars/restaurants. On the contrary, if one user chooses to walk directly, the feature “Topic_restaurant” should also be important. Our feature importance results for class of “walk” in Fig. 2 validates this assumption. Especially, the class of “Not go home” only has the highest importance values for feature “his_stay”, which represents the total stay time of historical visits, but this class is not very related with the topics of historical visits (relatively low importance values with “Topic_xxx”). Secondly, for the departure time estimation from workplace, the feature importance results also represent the same pattern with the class of “Not go home” in travel mode inference, which equally shows the high importance values in “his_stay” when considering the “HV” factors. In sum, this suggests that the decisions of mobile phone users were easily influenced by “his_stay”. But the topics of historical visits only show the feasibility in specific travel mode choice. For example, “Topic_Restaurant” only shows strong feature importance in terms of the classes of both “Train” and “Not go home” under travel mode choice inference (Fig. 3).



(a) No HV.

(b) HV.

Fig. 3. Feature importance analysis for departure time estimation from workplace.



(a) No HV.

(b) HV.

Fig. 4. The comparative feature importance analysis for total mobile phone users and those phone users whose work location were nearby Shinjuku station. Here, the analyzed task for feature importance is the departure time estimation from workplace.

4 Conclusion

In this study, we collected large-scale GPS log datasets and analyzed returning home strategy of mobile phone users during Tohoku earthquake. Our work emphasized the importance of understanding the topics of historical visit, to obtain a holistic view regarding the human behaviors of different topics/types. Then all influence factors from users' historical visit, historical travel model choice and locational features could be informative about the decision-making strategies inference during disasters. We define and evaluate two main prediction tasks (i.e., travel mode choice prediction and departure time estimation from company) for returning home strategy during disasters, showing the feasibility of using smartphone sensing to detect historical visits. In addition, we conduct an explainable analysis about the key factors that impel phone users to make decisions. We believe these findings could be useful for ubicomp and the government towards implementing future mobile disaster evacuation systems by effectively understanding human decisions during big disasters.

References

1. https://www.zenrin.co.jp/product/category/gis/contents/telpt/index.html
2. Cárdenas-Benítez, N., Aquino-Santos, R., Magaña-Espinoza, P., Aguilar-Velazco, J., Edwards-Block, A., Medina Cass, A.: Traffic congestion detection system through connected vehicles and big data. Sensors **16**(5), 599 (2016)
3. Chaoxu, X., Gaozhong, N., Xiwei, F., Junxue, Z., Xiaoke, P.: Research on the application of mobile phone location signal data in earthquake emergency work: a case study of jiuzhaigou earthquake. PLoS ONE **14**(4), e0215361 (2019)
4. Feng, J., et al.: Deepmove: predicting human mobility with attentional recurrent networks. In: Proceedings of the 2018 World Wide Web Conference, pp. 1459–1468 (2018)
5. Ghurye, J., Krings, G., Frias-Martinez, V.: A framework to model human behavior at large scale during natural disasters. In: 2016 17th IEEE International Conference on Mobile Data Management (MDM), vol. 1, pp. 18–27. IEEE (2016)
6. Hahm, J., Lee, J.H.: Human errors in evacuation behavior during a traumatic emergency using a virtual fire. Cyberpsychol. Behav. **12**, 98–98 (2009)
7. Horanont, T., Witayangkurn, A., Sekimoto, Y., Shibasaki, R.: Large-scale auto-GPS analysis for discerning behavior change during crisis. IEEE Intell. Syst. **28**(4), 26–34 (2013)
8. Kagan, Y.Y., Jackson, D.D.: Tohoku earthquake: a surprise? Bull. Seismol. Soc. Am. **103**(2B), 1181–1194 (2013)
9. Kuligowski, E.: Predicting human behavior during fires. Fire Technol. **49**(1), 101–120 (2013)
10. Kuligowski, E.D., Kuligowski, E.D.: The process of human behavior in fires. US Department of Commerce, National Institute of Standards and Technology (2009)
11. Lu, X., Bengtsson, L., Holme, P.: Predictability of population displacement after the 2010 Haiti earthquake. Proc. Natl. Acad. Sci. **109**(29), 11576–11581 (2012)
12. Mori, N., Takahashi, T., T.E.T.J.S. Group: Nationwide post event survey and analysis of the 2011 tohoku earthquake tsunami. Coastal Eng. J. **54**(1), 1250001–1 (2012)
13. Mori, N., Takahashi, T., Yasuda, T., Yanagisawa, H.: Survey of 2011 tohoku earthquake tsunami inundation and run-up. Geophys. Res. Lett. **38**(7) (2011)
14. Moussaid, M., Garnier, S., Theraulaz, G., Helbing, D.: Collective information processing and pattern formation in swarms, flocks, and crowds. Top. Cogn. Sci. **1**(3), 469–497 (2009)
15. Pan, Y., et al.: Quantifying human mobility behaviour changes during the covid-19 outbreak in the united states. Sci. Rep. **10**(1), 1–9 (2020)
16. Quddus, M.A., Ochieng, W.Y., Noland, R.B.: Current map-matching algorithms for transport applications: state-of-the art and future research directions. Transport. Res. Part C: Emerg. Technol. **15**(5), 312–328 (2007)
17. Ramadhan, M.I., et al.: An analysis of natural disaster data by using k-means and k-medoids algorithm of data mining techniques. In: 2017 15th International Conference on Quality in Research (QiR): International Symposium on Electrical and Computer Engineering, pp. 221–225. IEEE (2017)
18. Refonaa, J., Lakshmi, M., Vivek, V.: Analysis and prediction of natural disaster using spatial data mining technique. In: 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015], pp. 1–6. IEEE (2015)

19. Song, X., Zhang, Q., Sekimoto, Y., Horanont, T., Ueyama, S., Shibasaki, R.: Modeling and probabilistic reasoning of population evacuation during large-scale disaster. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1231–1239 (2013)
20. Song, X., Zhang, Q., Sekimoto, Y., Shibasaki, R., Yuan, N.J., Xie, X.: A simulator of human emergency mobility following disasters: knowledge transfer from big disaster data. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
21. Wang, J., Kong, X., Xia, F., Sun, L.: Urban human mobility: data-driven modeling and prediction. ACM SIGKDD Explor. Newsl **21**(1), 1–19 (2019)
22. Xiong, C., et al.: Mobile device location data reveal human mobility response to state-level stay-at-home orders during the covid-19 pandemic in the usa. J. R. Soc. Interface **17**(173), 20200344 (2020)
23. Yuan, J., Zheng, Y., Xie, X.: Discovering regions of different functions in a city using human mobility and pois. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 186–194 (2012)
24. Zhang, Z., Fan, Z., Song, X.: Returning home strategy analysis using mobile sensing data in Tohoku earthquake. In: Symposium on AI, Data and Digitalization (SAIDD 2023), p. 86 (2023)

Terminology Saturation Analysis: Refinements and Applications

Victoria Kosa, Oles Dobosevych, and Vadim Ermolayev(✉)

Ukrainian Catholic University, Kozelnytska st. 2a, Lviv 79076, Ukraine
{victoria.kosa,dobosevych,ermolayev}@ucu.edu.ua

Abstract. In this paper, we outline the results of our recent research on terminology saturation analysis (TSA) in subject domain-bounded textual corpora. We present the developed TSA method. We further report about the two use cases that proved the validity, efficiency, and effectiveness of TSA. Based on our experience of TSA use, we analyse the shortcomings of the method and figure out the ways to refinement and improvement. Further, we share our prognoses on how TSA could be used for: (i) generating quality datasets of minimal size for training large language models for performing better in scientific domains; (ii) iteratively constructing domain ontologies and knowledge graphs that representatively describe a subject domain, or topic; or (iii) detecting and predicting events based on the TSA of textual streams data.

Keywords: Terminological Saturation Analysis · Scientific Domain Ontology · Scientific Knowledge Graph · Large Language Model · Deep Learning · Transfer Learning · Event Detection · Event Prediction

1 Introduction

Our recent research has demonstrated [1] that TSA could help reveal terminological patterns in domain-bounded textual data – e.g. topical collections of scholarly publications[1]. We also discovered that it could instrument the discovery of the trends of technology adoption in industry [3]. We found out that the major factors hampering terminological saturation were: (i) the immaturity of the domain implying that the domain-bounded corpus is too small; (ii) the heterogeneity within the domain – e.g. the fragmentation due to the competition among different R&D strands; or (iii) the volatility of the domain terminology over time. Based on these findings, it was remarkable to notice that the existence of a terminologically saturated sub-collection in a corpus of texts – a terminological core sub-collection – indicates the maturity and stability of the respective topic or domain. On the other hand, the absence of terminological saturation points out that an opportunity window is open for the further development of the focal subject domain, including the mergers of competing strands. Application wise, our research was aimed at ensuring the completeness of a text corpus in a domain for ontology learning from texts. However, the results seem to have a broader potential R&D impact.

[1] Please see a comprehensive review of the related work (as for January 2022) in our chapter [2].

R. Akerkar (Ed.): SAIDD 2023, CCIS 1810, pp. 25–41, 2024.
https://doi.org/10.1007/978-3-031-53770-7_3

One promising use is in extracting the smallest possible, yet representatively complete, datasets for training machine learning models for natural language processing tasks. Furthermore, a knowledge graph, built using the terminology extracted from such a terminological core dataset, could be used as a structured representation of the set of features characterising the description of the domain. This might help make the training of the models more efficient and the outputs from the trained deep learning models better explainable, hence – trustworthy.

Another potential use case could be in event detection and prediction using social media text or document streams. We hypothesise that, if terminological saturation is detected in a timed topical stream of texts, it might point out that either (i) the stream is dominated by the authors that use coherent terminology; or (ii) the majority of the community around the topic is focused on something important, that already happened in the past or will happen soon. On the other hand, the lack of terminological saturation in a topical stream might indicate that the situation around the topic is stable in the democratic sense, which is characterised by the plethora of different competitive opinions and judgements on the topic.

The remainder of the paper is structured as follows. In Sect. 2, our approach to TSA [1] is outlined to make the paper self-contained. The review of the recent related work is provided in Sect. 3, with an aim to reveal the current research gaps regarding the potential use of TSA in the context of related open research questions. Section 4 outlines our results in answering some of these questions. Section 5 deliberates on the ways to refine the TSA method for making it more effective and efficient. In Sect. 6, our vision is presented of how the answers to the rest of the open questions, within our focus of interest, could be approached. This vision outlines the plans of our future work. Finally, the conclusions are drawn in Sect. 7.

2 An Outline of TSA

TSA seeks a local-optimal solution for the automated extraction of representative terminologies and respective document sub-collections using an iterative successive approximation approach. As an output, if the process converges to a solution, a terminological core sub-collection (*TCSC*) is extracted from the input collection of documents. *TCSC* carries the saturated set of terms (T_{sat}) that is representative for the subject domain. This set of terms is also extracted for further use. Each term in the set is supplemented with a significance value. The extracted set of terms is sorted in the decreasing term significance order.

Successive approximation starts with an empty set of documents – the dataset D_0. At the *i*-th iteration, several (*inc*) new documents are taken from the input collection as a plain text file, and appended to the dataset $D_{i\text{-}1}$ that has been processed at the *i*-1-st iteration, resulting in the dataset D_i. Hence, the datasets incrementally grow in the iterations. It is supposed that, while growing, the datasets successively become closer to the dataset D_{sat} that represents the *TCSC*. T_{sat} is finally acquired from D_{sat} if saturation is detected at the *i*-th iteration.

To evaluate how close we are to D_{sat}, terminological difference is measured between the sets of significant terms retained from the term candidates extracted from the datasets in successive iterations. At the *i*-th iteration, terminological difference (thd_i) is measured between the sets of retained significant terms T_i and $T_{i\text{-}1}$.

For generating T_i within the *i*-th iteration, the following steps are performed:

1 Extract the set of term candidates, with their significance values, from D_i as pairs $b =< t_i^k, sc_i^k >\in B_i$, where B_i is the set of terms b extracted from D_i, t is a term candidate, sc is its significance score reflecting the number of occurrences of b in D_i; B is ordered by the decrease in ns values

2 Compute the cut-off threshold eps_i to retain significant terms

3 Retain significant terms $t =< t_i^k, ns_i^k >$ from B_i into T_i and measure $thd_i = thd(T_i, T_{i-1})$; ns is a normalised significance score

For detecting terminological saturation, thd_i and eps_i values are observed. The process is stopped when thd_i reliably goes below eps_i (c.f. the chart in Fig. 1) – hence the following two conditions of terminological saturation (c.f. [5]) both hold true:

(i) $thd(T_i T_{i-1}) < eps_i$; and.

(ii) $\forall j > i, thd\left(T_j T_{j-1}\right) < eps_j$.

Significance scores for b in B are computed using the MPCV method [6], which is an optimised and scalable refinement of the C-Value method [7]. The *eps* function sets the individual term significance threshold for B based on the sc in B. The rationale behind *eps* is to regard the sum of the sc in the upper part of B as the simple majority opinion:$eps = sc_i : (\sum_{j=1}^{i} sc_j > 1/2 \sum_{j=1}^{\|B\|} sc_j)$.

The *thd* function measures the distance between the significance vectors of the sets of retained terms ($thd : \{< T_i, T_j >\} \to \Re^+$), which is the Manhattan distance (c.f. [8]) metric in the space of all possible sets of terms:

$$thd\left(T_i, T_j\right) = \sum_{k=1}^{\|int(T_i,T_j)\|} \left| ns_i^k - ns_j^k \right| + \sum_{k=1}^{\|dif(T_i,T_j)\|} ns_i^k + \sum_{k=1}^{\|dif(T_j,T_i)\|} ns_j^k.$$

3 Related Work and Open Problems

As mentioned in [4], TSA is not only a valid method for extracting terminological core sub-collections of texts and respective representative sets of terms for a particular subject domain. The method could be used to explore the patterns and trends that shape the community sentiments around and interpretations of (the semantics of) this domain. In particular, we envisioned that TSA could become an effective instrument for solving several open research problems. These problems relate to exploiting TSA in relevant use cases (Sect. 3.2 to 3.4). However, TSA needs to be further improved to become more effective in the settings where it currently falls short. These settings are discussed in Sect. 3.1. In this Section, we explore and analyse the most recent State-of-the-Art (SotA) research regarding these two strands. Via this analysis, we outline the corresponding open problems.

3.1 Shortcomings of TSA

TSA works well in the cases of established and mature subject domains. These domains are characterised by a well-shaped terminological consensus among the stakeholders presented in a representative body of the mainstream publications. Furthermore, new terms are contributed on time in a small proportion to the stable part of the terminology. As the fraction of these new terms is small, the induced volatility does not affect the saturation trend. Therefore, terminological saturation in such subject domains is quickly reachable and steady, as reported, for example, with regard to our experiments with the Springer Knowledge Management collection of journal articles [1, Ch. 5].

TSA is a hybrid domain-neutral method based on linguistic and statistical processing of data and designed to qualify mainstream terms as more significant. It falls short under several conditions. These conditions and the corresponding reasons for TSA shortcomings are as follows.

Immature, Hence Quickly Evolving, or Niche Domains. If a subject domain is new, immature, and evolves quickly, its terminology cannot be stable. Therefore, its mainstream interpretation by the knowledge stakeholders is blurred and volatile. This factor complicates the selection and collection of a collection of documents that reliably fall into the domain. It is also worth mentioning that immature domains often adopt terms from the other, topically neighbouring domains. This terminological re-use, though natural, is also a complication for a probabilistic topic modelling and snowball sampling approach [9] that is used in TSA for collecting relevant source documents. There might be two possible ways to approach this challenge. One could be to use a human-in-the-loop methodology for refining topic modelling [10]. Another, yet complementary way could be using a deep learning model (c.f. [11]) to gain more accuracy.

Innovative but not yet Frequently Cited Sources. One more deficiency of TSA is that it sacrifices recently introduced innovative terms in favour of the well-established terms in the domain, thus hampering terminological trend capture and analysis. This happens because the term significance measure in TSA is based on the frequency of term occurrence in the analysed document corpus. A balanced account for the citation-based frequency of use and innovativeness of a term could be a better indicator for assessing its significance. A recently published approach to measure the innovativeness of a term is reported in [12]. Their score is based on "how much the predicted publication year is ahead of or behind the actual publication year, which reflects whether the paper covers more topics researched by papers published in the past or more of its topics are covered by future papers" [12].

Quality of Terms Recognition. The C-Value terms recognition method that lies at the basis of the term recognition pipeline of TSA in its current implementation is known to be one of the most effective unsupervised hybrid methods in terms of its accuracy (c.f. [13]). However, its average precision on the mix of datasets (0.53) [13] is not sufficient and needs to be improved. One of the possible approaches is to employ a deep learning (DL) model in a refinement phase for the chunk of documents within the iteration of the TSA method. One of the most recent surveys of the use of DL transformer-based approaches for terms recognition is [14]. It could serve as a starting point for selecting a relevant DL model for domain-bound terms recognition.

Prognosis of Terminological Saturation. One of the shortcomings of TSA, lowering its performance, is the lack of the method for the prognosis of steady saturation after processing several successive approximation iterations (c.f. Sect. 2). A possible approach to remediate this deficiency might be the analysis of the statistical distributions of the terms extracted in these several iterations. A good reference for relevant distributions is provided by [15].

3.2 Representative Datasets of Minimal Size for NLP/I/U

A mainstream approach, currently, in Natural Language, Processing, Interaction, and Understanding (NLP/I/U) is the use of pre-trained Large Language Models (LLMs) (c.f. [16, 17]). LLMs are transformer-based neural models that are built and used following the pre-training approach. In this context, it is worth mentioning that LLMs are the larger successors of the Pre-trained Language Models (PLMs) of the smaller scale. LLMs demonstrate outstanding performance in several important NLP/I/U tasks (c.f. Table 1). Table 1 also shows the scale (number of parameters), pre-training dataset(s) and their size(s) of several prominent PLMs and LLMs, and their reported use in NLP/I/U tasks.

One important shortcoming, that hinders the development and use of LLMs in research, development, and practice, is the amount of resources needed to pre-train an LLM for achieving their SotA performance level. Indeed, a LLM has to be (pre-) trained on a huge language corpus to be as effective as indicated in Table 1. This complication has been recognized as important by the research community and funding bodies, e.g. the European Commission. For example, one of the expected outcomes in the recent

Table 1. Selected LLMs: NLP/I/U tasks and the volume of the language corpora used for (pre-) training. The information in the table is given based on [16]. The LLMs were selected as "landmark" according to [16].

(P/L)LM	Release	Tasks / Applications	No of Parameters	Language Corpus	Language Corpus Size
Pre-trained Language Models					
GPT-1 [18]	2018	NL: inference, text classification, paraphrase detection, question answering	117M	CommonCrawl[2]; BookCorpus [19]	Petabytes; 11,000 books
BERT$_{\text{Large}}$ [20]	2018	9 general NLP/I/U tasks (benchmarks)	340M	BookCorpus; English Wikipedia	800M words; 2,500M words

(*continued*)

[2] Https://commoncrawl.org/

Table 1. (*continued*)

(P/L)LM	Release	Tasks / Applications	No of Parameters	Language Corpus	Language Corpus Size
T5 [21]	2019	NL: sentence acceptability judgement, sentiment analysis, paraphrase detection, inference, co-reference resolution, sentence completion, word sense disambiguation, question answering	11B	C4[3]	1T tokens
Large Language Models					
GPT-3 [22]	2020	In addition to GPT-1: generate coherent text, write computer code, create art, produce natural-sounding text, ChatGPT	175B	Books1, Books2 [22], Common Crawl, Wikipedia	much bigger than BookCorpus
Codex [23]	2021	Based on GPT-3, fine-tuned on GitHub code, program synthesis	12B	GitHub code	100B tokens
LlaMA [24] open-source	2023	Commonsense reasoning, question answering, reading comprehension, mathematical reasoning, code generation	65B	CommonCrawl, C4, Github code, Wikipedia, Gutenberg[4] and Books3 [25], ArXiv	1.4T tokens

(*continued*)

[3] https://www.tensorflow.org/datasets/catalog/c4.

[4] Https://www.gutenberg.org/

Table 1. (*continued*)

(P/L)LM	Release	Tasks / Applications	No of Parameters	Language Corpus	Language Corpus Size
GPT-4 [26]	2023	In addition to GPT-3, adds multi-modality to prompts	N/A	N/A	N/A

Horizon 2020 Call for Proposals under the Advanced Language Technologies theme was developing "Language models, capable of learning from smaller language corpora"[5].

We also found out that LLMs have not been frequently used, so far, for topic modelling in scientific domains and domain-bounded terminology recognition in textual documents. Rare examples are [27] for scientific topic modelling and [28] for domain-bounded terminology recognition. A possible reason for this scarcity might be the need to fine-tune a basic pre-trained LLM for achieving acceptable quality. A necessary resource for fine-tuning is the representative corpus of documents that covers the domain of focus sufficiently completely.

In subject domain-bounded settings, TSA could serve as an effective instrument for generating the representative document collections of minimal size that could be further labelled and used for LLM training within these subject domains. We outline our results in using TSA for extracting these *TCSC*s in Sect. 4.1.

3.3 Building Scientific Domain Ontologies and Knowledge Graphs

A scientific knowledge graph (SKG) is an emerging representation for scholarly knowledge in a particular scientific field or domain. SKG complements the traditional way of representing and disseminating this knowledge in the form of scholarly papers by enabling machine processing of this knowledge for semantic search, querying, analysis involving reasoning, and visualisation.

As pointed out, e.g. in [29], a scientific "domain is characterised by its specific terminology and phrasing which is hard to grasp for a non-expert reader". Hence, extracting conceptual knowledge and building an SKG needs to be based on the human domain-specific professional expertise and specifically tailored extraction procedures. It also implies that an SKG for a scholarly subject domain needs to be built using a well-formed domain ontology that representatively covers the concepts (terminology) in the domain. The recent related work offers the SotA approaches for narrowing this gap.

A topical example of building and ontology for a broad scholarly domain is the extraction of the Computer Science Ontology (CSO) [30] using Klink-2 [31], and later building an SKG for Artificial Intelligence [32] using the CSO Classifier in the pipeline. This approach demonstrated impressive performance in terms of scalability – the dataset for building CSO covered 16M papers. However, it used only metadata in the combination

[5] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-DEtails/horizon-cl4-2023-human-01-03.

with external resources, like DBpedia[6] entries. This allowed only for shallow document analysis with regard to the conceptual/terminological coverage of the domain.

A domain-neutral approach for the extraction of scientific concepts from scholarly papers is proposed in [29]. This approach is based on the analysis of the paper abstracts using a DL pipeline. The pipeline is trained on a document corpus, covering 10 scientific domains, manually annotated by labelling generic scientific concepts. The experiments showed satisfactory transferability of these generic concepts across domains. The solution was reported to achieve "a fairly high F1 score".

The PLUMBER framework [33] is an integrative approach aiming at bringing "together the research community's disjoint efforts on KG completion". PLUMBER collects and dynamically integrates "40 reusable components for various KG completion subtasks" into the pipelines that fit best for SKG completion in a particular domain.

Our approach for building scientific domain ontologies from texts is OntoElect [34]. It uses TSA as the 1st phase, which ensures the sufficient coverage of the domain by extracted terminology that is further conceptualised into ontological fragments. The solution is sufficiently scalable [1, Ch. 5] to allow processing industrial-size full-text paper collections. Being a modular pipeline (c.f. PLUMBER), it allows including reusable components at different stages. However, such inclusions/replacements need to be done manually, which is a shortcoming against PLUMBER, that is automatically configurable.

3.4 Trend Analysis in Non-Stationary Time Series of Publications

Trend discovery and analysis are widely used techniques for the analysis of the history and prognosis of the future development in many fields and disciplines. The mainstream of these methods are based on processing time series data and discovering statistical patterns and turning points in it. Among the vast available body of research papers, a good recent overview of these methods, with the pointers to relevant applications, is [35].

In our work, we are focused on using TSA for analysing the trends and time lags in technology transfer to and adoption in industries [1, Ch. 6]. In this setting, the datasets containing relevant scientific paper sub-collections are regarded as non-stationary time series [36] with respect to the terms carried by the papers published at time points. It is also worth noting that the use of neural networks and machine learning has been gaining increasing popularity in non-stationary time series analysis for quite some time already [37]. Hence, it might be reasonable to combine these approaches with DL-based refinement of TSA.

3.5 Event Detection and Prediction

One of the recent reviews of the SotA approaches to event detection and prediction in online social networks, based on text analytics, is [38]; the paper also presents the timeline and taxonomy of existing methods. Outlines the major open issues in the field. An advanced approach to detect correlated events from individual documents using a graph model of event relationships is proposed in [39]. It also contains the review of

[6] Https://wiki.dbpedia.org/.

the SotA in document-level event detection. One of the recent contributions of a DL-based approach to fine-grained event detection from texts is [40]. This work proposes the BMRMC model. These SotA techniques and models could be put on the top of TSA to enable its use for event detection and prediction based on textual information streams.

4 Tested Uses of TSA

In this Section, we present two important application use cases in which TSA has been experimentally proven [1] being instrumental and effective. The first case covers the situation when a mature and stable body of scientific texts exist in the corresponding subject domain and describes this domain sufficiently fully. Therefore, steady terminology saturation could be expected in the document collection. The second case is for an immature, hence quickly evolving domain of scientific knowledge. In this case, a representative collection of scholarly publications cannot be easily collected or even does not exist. Hence, the saturation of terminology is not achievable.

4.1 Extracting a *TCSC* for a Mature Scholarly Domain

TSA has been used in the experiments with several well-established real-world document collections coming from different scientific subject domains with different breadths in the topic coverage [1, Ch. 5]. Please refer to Table 2 for the overview of these collections.

The results of TSA on these collections are summarised in Table 3. As a result, the extracted *TCSC*s in all the domains contain significantly less documents that have statistically the very similar terminological coverage of the domain. Hence, if these core sub-collections are used as the datasets for training DL models for NLP/I/U tasks, the effort to label these documents manually will be significantly lower. Remarkably, the quality of training could be expected to be the same as if the corresponding entire collections are used. Furthermore, as terminologically redundant documents are eliminated, it might be expected that the trained models will not be over-fitted toward a subset of terms.

The most substantial decrease in the numbers of documents and retained significant terms has been demonstrated with regard to the KM collection. This result is indeed encouraging as it proves that using TSA to extract ontologies and knowledge graphs from full-text domain-bounded document collections of industrial volumes is a feasible and scalable approach.

4.2 Analysing Trends in an Immature and Evolving Domain

In our industrial use case[7], TSA has been used [1, Ch. 6] to verify the prognosis, by Gartner[8] [41], of the Generative Adversarial Network technology adoption by the IT industry. That was the case to demonstrate the utility of observing the absence of terminological saturation in the collection of publications for an immature and rapidly

[7] SAGOIT-IT: Strategic Analysis of R&D Gaps and Opportunities for Industrial Uptake in Trending IT Fields –the project funded by Group BWT, llc.

[8] https://www.gartner.com/en

Table 2. The summary of the characteristics of real-world document collections and datasets

Collection	Domain	Type and Layout	No of Documents	Noise	Incre-ment (*inc*)	No Datasets (Iterations)
DMKD	Knowledge Management (DMKD journal)	journal, Springer 1-column	300	not cleaned, moderately noisy, regular noise not accumulated	20 papers	15
TIME	Time Representation and Reasoning	conference, IEEE 2-column	437	manually processed, moderately noisy, regular noise accumulated	20 papers	22
DAC	Engineering Design Automation	conference, IEEE 2-column	506	not cleaned, substantially noisy, regular noise accumulated	20 papers	26
KM	Knowledge Management (15 Springer journals)	journal, Springer 1-column	7 500	not cleaned, moderately noisy, regular noise not accumulated	100 papers	75

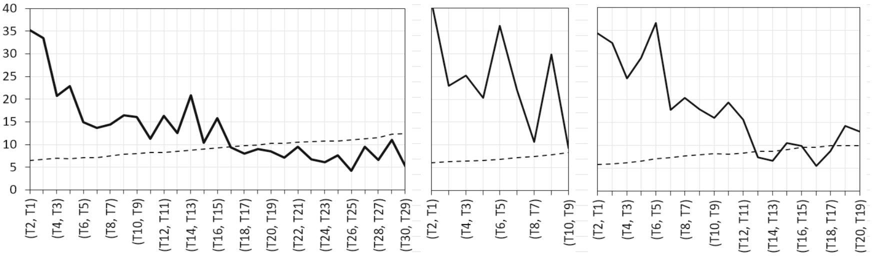
evolving domain. In this study, we also tried the iterative refinement of the process of collecting relevant publications to the collection. This refinement was done via the use of several most terminologically significant papers, among those collected at the previous iteration, as a refined seed for topic modelling and citation network analysis at the subsequent iteration.

Our research questions were: (1) Are the research contributions and respective professional community around the technology mature? and (2) Are there any gaps in the knowledge about the technology between the academic researchers and industrial adopters? To answer these questions, the terminological footprints, over the domain description, of different sub-collections of publications were examined: (i) authored by academics; (ii) authored by people from industry; and (iii) authored collaboratively by academics and industrialists. Figure 1 shows that only the academic part of the GAN community is mature, as it possesses a terminologically saturated body of publications. Hence, the GAN technology is ready to be transferred to industry.

Table 3. Compactness of *TCSC*s and saturated bags of retained significant terms (adapted from [1])

Collection	Number (%) of Documents		Number (%) of Retained Significant Terms		
	Entire Collection	TCSC	Entire Collection	TCSC	With Terms Grouping
DMKD	300	80 (26.7%)	313 506	3 399 (1.08%)	2 135 (0.68%)
TIME	437	180 (41.2%)	315 474	5 493 (1.74%)	3 456 (1.10%)
DAC Cleaned	506	220 (43.5%)	518 765	16 703 (3.22%)	7 496 (1.44%)
KM	7 500	500 (6.6%)	4 035 760	13 579 (0.34%)	---

On the other hand, Fig. 2 highlights substantial terminological differences between the sub-collections. Therefore, further collaboration between the academic and industrial parts of the community might contribute to the increase of the maturity in the industrial body of knowledge and practice.



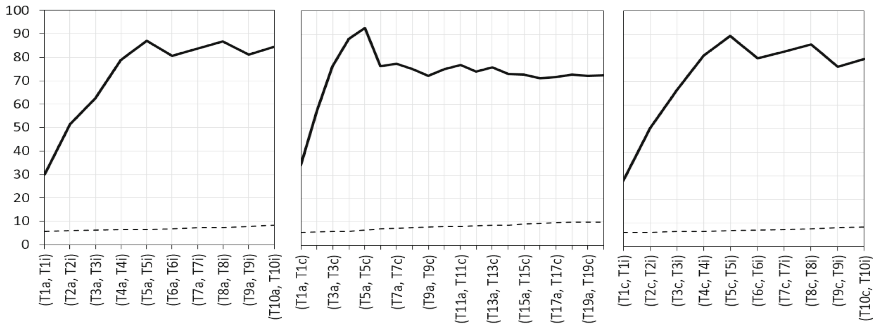
(i) Academic: saturated

(ii) Industrial: not saturated

(iii) Collaborative: not saturated

Fig. 1. Terminological saturation measurements for the Academic, Industrial, and Collaborative parts of the GAN collection. Adopted from [1]. Legend: - - - - - *eps*; **- - - - - - - - - - -** *thd*; vertical scale indicates the value of *eps* or *thd*; horizontal scale indicates the pair of compared datasets.

The combination of factors that has been discovered points out the opportunity window for the successful transfer of the GAN technology to industry, which might increase the competitive advantage of the participating companies.



(a) Academic – Industrial (b) Academic – Collaborative (c) Industrial - Collaborative

Fig. 2. Terminological difference curves for the pairs of collection parts. Adopted from [1]. Legend: - - - - - *eps*; **- - - - - - - - - - -** *thd*; vertical scale indicates the value of *eps* or *thd*; horizontal scale indicates the pair of compared datasets.

5 Toward Improving TSA

We plan to improve the quality of our baseline (probabilistic topic modelling and snowball sampling) pipeline for collecting relevant scholarly texts within a subject domain by exploiting a domain-neutral DL-based topic modelling approach (c.f. Sect. 3.3), especially in the niche and emerging domains. This will be done following an iterative bootstrapping method with a human in the loop. The baseline of this method has been initially tested in our GAN use case (c.f. Sect. 4.2). In the initial iteration the following steps of the documents collection workflow will be performed:

1. Collect the draft set of documents and extract the *TCSC* and T_{sat} using the TSA pipeline [1, Ch. 4]
2 Form the validated $TCSC_{val}$ by manually examining the *TCSC* using a domain expert and filtering out the irrelevant documents
3 Generate the validated set of relevant significant terms T_{val}^{sat} by applying the TSA term extraction pipeline to $TCSC_{val}$
4 Automatically label the documents of $TCSC_{val}$ using the terms from T_{val}^{sat}
5 Use $TCSC_{val}$ for training the DL model for topic modelling and relevant documents discovery (DL-TMDD)

In every subsequent iteration, the TSA pipeline in the 1-st step will be replaced by the use of DL-TMDD. The outlined workflow could be stopped at iteration *i* if $thd\left(T_{val}^{sat}[i], T_{val}^{sat}[i-1]\right) < eps_i$.

Another important issue is improving the accuracy of term recognition in the linguistic part of the TSA pipeline. Currently TSA uses NLTK API [42] to implement its linguistic part and extract the bag of candidate terms. To improve it, a similar iterative bootstrapping approach could be used for training a DL model to perform a classification task on the input text dataset for discovering term candidates (DL-TD). The initial iteration could be the use of the domain-neutral approach of [29] that builds upon the training of a DL model, on a manually labelled cross-domain corpus of scholarly articles, to discover generic scientific terms. The subsequent iterations will follow the incremental successive approximation pattern of TSA, using, however, the DL-TD as the substitute for NLTK and C-Value method in the pipeline, as follows for the *i*-th iteration.

1. Train DL-TD on the labelled TSA dataset $D_{i\text{-}1}$ of the (*i*-1)-th iteration
2. Add the increment of *inc* documents to $D_{i\text{-}1}$ to form D_i
3. Use DL-TD to discover terms in D_i
4. Validate the discovered additional terms by the inspection of a human domain expert
5. Automatically label D_i using the validated set of retained significant terms

This workflow should continue until either terminological saturation is detected or all the documents in the collection are processed.

As valuable side-effects, the labelled dataset ($TCSC_{val}$ and T_{val}^{sat}) for training 3-d party models (e.g. for PLUMBER) in the domain will be developed. This dataset could also be used for the iterative development of the domain ontology using the OntoElect methodology [34].

Yet further improvement of TSA performance could be sought via exploring the ways to predict terminology saturation, after initial iterations. This could be done by examining the collected knowledge about the statistical distributions of the terms in the collection under examination, in particular, to recognize, as significant, innovative [12] or emerging terms in addition to mainstream significant terms. The starting point for this work could be exploiting the distributions surveyed in [15].

Finally, the iterative bootstrapping approach, outlined above, could be made transferrable, as the subset of general scientific terms is domain-neutral. This subset of general terms could also be extended in the process of building the saturated sets of terms for different subdomains within a broader domain, other neighbouring or overlapping domains.

6 Potential Applications of TSA

Based on the already tested uses of the baseline TSA method (Sect. 4) and its planned refinements (Sect. 5), the following applications of TSA are envisioned as plausible.

Generating Datasets for the Focused Training of LLMs in Scientific Subject Domains. In Sect. 3.2, the shortcoming of LLMs has been pointed out with regard to the very substantial volume of the resources required for achieving the necessary accuracy in performing several signature tasks for these models (c.f. Table 1). Furthermore, as rightfully mentioned in [29], terminology and phrasing is specific to a scientific subject domain, which complicates the use of LLMs within such a domain. These complications raise the need for a tailored LLMs training for scientific domains in order to achieve acceptable performance. Hence, an efficient and effective method yielding quality and representative training datasets is of high demand, especially, if the method ensures that such a dataset has the minimal possible volume. The iterative bootstrapping approach, proposed in Sect. 5, might result in such a method.

Building Domain Ontologies to Support SKG Completion. A more advanced potential application of the refined TSA could be the use of the discovered domain-bounded saturated sets of retained significant terms as the input for constructing the ontologies describing the respective subject domains. As proposed in [34], TSA is the method underpinning the initial phase of our OntoElect methodology for ontology refinement. In the ontology refinement cycle, the method supplies the Conceptualization phase of OntoElect [43] with the T_{val}^{sat}, that are used as the building blocks for developing ontological fragments around the significant terms that represent concepts. The terms that represent properties are used to enrich these concepts semantically and connect ontological fragments in a semantic network. Using an evolving domain ontology-in-the-loop approach, which could be supported by OntoElect, might enable better performance of the frameworks like [31, 33] for semi-automatically completing SKG. This is especially relevant for immature and evolving domains, or for the domains of broad public interest covered by social media text streams.

Detecting and Predicting Events. Another promising idea for applying TSA arises from our observations of terminology volatility over time in real document collections [1, Ch. 5]. It was interesting to observe that, in scientific domains that we used for evaluating TSA (Sect. 4.1 and 4.2) there was a certain correlation between the appearance of an influential paper, that has further been well cited, and the rise of the terminological volatility. The volatility has been increased due to the appearance of a bunch of the new papers contributing new terms. If this observation is generalised to what happens, e.g., in social media texts within respective communities and topics, the following use case scenario could be thought of. Those who followed twitter on the agricultural exports from Ukraine in the spring of 2023, definitely noticed the rise of negative sentiment coming from Central-Eastern Europe, related to the prices of Ukrainian grain. This sentiment could have been detected if the volatility of used wordings had been analysed using TSA. This volatility could have been interpreted as an indicator of a potential event of grain exports ban in several involved countries. The event indeed happened after a short time lag. On the contrary, a noticeable decrease in terminology volatility over time, compared to the usual distributions, could indicate a coordinated campaign aimed at artificially forming desired sentiments on a topic. This is a potential indicator of propaganda, which might be useful for selecting messages for their origin verification and fact checking.

7 Conclusive Remarks

In this visionary paper, we presented our views and prognoses on how TSA could be refined and further exploited, for public good, in several important application fields. These include, but are not limited to, tailored training of LLMs in scientific subject domains, instrumenting the completion of SKG, trend analysis in immature and evolving domains, detecting and predicting events based on social media streams analysis. For enabling this spectrum of applications, we proposed to design an iterative bootstrapping approach within the refined TSA method, based on the development and use of domain-neutral DL models for topic modelling and terms recognition. This vision of the bootstrapping approach involves human-and-ontology-in-the-loop, as presented in Sect. 5. The envisioned applications of the improved TSA also constitute our plans for the future work. We also look forward to trying the refined TSA method as an instrument in our OntoElect ontology refinement methodology and the SKG development frameworks like Klink-2 or PLUMBER.

References

1. Kosa, V., Ermolayev, V.: Terminology saturation: detection, measurement, and use. Cognitive Science and Technology, Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-8630-6
2. Kosa, V., Ermolayev, V.: Related work and our approach. In: Terminology Saturation: Detection, Measurement, and Use. Cognitive Science and Technology, Springer, Singapore, pp. 7−39 (2022) https://doi.org/10.1007/978-981-16-8630-6
3. Kosa, V., Ermolayev, V.: Saturated terminology extraction and analysis in use. In: Terminology Saturation: Detection, Measurement, and Use. Cognitive Science and Technology. Springer, Singapore, pp. 155−170 (2022)
4. Ermolayev, V., Kosa, V.: Terminology saturation analysis for machine learning and event detection. In: Akkerkar, R. (ed.): Symposium on AI, Data and Digitalization (SAIDD 2023), Sogndal, Norway, 09–10 May 2023, Western Norway Research Institute (2023)
5. Tatarintseva, O., Ermolayev, V., Keller, B., Matzke, W.-E.: Quantifying ontology fitness in OntoElect using saturation- and vote-based metrics. Revised Selected Papers of ICTERI 2013. CCIS, vol. 412, pp. 136–162. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-319-03998-5_8
6. Kosa, V., Chaves-Fraga, D., Dobrovolskiy, H., Ermolayev, V.: Optimized term extraction method based on computing merged partial C-values. Revised selected papers of ICTERI 2019. CCIS, vol. 1175, pp. 24–49. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39459-2_2
7. Frantzi, K.T., Ananiadou, S.: The C-Value/NC-Value domain independent method for multiword term extraction. J. Natural Lang. Process. **6**(3), 145–179 (1999). https://doi.org/10.5715/jnlp.6.3_145
8. Gomaa, W.H., Fahmy, A.A.: A survey of text similarity approaches. Int. J. Comp. Appl. **68**(13), 13–18 (2013)
9. Dobrovolskyi, H., Keberle, N.: Collecting seminal scientific abstracts with topic modelling, snowball sampling and citation analysis. In: ICTERI 2018. Volume I: Main conference, Kyiv, Ukraine, 2018, CEUR-WS, vol. 2105, pp. 179–192 (2018)
10. Fang, Z., Alqazlan, L., Liu, D., et al.: A User-Centered, Interactive. Human-in-the-Loop Topic Modelling System. arXiv e-prints (2023). https://doi.org/10.48550/arXiv.2304.01774
11. Zhang, H., Chen, B., Cong, Y., Guo, D., Liu, H., Zhou, M.: Deep autoencoding topic model with scalable hybrid Bayesian inference. IEEE Trans on Patt. Anal. Mach. Intell. **43**(12), 4306–4322 (2021). https://doi.org/10.1109/TPAMI.2020.3003660
12. Savov, P., Jatowt, A., Nielek, R.: Identifying breakthrough scientific papers. Inf. Process. Manage. **57**(2), 102–168 (2020). https://doi.org/10.1016/j.ipm.2019.102168
13. Zhang, Z., Iria, J., Brewster, C., Ciravegna, F.: A comparative evaluation of term recognition algorithms. In: 6th International Conference on Language Resources and Evaluation, pp. 2108–2113 (2008)
14. Hanh, T.T., Martinc, M., Caporusso, J., Doucet, A., Pollak, S.: The Recent Advances in Automatic Term Extraction: A survey. arXiv:2301.06767 (2023)
15. Misuraca, M., Spano, M.: Unsupervised analytic strategies to explore large document collections. In: JADT 2018. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52680-1_2

16. Zhao, W.X., Zhou, K., Li, J., et al.: A Survey of Large Language Models. arXiv:2303.18223 (2023)
17. Fan, L., Li, L., Ma, Z., et al.: A Bibliometric Review of Large Language Models Research from 2017 to 2023. arXiv:2304.02020 (2023)
18. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. (2018)
19. Zhu, Y., Kiros, R., Zemel, R. et al.: Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of the ICCV 2015, Santiago, Chile, December 7–13, 2015, pp. 19–27. IEEE (2015)
20. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 (2019)
21. Raffel, C., Shazeer, N., Roberts, A., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 1–67 (2020)
22. Brown, T.B., Mann, B., Ryder, N., et al: Language Models are Few-Shot Learners. arXiv:2005.14165 (2020)
23. Chen, M., Tworek, J., Jun, H., et al.: Evaluating Large Language Models Trained on Code. arXiv:2107.03374 (2021)
24. Touvron, H., Lavril, T., Izacard, G., et al.: LLaMA: Open and Efficient Foundation Language Models. ArXiv, abs/2302.13971 (2023)
25. Gao, L., Biderman, S.R., Black, S., et al.: The Pile: An 800GB Dataset of Diverse Text for Language Modeling. arXiv:2101.00027 (2020)
26. OpenAI: Gpt-4 technical report. OpenAI (2023)
27. Grootendorst, M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 (2022)
28. Zheng, C., Deng, N., Cui, R., Lin, H.: Terminology extraction of new energy vehicle patent texts based on BERT-BILSTM-CRF. In: Barolli, L. (ed) Advances in Internet, Data & Web Technologies. EIDWT 2023. LNDECT, vol. 161. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-26281-4_19
29. Brack, A., D'Souza, J., Hoppe, A., Auer, S., Ewerth, R.: Domain-independent extraction of scientific concepts from research articles. In: Jose, J.M. et al. (eds.) Advances in Information Retrieval. ECIR 2020. LNCS, vol. 12035. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45439-5_17
30. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E.: The computer science ontology: a large-scale taxonomy of research areas. In: Vrandečić, D. et al. (eds) ISWC 2018: The Semantic Web, LNCS, vol. 11137. Springer, Cham, pp. 187–205 (2018). https://doi.org/10.1007/978-3-030-00668-6_12
31. Osborne, F., Motta, E.: Klink-2: Integrating multiple Web sources to generate semantic topic networks. In: Arenas, M, (ed.) The Semantic Web - ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I, pp. 408–424. Springer Int. Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-25007-6_24
32. Dessì, D., Osborne, F., Recupero, D.R., Buscaldi, D., Motta, E., Sack, H.: AI-KG: An Automatically Generated Knowledge Graph of Artificial Intelligence. In: Pan, J.Z., Tamma, V., d'Amato, C., Janowicz, K., Bo, Fu., Polleres, A., Seneviratne, O., Kagal, L. (eds.) The Semantic Web – ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II, pp. 127–143. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-62466-8_9
33. Jaradeh, M.Y., Singh, K., Stocker, M., Roth, A., Auer, S.: Information extraction pipelines for knowledge graphs. Knowl. Inf. Syst. **65**, 1989–2016 (2023). https://doi.org/10.1007/s10115-022-01826-x

34. Ermolayev, V.: OntoElecting requirements for domain ontologies. The case of time domain. EMISA Int. J Concept. Model. 13(Sp.I.): 86–109 (2018) https://doi.org/10.18417/emisa.si.hcm.9
35. Ghaderpour, E., Pagiatakis, S.D., Hassan, Q.K.: A survey on change detection and time series analysis with applications. Appl. Sci. **11**(13), 6141 (2021). https://doi.org/10.3390/app11136141
36. Rhif, M., Abbes, A.B., Farah, I., Martínez, B., Sang, Y.: Wavelet transform application for/in non-stationary time-series analysis: a review. Appl. Sci. **9**(7), 1345 (2019). https://doi.org/10.3390/app9071345
37. Kim, T.Y., Oh, K.J., Kim, C., Do, J.D.: Artificial neural networks for non-stationary time series. Neurocomputing **61**, 439–447 (2004). https://doi.org/10.1016/j.neucom.2004.04.002
38. Xiangyu, Hu., Ma, W., Chen, C., Wen, S., Zhang, J., Xiang, Y., Fei, G.: Event detection in online social network: Methodologies, state-of-art, and evolution. Comput. Sci. Rev. **46**, 100500 (2022). https://doi.org/10.1016/j.cosrev.2022.100500
39. Zhou, Ji., Shuang, K., An, Z., Guo, J., Loo, J.: Improving document-level event detection with event relation graph. Inf. Sci. **645**, 119355 (2023). https://doi.org/10.1016/j.ins.2023.119355
40. He, X., Yan, G., Si, C., et al.: General fine-grained event detection based on fusion of multi-information representation and attention mechanism. Int. J. Mach. Learn. & Cyber. (2023). https://doi.org/10.1007/s13042-023-01900-y
41. Trends appear on the Gartner hype cycle for emerging technologies 2019. Gartner Inc. https://www.gartner.com/smarterwithgartner/5-trends-appear-on-the-gartner-hype-cycle-for-emerging-technologies-2019/. Accessed 14 Oct 2021
42. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc. (2009)
43. Moiseyenko, S., Vasileyko, A., Ermolayev, V.: Building a feature taxonomy of the terms extracted from a text collection. In: Proceedings MS-AMLV 2019, CEUR-WS vol. 2566, 59–70 (2020)

How to Be a Well-Prepared Organizer: Studying the Causal Effects of City Events on Human Mobility

Jiyuan Chen[1], Hongjun Wang[1], Zipei Fan[2], and Xuan Song[1](✉)

[1] Southern University of Science and Technology, Shenzhen, China
songgx@sustech.edu.cn
[2] The University of Tokyo, Tokyo, Japan

Abstract. The analysis of how city events causally affect human mobility is of critical importance. The city government will be thrilled to know how an impending event will influence mobility beforehand, so that they can either decide specifically when and where the event will be held (or not), or be more prepared for some possible circumstances such as crowd collapses and crushes. Previous research on human mobility mainly focuses on simple future prediction based on data correlation, yet the study on the underlying causal effect is woefully inadequate. Motivated by the recent tragedy, the Itaewon Halloween disaster, in this paper we try to explore the causal effects of city events on human mobility using counterfactual prediction. The main technical challenge here lies in capturing and debiasing the time-varying unobservable confounders (e.g., people's willingness to go outdoors) that affect both the event organization and the number of event participants. Fortunately, the increasing sources of time-varying data offer the possibility to refactor such confounding effects from observation. To this end, we utilize multiple sources of observation data in New York City to construct a neural network-based causal framework, which automatically learns and balances the time-varying unobservable confounders representations and provides estimations for the ITE problem.

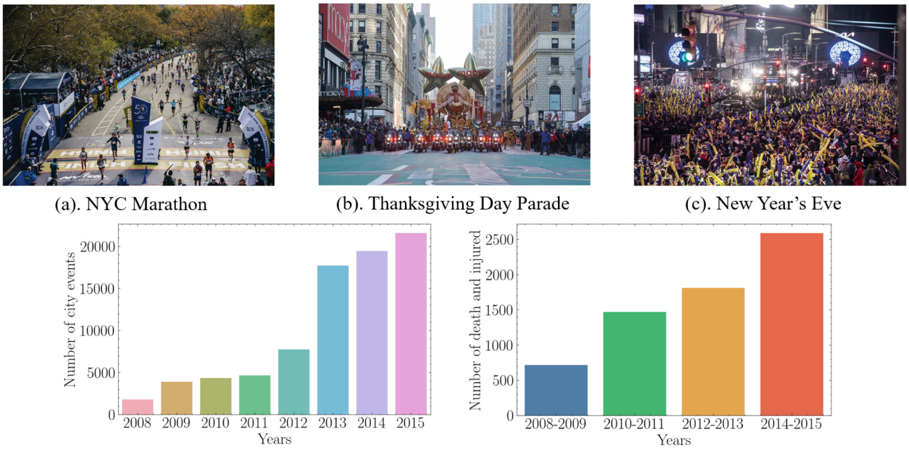
Keywords: Human Mobility · Urban Computing · Causal Inference · Smart City

1 Introduction

Seoul, a city popular for its diversity and nightlife, should have been filled with happiness and laughter for the Halloween celebration on October 29, 2022. However, it instead witnessed a terrible stampede accident at Itaewon, taking the lives of at least 156 people and causing more than 152 injured[1]. Just a day

[1] https://www.theguardian.com/world/2022/oct/30/itaewon-crowd-crush-felt-like-an-accident-was-bound-to-happen.

R. Akerkar (Ed.): SAIDD 2023, CCIS 1810, pp. 42–64, 2024.
https://doi.org/10.1007/978-3-031-53770-7_4



(a). NYC Marathon

(b). Thanksgiving Day Parade

(c). New Year's Eve

(d). Number of city events in NYC

(e). Number of victims of stampedes

Fig. 1. Part (a), (b), and (c) are some of the large gathering events in big cities (https://www.nyc.gov/site/cecm/gallery/photo-gallery.page). Part (d) shows the number of city events in New York City, with an obvious growing trend from 2008 to 2015 (https://data.cityofnewyork.us/City-Government/NYC-Permitted-Event-Information/tvpp-9vvx). Part (e) reports a number of people who fail victims to crowd stampedes over the years (https://en.wikipedia.org/wiki/List_of_human_stampedes_and_crushes). As we can observe from the statistical data, the increase in large gathering events in cities comes with an unwanted side-effect (i.e., growth in the number of stampedes victims).

later, another crowd crush accident took place in Kinshasa, Democratic Republic of the Congo, where 11 people were killed when attending a music concert[2]. While mourning for the dead, we should also be alerted by these consecutive tragedies. Over the years, the rapid development of society and economy has brought a significant increase to the number of large gathering activities, which can potentially lead to emergencies, especially the risk of the crowd stampedes [1–3], as shown in Fig. 1. Therefore, a question can be naturally thrown to technical researchers: *Can we estimate the influence of an incoming city event on human mobility given a certain context?* The study of this problem serves as an essential part for smart city development, and if solved properly, it could provide event organizers (e.g., government) with additional guidance [4,5], making them more prepared for what is about to happen. For example, a corresponding number of police officers can be sent to maintain order according to the estimated influence of the event. An appropriate time and location of the event can also be carefully selected to prevent excessive impact on normal mobility of people (e.g., city maintenance events during rush hours at busy cross-roads usually have great impacts on travelers).

[2] https://www.abc.net.au/news/2022-10-31/overcrowded-stadium-kills-11-people-in-congolese-capital/101596164.

There is already a myriad of work trying to incorporate city events into human mobility studies [6–10]. However, these researches only implicitly capture the data correlation between the events and human mobility, and can not generate the counterfactual outcome which deviates from the factual distribution. In this paper, we propose to estimate the causal effect (e.g., Individual Treatment Effect [11,12]) of city events on mobility with the help of counterfactual prediction [13,14], and to the best of our knowledge, no one has made such attempt in this direction. Traditionally, the ideal solution to obtain causal effect is by conducting Randomized Controlled Trials (RCTs) [15], which randomly divide samples into treated group and control group based on whether or not the treatments are given, and then the causal effect of treatment can be estimated via the outcome differences over the two groups. However, in reality RCTs are not always practical to conduct due to time and money overhead and ethical reasons [16], and a new line of machine-learning based methods [17–19] have been proposed to directly estimate causal effect from large amount of observation data. Since it is nearly impossible to put city events into experimental settings, we follow the new line of work to only utilize observation data for estimation. Nevertheless, most counterfactual prediction works focus on the setting without unobserved confounders [20], which is impractical in real-world circumstances. To provide a more meaningful and solid counterfactual analysis, in this paper we argue that a key challenge of estimating causal effects of city events from observation data is how to eliminate the unobserved confounding bias (e.g., in our case, the people's willingness to go outdoors) that both affect the treatments (e.g., whether the city event would be held or not) and the outcomes (e.g., how many people will come). For example, if people are unlikely to travel to a certain district because of weather condition or safety issues (e.g., several shooting accidents just happened in this region), the government might also avoid holding gathering events (e.g., farmer market) in this area. Even if it does, the number of participants will not be large. Here we can see that the people's willingness to go out is a confounder, and if not handled properly, we might incorrectly draw a statistical conclusion that the farmer market is not attractive to people. Moreover, the unobserved confounder in our case can be time-varying (e.g., people's safety awareness can significantly increase after a gun shot accident and thus they will be less willing to go to a certain district), and if not well addressed, such case can bring huge bias into the causal estimation [21]. Fortunately, an alternative way to circumvent this problem is to adopt a weaker form of the unconfoundedness assumption [22] (i.e., we can observe all the confounders directly), where proxy variables can be used to reform the unobserved confounders [11,23].

To this end, our paper utilizes multi-sources of data to study our problem. We selected web search results of some safety-related keywords (e.g., stampedes, gunshot) in Google, the crime data, and weather data in New York City as proxy variables to capture people's willingness to go out at a certain district (i.e., the unobserved confounder). Then, we build a simple neural network-based framework to learn data representations and help with counterfactual analysis. Finally, we estimate the ITE of city events using human mobility and event data in New York City. The main contributions of our work are summarized as follows:

1) We are the first to address the important problem of estimating individual treatment effect of city events on human mobility using counterfactual prediction. This study can utilize the power of machine learning technology to assist better decision making in smart city management.
2) We develop a neural network-based framework with multi-sources of city data to learn better data representations and help counterfactual analysis.
3) We assess the causal effect of different categories of events and illustrate them from both spatial and temporal aspects to give a comprehensive understanding of their causal effects.

The remainder of this paper is structured as follows: In Sect. 2, we provide a detailed description and analysis of the data. Section 3 introduces the basic notations and formulations. The model framework is introduced in 4. Experimental results and causal analysis are presented in Sect. 5. Section 6 summarizes the related works of our study and Sect. 7 concludes the paper.

Table 1. Examples of the selected events and their categories.

1st-Categories	2nd-Categories	Examples
Entertainment	Popular festival	Tribeca Film Festival; Celebrate Brooklyn
	Block party/event	Ninth Avenue International Food Festival; Chinese Lunar Festival
	Parade	St. Patrick's Day Parade; Halloween Parade
Sales&Market	Farmer market	South Village Farmers Market; Inwood Saturday Greenmarket
	Sidewalk sales	Old Cathedral Outdoor Market; THE MARKETPLACE AT ST ANTHONY'S
	Other sales	Ralph Lauren Home Sample Sale; Carl Schurz Park Conservancy Plant Sale
Special Event	Filming/photography	Permitted Film Event; Untitled Aziz Ansari Project-TV Shoot
	Outdoor sports	NYRR-NYC Half Marathon; Swoldier Run
	Construction	City maintenance; Hernshead Boat Landing Renovation

2 Data Description and Analysis

In this section, we introduce the multi-sources data we use to assess the causal impact of events on human mobility. Some preliminary data analyses are also presented to show their potential capability of capturing the (unobserved) confounders.

2.1 Treatment - Urban Event

There's always something interesting going on in New York City, such as, New York concerts, Broadway musicals and performances, as well as New York's famous operas, sporting events, museums and galleries exhibitions, which provide us abundant events to research. We here collect the record of events from the Office of Citywide Event Coordination and Management (CECM)[3], which contains general information on approved event applications. In this paper, we select the time period 2015/01/01-2015/06/30 as the research time range and manually categorize the activity types into three groups: Entertainment, Sale&Market, and Special Event. Several examples and event descriptions are shown in Table 1.

2.2 Outcome - Human Mobility

We here utilize the taxi data to represent the actual human mobility. The data is collected from the first half of 2015 and provided by New York City's official website NYC's Taxi Commission[4]. Each taxi trip data in the dataset includes the taxi ID, the time stamp of the taxi pickup, the time stamp of the taxi drop off, the latitude and longitude of the pickup location, the latitude and longitude of the drop off location, the duration of the trip, and the travel time. distance, and the number of passengers. It should be noted that this article only uses records where the interval between the pick-up timestamp and drop-off timestamp is greater than or equal to 1 min. In addition, if the longitude and latitude of the drop-off location and the pick-up location are too close, we also assume that they belong to the same urban area and such records will not be considered in our experiments.

2.3 Confounder - People's Willingness to Go Out

The people's willingness to go outdoors is a time-varying confounder in our study, which can potentially affect the decision of holding city events and the resulted human mobility. To avoid yielding biased estimation of the causal effect, we have to learn the representation of the confounders and integrate them into training to eliminate the confounding bias [18,21]. Although the confounder here is hard to be quantified, we can still use the proxy variables to capture the confounder under a weaker form of unconfoundedness assumption [22]. In this paper, we choose three proxy variables (i.e. gun shooting related keywords on Google, crime data in NYC, and weather data in NYC) to learn the representation of confounders. Note that the data sources we choose here are subject to data availability and further sources of data can be added if suitable.

[3] https://data.cityofnewyork.us/City-Government/NYC-Permitted-Event-Information/tvpp-9vvx.

[4] https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

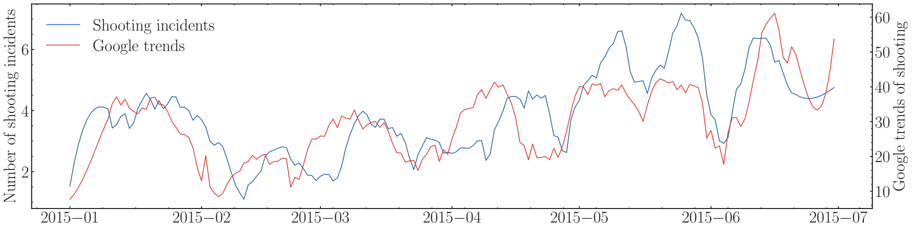


Fig. 2. As depicted, the Google trends regarding the gun shooting and the shooting incidents happening in NYC are shown together.

Google Trend. Google trends[5] is a completely free analysis tool based on Google search data. It analyzes billions of Google search engine search data every day and tells users the trend of a certain keyword or topic in various periods. The correlation between Google trends regarding to the gun shooting and the shooting incidents that occur in New York are shown in Fig. 2. The frequency of appearance in the Google search engine and its associated statistics can well establish the people's willingness to go outdoors.

NYC Crime. The Crime data were collected from the NYC OpenData portal, the New York City Open Dataset website[6]. This paper obtained the crime records from January 1, 2015 to June 30, 2015, each record contains the following attributes: crime category, latitude, longitude, and crime time. Similar to previous research, this paper divides New York City into 77 disjoint geographic regions based on New York City's police precincts because of their well-defined, historically recognized, and long-term stability. Several cases of distribution comparison between crowd flow and crimes are depicted in Fig. 4. We can observe that the nature of people always cling to avoid danger.

NYC Weather. There is a bunch of work [24–27], quantitatively measuring the impact of weathers on many aspects of transportation. Intuitively, people tend to stay home when extreme weathers happen. Thus, in this paper, the NYC weather is collected from website[7], including temperature, cloud cover, wind speed, etc., to reflect the people's willingness to go outdoors.

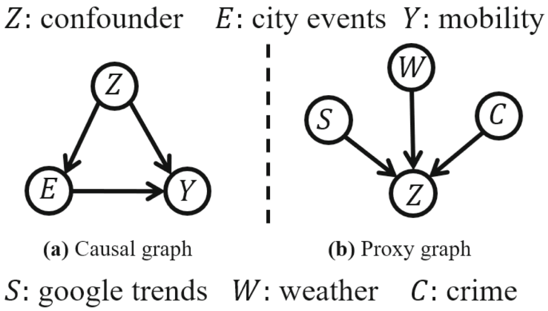
3 Preliminaries

In this section, we will briefly introduce the definition and assumption in causal inference.

[5] https://trends.google.com/trends/?geo=HK.

[6] https://data.cityofnewyork.us/Public-Safety/NYC-crime/qb7u-rbmr.

[7] https://darksky.net/forecast/40.7127,-74.0059/us12/en.



(a) Causal graph **(b)** Proxy graph

Fig. 3. Illustrating the causal diagram of city's events and human mobility.

Individual treatment effect (ITE) estimation aims to examine whether a treatment T affects the outcome $Y^{(i)}$ of a specific unit i. Let $\mathbf{x} \in \mathcal{R}^d$ denote the pre-treatment covariates of unit i, where d is the number of covariates. T_i denotes the treatment on unit i. In the binary treatment case, unit i will be assigned to the control group if $T_i = 0$, or to the treated group if $T_i = 1$. We follow the potential outcome framework proposed by [28,29]. If the treatment T_i has not been applied to unit i, $Y_0^{(i)}$ is called the potential outcome of treatment $T_i = 0$ and $Y_1^{(i)}$ the potential outcome of treatment $T_i = 1$. On the other hand, if the unit i has already received a treatment T_i, then Y_{T_i} is the factual outcome, and Y_{1-T_i} is the counterfactual outcome. In observational study, only the factual outcomes are available, while the counterfactual outcomes can never been observed. The individual treatment effect on unit i is defined as the difference between the potential treated and control outcomes:

$$\mathbf{ITE}_i = Y_1^{(i)} - Y_1^{(i)}. \tag{1}$$

The challenge to estimate $\mathbf{ITE}_i$ lies on how to estimate the missing counterfactual outcome. Existing counterfactual estimation methods usually make the following important assumptions.

Assumption 1: *(SUTVA). The potential outcomes for any unit do not vary with the treatment assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes [30].*

Assumption 2: *(Consistency). The potential outcome of treatment t equals to the observed outcome if the actual treatment received is t.*

Assumption 3: *(Ignorability). Given pretreatment covariates X, the outcome variables Y_0 and Y_1 is independent of treatment assignment, i.e., $(Y_0, Y_1) \perp\!\!\!\perp T|X$.*

Ignorability assumption makes the ITE estimation identifiable. In this paper, we make the assumption more plausible because we notice that the pretreatment covariates include the variables that affect both the treatment assignment and the outcome (i.e., confounders).

Assumption 4: *(Positivity). For any set of covariates x, the probability to receive treatment 0 or 1 is positive (i.e., $0 < P(T = t|X = x) < 1, \forall\ t$ and x).*

This assumption is also named as population overlapping [31]. If for some values of X, the treatment assignment is deterministic (i.e., $P(T = t|X = x) = 0$ or 1), we would lack the observations of one treatment group, such that the counterfactual outcome is unlikely to be estimated. Therefore, positivity assumption guarantees that the ITE can be estimated.

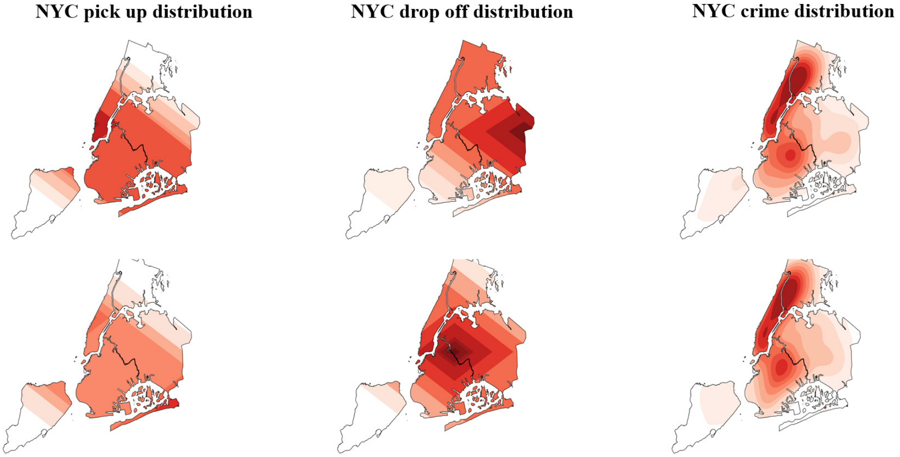


Fig. 4. Distributions comparisons between crowd flow and crime on different days are shown above. We can see that people consciously avoid encountering with that places with high rates of crime.

4 Methodology

With the aforementioned sources of data, in this section we aim to construct our causal framework and estimate the causal effect of city events on human mobility. Specifically, we first define our causal graph and capture the unobserved time-varying confounders from a set of observation data. Then a neural-network-based architecture is utilized to conduct counterfactual analysis which finally leads to the ITE estimation. The details are described as follows.

4.1 Causal Graph

The goal of this study is to understand the causal effect of city events on human mobility, using data on taxi pickups and drop-offs in NYC as an example. We propose a causal graph, as shown in Fig. 3, to represent the relationships between the different variables. Node E represents a list of city events, which have time ranges and locations and may influence human mobility. Node Y is the observed human mobility outcome, and the link $E \rightarrow Y$ represents the effect of events on human mobility. Node Z represents people's participation willingness, which is a confounder for both E and Y. People's participation willingness affects whether an event is held and the number of participants in an event. The link $Z \rightarrow E$ represents the effect of people's participation willingness on holding events, and the link $Z \rightarrow Y$ represents the effect of people's participation willingness on human mobility. Nodes S, W, and C represent features of Google search, weather, and crime, respectively. These variables are used as proxy variables to capture people's willingness to go outdoors, as they cannot be directly accessed. The links $S \rightarrow Z, W \rightarrow Z$, and $C \rightarrow Z$ represent the effect of these proxy variables on people's participation willingness (Fig. 5).

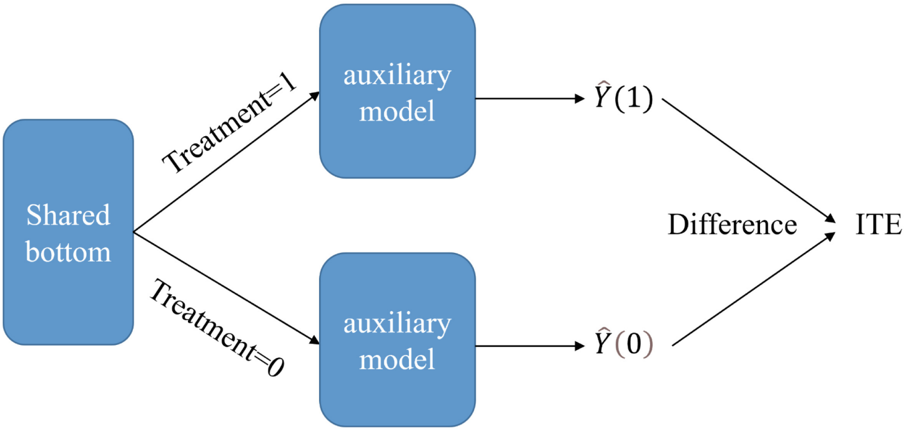


Fig. 5. Obtaining the causal effect via counterfactual analysis. E_t denotes the event list at t timestamp, and G is the features generator.

4.2 Learning the Representations of Time-Varying Confounders

We try to capture the time-varying confounders (e.g., people's participation willingness) through proxy variables. As shown in Fig. 3 (b), we assume the Google search of safety-aware keywords, weather conditions and crime rate can reflect people's participation willingness. Let n denotes the number of regions in NYC, and t denotes a certain time interval. We have $Z^t = \{z_i^t\}_{i=1}^n$, representing the unobserved confounders of people in different regions at time interval t. Similarly, we can define $S^t = \{s_i^t\}_{i=1}^n$, $W^t = \{w_i^t\}_{i=1}^n$, $C^t = \{c_i^t\}_{i=1}^n$ to represent the Google search results, weather condition and crime rate respectively. Specifically, $s^t, c^t \in R^n$ and their values are standardized to the interval of $[0, 1]$. Meanwhile, we let $w^t \in R^{n \times d}$, and d is the length of one-hot encoding which we use to represent weather condition. Inspired by the universal approximation theorem [32], we use a multi-layer perceptron with ReLU activation [33] as function $f(\cdot)$ to derive the unobserved confounders as follows:

$$z^t = f(s^t||w^t||c^t), \tag{2}$$

where || denotes the concatenate operator. However, as previously mentioned, the confounders are time-varying and have self-correlated temporal influences (e.g., people's participation willingness will not significantly increase just one day after a severe human stampede). Therefore, we adopt the idea of recurrent neural network [34] and propose the final formulation for z^t as:

$$z^t = RNN(\mathcal{H}^t, f(s^t||w^t||c^t), \tag{3}$$

where $\mathcal{H}^t$ is the hidden unit in the RNNs which captures the historical information.

4.3 Estimating ITE via Counterfactual Analysis

Counterfactual analysis asks: how will the outcome change if it's given an different treatment? In our case, we use E (events) to present the treatment. If E has a causal effect on Y, then a change in E will lead to a change in Y, keeping other variables as constants. The causal effect is equivalent to the magnitude of the change in Y caused by the intervention of E. Based on the values of the treatment variable E (E=1 and E=0), the outcome variable Y has two potential outcomes, $\hat{Y}(1)$ and $\hat{Y}(0)$, indicating the result with treatment or not. We propose a shared bottom architecture to handles the common variables, keeping them the same for auxiliary models, and the job of auxiliary models are to configure different parameters to simulate the situation of whether the treatment is given. In our paper, we use a simple two layers GCN [35] as the shared bottom model, and two MLPs for the auxiliary models. Specifically, let $X^t = \{x_i^t\}_{i=0}^n$ denotes the historical human mobility, the mathematical operation of the shared bottom model is shown as follows:

$$L^{t+1} = \hat{A} \cdot ReLU(\hat{A} \cdot (X^t||Z^t) \cdot W_1) \cdot W_0, \tag{4}$$

where $\hat{A}$ is the normalized adjacent matrix of the city graph, and W_1, W_0 are the parameters of GCN. After obtaining the latent variable L^{t+1}, we use two MLPs $(f_0(\cdot), f_1(\cdot))$ to get the two potential outcomes:

$$\hat{Y}(0) = f_0(L^{t+1})$$
$$\hat{Y}(1) = f_1(L^{t+1}).$$

Here $\hat{Y}(0)$ contains human mobility results with no city events and $\hat{Y}(1)$ contains the results with city events. The ITE can then be induced by the difference between the predicted potential outcomes: $ITE = |\hat{Y}(1) - \hat{Y}(0)|$. Finally, to supervise and train our model, we use the ground truth human mobility Y as supervise signals and formulate the MSE loss as $L_{mse} = \frac{1}{n}\sum_{t=1}^{T}|\hat{Y}^t - Y^t|$. Moreover, since the imbalanced confounders over different regions in the city will bring additional bias to the causal effect estimation [18], we also employ a distribution balancing constraint L_{cons} which is the Wasserstein-1 distance of the representation distributions. Overall, the training loss of our model is formulated as:

$$L = L_{mse} + \lambda L_{cons}, \tag{5}$$

where λ is a balance factor. After the model training, we can estimate the ITE of different city events using the differences between the two different potential outcomes.

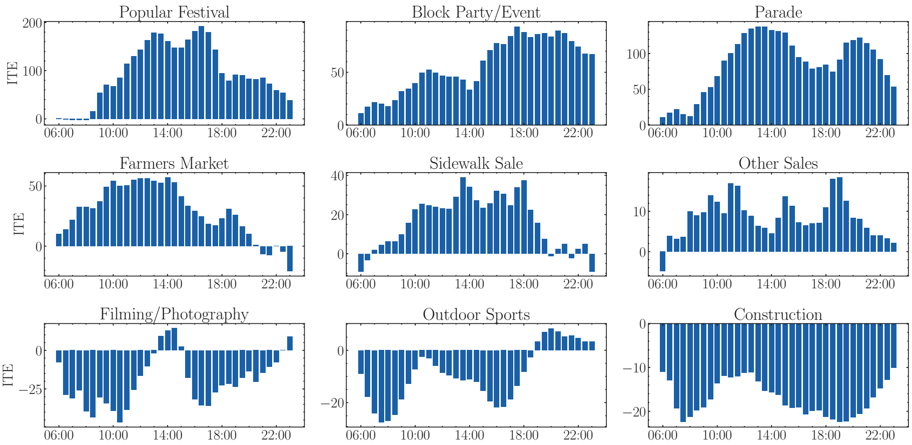


Fig. 6. Causal effect estimation of different event types on the workday.

5 Analyzing Estimated Treatment Effects

5.1 Entertainment

The category of entertainment, popular festival, block party/event, and parade, is shown in the top of Fig. 6 and Fig. 7. Comparing with the ITE values of other categories, entertainment tends to cause human gather. In terms of popular festival, the ITE values maintain in a high status overall day, regardless to workday and weekend. For example, the annual NYC Multicultural Festival[8] was held on May 30, 2015, the largest showcase of different cultures of the world in one place and at one time. In terms of a block party, it is a gathering where many members of a community come together and usually involves closing an entire city block to vehicular traffic or just one street. Many times, celebrations take place

[8] https://www.multiculturalfestival.nyc/.

in the form of music, games, dancing and food such as popcorn machines and barbecues. The peak of workday ITE is shown in the evening, due to the daily commuting. We can observe that the effect of block party significantly smaller than festivals. The last category: parade, is an activity in which a particular festival or event is celebrated by way of a mass procession and falls within the category of celebrations or festivals, it is very widespread in New York City. For instance, June 14, 2015, is the "National Puerto Rican Day Parade"[9] in the United States. On Fifth Avenue in Manhattan, New York, 80,000 people participated, and 2 million people watched the grand parade along the way. The parade continued along Fifth Avenue from 44th Street to 79th Street. For nearly 6 h, this is also one of the largest parades of the year in New York.

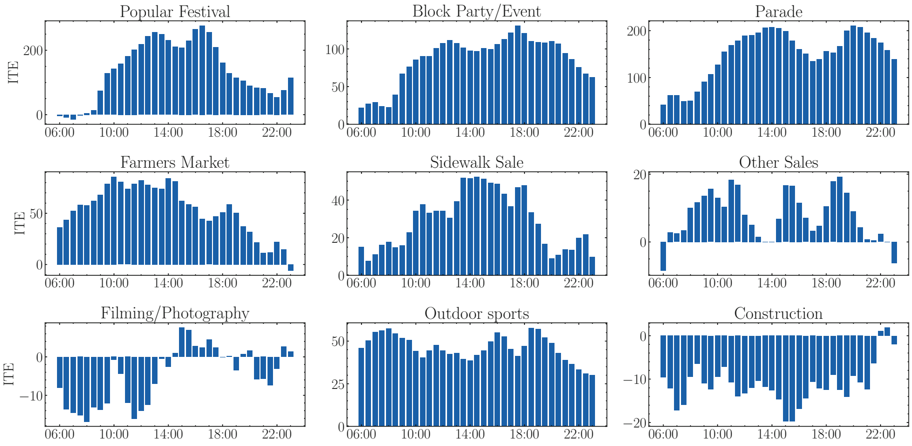


Fig. 7. Causal effect estimation of different event types on the weekend.

5.2 Sale and Market

Many supermarkets often hold special promotions during the day to attract people. As we can see in the middle of Fig. 6 and Fig. 7, a majority of people are inclined to go to the farmers markets to buy fresh products. Unlikely, the sidewalk sales, an outdoor sales event that retailers hold to get rid of end-of-season merchandise, more hold in an informal way. we can speculate from Fig. 6 that it will be greatly favored by those people who meet it causally. For instance, the Old Cathedral Outdoor Market can be found on a quiet street in Nolita, that quaint neighborhood north of Little Italy, against the backdrop of a brick wall on Prince Street.

5.3 Special Event

The causal effect of special events can be found at the bottom in Fig. 6 and Fig. 7. In fact, New York City is often considered the most filmed city in the

[9] https://www.nprdpinc.org/.

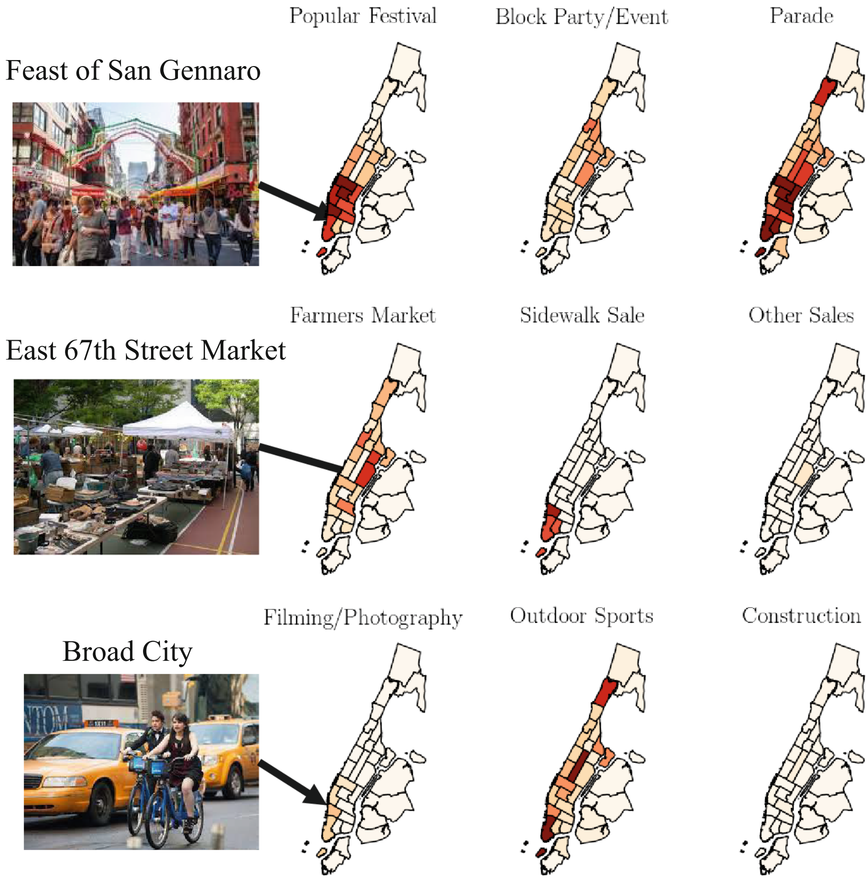
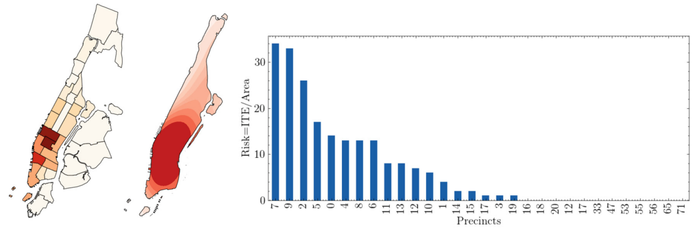


Fig. 8. Spatial ITE distribution in Manhattan with the case study.



(a) Average ITE of all categories and traffic collisions distribution (b) Rank of risk on precincts level

Fig. 9. The risk distribution calculated by ITE and precinct area, and traffic collision in (a), and rank of risk in (b).

Table 2. Performance comparison of MAE on NYC taxi Dataset on workday and weekend.

Method	Workday		Weekend	
	Pick up	Drop off	Pick up	Drop off
LR	420.64	501.72	32.52	55.28
HSIC-NNM	95.34	164.01	30.53	61.62
PSM	280.97	497.96	30.63	60.83
Causal Forest	184.97	279.72	24.34	44.38
BNN	149.60	332.93	24.34	47.55
TARNet	28.50	91.67	23.92	46.36
CFR-MMD	25.92	82.11	22.35	45.77
CFR-WASS	25.37	78.69	23.6	44.58
Ours	**23.61**	**44.81**	**23.5**	**43.39**
Ours-Trends	27.61	79.92	25.98	47.65
Ours-Crimes	29.42	80.46	28.44	44.58
Ours-Weather	29.73	78.84	24.30	44.99
Ours-Adj	26.33	81.89	26.07	45.89

world. Relive unforgettable film and television moments by visiting these iconic film locations. There are many authentic New York attractions featured in the famous Christmas movie Home Alone 2, such as the Empire Restaurant, Battery Park, Gapstow Bridge and Bethesda Terrace, as well as the Wolman Ice Rink in Central Park and, of course, Rockefeller Center. Moreover, Kelly's apartment in "Sex and the City" (66 Perry Street) attracts a lot of tourists every year. Nevertheless, the director usually locks the street or blocks to avoid a huge crowd. Thus, we can see the causal effect on filing/photography tends to be negative. From Fig. 6 and Fig. 7, we can speculate that people living in NYC have a great enthusiasm for sports since ITE values show positive results in the evening of the workday and in the general time of the weekend. For the last

Table 3. Estimated treatment effects of human mobility.

Treatment	Average Estimated ATE	Minimum	25% Quartile	75% Quartile	Maximum
Popular Festival	27.10	–3816.0	–78.0	184.0	3426.0
Block Party/Event	23.64	–3782.0	–36.0	78.0	3774.0
Parade	32.72	–3782.0	–29.0	91.0	3774.0
Farmers Market	56.38	–3816.0	–11.0	69.0	3770.0
Sidewalk Sale	143.67	–2219.0	–14.0	305.0	2278.0
Other Sales	–56.94	–3804.0	–368.0	242.0	2953.0
Filming/Photography	–6.29	–2015.0	–162.0	128.0	3347.0
Outdoor Sports	–0.78	–3905.0	–37.0	34.0	3431.0
Construction	3.87	–574.0	–17.0	23.0	387.0

attribution, without exception, people naturally sidestep the construction site (Table 3).

5.4 Spatial Distribution of ITE

For a clear understanding of the effect of ITE, here we illustrate the spatial distribution of ITE in Manhattan depicted in Fig. 8, since it contains the most holding events compared to others. We list three famous events, Feast of San Gennaro, East 67th Street Market, and Broad City, for the popular festival, farmer markets, and filming/photography, respectively, to better explain. We can notice that there is much crowd flow aggregation in several small blocks divided by the police precinct when the activities happen, which potentially increases the risk of emergencies. And, it is worth noting that people own the inertia to hold the events in the same place, especially for the festival and parade, which is challenging for public management. For the analysis of the implicit risk of aggregation of the crowd, we use the equation: $risk = \frac{ITE}{area}$ and show the result in Fig. 9, where $area$ is the area of the precinct. Intuitively, the risk of emergencies can be defined as how many events contribute to the density of the crowd. From Fig. 9, we can conclude that precincts 7, 9, and 2, have a relatively large risk than others since a majority of events prefer to hold in these places, such as popular festivals, parades, and outdoor sports. We also show the NYC traffic collision data collected from NYC open data[10]. The side confirming our generated causal effect is reasonable.

5.5 Experimental Setups

Our experiments are implemented with PyTorch 1.6.0 and Python 3.6, and trained with eight RTX2080Ti GPUs. The platform we run on is Ubuntu 16.04 OS. We train our model using Adam optimizer with an initial learning rate of 0.001. Dropout with 0.5 retaining rates is applied to the outputs of the graph convolution layer.

5.6 Baseline Methods

We evaluate our curriculum strategy with the following method:

- LR: Least square Regression.
- HSIC-NNM [36]: Nearest neighbor matching based Hilbert-Schmidt Independence Criterion based on nearest neighbor matching by learning subspaces that are predictive of the outcome variable for both the treatment group and the control group.
- PSM [37]: Propensity score, the conditional probability of assignment to a particular treatment given a vector of observed covariates that is consistent with logistic regression.

[10] https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/data.

- Causal Forest [38]: Tree and forest-based method for estimating heterogeneous treatment effects that extend Breiman's widely used random forest algorithm.
- BNN [39]: Balancing neural network for counterfactual inference which brings together ideas from domain adaptation and representation learning.
- TARNet [18]: Treatment-Agnostic Representation Network that captures nonlinear relationships underlying features to fit the treated and controlled outcome.
- CFR-MMD [18]: counterfactual regression with MMD metric that attempts to find balanced representations by minimizing the MMD metric between the treated and controlled individuals.
- CFR-WASS [18]: counterfactual regression with the Wasserstein metric that attempts to find balanced representations by minimizing the Wasserstein distance between the treated and controlled individuals

5.7 Result Analysis Comparing with Causal Baselines

To further investigate the capability of our framework, we conduct extensive experiments to compare the causal baselines, estimating human mobility over workday and weekend. Table 2 compares the performance of different methods in predicting the treatment assignment (whether the events are held at specific places) and outcomes (the numbers of pick up and drop off at these places). We also performed the ablation study by removing the variable of Google trends, crime, weather, and adjacent matrix, respectively, indicating that variable can well represent the cofounder [18]. We can see that our method can significantly outperform state-of-the-art methods.

5.8 Estimating Causal Effects on Synthetic Data

Generating synthetic data to evaluate the credibility of causal estimation is a common strategy [40,41]. To simplify, we here calculate the mean flow from 2015/01/01 to 2015/06/31 to eliminate the effect of events and assume it as the baseline, which presents it as the normal flow. Then, we applied the Gaussian distribution $\mathcal{N}(\mu, \sigma)$, where μ and σ denote the mean and variance, respectively. it's hard to obtain data that the actual number of participants. Therefore, we use the estimated result from Fig. 6 and Fig. 7 as the μ and define the values of σ. The estimation on synthetic data has shown in Table 4. We can see that the estimating error is still quite large, which can be blamed on two parts. 1) the data generating process may be inconsistent with the real world; 2) the causal estimation is unsupervised learning, which is challenging to learn precise prediction.

Table 4. Estimated treatment effects of human mobility in synthetic data.

Treatment	True ATE	Estimated ATE	Minimum	25% Quartile	75% Quartile	Maximum
Popula Festival	54.51	54.89	–3269.0	–157.0	92.0	2929.0
Block Party/Event	37.14	35.52	–3269.0	–72.0	63.0	2784.0
Parade	25.98	24.36	–3214.0	–58.0	64.0	2784.0
Farmers Market	66.58	64.96	–3269.0	–129.0	105.0	2784.0
Sidewalk Sale	7.87	6.25	–2112.0	–130.0	200.0	1857.0
Other Sales	9.23	7.61	–2230.0	–258.0	277.0	1882.0
Filming/Photography	165.91	162.29	–3214.0	–373.0	113.0	1552.0
Outdoor Sports	9.92	8.30	–3269.0	–17.0	32.0	2784.0
Construction	8.6	7.98	–574.0	–24.0	21.0	263.0

6 Related Work

6.1 Emergency Analysis

In the fields of urban planning, large-scale event management, building design, fire safety procedures, fire rescue and other fields, personnel safety issues in congested environments have received high attention. Safety capacity control is an effective measure to prevent excessive crowd gathering, avoid crowded and trampled and other vicious events, and is the key to ensuring urban public safety. For public places such as squares, parks, commercial blocks, leisure and entertainment spaces, and tourist attractions, there is a greater risk of crowd congestion. Therefore, it is of great significance to fundamentally grasp the temporal and spatial distribution characteristics of crowd gatherings, formulate safety capacity control standards that are more in line with the management needs of open public places, and effectively prevent and control personnel congestion. From the perspective of environmental factors, [42] studied the factors that affect the safety of crowds, and believed that in a crowded state, a small number of people had an accident. If the information in the crowd is asymmetric, the crisis will quickly spread to the surrounding crowd, and then the order will be out of control. This causes panic and confusion. [43] pointed out that the design defects of venues or public places may lead to hidden dangers to the safety of the crowds gathering in the area and aggravate the difficulty of crowd management. [44] proposed a real-time crowd density estimation method based on Markov random field (MRF), and introduced the application steps of the method in detail. [45] selected 13 subway stations in Beijing, conducted an empirical analysis from the feasibility and applicability of establishing a crowded stampede risk model, and applied the crowd aggregation risk research and judgment to practice.

6.2 Urban Flow Prediction

With the acceleration of the urbanization process, people's demand for travel is increasing, and traffic congestion is becoming more and more serious, which restricts the development of China's cities. At present, it is urgent to establish an intelligent transportation system to help people rationally plan traffic routes and alleviate traffic congestion. The core of the intelligent transportation system is to accurately predict future traffic flow conditions, so as to assist the tasks in the system. Urban traffic flow prediction is an important research problem in the field of spatiotemporal data mining, especially in the context of the existing big data era. [46] summarizes the latest progress in spatiotemporal data mining, and summarizes recent work in terms of research questions and corresponding methods. Spatiotemporal data is usually affected by other external environmental semantic information, such as weather, vacation, and surrounding environment. [47] proposed the ADAIN model to integrate air quality monitoring data and road network data.

6.3 Causal Inference

The goal of causal inference [48] is to discover the causal relationship behind variables (things). [49] divides causality into three levels, the first level is "ssociation", the second level is "intervention", and the third level is "counterfactual reasoning". He believes that the current research is only at the first level, which is "weak artificial intelligence". To achieve "strong artificial intelligence", intervention and counterfactual reasoning are needed, that is, causal inference. At present, the academic community generally believes that the ability of models to learn causal relationships is a key part of the road to strong artificial intelligence [50]. In the history of scientific research, causal learning has been applied in countless important fields, including education [51], medicine [52], economics [53], meteorology [54], and environmental health [55]. In machine learning, causal inference has had many applications, and some problems of learning causal relationships can be attributed to supervised learning problems. Once the data is labeled with causality, the problem of learning causality can be transformed into a prediction problem. The main difficulty here is obtaining labels for the causal direction. For some datasets, causality was confirmed to exist in the dataset [56].

6.4 Discussion

In addressing the confounding effects introduced by individuals' willingness to participate, our research extends beyond mere correlation to explore the causal relationship between social events and human mobility. We employ a counterfactual-based method to mitigate the interference of cofounding variables. A causal diagram is constructed to facilitate the understanding of these causal influences. Our analysis considers the impact of urban activities on mobility patterns during both workdays and weekends, revealing unique patterns associated with various activities (e.g., outdoor sports, community events) across different days. These insights have numerous implications for researchers and the mobility-oriented community and can be applied in a multitude of contexts.

Our causal framework bears significant relevance to the field of ubiquitous computing. As per our prior data analysis, the positive correlations between crime and mobility distribution intuitively align with expectations, although they are confounded by variables that concurrently influence both the propensity to conduct activities and human mobility. The causal analysis we've developed can alleviate these confounding effects in correlation analysis through counterfactual stratification. With advancements in ubiquitous computing and thanks to the NYC government, we've been able to gather data from a variety of sources, including wearable devices and mobile phones. Considering privacy and ethical concerns, we refrained from using user trajectory information, focusing instead on pickup and drop-off locations. The causal effects of various events we've identified may indirectly benefit downstream applications, such as traffic management and emergency prevention. By understanding mobility patterns during different events, adequate resources can be allocated in advance to prevent tragedies such as those that occurred in Itaewon, Seoul, South Korea, and the New Year's Eve Stampede at Chen Yi Square in Shanghai.

Our causal analysis uncovers the susceptibility of human mobility patterns to various social events. For example, in Manhattan, our estimations highlight that event locations and preferred venues tend to concentrate, particularly for festivals and parades, posing a significant challenge to public management. The distribution of traffic collisions is also depicted to corroborate our hypothesis, aligning with risk distribution patterns. Our findings suggest that the government should consider offering diverse activity venues to alleviate crowd pressure in Lower and Midtown Manhattan. On a personal level, individuals should exercise caution when participating in crowded events. We further validate our proposed model using synthetic data, which provides a ground truth for the average treatment effect.

7 Conclusion

In this paper, we quantify the causal effect of urban events on human mobility. We collect GPS data from the NYC taxi from January 1, 2015, to June 30, 2015. The corresponding events also are obtained from the Office of Citywide Event Coordination and Management, serve as the treatment from a causal aspect. We also consider the people's participation willingness as the cofounder and use crime, google trends, and weather as the proxy variable, which have been demonstrated by data analysis to evaluate the relationship. Three kinds of events categories are extracted from the data, including entertainment, sales&market, and special event. We propose a framework based on the counterfactual method to estimate the causal effect of urban events on human mobility by removing the bias brought by confounding effects. Our results provide new insight into understanding the role of urban events in influencing human mobility. Popular festivals and parade have significant estimated effects on human mobility. In the meanwhile, they also tend to hold in a specific region, which implicitly increases the risk of the crowd. We also qualitatively evaluate it by visualizing the traffic collision distribution that mainly concrete in Lower and Midtown Manhattan. We also calculate the risk of the precinct in Manhattan to provide quantitative values to the government. In terms of this, more event venues should be provided by the government to allow citizens diverse choices of holding places, which can maximize and reduce the burden of public law and order.

Acknowledgements. This work was partially supported by the grants of National Key Research and Development Program of China (2021YFB1714400) of China and Guangdong Provincial Key Laboratory (2020B121201001).

References

1. Illiyas, F.T., Mani, S.K., Pradeepkumar, A.P., Mohan, K.: Human stampedes during religious festivals: a comparative review of mass gathering emergencies in India. Int. J. Disaster Risk Reduct. **5**, 10–18 (2013)
2. Vanumu, L.D., Laxmikant, K., Rao, K.R.: Human stampedes at mass gatherings: an overview. Collect. Dyn. **5**, 502–504 (2020)
3. Ying, L., Qiu, L., Lyu, X., Jiang, X.: Human stampede causative factors and cluster risk: a multi-dimensional analysis based on isodata and fuzzy theory. Int. J. Disaster Risk Reduct. **66**, 102581 (2021)
4. Ho, T.-H., Lim, N., Reza, S., Xia, X.: Om forum-causal inference models in operations management. Manuf. Serv. Oper. Manag. **19**(4), 509–525 (2017)
5. Yusuf, F., Cheng, S., Ganapati, S., Narasimhan, G.: Causal inference methods and their challenges: the case of 311 data. In: DG. O2021: The 22nd Annual International Conference on Digital Government Research, pp. 49–59 (2021)
6. Fan, Z., Song, X., Xia, T., Jiang, R., Shibasaki, R., Sakuramachi, R.: Online deep ensemble learning for predicting citywide human mobility. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, no. 3, pp. 1–21 (2018)
7. Stange, H., Liebig, T., Hecker, D., Andrienko, G., Andrienko, N.: Analytical workflow of monitoring human mobility in big event settings using bluetooth. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness, pp. 51–58 (2011)
8. Tyagi, B., Nigam, S., Singh, R.: A review of deep learning techniques for crowd behavior analysis. Arch. Comput. Methods Eng. **29**(7), 5427–5455 (2022)
9. Huang, H., Yang, X., He, S.: Multi-head spatio-temporal attention mechanism for urban anomaly event prediction. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 5, no. 3, pp. 1–21 (2021)
10. Zhang, J., Feng, B., Yina, W., Pengpeng, X., Ke, R., Dong, N.: The effect of human mobility and control measures on traffic safety during covid-19 pandemic. PLoS ONE **16**(3), e0243263 (2021)
11. Louizos, C., Shalit, U., Mooij, J.M., Sontag, D., Zemel, R., Welling, M.: Causal effect inference with deep latent-variable models. Adv. Neural Inf. Process. Syst. **30**, 1–11 (2017)
12. van der Laan, M.J., Petersen, M.L.: Causal effect models for realistic individualized treatment and intention to treat rules. Int. J. Biostat. **3**(1), 1–55 (2007)
13. Bennett, J.: Event causation: the counterfactual analysis. Philos. Perspect. **1**, 367–386 (1987)
14. Ramachandran, M.: A counterfactual analysis of causation. Mind **106**(422), 263–277 (1997)
15. Deaton, A., Cartwright, N.: Understanding and misunderstanding randomized controlled trials. Social Sci. Med. **210**, 2–21 (2018)

16. Guo, R., Cheng, L., Li, J., Hahn, P.R., Liu, H.: A survey of learning causality with data: problems and methods. ACM Comput. Surv. (CSUR) **53**(4), 1–37 (2020)
17. Cui, P., et al.: Causal inference meets machine learning. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3527–3528 (2020)
18. Shalit, U., Johansson, F.D., Sontag, D.: Estimating individual treatment effect: generalization bounds and algorithms. In: International Conference on Machine Learning, pp. 3076–3085. PMLR (2017)
19. Yoon, J., Jordon, J., Van Der Schaar, M.: Ganite: estimation of individualized treatment effects using generative adversarial nets. In: International Conference on Learning Representations (2018)
20. Zou, H., Li, B., Han, J., Chen, S., Ding, X., Cui, P.: Counterfactual prediction for outcome-oriented treatments. In: International Conference on Machine Learning, pp. 27693–27706. PMLR (2022)
21. Rubin, D.B.: Bayesian inference for causal effects. In: Handbook of Statistics, vol. 25, pp. 1–16 (2005)
22. Donald, S.G., Hsu, Y.C., Lieli, R.P.: Testing the unconfoundedness assumption via inverse probability weighted estimators of (L) ATT. J. Bus. Econ. Stat. **32**(3), 395–415 (2014)
23. Imbens, G.W.: The role of the propensity score in estimating dose-response functions. Biometrika **87**(3), 706–710 (2000)
24. Maze, T.H., Agarwal, M., Burchett, G.: Whether weather matters to traffic demand, traffic safety, and traffic operations and flow. Transp. Res. Rec. **1948**(1), 170–176 (2006)
25. Cools, M., Moons, E., Wets, G.: Assessing the impact of weather on traffic intensity. Weather Clim. Soc. **2**(1), 60–68 (2010)
26. Hranac, R., Sterzin, E., Krechmer, D., Rakha, H.A., Farzaneh, M.: Empirical studies on traffic flow in inclement weather (2006)
27. Rakha, H., Farzaneh, M., Arafeh, M., Sterzin, E.: Inclement weather impacts on freeway traffic stream behavior. Transp. Res. Rec. **2071**(1), 8–18 (2008)
28. Splawa-Neyman, J., Dabrowska, D.M., Speed, T.P.: On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat. Sci. 465–472 (1990)
29. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**(5), 688 (1974)
30. Imbens, G.W., Rubin, D.B.: Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, Cambridge (2015)
31. D'Amour, A., Ding, P., Feller, A., Lei, L., Sekhon, J.: Overlap in observational studies with high-dimensional covariates. J. Econometr. **221**(2), 644–654 (2021)
32. Yulong, L., Jianfeng, L.: A universal approximation theorem of deep neural networks for expressing probability distributions. Adv. Neural. Inf. Process. Syst. **33**, 3094–3105 (2020)
33. Agarap, A.F.: Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375 (2018)
34. Medsker, L.R., Jain, L.C.: Recurrent neural networks. Design Appl. **5**, 64–67 (2001)
35. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
36. Chang, Y., Dy, J.: Informative subspace learning for counterfactual inference. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
37. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)

38. Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests. J. Am. Stat. Assoc. **113**(523), 1228–1242 (2018)
39. Johansson, F., Shalit, U., Sontag, D.: Learning representations for counterfactual inference. In: International Conference on Machine Learning, pp. 3020–3029. PMLR (2016)
40. Sun, W., Wang, P., Yin, D., Yang, J., Chang, Y.: Causal inference via sparse additive models with application to online advertising. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
41. Hasthanasombat, A., Mascolo, C.: Understanding the effects of the neighbourhood built environment on public health with open data. In: The World Wide Web Conference, pp. 648–658 (2019)
42. Baron, R.A., Richardson, D.R.: Human Aggression. Springer, Heidelberg (1994)
43. Tubbs, J., Meacham, B.: Egress Design Solutions: A Guide to Evacuation and Crowd Management Planning. John Wiley & Sons, Hoboken (2007)
44. Paragios, N., Ramesh, V.: A mrf-based approach for real-time subway monitoring. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, p. I. IEEE (2001)
45. Yan, L., Tong, W., Hui, D., Zongzhi, W.: Research and application on risk assessment dea model of crowd crushing and trampling accidents in subway stations. Procedia Eng. **43**, 494–498 (2012)
46. Atluri, G., Karpatne, A., Kumar, V.: Spatio-temporal data mining: a survey of problems and methods. ACM Comput. Surv. (CSUR) **51**(4), 1–41 (2018)
47. Cheng, W., Shen, Y., Zhu, Y., Huang, L.: A neural attention model for urban air quality inference: Learning the weights of monitoring stations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
48. Peters, J., Janzing, D., Schölkopf, B.: Elements of Causal Inference: Foundations and Learning Algorithms. The MIT Press, Cambridge (2017)
49. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic books (2018)
50. Pearl, J.: Theoretical impediments to machine learning with seven sparks from the causal revolution. arXiv preprint arXiv:1801.04016 (2018)
51. Dehejia, R.H., Wahba, S.: Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. J. Am. Stat. Assoc. **94**(448), 1053–1062 (1999)
52. Mani, S., Cooper, G.F:. Causal discovery from medical textual data. In: Proceedings of the AMIA Symposium, p. 542. American Medical Informatics Association (2000)
53. Imbens, G.W.: Nonparametric estimation of average treatment effects under exogeneity: a review. Rev. Econ. Stat. **86**(1), 4–29 (2004)
54. Ebert-Uphoff, I., Deng, Y.: Causal discovery for climate research using graphical models. J. Clim. **25**(17), 5648–5665 (2012)
55. Li, J., Zaïane, O.R., Osornio-Vargas, A.: Discovering statistically significant colocation rules in datasets with extended spatial objects. In: Bellatreche, L., Mohania, M.K. (eds.) DaWaK 2014. LNCS, vol. 8646, pp. 124–135. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10160-6_12
56. Lopez-Paz, D., Nishihara, R., Chintala, S., Scholkopf, B., Bottou, L.: Discovering causal signals in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6979–6987 (2017)

Modeling of the Human Cognition for the Metaverse-Oriented Design System Development

Yan Hong[1](✉), Zhonghua Jiang[1,2], Song Guo[3], Xianyi Zeng[4], and Xinping Li[1]

[1] Soochow University, 178 Ganjiang Road, 215021 Suzhou, China
yanhong@poly.edu.hk

[2] College of Textile Engineering, Taiyuan University of Technology, 030600 Jinzhong, China

[3] The Hong Kong Polytechnic University, 11 Kowloon Hung Hom Yuk Choi Road, Hong Kong 100872, China

[4] University of Lille, Métropole européenne de Lille, Roubaix 59100, France

Abstract. Metaverse can fully satisfy users' design scenarios, meeting their emotional and functional imaginations without any physical constraints, and its superior functionality makes it particularly suitable for the design domain. However, because the interactivity that the Metaverse can provide has not been fully exploited, its application in the design domain has not been studied. In this paper, we propose an interactive design system for the metaverse to enhance human interaction with the metaverse and to realize the interconnection between digital and real spaces. The system is expected to integrate the physical content of the design (including product components) with the designer's expertise. To this end, we modeled the product design process and its associated design knowledge, which is key to realizing the system. The designer's expertise was extracted and modeled as multiple perceptual-cognitive models. A fuzzy transformer method was innovatively developed for the computational modeling of the perceptual cognitive models. Using these models, user-system interactions and interactions between virtual and real products were enhanced. This work provides a conceptual framework for a Metaverse interaction design system based on computational modeling of human perceptual cognition. The proposed system greatly extends the scope of Metaverse applications for the development of various product design systems.

Keywords: Metaverse Design System · Design support system · Knowledge-based system · Human-centred design · Fuzzy transformer

1 Introduction

Metaverse can provide users with a connection between the real world and the virtual world, bringing them an immersive experience [2], and it is considered to

R. Akerkar (Ed.): SAIDD 2023, CCIS 1810, pp. 65–79, 2024.
https://doi.org/10.1007/978-3-031-53770-7_5

be another revolution in the Internet [1]. Meta-world can be regarded as an extension of today's Internet with a digital representation. It is a three-dimensional virtual space in which users can interact through avatars [3]. Metaverse space and the avatars inside it create a network of virtual worlds [4] in which users can socialize, trade, and other activities [5]. Current metaverse worlds are more commonly found in the gaming and entertainment industry [1], but the value of metaverse to the design field is also immense.

In terms of design, Metaverse provides a new virtual environment [6]. Through the use of augmented reality and virtual reality technologies, users' design solutions can be better met in this environment, fulfilling their emotional and functional imaginations without any physical constraints. In other words, Metaverse provides designers with an imaginative space to generate creations with a wide range of possibilities, and it is particularly suitable for design. In this space, designers can imagine their emotional and functional design concepts and evaluate their design solutions in a digital environment. Whether it's for various scenarios such as dressing an avatar [7], giving a lecture [8], building a house [9], or hosting a concert [10], designers' imaginations can be put to great use. In fact, the image space provided by the Metaverse immersive interactive environment enhances the interaction between designers and users. This interaction enables systematic human-product-environment interaction, which greatly facilitates human-centered design. This interaction can be between the present and the future or between the real world and the virtual world. This interaction blends digital environments and sensory experiences, taking digital design to a whole new level.

However, the interactive nature of the Metaverse has not yet been fully realized and exploited. This is because the digital and real spaces are not yet fully connected. The digital environment of the metaverse should have integrity and be able to fully integrate the physical content of the design, i.e., the product components and the designer's expertise. Therefore, design systems that can connect digital and real spaces are crucial for human interaction with the metaverse. Currently, the prevalent design system research is from a technology-oriented perspective, starting from the technical aspects of design system development and focusing on the creation and application of innovative technologies such as methodologies and software prototypes [11].

Design is an advanced human activity [12]. Design practices are accomplished based on a universal design process and its associated design knowledge [12]. Knowledge is becoming increasingly important in the development of systems in the meta-universe [7]. The realization of these systems in the meta-universe is only possible if the design knowledge is fully and rationally utilized in the different design processes; therefore, these digital tools are limited. Their limitations are mainly reflected in the fact that (1) they can only provide basic technical support for the design process [14]; (2) they are developed individually for specific design phases; and (3) these digital tools cannot be used systematically in conjunction or integrated into the complete apparel design process because the design knowledge is not sufficiently extracted and applied to the design process.

In order to realize the proposed Metaverse interactive and personalized design system, different design knowledge and its associated design processes should be fully extracted and utilized as the basic computational models supporting the proposed design system. With these computational models, designers can fully utilize their expertise in their imaginary world (digital environment) and easily translate their design solutions into a real manufacturing environment. Subsequently, human-product interactions as well as real-digital interactions in the digital environment can be optimized, and Metaverse's application scenarios can be expanded. As a result, the Metaverse design system is more advanced than all previous design systems.

The proposed system follows the general design process of "design-demonstrate-evaluate-adapt". Different computational perceptual cognitive models will be developed to extract relevant knowledge to support this process. The main parts of this work are as follows:

- We developed a knowledge-based interactive design system for Metaverse to enhance the interaction between users and Metaverse. Different computational perceptual cognitive models are developed to support the proposed system, ensuring that the system is able to generate design entities that correspond to user requirements.
- A general design process of "design-display-evaluate-adjust" is proposed and applied to the proposed system to realize the interaction mechanism of the system.
- In terms of computational modeling of perceptual-cognitive models, a fuzzy transformer approach is innovatively developed. These models form the basic models for design-related knowledge modeling.

2 The Proposed System and Its Disciplinary

2.1 Principle of the Proposed System

The real-world environment, the Metaverse virtual environment, the proposed system, and its interconnections are all included in the proposed system's structure, which is depicted in Fig. 1. The suggested system ties the outside world and the Metaverse together. The "Generate-Display Evaluate-Adjust" cycle is the general operating principle. All interactions between the real world and the Metaverse virtual environment are made possible by this mechanism, including those between user systems and the Metaverse and those between virtual and real products. To support these interactions, there is a design knowledge base that experts have already defined. It is a knowledge repository for all kinds of design expertise. Different perceptual cognitive computational models represent these many sorts of design information. They simulate the connection between product elements and design perceptions.

A user from the real world will utilize his or her avatar from the Metaverse to engage with the system in a genuine design case utilizing the suggested system. The user must first input their design specifications into the system. The

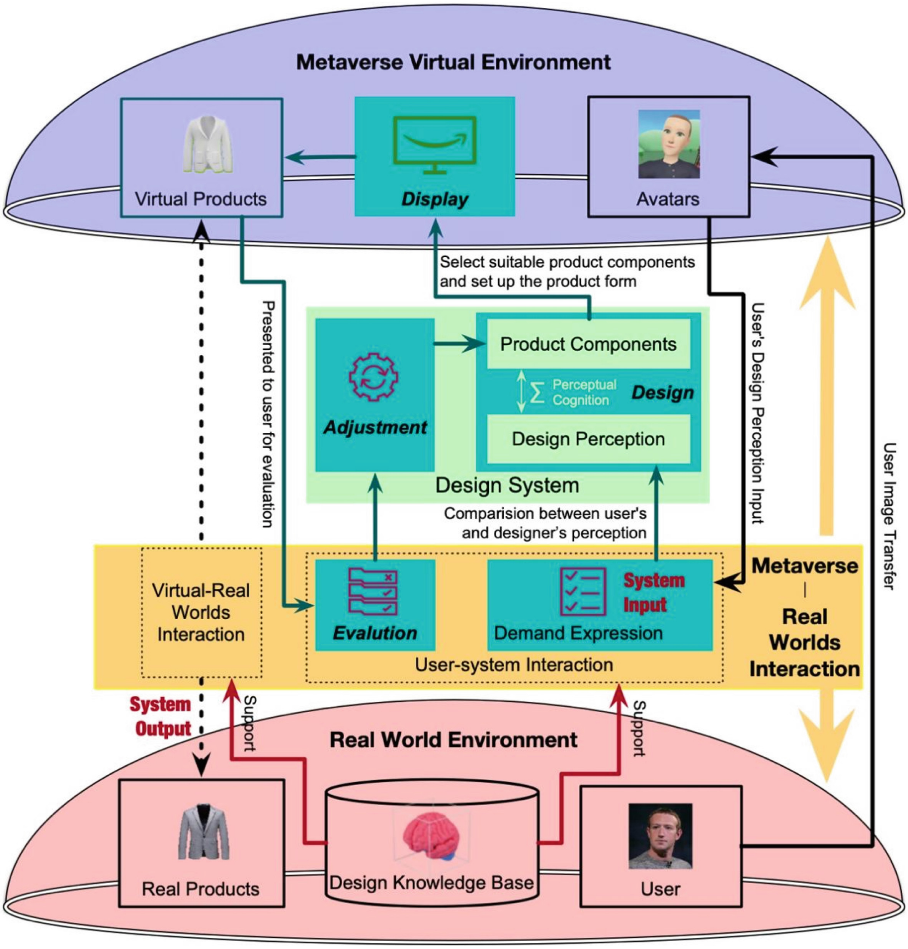


Fig. 1. Working flowchart of the proposed automatic pattern generation system.

design process will then become functional in the system. The user's design needs and the model's design perceptions of various product components will be compared using the perceptual cognitive computational model of the design knowledge base. The design outcome will be the pertinent product component with the highest design perception in relation to the consumer's requirement. The virtual product is then presented in Metaverse in a virtual form when the display software has been run. The user can indicate discontent with the generated design result because the assessment software is integrated within the system. The developed design result will be output and then produced in the real world if the evaluation result is satisfactory. The physical parameters of the suitable virtual product can be automatically developed and transferred to the smart factory in the actual world with the help of the provided design knowledge base. The product will be made swiftly and given to the customer. On the other hand, a method for adjustment will be carried out if the evaluation results are unfavorable. The user must recognize the problematic product elements.

The system will study customer dissatisfaction with the aid of the Design Knowledge Base before changing other product components. There will be new design outcomes produced. The Design Knowledge Base and the perceptual cognitive computing model will then be updated. A fresh set of customized product design results will be produced when the feedback is assessed by the model changing various knowledge base rules. Until suitable evaluation findings are guaranteed, this procedure is repeated.

2.2 Related Concepts and Methods

Design, Design Knowledge and Its Computational Modelling. Design is a brain activity focused on interaction that is predicated on that interaction [13]. Design demands, design perceptions, and their interactions are the fundamental phenomena that are primarily studied in relation to the phenomena and rules in the design process. The user's view of a design is called design demand, and it is typically semantically pre-pressurized, such as "modern style." Design perception is the result of a long history of design practice and is characterized as a "form-style" interaction that has a beneficial impact on the design process [15]. Form alludes to a product's tangible parts. For instance, in the design of clothing, the term "apparel form" refers to the strong physically functional elements that make up the garment's overall shape, such as the sleeve and collar shapes. Style (emotional expression of product components) is the subjective opinion of the designer regarding the aesthetics of the product components. For instance, style in the context of fashion refers to the appearance of a clothing detail. It has significant emotional characteristics and is articulated by generic semantics like "avant-garde style". Designers typically view "shoulder pad design" in this instance as being more avant-garde. Additionally, it may be said that although clothing style is a physical state, fashion style is a perceptual condition.

"Communicate-conceptualize-solve" is the fundamental design method. This design method is reflected in the "Generate-Display-Evaluate-Adjust" cyclic interactive structure utilized in this system. The designer will obtain the design requirements, or the user's perceptions, during the communication phase. The semantics of describing the user's perception will be understood by the designer. The designer will then work alone to complete the ideation stage. The designer will initially engage with this phase's design requirements and impressions mentally. To create a design solution, eliminate product forms with emotional semantics or user perception semantics that are similar. For each category of product component, there are several potential choices during the actual design process. Consider the term "sleeve type" in the context of fashion; examples include "lamb's leg sleeve," "shoulder insertion sleeve," and others. Sleeve type is a component of garment style. As a result, to create the final design solution, the designer will choose the product components in each category with the highest degree of matching.

It can be said that the “form-style” relationship forms the cornerstone of how design needs and design perceptions interact. Until a final, satisfactory design is generated, this process typically takes several rounds of engagement. This makes the perceptual perception application scenario very interactive and provides easy access to feedback. When considering the processes of its use, this aspect needs to be thoroughly taken into account.

Design Perception and Its Descriptive Space. A Perception-Cognition Description Space (PCDS) must be constructed in order to ascertain user and designer perspectives. It can be used to represent both the user’s design needs (user requirements) and the product’s design perception (style). Adjectives often signify the scale of expression that conveys the perception of something according to the kanji principle. So, a set of paired adjectives make up the proposed Perception-Cognition Description Space (PCDS). The kind and quantity of adjectives used are essential for ensuring the accuracy of the perception/need portrayal.

Assuming that the Perceptual-Cognitive Descriptive Space (PCDS) is defined as S with l dimensions in a left-right comparison format, and taking fashion as an example, the space can be expressed as:

$$S = \{SP_1, SP_2, ...SP_{l-1}, SP_l\} = \{\text{“}\textit{simple-complex}\text{”}, \text{“}\textit{formal-casual}\text{”}...\text{“}\textit{classical-modern}\text{”}\}.$$

In this research, we define F_c as a 7-segment semantic evaluation set denoted as:

$$\{F_1, F_2, F_3, F_4, F_5, F_6, F_7\}.$$

It can be used to organize the l dimensions into a perceptual evaluation table. Take “simple-complex” using 7 segment criteria as an example, it can be identified as “extremely simple-very simple-relatively simple-moderate-relatively complex-very complex-extremely complex”.

Product Components and Product Components Matrix (PCM). Ontology theory, which examines the generative and physical structure of products and their interrelationships, is the basic foundation for the study of product component matrix. It is determined utilizing techniques for morphological analysis. Products, product component categories, and product components are typically its three categories. They are set up in a tree-like form. The product component category and the relationship between its components, using fashion as an example, can be described as follows:

Product:

$X_1 = C_{1.1} = \{Product\}$

Product component category:

$X_2 = C_{2.1}, C_{2.2}.....C_{2.a} = \{structure, silhouette, color, ...\}$

Product components:
$X_3 = \{C_{3.1}, C_{3.2}.....C_{3.b}\} = \{{C^1}_{2.1}, {C^2}_{2.1}, ...{C^q}_{2.1}, {C^1}_{2.2}, {C^2}_{2.2}, ...{C^n}_{2.2}, ...{C^n}_{2.2}, ..., {C^1}_{2.a}, {C^2}_{2.a}, ...{C^r}_{2.a}\} = \{loosestructure, tight\ structure, ..., Casual\ structure, H\ type, Ttype, ..., A\text{-}type, ..., drop\ shoulder, ..., plunging\ sleeves\}$,
where $b{=}q{+}n{+}...{+}r$.

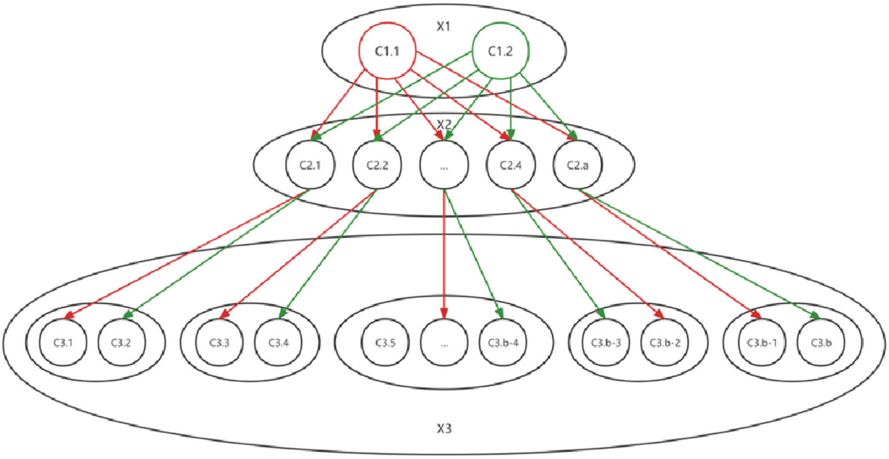


Fig. 2. The relationship between product, product component category, product components, and product family structure.

As is shown in Fig. 2, the hierarchical relationship of X_1, X_2, and X_3 reflects the product family structure of a certain product. The specific product X1 of the product is formed by combining the product component X_3 under each product component category of X_2. According to such affiliation and the mapping, the relationship network of the product family is formed. Figure 4 lists the product family structure of two products ($C_{1.1}$ and $C_{1.2}$). The actual product family relationship network is more complex and extensive than the figure.

Establishment of the "form-Style" Semantic Model Using Fuzzy Cognitive Map. Every component of the product has a corresponding perceptual picture expression. For instance, the "Bump Collar" reflects the "Formal Style" while the "Chest Patch Pocket" represents the "Casual Style" in terms of clothing style. The coordinate relationship between product components in the Perceptual-Cognitive Descriptive Space (PCDS) expresses the perceptual image of the Product Components Matrix (PCM). The fuzzy cognitive map can be used to replicate it, as seen in Fig. 3. It is possible to mimic as a semantic model the relationship between the product's constituent parts and the Perceptual-Cognitive Descriptive Space (PCDS) using the fuzzy cognitive map. A team of designers can execute this technique. The designer's expertise can be fully extracted using this approach.

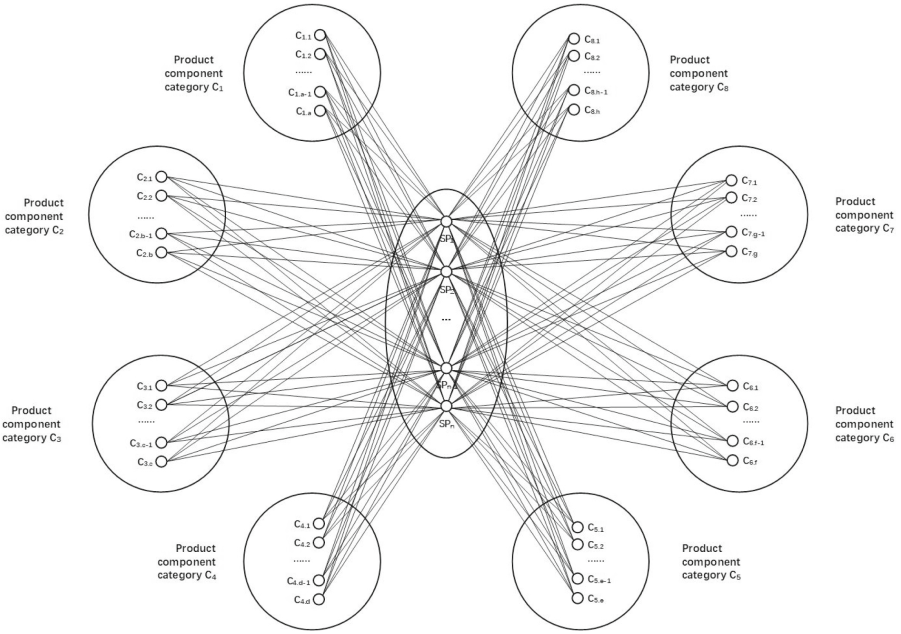

Product component category C1
C1.1
C1.2
......
C1.a-1
C1.a
Product component category C8
C8.1
C8.2
......
C8.h-1
C8.h
Product component category C2
C2.1
C2.2
......
C2.b-1
C2.b
Product component category C7
C7.1
C7.2
......
C7.g-1
C7.g
SP1
SP2
...
SPn-1
SPn
Product component category C3
C3.1
C3.2
......
C3.c-1
C3.c
Product component category C6
C6.1
C6.2
......
C6.f-1
C6.f
Product component category C4
C4.1
C4.2
......
C4.d-1
C4.d
Product component category C5
C5.1
C5.2
......
C5.e-1
C5.e

Fig. 3. An example of the fuzzy cognitive map of the "form-style" relationship.

Quantification of the "form-Style" Semantic Model Using a Fuzzy Transformer Method. It is required to quantify the "form-style" link and then model it because the data gathered for this study are semantic assessment data. As seen in Fig. 4, we suggest a fuzzy transformer strategy to achieve this goal. First, fuzzy logic is utilized to quantify the findings of the subjective judgment as a traditional way for fuzzy modeling. A mean-based transformer approach is suggested to process the quantified data in order to cluster the data and produce the desired outcomes.

For example, using the pair "*simple-complex*", the perceptual scale is $\{extremelysimple\text{-}very\,simple\text{-}relativelysimple\text{-}moderate\text{-}relativelycom - plex\text{-}verycomplex\text{-}extremelycomplex\}$, and the evaluation semantics can be quantified by the triangular fuzzy numbers (TFNs) $\{(0,0,1),(0,1,3),(1,3,5),(3,5,7),(5,7,9),(7,9,10),(9,10,10)\}$. Corresponding to the set of semantic evaluation results $\{F_1,F_2,F_3,F_4,F_5,F_6,F_7\}$, it can be symbolized as $TFN(Fc)(c = 1,2,3...,7)$, $i.e., TFN(F1) = (0,0,1), TFN(F2) = (0,1,3), ...,$ $TFN(F7) = (9,10,10)$ respectively.

$TFN(Fc)(c{=}1,2,3...,7)$ are denoted as $TFN(Fc1), TFN(Fc2)$, and $TFN(Fc3)$ for the lower, maximum possible, and upper limits, respectively, i.e., $TFN(Fc1){=}0, TFN(Fc2){=}0$, and $TFN(Fc3){=}1$ for $TFN(F1){=}(0,0,1)$. Then, a number of fashion industry professionals are requested to conduct private, in-depth interviews. They are asked to assess findings pertaining to l dimensions of the Perceptual-Cognitive Descriptive Space (PCDS) in relation to b product

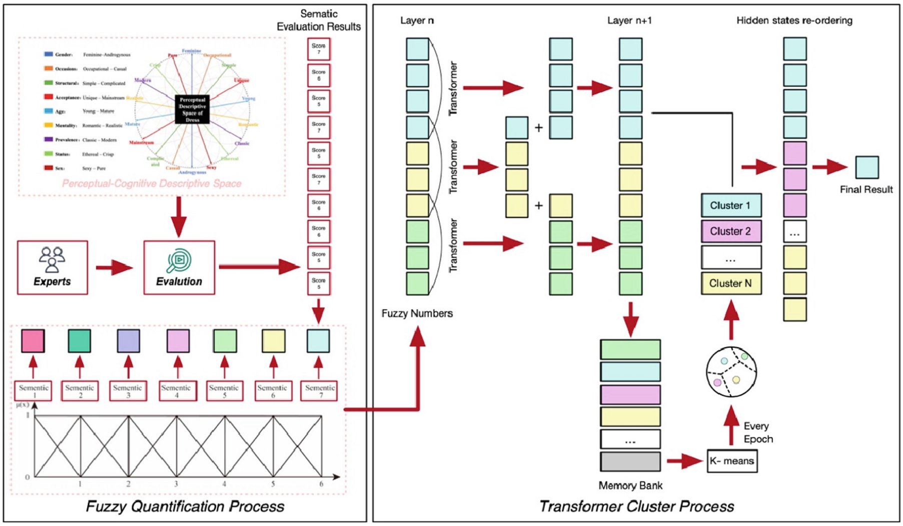


Fig. 4. The working principle of the proposed Fuzzy Transformer method.

components of the Product Components Matrix (PCM). Fuzzy logic will first be used to qualify all of the evaluation data for the data process. After clustering the data using the k-means-based transformer technique, the centroids of each product component are then determined on the various dimensions of the proposed Perceptual-Cognitive Descriptive Space (PCDS). After that, the quantified ”form-style” model is acquired. This method outperforms existing ones when it comes to ”form-style” semantic modelling.

3 Case Study and Related Experiments

Clothing style is a significant representative of product design due to its complicated fashion trend, rich form, and powerful personalisation. Clothing style design requires specific design expertise as well as a thorough design process. Clothing style design's cognitive mechanism is a significant representative object in the research on perceptual cognitive mechanism. In order to illustrate the concept of the proposed system, this study uses the perceived cognitive mechanism of clothing as an example.

3.1 Experiment I: Computational Dress Perceptual Cognition Modelling

Experiment I aims to create a Dress Design Knowledge Base (DDKB), which describes the connections between all Dress Perceptual-Cognitive Descriptive Space (D-PCDS) dimensions and all additional Dress Components Matrix (DCM) components.

Experiment I consists of three steps : (1) the development of the Dress Components Matrix (DCM), (2) the establishment of the Dress Perceptual-Cognitive Descriptive Space (D-PCDS), and (3) the Computational modelling of the Dress Design Knowledge Base (DDKB).

105 panelists (experiment fashion designers) are used as the experiment's expert panel for this aim. To make sure that they had the necessary information for this experiment, all of the panelists were trained for 6 h per week for 2 weeks, covering topics including descriptive terms, fashion style perceptual words, dress component element categories, and dress components for each dress component category. The methods used for subjective evaluation and the experimental protocols were both well-researched.

Step (1): the development of the Dress Components Matrix (DCM)

- **Collection:** The 105 panelists that are involved must do a brainstorming exercise to come up with as many types of outfit components as they can. Each of them creates a list of categories for dress parts based on their individual design expertise, such as the silhouette, sleeve style, dress length, etc.
- **Screening:** All panelists are given the opportunity to screen all of the mentioned dress component categories. Following that, a "round table" debate is undertaken to decide which dress component categories are most appropriate. The screening is based on two ideas: (a) the essential component of the dress is decomposed into a number of independent and variable dress component categories, and each independent and variable dress component category should be selected to be complete and not repeated; (b) recombination of individual and variable essential dress component categories is able to result in a new dress form.
- **Selection:** All panelists are required to list variable dress components in various dress component categories. Create and pick the various dress components of each dress component category. The stated dress components are then chosen by all panelists after a screening procedure. Similar to that, screening is based on two ideas: (a) the essential components of the dress component categories are decomposed into a number of independent and variable dress components, and each independent and variable dress component should be selected to be complete and not repeated; (b) recombination of individual and variable dress components of each dress component category is able to result in a new dress form. The final Dress Components Matrix (DCM) is shown in Fig. 5.

Step (2): the establishment of the Dress Perceptual-Cognitive Descriptive Space (D-PCDS)

In this phase, an additional 100 female consumers between the ages of 16 and 45 were invited to join the panel in addition to the 105 panelists from the previous phase. The purpose of this step is to use questionnaires to pick the perceptual-cognitive descriptions of the various outfit components.

- **Collection:** The panelists were given instructions regarding the aim of the experiment, which was to as fully as possible list the descriptive terms used in defining the both the fashion-oriented as well as the emotion-oriented dress needs. Each trained panelist is expected to come up with a variety of categories based on his personal experience and knowledge of the perceptual-cognitive descriptive phrases used to describe apparel. In the end, 178 words were produced.
- **Screening:** The group of experts held a held a group discussion in order to eliminate overly-emotional words. The key idea for this screening was to eliminate dress perceptual-cognitive descriptive words with similar meaning. For example, ‘contracted’, or ‘reduced’. After that, all words were paired. The assessment team matched and reserved 69 pair of words
- **Adjusting:** 100 female consumers were provided the aforementioned pairings in exchange for questionnaires. Both paper-based and internet surveys were employed. These 100 ladies were required to choose at least 28 specified word pairings that they felt were pertinent. The designers then selected those word pairs that appeared more frequently than 60 percent of the time in those questionnaires. Finally, these 105 designers in the panel selected 9 of these couples. Thus, Dress Perceptual-Cognitive Descriptive Space (D-PCDS) establishment is finished (Fig. 5).

Step (3): the computational modelling of the Dress Design Knowledge Base (DDKB)
105 designer panelists are taking part in this step. To begin, each participant was instructed to subjectively assess each component of the Dress Components Matrix (DCM) in the Dress Perceptual-Cognitive Descriptive Space (D-PCDS). On a scale of 1 to 7, each panelist was asked to offer the best score they could, using the sensory evaluation technology as a guide. When evaluating the picture functions, a score of 1 would be provided for the least functionality, and a score of 7 would be given for the most functionality. Similar to how 1 would be the lowest value and 7 would be the greatest for the pair of perceptual words.

The computational relationship between the Dress Perceptual-Cognitive Descriptive Space (D-PCDS) and the Dress Components Matrix (DCM) is then constructed using the suggested fuzzy transformer approach, and the desired Dress Design Knowledge Base (DDKB) is developed (Fig. 5).

3.2 Exploring the Application of the Computational Dress Perceptual Cognition Model: Development of a Personalized Dress Design System

A customized dress design system can be created using the computational dress perceptual cognition model created in this study, as shown in Fig. 5. This system can perform the duties of a real designer because it is a “intelligent designer” system. The system is meant to intelligently create tailored clothing styles based on user needs.

The system operates on the “Design - Display - Evaluation - Adjustment” cycle as a general rule. The user will repeat the cycle until they receive a final outcome that they are happy with. The computational dress perceptual cognition model, which is part of the fashion design knowledge base, supports the proposed system's operation.

Customers will initially submit their design requirements based on the Dress Perceptual-Cognitive Descriptive Space (D-PCDS) when the system is operational. The information will be compared to the perceptual cognition data from the Dress Design Knowledge Base (DDKB)'s Dress Components Matrix (DCM). Then, the most comparable dress elements from each category will be chosen and arranged together as the final design. The user will then be able to assess the result when the system has presented it. The system will stop functioning if the user is satisfied with the solution. If the new user is not pleased with the design solution, he or she must specify which part of the clothing is bothering them. The system should be able to tell which category the unsatisfactory clothing component belongs to, then change another garment component in the order of decreasing resemblance. Up until a good outcome is attained, the "Design - Display - Evaluation - Adjustment" cycle will be repeatedly carried out. It is clear that the system is a dynamic, interactive system for individualized design. The proposed system may realize any interaction that might occur between actual designers and consumers. The system can be used as a DIY design system for customers to fully meet their unique needs or as a design assistance system to help designers with less experience produce accurate and efficient designs.

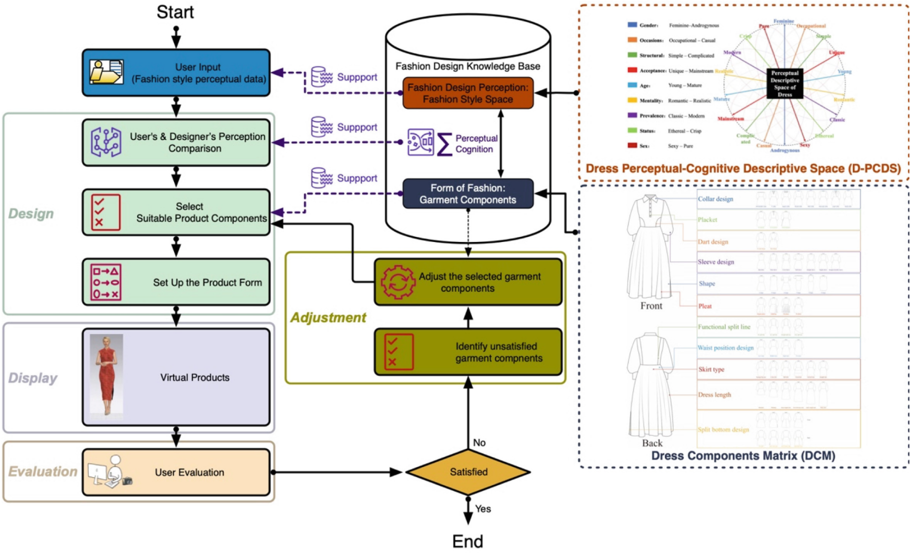
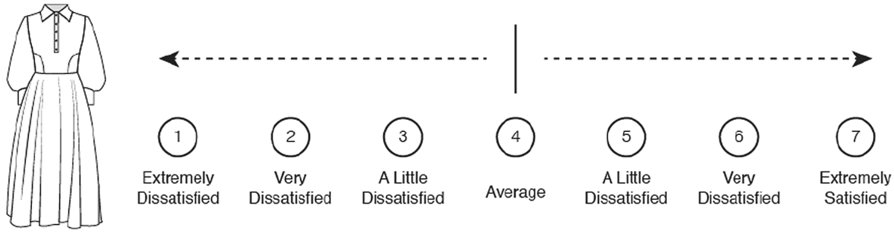


Fig. 5. Working flowchart of the proposed personalized garment style design system based on the proposed perceptual cognition computational model (CM-PCGSD).

3.3 System Evaluation

Using a group of 60 female users between the ages of 16 and 45, the results were examined to see if the proposed approach actually produces the desired results. Firstly They were all instructed to begin by evaluating the system-designed clothing subjectively. The dress pair of perceptual terms received the highest scores on a scale from 1 to 7, with 1 denoting "Extremely dissatisfied" and 7 denoting "Extremely satisfied," respectively. According to Fig. 6, the scores of 1 and 7 denote the highest and lowest levels of function for the evaluation of picture functions. The outcomes are displayed in Fig. 6. It is challenging to assess the reliability and stability of the 60 women who took part in this experiment because opinions regarding their satisfaction are considerably varying. So, to assess the data's stability, we utilized a variance test. The variation of the 60 involved users' satisfaction evaluation data, which covers 60 items, is 0.43, making it appear that the results are mostly steady. Based on this, we may conclude that the effectiveness of the experimental results has been verified.



User	Score	User	Score	User	Score	User	Score	User	Score
User 1	6	User 11	7	User 21	6	User 31	5	User 41	7
User 2	7	User 12	6	User 22	6	User 32	7	User 42	6
User 3	6	User 13	6	User 23	6	User 33	7	User 43	6
User 4	5	User 14	5	User 24	5	User 34	5	User 44	5
User 5	5	User 15	6	User 25	6	User 35	6	User 45	7
User 6	6	User 16	7	User 26	6	User 36	5	User 46	7
User 7	7	User 17	6	User 27	7	User 37	6	User 47	6
User 8	6	User 18	6	User 28	6	User 38	6	User 48	6
User 9	6	User 19	5	User 29	6	User 39	5	User 49	6
User 10	5	User 20	6	User 30	6	User 40	6	User 50	6

Fig. 6. A 7-point evaluation scale for dress design.

4 Conclusion

The goal of this effort is to broaden the use of the Metaverse to the realm of design. The development of an interactive knowledge-based design system for the Metaverse. The main difficulties in developing the system have been examined and overcome. The professional expertise of the designer is reproduced as a variety of computational perceptual cognition models. Ingeniously, a fuzzy transformer technique is created to assist this procedure. To support the system and ensure user-system-Metaverse interactions as well as virtual and physical product interactions, these models are kept in a design knowledge base. As a result, the proposed Metaverse design system's structural design has been abstracted from the general design process of Design-Display-Evaluation-Adjustment. The connection between the design knowledge base and the system can be realized using this procedure. It is also possible to improve user-system-Metaverse interactions. The effectiveness of the suggested system has been confirmed based on a genuine application scenario for using it in customized outfit creation. This paper offers a conceptual framework for the interactive Metaverse design system based on computer modeling of human perceptual cognition. The proposed system significantly broadens the Metaverse's scope of use and has a variety of product design system development applications. The connection between the system and the smart factory in the real world, as well as the modeling of product technical parameter modeling and its application to the proposed system, are future research directions.

Acknowledgments. This research was funded by the National Natural Science Foundation of China (Grant Number: 61906129), the China Association for Science and Technology Youth Support Talent Project (Grant Number: 2021-298), and the Hong Kong Polytechnic University GBA Startup Postdoc Programme 2022 (Grant Number: SP-22-13).

Author contributions. First Author and Second Author contribute equally to this work.

References

1. Cheng, R., Wu, N., Chen, S., Han, B.: Will metaverse be nextg internet? vision, hype, and reality. IEEE Network, pp. 1–9 (2022). https://doi.org/10.1109/MNET.117.2200055
2. Aloqaily, M., Bouachir, O., Karray, F., Ridhawi, I.A., Saddik, A.E.: Integrating digital twin and advanced intelligent technologies to realize the metaverse. IEEE Consumer Electronics Magazine, pp. 1–8 (2022). https://doi.org/10.1109/MCE.2022.3212570
3. Han, B., Pathak, P., Chen, S., Yu, L.F.C.: CoMIC: a collaborative mobile immersive computing infrastructure for conducting multi-user XR research. IEEE Network, pp. 1–9 (2022). https://doi.org/10.1109/MNET.126.2200385
4. Qin, H.X., Wang, Y., Hui, P.: Identity, Crimes, and Law Enforcement in the Metaverse. arXiv preprint arXiv:2210.06134 (2022)
5. Wang, F.Y., Qin, R., Wang, X., Hu, B.: MetaSocieties in metaverse: metaeconomics and metamanagement for metaenterprises and metacities. IEEE Trans. Comput. Social Syst. **9**(1), 2–7 (2022). https://doi.org/10.1109/TCSS.2022.3145165

6. Wang, Y., Chardonnet, J.R., Merienne, F., Ovtcharova, J.: Using fuzzy logic to involve individual differences for predicting cybersickness during VR navigation. In: 2021 IEEE Virtual Reality and 3D User Interfaces (VR), pp. 373–381 (2021)
7. Yang, W., et al.: Semantic communications for future internet: fundamentals, applications, and challenges. IEEE Communications Surveys & Tutorials, p. 1 (2022). https://doi.org/10.1109/COMST.2022.3223224
8. Wang, Y., Lee, L.H., Braud, T., Hui, P.: Re-shaping post-covid-19 teaching and learning: a blueprint of virtual-physical blended classrooms in the metaverse era. arXiv preprint arXiv:2203.09228 (2022)
9. Musamih, A. et al.: Metaverse in healthcare: applications, challenges, and future directions. IEEE Consumer Electronics Magazine, pp. 1–13 (2022). https://doi.org/10.1109/MCE.2022.3223522
10. Zhang, X., Wang, J., Cheng, N., Xiao, J.: MetaSID: singer identification with domain adaptation for metaverse. In: 2022 International Joint Conference on Neural Networks (IJCNN), 18–23 July 2022, pp. 1–7(2022). https://doi.org/10.1109/IJCNN55064.2022.9892793
11. Fettke, P., Houy, C., Loos, P.: On the relevance of design knowledge for design-oriented business and information systems engineering. Business Inform. Syst. Eng. **2**(6), 347–358 (2010)
12. Leaf-nosed bat. Encyclopædia Britannica, ed: Encyclopædia Britannica Online (2009)
13. Hong, Y., Zeng, X.Y., Wang, Y.Y., Bruniaux, P., Chen, Y.: CBCRS: an open case-based color recommendation system. Knowl.-Based Syst. **141**, 113–128 (2018). https://doi.org/10.1016/j.knosys.2017.11.014
14. Hong, Y., Zeng, X.Y., Bruniaux, P., Liu, K.X.: Interactive virtual try-on based three-dimensional garment block design for disabled people of scoliosis type. Text. Res. J. **87**(10), 1261–1274 (2017). https://doi.org/10.1177/0040517516651105
15. Xue, L., Jin, Z.Y., Yan, H., Pan, Z.J.: Development of novel fashion design knowledge base by integrating conflict rule processing mechanism and its application in personalized fashion recommendations. Textile Research Journal. https://doi.org/10.1177/00405175221129868

Patient Self-reports for Explainable Machine Learning Predictions of Risks to Psychotherapy Outcomes

Hans Jacob Westbye[1,2](✉), Andrew A. McAleavey[1,3,4], and Christian Moltu[1,3]

[1] District General Hospital of Førde, Førde, Norway
hans.jacob.westbye@helse-forde.no
[2] Faculty of Medicine, The University of Bergen, Bergen, Norway
[3] Western Norway University of Applied Sciences, Bergen, Norway
[4] Weill Cornell Medicine, New York, USA

Abstract. Prioritizing the right patients and providing personalized treatment in a timely manner is crucial to improve access to healthcare. In psychotherapy, at least 1 in 3 patients drop out of treatment, with therapeutic alliance among the common predictors. Recommendations to safeguard retention include strengthening the patient-therapist bond through developing shared goals and checking in on progress and treatment path. Using a sample of 11,095 mental health patients from the USA, we used machine learning to develop a clinical support tool for treatment personalization. A gradient-boosted decision tree was trained on patient-reported data to establish global and individual predictions/predictors for early treatment dropout, treatment length, and symptom outcomes conditional on different treatment lengths in out-of-sample patients. The models demonstrated marginal to moderate improvements in performance versus baseline predictions. The resulting decision support tool could assist in the collaborative selection of treatment goals, appropriate treatment intensity, and optimal allocation of resources. Results are discussed in the context of explainable AI emphasizing interpretability in a clinical context.

Keywords: Artificial Intelligence · Machine Learning · eXplainable AI · Psychotherapy · Routine Outcome Monitoring · Outcome Prediction

1 Introduction

Psychological interventions necessitate a high degree of personalization tailored to individual patients' unique characteristics and needs [1]. Therapists most often rely on their clinical experience to adapt their approach. This adaptability, however, results in substantial variation in treatment approaches and outcomes, with therapists often being unaware of the consequences of this variation for the individual patient. For example, therapists may underestimate patients' negative experiences and fail to anticipate adverse outcomes, such as dropout from treatment programs [7, 12, 18]. This is unwanted on both an individual and systemic level, calling for the optimization of treatment personalization processes.

R. Akerkar (Ed.): SAIDD 2023, CCIS 1810, pp. 80–105, 2024.
https://doi.org/10.1007/978-3-031-53770-7_6

To facilitate treatment individualization, mental health systems have increasingly turned to routine measurements obtained through self-report questionnaires, administered throughout treatment processes [9]. These questionnaires enable the tracking of patient progress and have been demonstrated to improve symptomatic outcomes when combined with feedback to therapists [1]. This approach produces vast amounts of patient generated data, providing a valuable resource for developing data-driven decision support tools. These tools can empirically identify patients at risk of negative outcomes, that allow therapists to intervene proactively and adjust their approach.

Machine learning (ML) algorithms present a promising avenue for leveraging the data derived from self-report questionnaires to create predictive models that enhance personalization and effectiveness in psychological treatments [4]. In a comparative study of 45 ML algorithms and ensembles, researchers aimed to predict dropout from cognitive-behavioural therapy before the first session in clinically labelled data. The top-performing ensemble model achieved an AUC (Area Under the Receiver-Operator Characteristic Curve) of 0.6581, indicating modest accuracy [3]. The authors note that this result might not seem very precise, but that the best model significantly outperformed a generalized linear model (GLM). However, tools with limited accuracy may still provide clinical value to therapists when they are used to involve patients collaboratively to modify treatment strategies to mitigate the risk of adverse outcomes. However, it is key that both therapists and patients understand the reasoning behind these predictions to jointly derive benefits from such models. Consequently, the explainability and interpretability of models becomes a crucial component when designing ML decision support tools for clinical use - an emerging field known as eXplainable AI (XAI) [2].

This study describes the development and validation of a ML model using granular data from patient self-reports to predict treatment outcomes in a historical dataset. Further, it examines the potential of ML-based predictive tools to enhance therapists' capacity to personalise treatment strategies, decrease dropout rates, and improve patient outcomes. Results are presented using visualisations to demonstrate possible routes to increase explainability and interpretability.

2 Methods

2.1 Sample

The data used in this study was sourced from the digital measurement-based care provider, Mirah, Inc, as part of an international research project [15]. This anonymized data was collected during routine practice in various outpatient clinics across the United States. The sample consisted of patients receiving treatment between March 15th, 2016, and February 17th, 2022. The data set encompasses a diverse patient group, each seeking help for their unique mental health challenges.

Mirah routinely gathers anonymized data to drive improvements in their software and contribute to ongoing research. As the data was fully anonymized at the patient level, there was no need for written informed consent in this instance.

2.2 Instruments

Norse Feedback (NF) is a clinical feedback system developed at the District General Hospital of Førde, designed to facilitate personalized therapeutic interventions based on clinicians' and patients' needs [14, 16, 17]. The development of NF involved a rigorous process of item generation, testing, and refinement through clinical implementation studies, ensuring its relevance and effectiveness in clinical practice [8].

The NF system used in this study (version 1) comprises a maximum of 88 items (Appendix 1), which load onto 18 dimensions[1]. Patients respond to the items using a seven-point Likert scale (Not at all true for me – True for me), allowing for nuanced feedback on their experiences and symptoms. The measurement provider did not use NF item 55 for the administrations comprising this dataset.

A unique feature of NF is its dynamic and adaptive nature, employing patient-adaptive computer logic to determine which dimensions are relevant to a particular patient. This is achieved by utilizing trigger items for each dimension; based on the patient's responses, certain dimensions may be opened or closed, ensuring the assessment remains focused on the patient's specific needs. This adaptive approach results in some dimensions being absent during certain administrations, as they are deemed less relevant to the patient's current condition.

2.3 Description and Preparation of Data

The initial dataset consisted of $n = 11{,}470$ patients and $k = 2{,}065$ variables with a high degree of missingness. These variables included various process data and measures other than NF that only select patients had completed making them unsuitable for inclusion in further analysis. Following a rigorous process of dataset cleaning and variable selection, the final dataset included $k = 110$ variables (88 NF items, 18 NF dimensions, recent life changes (binary), in-treatment status at the first assessment (binary), age, and gender) for $n = 11{,}663$ patients that were 18 years or older at the time of assessment. For the recent life changes variable, a change in one or more of the multiple-choice alternatives (employment status, ER visit, housing status, medications taken, and relationship status) resulted in the variable being coded as 1. The data was received unlabelled from the data provider and contained no clinical information about patient outcomes.

We applied a practical approach to labelling the outcome variables estimated treatment length and dropout in collaboration with clinical expertise. The dataset was randomly divided by subject IDs into a training dataset and a test dataset where 20% of patients were reserved for the test dataset. Patients missing more than 90% of observations across all variables for the first assessment were excluded resulting in $n = 8850$ in the training dataset and $n = 2245$ in the test dataset. The estimated length of treatment was encoded as a count variable, represented by the number of unique assessments for

[1] Alliance, Attachment, Avoidance, Connectedness, Demoralization, Eating Problems, Emotional Distancing, Hurtful Rumination, Hypervigilance, Perfectionism-Control, Pressure from Negative Affect, Psychosis, Relational Distress, Resilience and Personal Coping, Social Role Functioning, Somatic Anxiety, Substance Use, and Suicide Risk.

each patient. A binary outcome variable, “dropout,” was encoded as 1 when the estimated length of treatment was less than three. Values for each of the NF dimensions were calculated as the average of the corresponding item scores, ignoring missing values. Dimensions with reversed scores (Alliance, Attachment, Connectedness, Resilience and Social Role) were transformed so low scores always represent best outcomes to enhance explain ability and visualizations.

2.4 Machine Learning Algorithm and Missing Data Strategy

While variable selection helped reduce the overall missingness in our dataset, a considerable number of missing observations persisted. To address this issue, we opted for a machine learning (ML) algorithm that can effectively handle missing data. Among the various ML algorithms, Gradient Boosted Decision Trees (GBDT) have shown remarkable capabilities in minimizing errors through the gradient descent algorithm employed in sequential models [6]. One such GBDT algorithm, eXtreme Gradient Boosting (XGBoost), has consistently demonstrated superior performance in numerous problems involving tabular data [5].

In the study by Bennemann et al., GBDT algorithms ranked among the top-performing single models when comparing 45 ML models and ensembles for dropout prediction [3]. The XGBoost algorithm offers several advantages, including consistent performance, high speed, lack of necessity for regularization, interpretability, and efficient handling of high-dimensional data, which reduces the need for feature selection. Additionally, XGBoost incorporates an in-built mechanism for handling missing data, making it particularly suitable for our dataset. Another important feature of GBDTs relevant to our data is the handling of correlated variables. As we use both NF items and the NF dimensions in the dataset there will be multicollinearity between variables. GBDTs are particularly robust to multicollinearity due to the decision tree design [22].

However, the optimal performance of the XGBoost algorithm requires the tuning of numerous hyperparameters, which can be computationally demanding. In order to strike a balance between the model’s performance and computational efficiency, we employed a systematic hyperparameter optimization strategy [21].

To enhance the explainablility of the XGBoost models, we incorporated the SHapley Additive exPlanations (SHAP) method for variable impact analyses in the final models and visualization of the ML model prediction process [11]. SHAP values offer a unified measure of feature importance grounded in cooperative game theory, providing a consistent and locally accurate interpretation of the predictions made by complex ML models. By implementing SHAP values in our analysis, we aimed to provide insights into the variables that significantly contribute to predicting dropout risk and symptomatic outcomes, thereby facilitating interpretability and a deeper understanding of the factors influencing treatment personalization and informing clinical decision-making.

2.5 Development and Validation of Prediction Models

Defining Clinical Objectives. In collaboration with clinical psychologists at the District General Hospital of Førde, we defined four prediction tasks using data from the first patient assessment:

- Risk of dropout
- Length of treatment
- Probability of completing the predicted treatment length
- Outcomes on the NF clinical dimensions, given a completed treatment length

Software. All data analyses were performed using R [19] in the RStudio software for Windows [20]. Individual packages used for data analysis are described in the following sections.

Model Development and Baseline Comparisons. For each prediction task, we established baseline models and trained an ML model. Hyperparameter optimization and cross-validation were performed using the caret package [10], and the final models were trained using the xgboost package [5]. The performance of the ML models was assessed in the out-of-sample test dataset using appropriate evaluation metrics and compared with the baseline model.

Task 1: Dropout Risk. We calculated the baseline probability of dropout using the overall dropout rate for the full training dataset. The baseline model was validated using a Monte Carlo simulation with 1,000 repetitions on the test dataset. An XGBoost model was trained and validated, with performance compared to the baseline model using the area under the receiver operating characteristic curve (AUC), Positive Predictive Value (PPV) and Negative Predictive Value (NPV).

Task 2: Length of Treatment. The training dataset was filtered for patients not labelled as dropouts. Outliers over the 95th percentile on the estimated treatment length variable were removed. The mean estimated treatment length in the resulting training dataset was used as the baseline prediction. An XGBoost model was trained and validated using the root mean squared error (RMSE).

Task 3: Probability of Completing Predicted Treatment Course. Patients were deemed to have completed a treatment course if the estimated treatment length was equal to or surpassed the predicted treatment length minus one. Those with eight or more assessments were considered to have completed the series. Predicted treatment lengths were adjusted to fall within a clinically relevant range, set to five for predictions under five and capped at 12 for predictions over 12. The baseline model employed the probability of treatment completion for patients with three or more assessments from the full training set. For baseline model validation on the test dataset, we utilised a Monte Carlo simulation. We compared the baseline model to the performance of the trained and validated XGBoost model using AUC, PPV and NPV.

Task 4: NF Clinical Dimensions Outcomes. For this task, we excluded NF items from model training due to the high computational demand of a high number of variables. Consequently, after removing the 88 NF items, the training and test datasets contained

22 variables. For the baseline model, we forecasted the average outcome for all patients in the training set undergoing a specific treatment length. For example, to predict the Attachment outcome for a patient with a predicted treatment length of eight sessions, we utilised the mean Attachment outcome for all patients with an estimated treatment length of eight. This resulted in 108 mean predictions for all combinations of predicted treatment length ($i = 6$, data for treatment lengths 8/9 and 10/11 were grouped for adequate training data) and outcome variables (dimensions = 18). We then trained 108 XGBoost models on the training set for each combination of outcome and predicted treatment length. A key aspect of the NF system is its ability to adapt to patient needs, meaning less relevant dimensions for patients may be closed. To account for this, outcome predictions for dimensions with a first assessment score of 1 were set to 1, and those with a score below 1.5 were set to 1.5 for both models. We assessed model performance using RMSE.

Model and Hyperparameter Optimization Strategy. For this project we tuned the following XGBoost hyperparameters for optimal performance: *eta* – step size shrinkage used in the boosting process to reduce feature weight to avoid overfitting (learning rate - more conservative with lower values), *max_depth* – maximum depth of a tree, *min_child_weight* – minimum number of instances needed to be in each node to implement partition (more conservative with higher values), *colsample_bytree* – the subsample ratio of dataset columns to use when constructing each tree (more conservative with lower values), *subsample* – subsample ratio of the training instances for constructing each tree (more conservative with lower values), and *gamma* - minimum loss reduction required to make a further partition on a leaf node of the tree (more conservative with higher values) [23]. For tasks 1 and 3, we optimized XGBoost hyperparameters in the training dataset using 10-fold cross-validation through a grid search, conducted in a stepwise fashion to minimize computational expense. Each step tested all combinations of a set of parameters to identify optimal values. We adjusted the parameters in the following order: (1) eta and max_depth, (2) min_child_weight, (3) colsample_bytree and subsample, (4) gamma, (5) eta, and (6) nrounds. For task 2, we began with parameters resulting from tuning task 1 and manually adjusted for optimal performance. For task 4, we adopted a pragmatic approach, selecting a single set of hyperparameters for all models. We applied scaled weighting to binary outcomes during model training to adjust for class imbalances. The loss functions for optimization were the default ones for the prediction task— error for tasks 1 and 3 and RMSE for tasks 2 and 4.

Model Explainability. We used the shapviz package [13] to calculate SHAP values for the trained models. Both global and local explanations are presented, with global explanations assessing the average impact of variables on predictions, and local explanations gauging the influence of the variable values for a single patient on a particular outcome prediction.

Table 1. Values for hyperparameters provided to grid search.

Parameter	Values
eta	0.01, 0.05, 0.1
max_depth	2, 3, 4, 6
min_child_weight	1, 3, 5, 7
colsample_bytree	0.4, 0.6, 0.8, 1.0
subsample	0.5, 0.75, 1.0
gamma	0, 0.05, 0.1, 0.5, 0.7, 1.0

3 Results

The modelling strategy resulted in two datasets used in the analyses, a summary of the sample characteristics for these datasets is provided in Table 2. Notably, about half of the patients were already in treatment when they responded to their first assessment.

Table 2. Dataset characteristics

	Training	Test
Characteristic	N = 8,850[1]	N = 2,245[1]
Age (median (IQR))	34 (25, 47)	34 (25, 48)
Dropout	4,507 (51%)	1,172 (52%)
Gender		
Female	5,143 (58%)	1,280 (57%)
Male	2,664 (30%)	698 (31%)
Other	21 (0.2%)	2 (<0.1%)
Unknown	1,022 (12%)	265 (12%)
Assessments (mean (SD))	5.0 (8.5)	4.9 (7.6)
In Treatment		
No	3,940 (45%)	1,003 (45%)
Yes	4,071 (46%)	1,031 (46%)
Unknown	839 (9.5%)	211 (9.4%)
Recent changes	3,833 (43%)	390 (41%)
[1] n (%)		

Description of Training Data. The dataset exhibited a high rate of attrition, with 51% of patients in the training data completing fewer than three assessments. The estimated length of treatment within the training dataset demonstrated a widely dispersed distribution, with values ranging from a minimum of 1 to a maximum of 148. The amount of data missing from the training set was substantial. Despite a reduction in overall missing data following data cleaning, missingness persisted especially among certain variables, some of which displayed over 90% missingness. Some of this is due to the adaptive nature of the NF. For instance, the Alliance scale is not asked at the first administration, so only patients already in treatment responded to these items.

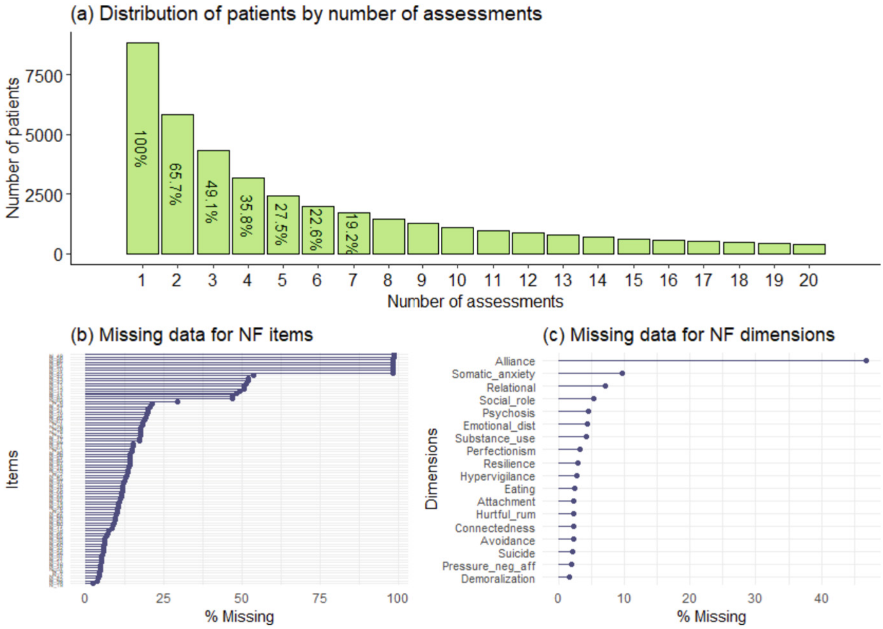


Fig. 1. Training dataset characteristics: (a) Distribution of estimated treatment length for patients, (b) proportion of missing data for Norse Feedback (NF) items (items with more than 90% missingness: N_49, N_56, N_86, N_54, N_84, N_10, N_21, N_25), (c) proportion of missing data for NF dimensions. (Color figure online)

Task 1: Dropout Risk. The predictive accuracy of the XGBoost model in the test dataset was modest (AUC 0.624, PPV 0.612, NPV 0.558), but outperformed the baseline model (AUC 0.505, PPV 0.522, NPV 0.478). The ROC curve for this model is presented in Fig. 4 (a). The three most important variables in the model were Norse Item 12 (N_12) – "I feel that my therapist understands me and understands why I am in treatment now", the Resilience dimension and Norse Item 16 – "People have told me they are worried about my drinking and/or drug use" (N_16). Refer to Appendix 1 for a full description of Norse Items and dimensions.

Task 2: Length of Treatment. After filtering for dropout (n = 4507) and estimated treatment length outliers (*n* = 229; identified at the 95th percentile post-dropout removal), 4114 patients remained within the training dataset for this task. The XGBoost model's predictive accuracy was modest (RMSE = 7.731), only marginally outperforming the baseline model (RMSE = 7.88), which predicted the mean estimated length

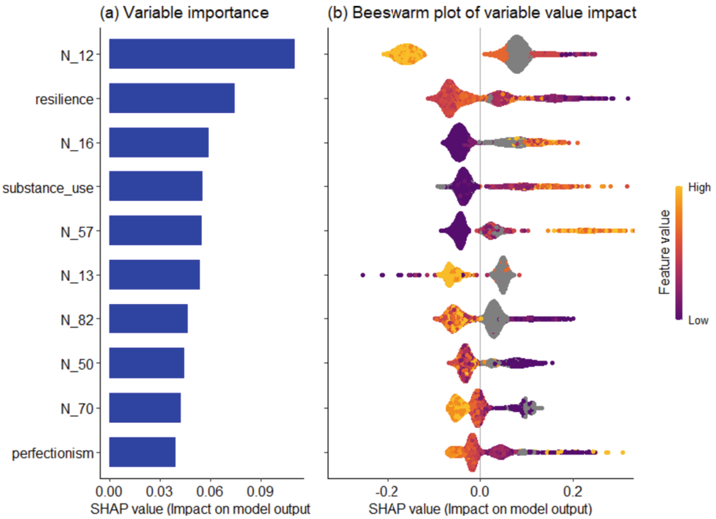


Fig. 2. Task 1 - Dropout prediction: (a) Overall 10 most important variables for predictions and (b) impact on global model of variable values. High values (dot colour) indicate patients' endorsement of problem or item statement, low values represent absence of problems or disagreement with item statement. Missing values are visualised as grey dots. Negative SHAP values indicate decreased risk of dropout, positive SHAP values indicate increased risk prediction.

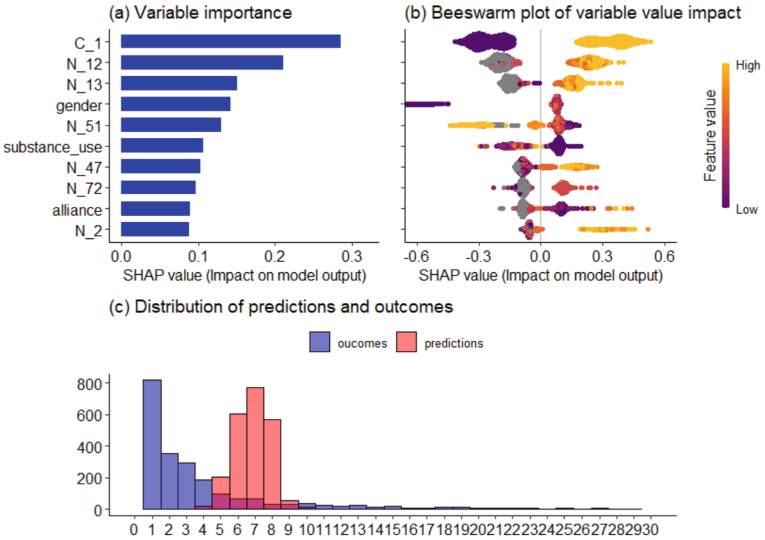


Fig. 3. Task 2 – Treatment length prediction: (a) Overall 10 most important variables predictions and (b) impact on global model of variable values, (c) distribution of predictions and test set outcomes. Only outcomes < 30 assessments are included.

of treatment (6.93 assessments). The three most important variables in the model were recent changes (C_1), Norse item 12 (N_12), and Norse item 13 – "I feel that my therapist accepts me as a person" (N_13). Figure 3 (c) presents the distribution of the XGBoost treatment length predictions alongside the distribution of estimated treatment length outcomes from the test set.

Task 3: Probability of Completing Predicted Treatment Course. Data attrition prevailed after the first two sessions, only 27.5% of patients completed five or more assessments. The probability for completing the predicted treatment series in the training data was 0.164. The predictive accuracy of the XGBoost model in the test dataset was modest (AUC 0.655, PPV 0.238, NPV 0.879), but outperformed the baseline model (AUC 0.5, PPV 0.16, NPV 0.841). The most important variable in the model was the predicted treatment length (xgb_pred). The other variables were considerably less influential.

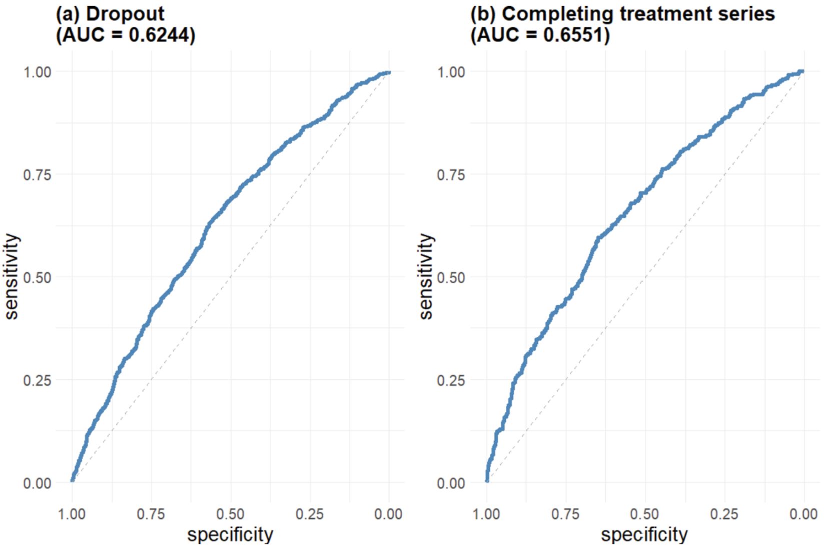


Fig. 4. (a) ROC Curve for XGBoost dropout prediction model, (b) ROC Curve for XGBoost completing treatment series prediction model.

Task 4: Outcomes on NF Clinical Dimensions. Each dimension and treatment length necessitated a unique training dataset for this prediction task, resulting in 108 training datasets (Appendix 2). Table 3 showcases the average treatment outcomes for all permutations of dimensions and estimated treatment lengths used as the baseline predictions. The average RMSE of all 108 XGBoost models was 1.23 (95% CI 1.15–1.31), which significantly outperformed the baseline models' average RMSE of 1.391 (95% CI 1.3–1.49) with a p-value of < 0.05 (Welch Two Sample t-test). Appendix 2 provides detailed results of the validation of predictions in the test set. Overall, most outcomes did not appear to systematically vary with increasing treatment length. Patients who completed different numbers of sessions had similar outcomes on most NF dimensions.

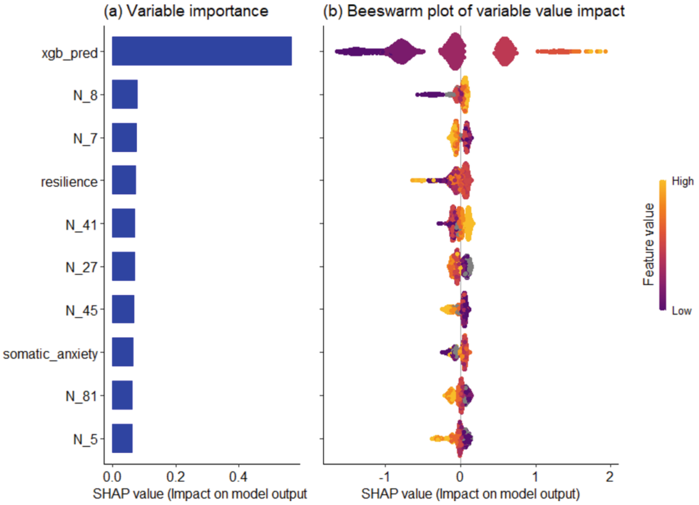


Fig. 5. Task 3 – Probability of completing predicted treatment course (a) Overall 10 most important variable for predictions and (b) impact on global model of variable values

Table 3. Mean outcomes (baseline predictions) for NF dimensions for patients with estimated treatment lengths of 5–12 in the training dataset. Patients with treatment lengths of 8 and 9, and 10 and 11 were binned. Estimated treatments lengths > 12 were binned as 12.

Mean values for NF dimension outcomes						
	Estimated treatment length					
Variable	5	6	7	8	10	12
alliance	1.91	1.91	2.01	1.94	1.94	1.83
attachment	3.28	3.50	3.36	3.40	3.33	3.26
avoidance	3.81	3.71	3.77	3.79	3.77	3.61
connectedness	3.28	3.29	3.25	3.25	3.33	3.37
demoralization	3.42	3.40	3.47	3.41	3.42	3.31
eating	2.73	2.65	2.66	2.52	2.60	2.41
emotional_dist	3.79	3.76	3.81	3.73	3.88	3.69
hurtful_rum	3.90	3.79	3.82	3.84	3.79	3.61
hypervigilance	4.19	3.83	4.11	3.93	3.94	3.64
perfectionism	3.68	3.54	3.58	3.64	3.63	3.50
pressure_neg_aff	3.02	2.94	3.00	2.91	2.98	2.75
psychosis	1.41	1.40	1.39	1.40	1.35	1.26

(*continued*)

Table 3. (*continued*)

Mean values for NF dimension outcomes						
	Estimated treatment length					
relational	3.98	3.92	3.94	4.00	4.26	4.00
resilience	3.38	3.45	3.30	3.38	3.35	3.44
social_role	3.56	3.61	3.47	3.58	3.71	3.56
somatic_anxiety	3.57	3.62	3.59	3.66	4.10	3.61
substance_use	1.58	1.59	1.45	1.51	1.38	1.47

Hyperparameters. The resulting hyperparameters from the grid-search utilizing 10-fold cross validation varied only slightly between the models.

4 Discussion

We have delineated a methodology leveraging patient self-reported data alongside ML techniques to predict psychotherapy outcomes. Our aim was to elucidate how patients' responses influence model predictions, employing methods drawn from XAI. Although the ML models derived from our four designated prediction tasks outperformed baseline predictions, the overall performance remained modest. While Bennemann et al. (2022) examined various ML algorithms and ensembles to predict patient dropout, emphasising maximising predictive performance, our focus has been on establishing methodologies that facilitate clinical implementation of predictive models. However, our findings align with those of Benneman et al. who reported a range in performance, (AUC – Area Under the Receiver-Operator Characteristic Curve) after training 21 single algorithm ML models on 77 variables to predict dropout, from 0.52 to 0.653, and for algorithm ensembles (using various stacking and variable selection methods), from 0,547 to 0,658. With a dropout prediction AUC of 0.624, our model outperformed 12 of the 21 models tested, and 13 of 30 ensembles, achieving this with a more parsimonious dataset and modelling methodology. In contrast to Benneman et al., who required patients to respond to more than 430 items to construct the model variables, our model only require patients to respond to up to 92 items. For most patients the number of items was lower due to the patient-adaptive nature of the data collection. This suggests that the NF assessment alone could suffice for predictive models with reduced patient burden.

Table 4. Hyperparameters used for the final XGBoost models

Prediction task	eta	max_depth	min_child_weight	colsample_bytree	subsample	gamma	nrounds	scale_pos_weight
Task 1: Dropout	0.1	2	7	0.8	1	0.05	296	0.964
Task 2: Session length	0.025	2	7	0.4	0.75	0.1	317	NA
Task 3: Treatment completion	0.025	3	3	0.8	0.5	0	934	5.099
Task 4: Clinical outcomes	0.025	2	7	0.4	0.75	0.1		

To realise successful clinical implementation, perceived benefit from new technology use is essential. Consequently, fostering explanations that facilitate interpretation has been prioritised over optimising prediction performance. For therapists and patients, identifying areas where self-reports, coupled with ML, reveal risk of poor outcomes, can prompt discussions about mutual goals, thereby strengthening the patient-therapist alliance. In our data, the ML models highlighted NF items that concentrate on patient-therapist relations as crucial predictors.

The four prediction tasks outlined in our study could be instrumental in reducing patient attrition from treatment and enhancing patient outcomes. Our results do not provide therapists with definitive answers about patients at risk of poor outcomes but can enlighten therapists on areas needing particular focus to retain the patient in treatment.

Handling missing data was a critical component of our process, given the adaptive nature of the data-collection technology. Traditional approaches to missing data, such as eliminating incomplete data or imputing missing data, were untenable, thereby restricting our ML algorithm options. The Gradient Boosted Decision Tree algorithms are uniquely equipped to handle this issue with their inherent capacity for managing missing and correlated data. Our findings reveal that missing data, a byproduct of patient-adaptive data collection, carry predictive value.

Our study is not without limitations, most notably, the lack of clinical labelling of the dataset necessitated using the information contained in the dataset to label outcomes such as dropout. This knowledge gap regarding reasons for patient dropout may have led to mislabelling patients as dropouts when they, in reality, continued treatment but ceased further assessments. Future data collection will ideally include this information, potentially enhancing predictive performance when training new ML models. Moreover, we have solely used data from the first assessment for predictions. Including data from additional assessments and session changes might improve predictions, providing avenues for future research. Lastly, although we have concentrated extensively on using techniques and visualisations to boost explainability and interpretability, we have not yet obtained end-user feedback. We anticipate that further enhancements in presenting predictive model results can be realised via an iterative process involving end-users.

5 Conclusion

Our study indicates that Machine Learning (ML) models, when applied to self-reported patient data, can assist in predicting clinical outcomes with improved predictive performance over baseline models. Additionally, our findings highlight that a meticulously designed and patient-adaptive data collection method can minimise the required number of item responses, alleviating the burden on patients without compromising the efficacy of predictive tasks. We have also demonstrated how such predictive model results can be visualised in a clinical context implementing the principles of eXplainable AI (XAI). For successful clinical implementation of predictive models, it is imperative to provide end-users with a clear and understandable pathway from input data to recommendations to foster trust and understand ability. XAI provides a strategy to achieve this that will be essential for future work.

This is the first study that use Norse Feedback (NF) patient-reported data for predictive modelling. The implications of this study are noteworthy; the NF assessment, capable of integration into all levels of mental healthcare, could undergo further refinement to incorporate ML predictions, thus supporting therapists in making clinical decisions. However, embedding ML outputs in a clinical scenario will necessitate continued efforts to enhance predictive accuracy and to refine the representation of results, aimed at improving end-user understanding and interpretation.

Appendix 1

Table 5. Norse Feedback items and dimensions

Item	Question	Dimension 1	Dimension 2	Dimension 3
1	It's easy for me to care about other people	Attachment		
2	I spend a lot of energy deciding if situations are safe	Hypervigilance		
3	My body is so tense that it hurts	Hurtful Rumination	Somatic Anxiety	
4	My alcohol and/or drug use interferes with my ability to function	Substance Use		
5	I am valued by my community	Social Role Functioning		
6	I enjoy my job/school	Resilience and personal coping		
7	I constantly think I need to be ready in case something bad happens	Hypervigilance		
8	I go out of my way to avoid certain places or experiences	Avoidance		
9	I feel safe in my own home	Hypervigilance		
10	My emotions help me know what is important/right for me	Avoidance		

(*continued*)

Table 5. (*continued*)

Item	Question	Dimension 1	Dimension 2	Dimension 3
11	I understand how treatment can help me	Alliance		
12	I feel that my therapist understands me and understands why I am in treatment now	Alliance		
13	I feel that my therapist accepts me as a person	Alliance		
14	I understand what I need to do or work on to get better	Alliance		
15	I feel hope that things are going to get better	Resilience and personal coping	Alliance	
16	People have told me they are worried about my drinking and/or drug use	Substance Use		
17	I am so afraid of something (e.g. insects, flying) that I do everything I can to avoid it	Avoidance		
18	I spend excessive time thinking about food and planning my meals	Eating problems		
19	I think it would be better if I were dead	Suicide risk		
20	I worry too much about being careless or sloppy	Perfectionism-Control		
21	My anger frightens me	Pressure from Negative Affect		
22	I am good at relaxing when I need to	Resilience and personal coping	Social Role Functioning	
23	I can sometimes “zone out” during intense or emotional moments	Emotional Distancing		

(*continued*)

Table 5. (*continued*)

Item	Question	Dimension 1	Dimension 2	Dimension 3
24	I constantly feel that I can't handle things in my life	Demoralization		
25	I care too much about what other people think of me	Relational distress		
26	I care so much about doing things right that it gets in the way	Perfectionism-Control		
27	I feel good about the amount of work that I do	Resilience and personal coping	Social Role Functioning	
28	I feel that if I started crying, I wouldn't be able to stop	Perfectionism-Control	Pressure from Negative Affect	
29	I am very afraid that a secret organization is watching me	Psychosis		
30	I fear there is something wrong with my body or physical health	Pressure from Negative Affect	Somatic Anxiety	
31	During stressful events, I sometimes feel like I'm watching the event unfold from outside my body	Emotional Distancing		
32	I have been hurting myself on purpose	Pressure from Negative Affect		
33	I get enough sleep to wake feeling rested	Pressure from Negative Affect	Resilience and personal coping	
34	I try very hard not to feel certain emotions	Avoidance		
35	I am concerned that I'm dependent on alcohol/drugs	Substance Use		

(*continued*)

Table 5. (*continued*)

Item	Question	Dimension 1	Dimension 2	Dimension 3
36	Other people don't seem to be able to understand me anyway, so I have given up trying	Demoralization	Relational distress	
37	I am easily annoyed by other people	Relational distress		
38	I constantly tell myself all the things I do wrong	Hurtful Rumination	Perfectionism-Control	
39	I can't stop worrying, even when I try	Hurtful Rumination		
40	I'm good at letting others know what's important to me	Resilience and personal coping		
41	I am someone who can form strong attachments to others	Attachment		
42	I am feeling depressed	Demoralization	Pressure from Negative Affect	
43	I find it easy to be myself with my friends	Connectedness		
44	I have thoughts of killing myself	Suicide risk		
45	I am generally satisfied with my love life	Connectedness		
46	I am afraid I will lose control when it comes to food	Eating problems		
47	I have difficulties with my stomach and digestion	Eating problems	Somatic Anxiety	
48	I absolutely have to do some things in order to manage my anxiety. (wash hands repeatedly, check locks many times)	Perfectionism-Control		

(*continued*)

Table 5. (*continued*)

Item	Question	Dimension 1	Dimension 2	Dimension 3
49	If I could I would rather stay alone the rest of my life	Psychosis		
50	I find it easy to trust people	Attachment	Connectedness	Hypervigilance
51	I feel restless and uneasy most of the time	Pressure from Negative Affect	Somatic Anxiety	
52	I need to feel in control at all times	Perfectionism-Control		
53	I experience shortness of breath, racing heart, and/or numbness and tingling in my hands or face	Somatic Anxiety		
54	I feel like everything just happens to me, and I have no control over my life	Connectedness		
56	It feels good to get my heart rate up	Resilience and personal coping		
57	My eating habits restrict my ability to be social	Eating problems		
58	I am very worried that other people can hear my thoughts	Psychosis		
59	I think that I need to cut down on my drinking / drug use	Substance Use		
60	Most of the time I feel…. DownExcited	Pressure from Negative Affect		
61	I have given up hope for a better future	Demoralization		
62	I have good friends who really know me	Connectedness		
63	It is very important to control what and how much I eat	Eating problems		

(*continued*)

Table 5. (*continued*)

Item	Question	Dimension 1	Dimension 2	Dimension 3
64	I feel so uncomfortable around others that I often prefer to be alone	Avoidance		
65	I have a lot of conflict in my personal relationships	Relational distress		
66	I'd have many fewer problems in my life if it weren't for other people	Relational distress		
67	I generally like who I am	Resilience and personal coping		
68	It is hard for me to allow others to be in control of some aspects of my life	Perfectionism-Control		
69	I feel alone even when I'm with other people	Connectedness	Relational distress	
70	I am uncomfortable asking other people for help or support, even when I need it	Resilience and personal coping		
71	I would like my therapist to teach fewer/more skills and strategies	Expressed needs in therapy		
72	I would like my therapist to show their personality and humor more / be more formal	Expressed needs in therapy		
73	I would like my therapist to focus more on my feelings / more on my thoughts	Expressed needs in therapy		
74	I would like my therapist to focus less on the relationship between us / more on the relationship	Expressed needs in therapy		

(*continued*)

Table 5. (*continued*)

Item	Question	Dimension 1	Dimension 2	Dimension 3
75	I'm often afraid but don't know why	Somatic Anxiety		
76	I am worthless	Hurtful Rumination		
77	I am scared that I might lose control and kill myself	Suicide risk		
78	I spend a lot of energy trying not to think about things that hurt	Avoidance		
79	I have plans for how to kill myself	Suicide risk		
80	I have people I can turn to if I need support	Connectedness		
81	Relationships cause me a lot of stress	Relational distress		
82	I think that my psychiatric medication helps me	Independent (Not a subscale)		
83	I have concerns about my medication that I would like to discuss	Independent (Not a subscale)		
84	I look after my health through being active	Resilience and personal coping		
85	I need to worry less	Hurtful Rumination		
86	I don't have the body I should have	Eating problems		
87	I matter to the people around me	Social Role Functioning		
88	I feel like I am trapped and don't know what to do	Demoralization	Pressure from Negative Affect	
89	I'm comfortable sharing my emotions when appropriate	Attachment	Resilience and personal coping	

Appendix 2

Table 6. Results for 108 models trained and validated for task 4 – outcome prediction.

Model	nrounds	RMSE_base	RMSE_xgboost
Session_5_alliance	109	0.88490273	1.22635560
Session_6_alliance	122	1.27427599	1.43939047
Session_7_alliance	111	1.00948297	1.38263836
Session_8_alliance	172	0.85979575	1.07047370
Session_10_alliance	188	0.50263435	0.49791166
Session_12_alliance	93	0.25000000	0.25000000
Session_5_attachment	156	1.16458472	1.10580248
Session_6_attachment	207	1.46729070	1.23818947
Session_7_attachment	140	1.16759628	0.99588591
Session_8_attachment	143	1.26645217	1.16962573
Session_10_attachment	195	1.14210023	1.19546417
Session_12_attachment	142	0.01401869	0.06447458
Session_5_avoidance	155	1.24547662	1.02691426
Session_6_avoidance	154	1.73733870	1.46613213
Session_7_avoidance	168	1.51374925	1.31539511
Session_8_avoidance	158	1.56931191	1.48056416
Session_10_avoidance	238	1.75890572	1.34093641
Session_12_avoidance	144	1.38942901	0.47525406
Session_5_connectedness	188	1.16963887	1.22349417
Session_6_connectedness	159	1.45280230	1.33742691
Session_7_connectedness	158	1.22325648	1.12908345
Session_8_connectedness	172	1.29888496	1.12753913
Session_10_connectedness	215	1.22854907	1.22757250
Session_12_connectedness	148	1.62928135	0.98345041
Session_5_demoralization	148	1.49942032	1.34909431
Session_6_demoralization	134	1.65506086	1.40923371
Session_7_demoralization	171	1.58916925	1.42956305
Session_8_demoralization	154	1.57641044	1.43627873
Session_10_demoralization	125	1.57067557	1.57262366
Session_12_demoralization	123	2.09212538	0.79871311

(*continued*)

Table 6. (*continued*)

Model	nrounds	RMSE_base	RMSE_xgboost
Session_5_eating	199	1.13620876	1.09966767
Session_6_eating	146	1.43147652	1.41868659
Session_7_eating	118	1.48684035	1.46692999
Session_8_eating	137	1.42950711	1.42347914
Session_10_eating	115	1.23396355	1.02910855
Session_12_eating	132	0.79444444	0.23697095
Session_5_emotional_dist	150	1.80493449	1.65557510
Session_6_emotional_dist	161	1.94035270	1.71581795
Session_7_emotional_dist	142	2.04121654	1.90072939
Session_8_emotional_dist	182	1.74706029	1.68945820
Session_10_emotional_dist	181	2.05872340	1.93973419
Session_12_emotional_dist	196	2.80751174	1.56455517
Session_5_hurtful_rum	188	1.18385554	1.03755300
Session_6_hurtful_rum	167	1.56790256	1.30349294
Session_7_hurtful_rum	153	1.67793227	1.38121805
Session_8_hurtful_rum	146	1.56493979	1.36515091
Session_10_hurtful_rum	144	1.32385279	1.16126696
Session_12_hurtful_rum	258	2.18952599	0.60770254
Session_5_hypervigilance	165	1.33607959	1.44511182
Session_6_hypervigilance	190	1.74237463	1.59433736
Session_7_hypervigilance	167	1.52602371	1.45252106
Session_8_hypervigilance	159	1.82612850	1.86906110
Session_10_hypervigilance	171	1.81068982	1.76221864
Session_12_hypervigilance	168	0.69147287	0.99951760
Session_5_perfectionism	186	1.31309171	1.18946341
Session_6_perfectionism	155	1.70181921	1.45088513
Session_7_perfectionism	133	1.53226460	1.30476560
Session_8_perfectionism	182	1.50418783	1.39961014
Session_10_perfectionism	158	1.52369628	1.29406311
Session_12_perfectionism	159	2.33819444	1.22298559
Session_5_pressure_neg_aff	130	1.20549209	1.07999931
Session_6_pressure_neg_aff	150	1.35364763	1.28372336
Session_7_pressure_neg_aff	126	1.10312879	1.14186338

(*continued*)

Table 6. (*continued*)

Model	nrounds	RMSE_base	RMSE_xgboost
Session_8_pressure_neg_aff	152	1.08359559	1.14312227
Session_10_pressure_neg_aff	139	0.74650762	0.66413678
Session_12_pressure_neg_aff	176	0.99961598	0.28908825
Session_5_psychosis	147	1.37610037	1.94802328
Session_6_psychosis	95	1.58024890	1.54538092
Session_7_psychosis	138	1.02708451	1.18543778
Session_8_psychosis	80	0.85019869	0.96459862
Session_10_psychosis	128	1.53934271	1.20251818
Session_12_psychosis	195	0.50000000	0.50000000
Session_5_relational	183	1.18693573	1.11133344
Session_6_relational	143	1.42915033	1.30450169
Session_7_relational	214	1.29423373	1.15571328
Session_8_relational	219	1.39328412	1.27289435
Session_10_relational	183	1.39698238	1.36214004
Session_12_relational	122	2.49796748	1.54161549
Session_5_resilience	171	1.19028971	1.12723237
Session_6_resilience	139	1.33915740	1.17299660
Session_7_resilience	133	1.29325005	1.11979592
Session_8_resilience	151	1.31405538	1.17027009
Session_10_resilience	162	1.21347478	1.14034979
Session_12_resilience	183	0.27583774	0.10739814
Session_5_social_role	239	1.46334982	1.20509086
Session_6_social_role	217	1.64219660	1.44072654
Session_7_social_role	227	1.65363272	1.54783089
Session_8_social_role	141	1.46441083	1.22191446
Session_10_social_role	124	1.24078998	1.00802087
Session_12_social_role	138	3.44339623	2.57707787
Session_5_somatic_anxiety	125	1.47805224	1.33258328
Session_6_somatic_anxiety	148	1.92655998	1.93686425
Session_7_somatic_anxiety	155	1.53012036	1.47741274
Session_8_somatic_anxiety	143	1.65119919	1.59082473
Session_10_somatic_anxiety	136	1.68790530	1.39074806

(*continued*)

Table 6. (*continued*)

Model	nrounds	RMSE_base	RMSE_xgboost
Session_12_somatic_anxiety	139	2.38513514	1.15141821
Session_5_substance_use	130	0.82679676	1.71279060
Session_6_substance_use	103	1.26187910	1.51394106
Session_7_substance_use	135	1.28018228	1.27816775
Session_8_substance_use	91	0.60071523	0.91752323
Session_10_substance_use	107	0.75020174	0.31216782
Session_12_substance_use	149	NA	NA
Session_5_suicide	128	1.65701306	1.89081712
Session_6_suicide	97	1.15689694	1.34000627
Session_7_suicide	337	1.32186006	1.37823519
Session_8_suicide	117	0.78848708	0.78736943
Session_10_suicide	112	0.57946713	1.22187845
Session_12_suicide	79	1.35113636	0.09664583

References

1. Barkham, M., De Jong, K., Delgadillo, J., Lutz, W.: Routine outcome monitoring (rom) and feedback: research review and recommendations. Psychother. Res. **33**(7), 841–855 (2023). https://doi.org/10.1080/10503307.2023.2181114
2. Arrieta, A.B., et al.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inform. Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
3. Bennemann, B., et al.: Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms. Br. J. Psychiatry **220**(4), 192–201 (2022). https://doi.org/10.1192/bjp.2022.17
4. Chekroud, A.M., et al.: The promise of machine learning in predicting treatment outcomes in psychiatry. World Psych. **20**(2), 154–170 (2021). https://doi.org/10.1002/wps.20882
5. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 ACM, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785
6. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)
7. Hatfield, D., et al.: Do we know when our clients get worse? An investigation of therapists' ability to detect negative client change. Clin. Psychol. Psychother. **17**(1), 25–32 (2009). https://doi.org/10.1002/cpp.656
8. Hovland, R.T., Moltu, C.: The challenges of making clinical feedback in psychotherapy benefit all users: a qualitative study. Nord. Psychol. **72**(3), 248–262 (2020). https://doi.org/10.1080/19012276.2019.1684348
9. de Jong, K., et al.: Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: a multilevel meta-analysis. Clin. Psychol. Rev. **85**, 102002 (2021). https://doi.org/10.1016/j.cpr.2021.102002

10. Kuhn, Max: Building predictive models in R using the caret package. J. Stat. Softw. **28**, 5, 1–26 (2008)https://doi.org/10.18637/jss.v028.i05
11. Lundberg, S., Lee, S.-I.: A Unified Approach to Interpreting Model Predictions. (2017). https://doi.org/10.48550/ARXIV.1705.07874
12. Macdonald, J., Mellor-Clark, J.: Correcting psychotherapists' blindsidedness: formal feedback as a means of overcoming the natural limitations of therapists: correcting psychotherapists' blindsidedness. Clin. Psychol. Psychother. **22**(3), 249–257 (2015). https://doi.org/10.1002/cpp.1887
13. Mayer, M., Stando, A.: shapviz: SHAP Visualizations. https://cran.r-project.org/web/packages/shapviz/index.html (2023)
14. McAleavey, A.A., et al.: Initial quantitative development of the Norse Feedback system: a novel clinical feedback system for routine mental healthcare. Qual. Life Res. (2021). https://doi.org/10.1007/s11136-021-02825-1
15. Moltu, C.: NORSE: Building bridges between psyche and soma through personalized and dynamic mental health systems (2017–2023). Norwegian Research Council, grant 269097
16. Moltu, C., et al.: This is what I need a clinical feedback system to do for me: a qualitative inquiry into therapists' and patients' perspectives. Psychother. Res. **28**(2), 250–263 (2018). https://doi.org/10.1080/10503307.2016.1189619
17. Nordberg, S.S., et al.: Continuous quality improvement in measure development: lessons from building a novel clinical feedback system. Qual. Life Res. (2021). https://doi.org/10.1007/s11136-021-02768-7
18. Probst, T., et al.: Attitudes of psychotherapists towards their own performance and the role of the social comparison group: the self-assessment bias in psychodynamic, humanistic, systemic, and behavioral therapists. Front. Psychol. **13**, 966947 (2022). https://doi.org/10.3389/fpsyg.2022.966947
19. R Core Team: R: A language and environment for statistical computing. https://www.R-project.org/. Accessed 12 Nov 2022
20. RStudio Team: RStudio: Integrated Development for R., http://www.rstudio.com/. Accessed 12 Nov 2022
21. Zahedi, L. et al.: Search algorithms for automated hyper-parameter tuning (2021). https://doi.org/10.48550/ARXIV.2104.14677
22. Understand your dataset with XGBoost — xgboost 1.7.6 documentation. https://xgboost.readthedocs.io/en/stable/R-package/discoverYourData.html. Accessed 21 Jun 2023
23. XGBoost Parameters — xgboost 2.0.0 documentation. https://xgboost.readthedocs.io/en/stable/parameter.html. Accessed 20 Sept 2023

I-KNOW-FOO: Interlinking and Creating Knowledge Graphs for Near-Zero CO_2 Emission Diets and Sustainable FOOd Production

Görkem Simsek-Senel[1], Hajo Rijgersberg[1], Bengü Öztürk[1], Jeroen Weits[2], and Anna Fensel[2,3](✉)

[1] Food Informatics, Wageningen Food and Biobased Research, Wageningen University and Research, Bornse Weilanden 9, Wageningen 6708 WG, the Netherlands
{gorkem.simsek-senel,hajo.rijgersberg,bengu.ozturk}@wur.nl
[2] Consumption and Healthy Lifestyles, Wageningen University and Research, Hollandseweg 1, Wageningen 6706 KN, the Netherlands
{jeroen.weits,anna.fensel}@wur.nl
[3] Wageningen Data Competence Center, Wageningen University and Research, Droevendaalsesteeg 2, Wageningen 6708 PB, the Netherlands

Abstract. It is already known that the diet of the world's population has a massive impact on climate change. However, how climate change affects the growing conditions of ingredients for different foods and beverages, and emission rates due to, for example, production and logistics are still not known. In this work, different datasets have been explored to study the feasibility of interlinking datasets to automatically generate alternatives for climate change-sensitive food items selection and substitution. A core question to be answered is what the alternatives of the mostly consumed crops in current diets in the Netherlands in case of a climate change can be. The main crop attributes taken into account are nutritional composition and the growing conditions. The growing conditions of three most-consumed crops in the Netherlands have been linked manually to their nutritional composition data and a corresponding knowledge graph is created. This study shows that linking various data semantically promises to generate alternatives automatically.

Keywords: Climate change · food replacement · knowledge graph · reasoning · decision making

1 Introduction

It is already known that the diet of the world's population has a massive impact on climate change [1, 2]. However, still too little attention is being paid to the climate change's impact on the growing conditions of ingredients for different foods and beverages and further, similarly, to emission rates due to, for example, production and logistics. The provenance and climate change impact on various foods are often not clearly known or accessible, both for end consumers as well as for the whole supply chain elements.

R. Akerkar (Ed.): SAIDD 2023, CCIS 1810, pp. 106–119, 2024.
https://doi.org/10.1007/978-3-031-53770-7_7

To give an example, many food options are untrivial and interdependent in terms of sustainability, for example, it may be unclear to consumers that production of mineral water (due to the packaging materials used) may be more damaging to the climate than the production of rice, and further aspects (e.g. logistics and prices) become relevant. As in all information-intensive environments, food producers and consumers continuously face complex decisions on which ingredients or products to choose, in which amounts and how to process them or which alternatives to select for the products that are consumed regularly. To make decisions, the food producers, providers and consumers need access to data about these food items, for example their nutritional value, taste, sustainability characteristics as well as the needed nutrients and logistics information. This information is still scattered, and the quality of the data varies. Meanwhile, data indicating climate change impact of different foods and beverages exists (or can be collected) as well as data on the supply chains. However, these data are still often not easily available and discoverable, and have no explicit representation and connections between them, in a way these can be achieved with semantic technology (ontologies and knowledge graphs). Generally, ontologies are data models representing a set of concepts and their relationships in a domain. Knowledge graphs include domain-specific data points (instances of these concepts) and their specific data and object properties. Knowledge graphs are highly scalable and flexible data structures which allow us to develop a linked data model to illustrate these explicit linkages between the crops, nutritional contents and also growth temperature conditions.

In this work we aim to reach a clear understanding of the diets and how to make them equivalently nutritious, sustainable and approaching near-zero CO_2 emission, as well as also change the diets considering and adapting to climate change characteristics taking into account the growing conditions. The main question that is to be answered with this research is: ***"How can we interlink datasets so that an alternative to the current consumed products can be (automatically) found by taking into account the nutritional composition, growing conditions which will be effected by climate change and sustainability information?"*** and ***"To what extent can that process be automated?"***.

The objective of the research is to identify the relevant data and make them more accessible for discoveries and supporting (automatic) decision making in the food supply chain and for end consumers. Thus, the goal is to develop and generate knowledge graphs, benefiting from semantic technology which helps interlinking scattered information using standardized concepts. With employment of knowledge graphs and using them for interlinking and our I-KNOW-FOO project approach presented in this work, one will be able to create a web-like large-scale data infrastructure and tools to easily explore the domain for everyone (researchers, businesses, policy makers, manufacturers, consumers), as well as to assist in making estimates of CO_2 footprints of various foods and beverages and in adaptation of the diets given the climate change.

The remainder of this document is organized as follows. Section 2 explains the methodology of the work and the starting points with relation to the available data. In Sect. 3, the results are presented, Sect. 4 evaluates the approach, and Sect. 5 describes the conclusions and future work.

2 Data and Methods

For the purpose of the study, data on import, growing conditions and nutritional value of crops was required. The datasets were searched on the web, through data repositories, government websites, academic databases, and open data portals. Preferably, the data on growing conditions and nutritional values would be accessed through existing ontologies as this saves steps in data management and it implies that a shared vocabulary already exists. Alternatively, datasets and databases were converted to linked data and a knowledge graph was constructed manually. The following tables (Tables 1 and 2) include the ontologies and databases that have been collected on both crop growing conditions and food items, respectively. Data inquiry has been conducted based on the following keywords: *Crop ontology*, *Plant ontology*, *Agriculture ontology*, *Growing conditions ontology*, *Nutritional profile ontology*, *Crop traits ontology*, *Crop characteristics ontology*.

Table 1. Ontologies on crop growing conditions.

Name	Description/Contents	Availability	Reference
Agronomy Ontology (AgrO)	Agronomic practices, agronomic techniques, and agronomic variables used in agronomic experiments; 350 variables	Publicly available	Aubert et al., 2017 [3]
AgroPortal	A vocabulary and ontology repository for agronomy; 77 ontologies	Publicly available	Jonquet et al., 2018 [4]
Crop Ontology (CO)	Species-specific phenotypic plant traits; 4,235 traits and 6,151 variables for 31 plant species traits	Publicly available	Matteis et al., 2013 [5]
Crop Planning and Production Process Ontology (C3PO)	A new model to assist diversified crop production	Publicly available	Darnala et al., 2021 [6]
Design and application of an ontology to identify crop areas and improve land use	Facilitates the identification of cultivation areas and improvement of land use	Publicly available	Riaño et al., 2022 [7]
ECOCROP	Data on plant characteristics and crop environmental requirements; more than 2000 plant species	Ontology is not publicly available, only dataset	ECOCROP, 2016 [8]
ISO-FOOD ontology	Food items related to isotopes; 1323 classes related to geolocations	Publicly available	Eftimov et al., 2019 [9]
Plant Experimental Conditions Ontology	Treatments and growth conditions used in plant science experiments; 95 plant taxa	Publicly available	Cooper et al., 2018 [10]
Plant Trait Ontology (TO)	Phenotypic traits in plants	Publicly available	Plant Ontology Consortium, 2002 [11]

Table 2. Ontologies on food items and nutritional composition.

Name	Description/Contents	Availability	Reference
FoodOn	Food product analogue subhierarchy; over 9,600 generic food product categories	Publicly available	Dooley et al., 2018 [12]
FoodKG	Heuristics based on ex-plicit semantics and embeddings. Dietary restrictions, nutritional change. No ontological conceptualization; approximately 67 million triples	Publicly available	Haussmann et al., 2019 [13]
Food Recipe Ingredient Substitution Ontology Design Pattern	Includes the substitution of food on all kinds of different criteria	Publicly available	Ławrynowicz et al., 2022 [14]
Food Item Ontology	Food items with nutritional composition based on the NEVO database; 2346348 statements	Not publicly available	FIO [15]

As the datasets are scattered, the research started initially by making an inventory of the available datasets, ontologies and knowledge graphs on food products, and the impact of climate change on the food products availability. The datasets and databases screened were the SHARP Indicators Database, Food Consumption Impact datasets (Optimeal-Blonk Sustainability Datasets), RIVM Sustainability dataset, the Pizza dataset, the Eaternity Database, Data Explorer: Environmental Impacts of Food, Dataset on potential environmental impacts of water deprivation and land use for food consumption in France and Tunisia. Food environmental impact UK database (by Clark et. al.), and the World Food LCA Database. Among those, a few are publicly available [16–24].

As the databases and datasets are from all over the world, the food products vary from one database to another and it is not straightforward to map or relate them. Additionally, the existing food databases provide information about impact of food consumption on sustainability and they do not have a direct link to changing climate conditions which is required to determine alternatives for the original products that are currently part of the diets.

We then defined a use case that focuses on the most-imported crops in the Netherlands to connect the consumption to the changing climate. For most-imported crops, their important nutritional values are determined and alternative crops are found (e.g., for the case when the Netherlands may run out of these crops in a changing climate over years). For this goal (to determine the most-imported crops) we use the FAOSTAT

Database [25]. The crop information was manually interlinked to growing conditions. The most useful information was considered to be found in the ECOlogical CROP Database (ECOCROP) [8], as it contains information on the growing conditions of more than 3000 crops. However, unfortunately, it was not represented as linked data, so we had to uplift it to this format.

3 Results

The use case focuses on the most-imported crops in the Netherlands. Our aim was to determine the most imported crops to the Netherlands in order to evaluate their important nutritional characteristics and to find alternative crops to these crops for the Netherlands, if these crops become unavailable (such as due to climate change). Top 10 commodities that were imported to the Netherlands within the last 5 years (2016–2020) were screened using the TRADE Datasets for Crops and livestock products in the FAOSTAT database. Moreover, commodities supplied to the Netherlands were assessed using Food Balance Datasets in terms of Domestic Food Supply Quantity (1000 t/yr) and Food Supply Quantity (kg/capita/year). These commodities are listed in terms of their import quantity, import values, supply quantities, in descending order (see Table 3).

Table 3. Comparison of imported food groups versus food supply data (from 2016–2020).

Imported Quantity (t)	Imported Value (1000 US$)	DFSQ	FSQ
Maize	Cocoa Beans	Milk - Excluding Butter	Milk - Excluding Butter
Soybeans	Palm oil	Sugar beet	Potatoes and products
Wheat	Soybeans	Wheat and products	Wheat and products
Rapeseed	Wine	Maize and products	Vegetables, other
Barley	Chocolate products	Potatoes and products	Beer
Potatoes	Maize	Soybeans	Sugar (raw equivalent)
	Cheese	Barley and products	Apples and products

DFSQ: Domestic Food Supply Quantity (1000 t/yr); FSQ: Food Supply Quantity (kg/capita/yr)

We focused on three main commodities that are imported in high quantities and supplied to the Dutch population, and identified *soybean, wheat* and *potato* as the mostly imported and consumed food products. The next step was to find nutritionally similar alternative crops using the NEVO Dutch Food Composition Database. We have also searched for growing conditions of original and alternative crops, and developed knowledge graphs to link these data and reuse parts of the existing knowledge graphs. The alternatives are generated by manually processing the intersection of different result sets of queries on the knowledge graphs (either manually or automatically) for nutritional equivalent (or nutritionally better) food items and crops that are more climate-resistant. In the following parts of the section, we will describe the resulting ontologies and knowledge graphs and the querying in our approach.

3.1 I-KNOW-FOO Ontologies and Knowledge Graphs

To be able to answer basic queries for our problem setting, we have prepared the data as follows, applying manual and automated uplifting and extension to knowledge graphs and ontologies.

Manually Generated Alternatives. To find alternatives to the three crops, we have focussed on parameters of climate resilience, nutrient-rich comparable crops and food products that have been screened using knowledge rules provided by a dietary expert using the NEVO database. These possible alternative crops have then been evaluated in terms of their resistance to temperature increases in a changing climate using crop growth temperatures from the ECOCROP database.

Generating an ECOCROP Ontology. The ECOCROP database is transformed into a knowledge graph manually. First, the dataset has been cleaned. The measurementType 'optimalGrowthTemperature' has been subdivided into *maxGrowthCelsiusTemperature* and *minGrowthCelsiusTemperature* to distinguish between the two as well as add a unit into the predicate. The triples consist of the *occurenceID* as subject, *measurementType* as predicate and *measurementValue* as object. These have been transformed using the OntoText Refine tool and have been loaded into an RDF repository in RDF4J. OntoText Refine is a software tool that supports the transformation of string data into knowledge graphs [26].

ECOCROP Extension and Interlinking to FIO and FoodOn. We have extended ECOCROP manually by adding triples linking some of the *occurenceID*s in ECOCROP to the IDs of crops in FoodOn (including NCBITaxon [27]) and food items in FIO (Food Item Ontology [15]), based on the RIVM NEVO IDs. In FoodOn, we have chosen for instances of the organism class, because these represent the plants rather than the different foods that may originate from these plants. The plants namely are grown under (climate-changing or not) temperatures, not so much the foods. The relations used for linking the concepts are the *skos:closeMatch* and the *owl:sameAs* relations.

The open access ECOCROP ontology and knowledge graphs created in our project are available at: https://git.wur.nl/FoodInformatics/i-know-foo.git.

3.2 Querying the Knowledge Graphs

Subsequently, we have loaded the triples in the triple repository, where the information can be queried using SPARQL. In the future, this could be done by an automated tool. The query that we have formulated searches for crops that are more resilient to a warmer climate, being candidates to replace the current crop. So far in this exercise, we have only focused on the maximum growing temperature being one of the important factors in climate change on crop growth [28]. In our examples, the maximum optimal growing temperatures are 33 °C for soybean, 23 °C for wheat and 25 °C for potato. Combining this information with nutritional values information, still leaves multiple but often restricted options for food alternatives with similar nutrition characteristics. For example, for potatoes, possible alternatives are beans white/brown dried, peas green dried, chestnuts raw, tapioca, cassava raw, taro raw, yam raw, tannia raw, beans black eyed dried, peas split yellow/green dried, tamarind, flour cassava. The alternatives are found by intersecting the different result sets, i.e., the climate-resilient crops from ECOCROP and the alternatives as defined by nutritionists.

To obtain the solutions, queries have been written to find alternatives based on growing temperature and these alternatives have been superimposed onto the nutritional results from NEVO. This identifies the alternatives that are more climate resilient as well as nutritional equivalent. For each crop, a SPARQL query can be written for finding alternatives when altering the optimal maximum growing temperature. For example, for wheat, the maximum optimal growing temperature is 23 °C. The query will therefore be:

```
      <http://example.com/base/test_jeroen/>
    PREFIX xsd: <http://www.w3.org/2001/XMLSchema>
    SELECT ?a ?b ?c
    WHERE    {?a    <http://example.com/base/http://www.w3.org/2000/01/rdf-
  schema#label> ?b.
  ?a   <http://example.com/base/http://eol.org/schema/terms/maxGrowthCelsi-
  usTemperature> ?c
  FILTER(?c > "23"^^<http://www.w3.org/2001/XMLSchema#integer>)
     }
```

This query returned 1,790 crops with a maximum optimal growth temperature greater than 23. These alternatives were then superimposed onto the nutritional alternatives for wheat from NEVO, resulting in the four alternatives as listed in Table 4.

We have attempted to automate the intersecting (superimposing) of the different result sets (temperature-resistant crops, food items with equivalent or improved starch, pyridoxine, ascorbic acid and potassium levels), but unfortunately that has not worked out. The SPARQL query, given below, appeared to be too heavy due to the five filters that were required. With four filters (one removed) it was still possible to obtain results (in a regular desktop computer set up), but the processing time went up unacceptably high at increasing query complexity (given number of filters included):

```
PREFIX fio: <http://www.foodvoc.org/resource/fio#>
PREFIX nevo: <http://www.foodvoc.org/resource/nevo#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX om: <http://www.ontology-of-units-of-measure.org/resource/om-
 2/>

SELECT ?ecocropID ?ecocropLabel ?ecocropMaxGrowthCelsiusTemperature
 ?fioID ?fioLabel ?fioStarchvalue ?fioPyridoxinevalue ?fioAscorbicAc-
 idvalue ?fioPotassiumvalue
WHERE {
  ?ecocropID <http://example.com/base/http%3A//www.w3.org/2000/01/rdf-
 schema#label> ?ecocropLabel.
  ?ecocropID <http://eol.org/schema/terms/maxGrowthCelsiusTemperature>
 ?ecocropMaxGrowthCelsiusTemperature .
  FILTER(?ecocropMaxGrowthCelsiusTemperature >
 "25"^^<http://www.w3.org/2001/XMLSchema#integer>).

  ?ecocropID skos:closeMatch ?fioID .
  ?fioID a fio:FoodItem;
    skos:prefLabel ?fioLabel .

  ?starchnn fio:hasNutrient ?starchnutrient;
    om:hasValue ?starchmeasure.
  ?starchnutrient skos:prefLabel "starch, total"@en.
  ?starchmeasure om:hasNumericalValue ?fioStarchvalue.
  FILTER ( ?fioStarchvalue > 18 ).

  ?pyridoxinenn fio:hasNutrient ?pyridoxinenutrient;
    om:hasValue ?pyridoxinemeasure.
  ?pyridoxinenutrient skos:prefLabel "vitamin B-6, total"@en.
  ?pyridoxinemeasure om:hasNumericalValue ?fioPyridoxinevalue.
  FILTER ( ?fioPyridoxinevalue > 0.1 ).

  ?ascorbicAcidnn fio:hasNutrient ?ascorbicAcidnutrient;
    om:hasValue ?ascorbicAcidmeasure.
  ?ascorbicAcidnutrient skos:prefLabel "vitamin C"@en.
  ?ascorbicAcidmeasure om:hasNumericalValue ?fioAscorbicAcidvalue.
  FILTER ( ?fioAscorbicAcidvalue > 20 ).

  ?potassiumnn fio:hasNutrient ?potassiumnutrient;
    om:hasValue ?potassiummeasure.
  ?potassiumnutrient skos:prefLabel "potassium"@en.
  ?potassiummeasure om:hasNumericalValue ?fioPotassiumvalue.
  FILTER ( ?fioPotassiumvalue > 400 ).

}
```

We have also converted the query to a nested query with subqueries for each of the filters, with the aim to retrieve a relatively small part of the data per subquery and hence reduce the amount of data processed in the overarching main query, but that had no effect. Future research should focus on (further) query optimization.

What is more, the ECOCROP ontology should be extended with candidate alternatives for crops provided by the dietary expert. Presently, *all* crops/food items in the ontology are considered as alternatives (i.e., only based on higher growing temperature and equivalent or better nutritional values), rather than a specific set that is really *suitable* as alternative for the specific food item focused on, replacing the current food item in a meal or recipe at a specific moment of the day.

4 Evaluation

The following section describes the results of the evaluation of the approach. The interoperability was tested through a use-case scenario in which new data was linked to the knowledge graph and queried for results. The section contains a description of the use case scenario, the approach to linking new data, and the results from the query.

Suppose a certain area of cropland is being affected by an increase in average annual temperature, rendering it increasingly more difficult for wheat to grow as it requires not to exceed a certain maximum temperature throughout the year. A farmer may want to find other crop options to cultivate on the farmland in order to increase efficiency and climate resilience. However, besides searching for alternative crops that can withstand higher temperatures, the farmer is also concerned about the change in profits when switching to alternatives. When changing to a different crop, producer prices for yearly yield will also change. Therefore, if the farmer wants to identify climate-resilient crops and prioritize these results based on producer prices per tonne, a new dataset should be added to the knowledge graph.

Table 4. Alternatives for wheat based on nutritional profile and growing temperatures.

Food	Starch (g)	Fibre (g)	Iron (mg)	Zinc (mg)	Vit B1 (mg)	Niacin (mg)	Folate DFE (μg)	Scientific Name	Maximum growing temperature (°C)
Flour wheat wholemeal	53	10.3	4	2.9	0.4	5.7	37	Triticum aestivum	23
Flour wheat white	67.9	3.2	0.8	0.64	0.07	1	22	Triticum aestivum	23
Flour buckwheat	70	5.6	2	2	0.2	2	54	Fagopyrum esculentum	27
Beans black eyed dried	60	4.4	5.8		1.05	2.2	630	Vigna unguiculata	35
Cornmeal	75.5	3.2	0.1	0.3	0.33	0.1	20	Zea mays	33
Flour rice	78	1.3	0.4	1.17	0.04	1	10	Oryza sativa	30

In order to add pricing as a further prioritizing variable for the identified crops, a new dataset was also added to the knowledge graph as a part of evaluation. Note that the data for this use case is synthetic, for evaluation purposes, and it does not represent actual market data. All results and figures from this validation should not be interpreted as real crop pricing data. As data on producer pricing of crops is difficult to find due to frequent changes and a lack of accessibility, a synthetic database provides a viable alternative for a validation use case.

Synthetic data was generated to create a simulation for producer prices on the crops that have been identified as alternatives for wheat (see Table 5). Subsequently, the synthetic data was added to the repository, and the original SPARQL query for wheat was extended and run in the repository as follows.

```
   <http://example.com/base/test_jeroen/>
  PREFIX xsd: <http://www.w3.org/2001/XMLSchema>
  SELECT ?a ?b ?c ?d

  WHERE {
?a <http://example.com/base/http%3A//www.w3.org/2000/01/rdf-
schema#label> ?b.
?a <http://example.com/base/hasEuroPricePerTonne> ?c.
?a <http://eol.org/schema/terms/maxGrowthCelsiusTemperature> ?d
  FILTER(?d >"23"^^<http://www.w3.org/2001/XMLSchema#integer>)
}
```

Table 5. Synthetic data on producer prices for wheat alternatives (€/t).

Food	scientificName	hasLocalCurrencyUnit	producerPricePerTonne (in €)
Flour wheat wholemeal	Triticum Aesticum	euro	205.00
Flour wheat white	Triticum Aesticum	euro	205.00
Flour buckwheat	Fagopyrum Esculentum	euro	540.00
Beans black eyed dried	Vigna Unguiculata	euro	800.00
Cornmeal	Zea Mays	euro	175.00
Flour rice	Oryza Sativa	euro	330.00

After the query output was superimposed on the manually created NEVO alternatives, it resulted in the data as shown in Table 6.

Table 6. Query result for wheat alternatives including producer prices.

taxaID	scientificName	hasEuroPricePerTonne	maxGrowthCelsiusTemperature
http://example.com/base/1574	Oryza sativa	330	30
http://example.com/base/2153	Vigna unguiculata	800	35
http://example.com/base/2175	Zea mays	175	33
http://example.com/base/2285	Fagopyrum esculentum	540	27

5 Conclusions and Future Work

More sustainable food production, distribution and consumption options can be discovered by all stakeholders, eventually leading to near-zero CO_2 emission diets and sustainable food production that will have a positive impact on climate change and will also be adaptive to it. Linking datasets and unchaining the information about crops and food products allow automatically finding nutritionally similar alternatives in case of changing climate. This research demonstrates that automation is possible. In this work, alternatives are generated manually for the three most-imported crops in the Netherlands to showcase the feasibility of automatic generation. The growing conditions of the crops are defined in the ECOCROP ontology which we based on the open ECOCROP data, with the nutritional values available from FIO and based on NEVO. The linking between NEVO database and the ECOCROP ontology is done through the NEVO codes (inserted in the ECOCROP ontology).

The findings demonstrate the effectiveness of linking structured datasets and ontologies to facilitate automated decision-making. By querying the knowledge graph, nutritionally similar alternatives can be identified to adapt to changing climate conditions. Importantly, the use of the developed knowledge graph is not limited to this study alone. It serves as a foundation for further development, inviting a multitude of stakeholders to contribute and integrate additional data sources. Furthermore, future enhancements can involve the integration of an advanced ontology into multiple infrastructures and data platforms, for example, for ontology-enabled food ingredient substitution [14], thereby increasing its utility and impact in the field of sustainable food production. However, extensive querying may be reaching computational performance bottlenecks in usual computational settings of regular users.

Furthermore, a user interface might increase the usability for other stakeholders in the future, besides researchers. It has been demonstrated before that visual elements, including graphs and images, are more easily understood than text and numbers [29]. Earlier research shows techniques for visualizing SPARQL query outputs from GraphDB, with the goal of increasing the understandability of vast knowledge graphs and complex queries. Besides ontology visualization tools such as Nitelight and FedViz, other studies have constructed visualization tools and frameworks to increase understandability with end-users such as in studies on raising awareness of data sharing consent [30, 31]. In these cases, a framework for an application is created where the user communicates with the front end that links to GraphDB through several APIs and visualizes the resulting data for increased understandability. Similar user interfaces could be developed for the current knowledge graph when implemented in non-academic situations.

Acknowledgements. This work has been partially funded by WUR investment theme "Data Driven discoveries in a Changing Climate". The authors would like to thank Edith Feskens, Sander de Leeuw, and Karin Borgonjen for their contributions to this study.

References

1. Stehfest, E., Bouwman, L., van Vuuren, D., den Elzen, M., Eickhout, B., Ka-bat, P.: Climate benefits of changing diet. Clim. Change **95**(1), 83–102 (2009). https://doi.org/10.1007/S10584-008-9534-6
2. Neha, B., Hills, T., Sgroi, D.: Climate Change and Diet. No. 13426. Institute of Labor Economics (IZA) (2020)
3. Aubert, C., Buttigieg, P.L., Laporte, M.A., Devare, M., Arnaud E.: CGIAR Agronomy Ontology, http://purl.obolibrary.org/obo/agro.owl, licensed under CC BY 4.0 (2017)
4. Jonquet, C., et al.: AgroPortal: a vocabulary and ontology repository for agronomy. Comput. Electron. Agric. **144**, 126–143 (2018). https://doi.org/10.1016/j.compag.2017.10.012
5. Matteis, L., et al.: Crop ontology: vocabulary for crop-related concepts. Proceedings of the First International Workshop on Semantics for Biodiversity. CEUR-WS.org (2013)
6. Darnala, B., Amardeilh, F., Roussey, C., Jonquet, C.: Crop planning and production process ontology (C3PO), a new model to assist diversified crop production. In: IFOW 2021-Integrated Food Ontology Workshop @ 12th International Conference on Biomedical Ontologies (ICBO) (2021). hal-lirmm.ccsd.cnrs.fr/lirmm-03389513
7. Riaño, M.A., Rodriguez, A.O.R., Velandia, J.B., García, P.A.G., Marín, C.E.M.: Design and application of an ontology to identify crop areas and improve land use. Acta Geophys. **71**, 1409–1426 (2023). https://doi.org/10.1007/s11600-022-00808-5
8. Ecocrop: Ecocrop Database. FAO, Rome, Italy (2016)
9. Eftimov, T., Ispirova, G., Potočnik, D., Ogrinc, N., Seljak, B.K.: ISO-FOOD ontology: a formal representation of the knowledge within the domain of isotopes for food science. Food Chem. **277**, 382–390 (2019). https://doi.org/10.1016/j.foodchem.2018.10.118
10. Cooper, L., et al.: The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Res. **46**(D1), D1168–D1180 (2018). https://doi.org/10.1093/nar/gkx1152
11. Plant Ontology™ Consortium.: The Plant Ontology™ consortium and plant ontologies. Comp. Func. Genom. **3(2)**, 137–142 (2002). https://doi.org/10.1002/cfg.154
12. Dooley, D.M., et al.: FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. NPJ Sci. Food **2**(1), 23 (2018). https://doi.org/10.1038/s41538-018-0032-6

13. Haussmann, S., et al.: FoodKG: a semantics-driven knowledge graph for food recommendation. In: The Semantic Web– ISWC 2019: 18th International Semantic Web Conference Proceedings, Part II 18, pp. 146–162. Springer International Publishing Auckland, New Zealand, (2019). https://doi.org/10.1007/978-3-030-30796-7_10
14. Ławrynowicz, A., Wróblewska, A., Adrian, W.T., Kulczyński, B., Gramza-Michałowska, A.: Food recipe ingredient substitution ontology design pattern. Sensors **22**(3), 1095 (2022). https://doi.org/10.3390/s22031095
15. Food Item Ontology. https://git.wur.nl/FoodInformatics/foodontology.git
16. Mertens, E., Kaptijn, G., Kuijsten, A., van Zanten, H., Geleijnse, J. M., van 't Veer, P.: SHARP-Indicators Database towards a public database for environmental sustainability. Data Br. **27**, 104617 (2019). https://doi.org/10.1016/j.dib.2019.104617
17. Blonk Sustainability | Databases. https://blonksustainability.nl/tag/Databases
18. RIVM Life Cycle Assessment (LCA) database. https://www.rivm.nl/life-cycle-assessment-lca
19. Cortesi, A., Pénicaud, C., Saint-Eve, A., Soler, L.G., Souchon, I.: Life cycle inventory and assessment data for quantifying the environmental impacts of a wide range of food products belonging to the same food category: a case study of 80 pizzas representatives of the French retail market. Data Br. **41**, 107950 (2022). https://doi.org/10.1016/j.dib.2022.107950
20. Eaternity Database. https://eaternity.org/foodprint/database
21. World Food LCA Database. https://ourworldindata.org/explorers/
22. Sinfort, C., Perignon, M., Drogué, S., Amiot, M.J.: Dataset on potential environmental impacts of water deprivation and land use for food consumption in France and Tunisia. Data Br. **27**, 104661 (2019). https://doi.org/10.1016/j.dib.2019.104661
23. Clark, M., et al.: Estimating the environmental impacts of 57,000 food products. Proc. Natl. Acad. Sci. **119**(33), e2120584119 (2022). https://doi.org/10.1073/pnas.2120584119
24. Notarnicola, B., et al.: Life cycle inventory data for the Italian agri-food sector: background, sources and methodological aspects. Int. J. LCA., 1–16 (2022). https://doi.org/10.1007/s11367-021-02020-x
25. FAOSTAT. https://www.fao.org/faostat/en/#home
26. Ontotext Refine tool. https://www.ontotext.com/products/ontotext-refine/
27. NCBITaxon. http://obofoundry.org/ontology/ncbitaxon.html
28. Hatfield, J.L., et al.: Climate impacts on agriculture: implications for crop production. Agron. J. **103**(2), 351–370 (2011). https://doi.org/10.2134/agronj2010.0303
29. Passera, S.: Enhancing contract usability and user experience through visualization-an experimental evaluation. In: 16th International conference on information visualization, pp. 376–382. IEEE (2012). https://doi.org/10.1109/IV.2012.69
30. Bless, C., et al.: Raising awareness of data sharing consent through knowledge graph visualization. In: Further with Knowledge Graphs, pp. 44–57. IOS Press (2021). https://doi.org/10.3233/SSW210034
31. Rasmusen, S.C., et al.: Raising consent awareness with gamification and knowledge graphs: an automotive use case. Int. J. Semantic Web Inf. Syst. (IJSWIS), **18**(1), 1–21. Igi-global.com (2022). https://doi.org/10.4018/IJSWIS.300820

Extreme and Sustainable Graph Processing for Green Finance Investment and Trading

Laurenţiu Vasiliu[1], Dumitru Roman[2], and Radu Prodan[3](✉)

[1] Peracton Ltd., Galway, Ireland
laurentiu.vasiliu@peracton.com
[2] SINTEF AS, Oslo, Norway
dumitru.roman@sintef.no
[3] Institute of Information Technology, University of Klagenfurt, Klagenfurt, Austria
radu.prodan@aau.at

Abstract. The Graph-Massivizer project, funded by the Horizon Europe research and innovation program, aims to create a high-performance and sustainable platform for extreme data processing. This paper focuses on one use case that addresses the limitations of financial market data for green and sustainable investments. The project allows for the fast, semi-automated creation of realistic and affordable synthetic (extreme) financial datasets of any size for testing and improving AI-enhanced financial algorithms for green investment and trading. Synthetic data usage removes biases, ensures data affordability and completeness, consolidates financial algorithms, and provides a statistically relevant sample size for advanced back-testing.

Keywords: Biases · Computing continuum · Ethical concerns · Extreme data · Graph processing · Green finance · Knowledge graphs · Serverless computing · Sustainability · Synthetic data

1 Introduction

The wide use, availability, accessible costs, interoperability, and analytical exploitation of financial data are essential for the European data strategy. Graphs or linked data are crucial to innovation, competition, and prosperity and establish a strategic investment in technical processing and ecosystem enablers. Graphs are universal abstractions that capture, combine, model, analyze, and process knowledge about real and digital worlds into actionable insights through item representation and interconnectedness. In this context, graphs are extreme data enablers that require further technological innovations to meet the needs of the European data economy. A study by IBM [1] revealed that the world generates nearly 2.5 quintillion bytes of financial data daily, posing extreme business analytics challenges. Graph-based technologies help pursue the United Nations Sustainable Development Goals [2] by enabling better value chains, products, and services for green financial investments and deriving trustworthy insights for creating sustainable communities.

R. Akerkar (Ed.): SAIDD 2023, CCIS 1810, pp. 120–135, 2024.
https://doi.org/10.1007/978-3-031-53770-7_8

The improvement and optimization of green investments and trading face significant barriers. Historical securities' data, particularly on environmental, social, and governance data (starting from the early 2010s, is insufficient for in-depth testing, derisking financial algorithms, and training AI models. Unfortunately, financial data is often difficult and expensive to access for training AI-driven financial algorithms. One historical record per security is typical to optimize a financial strategy, but this can lead to deficiencies in data training and losses during live trading.

The *Graph-Massivizer project* [1] aims, among others, to remove the limitations of financial market data (limited volume, reduced accessibility, price barriers) by enabling fast, semi-automated creation of realistic and affordable synthetic extreme financial datasets, unlimited in size and accessibility. The extreme synthetic data goes one order of magnitude beyond the current big financial data features, aiming for petabytes in volume and affordable prices. The project researches and develops a high-performance, scalable, gender-neutral, secure, and sustainable platform based on the *massive knowledge graph (KG)* representation of extreme financial data. It delivers the *Graph-Massivizer toolkit* of five open-source software tools and findable, accessible, interoperable, and reusable (FAIR) graph datasets that cover the sustainable lifecycle of processing extreme data as massive graphs, as shown in Fig. 1.

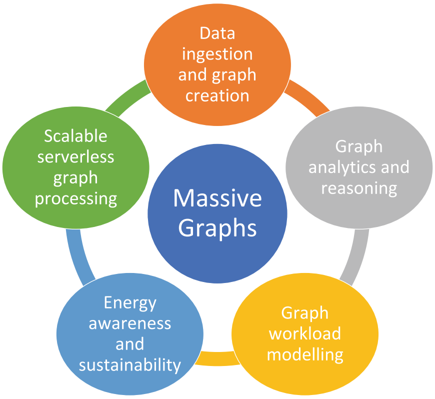

Data ingestion and graph creation
Graph analytics and reasoning
Graph workload modelling
Energy awareness and sustainability
Scalable serverless graph processing
Massive Graphs

Fig. 1. Sustainable massive KG operation lifecycle [1].

This paper provides a comprehensive introductory overview of the green and sustainable finance pilot use case researched in the Graph-Massivizer project. Section 2 outlines the related work structured targeting functional gaps in historical financial data offerings as motivation for synthetic financial data alternative to actual data and concludes with a short list of existing relevant financial synthetic data companies highlighting market innovation potential. Section 3 presents the green and sustainable finance use case comprising its conceptual architecture, objectives, and scientific challenges. Section 4 outlines the conceptual architecture of the Graph-Massivizer platform and toolkit, followed by a summary of the planned financial use case integration in Sect. 5. Section 6 concludes the paper.

2 State-of-the-Art in Green and Sustainable Finance

Intensive testing, data accuracy, quality, and quantity are paramount to investment and trading. Strict adherence to statistical relevance is necessary, such as conducting back-tests on financial algorithms with a minimum of 10,000 out-of-sample data points. Testing becomes even more challenging when various machine learning (ML) models are employed or benchmarked against each other to enhance existing financial algorithms. While sourcing and preparing data has become more accessible for companies in recent years, numerous challenges persist. We identify several gaps that exist in commercial financial data.

2.1 Functional Gaps in Historical Financial Data

According to Appen's 2022 report [4], 42% of technologists find *data sourcing* challenging, 34% find data preparation, and 38% find model testing and deployment. Furthermore, 51% find *data accuracy* critical for artificial intelligence (AI) use cases. However, 78% find that the training data accuracy varies widely between 1% and 80%. Training and testing AI models with low-quality data make the model predictions and results inaccurate and inapplicable to financial transactions.

Regarding *data volumes*, Capital Fund Management [5] used all historical data from 1800 for back-testing to achieve statistical significance. Time-series data for futures and equities extends back at least to the 1960s and 1970s and, where possible, as far back as 1800 (e.g., monthly data for many indices, commodities, bonds, and various interest rates). To show the scale of their current data acquisition, they have over 1,500 servers enabling daily collection and presentation of over three terabytes of information. However, historical financial data may be irrelevant to financial model testing for several reasons, as discussed in the following paragraphs.

Changes in the Business Environment. The business environment can change rapidly, and old financial data may not reflect the current market conditions. Economic factors such as inflation, interest rates, and trade policies can significantly impact a company's financial performance, and these factors can change over time.

Changes in the Company's Operations. Companies can change operations over time, affecting their financial performance. For example, a manufacturing company may shift to a service-oriented business model, impacting its financial statements.

Changes in Accounting Standards. Accounting standards can change over time, making reporting, recording, and comparing from different periods difficult, as the accounting methods may differ.

Outdated Technology. The methods used to collect and process financial data can become outdated. For example, financial data collected manually and recorded on paper may be less reliable than data collected electronically.

Data Quality. Financial data can be subject to errors and inconsistencies, which become more prevalent over time as the data becomes outdated and challenging to use for accurate testing of financial models.

Statistical significance and long periods of back-tests are critical. However, going back in time, even to the 1970s and 1960s, the richness of the data decreased dramatically compared to our days. Furthermore, the market conditions back then were different, and algorithms designed for today's markets do not necessarily fit the conditions from the 1960s and 1970s.

Data purchase and storage costs can be another critical barrier, except for big and rich financial institutions that can theoretically afford to purchase any volume if needed. As big financial data refers to terabytes of structured and unstructured data, it is costly for any financial player except the big institutions to source, store, prepare, and test models. Therefore, most companies use smaller, more affordable sets that are insufficient, incomplete, or biased, affecting the models' results and performance.

Another critical issue is the *lack of accuracy and audit* of some real-world financial data, particularly the *environmental, social, and governance (ESG)* data, briefly detailed in Table 1. Unlike accounting data that undergo auditing to ensure its accuracy and integrity, there are no clear regulations for verifying ESG data accuracy [6]. Nevertheless, regulatory efforts are underway, as evident from the shift of ESG issues from being voluntary disclosure-oriented to becoming regulatory. This development has significant implications for organizations collecting, verifying, and utilizing ESG information. As per Thomson Reuters, while the current regulations are still incomplete, the direction is towards greater regulation of ESG issues (see Table 1).

Table 1. Environmental, social, and governance parameters.

Environmental Parameters	Social Parameters	Governance Parameters
- Carbon emissions [7] - Water use [8] - Waste production [8] - Energy efficiency [9] - Renewable energy [9] - Biodiversity and habitat protection [10] - Hazardous materials management [10] - Air pollution and emissions reduction [7] - Climate change strategy [11]	- Employee turnover [9] - Employee diversity and inclusion [9] - Employee health and safety [12] - Employee compensation and benefits [9] - Human rights policies [13] - Community relations and impact [8] - Customer privacy and data security - Product safety and quality [8] - Supply chain management [8]	- Board diversity and independence [8] - Executive compensation and incentives [8] - Shareholder rights and engagement [14] - Political contributions and lobbying [8] - Transparency and disclosure [8] - Anti-corruption policies and practices [15] - Business ethics and code of conduct [8] - Risk management and oversight [8]

One significant challenge associated with using limited or reused real datasets is the issue of financial algorithms' *overfitting*. This phenomenon occurs when designing algorithms to fit the real datasets too closely, resulting in high performance on test data but poor performance on previously unseen data. Overfitting is a common problem in ML and data mining, mitigated by regularization, cross-validation, and ensemble learning techniques. Nevertheless, overfitting remains a critical issue that needs careful consideration when working with real datasets in algorithm development (Table 2).

Table 2. Environmental, social, and governance data challenges.

ESG Data Challenge	Explanation
Inconsistency and unreliability	- ESG data can have a variety of inconsistent or unreliable provider sources - Companies may not provide complete or accurate data, impacting the reliability of ESG ratings and scores
Lack of standardization and comparability	- ESG data often have different reporting formats, which makes it challenging to compare across companies and industries - No standardized methodology for calculating ESG ratings results in discrepancies and inconsistencies
Limited scope and coverage	- ESG data typically focuses on a subset of relevant issues by rating agencies or investors, which may not capture all ESG risks and opportunities for a company
Lack of correlation with financial performance	- There is limited correlation evidence between high ESG scores or ratings and better financial performance, making it difficult to justify using ESG data in investment decision-making

Although historical financial data represent quantified actual events, history never repeats itself. Even for recurring market booms and crashes [16], the underlying economic factors are distinct and unique each time, exposing the models and algorithms to *new and unencountered financial situations*. As a result, linking historical data with performance can be challenging. The quote commonly associated with mutual funds warning investors that past performance is not a reliable indicator of future success exemplifies this challenge.

2.2 Synthetic Financial Data

Given the limitations of historical data, *synthetic data*, which closely mimics real-world data, emerges as an alternative, complementary solution for testing financial models and algorithms. According to Johnathan Kinlay, head of quantitative trading at Systematic Strategies LLC [17], synthetic data addresses one primary concern about using real data series for modeling purposes. Concretely, models designed to fit the historical data produce test results that one is unlikely to replicate [18]. Such models are not robust to changes likely occurring in any dynamic statistical process and will perform poorly out of sample. Synthetic data will expose and stress-test financial models to new situations, thus validating or invalidating their assumptions and exposing their strengths and weaknesses.

Linden [19] raised many interesting points about the *"usability and future of synthetic data, increasing the accuracy of ML models."* Real-world data is happenstance and does not contain all permutations of conditions or events possible in the real world. Synthetic data can help generate data at the edges or for unseen conditions.

Deployed correctly, data and analytics leaders can use synthetic data to create more efficient AI models, taking their organizations' applications to the next level [20], according to a Gartner analyst. Gartner further estimates that by 2030, synthetic data will overshadow actual data in a wide range of AI models and will help organizations understand the technology's potential [21]. Nevertheless, risks are also present when using synthetic data whose quality relies on the quality of the model that created it and the resulting dataset. Therefore, synthetic data requires additional verification steps, such as a comparison with human-annotated, real-world data, to ensure its validity. The widespread of synthetic data raises questions about the transparency and explicability of the techniques used for generating it.

According to Datanami [28], hedge funds and banks deploy KGs as a powerful option to meet growing functional data management challenges. Using KGs gets increased traction in the financial space.

2.3 Green and Sustainable Finance Market

Statice.ai [22] presents a comprehensive list of 56 synthetic data vendors and a thriving ecosystem covering many industries. However, few vendors target financial markets with dedicated KG applications. Table 3 summarizes several relevant synthetic data companies focused on finance and banking selected from an extensive directory [23]. We identified a large ecosystem of synthetic data producers and users but very few in the green financial market. The analysis shows a strong drive for synthetic data and some competition but enough space for innovative newcomers.

Table 3. Relevant financial synthetic data companies.

Company	Offering	Analysis
Gretel [24]	- Unlimited synthesized datasets - Privacy-preserving transformations on sensitive data - Advanced NLP detection of personally identifiable information	- Unlimited synthetic data generation and delivered as-a-service - Automated transformations to anonymize data with privacy guarantees - Automatic and continuous detection and labeling of sensitive data - Financial study: time series-focused, banking customer accounts, synthetic data pipeline [24]

(*continued*)

Table 3. (*continued*)

Company	Offering	Analysis
Statice [25]	- Synthetic data for financial applications - Predictive analytics algorithms - Secondary data use with high compliance overhead reduction - Unified data use across jurisdictions	- Synthetic financial data with a statistical value similar to the original data, fit to use as a drop-in replacement for data science operations - Large volume of synthetic data, easily accessed, shared, and used for business intelligence, analysis, and ML
Synthesized [26]	- High-quality data for ML, application development, and testing	- Easy-to-use configuration files
DataCebo [27]	- PAR class implementation of a probabilistic autoregressive model for learning multi-type, multivariate time series data	- Synthetic data generation with identical format and properties - Synthetic time series generation conditioned on the properties of the associated entity

3 Graph Processing for Green Finance

Green finance targets financial products and services that direct investments into green-oriented enterprises. It aims for economic growth while reducing waste, pollution, and greenhouse gas emissions and improving overall efficiencies. Sustainable finance considers environmental, social, and governance factors for investment decisions for long-term sustainable economic activities.

Financial KG Use Cases. Deloitte [29] lists several use cases that apply KG to the financial services sector, summarized in Table 4. Linked data standards such as the hypertext transfer protocol, uniform resource identifiers, and Resource Description Framework (RDF) [30] models represent data in a single interchangeable format that machines and humans understand. Additionally, multi-model graph databases support multiple data models against a single, integrated backend.

Financial KG. Graph-Massivizer aims to remove the limitations of financial market data providers (limited volume, reduced accessibility, high costs) by enabling fast, semi-automated creation of realistic and affordable synthetic extreme financial data sets, unlimited in size and accessibility. The extreme financial datasets will enable improved ML-based green investment and trading simulations, free of critical biases such as prior knowledge, overfitting, and indirect contaminations due to present data scarcity. We plan to research a *financial knowledge graph* as a fundamental data structure representing a hybrid, graph-based financial metadata structure (time series, values, boolean, monetary, securities taxonomies, statistical factors, rules) that helps research improved financial algorithms operating in five high-level steps:

Table 4. Financial KG use case examples.

Compliance Management	- KGs leverage the power of semantic technologies to unify and interlink various sources of compliance data and apply complex rules and patterns for (semi-) automated compliance monitoring - KGs combine contextual domain knowledge with natural language processing (NLP) and ML
Data Lineage and Metadata Management	- The combination of detailed metadata and relationships between data lifecycle phases results in a semantic data layer - A semantic data layer (data fabric) enables data experts and business stakeholders to take advantage of any data asset they access
Fraud Detection and Financial Crime Analytics	- KGs empowered by ML and reasoning capabilities allow companies to better identify fraudulent patterns by traversing many hops on vast amounts of interconnected data in real time
Recommender Systems	- A KG built as an extensive semantic network of entities and attributes allows finding best-matching entities based on semantic similarities - KGs also enrich the data context by incorporating domain-specific knowledge vocabularies, taxonomies, or ontologies

- Historical financial data structure mapping into a financial KG;
- Synthetic data generation by preserving the original historical statistical features;
- Missing data interpolation using ML inference and reasoning methods;
- Green financial investments and trading simulation;
- Recommendation of the “greenest” investments and trading opportunities.

Scientific Challenges. To define our goals and methodology, we first conducted a comprehensive analysis to identify the scientific challenges of using KGs for green and sustainable finance, as summarized in Table 5.

Green Financial Data Multiverse. Graph-Massivizer targets energy-efficient synthetic financial data generation (Fig. 2) in a 1–75 petabytes range, validated by standard green financial investment and trading algorithms. The developed technology promises a 90% energy consumption accountability for extreme data creation streamed to clients. Samples of it will be available as open data for internal testing. The availability of cheaper synthetic financial data for testing in extreme quantities allows more fintech companies, funds, and investors to test and derisk investment models.

Greener Financial Algorithms and Better Investments. Peracton Ltd. Aims to use the financial multiverse for improved green AI-enhanced financial algorithms with reduced bias and risk and an increased investment return by a realistic 2%–4%. It further targets

Table 5. Scientific challenges of KGs for green and sustainable finance.

Scientific Challenge	Description
Complex relationship representation	- Create synthetic financial data using KGs accurately representing financial market relationships and interactions - Obey rules and correlation requirements relative to the original historical data they model - Model relationships between highly dynamic asset classes, such as equities, fixed income, and commodities, as dynamic KGs - Investigate and eliminate biases and assumptions in training data used by ML algorithms that generate synthetic financial data
Bias and ethical concerns	- Consider ethical concerns when using synthetic data for financial applications (e.g., for decisions that affect people's lives) - Generate and use high-quality and complete synthetic financial data for the KG
Data quality and completeness	- Apply data cleaning and preprocessing techniques to ensure the accuracy and completeness of the data for building the KG
Interoperability	- Develop visualization and interpretation tools that help understand and use synthetic graph data with complex relationships
Reasoning and inference scalability	- Perform real-time inference or reasoning on the large and complex KG data
Reasoning and inference uncertainty	- Infer relationships and make predictions based on noisy financial data containing errors and outliers
Data integration for reasoning and inference	- Integrate dispersed financial data in different formats into KG - Perform inference and reasoning on inconsistent graphs
Reasoning and inference explain ability	- Explain the inferences or predictions the KG makes for complex financial models
Stream partitioning	- Research adequate graph partitioning to reduce overall graph processing times and enable distributed processing

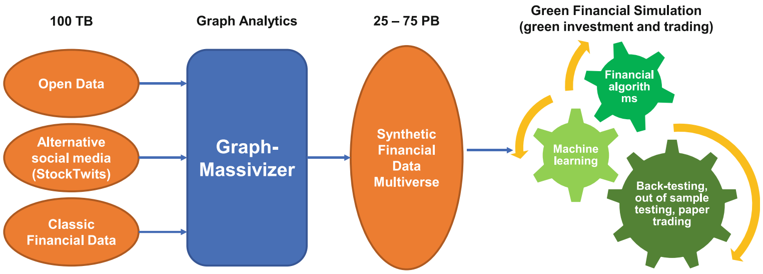

100 TB
Graph Analytics
25 – 75 PB
Green Financial Simulation (green investment and trading)
Open Data
Alternative social media (StockTwits)
Classic Financial Data
Graph-Massivizer
Synthetic Financial Data Multiverse
Financial algorith ms
Machine learning
Back-testing, out of sample testing, paper trading

Fig. 2. Green financial data multiverse in Graph-Massivizer [1].

an increase in excess return (alpha) by 1% – 2% with a quick ratio higher than 1.5, reflecting healthy investments with lower risk and higher returns.

4 Graph-Massivizer Toolkit

The Graph-Massivizer toolkit architecture consists of five tools, depicted as a simplified C4 container diagram in Fig. 3 and published in [1]: Graph-Inceptor, Graph-Scrutinizer, Graph-Optimizer, Graph-Greenifier, and Graph-Choreographer.

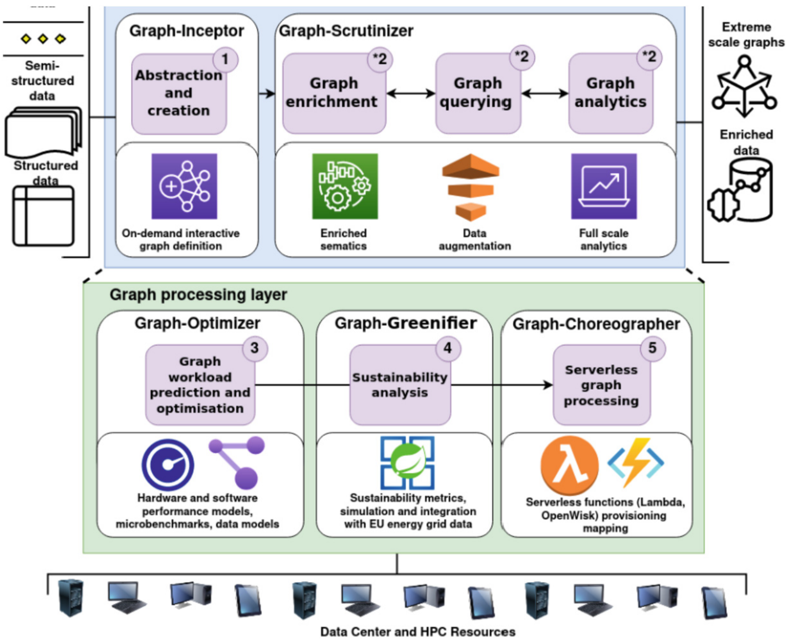

Semi-structured data
Structured data
Graph-Inceptor
1
Abstraction and creation
On-demand interactive graph definition
Graph-Scrutinizer
*2
Graph enrichment
Graph querying
Graph analytics
Enriched sematics
Data augmentation
Full scale analytics
Extreme scale graphs
Enriched data
Graph processing layer
Graph-Optimizer
3
Graph workload prediction and optimisation
Hardware and software performance models, microbenchmarks, data models
Graph-Greenifier
4
Sustainability analysis
Sustainability metrics, simulation and integration with EU energy grid data
Graph-Choreographer
5
Serverless graph processing
Serverless functions (Lambda, OpenWisk) provisioning mapping
Data Center and HPC Resources

Fig. 3. Graph-Massivizer conceptual architecture using a simplified C4-context diagram [1].

Operational graph layer generates, transforms, and manipulates extreme data through BGOs, which comprise graph creation, enrichment, query, and analytics.

Graph creation implemented by the *Graph-Inceptor* tool translates extreme data from various static and event streams or follows heuristics to generate synthetic data, persist it, or publish it within a graph structure.

Graph enrichment, graph query, and graph analytics are three operation types implemented by the *Graph-Scrutinizer* tool. They analyze and expand extreme datasets using probabilistic reasoning and ML algorithms for graph pattern discovery, low memory footprint graph generation, and low latency error-bounded query response. The output of this phase is a new graph, a query, or an enriched structured dataset.

Graph processing layer provides sustainable, energy-aware, serverless graph analytics on the heterogeneous HPC infrastructure.

Graph workload modeling and optimization, represented by the Graph-Optimizer tool, analyses and expresses the given graph processing workload into a workflow of BGOs. It further combines parametric BGO performance and energy models with hardware models to generate accurate performance and energy consumption predictions for the workload running on a given multi-node, heterogeneous infrastructure of CPUs, GPUs, and FPGAs. The predictions indicate the most promising combinations of BGO optimizations and infrastructure, representing a codesigned solution for the given workload while guaranteeing its performance and energy consumption bounds.

Sustainability analysis, implemented by the *Graph-Greenifier* tool, collects, studies, and archives performance and sustainability data from operational data centers and national energy suppliers on a large scale. It simulates multi-objective infrastructure sustainability profiles for operating graph analytics workloads, trading off performance and energy (e.g., consumption, CO_2, methane, GHG emissions) metrics. Its purpose is to model the impact of specific graph analytics workloads on the environment for evidence-based decision-making.

Serverless BGO processing, implemented by the *Graph-Choreographer* tool, uses performance and sustainability models and data to deploy serverless graph analytics on the computing continuum. It relies on novel scheduling heuristics, infrastructure partitioning, and environment-aware processing for scalable orchestration of serverless graph analytics with accountable performance and energy tradeoffs.

Hardware infrastructure layer considered by Graph-Massivizer consists of geographically distributed data centers across the Cloud HPC, mid-range Fog, and low-end Edge computing continuum.

5 Green and Sustainable Finance in Graph-Massivizer

The container diagram in Fig. 4 depicts the Graph-Massivizer tool pipeline implementing the green and sustainable finance use case for fast and semi-automated creation of realistic and affordable synthetic financial datasets of extreme size. Peracton Ltd. Will test the produced extreme synthetic financial data on its platform to evaluate their quality and the impact on green financial algorithms training and testing results. The extreme data sets for training allow (by design) various data distributions and market scenarios, including extreme variations that never occurred in the past. They will expose financial algorithms and models to new conditions and tune them with appropriate stress tests.

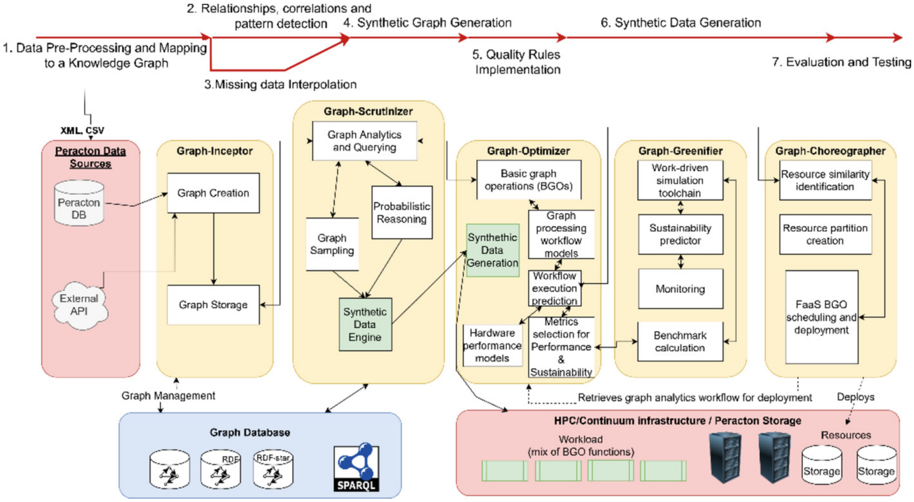
Table 6. Bias and ethical concerns in green and sustainable finance.

Concern	Description	Implication
Bias amplification	Synthetic data can inherit biases from the training data used to create and potentially amplify them	Financial historical data can inherit biases from accounting metric calculation or interpretation, which can vary with geographic accounting practices and are unalterable or uncorrectable if produced by specialized providers such as Bloomberg or Thomson Reuters
Representativeness bias	Synthetic data may not accurately represent real-world diversity and complexity, leading to skewed or incomplete representations	The design phase must consider quality and existence rules and the diversity and complexity of the actual historical data and reflect it in the synthetic data
Unintentional leakage	Synthetic data generation methods could inadvertently include private or sensitive information from the original dataset, violating privacy and confidentiality	There is no private or sensitive information in the historical financial data publicly disclosed by companies that could leak in the synthetic generation
Over-generalization	Synthetic data may oversimplify complex real-world scenarios, leading to generalizations and assumptions that may not always hold	The synthetic data design must capture historical data complexities need to be captured in individual patterns and replicate them
Unintended correlations	Synthetic data generation may introduce artificial nonexisting correlations between variables, leading to biased or misleading analysis and predictions	New correlations are not necessarily wrong because the purpose is to train algorithms and models on new unencountered situations and test in extreme and unusual circumstances
Data sparsity	Synthetic data may not capture rare occurrences or tail-end events in actual data depending on the generation method, limiting its specific applicability	The design must consider data sparsity by closely matching the historical series as non-Gaussian with fat tails

(*continued*)

Table 6. (*continued*)

Concern	Description	Implication
Validation	It can be challenging to validate synthetic data's quality, accuracy, and generalisability due to the absence of ground truth	The design and open-source tools like the SDMetrics library can control the quality of synthetic data by analyzing statistical properties
Lack of consent	Synthetic data based on real individuals' data raises concerns about consent	Historical financial data is publicly available under various academic and commercial licenses and contains no private individual data
Accountability and responsibility	Accountability and responsibility for potential errors or biases in data generation are crucial in decision-making with complex liability	It is essential to connect trained financial algorithms to live data streams for advanced testing with virtual money under human supervision before employing them in real scenarios
Unintended consequences	Using synthetic data may have unintended consequences, such as reinforcing biases or influencing decisions, perpetuating discrimination or unfairness	Observing and monitoring the algorithms for unusual or novel patterns in their decision-making during real testing with streaming data under human supervision is essential



1. Data Pre-Processing and Mapping to a Knowledge Graph
2. Relationships, correlations and pattern detection
3.Missing data Interpolation
4. Synthetic Graph Generation
5. Quality Rules Implementation
6. Synthetic Data Generation
7. Evaluation and Testing
XML, CSV
Peracton Data Sources
Peracton DB
External API
Graph-Inceptor
Graph Creation
Graph Storage
Graph-Scrutinizer
Graph Analytics and Querying
Probabilistic Reasoning
Graph Sampling
Synthetic Data Engine
Graph-Optimizer
Basic graph operations (BGOs)
Synthethic Data Generation
Graph processing workflow models
Workflow execution prediction
Hardware performance models
Metrics selection for Performance & Sustainability
Graph-Greenifier
Work-driven simulation toolchain
Sustainability predictor
Monitoring
Benchmark calculation
Graph-Choreographer
Resource similarity identification
Resource partition creation
FaaS BGO scheduling and deployment
Graph Management
Graph Database
RDF
RDF-star
SPARQL
Retrieves graph analytics workflow for deployment
Deploys
HPC/Continuum infrastructure / Peracton Storage
Workload (mix of BGO functions)
Resources
Storage
Storage

Fig. 4. Green and sustainable finance architectural component diagram in Graph-Massivizer.

Data Preprocessing and KG Creation. The *financial data sources* component provides historical data samples purchased by PER as input to the platform in XML and CSV formats. Then, the *Graph-Inceptor* tool extracts the historical financial data, maps it on a massive financial KG structure using its graph creation component, and stores it in the graph database for later use. The later use can also support auditing these processes.

Relationships and Pattern Detection. The *Graph-Scrutinizer* tool runs probabilistic reasoning on the financial KG to identify financial data patterns and correlations that are hard to identify on the raw data. Then, the *inter-company/product graph (ICG)* creation component driven by a graph convolutional network engine records the identified patterns and correlations.

Synthetic Data Generation. The *missing data interpolation* component handles incomplete data and gaps encountered in the historical data and generates synthetic data ranges to fill them. Then, the *synthetic graph generation* component uses the ICG to generate a synthetic financial KG as a template for generating further synthetic data;

Quality Rules Implementation. The *Graph-Optimizer* tool executes BGOs and implements synthetic data quality rules with the generated synthetic financial KG to instantiate synthetic data further.

Synthetic Data Generation. The *financial storage* (and the continuum infrastructure) host the generated synthetic data for further use in financial model simulations.

6 Conclusions

Graph-Massivizer allows European green financial investors to avail of a massive financial data synthetic multiverse and a proven competitive, sustainable advantage. In comparison, other forms of analysis rely on present assumptions about *"what happened"* or "*what happens"* correctly building and employing financial graphs to generate synthetic data can further reveal patterns suggesting what *"might happen"* with clear evidence for each connection or inference step. Graph processing facilitates problem-solving driven by metrics related to costs or inefficiencies. The large-scale financial graph analytics market still traverses a developing phase, hampered by the lack of technology research and use case adoption. Graph-Massivizer provides for Europe these missing links.

Finally, we constantly monitor and address addresses biases and ethical concerns arising from using various AI, IT technologies, and data, as summarized in Table 6.

Acknowledgments. The European Union funds this work through the Graph-Massivizer project (Horizon Europe 101093202).

References

1. Prodan, R., et al.: Towards extreme and sustainable graph processing for urgent societal challenges in Europe. In: 2022 IEEE Cloud Summit, pp. 23–30. IEEE (2022)

2. United Nations. Sustainable Development Goals. https://www.un.org/sustainabledevelopment/. Accessed 2023/09/19
3. Oriel, A.: Accelerating the process of financial trading with big data analytics. https://www.analyticsinsight.net/accelerating-the-process-of-financial-trading-with-big-data-analytics/. Accessed 14 July 2023
4. Appen. State of AI and Machine Machine Learning Report. https://appen.com/stateofai2022/. Accessed 14 July 2023
5. Capital Fund Management. Our approach. https://www.cfm.com/our-approach/. Accessed 14 July 2023
6. Thomson Reuters. How regulations are moving ESG into the risk and compliance field. https://www.thomsonreuters.com/en-us/posts/investigation-fraud-and-risk/esg-regulations-compliance/. Accessed 14 July 2023
7. CDP. Disclosure insight action. https://www.cdp.net/. Accessed 14 July 2023
8. GRI. The global leader for impact reporting. https://www.globalreporting.org/. Accessed 14 July 2023
9. IEA. https://www.iea.org/. Accessed 14 July 2023
10. UN Environment Programme. https://www.unep.org/. Accessed 14 July 2023
11. Task Force on Climate-related Financial Disclosures. https://www.fsb-tcfd.org/. Accessed 14 July 2023
12. United States Department of Labor. Occupational Safety and Health Administration. https://www.osha.gov/. Accessed 14 July 2023
13. United Nations Global Compact. https://unglobalcompact.org/. Accessed 14 July 2023
14. Principles of responsible investment. https://www.unpri.org/. Accessed 14 July 2023
15. United Nations Office on Drugs and Crime. United Nations Convention against Corruption. https://www.unodc.org/unodc/en/treaties/CAC/. Accessed 14 July 2023
16. The Economic Times. Does history repeat or rhyme in financial markets? https://economictimes.indiatimes.com/markets/stocks/news/does-history-repeat-or-rhyme-in-financial-markets/articleshow/78144154.cms. Accessed 14 July 2023
17. Systematic Strategies. A quantitative investment management firm. https://systematic-strategies.com/, 14 July 2023
18. Kinlay, J.: A new approach to generating synthetic market data. https://jonathankinlay.com/2022/07/a-new-approach-to-generating-synthetic-market-data/. Accessed 14 July 2023
19. Linden, A.: Is synthetic data the future of AI? https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai. Accessed 14 July 2023
20. Gartner: Can synthetic data drive the future of AI? https://aibusiness.com/data/gartner-can-synthetic-data-drive-the-future-of-ai-. Accessed 14 July 2023
21. Goasduff, L.: Adaptive artificial intelligence (AI) systems, data sharing, and data fabrics are among the trends that data and analytics leaders need to build on to drive new growth, resilience and innovation. https://www.gartner.co.uk/en/articles/12-data-and-analytics-trends-to-keep-on-your-radar. Accessed 14 July 2023
22. Devaux, E.: List of synthetic data vendors – 2022. https://elise-deux.medium.com/new-list-of-synthetic-data-vendors-2022-f06dbe91784. Accessed 14 July 2023
23. Browse a collection of synthetic data tools and companies. https://syntheticdata.carrd.co/. Accessed 14 July 2023
24. Gretel. Creating Synthetic time series data for global financial institutuons. https://cdn.gretel.ai/case_studies/gretel_time_series_case_study.pdf. Accessed 14 July 2023
25. Statice by Anonos. https://www.statice.ai/. Accessed 14 July 2023
26. Synthesized. https://www.synthesized.io/. Accessed 14 July 2023
27. Datacebo. Make synthetic data a reality. https://datacebo.com/. Accessed 14 July 2023

28. Datanami. Why knowledge graph for financial services? Real use cases. https://www.datanami.com/2022/03/28/why-knowledge-graph-for-financial-services-real-use-cases/. Accessed 14 July 2023
29. Deloitte. Knowledge graphs for financial services. The path to unlock new insights from your data. https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/risk/deloitte-nl-risk-knowledge-graphs-financial-services.pdf. Accessed 14 July 2023
30. W3C, RDF 1.1 Concepts and Abstract Syntax. https://www.w3.org/TR/rdf11-concepts/. Accessed 14 July 2023
31. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
32. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016)
33. Author, F., Author, S., Author, T.: Book title, 2nd edn. Publisher, Location (1999)
34. Author, F.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
35. LNCS Homepage. http://www.springer.com/lncs. Accessed 20 Jun 2023

A Comprehensive Framework for Detecting Behavioural Anomalies in the Elderly

Ankit Jain(✉) and Abhishek Srivastava

Indian Institute of Technology Indore, Madhya Pradesh, India
{phd1801101002,asrivastava}@iiti.ac.in

Abstract. The world is seeing a rapid increase in the population of the aged. This, combined with a shortage of affordable care-giving manpower, leads to a dependence on automated systems for monitoring the well-being of the elderly and detecting abnormalities. There exist techniques based on sensors of various types to detect and recognize the daily activities of the elderly and detect anomalies. While such sensor-based techniques are effective at detecting immediate exigencies, they are unable to comprehend gradual deterioration in the behavior of the elderly indicating conditions like dementia and Alzheimer's, for example. This aspect is also not properly addressed in the literature. This paper introduces an approach for the comprehensive detection of anomalies in the activities of the elderly using a graph-based approach. The approach employs dynamic activity graphs where anomalies are detected using a dissimilarity score. It is capable of detecting both short-term and long-term anomalies in the daily activities of the elderly.

Keywords: Anomaly Detection · Activity Graph · Graph Matching · Machine Learning

1 Introduction

The world continues to age rapidly, and the number of persons aged above 65 years is expected to reach 1.5 billion in 2050 [14]. Countries are therefore grappling with the issue of facilitating independent living for the elderly. With labour shortage being another issue, substantial investment is directed toward developing automated monitoring capabilities for the elderly. There are numerous approaches in literature for detecting the daily activities of the elderly (and also in general) in indoor settings through the use of wearable sensors [7], ambient sensors [16], and vision sensors [17,18].

Anomalies in the daily activities of the elderly may be classified as 'sudden', including phenomena like falls, and 'behavioral', including variations in the normal routine. Behavioral anomalies are further classified as 'short-term' and 'long-term' based on the nature of the variations. Short-term anomalies include: variations in performing a single activity; variations in the order of performing a set of activities; and variations in the activities of the entire day. Long-term

R. Akerkar (Ed.): SAIDD 2023, CCIS 1810, pp. 136–150, 2024.
https://doi.org/10.1007/978-3-031-53770-7_9

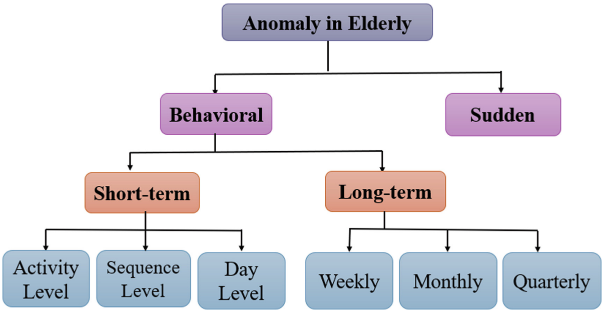


Fig. 1. Classification of anomalies in elderly care.

anomalies, on the other hand, encompass variations in the daily routine of an individual over time. A few examples of long-term anomalies in the elderly are: a person starts spending more/less time in bed; starts using the toilet more often than usual; stops going out; misses important activities often. Some of these anomalies fall in the category of short-term anomalies, but the consequence of such anomalies in the short term is not severe. When these patterns repeat for a longer duration, they have severe consequences and are termed long-term anomalies. A brief taxonomy of anomalies in the elderly is included in Fig. 1.

Approaches in literature are mostly capable of recognizing sudden anomalies like falls [2,11] using different types of sensors. Few approaches analyze changes in behavior to detect short-term anomalies [1,3,8,9,12,15]. The few endeavors in the literature that focus on short-term behavioral anomaly detection take one of the following approaches: activity-wise anomaly detection [8,12] that detects abnormalities in the execution of an activity, i.e., changes in the duration of the activity, or the location where the activity is performed; and sub-sequence wise anomaly detection [3,9] that looks at anomalies in the ordering and sequencing of activities over a short period of time, and day-wise anomalies [1,15] that is also based on ordering and sequencing of activities but over the duration of an entire day. Most of the existing activity-level anomaly detection systems ignore the ordering aspect of the activities; on the other hand, sub-sequence and day-level anomaly detection approaches do not consider the abnormality in the execution of individual activities. Furthermore, they miss out on detecting and comprehending patterns in behavioral change that may not require immediate intervention but can indicate long-term pernicious health issues.

These approaches classify behavioral anomalies either at a fine-grained level and look at individual activities separately or at a more coarse-grained level and focus on the sequence of activities over the whole day. There needs to be, therefore, a robust mechanism that immediately alerts the concerned caregiver for all types of short-term anomalies and also keeps track of long-term anomalies that eventually, over a period of time, indicate a behavioral change and perhaps the build-up of an adverse medical condition. This paper aims at comprehensive anomaly detection over both the short-term and the long term.

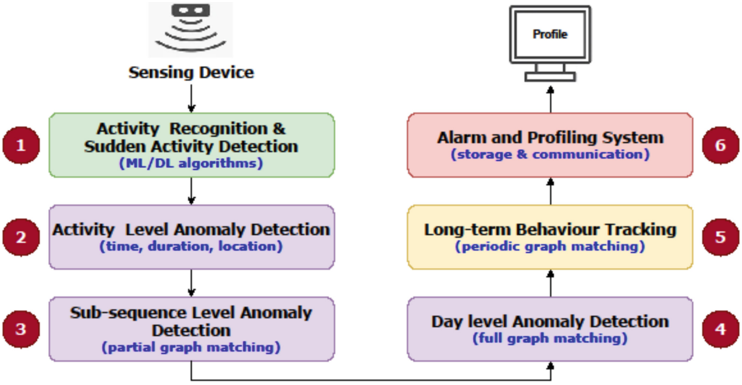

Sensing Device
Profile
1
Activity Recognition & Sudden Activity Detection
(ML/DL algorithms)
2
Activity Level Anomaly Detection
(time, duration, location)
3
Sub-sequence Level Anomaly Detection
(partial graph matching)
4
Day level Anomaly Detection
(full graph matching)
5
Long-term Behaviour Tracking
(periodic graph matching)
6
Alarm and Profiling System
(storage & communication)

Fig. 2. Workflow of the proposed Comprehensive Anomaly Detection System(CADS).

2 Anomaly Detection System

Existing approaches for activity monitoring and anomaly detection, as discussed earlier, look at limited portions of the activity network and are unable to study behavioral patterns comprehensively. Our objective is to develop a complete tool for monitoring the behavioral patterns of the elderly over a short and long duration. The proposed system detects anomalies at the level of: 1) individual activities; 2) within sub-sequences of activities; 3) within longer sequences of activities covering the entire day; and 4) finally, within activity networks spanning multiple days. Such comprehensive analysis and anomaly detection at varying levels of granularity lead to a good understanding of the behavioral patterns of the individual. This facilitates the detection of anomalies in behavior and helps draw sound predictions on potentially adverse medical conditions. The systematic workflow of the proposed system is shown in Fig. 2. The system comprises six modules starting from activity detection and extending up to the point that the system informs concerned personnel about the well-being of the monitored individual.

2.1 Activity Recognition and Sudden Anomaly Detection

The activity detection and recognition module detects and classifies activities from streaming data emanating from various sensors (i.e., wearable, ambient, or vision). The proposed framework works at the activity level and can utilize any of the existing activity detection and recognition systems depending on the sensors involved.

2.2 Activity Level Anomaly Detection

Activities are the basic building blocks of the behavioral patterns of the elderly and describe how an elderly person performs his/her daily activities. Let $A =$

$\{a_1, a_2, ..., a_n\}$ be a set of n activities that a person performs on a daily basis. The daily activity routine is defined as a tuple:

$$R = \{a_i, S_i(t), D_i(t)\} \tag{1}$$

where

- R: Sequence of activities performed on any given day.
- a_i: i^{th} activity of the day.
- $S_i(t)$: Start time of the i^{th} activity.
- $D_i(t)$: Duration of the i^{th} activity.

Each activity a_i is characterized by two features $S_i(t)$ and $D_i(t)$. An activity is classified as normal or anomalous based on significant variations in $S_i(t)$ and/or $D_i(t)$. For example, suppose the normal time of having breakfast for an individual is around 10:00 AM, but on a given day, he/she takes breakfast at 12.00 PM; this should be treated as abnormal. Similarly, the usual duration for taking a bath for a person is 15–20 minutes, but on a given day, the person takes 1 h; this should also be treated as abnormal.

The first step of any anomaly detection system is defining what constitutes a normal pattern. For activity level anomaly detection, the normal pattern is defined separately for each activity based on the available historical data over several days. We harness a One-Class Support Vector Machine (OC-SVM) algorithm [13] to model the normal pattern of activities. OC-SVM is an unsupervised learning approach that creates a hyper-sphere for normal patterns. Figure 3 demonstrate a hyper-sphere of radius 'r' formed by the OC-SVM algorithm using arbitrary data points. The instances belonging to the normal category are shown by green color points, and the abnormal instances are shown in red. The computation of the hyper-sphere is done using Eq. 2 during training.

$$\min_{r,c} r^2 \text{ subject to, } ||\phi(x_i) - c||^2 \leq r^2 \qquad \forall i \in 1, 2, ...n \tag{2}$$

where r and c are the radius and center of the sphere and x_i are the training samples. ϕ is the feature transformation function. A testing instance x_t is classified as anomalous if its distance from the center c is greater than r (i.e., x_t lies outside the hyper-sphere).

The anomaly score for an activity instance is calculated based on the distance of the activity instance from the center of the hyper-sphere as described in Eq. 3.

$$a\_score_t = \varphi(dist(x_t, c)) \tag{3}$$

where dist(*) is a function calculating the distance between a test sample and the hyper-sphere center, and the φ is a function that maps the distance to an anomaly score.

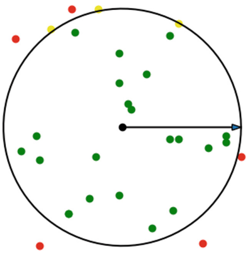


Fig. 3. One Class Support Vector Machine (OC-SVM) hyper-sphere.

2.3 Day Level Anomaly Detection

Activity level anomaly detection considers the deviations in the execution of an individual activity only without looking at its correlations with other activities. We propose a graph-based approach to model the relationships between several activities. Anomalies in the order of performing activities are then detected using graph-matching algorithms.

The daily activities of an individual are represented using a dynamic graph $G_t = (A_t, E_t, A_s, F, W)$ called the Daily Activity Graph (DAcG). The nodes of the graph denote features describing an activity, and the edges from a node denote the probability of occurrence of subsequent activities. G_t is defined as a tuple, where:

1. A_t is the set of activities performed till time 't'.
2. E_t is the set of edges between the activities, denoting the order.
3. A_s is the first activity of the day.
4. F is a feature vector corresponding to each activity, s.t., $F : A_t \rightarrow R^{d_a}$.
5. W is a weight matrix corresponding to each edge denoting transition probability, s.t., $W : E_t \rightarrow eVal$.

The DAcG for several days together indicates the 'normal' behavior of a monitored individual. We call this G_{ref} the Reference Graph of daily behavior. A sample G_{ref} for an elderly person is shown in Fig. 4. The tuple (p, q) on bidirectional edge $E(A, B)$ denotes the transition probability, where p is the probability of 'B after A' and q is the probability of 'A after B'. A probability value 'p=0' indicates that the elderly person never performs activity 'B after A'.

A *day level* anomaly is detected by matching the current DAcG (G_t) with the Reference Graph (G_{ref}). Graph matching is performed using two similarity metrics, namely, Jaccard distance and Hausdorff distance. The adjacency matrix of the DAcG is a matrix W of size $n \times n$, where $W[i, j]$ is the transition probability between activity A_i and A_j.

The Jaccard distance [6] is the measure of dissimilarity between two sets/vectors and is based on the common elements in the two sets. The Jaccard distance between two adjacency matrices X and Y with 'n' nodes is defined

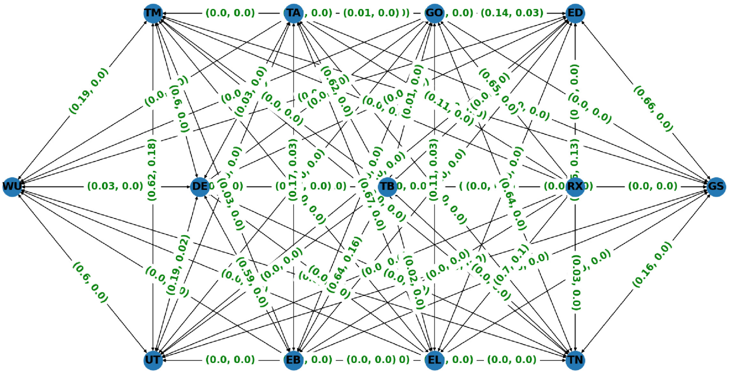


Fig. 4. Daily Activity Graph for Elderly (Each edge A-B is represented by a tuple (p,q), where p is the transition probability from A to B and q is the transition probability from B to A).

in Eq. 4. Jaccard distance is ’0’ for similar matrices and is a good measure of similarity/distance between two adjacency matrices of similar dimensions.

$$j\_dist(X, Y) = 1 - \frac{\sum_{k=1}^{(n\times n)} \min(X_k, Y_k)}{\sum_{k=1}^{(n\times n)} \max(X_k, Y_k)} \tag{4}$$

The Hausdorff distance [4] between two matrices is another useful measure of dissimilarity. This distance is different from the Jaccard distance because it forms a many-to-many mapping between the matrices instead of a one-to-one correspondence. This is also known as the maximum distance between the two. The Hausdorff distance between two adjacency matrices X and Y is defined in Eq. 5.

$$h\_dist(X, Y) = \max_{x\in X} \min_{y\in Y} ||a - b|| \tag{5}$$

Jaccard distance and Hausdorff distance are adept at detecting anomalies that occur due to incorrect ordering and missing activities, respectively. However, these distances are not capable of detecting anomalies due to variations in the execution of individual activities.

An anomaly score is established by combining the distances and anomaly scores of all the individual activities. Equation 6 describes the calculation of the overall anomaly score for a day. A higher anomaly score indicates a greater chance of deviations from the normal behaviour.

$$anomaly\_score = w_1 * \left[\frac{\sum_{k=1}^{n} a\_score_k}{n}\right] + w_2 * j\_dist(G_{ref}, G_t) + w_3 * h\_dist(G_{ref}, G_t) \tag{6}$$

where n is the number of activities in a day, G_t is the DAcG of the current day, and G_{ref} is the reference graph of the normal behavior of an individual.

a_score_k is the anomaly score of the k^{th} activity of the current day. w_1, w_2, w_3 are the weights assigned to each metric. A day is classified as normal or anomalous based on a threshold on the anomaly score. Both weights and the threshold are established during the training process.

2.4 Sub-sequence Level Anomaly Detection

DAcG is dynamic in nature, where a new activity is added as a new node whenever one is detected. A dynamic graph $G_{t'}$ at any time $t^{'}$ is a sub-graph of the Daily Activity Graph G_t, where $t^{'}$ is the time of the day. $G_{t'}$ comprises activities performed till time $t^{'}$ (e.g., $t^{'}$=1.00 PM can include all activities until lunch).

A sub-sequence level anomaly is detected by matching [5,10] a sub-graph $G_{t'}$ with the Reference Graph G_{ref} with the former being classified as anomalous or normal based on the anomaly scores. The sub-graph $G_{t'}$ containing 'k' vertices is first searched in the reference graph and then compared with the corresponding part of the reference graph.

The anomaly score for the sub-graph is calculated in a manner similar to that described in the last sub-section for day-level anomalies.

2.5 Long Term Behavior Tracking

The long-term behavior tracking module is responsible for tracking changes in the behavior of a monitored elderly person based on a comprehensive assessment of his/her daily activities over a period of time (i.e., weeks, months, quarters, or years). This module involves generating periodic graphs and comparing successive periodic graphs to identify changes in the activity patterns, if any. These changes in behavior are analyzed and potentially indicate a gradual deterioration in the health and well-being of the individual.

As mentioned earlier, the DAcG is defined as a tuple $G_t = (A_t, E_t, A_s, F, W)$. The periodic graph $G_{period} = (A_p, E_p, A_{sp}, F_p, W_p)$ for a certain number of days is generated by combining the DAcGs of these days. G_{period} is calculated as described in Eq. 7.

$$\begin{aligned} A_p &= \bigcup_{i=1}^{m} A_{ti} \\ E_p &= \bigcup_{i=1}^{m} E_{ti} \\ A_{sp} &= \min_{i=1}^{m} A_{si} \\ F_p &= \sum_{i=1}^{m} F_i \\ W_p &= \sum_{i=1}^{m} W_i \end{aligned} \tag{7}$$

where 'm' is the number of days in the period (i.e., $m = 7$ for a weekly graph). A_{t_i} denotes the set of activities performed on the i^{th} day. Similarly, E_{t_i} denotes the sequence of the activities on the i^{th} day.

The change in behavior in a given period of time is computed by comparing the periodic graphs G_{period} of two successive periods (i.e., week-1 and week-2). Activity level anomaly scores between the two periodic graphs indicate variations in the execution of different activities. Activity level anomaly scores are calculated as per Eq. 3. The distance matrix between the two periodic graphs indicates the variations in the ordering of activities or the number of missing activities.

2.6 Alarm and Profiling System

The alarm and profiling module primarily works towards establishing a robust system for effective communication of an adverse medical condition to designated caregivers of the monitored individual. The communication may be urgent such that it requires immediate intervention (e.g. anomalies like falls), or it can be routine wherein the caregiver is informed about long-term and gradual behavioral changes in the monitored individual, indicating the possible development of an adverse medical condition. This module enables caregivers and doctors to track the health conditions of the elderly periodically and make recommendations accordingly.

Alerts are sent using existing technologies like automated SMS/calls in case of sudden anomalies and through email in case of short-term and long-term anomalies. Profiles are created based on the results obtained from the Long-Term Behavior Tracking module. The profile includes the changes in individual activities and their ordering.

3 Results and Discussion

The evaluation of the proposed comprehensive anomaly detection system is presented in this section. The details of the dataset, including the set of activities and their attributes, are discussed first. Subsequently, the validation of anomaly detection at various level days is presented.

3.1 Dataset

Most of the existing works on anomaly detection use their own dataset and are not publicly available. The few datasets that are available are very small and do not contain a complete set of daily activities and types of anomalies. Therefore, we created a dataset synthetically containing thirteen daily activities of the elderly. To imitate the real-world behaviour of multiple individuals, the start time and duration of each activity is varied randomly. Our dataset contains daily activity patterns of around 350 normal days and 180 anomalous days. A sample of daily activities for the day is included in Table 1. *WakeUp* is always the

Table 1. A sample of daily activity patterns.

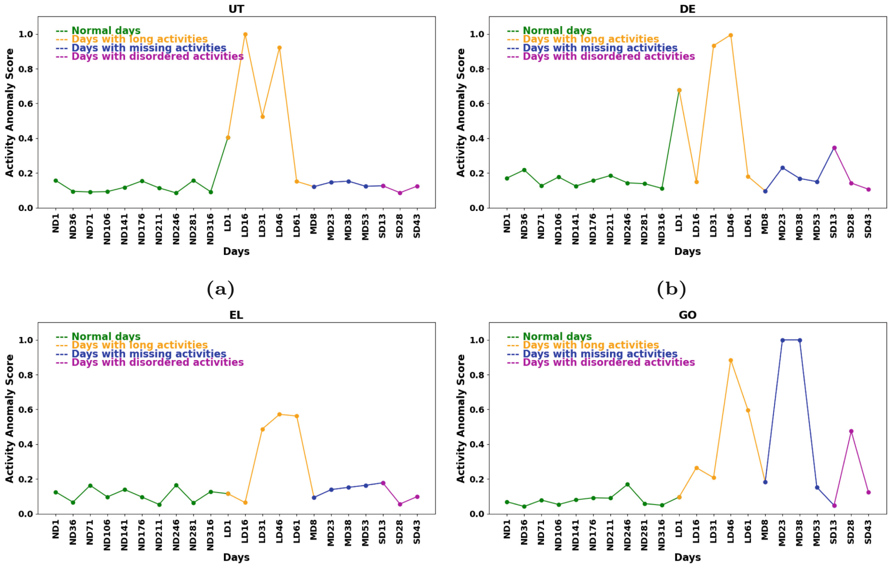
ID	Name of Activity	Activity Abbr	Start Time	End Time	Duration (mins)
1	wakeUp	WU	8:02	8:02	0
2	useToilet	UT	8:05	8:22	17
3	takeMedicine	TM	8:24	8:28	4
4	doExercise	DE	8:35	9:31	56
5	takeBath	TB	9:37	10:14	37
6	eatBreakfast	EB	10:18	11:32	74
7	watchTV	TA	11:35	13:2	105
8	eatLunch	EL	13:25	14:1	45
9	relax	RX	14:11	16:31	140
10	goOut	GO	16:41	18:21	100
11	watchTV	TN	18:26	20:47	141
12	eatDinner	ED	20:47	22:03	76
13	goSleep	GS	23:07	7:37	510

first activity, and its duration is kept at zero. *goSleep* is always the last activity of the day, and the end time of this activity is the next morning.

A random degree of variation is incorporated into the activities on normal days by varying the start time and the end time of the activities. The sequence of some of the activities is also altered, assuming that the elderly can perform some activities in an interchangeable order (e.g., *takeMedicine* and *useToilet*, or *eatBreakefast* and *takeBath*). In the anomalous patterns, the three most common types of anomalies considered are Long Day(LD), Missing Day (MD), and Swapped Day (SD). A long day means one or more activities took a significantly longer/shorter duration than usual, a missing day means one or more activity was not performed on that day, and swapped day means the order of one or more activities is changed, unlike usual. The dataset contains patterns of around 60 d for each type of anomaly.

3.2 Validation of Activity Level Anomaly

The method described in Sect. 2.2 is used to generate the anomaly score for each activity. The daily activity data of normal days is used to train the OC-SVM classifier. Two features, start time and duration corresponding to each activity, are transformed to a similar scale using Min-Max normalization. Before normalization, the start times are converted into numbers. The learned OC-SVM model is then utilized to generate the anomaly scores of the activities of both normal and anomalous days. Each activity is classified as normal or anomalous based on the activity anomaly score. Figure 5 shows the activity anomaly scores of four activities, useToilet, doExercise, eatLunch, and goOut. A small number of days selected from all the days are plotted for better visualization. Figure 5 shows the activity level anomaly score for each activity based on the prediction of OC-SVM classifier trained for the activities of multiple individuals over multiple days.



(a)

(b)

(c)

(d)

Fig. 5. Activity anomaly score of four activities: a) Use Toilet; b) Do Exercise; c) Eat Lunch; d) Go Out

A higher anomaly score indicates more variations in the activity and is termed an anomaly beyond a threshold. Activity level anomaly occurs mostly due to missing activity or significantly longer/shorter duration of the activity. In Fig. 5(a), the anomaly score of the *useToilet* activity is abnormal on LD16 and LD46 due to longer duration and an earlier start-time, respectively. Similarly, the bigger spikes in the *goOut* activity are shown on MD23 and MD38, indicating the abnormality due to the missing *goOut* activity.

3.3 Validation of Day Level Anomaly

Day-level anomalies are detected based on the overall anomaly score of the given day as described in Eq. 6. The weights and the thresholds to classify a day as normal or anomalous are tuned empirically using the training data. Figure 6 shows the overall anomaly scores of all the normal and abnormal days, and the threshold ('Th') separates normal days from the anomalous ones.

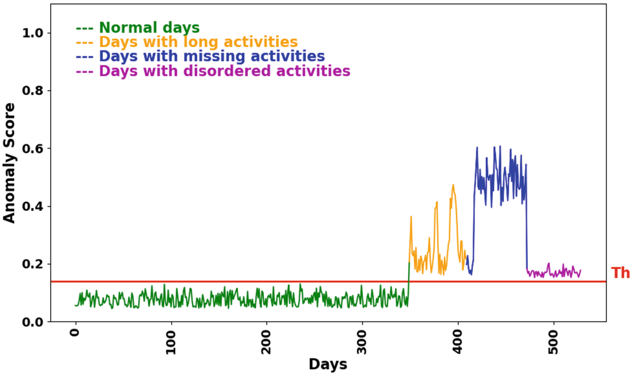
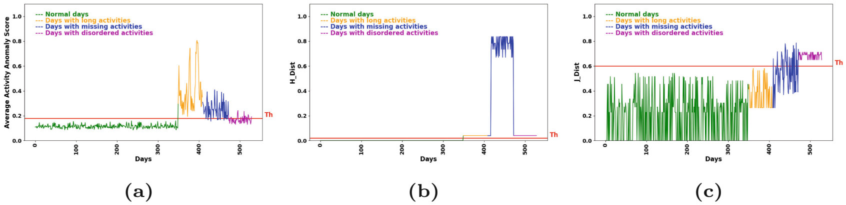


Fig. 6. Overall anomaly score of normal and anomalous days.

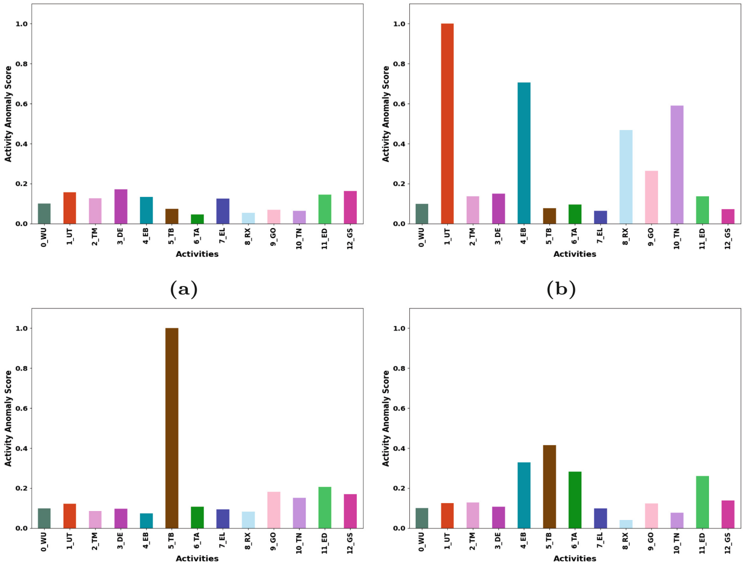
Figure 7 further specifies the types of anomaly that lead to the day being anomalous. Anomaly due to abnormal start time or duration is effectively classified using the average activity anomaly scores as shown in Fig. 7(a). All the anomalous days with longer activities (shown in orange color) are comfortably distinguished from normal days. The H_dist, as given in Eq. 5, is near zero for normal days and has a much larger value for a missing day. Hence, H_dist is an effective indicator of days with missing activities, as shown in Fig. 7(b). J_dist, as per Eq. 4, is high for the days with disordered activities. J_dist comfortably classifies days with unusually swapped activities from normal ones.



(a) (b) (c)

Fig. 7. Categorization of anomalies using separate scores: a) Average anomaly scores of all the activities in the day; b) H_Dist; c) J_Dist

In the case of an anomalous day, the activity anomaly scores of each activity on the given day also provide insights into the cause of the anomaly. A sample day is shown for each category (normal, long, missing, and swapped) in Fig. 8. Activity anomaly scores of each activity on a normal day are smaller, as shown in Fig. 8(a). In Fig. 8(b), the activity anomaly scores of certain activities are high, indicating abnormality in these activities (i.e., *useToilet*, *eatBreakfast*, and *watchTV* at night). In the case of a missing activity, the activity anomaly score is very high, as shown in Fig. 8(c). A set of activities performed in an incorrect order is indicated in Fig. 8(d) for a day with disordered activities. However, the anomaly score is a comparatively weak indicator of a day with disordered activities.



(a)

(b)

(c)

(d)

Fig. 8. Analysis of activity anomaly score of all the activities on a day: a) Normal day; b) Anomalous day where some activities performed abnormally; c) Anomalous day where some activities missed; d) Anomalous day where some activities performed in incorrect order.

3.4 Validation of Sub Sequence Level Anomaly

Sub-sequence level anomaly is a subset of a day-level anomaly and has fewer activities. The anomaly scores for sub-sequence level anomalies are calculated in a manner similar to day-level anomaly but using a subset of activities. To validate the effectiveness of the proposed approach in sub-sequence level anomaly detection, we experimented with different subsets of the activities. For example, we considered only the first three ($k = 3$) activities (i.e., *wakeUp*, *useToilet*, and *takeMedicine*) for any given day. A sub-graph consisting of three activities G_k is then searched in the reference graph G_{ref}, and the selected part of G_{ref} is termed as G_{ref_k}. The average activity score, distance, and anomaly scorse are computed for G_k with respect to G_{ref_k} in a manner similar to that described in Sect. 3.3.

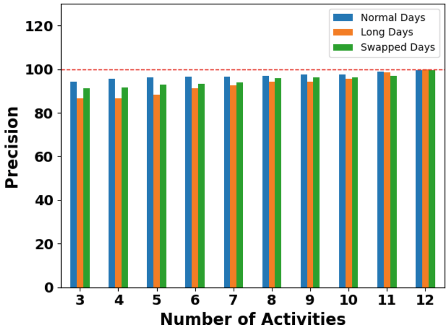


Fig. 9. Precision of anomaly detection on the sub-sequences of different lengths.

Sub-sequences are classified as normal or anomalous based on the anomaly score and other indicators. The precision of the classification for different types of days (i.e., normal, long, missing, swapped) is calculated based on the true class of the sub-sequence. It is important to note that many sub-sequences with the first 'k' activities (i.e., $k = 3$) for an anomalous day can be normal as an anomaly may lie in the upcoming activities. Figure 9 shows the precision of the anomaly detection for the sub-sequences of different lengths. For now, the analysis on missing days is excluded as it is difficult to decide whether the activity is missing or will be performed later in the sequence.

3.5 Results on Long Term Behavior Tracking

Long-term behavior tracking involves the changes in activity patterns over a certain period of time. To simulate the results of long-term behavior changes, both normal and abnormal days are mixed and shuffled. Periodic graphs are created by combining seven normal days, termed normal weeks, and seven anomalous days with similar kinds of anomalies, termed anomalous weeks. Figure 10 shows the behavioral changes in the execution of various activities in four weeks. This contains two normal weeks where the activity patterns are almost similar. A long week was created by combining days with a specific type of anomaly (i.e., *doExercise* activity with a longer duration), and the spike indicates that the *doExercise* activity is performed abnormally in the given week. The missing week is created with days having another specific anomaly (i.e., missing the *eatLunch* activity), and the abrupt change in the plot indicates the missing *eatLunch* activity in the given week. Such information enables the caregivers to introspect into the issues and take necessary action if and when needed. Similar analysis can be done for any period, such as months, quarters, or years.

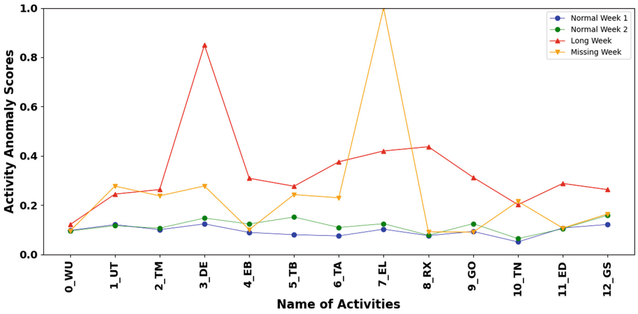


Fig. 10. Precision of anomaly detection on the sub-sequences of different lengths.

4 Conclusion

In this paper, we proposed a comprehensive system for anomaly detection in the daily activities of the elderly. The proposed system is unique because it not only enables the detection of immediate anomalies but also enables analysis and comprehension of the behavioral patterns of an individual over the long term. An understanding of the latter facilitates early detection and diagnosis of pernicious developments like dementia and Alzheimer's. We expect the system to be of utility to the elderly and also assist caregivers in providing better care. As the part of future work, we will create a real-world dataset of daily activities of elderly, including adequate number of elderly. Also, adaptability of the framework for different individuals behaviour can be explored in the future work.

References

1. Azefack, C., et al.: An approach for behavioral drift detection in a smart home. In: 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), pp. 727–732. IEEE (2019)
2. Ding, J., Wang, Y.: A wifi-based smart home fall detection system using recurrent neural network. IEEE Trans. Consum. Electron. **66**(4), 308–317 (2020)
3. Gao, H., Zhou, L., Kim, J.Y., Li, Y., Huang, W.: The behavior guidance and abnormality detection for a-mci patients under wireless sensor network. ACM Transactions on Sensor Networks (2021)
4. Guthe, M., Borodin, P., Klein, R.: Fast and accurate hausdorff distance calculation between meshes. vol. 13, pp. 41–48 (01 2005)
5. Han, M., Kim, H., Gu, G., Park, K., Han, W.S.: Efficient subgraph matching: Harmonizing dynamic programming, adaptive matching order, and failing set together. In: Proceedings of the 2019 International Conference on Management of Data, pp. 1429–1446 (2019)
6. Hancock, J.: Jaccard Distance (Jaccard Index, Jaccard Similarity Coefficient) (10 2004). https://doi.org/10.1002/9780471650126.dob0956
7. Huang, W., Zhang, L., Gao, W., Min, F., He, J.: Shallow convolutional neural networks for human activity recognition using wearable sensors. IEEE Trans. Instrum. Meas. **70**, 1–11 (2021)

8. Parvin, P., Chessa, S., Manca, M., Paterno', F.: Real-time anomaly detection in elderly behavior with the support of task models. In: Proceedings of the ACM on human-computer interaction 2(EICS), pp. 1–18 (2018)
9. Poh, S.C., Tan, Y.F., Guo, X., Cheong, S.N., Ooi, C.P., Tan, W.H.: Lstm and hmm comparison for home activity anomaly detection. In: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), pp. 1564–1568. IEEE (2019)
10. Ravindra, V., Sanders, G., Grama, A.: Identifying coherent subgraphs in dynamic brain networks. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 121–125. IEEE (2021)
11. Sadreazami, H., Bolic, M., Rajan, S.: Contactless fall detection using time-frequency analysis and convolutional neural networks. IEEE Trans. Industr. Inf. **17**(10), 6842–6851 (2021)
12. Saqaeeyan, S., Amirkhani, H., et al.: Anomaly detection in smart homes using bayesian networks. KSII Trans. Internet Inform. Syst. (TIIS) **14**(4), 1796–1816 (2020)
13. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. **13**(7), 1443–1471 (2001)
14. United Nations Department of Economic and Social Affairs, Population Division: World population ageing 2020 highlights (2020). www.un.org.development.desa.pd/files/undesa_pd-2020_world_population_ageing_highlights.pdf
15. Wang, L., Zhou, Y., Li, R., Ding, L.: A fusion of a deep neural network and a hidden Markov model to recognize the multiclass abnormal behavior of elderly people. Knowl.-Based Syst. **252**, 109351 (2022)
16. Yatbaz, H.Y., Eraslan, S., Yesilada, Y., Ever, E.: Activity recognition using binary sensors for elderly people living alone: scanpath trend analysis approach. IEEE Sens. J. **19**(17), 7575–7582 (2019)
17. Yin, J., et al.: Mc-lstm: Real-time 3D human action detection system for intelligent healthcare applications. IEEE Trans. Biomed. Circuits Syst. **15**(2), 259–269 (2021)
18. Zhang, J., Shen, F., Xu, X., Shen, H.T.: Temporal reasoning graph for activity recognition. IEEE Trans. Image Process. **29**, 5491–5506 (2020)

Violence-Inducing Behavior Prevention in Social-Cyber Space

Yasas Senarath[1], Hemant Purohit[1], and Rajendra Akerkar[2](✉)

[1] George Mason University, Fairfax, USA
{ywijesu,hpurohit}@gmu.edu
[2] Western Norway Research Institute, Sogndal, Norway
rak@vestforsk.no

Abstract. Hate speech, radicalization, and polarization in online social environments are some of the leading global societal challenges today. How to respond to online hate speech leading to violence and social threats is a question troubling many democracies – including Norway. Such malicious online behaviors not only impede the universal right to a free and peaceful existence, they also negatively affect response efforts of both public and official agencies during disasters, and thus, local community services require tools to monitor risks to community resilience from the social environment. In this paper, we will elaborate on our ongoing research project "SOCYTI" about developing methods and tools to timely inform local community services for proactive interventions at scale regarding violence-inducing social behaviors by individuals online.

Keywords: Social Cyber Threats · Resilience · Malicious Online Behavior · Social Media

1 Introduction

This paper describes SOCYTI, which is a research and innovation project currently engaging with users, gathering requirements, and writing initial technical specifications and which deals with assessing key behaviors and risks to local communities and mitigating the identified risks by developing advanced and real-time methods for monitoring and preventing such malicious behavior on the social cyber-space. Online violence-inducing behaviors (e.g., hate speech) and information disorder have become a global threat for information integrity and are driving distrust towards individuals, communities, and governments worldwide [1]. Hate speech is not a new phenomenon, but social and technological developments, comprising the persistent spread of social media, malicious social and political discourse, political polarization and deepening economic inequality have determined both an increase in its incidence and the ease with which it spreads. This is an alarming trend that is undermining democratic discourse, fueling discrimination, and stirring violence across the world [1].

Community resilience [2, 3] (the ability of a community to cope with disasters) has increasingly become a priority for local to national governments, especially in the

R. Akerkar (Ed.): SAIDD 2023, CCIS 1810, pp. 151–159, 2024.
https://doi.org/10.1007/978-3-031-53770-7_10

times of COVID-19 pandemic when even a small-scale hazard could threaten the already resource-scarce community/municipality services. Community resilience is defined in many ways in the literature across multiple disciplines [2, 3]; in simple terms, it is the ability of a community to bounce back together [3]. A whole community approach to resilience requires efforts from all the stakeholders of a community including both individual members of the public and government services [4]. Further, research literature on community resilience indicates that an important yet underexplored factor to building community resilience is social connectedness, harmony, and cohesion among community members [2, 5]. Any attempt for community resilience assessment by local community services needs to include the factors for social atmosphere of the community. Some reports show that existing tools to support resilience initiatives of the local community services lack the ability to assess and mitigate the dynamic risks associated with the social atmosphere through malicious behavior online such as spreading hate and misinformation on social media [6, 7]. Such behaviors significantly harm the response efforts of community members during disasters, and thus, local community services require tools to timely monitor risks to resilience from social atmosphere [8]. A variety of large-scale social media datasets, collaborative mapping technologies, and data science approaches have emerged that can facilitate computational social science research to gain a better understanding of community resilience processes accounting for risks associated with social behavior of community members and design actionable tools for local community services. The SOCYTI project will analyze the big data sources at scale, by taking in account all ethical, social, and legal (privacy and data protection) challenges and considerations (e.g., compliance with human rights, respect right of privacy, compliance with data protection principles such as transparency, necessity, and proportionality etc.), in contrast to only existing approaches of small-scale human observations, or survey-based analytical approaches.

In order to achieve the objectives a SOCYTI system will be developed, which will act as a toolset for further analysis.

2 Objectives of SOCYTI Project

The SOCYTI has three main objectives.

First, we aim to analyze the types of violence-inducing social behaviors expressed by the public online that are critical to understanding the risk factors associated with social atmosphere of a community that harm the community resilience, in order to advance the existing community resilience index models.

Second, we aim to infer violence-inducing social behavior from the multimodal and multilingual posts shared by public on social media considering an appropriate trade-off between individual interests (maintaining individual privacy) and the legitimate concerns of achieving and sustaining community resilience.

Third, we aim to design and develop a real-time violence-inducing social behavior detection system for online social media content that could inform the risks to local community services for conducting proactive intervention such as through crisis communication strategies and campaigns.

3 Background and Challenges

Many methods and tools have been developed for the early detection of online malicious activities and actors, for instance, by using natural language processing and social network analysis, or by identifying bots and various patterns [9–11]. But there is a lot more that needs to be done. The SOCYTI is trying to address four most important challenges relevant to the research domain.

Lack of Human Behavior Indicators in Community Resilience Models: There is an increased interest in community resilience by the local government and community/municipality services. Recent studies show that community resilience is a highly local culture-bound phenomenon and related to the kind and frequency of hazards experienced by a local community [2, 3]. Further, research literature on community resilience indicates that an important but underexplored factor to building community resilience is social connectedness, harmony, and cohesion among community members, which have often been seen to help people deal with uncertainty after a disaster [2–5]. However, any threats (e.g., spreading hate speech and disinformation against target demographics online) that break such social harmony present risks to the wholistic community resilience. Thus, any attempt for community resilience assessment should be based on both the historic factors of local hazard experiences in the community as well as the emerging risks in the social atmosphere of the community. Resilience indicators from local to global levels have been developed by the U.S. Federal Emergency Management Agency (FEMA) [12] and the United Nations Office for Disaster Risk Reduction (UNISDR) [13]. New research initiated by the EU Horizon 2020 projects such as Resiloc also addresses the issue of developing resilience methods [14]. However, models or indexes as well as tools developed in these projects and initiatives do not factor in the dynamic risks associated with the human/social behavior of the local community members, especially in the era of social media where public could be manipulated [5, 12]. We will address this challenge of first characterizing the types of risks caused due to the violence-inducing behavior in social media that are harmful for the local community resilience, by taking a social cybersecurity approach. In this process, we will construct extendible artefacts of such characterization of malicious behaviors via an ontology and defining novel indexes with the resulting risks from social atmosphere for community resilience to empower community services.

Limited Studies on Social Cybersecurity and Violence-inducing Behaviors Online: Online content platforms, primarily social networking sites are observing an increasing trend of various types of aggressive behavior such as racism and sexism in the shared content, often manifesting through offensive or malicious language, or multimedia [9, 11, 12, 15]. Such behaviors pose a risk to the civil foundation of our communities, by promoting negative social construction of the diverse social identities such as race and gender that, in turn, divides our society and harms the social connectedness of local community members during disasters [5, 9, 10, 12].

Social cybersecurity has emerged as a new area of applied computational social science research to study malicious online behavior with two objectives [5, 12]: "(a) characterize, understand, and forecast cyber-mediated changes in human behavior and in social, cultural, and political outcomes; and (b) build a social cyber infrastructure that will allow the essential character of a society to persist in a cyber-mediated information environment that is characterized by changing conditions, actual or imminent social cyberthreats, and cyber-mediated threats." In the context of our proposed research, we investigate how to characterize, understand, and detect malicious human behavior with implications to community resilience. While hate speech [1] is one type of relevant social behavior to our research, our aim is to specifically identify the comprehensive set of behavior classes [16], which could be associated to the varied types of risks posed to the community resilience. Such malicious behavior could lead to harmful implications for the local community services and their limited resources for interventions.

There is a growing interest in automatically detecting malicious social behavior online in social cybersecurity field, however, there are a variety of related conceptual definitions investigated in the literature [1, 9–11], such as 'cyberbullying', and 'online harassment'. Schmidt & Wiegand [9] summarizes the diverse definitions of such malicious behaviors studied in the last two decades as follows: *abusive* messages, *hostile* messages or *flames*, *cyberbullying, hate speech*, *insults and profanity*, *offensive language*, and *teasing messages*. Therefore, the challenge to advance the research studying such behaviors in online social spaces from the community resilience perspective has been the definition of what is a violence-inducing behavior with risk implications for local community services. It is, therefore, crucial to develop corresponding novel datasets while preserving the privacy of individuals for building advanced computational methods to detect a comprehensive set of violence-inducing behavior, whether offensive, abusive, insulting, or other malicious behavior, as summarized above. We will focus on the content, rather than individual profile, for analyzing the malicious social behavior patterns. We scope the malicious social behaviors online to the most severe types of violence-inducing behaviors and focus on developing an ontology of such behaviors as well as the automated methods to detect them.

Next, we have identified the following two primary technical challenges that limit the performance of the AI methods for the detection of such malicious behaviors.

Sparsity of labelled datasets and major focus on single language: AI methods for online malicious behaviors are primarily dependent on the labelled datasets for employing supervised or semi-supervised machine learning techniques to infer such behaviors from the online message content. It is difficult to capture all types of malicious behaviors in the labelled data despite the large-scale annotation tasks, due to the sparsity of the presence of such malicious behavior in the content and multiple ways in which these behaviors are expressed. For instance, Founta et al. [10] found the presence of only about 7% content containing the malicious behavior versus normal/harmless behavior in a large-scale annotation task conducted for the social media posts from Twitter. Additionally, recent literature has shown that performance of hate speech detection models trained on one targeted identity does not generalize to other targeted identities [17]. Therefore, datasets that contain a limited set of targeted identities may not be sufficient to train models that are capable of general hate speech detection. Furthermore, it could significantly affect the training process of models for low-resource languages, as the limited data may not represent all target identities. Furthermore, the existing labelled datasets [18] are primarily available for a few languages and only recent efforts include multilingual annotation tasks but focusing only on hateful behavior detection [19, 20]. It is due to the complexity of inferring diverse malicious behaviors from the natural language content that has led researchers to focus on a single language, mainly English [21]. Yet to the best of our knowledge, there is no effort made to develop Norwegian language datasets to encourage the design of AI methods for online violence-inducing behavior detection. This not only requires just an annotated dataset development but rather, a careful multidisciplinary understanding of qualitative context of malicious social behavior in Norwegian culture and society.

Proliferation of multimodal malicious behaviors and lack of AI detection methods: The online content sharing has increasingly become multimodal in nature over the last few years—as per the leading network provider Cisco, online multimedia will make up more than 82% of all consumer internet traffic by 2022 [20]. Thus, we expect the likelihood of malicious behavior being expressed via multimodal content will increase over time. Currently, natural language text is the dominant modality to share content online across geographies, given the ease of sharing text across any online platform whether on social networking sites, news comment sections, or forums. Therefore, the existing literature has primarily focused on exploiting textual content using NLP methods [1, 11] for malicious behavior detection. There is a critical need to design AI methods for exploiting the multimodal content (e.g., Twitter posts with both text and images) to detect diverse malicious social behaviors that could easily get viral across the language boundaries of geography and culture, presenting a greater risk to the local communities across multiple geographical areas.

4 SOCYTI Approach

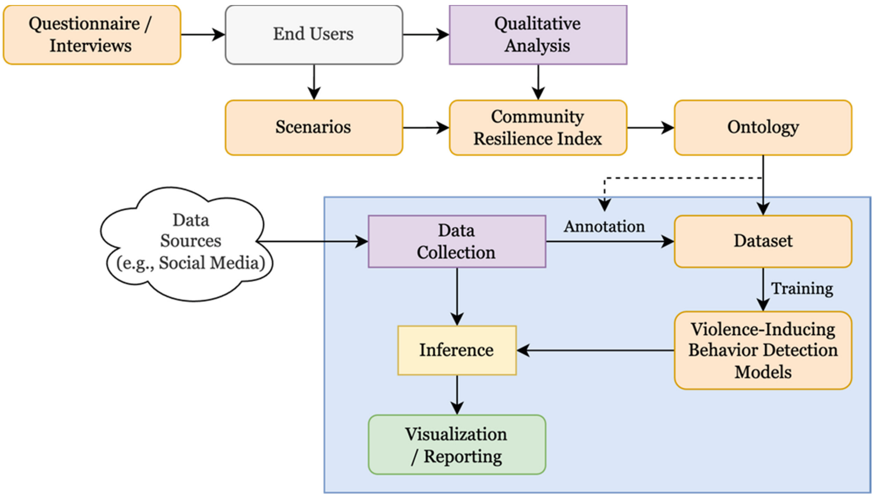

Questionnaire / Interviews
End Users
Qualitative Analysis
Scenarios
Community Resilience Index
Ontology
Data Sources (e.g., Social Media)
Data Collection
Annotation
Dataset
Training
Violence-Inducing Behavior Detection Models
Inference
Visualization / Reporting

Fig. 1. Overview of the SOCYTI approach.

The project aims at reaching its objectives by thoroughly studying various aspects of community resilience involving social connectedness and risks to social atmosphere through big data analytics, AI methods and by innovating on several scientific/technical components. Figure 1 illustrates the overarching approach used in the SOCYTI project. Specifically, we propose the following novel scientific advancement aligned to our three objectives:

Social scientists in SOCYTI team have conducted several questionnaires and interviews with the end users to identify scenarios that drive the next research activities. Furthermore, they have performed a qualitative analysis of the results of the questionnaires and interviews. This analysis and learnings from past research have provided us with the necessary knowledge about the risk factors associated with the social atmosphere to develop a community resilience index and an ontology for violence-inducing behaviors.

Next task involves creating a system to detect potentially violent social behavior from social media posts. More precisely, we are developing models to identify various forms of violent social behavior observed in online social media such as hate speech. As previously noted, there are numerous datasets available for detecting violent social behavior. Unfortunately, no such datasets are available in Norwegian, one of the core languages we aim to support in the SOCYTI project. Hence, we will undertake the annotation of social media data in the Norwegian language to broaden the linguistic coverage of our proposed system. This will enable us to train and assess our multilingual models. To do this, we have gathered social media data from Twitter posts that contain keywords associated with hate (as identified in Hurtlex [22]). We will label these posts as either hateful or non-hateful through a combination of manual review and predictions generated by a Large Language Model (LLM) [23].

Next, we will train hate speech detection models with the help of all collected datasets. Once we've trained these models, we'll apply them for ongoing, real-time detection of hate speech in online social media. Following that, we will employ visualizations and reports to present a more comprehensible summary of these real-time instances of violence-inducing behaviors. This will empower the local community services with a real-time violence-inducing behavior system for online social media content. It allows better risk management for the public safety community services to improve their crisis communication strategies and campaigns to counter violence implications.

5 The Proposed SOCYTI System

Our team proposed to develop the SOCYTI system as a toolset comprising various components for detecting and monitoring violence-inducing social behavior, specifically through hate speech detection methods as the foundation. Figure 2 illustrates the high-level architectural approach proposed for the system design.

We note that detecting hate speech using static models has the disadvantage of not being robust to domain shifts particularly when behaviors such as hate speech are targeted at different identity groups [17]. By leveraging knowledge bases (such as WordNet) we can augment the existing datasets to contain a substantial number of training examples for all identities. For example, we can find synonyms of a given identity from WordNet and

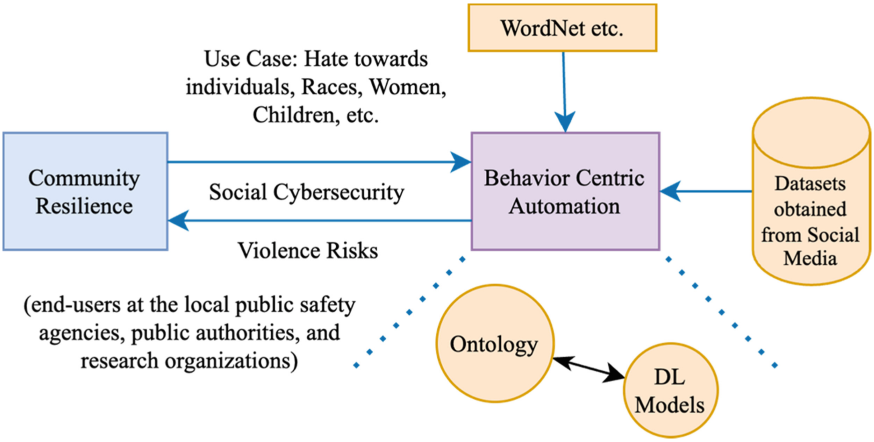

WordNet etc.
Use Case: Hate towards individuals, Races, Women, Children, etc.
Community Resilience
Behavior Centric Automation
Social Cybersecurity
Violence Risks
Datasets obtained from Social Media
(end-users at the local public safety agencies, public authorities, and research organizations)
Ontology
DL Models

Fig. 2. The architectural approach to SOCYTI.

replace the occurrences of the identity term in the examples with its synonyms to create more examples of similar ground truth label. By such augmentation, we aim to enhance the performance of hate speech detection for all identities. Nonetheless, a limitation of WordNet is its static nature. To address this issue, we will develop our own knowledge base, which supplements WordNet with additional information obtained through automated data mining techniques, leveraging the assistance of large language models. Furthermore, with multilingual WordNet and multilingual language models, we can extend the aforementioned process to support multiple languages. Here, a multilingual knowledge graph such as Open Multilingual Wordnet (OMW) [24], and DBpedia [25] provide cross-lingual interpretations for similar entity mentions in multilingual social media content, which improves explicit context representation (e.g., a religious group of a violence-target mention in a text) to let the algorithms learn patterns of violence-inducing behavior across languages efficiently. Furthermore, we will use those knowledge bases and continuous training of models to adapt to evolving social discourse. Apart from the textual content related to violence-inducing behavior, algorithms can perform better for multilingual online content when exploiting the multimodal information from both text and multimedia objects, which helps augment feature space and capture the context of malice more effectively than the approach of inferring the behavior using features from only text.

Furthermore, the SOCYTI system will provide customizable components for detecting violence-inducing behavior expressed through hate speech in different modalities of data including text and multimedia data. To facilitate multimodal hate speech detection, our proposal involves training a model that encodes various types of inputs, such as images, into their respective vector representations. These representations are then concatenated, and classification is performed on the combined input.

6 Conclusion

In conclusion, this work provides an overview of ongoing SOCYTI project, outlining its scope and objectives, and emphasizing the significance of hate speech detection in reducing violence-inducing behaviors within online communities. We also explore the background literature for this research and highlight the limitations and challenges associated with the existing methods. Lastly, we explain the approach employed in the SOCYTI project and provide insights into the proposed hate speech detection system that will be utilized for identifying violence-inducing behaviors in social media.

Acknowledgement. The SOCYTI project is owned by Western Norway Research Institute and has received funding from the Research Council of Norway as a Researcher Project for Technological Convergence related to Enabling Technologies under grant agreement no 331736.

References

1. Kursuncu, U., Purohit, H., Agarwal, N., Sheth, A.: When the bad is good and the good is bad: understanding cyber social health through online behavioral change. IEEE Internet Comput. **25**, 6–11 (2021)
2. Nguyen, H.L., Akerkar, R.: Modelling, measuring, and visualising community resilience: a systematic review. Sustainability. **12**, 7896 (2020)
3. Patel, S.S., Rogers, M.B., Amlôt, R., Rubin, G.J.: What do we mean by'community resilience'? A systematic literature review of how it is defined in the literature. PLoS Curr. **9**, (2017). https://pubmed.ncbi.nlm
4. FEMA: Whole community approach to emergency management: Principles, themes, and pathways for action. Fed. Emerg. Manag. Agency US Dep. Homel. Secur. Wash. DC (2011)
5. Aldrich, D.P., Meyer, M.A.: Social capital and community resilience. Am. Behav. Sci. **59**, 254–269 (2015)
6. Zadrozny, B., Collins, B.: West Coast officials are already fighting wildfires. Now they're fighting misinformation, too. https://www.nbcnews.com/tech/security/wildfires-rage-false-antifa-rumors-spur-pleas-police-n1239881. Accessed 22 Sept 2023
7. Centre for the new economy and society: Chief risk officers outlook (2023). https://www3.weforum.org/docs/WEF_Chief_Risk_Officers_Outlook_2023.pdf
8. Hunt, K., Wang, B., Zhuang, J.: Misinformation debunking and cross-platform information sharing through Twitter during hurricanes Harvey and Irma: a case study on shelters and ID checks. Nat. Hazards **103**, 861–883 (2020)
9. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pp. 1–10 (2017)
10. Founta, A., et al.: Large scale crowdsourcing and characterization of twitter abusive behavior. In: Proceedings of the International AAAI Conference on Web and Social Media (2018)
11. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. CSUR. **51**, 1–30 (2018)
12. FEMA: FEMA Strategic Plan 2014–2018 (2014)
13. UN General Assembly: The Sendai framework for disaster risk reduction 2015–2030. UN Gen. Assem. Geneva Switz. (2015)

14. Akerkar, R., Nguyen, H.L.: Resilient Europe and societies by innovating local communities (Resiloc) (2019). https://www.vestforsk.no/en/project/resilient-europe-and-societies-innovating-local-communities-resiloc
15. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media, pp. 512–515 (2017)
16. Purohit, H., Pandey, R.: Intent mining for the good, bad, and ugly use of social web: concepts, methods, and challenges. In: Agarwal, N., Dokoohaki, N., Tokdemir, S. (eds.) Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining. LNSN, pp. 3–18. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-94105-9_1
17. Yoder, M., Ng, L., Brown, D.W., Carley, K.M.: How hate speech varies by target identity: a computational analysis. In: Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL), pp. 27–39 (2022)
18. Vidgen, B., Derczynski, L.: Directions in abusive language training data, a systematic review: garbage in, garbage out. PLoS ONE **15**, e0243300 (2020)
19. Basile, V., et al.: Semeval-2019 task 5: multilingual detection of hate speech against immigrants and women in twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 54–63 (2019)
20. Cisco: Cisco visual networking index: forecast and trends, 2017–2022 white paper – cisco (2020). https://web.archive.org/web/20200215211855/https:/www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html
21. Vitiugin, F., Senarath, Y., Purohit, H.: Efficient detection of multilingual hate speech by using interactive attention network with minimal human feedback. In: Proceedings of the 13th ACM Web Science Conference 2021, pp. 130–138 (2021)
22. Bassignana, E., Basile, V., Patti, V., et al.: Hurtlex: a multilingual lexicon of words to hurt. In: CEUR Workshop proceedings, pp. 1–6. CEUR-WS (2018)
23. Wang, S., Liu, Y., Xu, Y., Zhu, C., Zeng, M.: Want to reduce labeling cost? GPT-3 Can help. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 4195–4205 (2021)
24. Bond, F., Paik, K.: A survey of wordnets and their licenses. In: Proceedings of the 6th Global WordNet Conference (GWC 2012), pp. 64–71 (2012)
25. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52

Artificial Intelligence in the Public Sector in Norway:

AI Development as a Hop-on-Hop-off Journey

Hilde G. Corneliussen[1](✉), Gilda Seddighi[2], Aisha Iqbal[3], and Rudolf Andersen[3]

[1] Western Norway Research Institute, 6851 Sogndal, Norway
hgc@vestforsk.no
[2] NORCE, Nygårdsgaten 112, 5008 Bergen, Norway
[3] Rambøll Management Consulting, Harbitzalléen 5, 0275 Oslo, Norway

Abstract. This paper presents a study of the use of artificial intelligence (AI) in the Norwegian public sector. The study focused particularly on projects involving personal data, which adds a risk of discriminating against individuals and social groups. The study included a survey of 200 public sector organizations and 19 interviews with representatives for AI projects involving personal data. The findings suggest that AI development in the public sector is still immature, and few projects involving personal data have reached the stage of production. Political pressure to use AI in the sector is significant. Limited knowledge and focus on AI development among managements has made individuals and units with the resources and interest in experimenting with AI an important driving force. The study found that the journey from idea to production of AI in the public sector presents many challenges, which often leads to projects being temporarily halted or terminated. While AI can contribute to the streamlining and improvement of public services, it also involves risks and challenges, including the risk of producing incorrect or discriminatory results affecting individuals and groups when personal data is involved. The risk of discrimination was, however, not a significant concern in the public sector AI projects. Instead, other concepts such as ethics, fairness, and transparency took precedence in most of the project surveyed here.

Keywords: Artificial Intelligence · Public Sector · Discrimination

1 Introduction

Artificial intelligence (AI) has made significant progress in recent years. It has been identified as one of the most important technologies of the 21st century [1], expected to have a major impact on solving small and large societal challenges with an effect for private and public sectors and for individuals [2]. AI is considered an important tool for improving services and making the Norwegian public sector more efficient [3]. Adopting AI, however, also entails challenges and risks. One of the most worrying risks for citizens is the risk of incorrect, unfair, or discriminatory results when AI is used in public services to draw conclusions based on personal data.

R. Akerkar (Ed.): SAIDD 2023, CCIS 1810, pp. 160–172, 2024.
https://doi.org/10.1007/978-3-031-53770-7_11

This paper reports from a study of AI projects in the Norwegian public sector. A survey helped mapping AI projects in the public sector. The survey included questions about what challenges and possible risks AI projects faced, and about efforts to prevent potential risks, in particular regarding unfair or discriminatory treatment of individuals. Through follow-up interviews with AI projects involving information about individuals and the use of personal data, we further explored in-depth the experiences involved in developing AI systems in the public sector, including a special focus on how the risk of discrimination was perceived and dealt with.

In this paper we present the results of this study, illustrating several challenges, obstacles and stumbling blocks that AI projects in the public sector had experienced. Considering these challenges in relation to requirements for AI projects, we use the metaphor of a *hop-on-hop-off journey* as a way of understanding the public sector organizations' experiences.

We start with a literature review of relevant aspects of the public sector engagement with AI before describing the methodological framework and the empirical data. Then we present the findings with an emphasis on the elements producing barriers, or "stops", on the hop-on-hop-off journey.

2 Literature Review

The public sector in Norway is working purposefully to contribute to the development and improvement of public services in an efficient and sustainable manner. Digitalization and new technologies including AI are important tools to achieve this. AI can make a significant contribution for the wish of creating public digital services tailored to users' needs [3]. Used at its best, AI can contribute to the streamlining and improvement of public services, to make services easier to use and to reduce unfair differentiation in services offered to different social groups. There are good examples of the use of AI in Norway, for instance, in the healthcare sector. The benefit of using AI is emphasized and further development encouraged in political strategy documents such as the National Strategy for Artificial Intelligence [4]. The national strategy points out that Norway has the socio-technical infrastructure in place for succeeding with AI. This includes elements such as a high level of public trust in authorities and the public sector in Norway, a high degree of digital competence in the population, a well-developed technological infrastructure, and a public sector that has come a long way in developing a digital administration. Registry data developed over a long period of collecting data, including personal data, can provide important access to data for developing AI for citizen services [5]. The national AI strategy also emphasizes that public organizations "have the capacity and competence to experiment with new technologies," which can be crucial for the public sector organizations' ability to adopt new technology, such as AI.

However, the risks for errors, unfair, or biased results of AI systems that have concerned researchers the last decades [6–12], also concern the public sector in Norway. National and international initiatives are developing guidelines for minimizing such risk [4, 13–16]. One step on the way is regulation of the use of personal data through the European Union's framework GDPR – the General Data Protection Regulation, which has been in work since 2018 in Norway. The introduction of GDPR received massive attention and succeeded in putting the issues of protecting and limiting storage and use of personal data on the agenda, according to the Norwegian Data Protection Authority's yearly report of 2018 [17]. The overarching principle of GDPR is to store as little personal data for the shortest time possible [18]. This principle, however, is not in line with requirements for developing high quality AI systems, which rather requires large amounts of data. Such paradoxes are the subject of assessment and prioritization, and the recent EU AI Act aims to balance these considerations. A draft for the EU AI Act was approved in June 2023, pointing towards the first European law on AI.[1] The AI Act aims to regulate the use of across sectors in terms of risks it poses. The use of AI in ways that are considered a threat to people, such as biometric surveillance, emotion recognition, and predictive policing, is banned as an "unacceptable risk". High-risk application of AI that involves using personal data, for instance for recruitment or for providing services, will be regulated and subject to specific legal requirements. AI systems that do not belong in the two first categories are mainly left outside this regulation [16, 19].

As reflected in the EU AI Act, a critical risk of AI producing discriminatory results for social groups and individuals is related to the use of personal data in an AI system. While this is a risk in private as well as public sectors, AI involves a particular set of challenges for the public sector, which relies on accessing a large amount of personal data in order to deliver its services to the citizens. The public sector has a particular obligation to provide good and fair services [3]. Compared to many other western countries, Norway has a large public sector managing public resources and providing vital services that depend on sufficient and updated information about citizens [20]. Therefore, the Norwegian government and public sector have a unique access to a large amount of data about citizens that has been collected over time. This data includes a wide set of information about citizens, from sensitive personal data, health data, to data about employment, education, and finances [5, 20]. While the Norwegian population has a high level of trust in the public sector [21], responsible use of technology, including AI, is crucial for maintaining the trust-based relation between citizens and authorities and, ultimately, for democracy [22]. Experts have expressed concern that the public sector's access to large amounts of data, along with increased digitalization, "can create pressure to use data in new and intrusive ways," and see a risk that the development may point towards greater control, for example, through an increasing use of AI to "uncover fraud and deception" [12, 23]. While negative effects of AI are often unintended [24], clear goals and strategies are also necessary to avoid a development towards "public surveillance capitalism" [12, 25].

United Nations has warned about the combination of AI with large amounts of data about citizens collected through government agencies, referring to this as "a digital welfare dystopia" [23]. Broomfield and Lintvedt claim that also in Norway there is a risk of stumbling into this digital welfare dystopia [12], despite the Norwegian public sector organizations being subjects to strict laws and regulations for handling personal data. This is particularly critical as public organizations are obligated to ensure equality for a diverse population in public services. Equally critical for the public sector organizations is the fact that their customers, i.e. the citizens, cannot choose another service provider. Thus, while market forces might punish private companies that produce unfair results

[1] The EU AI Act was approved after the data collection in our study had been concluded.

for their clients, there need to be other mechanisms, for instance an auditing system, to secure fairness across public sector services.

In this study we aimed to learn more about challenges as well as potential risks of using AI in the public sector, and in particular how potential risks of discrimination were dealt with to ensure the fairness and high level of trust expected by the public sector.

3 Methods

The main aim of this study was to map the use of AI in the public sector, and to learn more about challenges and risks associated with AI projects in the sector. AI is not operating in isolation, therefore, it must be understood within the context in which it operates. In order to capture and understand the relationship between society and technology, and between opportunities, challenges, and risks, in particular related to discrimination, we worked in a cross-disciplinary team including sociological, technical, and feminist perspectives.

We also engaged a mixed methods strategy that involved initial meetings with AI experts from public and private sector as well as from academia and NGOs. Based on literature review and these meetings, we developed a survey to map the situation, before inviting a selection of public sector organizations to participate in in-depth interviews.

3.1 Online Survey

The online nation-wide survey was sent to nearly 500 governmental and municipal organizations from various sectors including health care, education, labour and welfare administration, tax authorities and more. The main goal of the survey was to map the status for plans and projects related to AI in the public sector. The survey included questions about aim of using AI in the organization, competence involved, challenges encountered in relation to access and use of data, competence about technical, organizational, and juridical issues, and how risks of discrimination were perceived and dealt with. The response rate was 40%, with 200 organizations responding to the survey. Among the 200 received responses, 59 organizations had active projects or plans for using AI, and out of these, 39 AI projects involved the use of personal data.

3.2 Interviews

The next phase involved an in-depth study including interviews with public sector organizations that had AI projects and plans, identified through the survey. Here we were mainly interested in the projects involving personal data, as that brings up some of the special issues for public sector as a service provider for citizens and raise critical questions regarding risks of discrimination when engaging AI in the public sector. The interview questions focused on the concrete AI projects and plans and the organizational setting; technical, organizational, and juridical competences and challenges; and risks of discrimination and mitigation of such risks.

We invited the organizations that had AI projects or plans involving personal data to participate in interviews. A total of 19 interviews were organized with organizations representing a wide spectrum of governmental, municipal, and inter-municipal organizations of different sizes and localized in different regions of the country.

In the invitation for the interviews, we encouraged the organizations to include a team of individuals with diverse roles in their AI project. The informants included 18 men and 9 women. Some were managers, while the majority had a professional background in information technology and AI.

3.3 Analysis

The survey was used to map the status of AI projects and to identify which issues that appeared most challenging for the public service organizations. The qualitative interviews allowed us to explore more in-depth the challenges and risks experienced in the AI projects.

In the qualitative analysis we used an image of an AI project lifecycle based on relevant research literature emphasizing different stages of an AI project, various aspects of maturity and requirements of such a lifecycle from start to finish [26–28] to map and understand the development of the projects. Due to how the AI projects had started, developed, and temporarily or permanently ended, many of them before reaching the expected end point, we pictured this as a journey with a set of stops. The metaphor of a journey guides the presentation of the main findings below. We share some of the findings from the survey, while the main part of the analysis is based on the in-depth interviews focusing on the projects that involved personal data.

4 Findings

The National AI Strategy [4] encourages public sector to develop AI as a tool supporting decision making and as a part of public digital services, including the interaction between government bodies and citizens. The survey mapping AI activity in public sector showed that many organizations are currently exploring how AI can be used to improve and make public services more efficient. The interviews further showed that the motivation varied from a curiosity about what technology can accomplish, to an argument of spending taxpayers' money wisely. This could be interpreted as a duty to explore how to become more efficient with the use of new technology such as AI, one informant pointed out.

The National AI strategy employs a wide definition of AI, from machine learning (ML) to automatised processes. We started the survey with a similar wide definition to let the respondents decide what they included in a definition of AI, thus, among the AI projects involved in this study were both ML and simple automated procedures.

The survey showed that many organizations had plans for using AI. However, less than 20% of the organizations responding included personal data in their AI project. These AI projects had a variety of goals, from improving the quality of data, detecting suspicious patterns and errors in the data, predicting needs in the organization or users' behaviour, and more. Some of these projects had been initiated to explore the possibilities of AI for the services, or for testing AI models.

Most of the AI projects involving personal data were, however, still in an early stage, exploring and developing the possibility of using AI, and only a handful of these projects confirmed that they had reached the final stage of being in production. Many of the projects had encountered challenges, which will be further elaborate below.

4.1 AI Development as a Hop-on-Hop-off Journey

The survey as well as the in-depth interviews showed that the journey from idea to production of AI in public sector presents many challenges that often lead to AI projects being temporarily halted or terminated. The result was that only a few AI projects operating on personal data had reached the production stage, however, their stories involve important lessons. Figure 1 illustrates identified challenges of AI projects as a series of elements that need to be in place to safely navigate from start to end, starting with the design of the project to the end of the journey where the AI system is in use, or "in production".

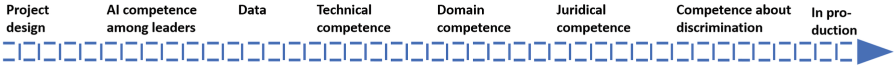


Fig. 1. A hop-on-hop-off journey of AI projects.

Figure 1 illustrates the development of an AI project with the different elements representing stops on an imaginary hop-on-hop-off journey. The elements of Fig. 1 do not represent a perfect or necessary chronological development.

4.2 Project Design

Project design is an important stage of an AI project where crucial decisions and choices are made. However, not all the AI projects had started with an overall design or management strategy anchored at the organizational level. In several cases the origin of the project had been tech people finding space and resources for exploring AI in context of the organization.

4.3 Leader Engagement and Competence

Most projects had a clearly defined goal and objective within the organization. However, the previous examples of exploring AI as the main driving force illustrated a weak leadership involvement in some organizations. AI is still a new technology for most public sector organizations, and it is dependent on those who want to participate, one of the informants told us. Similar to other studies among leaders in Norway [29], most respondents to the survey agreed that there was a need for more knowledge about AI among leaders.

4.4 Data

Data is a crucial element of AI, and Gröger claims that there will be no AI without data [30]. Data, however, also introduced a number of challenges for the public sector organizations, including navigating the European General Data Protection Regulation (GDPR) and Norwegian juridical frameworks as well as known and unknown bias in data [31, 32]. While many of these challenges are not specific to the public sector, some of them are, in particular due to the position and responsibility that the public sector organizations have as service providers for citizens. This, for instance, includes the public sector's access to a huge amount of data collected from the service recipients. While this suggests that the public sector has access to valuable data that would be useful for developing good and precise AI systems, the juridical framework for public sector challenges this use of the data. The Norwegian government established a regulatory sandbox for AI systems in 2020, under the Norwegian Data Protection Authority [33]. One of the early public sector organizations to be assessed through the sandbox was the Norwegian Labour and Welfare Organization (NAV), which has access to a large amount of personal data from the service recipients. The concluding report suggests that NAV has the right to use such data in an AI system, however, using the same data for developing and training the AI system was questioned. The concluding report points to the need for developing laws that also take into account a responsible development of AI projects in the public sector [34].

4.5 Technical Competence

Access to technical expertise varied with the size of the organization. Norway has many small and medium sized municipalities that do not have the same access to technical expertise as the bigger municipalities have. The long-stretched country also adds challenges for rural organizations to access networks of competence gathered in and around the bigger cities. Access to technical competence influenced whether the AI system in question had been developed in-house or bought from external companies, sometimes specially designed, sometimes referring to shelf-ready AI algorithms and systems. This could introduce doubt regarding what kind of data the models were based on, and thus doubt to whether the models would fit the organization's needs. This could also make it more challenging to assess the risk of discrimination from a particular AI system.

4.6 Domain Competence

It might seem like an obvious statement that digital systems must meet the requirements of the organization implementing them, implying that a certain level of domain competence is necessary. Some of the projects that had bought or received external tech support, had, however, experienced that lack of domain competence and insights into the organization's practical and legal requirements had introduced weaknesses and challenges to the AI system. Here we also saw how lacking multi-disciplinary competence could result in an unclear situation, for instance by making it difficult to judge whether the AI system complied with objectives of the organization and relevant regulations.

4.7 Juridical Competence

Juridical competence is vital for establishing how personal data can be used in an AI project. While this is true for any AI project, it includes an extra set of regulations for public sector organizations that depend on collecting personal data from its clients. In general, permission for such data collection is strictly limited to the kind of data needed for providing services. Thus, as the case of NAV above illustrates, some of the relevant laws and regulations are challenged by new digital technologies, such as AI, that require public organizations to consider a new set of juridical questions. Some of the informants had experienced that failing to involve juridical competence at an early stage of their AI project had led to an abrupt stop. It was typical for such cases that the AI project had started without an overall design for mapping the socio-technical opportunities and requirements. Others struggled with strict interpretations of GDPR and risk assessments, and for some, this was perceived as barriers for exploring AI.

4.8 Competence About Discrimination

Finally, competence about discrimination is not positioned as the last stop before the end point of this AI development journey because it is the least important, but rather because few of the AI projects had engaged with questions of discrimination. The public sector is required to provide similar services for all citizens. Thus, considering the risk of discriminating between groups or individuals when introducing new technical systems including AI, is crucial. Research has shown that when using personal data in AI systems, it is necessary to deal with known as well as unknown biases in the data [31, 32]. Historic as well as recent data might portray different social groups' education, employment, social positions, economy and more, in ways that can appear discriminating if they were to be reproduced. These issues have been illustrated in examples using AI in recruitment processes in which, for instance, current gender imbalance in working life has been turned into preferences for candidates [35].

Thus, even when everything is "correct", bias in data can still produce wrong, unfair, and discriminating results [6]. In our survey, however, only 3% of the respondents believed that AI would increase the risk for discrimination. In the interviews we found that few of the organizations had engaged with this topic: "We haven't thought about that, thank you for reminding us", one informant said. We also found a tendency for other considerations and concepts such as AI ethics, fairness, and transparency, or privacy taking precedence while discrimination in line with the definition of the Norwegian Anti-Discrimination Act as unwanted discrimination, was a topic only for a handful of AI projects. In some cases, we found that discrimination was translated into a technical concept of "differentiation". This made discrimination appear as natural and as a wanted discrimination, quite different from the Anti-Discrimination Act, which aims to prevent unwanted discrimination. While some of the alternative concepts that were introduced when we asked about discrimination reflect important discussions of AI [36, 37], they also move the question further away from the issue of preventing unwanted discrimination and consequently weaken the perceived need for dealing with these issues. The interviews demonstrated that lack of knowledge about discrimination resulted in little attention devoted to the risk of discrimination with AI.

4.9 In Production

The many hurdles and barriers experienced in the development phase of the AI projects in the public sector had resulted in only a handful of projects involving personal data reaching the end point of putting AI in production. The interviews illustrated that also this stage could introduce challenges for the AI system. In this phase, however, the challenges were more directly affecting users, for instance, a risk of differentiating users based on digital competence, and posing a threat towards the citizens' trust to the public sector if something was unclear or error happened.

5 Discussion

The analysis above focused on the public sector organizations with ongoing AI projects that involved personal data. The findings illustrate that the AI development is still in an early stage in large parts of the Norwegian public sector. The development seems in some cases to be less driven by management and more by interest and willingness to experiment with the technology. In addition, there is a notable political pressure to use AI in the public sector, which was mentioned several times in the interviews. Thus, we can identify three levels of driving forces for the current AI development in the Norwegian public sector:

- The political level with strategies for digitalization in the public sector and for AI in Norway created expectations about the use of AI [3, 4].
- The management level: The management level in the public sector, with some exceptions, was considered to have limited knowledge of AI and a limited focus on establishing and directing the development of AI in the sector.
- Individuals and units within each organization that expressed a willingness to contribute to AI development were important for initiating AI projects.

Thus, the political level contributed to pushing AI into the sector, the management level was less visible in many of the AI projects, while individuals and units with knowledge about and the resources to experiment with AI were important driving forces in many organizations. This indicates that the public sector is still in an early stage of exploring and learning about AI, also reflected in other studies [5, 38].

Another critical issue for engaging AI in the public sector was the confusion arising from the wide definition of AI in the National AI strategy, involving everything from simple automation to machine learning (ML) techniques. Only automated procedures can be used for decision making in public sector, while AI techniques involving ML can only be used for supporting decisions where humans have the final say. Putting these different technologies in the same pot creates a confusion both inside and outside the public sector, about whether it is technology or humans making the decisions.

This also highlights the importance of interdisciplinary competence for successful AI projects, as these projects do not operate in a digital vacuum but must interact with a range of different social, cultural, political, and legal rules and regulations [13, 37].

Many of the challenges we found revolve around paradoxes that arise when AI is introduced in the public sector. Firstly, the public sector has different requirements for accuracy in services compared to the private sector, as the Public Administration Act requires all individual decisions to be justified [4]. Therefore, AI in the public sector must be transparent and explainable to avoid producing errors. Such issues are further complicated by the fact that not all decisions in the public sector warrant full transparency, especially concerning the government's control functions. This creates room for interpretation and discussion, illustrating that regulatory guidelines and legislation do not guarantee a shared understanding among all parties involved. Secondly, the regulations and legal framework are not well-adapted to the digital reality, leaving many questions unanswered or open to multiple and conflicting interpretations. Thirdly, the different regulations that intersect in this field are partly in conflict with each other. The overarching principle in GDPR legislation is to store as little data for the shortest period possible, while AI requires a significant amount of data, sometimes more than the original registration of data, if discrimination is to be avoided. Such paradoxes become subject to assessment and prioritization: Which carries more weight, development, and efficiency on one side, or the risk of errors on the other side?

While there were many challenges and barriers for the public sector AI projects, a reflection on the risk of discrimination was not particularly prominent in the landscape of challenges described by the public sector organizations. Other concepts such as AI ethics, fairness, and transparency appear to take precedence before discrimination. The concept of discrimination functions as what theorists Laclau and Mouffe refer to as a "myth," as a term we can discuss together without having agreed on a specific definition, thus we can put different content into it [39]. "Bias" was often discussed, usually as unconscious bias, while the term "discrimination" was less frequently used in this field. If questions of discrimination are translated into alternative concepts that make it appear less of a problem, it is less likely to be addressed as a challenge. The lack of focus on unwanted discrimination reflects a gap to be filled by future policy and practice in the public sector.

There is an increasing number of guidelines, frameworks, and models aimed at countering AI from producing biased, unfair, or harmful outcomes [1], and more are currently being developed. Many of these resources target technologists who are responsible for the technological development of an AI system [6]. However, AI researchers warn that AI developers alone should not be held accountable for the broad set of considerations that need to be made in AI development, including those necessary for avoiding discrimination [32].

6 Conclusion

Artificial intelligence is still a young technology in the Norwegian public sector. While a handful of larger organizations are well advanced, many smaller public sector organizations do not have adequate resources or access to the necessary competences for developing AI systems. Our study illustrates how the process of developing AI can be described with the metaphor of a hop-on-hop-off journey, where the stops represent necessary elements that sometimes turn into barriers. Thus, it was not all the AI projects that had entered via the first stop of project design. Many of the projects had left, temporarily or permanently, on various stations. Only a handful of projects involving personal data had reached the final stage of putting their AI system into production. The metaphor of the journey illustrates the many elements and questions that need to be dealt with during this process. Technical, juridical, and domain competence together with competence about discrimination, are all vital for an AI project to make it successfully to the final stage of production with as little risk as possible involved.

While some of the public organizations had succeeded and had a good structure for developing AI with a multi-disciplinary team covering the different required competences, most of the smaller units recognized that they were lacking in one or more competences. Questions regarding juridical regulation and domain competence are necessary for developing a legal and precise AI system, and risk of discrimination should not be left behind as the final stop on this journey. Thus, our study confirms the importance of the recommendations from the Council of Europe's study of the impact of artificial intelligence on gender equality, as the authors conclude that "the regulatory subject is not AI taken in isolation but rather the broader socio-technical apparatus constituted by the interaction of social elements with algorithmic technologies" [13]. AI makes a good example of how technology and society are interwoven, and thus why leaving technology to tech people alone is not a good strategy for developing technology that can support societal needs.

Acknowledgements. This paper reports from the project "Use of artificial intelligence in the public sector and risk of discrimination", which was carried out on behalf of the Norwegian Directorate for Children, Youth and Family Affairs (Bufdir) by Rambøll Management Consulting and Vestlandsforsking (Western Norway Research Institute) in the period from November 2021 to December 2022.

We want to thank all contributors including Norwegian and international AI experts participating in the reference group and in meetings to discuss the project. We also want to thank the contributors from the public sector responding to the survey and participating in interviews.

The paper is developed from the project report, available in Norwegian: https://www.vestforsk.no/en/project/artificial-intelligence-public-sector-and-risk-discrimination.

References

1. Di Noia, T., Tintarev, N., Fatourou, P., Schedl, M.: Recommender systems under European AI regulations. Commun. ACM **65**(4), 69–73 (2022)
2. Sousa, W.G., Melo, E.R.P., Bermejo, P.H.D.S., Farias, R.A.S., Gomes, A.O.: How and where is artificial intelligence in the public sector going? A literature review and research agenda. Gov. Inform. Q. **36**(4), 101392 (2019). https://doi.org/10.1016/j.giq.2019.07.004
3. KMD. Én digital offentlig sektor: Digitaliseringsstrategi for offentlig sektor 2019–2025. Kommunal- og moderniseringsdepartementet (2019). https://www.regjeringen.no/no/dokumenter/en-digital-offentlig-sektor/id2653874/
4. KMD. Nasjonal strategi for kunstig intelligens. Kommunal- og moderniseringsdepartementet (2020). https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/
5. Broomfield, H., Reutter, L.M.: Towards a data-driven public administration: an empirical analysis of nascent phase implementation. Scand. J. Public Adm. **25**(2), 73–97 (2021)

6. Belenguer, L.: AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. AI and Ethics **2**, 771–787 (2022). https://doi.org/10.1007/s43681-022-00138-8
7. Barbieri, D., Caisl, J., Lanfredi, G., Linkeviciute, J., Mollard, B., Ochmann, J., et al.: Artificial intelligence, platform work and gender equality. European Institute for Gender Equality (EIGE) (2022)
8. White, J.M., Lidskog, R.: Ignorance and the regulation of artificial intelligence. J. Risk Res. **25**, 488–500 (2021)
9. Lepri, B., Oliver, N., Pentland, A.: Ethical machines: the human-centric use of artificial intelligence. IScience **24**(3), 102249 (2021). https://doi.org/10.1016/j.isci.2021.102249
10. Mannes, A.: Governance, risk, and artificial intelligence. AI Mag. **41**(1), 61–69 (2020)
11. Zuiderveen B.F.: Discrimination, artificial intelligence, and algorithmic decision-making (2018)
12. Broomfield, H., Lintvedt, M.N.: Is Norway stumbling into an algorithmic welfare dystopia? Tidsskrift for velferdsforskning **25**(3), 1–15 (2022). https://doi.org/10.18261/tfv.25.3.2
13. Bartoletti, I., Xenidis, R.: Preliminary draft Council of Europe study on the impact of artificial intelligence, its potential for promoting equality, including gender equality, and the risks to non-discrimination. The Gender Equality Commission (GEC) and the Steering Committee on Anti-Discrimination, Diversity and Inclusion (CDADI), The Council of Europe (2022). https://rm.coe.int/gec-2022-9-study-on-ai-211022/1680a8ad89
14. Xenidis, R., Senden, L.: EU non-discrimination law in the era of artificial intelligence: mapping the challenges of algorithmic discrimination. In: Bernitz, U., Groussot, X., Paju, J., de Vries, S.A., (eds.) General Principles of EU law and the EU Digital Order. Kluwer Law International, pp. 151–82 (2020)
15. UNESCO: Artificial intelligence and gender equality: key findings of UNESCO's Global Dialogue. Division for Gender Equality, UNESCO2020 (2020)
16. European Commission: Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. Office for Official Publications of the European Communities Luxembourg (2021)
17. Kommunal- og moderniseringsdepartementet. Datatilsynets og Personvernnemndas årsrapporter for 2018. (Meld. St. 28 2018–2019)
18. Lovdata. Act relating to the processing of personal data (The Personal Data Act) (2018)
19. European Parliament. MEPs ready to negotiate first-ever rules for safe and transparent AI (2023). https://www.europarl.europa.eu/news/en/press-room/20230609IPR96212/meps-ready-to-negotiate-first-ever-rules-for-safe-and-transparent-ai. Accessed 11 July 2023
20. Parmiggiani, E., Mikalef, P.: The case of Norway and digital transformation over the years. In: Mikalef, P., Parmiggiani, E., (eds.) Digital Transformation in Norwegian Enterprises. Springer International Publishing, Cham, pp. 11-8 (2022). https://doi.org/10.1007/978-3-031-05276-7_2
21. OECD: Drivers of trust in public institutions in Norway, building trust in public institutions. OECD Publishing, Paris (2022). https://doi.org/10.1787/81b01318-en. Accessed 11 July 2023
22. Andreasson, U., Stende, T.: Nordiske kommuners arbeid med kunstig intelligens. Nordic Council of Ministers (2019)
23. Alston, P.: Report of the special rapporteur on extreme poverty and human rights. UN General Assembly A/74/493 (2019). https://documents-dds-ny.un.org/doc/UNDOC/GEN/N19/312/13/PDF/N1931213.pdf?OpenElement
24. Redden, J.: Democratic governance in an age of datafication: lessons from mapping government discourses and practices. Big Data Soc. **5**(2) (2018). https://doi.org/10.1177/2053951718809145
25. Jørgensen, R.F.: Data and rights in the digital welfare state: the case of Denmark. Inf. Commun. Soc. **26**(1), 123–138 (2023). https://doi.org/10.1080/1369118X.2021.1934069

26. Suresh, H., Guttag, J.: A framework for understanding sources of harm throughout the machine learning life cycle. Equity and access in algorithms, mechanisms, and optimization, pp. 1–9 (2021)
27. OECD: Scoping the OECD AI principles (2019). https://doi.org/10.1787/d62f618a-en
28. SILO: The Nordic state of AI (2022). https://www.silo.ai/ebooks-reports/nordic-state-of-ai-2022
29. Norwegian cognitive center: Bergen Næringsråd. Digital Modenhet på Vestlandet. Delrapport 1: Kunstig intelligens (2022)
30. Gröger, C.: There is no AI without data. Commun. ACM **64**(11), 98–108 (2021). https://doi.org/10.1145/3448247
31. Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM transactions on information systems (TOIS) **14**(3), 330–347 (1996)
32. Srinivasan, R., Chander, A.: Biases in AI systems. Commun. ACM **64**(8), 44–49 (2021)
33. The Norwegian data protection authority. Sandbox forever (2022). https://www.datatilsynet.no/en/news/aktuelle-nyheter-2022/sandbox-forever/
34. Datatilsynet. Sluttrapport fra sandkasseprosjektet med NAV. Temaer: rettslig grunnlag, rettferdighet og forklarbarhet (2022). https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/nav-sluttrapport/
35. Köchling, A., Wehner, M.C.: Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Bus. Res. **13**(3), 795–848 (2020). https://doi.org/10.1007/s40685-020-00134-w
36. Gjerdsbakk, T.C.G.: Åpen og rettferdig kunstig intelligens. Lov & Data **150**(3), (2022)
37. Gerards J, Xenidis R. Algorithmic discrimination in Europe: challenges and opportunities for gender equality and non-discrimination law. European commission (2021)
38. Andréasson, U., Stende, T.: Nordic municipalities' work with artificial intelligence (2019)
39. Laclau, E., Mouffe, C.: Hegemony and Socialist Strategy: Towards a Radical Democratic Politics. Verso, London (1985)

Challenges in Regulating Online Hate-Speech Within the Norwegian Context

Carol Azungi Dralega[1](✉), Torborg Igland[1], and Gilda Seddighi[2]

[1] NLA University College, Bergtorasvei 120, 4633 Kristiansand, Norway
{carol.dralega,Torborg.Igland}@nla.no
[2] NORCE Norwegian Research Centre, Nygårdsgaten 112, 5008 Bergen, Norway
gsed@norceresearch.no

Abstract. Recent research shows that online hate-speech is on the rise in western societies, including in Norway. This submission builds on a broader 4-year research initiative on: "Violence-inducing Behavior Prevention in Social-Cyber Space of Local Communities" aimed at gaining deeper understanding of the complexities surrounding online hate as well as developing radical technological solutions (i.e., real-time artificial intelligence tool) to support authorities in their work against the vice. This paper addresses the former with a focus on the work of the police as one of several other community resilience stakeholders engaging in the fight against online hate-speech. The paper posits a socio-cultural, technological, and ethical analysis of the challenges, posed particularly in the enforcement of legislation on hate-speech (paragraph 185). Reflections are also included on the implication of the findings for AI development for community resilience against online hate speech.

Keywords: Online hate-speech · legal grey areas · racism clause · law enforcement · freedom of expression

1 Introduction

Online hate speech has profound consequences for individuals, groups and community resilience. Research shows that hate speech can erode social cohesion, fracture community bonds, and create divisions within online communities (Citron, 2014). Hate speech can perpetuate discriminatory attitudes, reinforce stereotypes, and marginalize targeted groups, undermining the inclusive fabric of communities (Baumgartner et al., 2019). Studies show that the vice of online hate speech is not only multi-disciplinary and complex (Paz et al. 2020), but also on the rise globally (Tontodimama et al. 2021, Laub 2019) including in Norway (Medietilsynet, 2022; Wigh 2022; Hatkriminalitet 2022; 2021 and 2019). A recent study on the proliferation of online hate speech covering 10.5 million Norwegian comments on Facebook over a period of 18 months, shows that 1, 7 percent of the comments are defined as attacks and 0,4 percent as hate speech (Nordic Safe Cities, 2023).

R. Akerkar (Ed.): SAIDD 2023, CCIS 1810, pp. 173–186, 2024.
https://doi.org/10.1007/978-3-031-53770-7_12

Given the steady growth over the last years, of online hate-crimes in Western societies, combating it has become crucial for governments, organizations, local communities, and stakeholders (Kalsnes & Ihlebæk 2021, NOU, 2022:9i1). Legislation has been one way of fighting the vice of online hate speech globally (UN, 2023) and in Europe (See Council of Europe recommendations[1] and frameworks[2]). Just as in several countries in Europe, in Norway, the constitution and several legislations prohibit hate-based discrimination and the abuse of individual and group rights. In Norwegian legislation, the so-called "hate-speech" paragraph (§ 185), placed in Chapter 20 in the Norwegian legislation, highlights the protection of society, public peace, order, and security. The § 185 deals with hate speech targeted at vulnerable groups, based on ethnicity, skin color, gender, religion, disability, sexual orientation, or worldview. The ongoing debate and scholarship on online hate-speech in Norway gained momentum in 2020, after a legal precedence was set with its first guilty verdict on racial online hate crimes (Nguyen, 2020). In August 2022, a special committee on freedom of expression released a new compendium that further defines and gives guidance on freedom of expression including online expressions (NOU, 2022: 9). Despite the Norwegian Constitutional and legislative recourse, there are still lingering tensions/grey areas including when freedom of speech collides with other rights and interests (Kierulf, 2021). We revisit some of these debate by exploring current police force encounters with this evolving § (185).

As a key stakeholder in building community resilience against online hate speech and for their close work with the § 185 in Norwegian contexts, the Police have been selected as the main target group for this paper. Police in Norway engage in several activities in their work on online hate speech. Their work includes (and is not limited to): outreach programs meant for raising awareness and education (such as making TikTok, reaching out to minority social arenas, mosques, etc.), partnerships with other stakeholders, net patrol, tips portal, legislative processes, and so on. More centrally, their task involves applying/interpreting the legislation (§ 185) to establish hate-crime cases. This paper seeks answers to the research question(s):

W*hat are the challenges in the work against online hate speech generally, and more specifically, how do ambiguous aspects within legislation (§ 185) affect law enforcement's ability to regulate online hate speech?*

The paper's focus on § 185, is not intended as a legal analysis or undertaking per se, but rather to help highlight how the interviewees in police force reflect among other issues on the importance of the socio-cultural context of hate speech. The socio-cultural lens of analysis is vital for deepening our understanding of societal implications of legislative processes and outcomes (a focus we also recommend later in the chapter when developing AI tools). It is a lens well-articulated by Gagliardone, Gal, Alves and Martinez who caution that: "…purely legal lens can miss out on how societies evolve through contestation and disagreement" (Gagliardone et al., 2015, p.15). To highlight the existing tensions of § 185, we bring to view three recent public debates we refer to as the Sumaya Facebook case, the Sumaya-Atle case and the Swastika case.

[1] https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680a67955#_ftn1.

[2] https://www.coe.int/en/web/combating-hate-speech/council-of-europe-on-hate-speech.

With regards to AI. It was established early on in our exploratory study, that, although highly desired, the police did not yet employ AI tools in their work. Given that our overall computational social science project aims to also develop real time AI tools to support community resilience work such as the work police do, we include, brief reflections on how our findings here are relevant for and how they can inform the development of AI and what measures ought to be integrated in the AI development processes.

Methodologically, the study adopts a qualitative approach encompassing 'bottom-up', dialogical, interdisciplinary tenents of research. What follows is a brief contextualization of the paper, conceptual framework, followed by methodology, presentation of analysis, and discussion of findings. Concluding remarks include recommendation for police work, within the framework of the § 185, on community resilience against hate speech.

2 Paragraph 185 Contextualized Within Three Use Cases

The § 185 on hateful utterances was enacted in 1970, but has been changed several times, latest in 2020. The § 185 does not protect every hateful utterance, but the hate speech must affect a certain vulnerable group, based on ethnicity, skin color, religion, disability, sexual orientation or worldview. This right to protection by the law, sometimes collide with the freedom of speech (Jakubowicz et al., 2017). Conditions for conviction include: the act has to be public (Lovdata, § 185, 1st section, "in the presence of others" [our translation]), there has to be an intent to cause harm behind the act, or to reduce the value of someone, (Lovdata, § 185, 1st section, "intentionally or grossly negligent" [our translation]), the act has to have a certain degree of offensive impact (Lovdata, § 185, 2nd section, "threaten or insult somebody, or promote hatred, persecution or contempt towards somebody" [our translation]), and the target must, as mentioned above, be a person a minority based on color, race, sexual orientation, religion, gender, 'handicap' (Lovdata, § 185, 2nd section). The context also needs to be taken into consideration (ECHR, 2022, p. 4).

The following use-cases illustrate the tensions of the § 185. These cases will be referred to as Sumaya Facebook case, Sumaya-Atle case, and swastika case. In the Sumaya-Facebook case, from 2019, a woman wrote to Sumaya Jirde Ali (henceforth referred to as Sumaya) in a public Facebook group: *Bloody black offspring go back to Somalia and stay there, you corrupt cockroach.* Sumaya is a Norwegian-Somalian award-winning writer and debater. She has lived with threats and harassment in many years, because of her religion and origins (Somalia). The woman, who wrote the above-mentioned sentence, was convicted because the case met all the requirements in the hate speech paragraph. In the Sumaya-Atle case, from 2022, is when a famous comedian Atle Antonsen told Sumaya in a bar: *You are too black to be here.* After repeatedly yelling at her to: 'Shut the f**k up. The violation was reported but was rejected. The two court cases on hate speech towards her, received respectively a conviction and a rejection. Both cases reveal tensions within the hate crime legislation, and between the expectations in the public on what the law covers – and the law (Lovdata, § 185). As a comedian, Atle's intension was considered not to harm, and the context around the expressions - they were colleagues, Atle had earlier sent Sumaya support in a similar situation - supported that interpretation.

The intention criterion is often difficult to prove in hate speech cases (Lovdata, § 185, 1st section). The context of the hate speech, as illustrated above, needs to be taken into consideration. The case of the swastika flag illustrates this point. But it also illustrates the wide interpretation space there is in these cases. The swastika case received much attention in the media. A white supremacy group put up the swastika flag in the Norwegian Town Kristiansand. In the lower court, these men were sentenced and fined. The context, in reference to the date, place and symbolisms, made the act even more vile and threatening. The date was 9th of April, the date of the German invasion in Norway. The place was the *Arkivet*, a place where opponents of the nazis were tortured and imprisoned. However, the conviction was appealed, and the court of appeal acquitted the defendants, saying that the utterance was not aimed at one of the protected groups, since it also affected many other groups. In addition to the three use cases above, the Supreme court has stated that the expressions need to have a good "margin" for tastelessness or offensiveness (Kierulf, 2021).

3 Conceptual Framework

Online hate speech and cybercrime calls for a new discipline, social cyber security, and a need to build an infrastructure that allows the essential character of a society to persist through a cyber-mediated information environment (Carley, 2020). AI can be part of such infrastructure. Social cyber security is focused on how humans, as well as communities and narratives, can be compromised, and how this can be stopped. Within the social cyber security area, AI can provide new tools to support good decision-making. But such tools will have limitations, according to Carley (2020). AI and machine learning are immensely valuable for managing extensive datasets. Nevertheless, these systems depend on data that is structured around signs and vocabularies commonly recognized as indicative of hate. Consequently, AI may struggle to understand more fluid socio-cultural contexts and nuances in language and sentiment (Carley, 2020). The tensions and uncertainties within the legal framework for hate speech underscore the need to take the socio-cultural context into account when developing AI-based tools to combat hate speech (we return to this later). Firstly, the term 'hate speech' lacks a *universal* definition (Assimakopoulos et al., 2017). Hate speech may be seen as "[...] the expression of hatred towards an individual or group of individuals based on protected characteristics," (ibid, 2017, 12). But again, these characteristics are also open to definition. Hate speech can be conveyed through any form of expression, online or offline, it is discriminatory or pejorative of an individual or group and it "calls out real or perceived 'identity factors' of an individual or a group, including 'religion, ethnicity, nationality, race, color, descent, gender', but also characteristics such as language, economic or social origin, disability, health status, or sexual orientation", among others" (ECHR-KS, 2022; UN.org[3]).

Secondly, there is no *comprehensive* definition of the concept hate speech. Instead, for instance, the European Court of Human Rights approach the problem on a case-by-case basis (ECHR, 2022, p. 1). This is consistent with the Norwegian case law (Kierulf, 2021). They acknowledge National courts' responsibilities to interpret and apply domestic laws.

[3] https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech. Last accessed 07.06.2023.

The ECHR has general principles drawn from the case-law (ECHR, 2022, p. 1). One is that freedom of expression constitutes not only sharing ideas or information, but also offensive utterances that may disturb the State and sections of the population. Hate speech is often categorized as either hard or soft hate speech, according to Assimakopoulos et al. (2017). Hard hate speech is prohibited by law, and soft hate speech is not illegal, but still have serious consequences when it comes to intolerance and discrimination. The line between the two is drawn differently from country to country. In addition, there is a grey zone of ambiguity where human values and human rights, freedom of expression and the right to be protected from discrimination might collide. Freedom of speech is the right to utter one's opinion without interference from the authorities. The term "speech" includes expression more than just words, but also clothing, what a person reads, performs, protests and so on (CSUSM, u.å.) (Jakubowicz et al. 2017, p. 26). Hate speech such as racism "can be embedded in structures of societies" (Jakubowicz et al. 2017, p. 26), in the benefits that flow and the disadvantages that inhibit. To build community resilience against cyber racism, there should be established online communities that offer support against racism. When talking about resilience, Jakubowicz et al. (2017) are not talking about victims needing to get tougher, nor is it right of others to be bigots, rather it is about "enabling citizens to discern and name racism, and ultimately resist it", to unpack and challenge hateful behavior. (Bodkin-Andrews et al. 2013 in Jakubowicz et al., 2017, p. 277).

The internet is argued to be a non-governable space as it might create a favorable setting for unwanted activities due to the high quantity of online activities (Bromell, 2022); the lack of social and cultural consensus on what content is acceptable and what should be requirements for intervention (Kierulf, 2021), lack of harmonization of legislation across borders as well as how easy it is to be anonymous (Citron & Norton, 2011; Jakubowicz et al., 2017). The technological provisions that allow the use of pseudonyms, fake accounts, or anonymous browsing tools hinder law enforcement agencies' ability to trace, gather evidence and attribute hateful content to specific individuals (Citron & Norton, 2011). The lack of identifiable information (for instance when someone deletes their initial hateful utterances) makes it challenging to initiate legal proceedings or pursue criminal charges (Kovacs, Alonso and Saini, 2021). Legal frameworks often require evidence that can establish a direct connection between the hate speech and its author, which anonymity obstructs. The practical limitations imposed by anonymity can impede the enforcement and hinder the pursuit of justice. The absence of personal accountability or disinhibition may contribute to a culture of impunity and increase the prevalence of hate speech online. On the other hand, the fear of retaliation or retribution faced by victims and potential whistleblowers can deter them from reporting instances of hate speech, further enabling anonymous perpetrators to operate freely (Citron & Norton, 2011).

Further, there are tensions between different human rights, like freedom of expression and freedom of religion and between freedom of expression and freedom from discrimination. Rogstad (2014 in Colbjørnsen 2016) points out that discussions on where these borderlines should be drawn, often are turned into issues of principle, that often may be driven by a news event, covered by the media, and further discussed in the social media. Rogstad calls these events collective references (Colbjørnsen, 2016).

While societal norms often sanction overt racism in offline spaces, anonymity of online spaces provides opportunities for online racist hate speech, the study points out (Ortiz, 2019). One of the strategies of the men of color in the study was *desensitizing* the racism. Soral et al. (2018) suggest that regularly being exposed to online hate speech "reduces automatic triggering of negative emotional reactions to images, words, or thoughts of violence" (p 137). This lower sensitivity has been measured by physiological tests, as decreased attention to violence, increased belief that violence is normal and so on. Jakubowicz et al. (2017) reflects on whether cyber racism should be seen as a civil or a criminal wrong. The discussion is relevant for several types of hate speech. Criminalization has the advantage that the victim does not have to enforce the matter. Moreso, criminal sanctions send the signal that the state condemns this kind of behavior. On the negative side, criminalization tends to individualize the problem, with the risk of re-producing the problem (McNamara, 2002).

4 Methodology

The focus of this paper is the police work against online hate speech that intersect with other works towards building community resilience such as investigating the punishable utterances (i.e., penal code § 185), preventing radicalization to violent extremism; preserving the democracy; polarization; working to prevent lack of feeling of safety, i.e. religious minorities, sexual minorities and other vulnerable groups (§ 185); working to prevent lack of trust from vulnerable groups towards the police.

Empirical insights for this paper were obtained through qualitative methodologies, including document analysis, a workshop with stakeholders, interviews, and focus group discussions with individuals within the Norwegian Police force between October 2022 and February 2023.

Workshop: In the early phase of the research, a workshop was organized with various public and private stakeholders to gain insights into how stakeholders working against online hate speech in different societal arenas perceive their roles and challenges. We also wanted their input for an AI tool. These stakeholders included the Norwegian Media Authority, the Police, voluntary organization, and local municipality. The workshop began with a presentation of the research project and then proceeded with stakeholders sharing their work experiences. The stakeholders' presentations were transcribed. The stakeholders invited to the workshop served as gate openers, helping us connect with the appropriate participants for individual and group interviews.

Interviews: In the subsequent phase of the research, three key informants from the relevant police units from Bergen and Oslo were invited to interview. An interview-guide was administered encompassing five primary areas of discussion: 1) projects focused on combating online hate speech, 2) organizational structure and workflow, 3) the specific project's definition of hate speech, 4) the resulting products such as statistics, reports, and measures, and 5) the challenges encountered in the project. All interviews took 45–60 min.

Document Review: During the fieldwork, we collected publicly available reports and notes that were suggested or provided by the participants. Some of these documents included the Police annual report on hate criminality and Norway's Supreme Court rulings (Hatkriminalitet 2022; 2021 and 2019; Borgarting lagmannsrett Dom – LB-2019-177188). The reports were examined to provide context for the research, while the rulings from the Supreme Court were studied and analyzed in the context of the discussions that came up in the interviews and workshop.

This study gained ethical approval from the Norwegian Centre for Research Data (SIKT) and followed their rules for data security. All informants are anonymized. However, there are limitations of anonymity as there are a few departments in the Norwegian police force involved in combating online hate speech. For anonymity, we use letters for identification of informants.

Analytical Framework

Inspired by grounded theory (Charmaz, 2014), we have applied an inductive qualitative analysis inquiry. The process started with coding the data material from the interviews and workshop. At this phase the research team aimed to get familiar with the data. This developed to find patterns across the interviews and going back and forth between notes from the Supreme Court, the interviews and workshop discussions. In this process, codes were categorized. The relevant categories for this paper were interpretation of law, definition of hate speech, identifying online hate speech, challenges in combating online hate speech. Following the steps of grounded theory, the identified categories or central themes are discussed in the form of findings and discussed in the context of earlier research and literature.

5 Presentation of Key Findings on Challenges of Regulatory Work on Online Hate Speech

It is important to note that during the fieldwork, the police did not utilize AI tools. However, they did employ digital/online technologies, primarily open source, in conjunction with outreach efforts such as talks in cafes and mosques. Interviews frequently highlighted the use of web-based tips portals, net patrols, and the police's presence on popular youth-oriented social media platforms like TikTok. The informants saw the need for an AI tool that could support their existing efforts through digital technology. In this section and following pages, we will delve deeper into the challenges faced by the police in regulating online hate speech, shedding light on the implications of paragraph 185 legislation and the areas of ambiguity it presents.

Police reports (Hatkriminalitet I Oslo pd 2021 p.4; 2020 p.5; 2019, p.4) highlight a rise in online hate speech. Despite this, fewer online hate cases were reported to the police (informant A). 'Tips' portal (Tipsportalen), a web-based scheme, is created by Police to opens access for the public to report cases of hate speech including online hate speech. However, for instance informant C underscores the fact that "there is a massive underreporting of hate crime to the police. Both online and offline.

Difficulties in Understanding the Parameters of § 185

All our informants reflected on whether and how § 185 is challenging. For instance, when the public register their experiences of hate speech on police website (Tipsportalen), the police hate crime team interpret the paragraph to find out whether they must open a case of a crime. Informant C argues that the paragraph is not easy to understand and/or interpret by the public. According to her:

> I do not think people walk around with Supreme Court judgements in their heads[4]. We experience through the public debate that hate crime is more on the agenda, but the threshold for reporting to the police is far too high. It is not the public's task to assess whether its punishable or not, but only to submit incidents. Informant C.

Commenting on the challenges relating to 'intention,' 'specificity' and 'public' nature of utterances of online hate speech, informant A argues that the line between punishable hate speech and freedom of speech is vague and the requirements for a conviction by § 185 are strict. "Often, one wonders, is the case public enough? Specific enough? "Loosening up the requirements a little could solve a lot." On the question of specificity, Informant A offers an illustration:

> There are cases where people of ethnic minority backgrounds have tried to file a report and are rejected at the front desk. This is because first reports must be registered, and the hate motive has to be registered. Identifying the motives behind hate speech and who are these people who perform the hate speech online could be difficult for victims to identify. A rejection at the front desk can be a demoralizing factor for further reports.

Most online hate speech occurs on social media, which challenges the general traditional public-private boundaries. In addition, such hate speech might be operating in small and closed platforms different from others that are open and "public" – challenges that hinder the assessment of the online hate speech. If you want to make it relevant to § 185, you need to remind the reader that the fact that the hate is happening in public is one of the criteria of the hate speech as explained by the Sumaya case. Informant A highlights that the interpretation of what can be considered as public in the context of § 185 is difficult when we operate in a digital world where the boundaries between public and private are unclear. Informant A ponders:

> Interpretation of the § 185 is difficult, for instance, what is 'in public'? Would a private group on Facebook be considered public? What about a crowd of 10–15 people?

In addition to the challenges of lack of simplicity, difficulty of interpretating the paragraph, underreporting of incidences of online hate speech by the public, the grey areas associated with specificity and public nature of utterances, another challenge relates to diversity and interpretation from multi-cultural contexts. Two of our informants explain below:

> The police used translation programs or colleagues with language skills, but these come with challenges as some nuances and contexts may be lost along the way. Informant A.

[4] Supreme Court judgements represent the most important case law.

For informant C, the socio-cultural challenge is more complex as it involves a good understanding of language, context, culture, and rhetoric. It also comes with a resource aspect:

> For things said in other languages, we need both the immediate understanding of a text and seeing it in the context of what has been said before. The meaning also depends on culture. We have not had many of those cases, but we have had some. Then we try to identify (human) resources within the police force, read and assist with how the statement can be interpreted, then we contact the interpreting services. We have also been in contact with the Mosque to get help in understanding the context and culture. That also means understanding rhetoric and which words we use in everyday speech."

We understand that some of the underreporting is related to difficulties in understanding the parameters of §185, or lack of trust in the authorities or fear that the case will be rejected. However, Informant A, highlights another challenge as a possible explanation for underreporting:

> We are witnessing a rise in hostile opinions and utterances becoming more socially accepted than before. Tendencies are that hostile debates are increasing in society, and there is more acceptance and less social control on this area. According to Informant A.

Informant A argues that in today's online spaces, hate speech is normalized and people are becoming desensitized to it. The statistics are not accurate nor representative of the scope of hate speech violations. Accordingly, the opportunity by the police to convict haters is reduced.

The issue of anonymity and its impact on identifying individuals responsible for engaging in online hate speech poses a major obstacle to linking hate speech to real-world identities which undermines efforts to hold individuals responsible for their actions. It is also not possible to assess the intention of the act, nor the full context or whether it occurred in 'public', which are essential when it comes to assess a case legally.

The police further argue that the affordances of anonymity and privacy on social media platforms makes their work challenging in the following ways:

> Anonymous profiles online can be a problem. Some suggest that you log in with a bank ID etc. But there are many opportunities to make yourself invisible, which makes it difficult for us. Most people who engage in hate speech are around 50 years old. Also, a lot is deleted, then we struggle if we do not have a screenshot. If it has happened on Facebook, Facebook probably owns it, with a view to tracking it down afterwards. According to informant C.

We see that the inherent ambiguity in hate speech laws may be a result of the complex nature of hate speech and legislation cannot effectively guide people without a proper understanding of the context within which it occurs. But, as a novelty, it gets more complicated especially in discerning multimodalities (i.e., images, sound, graphics and videos) as our informant laments:

> While the law is sometimes difficult to interpret, online hate speech perpetuators are creative and find ways to go around the law. For instance, they may post an image of monkeys instead of texting insults or a hateful utterance. Informant A.

So, we see the police work towards building community resilience against online hate speech are hindered by several challenges ranging from legalistic, to socio-cultural, to ethical (privacy and anonymity issues), as well as technological affordances.

Discussion

The § 185 which falls under Chapter 20 of the Norwegian Penal Code is primarily for societal protection, the aim is to prevent expressions that can foster hatred or harm against minorities in society. While acknowledged as a positive measure towards curbing online hate speech, our informants also reflected on the limitations and challenges – which is the focus of this paper. The challenges are varying from underreporting, lack of trust, fear of harassment, difficulties in using the 'tipsportalen', privacy on social media platform and difficulties in finding and defining contexts and intentions behind the online utterances, which goes directly to the core of the challenge regarding legislation, §185. Some of these challenges are about technological infrastructure, and some are about the victim's experiences. The challenge about legislation covers language, interpretation and judicial precedent and discretion. The grey areas challenge the informants to evaluate the registered online hate speech and open cases on hate-crime. Their arguments among others range from: the § 185 has high threshold which makes it be experienced by offended as out of reach and too strict and must be loosened (informant A), the parameters for what is hate speech and freedom of expression are blurry or not specific enough (informant C) and the requirement for what is considered a publicly performed hateful utterance are unclear (informant A). The illustrative use-cases presented earlier i.e. Sumaya Facebook case, Sumaya-Atle case, Swastika cases derived from the paragraph's requirements for intent ((Lovdata, § 185, 1st section), public (Lovdata, § 185, 1st section), context (ECHR, 2022, p.4) and level of offensiveness and tastelessness of hateful utterances (Lovdata, § 185, 2nd section).) further illustrate how the lack of clarity poses challenges in the interpretation/operationalization of the paragraph (Kierulf, 2021).

Further, according Kierulf (2021) the human language will always be more open to interpretation than human actions. § 185 says that the meaning of an utterance must never be interpreted beyond the actual phrase – to avoid convictions for something suspects did not utter or intend to utter. This principle needs to be made clearer or literal in the legislation (Kierulf, 2021). When it comes to balancing freedom of expression against hate speech, the Norwegian legislation and jurisdiction is bound by the European Court of Human Rights, which states that the two must be assessed with each case, in the same way as the utterance must be evaluated in the specific context in which it is made (ECHR-KS, 2022). Concrete boundaries are easier to establish against actions than against utterances, but even with actions, there exist gray areas. You will always need an interpretive space, and more so when it comes to language – this will often be perceived as vagueness. Also, when it comes to building community resilience, for instance, through an AI tool and machine learning, which can work to identify hate speech terms and expressions, it is worth reminding ourselves on the limitations on AI on interpreting more volatile socio-cultural contexts and sentiments in human language

(Carley, 2020). It is in other words difficult to avoid grey areas when interpreting and regulating utterances, and to achieve community resilience, we need humans in the loop.

The 'Chaotic' Terrain of Online Hate Speech

The findings highlight how community resilience efforts by the police must be placed within the chaotic contextual terrain of digitalization and online affordances within which hate speech survives and thrives. 'Chaotic' terrain because, it lacks coherence which creates confusion and makes it difficult to establish clear guidelines or solutions. Informant A points to the concerning trends of the normalization of hate speech through repeated exposure, possibilities of anonymity and limited disinhibition echoing researcher warnings on how repeated exposure to online hate speech, leads to desensitization and negative effects on empathy and attitudes according to scholars such as Kircaburun et al. (2018) and Tynes et al. (2018). Duffy and Chan also point to ethical concerns about privacy and the fear of being incorrectly associated with hate speech may lead individuals to self-censor or refrain from engaging in legitimate expression (2019).

Novelty and Trust

As the informants indicated, the police's challenges of regulation of online hate speech are also due to its novelty in comparison to regulation of traditional/offline forms of hate speech and other forms of discrimination (such as the gender equality law). Our study supports the existing research that shows novelty in legal frameworks and definitions (Trottier & Fuchs, 2014); novelty in jurisdictional complexity (Feldman, 2019), technological advancement and evasion tactics (Citron & Norton, 2011), dynamic nature of online platforms (Van Dijck, Poell, & De Waal, 2018) as legitimate challenges which may be relevant in Norwegian contexts. The novelty of these practices and laws presents challenges in trust-building, credibility of enforcement mechanisms, user engagement in reporting, and platform cooperation. Our findings highlight the challenge from the novelty of the fight against online hate speech particularly pointing to the issue of trust among the public especially among victim groups and communities as particularly indicated by informant C.

Social-Cultural Sensitivities

The work that the police undertake particularly with the interpretation of the law is arguably limited considering the victims' perspectives, experiences, and contexts of online hate speech. Given that the threshold is high for conviction from paragraph 185 means that several victims (often minorities), in this case, ethnic minorities, cannot obtain legal recourse from the racism clause and yet as research shows, hate speech whether soft racism or hard has consequences to victims and victim communities varying from: health and wellbeing (Burks et al., 2018; Gagliardone et al., 2015) and their right to living in peace and security (UN Declaration of Human rights article 28). That conviction for the online hate speech must among other things be highly offensive (Kierulf, 2021) as understood by the public (Kierulf, 2021) calls for a debate about power and hegemony over marginalized groups. As informants C highlights, although there might be intersections with majority (Rios & Cohen, 2023), majority and minority understandings and experiences of online hate may not be similar given differences in culture, language, contexts).

6 Concluding Remarks and Implications for AI Development

This article explores the challenges faced by the police in their efforts to combat online hate speech. The focus on and purpose of § 185 is to protect society from hate speech directed at particularly vulnerable groups. However, the qualitative interviews conducted with stakeholders in this study indicate that while necessary and useful in some cases, the legislative provisions are not sufficient nor efficient enough to cover the scope, complexity and nuances of victims' experiences, desired societal perspectives, and varied contexts. The police, responsible for enforcing the law aimed at protecting society from online hateful expressions, point out that the wide interpretive scope for online expressions, which is much broader than for actions of hate, affects the police's ability to address online hate speech. For instance, no one should be convicted for anything more than they have intended and language and messages can be interpreted differently depending on the context. With comprehensive language (including multimodal) training, AI can assist the police in identifying possible hateful expressions, but the many, ever-changing layers of complexity and nuanced aspects of language, different socio-cultural codes, traditions, and contexts, imply that an AI tool may not fully comprehend it all on its own. Humans are required in the loop, to provide context and accurate language interpretation.

Acknowledgements. This 4-year project was funded by the Norwegian Research Council, to whom we are grateful. We are also grateful to the informants for their generosity with information and insights. Appreciation also goes to the project manager Professor Rajendra Akerkar from Western Norway Research Institute as well as our colleagues from NTNU in Norway and George Mason University, USA for the critical discussions that have enriched this work.

References

Assimakopoulos, S., Baider, F. H., Millar, S.: Online hate speech in the European Union: A Discourse-Analytic Perspective. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-72604-5

Baumgarten, N., et al.: Towards balance and boundaries in public discourse: expressing and perceiving online hate speech (XPEROHS). RASK: Int. J. Lang. Commun. **50**(Autumn 2019), 87–108 (2019)

Bromell, D.: Regulating Free Speech in a Digital Age: Hate, Harm and the Limits of Censorship. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-030-95550-2

Burks, A.C., Cramer, R.J., Henderson, C.E., Stroud, C.H., Crosby, J.W., Graham, J.: Frequency, nature, and correlates of hate crime victimization experiences in an urban sample of lesbian, gay, and bisexual community members. J. Interpers. Violence **33**(3), 402–420 (2018). https://doi.org/10.1177/0886260515605298

Carley, K.M.: Social cybersecurity: an emerging science. Comput. Math. Organ. Theory **26**(4), 365–381 (2020). https://doi.org/10.1007/s10588-020-09322-9

Charmaz, K.: Constructing grounded theory. Sage (2014)

Citron, D.K.: Hate Crimes in Cyberspace. Harvard University Press (2014)

Citron, D.K., Norton, H.: Intermediaries and hate speech: fostering digital citizenship for our information age. BUL Rev. **91**, 1435 (2011)

Colbjørnsen, T.: Kritiske øyeblikk i norsk ytringsfrihetsdebatt En analyse av forekomster av omtaler av ytringsfrihet og pressefrihet i norske aviser 1993–2015. Sosiologisk tidsskrift nr 3 2016 © Universitetsforlaget, **24** 170–199 (2016). https://doi.org/10.18261/issn.1504-2928-2016-03-01, www.idunn.no/ts/st

European Court of Human Rights. KEY THEME1 Article 10 Hate speech. Last updated: 12/07/2022. Accessed 07 April 2023

Feldman, S.M.: Free-Speech Formalism and Social Injustice. Wm. & Mary J. Race Gender & Soc. Just. **26**, 47 (2019)

Gagliardone, I., Gal, D., Alves, T., Martinez, G.: Countering online hate speech. Unesco Publishing (2015)

George, C.: Hate speech law and policy. Int. Encycl. Digital commun. Soc. 1–10 (2015)

Hatkriminalitet. Anmeldt hatkriminalitet I Oslo politidistrikt 2021. Mars 2022. Politiet

Hatkriminalitet. Anmeldt hatkriminalitet I Oslo politidistrikt 2021. Mars 2021. Politiet

Hatkriminalitet. Anmeldt hatkriminalitet I Oslo politidistrikt 2021. Mars 2019. Politiet

Jakubowicz, A., et al.: Cyber racism and community resilience. PHS, Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64388-5

Kalsnes, B., Ihlebæk, K.A.: Hiding hate speech: political moderation on Facebook. Media Cult. Soc. **43**(2), 326–342 (2021)

Kierulf, A.: Hva er ytringsfrihet? Universitetsforlaget (2021)

Kircaburun, K., Jonason, P.K., Griffiths, M.D.: The Dark Tetrad traits and problematic social media use: The mediating role of cyberbullying and cyberstalking. Pers. Individ. Differ. **135**, 264–269 (2018)

Kovács, G., Alonso, P., Saini, R.: Challenges of hate speech detection in social media: data scarcity and leveraging external resources. SN Comput. Sci. **2**, 1–15 (2021)

Laub, Z.: Hate speech on social media: global comparisons. Council on foreign relations, 7 (2019)

McNamara, L.: Regulating racism: racial vilification laws in Australia (2002)

Man må ha tykk hud eller ungå å være på nettet: en undersøkelse om unges erfaringer med hatefulle ytringer (2022). https://www.medietilsynet.no/globalassets/publikasjoner/kritisk-medieforstaelse/2022-rapport-hatefulle-ytringer.pdf

Nguyên Duy, I.: The limits to free speech on social media: on two recent decisions of the supreme court of Norway. Nordic J. Hum. Rights **38**(3), 237–245 (2020)

Nordic safe cities. Angrep & hat i den offentlige debatten på Facebook (2023). (nordic-safecities.org)

Et helhetlig diskrimineringsvern— Diskrimineringslovutvalgets utredning om en samlet diskriminerings lov, grunnlovsvern og ratifikasjon av tilleggs protokoll nr. 12 til EMK (2009)

En åpen og opplyst offentlig samtale: ytringsfrihetskommisjonens utredning. Norges offentlige utredninger 2022:9. Oslo (2022)

Ortiz, S.M.: "You can say i got desensitized to it": how men of color cope with everyday racism in online gaming. Sociol. Perspect. **62**(4), 572–588 (2019). https://doi.org/10.1177/0731121419837588

Paz, M.A., Montero-Díaz, J., Moreno-Delgado, A.: Hate speech: a systematized review. SAGE Open **10**(4), 2158244020973022 (2020)

Rios, K., Cohen, A.B.: Taking a "multiple forms" approach to diversity: an introduction, policy implications, and legal recommendations. Soc. Issues Policy Rev. **17**(1), 104–130 (2023)

Soral, W., Bilewicz, M., Winiewski, M.: Exposure to hate speech increases prejudice through desensitization. Aggressive Behav. **44**, 136–146 (2018). https://doi.org/10.1002/ab.21737

Tontodimamma, A., Nissi, E., Sarra, A., Fontanella, L.: Thirty years of research into hate speech: topics of interest and their evolution. Scientometrics **126**, 157–179 (2021)

Tynes, B.M., Lozada, F.T., Smith, N.A., Stewart, A.M.: From racial microaggressions to hate crimes: a model of online racism based on the lived experiences of adolescents of color. Microaggression Theory: Influence Implications 194–212 (2018)

Trottier, D., Fuchs, C.: Theorising social media, politics and the state: an introduction. In social media, politics and the state, pp. 3–38. Routledge (2014)

Countering and Addressing online hate speech: a guide for policy makers and practitioners (2023). https://www.un.org/en/genocideprevention/documents/publications-and-resources/Countering_Online_Hate_Speech_Guide_policy_makers_practitioners_July_2023.pdf

Van Dijck, J., Poell, T., De Waal, M.: The platform society: public values in a connective world. Oxford University Press (2018)

Wigh, K.: Hvordan bruker politiets nettpatruljer sosiale medier? (Master's thesis, University of Agder) (2022)

Digital and AI Maturity of Enterprises in Sogn Og Fjordane, a Rural Region of Norway

Malin Waage(✉), Bjørn Christian Weinbach, and Øyvind Heimset Larsen

Vestlandsforsking, Sogndal, Vestland, Norway
mwa@vestforsk.no

Abstract. This paper, which is based on results of a questionnaire sent out to employees, aims to evaluate the level of digital and artificial intelligence (AI) maturity among businesses in a rural Norwegian region (Sogn og Fjordane), identify challenges, and recommend potential opportunities within important regional sectors. Western Norway's Sogn og Fjordane is significantly dependent on its small- to medium-sized enterprises (SMEs). Despite the fact that many businesses in the area are aware of the benefits of using data and AI, implementing these technologies into their daily operations seems to present a number of difficulties. Companies that are gathering large enough data-sources, may encounter challenges in effectively leveraging data-driven technology due to a lack of long-term strategies, knowledge, skills, and finance. Recommended tactics to adopt AI-techniques or implement specialized AI solutions and enhance internal skills can rely on training in specific abilities, knowledge exchange across disciplines or industries, and through research collaborations. According to the study results, the maturity is comparable to the larger area of western Norway; hence, the region's rurality and SME dominance might not prevent AI adaptation.

Keywords: Digital maturity · business life · Industry 4.0 · Artificial Intelligence

1 Introduction

Adopting new technology and AI-driven techniques may have the potential to, i.e., optimize production, increase insights and gain a competitive edge thus leading to economic growth for companies as well as making them more sustainable [1]. This paper addresses the willingness, ability, and plans to adopt digital technologies, or AI (artificial intelligence) in a rural region of western Norway called "Sogn og Fjordane". We wanted to find out to what extent SMEs in the region use data analysis and AI applications and the companies understanding and skills of how AI and data utilization can change their businesses. We also wanted to find out if they had plans for technology development or further data utilization.

The small population density of Sogn og Fjordane (5.9 people per km^2) makes it a rural region. The business life is dominated by small-medium sized enterprises (SMEs) [2]. However, despite being a rural region reliant on SMEs, it is also renowned for its robotics development through some companies and university groups, and several

R. Akerkar (Ed.): SAIDD 2023, CCIS 1810, pp. 187–202, 2024.
https://doi.org/10.1007/978-3-031-53770-7_13

nationally and internationally known digital companies and companies known for being at the forefront of technology have their primary location there. Whether these forefront companies and other activities have affected the general business life in terms of digitalization, AI, and technology development is unknown, and another interesting research question to assess.

The contribution of this paper is to uncover the digital maturity state, including AI usage among SMEs in a region, and, from a regional perspective, discuss some potential key future aspects and opportunities for sustainable data and AI adaptation.

2 Literature Review

2.1 Industry 4.0 and Value of AI in Future Business Life

There are numerous definitions of AI depending on the targeted usage of it. The European AI Strategy defines AI as: "refers to any machine or algorithm that is capable of observing its environment, learning, and based on the knowledge and experience gained, taking intelligent action or proposing decisions." [3]. Like humans' ability to learn from past experiences, AI can learn from historical data. By analyzing data, AI-powered systems can detect patterns and make informed decisions that lead to optimal outcomes. These outcomes can range from increased efficiency and process optimization to the prediction of future outcomes, providing decision support, and automating traditional manual and repetitive work.

The fourth industrial revolution (Industry 4.0), or "the digital shift" [4] is upon us and over the last two decades many businesses have undergone a drastic digital shift, starting with less paperwork and more digital data to AI driven business models. As the amount of data captured, collected, and used globally along with AI-applications continues to increase rapidly [5], exploiting data and such technology is becoming an increasingly important asset for governments and enterprises. In efforts of meeting global standards on quality, technology, sustainability, and pricing, leveraging data and Artificial Intelligence (AI) seems to be a continued growing key ingredient of Industry 4.0 [6]. Some studies suggest that AI-technologies may contribute to boost a company's revenue by 10–70% [7, 8]. By 2030, projections made by PwC suggest that AI-technologies (including robotics) has a potential to create 15.7 trillion dollars in annual value globally, making it the largest commercial opportunity in today's changing economy [1].

The current focus on circular economy and sustainability also creates a huge pressure on making manufacturing operations more ethical and sustainable. The application of Industry 4.0 technologies has been observed to have positive influences on achieving those goals [9].

2.2 SMEs and Challenges of AI Adoption

In the race to adopt new digitalized methods, sustainability, and a circular economy, some companies are at the forefront and others are lagging [1]. A larger study of AI usage in Nordic countries based on interviews conducted in 2021 found that 70% of Nordic companies' overall use of AI as part of their product and 55% were planning to experiment with AI techniques within the next 6 months [12]. However, the threshold of market entry is rising, and business development is dominated by large enterprises. In Norway, only 9% of small enterprises used at least one AI technology, while the adoption rate was twice as high for medium-sized enterprises. For large enterprises (250 + employees), 43% used more than one AI application. Roughly 76% of companies answered that they have plans to hire new AI professionals within the next 6 months.

Small and medium enterprises (SMEs) often play a crucial economic role in many countries, especially in less urban areas, as they stimulate innovation, provide jobs, foster competitiveness, and contribute to overall economic growth [10]. Several studies have assessed and reviewed why SMEs face various challenges in adopting emerging technologies such as artificial intelligence (AI) [e.g., 11, 13]. Here, key challenges in SMEs are found to be related to a lack of necessary strategies and knowledge (often due to a lack of knowledge among leaders), lack of talents and skills, and a lack of resources. The latter is often found to be related to economic shortage or being unable to obtain large amounts of high-quality data and use cases in small companies [11, 13].

3 Methodology and Empirical Data

A 6-year-long research project, "Teknoløft Sogn og Fjordane", was financed to build competence in digitalization and automation, robotics, and big data at the Western Norway University of Applied Sciences (HVL) and the Western Norway Research Institute (Vestlandsforsking) and conduct research in close cooperation with local enterprises in the region of Sogn og Fjordane. The overall target of the project is to increase the role of research in local industry innovation.

Sogn og Fjordane has a population of 109 000 spread over an area of 18 623 km^2 [14]. This makes it one of the most rural regions in Norway in terms of population. According to Statistics Norway, 99% of companies in Sogn og Fjordane have less than 49 employees, and 90% of companies have 9 or fewer employees [15] (Fig. 1). Agriculture, forestry, aquaculture, and domestic trade occupy the largest sector among small companies (0–20 employees). The latter sector has an annual turnover of almost 14 billion (as of July 26, 2021), which is ~30% of Sogn og Fjordane's total annual turnover. Some larger businesses with more than 100 employees are in education, human health and social work, manufacturing, transportation, and storage, in addition to maritime industries. The region is abundant in natural resources, which has enabled the development of industries within agriculture, aquaculture, hydroelectricity, aluminium, and petroleum production. Several of the oil and gas fields that contribute to the country's welfare are located right outside the coastline of the region. The coastal city of Florø is, for example, Norway's largest supply base for the petroleum industry [16]. The landscape of the region is otherwise largely characterized by lush green fields, mountains, fjords, and glaciers, the most notable of which is the Sognefjorden, the longest fjord in Norway and the second longest in the world. The area has seen an influx of people and businesses due to the trend of younger generations seeking a more rural lifestyle close to nature. Analytics suggest that the region is expected to experience an increase in population over the next two decades, contradictory to many other rural regions in Norway [17].

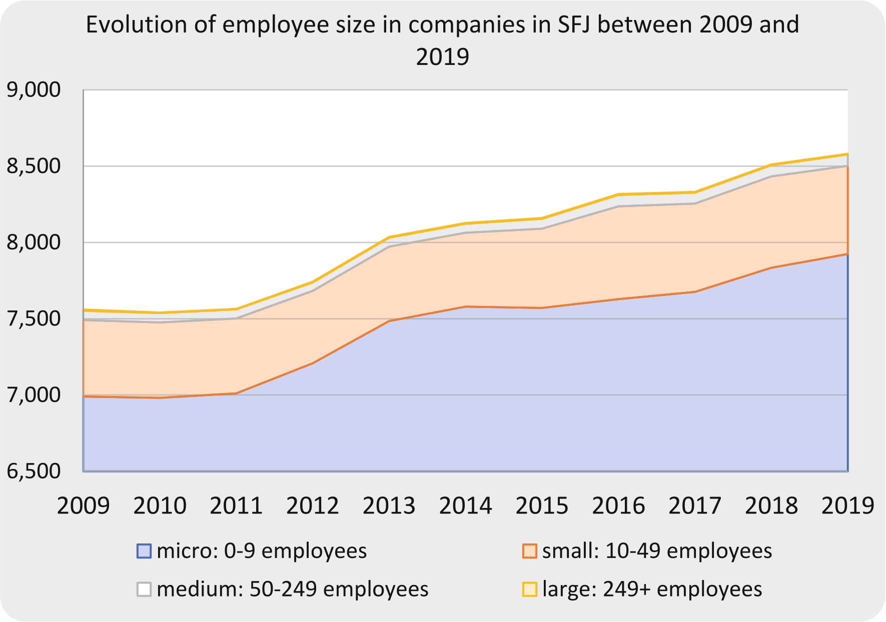


Fig. 1. Evolution of employee size in companies in SFJ between 2009 and 2019 (pre-COVID evolution) 2019 was also a year when Sogn og Fjordane became a part of the larger county of Vestland. As seen, the number of active micro-companies has grown significantly during this period.

3.1 Survey

To map the status of digital maturity and AI adoption among enterprises based in Sogn og Fjordane, we collected answers through a questionnaire. The first part of the questionnaire entails general information about the company and the work position of the respondent. The main part presents many allegations essentially about the company's thoughts and plans on digitalization and AI, while in the last voluntary part, the respondent can explain with their own words their plans, challenges, and competence needs regarding digitalization. The last part is an addition to the otherwise identical questionnaire sent out to businesses in other nearby regions.

We got answers from 25 respondents from various companies across a range of sectors (Fig. 2). However, three companies that answered are based outside the region and therefore excluded from the survey responses.

An almost identical questionnaire, "Digital maturity in western Norway" was collected by Bergen Næringsråd and others in 2021 (BN survey), where they got ~350 replies [18]. However, even though the region of "Sogn og Fjordane" is a significant part of western Norway, no businesses from that region were represented in that survey [18]. Hence, when collecting answers from this missed region, the similarity of the questionnaires was important to enable a future comparison and a merge of the answers from the greater region.

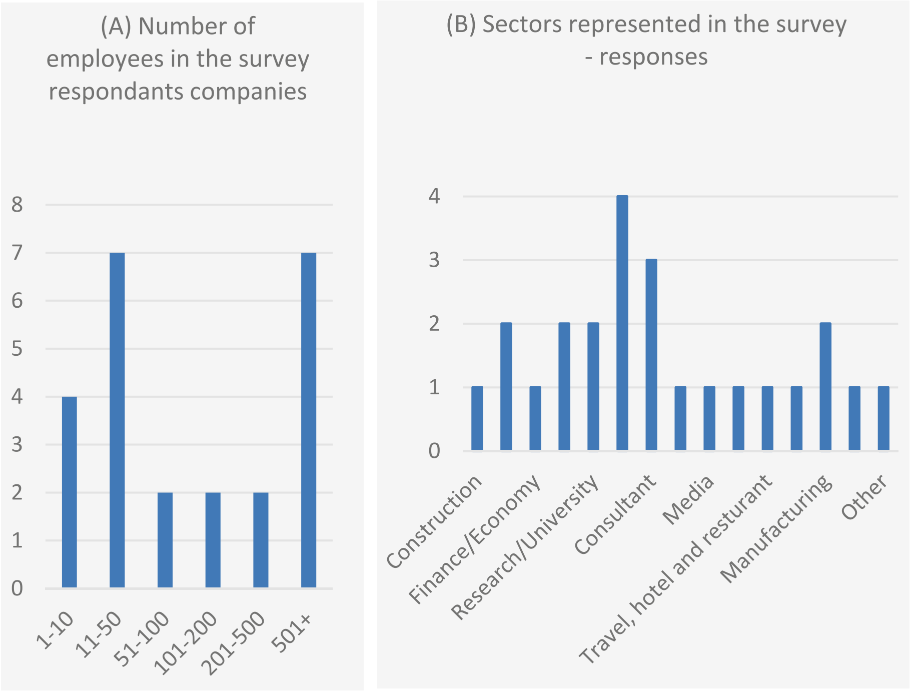


Fig. 2. (A) Employee-size and (B) sectors of enterprises that joined the survey

4 Hurdles of Data Leverage and AI Implementation Among Regional SMEs

Figures 3 and 4 present the survey results. 10 of 21 respondents answered positively regarding their company's understanding of how AI, machine learning, or robotics can affect their business development. Furthermore, 11 agrees that their company is aware of how they could change or improve with data utilization, and all of those agreeing to that statement also state that their company has concrete plans or strategies related to exploiting their data. Five companies answered that they are lacking concrete plans to leverage their data, and all of those also disagreed that their companies understood the value of AI from an enterprise perspective. Nine respondents answered that AI is complicated and that it is hard to know where to start with it. When comparing small-medium-sized and large enterprises regarding future and AI understanding, the answers are comparable, meaning that we cannot identify a trend of less or more digital maturation depending on the company size based on the respondents' answers (Fig. 3).

7 of the respondents expressed apprehension that the key employees may not have enough understanding about the significance of upcoming technological challenges, and 6 confirm that there is limited or no financial framework assigned for digital development within their companies. Furthermore, 10 companies answered that they only prioritize simple digitalization work (Fig. 3).

A noticeable inclination was towards inadequate regular competence training to enhance employees' digital competency, with 11 companies agreeing to this statement. Maybe a significant point, as 8 expressed difficulty themselves in comprehending AI and machine learning, feeling that it is a complex area to initiate with. Five of the respondents were neutral to that question. In addition, five respondents anticipate that AI and robots may become their biggest rivals in the future.

Despite these findings, 16 of the respondents reported that their company has definite plans to tackle future technological challenges (Fig. 4). The three that disagree that their company has plans for tackling future technological challenges partly or strongly agree that they are worried if enough of the key employees in their company understand the seriousness of future technological challenges.

As much as 6 of 7 leaders and managers agree that their company understands how AI and/or robotics can change their business, while significantly less than 4 of 12 regular employees agree with the same statement (Fig. 4). The leaders that agree state that their company also understands the value they can create with their data and has plans or strategies to exploit their data better in the near future. Interestingly, 3 of these leaders also agree that AI is an area that mainly concerns the IT department or IT-responsible and 2 of the respondents who agree that the company understands how AI and/or robotics can change their business also agree that AI is complicated and it's hard to know where to start.

The ones disagreeing that their company has plans or strategies regarding the exploitation of data in the future (6 respondents) are either not aware of the value they can create with their data (5) or neutral to that (1), and 5 of 6 are worried if enough of the key employees in their company understand the seriousness of future technological challenges.

To conclude, it seems that a significant portion of the respondents, especially leaders, believe that their companies have a good understanding of how AI and data utilization can impact their business development, and these are companies that also have plans for the exploitation of their data and are facing future technological challenges. Some companies lack concrete plans or strategies to exploit their data, and those are also the companies that express apprehension or limited understanding of upcoming technological challenges, including AI, and would only prioritize cost-efficient and simple digitalization work. There is an inclination towards inadequate regular competence training to enhance employees' digital competency, and there seems to be a discrepancy between leaders, managers, and employees in terms of their perception of the company's understanding of AI and robotics, with leaders being more positive.

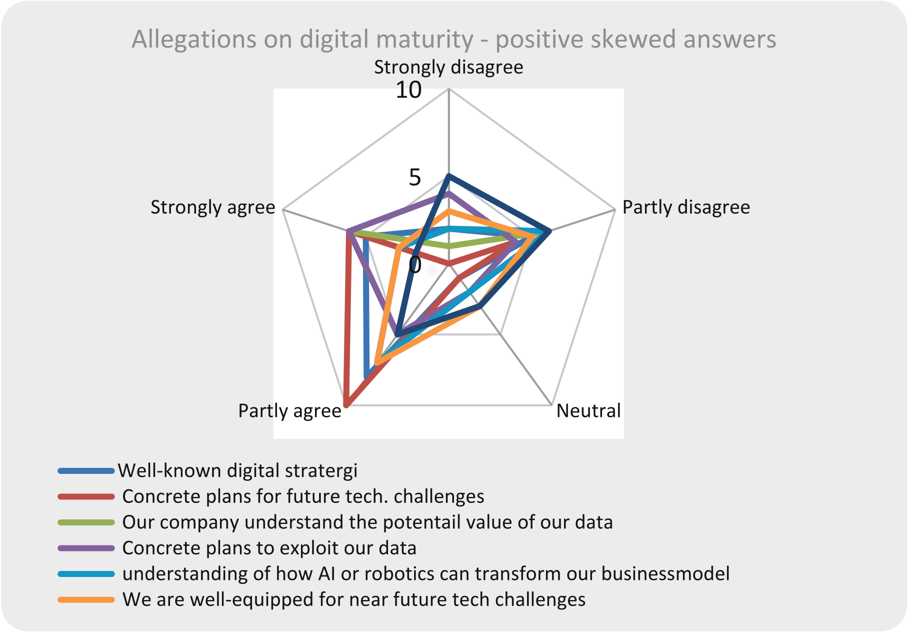


Fig. 3. Allegation answers from the survey on digital maturity - positive skewed answers

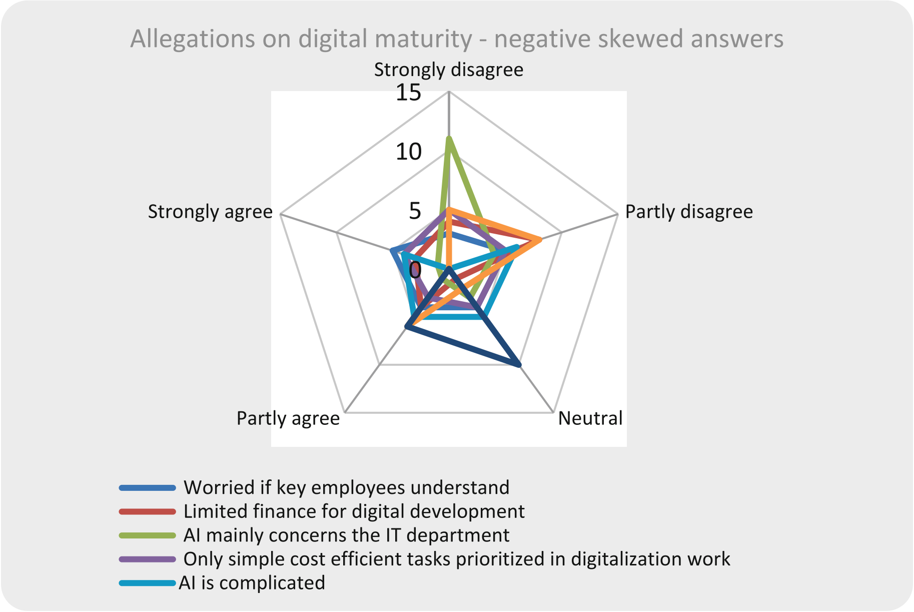


Fig. 4. Allegation answers from the survey on digital maturity - negative skewed answers.

5 Usage of AI in Regional Key Sectors

Here, we will present some examples of AI usage in key SMEs—sectors of the region—to shed some light on the possibilities within these sectors regarding data utilization and AI.

5.1 AI Potential in Agriculture

The experience of AI implementation in agriculture and aquaculture worldwide has shown great potential for higher value creation as well as often affecting sustainability positively [19]. The agriculture sector is increasingly collecting data that can be used for AI applications, or so-called precision agriculture. Sensor data, images, and satellite images are used to classify and detect different phenomena and objects, as well as for performing sorting procedures. Some examples of the monitored objects include weed and disease detection, yield predictions, soil and water management, species recognition, crop quality, animal welfare, and livestock production [20].

A specific target known as one of the main challenges in agriculture's crop production is weeds and diseases [20]. Accurate weed and disease detection is necessary for sustainable agriculture because weeds are competing for the same resources as crops in terms of nutrients in the soil and sunlight. Today, pesticides are one of the primary methods for combating weeds and diseases; however, pesticides contaminate soil and drinking water, especially in Norway, where the low temperature makes the breakdown of pesticides slower [21]. Machine learning algorithms, in conjunction with sensors and images, can enable the detection and identification of weeds and diseases without causing environmental issues or secondary effects [20]. This may be executed with cameras mounted on drones, four-wheels, or tractors in the field, which automatically take pictures of the same location and of the same objects in the field.

Moreover, image processing and analysis of the pictures may involve counting fruits, classification of colors, and eventually appearance or changes in fine-scaled objects on the fruits, such as diseases. While some manufacturers invest in AI-based sorting machines, most sorting today is still done manually [13]. AI-based apps integrated with machines have the potential to be more efficient, less biased, and more accurate, in addition to replacing boring, repetitive work at workplaces [22, 23]. At fruit and greens reception sites, fresh food turnover time, product quality, and waste reduction are essential. An AI-based sorting machine can use automated image recognition for quality checks and sorting and significantly speed up the sorting process, as well as making it more accurate. The algorithms can be trained to identify flaws, contaminants, or product defects [23]. In Sogn og Fjordane, several such sorting machines are already placed at the region's largest fruit and greens reception site. Experiences of AI implementation in agriculture and aquaculture worldwide has shown great potential for higher value creation as well as often affected sustainability positively [19]. The agriculture sector is increasingly collecting data that can be used for AI applications, or so-called precision agriculture. Sensor data, images, and satellite images are used to classify and detecting different phenomena and objects, as well as for performing sorting procedures. Some examples of the monitored objects include weed and disease detection, yield predictions, soil and water management, species recognition, crop quality, and animal welfare and livestock production [20].

A specific target known as one of the main challenges in agriculture of crop production, is weeds and diseases [20]. Accurate weed and disease detection is necessary for sustainable agriculture because the weeds are competing for the same resources as crops in terms of nutrients in the soil and sunlight. Today pesticides are one of the primary methods for combating weeds and diseases, however pesticides contaminate soil and drinking water, especially in Norway as the low temperature makes the breakdown of pesticides slower [21]. Machine learning algorithms in conjunction with sensors and images can enable detection and identification of weeds and diseases without causing environmental issues or secondary effects [20]. This may be executed with cameras mounted on drones, four-wheels or tractors in the field which automatically takes pictures on the same location and of the same objects in the field.

Moreover, image processing and analysis of the pictures may involve counting fruits, classification of colors and eventually appearance or changes in fine-scaled objects on the fruits such as diseases. While some manufactures invest in AI- based sorting machines most sorting today is still done manually [13]. AI-based apps integrated with machines have potential to work more efficient, less biased more accurate, in addition to replace boring repetitive work at workplaces [22, 23]. At fruit and greens reception sites, fresh food turnover time, product quality and waste reduction are essential. An AI-based sorting machine can use automated image recognition for quality checks and sorting and significantly speed up the sorting process as well as making it more accurate. The algorithms can be trained to identify flaws, contaminants, or product defects [23]. In Sogn og Fjordane, several such sorting machines are already placed at the region's largest fruit and greens reception site.

5.2 AI Potential in Aquaculture

One of the largest industry sectors in Norway is the aquaculture sector [24]. An already-published report presents an overview of the most common use of big data in aquaculture and comes up with recommendations that can be applied to businesses within the Sogn og Fjordane region. The report also highlights areas of research needs that could potentially be converted into research projects for businesses in the region [25]. These recommendations are, i.e., water quality combined with fish behavior monitoring using sensors and underwater cameras, techniques to make fish feeding more efficient and reduce feed waste (identified as one of the largest challenges in aquaculture), and disease identification using underwater cameras and biosensors. It has been shown to contribute to improved operations, leading to faster production rates, better asset control and prediction, and enhanced safety and efficiency. Due to the implementation of data-driven computing, it is possible to predict sea lice two weeks in advance.

Additionally, enormous quantities of data are also currently either in the planning phase or being collected related to a global push to track small-scale fishing farms and vessels [25]. These quantities of data are of great variety, ranging from fishways, fishing boat administration, marine satellite images, readouts of fish farms, tides, weather, etc. While this data is increasing and becoming available as sensors and tracking become more widespread, companies are also concerned with managing the substantial complexity of their data [25].

5.3 AI Potential in the Hydroelectric Energy Sector

Like aquaculture, the energy sector is experiencing a similar increase in sensory data that you also see in aquaculture. Continuous measurements of grid equipment and related parameters can be used in predictive analyses for developing maintenance strategies [26]. The most common and well-known source is the smart meter. Prediction of outages in the power grid is essential to having a reliable power grid. Predicting the influence of weather, smart meter events, and location will help identify root causes. For example, researchers at GE Research used GIS (geographic information systems) data and satellite imagery together with survival models to assess outage risk [27]. This was used to find outages related to vegetation. It has been found that vegetation events are not only due to growth but also to extreme weather events. Integrating weather forecasts into the systems for outage prediction is necessary for even better predictions. Other data sources more relevant to the region, such as SFJ, may also be a point of research in the future. Another approach may be to use machine learning and Bayes decision theory to find the optimal decision boundary. The proposed classifier provides an effective framework that not only minimizes outage prediction errors for power system components but also considers the cost of each preventive action according to its implication in extreme events [28].

Smart meters are also installed at the end-user's location and track energy consumption with a resolution as high as minutes or even seconds [29]. The smart electricity grid enables a two-way flow of power and data between suppliers and consumers to facilitate power flow optimization in terms of economic efficiency, reliability, and sustainability. This infrastructure permits consumers and micro-energy producers to take a more active role in the electricity market and dynamic energy management. Robust data analytics, high-performance computing, efficient data network management, and cloud computing techniques are critical to the optimized operation of smart grids [30]. Weather data, thermostats, real estate data, and energy behavior integrated with energy demands can provide better forecasting and prediction services [26]. GIS also has an important role in the sector. Data from GIS sources can provide valuable information that can be used in decision-making systems since it provides local geographic information for many issues, such as identifying optimal locations of solar farms [26].

Furthermore, forecasting the power generation potential, or streamflow to power plants, forward in time using AI-based time-series methods and weather forecasts has shown great accuracy and uses, providing a potentially powerful new tool for renewable power companies that have not already implemented the method [31].

5.4 AI Potential in Manufacturing

A food manufacturer company called Danone Group started to use AI to improve their demand forecast accuracy, which they report has led to a 30% decrease in lost sales and a 50% reduction in the demand planner's workload [32]. Manufacturing, which is also large in Sogn og Fjordane, is another interesting sector as it has great potentials regarding circular economy, sustainability, and otherwise high AI potentials [33]. By using AI, businesses can react and have better control of market responses, identify trends in demands and consumer behaviors, conduct forecasts, and automate complex tasks and decision support, therefore maybe freeing up employees to more interesting innovation tasks [34].

Manufacturing also involves product lifecycle management, operation and maintenance, energy management, and supply chain management. The ability to track assets such as products, equipment, tools, and inventory in real-time is an essential component of AI manufacturing optimization in I4.0 [35]. A natural first step towards AI implementation is adapting the existing asset tracking system, transforming it into a real-time location system, and preparing it for the Industrial Internet of Things. The addition of RTLS to the IIOT environment of smart manufacturing can provide not only accurate asset tracking but also additional benefits for parts inventory, tool management, and personnel supervision [35].

With dataflow and product monitoring in place, various AI methods can be implemented to boost productivity, storage optimization, and sales.

Augmented reality has further been proposed as a disruptive and enabling technology within the I4.0 manufacturing paradigm. The term augmented reality (AR) has been defined as a system that "supplements the real world with virtual (computer-generated) objects that appear to coexist in the same space as the real world" [36]. Augmented reality may be used for troubleshooting and support. It allows experts to remotely assist in cases where they traditionally needed to be on location.

In contrast with maintaining equipment when the need arises, predictive maintenance, together with intelligent sensors, has also emerged as a new approach to maintenance in I4.0. Intelligent sensors make it possible to obtain an ever-increasing amount of data, which must be analyzed efficiently and effectively to support increasingly complex systems' decision-making and management [37]. Using AI methods to predict when production machinery needs maintenance and for which parts, many benefits arise. For instance, predicting what spare parts need to be in stock at what time [37].

5.5 AI Potential in the Human Health Sector

The rapid innovation of analytics and data-driven technology within the human health sector is taking place worldwide. This holds significant potential for medical diagnosis by AI, natural language processing for assessing mental health [38, 39], and identifying high-risk and high-cost patients. Big data spaces offer the ability to establish an observational evidence base for clinical questions that couldn't otherwise be answered. This could be particularly useful with issues of generalizability, allowing data-driven clinical decision support tools that may lead to cost savings and promote appropriate standardization of care. Clinicians could receive messages that inform them of the diagnostic and therapeutic choices made by respected colleagues facing similar patient profiles [40].

6 Discussion and Outlook

According to both our and the BN survey, many companies still find it challenging to incorporate AI, machine learning, and robotics, despite some progress. A slightly lower percentage of our respondents (9 of 21, 43%) compared to the BN survey (45.6%)

answered that AI is complicated and it is hard to know where to start. Lack of understanding how AI can help their business may come from a lack of knowledge, but it can also be that AI can't improve their business. Whereas 1 of 2 respondents in the SFJ agrees that they have plans for more data utilization, 2 of 3 respondents in the BN survey agree with the same statement. Interestingly, the BN survey found that leaders have a significantly lower understanding than employees regarding the criticality of AI and data for future business operations. However, our survey of businesses in Sogn og Fjordane showed the opposite result. Leaders have a better understanding of the criticality of AI in their company's future business development compared to employees. Is it easier to convey change either in a bottom-up or top-down way in the district? Nevertheless, further research is necessary to determine if this difference exists for the region of Sogn og Fjordane, if it is a sample-size issue, or if in only one year the opinions of leaders have changed due to rapid changes in the markets and new understanding.

Although 12% less of the SFJ respondents compared to the BN survey agree that they are worried they don't have enough competence to meet future technological challenges, a large number of companies in both surveys do not have regular competence training or finance for digital development. These aspects may be hindering enterprises from fully utilizing data-driven technologies and taking full advantage of their potential data. The absence of regular training to improve digital competency may also hinder the company's ability to adapt and take advantage of new digital opportunities.

The study of various regions and their digital maturity is important since regions may have varied bases for the advancement of technology, adoption of digitization, and implementation of AI. The potentially large difference between the degree of leaders versus employees that understand how AI can change their business compared to the BN survey and our survey may indicate a regional difference. In smaller, more rural areas, the competition among businesses is generally less than in cities, and the work structure may be flatter due to smaller offices and less work-private distance. The latter may make it easier to convey change in a bottom-up or top-down way. The region also has some large and critical national digital services that are developed and operated by public entities located in Sogn og Fjordane (i.e., the digitalization directory in Norway). SMEs may be facing these challenges due to low capacity and a lack of financial and skilled human resources, preventing them from competing internationally and investing in training to ensure that employees are equipped with the skills and understanding required to increase the business's competitiveness and robustness for future challenges.

Currently, there is limited research being conducted in companies located in Sogn og Fjordane, as evidenced by the low number of companies receiving support from Norwegian and European research funds. However, this does not necessarily indicate a lack of development in companies. Western Norway, which includes Sogn og Fjordane, is ranked highest in Norway for business-related development. However, development can also be achieved through investments in new machines, systems, and consulting services that have been launched in the market. Instead, networking, consulting with non-profit organizations, and investing in more research for innovation, growth, and development may be a good strategy to develop affordable skills internally, make collaborations across nations, and implement tailored AI solutions. Zooming out to a Nordic perspective on AI adaptation, key challenges found involve data management, transitioning from pilots to production, understanding regulations, and developing ethical and fair AI solutions, which are also likely aspects of regional challenges [12].

A generally low financial allocation for digital development may also indicate that companies may be reluctant to invest in these technologies and their training. Challenges faced among the surveyed companies may have more economical than "limited understanding of AI's potential or digital readiness" aspects. Documentation of the impact of specific AI-use cases across sectors will likely be valuable towards a more predicable understanding of how data collection and AI may change businesses. Furthermore, incremental steps and a focus on cost-efficient actions will help gain knowledge and insights into data management and applicable AI by, for example, utilizing non-profit agencies for support and research engagements and collaborations. Clustering and networking with companies that face similar challenges may also be good strategies to bolster SMEs positions and effectively navigate the challenges that lie ahead. Companies that may utilize the same technology and that are not in direct competition with each other can benefit from co-joining projects and collaborating to learn from each other's successes and failures. The company's willingness to invest in AI technologies essentially depends on the economic output, which answers from this survey may also suggest: the company must generally be able to see a concrete positive impact and how big that impact is [41]. Such a realistic perspective is healthy, but there will also be cases where AI is not having enough impact to make a difference for the company or may in some way have a negative impact. However, few companies measure the success of AI implementation [13], and there is a great need for future research and companies to document the effects of AI adoption. Making more user stories and how-to's with successes and failures of digital developments and AI implementation available in "safe/non-competitive" spaces can be valuable for faster and leaner change among SMEs.

7 Conclusions

The respondents see challenges in investing in AI but have a slightly greater understanding and ambition for it. We found that most companies surveyed have concrete plans to tackle future technological challenges, but not necessarily for data utilization. Our survey may indicate that enterprises in SFJ are not performing less regarding digital maturity, capacity, and human resources than enterprises in the larger western Norway more urban regions. Our survey, in fact, shows a better understanding among the managers and leaders surveyed. "The more you know, the more you don't know", is a common expression. Seeing potential and seeing what one must mobilize to solve a challenge is maybe a better way than thinking you have everything you need. The findings of the survey seem generally to be in line with other large surveys in the western Norway region, the Nordics, and Germany, which is positive news for a rural but rich region.

The two quantitative surveys in Vestland need to be followed up on with a new BN survey that better incorporates SFJ and rural areas as well as with a number of qualitative interviews that focus on the topics in-depth. The information provided by this strategy will be more accurate and thorough, and stakeholders can use it decision-making processes.

References

1. PwC: Sizing the Prize. What's the real value of AI for your business and how can you capitalise? (2020). https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf
2. Kyrkjebø, E.: Teknoløft Sogn og Fjordane (2018). https://www.hvl.no/en/project/573142/
3. Samoili, S., Cobo, M.L., Gómez, E., De Prato, G., Martínez-Plumed, F., Delipetrev, B.: AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence (2020)
4. Hermann, M., Pentek, T., Otto, B.: Design principles for industrie 4.0 scenarios. In: Presented at the 2016 49th Hawaii International Conference on System Sciences (HICSS) (2016)
5. Taylor, P.: Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 (2023). https://www.statista.com/statistics/871513/worldwide-data-created/
6. Singh, M., Goyat, R., Panwar, R.: Fundamental pillars for industry 4.0 development: implementation framework and challenges in manufacturing environment. TQM J. ahead-of-print (2023). https://doi.org/10.1108/TQM-07-2022-0231
7. Rüßmann, M., et al.: Industry 4.0: the future of productivity and growth in manufacturing industries. Boston Consulting Group **9**, 54–89 (2015)
8. Bauernhansl, T., Krüger, J., Reinhart, G., Schuh, G.: WGP-Standpunkt Industrie 4.0. (2016)
9. Rajput, S., Singh, S.P.: Connecting circular economy and industry 4.0. Int. J. Inform. Manage. **49**, 98–113 (2019)
10. Gherghina, Ștefan C., Botezatu, M.A., Hosszu, A., Simionescu, L.N.: Small and medium-sized enterprises (SMEs): the engine of economic growth through investments and innovation. Sustainability **12** (2020). https://doi.org/10.3390/su12010347
11. Lu, X., Wijayaratna, K., Huang, Y., Qiu, A.: AI-enabled opportunities and transformation challenges for SMEs in the post-pandemic era: a review and research agenda. Front. Public Health **10**, 885067 (2022)
12. Silo AI: Nordic state of AI 2022 (2020). https://static1.squarespace.com/static/5dd533f44eb1de01971d74a0/t/636914027a68e03a5fbb8f2a/1667830802535/Report_Nordic_State_of_AI_2022.pdf
13. Bunte, A., Richter, F., Diovisalvi, R.: Why it is hard to find AI in SMEs: a survey from the practice and how to promote it. In: Presented at the ICAART, vol. 2 (2021)
14. Askheim, S.: Sogn og Fjordane (tidligere fylke) (2021). https://snl.no/Sogn_og_Fjordane_-_tidligere_fylke
15. SSB, 07091: Virksomheter, etter næring (sn2007) og antall ansatte (k) 2009 – 2021 (2021). https://www.ssb.no/statbank/table/07091/
16. Wikipedia: Fjord base (2020). https://no.wikipedia.org/wiki/Fjord_Base
17. Vareide, K.: Nye scenarier med oppdaterte 2020-tall (2022). https://regionalanalyse.no/artikkel/prognose2021
18. Warnacke et al.: Digital Modenhet på Vestlandet Delrapport 1: kunstig intelligens (2020). https://d3gkcpa86cdznk.cloudfront.net/1669293750/kartlegging-digital-modenhet-delrapport-1.pdf
19. Eli-Chukwu, N.C.: Applications of artificial intelligence in agriculture: a review. Engineering. Technol. Appl. Sci. Res. **9** (2019)
20. Cravero, A., Sepúlveda, S.: Use and adaptations of machine learning in big data—applications in real cases in agriculture. Electronics **10**, 552 (2021)
21. Almvik, M., Eklo, O.M., Stenrød, M., Nyborg, Å.A., Hole, H.: Plantevernmidler i miljøet i jordbruket i Norge (2016)

22. Strollo, E., Sansonetti, G., Mayer, M.C., Limongelli, C., Micarelli, A.: An AI-Based approach to automatic waste sorting. In: Presented at the HCI International 2020-Posters: 22nd International Conference, HCII 2020, Copenhagen, Denmark, 19–24 July 2020, Proceedings, Part I 22 (2020)
23. Kumar, I., Rawat, J., Mohd, N., Husain, S.: Opportunities of artificial intelligence and machine learning in the food industry. J. Food Qual. **2021**, 1–10 (2021)
24. Afewerki, S., Asche, F., Misund, B., Thorvaldsen, T., Tveteras, R.: Innovation in the Norwegian aquaculture industry. Rev. Aquac. **15**, 759–771 (2023)
25. Akerkar, R., Hong, M.: Big data in aquaculture (2021). https://www.vestforsk.no/nn/publication/big-data-aquaculture
26. Akerkar, R., Hong, M.: Big data in electric power industry (2021). https://www.vestforsk.no/nn/publication/big-data-elektric-power-industry
27. Jain, A., Shah, T., Yousefhussien, M., Pandey, A.: Combining remotely sensed imagery with survival models for outage risk estimation of the power grid. In: Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
28. Mohammadian, M., Aminifar, F., Amjady, N., Shahidehpour, M.: Data-driven classifier for extreme outage prediction based on Bayes decision theory. IEEE Trans. Power Syst. **36**, 4906–4914 (2021)
29. Stegner, C., Bogenrieder, J., Luchscheider, P., Brabec, C.J.: First year of smart metering with a high time resolution—realistic self-sufficiency rates for households with solar batteries. Energy Procedia. **99**, 360–369 (2016)
30. Diamantoulakis, P.D., Kapinas, V.M., Karagiannidis, G.K.: Big data analytics for dynamic energy management in smart grids. Big Data Res. **2**, 94–101 (2015)
31. Niu, W., Feng, Z.: Evaluating the performances of several artificial intelligence methods in forecasting daily streamflow time series for sustainable water resources management. Sustain. Cities Soc. **64**, 102562 (2021)
32. Khan, H., Kushwah, K.K., Singh, S., Thakur, J.S., Sadasivuni, K.K.: Machine learning in additive manufacturing. Nanotechnol.-Based Addit. Manuf. Prod. Des. Properties Appl. **2**, 601–636 (2023)
33. Bag, S., Gupta, S., Kumar, S.: Industry 4.0 adoption and 10R advance manufacturing capabilities for sustainable development. Int. J. Prod. Econ. **231**, 107844 (2021)
34. Kagermann, H., Wahlster, W.: Ten years of Industrie 4.0. Sci. **4**, 26 (2022)
35. Krishnan, S., Santos, R.X.M.: Real-time asset tracking for Smart Manufacturing. In: Toro, C., Wang, W., Akhtar, H. (eds.) Implementing Industry 4.0: The Model Factory as the Key Enabler for the Future of Manufacturing, pp. 25–53. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-67270-6_2
36. van Lopik, K., Sinclair, M., Sharpe, R., Conway, P., West, A.: Developing augmented reality capabilities for industry 4.0 small enterprises: lessons learnt from a content authoring case study. Comput. Ind. **117**, 103208 (2020)
37. Pech, M., Vrchota, J., Bednář, J.: Predictive maintenance and intelligent sensors in smart factory. Sensors. **21**, 1470 (2021)
38. Bates, D.W., Saria, S., Ohno-Machado, L., Shah, A., Escobar, G.: Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff. **33**, 1123–1131 (2014)
39. Silahtaroğlu, G., Yılmaztürk, N.: Data analysis in health and big data: a machine learning medical diagnosis model based on patients' complaints. Commun. Stat. –Theor. Methods **50**, 1547–1556 (2021)
40. Murdoch, T.B., Detsky, A.S.: The inevitable application of big data to health care. JAMA **309**, 1351–1352 (2013)
41. Bughin, J., Hazan, E.: The new spring of artificial intelligence: a few early economies. VoxEU.org. 21 (2017)

Author Index

R. Akerkar (Ed.): SAIDD 2023, CCIS 1810, p. 203, 2024.
https://doi.org/10.1007/978-3-031-53770-7