TU WIEN Informatics

Efficient Temporal Graph Analytics

Using large scale telecommunication data for mobility modeling and infrastructure maintenance

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktor der Technischen Wissenschaften

by

Dipl.-Ing. Georg Heiler, BSc
Registration Number 1225063

to the Faculty of Informatics

at the TU Wien

Advisor: Prof. Allan Hanbury
Second advisor: Prof. Stefan Thurner

The dissertation has been reviewed by:

Leo Ferres	Axel Polleres

Vienna, 9th September, 2022

Georg Heiler

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Dipl.-Ing. Georg Heiler, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 9. September 2022

Georg Heiler

Acknowledgements

I want to thank everyone who has contributed directly or indirectly to the success of this work.

I would particularly like to thank my parents, Barbara and Klaus, who laid the foundation for my education. Over the years, they have confirmed my decisions, supported and encouraged me, and given me valuable strength for my life path through their love. I would also like to thank my grandmother Gertrud, who motivated me to never give up, especially in difficult times.

Furthermore, Prof. A. Hanbury (E-Commerce, TU Wien) should be mentioned, who, as supervising professor of the dissertation, made an essential contribution to the success of this work through fruitful discussions. I want to thank Prof. S. Thurner (Medical University Vienna, Section for Complex Systems) for the second supervision, the orientation and the many wonderfully exciting projects on which we have worked together. I want to thank A. Prof. P. Klimek (Medical University Vienna, Section for Complex Systems) for excellent feedback and calm despite the time pressure during the many analyses about COVID-19. I want to thank Prof. P. Filzmoser (Computational Statistics Group, TU Wien).

I want to thank my colleagues Tobias Reisch and Christian Diem for their excellent cooperation and the many late-night discussions.

For more than five years, besides my work at the university (TU Wien and the Complexity Science Hub), I have been working at T-Mobile Austria as a data scientist. In particular, I would like to thank Dominik Steinschaden (T-Mobile Austria) for the successful cooperation that lasted for over three years. Many valuable ideas have emerged in countless discussions and new approaches have been implemented. I want to thank all Magenta colleagues for the friendly working atmosphere and the opportunity to participate in exciting projects. Many thanks also to the data engineering team, who always gave terrific support and the operations and network technology departments, who always patiently satisfied my thirst for knowledge.

A special thank you goes to my friends who have shown a lot of understanding for me and my work, especially in the last few months.

Furthermore I want to thank the people who supported during the publications: Eva Bauer, Barbara Prainsack and Hannah Metzler for helpful discussions.

This work was supported by: The Austrian Science Fund FWF under projects P29252, 3073-N32 and I3073, the Austrian Research Promotion Agency (FFG) under projects 857136 and 873927, the WWTF under projects COV 20-017 and COV 20-035 and the Medizinisch-Wissenschaftlicher Fonds des Bürgermeisters der Bundeshauptstadt Wien under project CoVid004 and Hochschuljubiläumsstiftung of the Austrian National Bank OeNB under P17795 2018-2021.

A big thank you also goes to the various partners who made anonymized data available for research.

Georg Heiler

Kurzfassung

Daten mit einer Graphenstruktur und Dimensionen in Raum und Zeit stellen die Verbindung zwischen Ereignissen in der realen Welt und ihrer abstrakten Darstellung dar, wie in etwa Verhaltensänderungen in Gesellschaft oder Technologie. Durch die Verarbeitung solcher Daten werden Erkenntnisse gewonnen, die Auswirkungen auf die reale Welt haben können.

Die bei einem Internet Service Provider vorhandenen Datensätze enthalten üblicherweise derartige Dimensionen. Solche Daten können beispielsweise durch die Nutzung von Mobiltelefonen oder die Telemetriedaten eines Kabelnetzes erzeugt werden. Ersteres kann sehr nützlich sein, um die Veränderung der Eigenschaften unserer Gesellschaft auf nationaler Ebene zu bestimmen, und letzteres für die vorausschauende Wartung des Netzwerks. Beim Umgang mit solchen Datensätzen ist die Skalierbarkeit insbesondere bei kostenintensiven Operationen wie Geodaten- oder Graph-algorithmen wichtig. Wir entwickeln verteilte skalierbare Bausteine für Geo-operationen oder führen intelligente Aggregationen durch. Diese Nutzen wir zur Analyser der Auswirkungen der Non Pharmaceutical Interventions auf die Gesellschaft.

Das systemische Risiko von Unternehmen waren bisher nicht quantifizierbar, da Liefernetzwerke auf Unternehmensebene mit Ausnahme einiger weniger Länder nicht existierten. Hier rekonstruieren wir aus Telekommunikationsdaten flächendeckende unternehmensweite Versorgungsnetze. Die daraus resultierenden Netzwerke erlauben es uns, das systemische Risiko einzelner Unternehmen zuverlässig zu quantifizieren und damit die wirtschaftliche Widerstandsfähigkeit eines Landes abzuschätzen. Die Methode kann zur objektiven Analyse von Veränderungen in Produktionsprozessen eingesetzt werden, die für die grüne Wende unerlässlich werden könnten.

Durch die vorausschauende Wartung des Kabelnetzes könnten Auswirkungen im Unternehmensbereich erzielt werden. Für hybride Glasfaser-Koaxial-Netzwerke war die Suche nach starkem Rauschen im Upstream Kanal in der Vergangenheit umständlich und zeitaufwändig. Wir präsentieren die Automatisierung einer einfachen Geschäftsregel (größte Änderung eines bestimmten Werts) und vergleichen ihre Leistung mit modernsten maschinellen Lernmethoden und kommen zu dem Schluss, dass die top-1 Genauigkeit um das 2,3-fache verbessert werden kann.

Abstract

Behavioral changes in society or technology can be represented as a graph with dimensions in space and time. Such graphs represent the link between events in the real world and their abstract representation. By analyzing such data, insights are derived, impacting decisions taken in the real world.

The datasets collected at a telecommunication company commonly contain these dimensions; for example, the usage of mobile phones or the telemetry of a cable modems in a network. The former can be helpful to determine the change of characteristics of society and its behavior at the scale of whole countries and the latter for predictive maintenance of the network. The scalability of particularly costly operations such as geospatial or graph algorithms is essential when handling such data sets. We develop distributed scalable primitives here for geospatial operations or perform smart aggregations. These primitives are applied to analyze the impact of non-pharmaceutical interventions (e.g. lockdowns) on society.

Systemic risk is the possibility that an event at the company level could trigger severe instability or collapse an entire industry or economy. The Systemic risk contribution of companies was hitherto not quantifiable since supply networks on the company-level did not exist except for very few countries. Here we use telecommunication data to reconstruct nationwide company-level supply networks. The resulting networks allow us to quantify the systemic risk of individual companies reliably and thus estimate a country's economic resilience. The method can be used for objectively monitoring change in production processes which might become essential in the green transition.

We could achieve impact in the corporate domain for the predictive maintenance of the cable network. For hybrid fiber-coaxial (HFC) networks, searching for upstream high noise in the past was cumbersome and time-consuming. Even with machine learning, the task remains challenging due to the heterogeneity of the network and its topological structure and noisy data. We solve the task by sessionizing the data per-incident and reformulating the classification into a ranking job. We present the automation of a simple business rule (largest change of a specific value), compare its performance with state-of-the-art machine-learning methods, and conclude that the precision@1 can be improved by 2.3 times using the developed machine learning approach.

Contents

CHAPTER 1

Introduction

Behavioral changes in society or technology can be represented as a graph with dimensions in space and time. Such graphs represent the link between events in the real world and their abstract representation. By analyzing such data, insights are derived, impacting decisions taken in the real world.

Interactions and communication between humans ever more frequently take place online all over the world and generate traces of data. Often, the data is collected for billing purposes or to improve the maintenance of the underlying communication infrastructure. But there is more to it - this data may also be used to support solving societal problems:

- Mobile phone usage data permits the moment-by-moment quantification of mobility behavior for Austria. Such data allows empowering rapid response to combat COVID-19 and potential future pandemics. We analyze the impact of gender differences and relative changes with regards to mobility to protect especially susceptible cohorts of our society. During the COVID-19 pandemic, we had the opportunity to perform several scientific analyses for social good where some of them were featured on national broadcast television and other media soon after publication. In particular, we analyzed the mobility of smartphones in Austria to calibrate the official epidemiological forecasting models[1] and to measure the impact of the non pharmaceutical interventions (NPI) in real time for a whole country. By using such data, informed decisions can be made where feedback is available only with a short delay, and potentially needed corrective measures can be applied. The pipeline created in this project and outlined in Figure 1.1 illustrates a fruitful collaboration between government administration, business, and research in obtaining insights. It should serve as a starting point for further collaborations also in non-crisis times.

[1] https://syd19.netlify.app

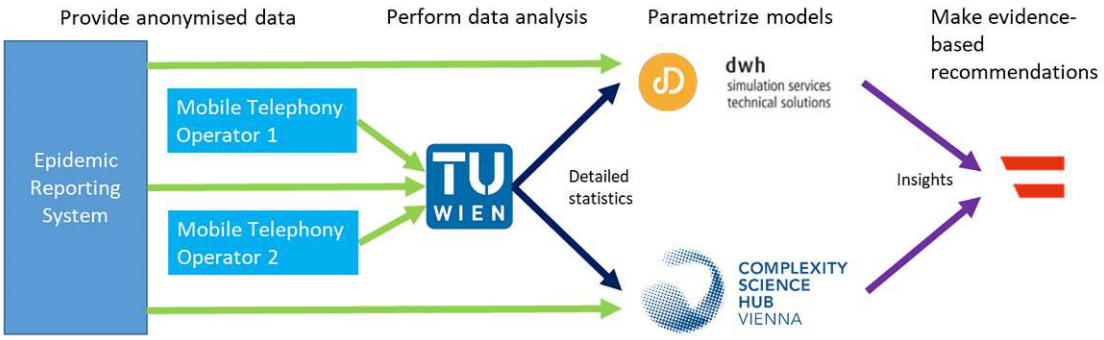


Figure 1.1: We collaborate with multiple mobile phone operators and the epidemic reporting system in Austria. Detailed mobility statistics were delivered to two organizations (DWH and CSH) for modelling the pandemic. The results of the simulations were used as the basis for evidence-based recommendations to the Austrian government through the COVID-19 Forecast Consortium.

- Such mobile-phone usage data sets cannot only be used to estimate the amount of mobility or flow of people between locations. A graph could potentially be created to determine probable supply links between firms by analyzing the interactions. Furthermore the interest in analyzing supply chains increased with the exit of the United Kingdom (UK) from the European Union (EU) as well as the COVID-19 pandemic, as a supply chain crisis could be observed[2]. By analyzing the interactions between mobile phones, we can estimate such supply chain networks and help to suggest potentially stabilizing measures by outlining the weak points in these networks. The monetary support could potentially be steered to the most important firms in the graph to prevent the meltdown of the supply network as remarkably little is known about the structure, formation, and dynamics of supply and production networks that are one foundation of society. Neither the resilience of these networks is understood, nor do we have ways to monitor their ongoing change systematically. Systemic risk contribution of companies was hitherto not quantifiable since supply network representations on the company-level did not exist except for very few countries. We use telecommunication data (calls) to reconstruct nationwide company-level supply networks. We find the conditional probability of observing a supply-link, given that a communication-link exists, to be about 90%. The method can be used for objectively monitoring change in production processes which might become essential in the green transition.

However, internet service providers (ISPs) usually do not have an intrinsic incentive to support studies on such datasets.

[2] https://www.instituteforgovernment.org.uk/publication/supply-chains, 27th Nov 2021

The infrastructure powering the communication network requires maintenance. Usually, the cable network is structured like a graph where various devices are connected in a topology. Finding the root cause of a specific problem can be cumbersome and lengthy. Even with machine learning, the task remains challenging due to the heterogeneity of the network and its topological structure. When analyzing the root cause of problems in the cable network to prevent outages of the internet for subscribers by automatically directing technicians to the source of the problem, we hope to support better internet quality for subscribers.

1.1 Problem statement

Analyzing massive quantities of spatio-temporal or interconnected (graph) data requires optimized tools and algorithmic implementations that rise to the challenge. The scalability of particularly costly operations such as geospatial or graph algorithms is essential when handling such data sets. The classical big data tools like Apache Spark or Hive did not offer geospatial primitives out of the box. We develop distributed scalable primitives for geospatial operations and perform smart aggregations.

Analyzing supply chains on a sufficiently fine-grained level is rarely possible due to data availability. However, mobile-phone usage data in the past frequently was accessible for researchers. We develop a suitable anonymization methodology and processing framework to utilize such data for economic shock simulations.

The cable network infrastructure distributing internet connectivity on the last-mile to end consumers is sometimes unreliable. In particular certain problem characteristics lead to a problem amplification where a problem in one single device quickly spreads to a whole region of the network as parts of the cable infrastructure and frequencies are shared by multiple devices. Therefore, after the monitoring system has identified that a problem exists in one region of the network it would be beneficial if a technician could be quickly directed the location of the root cause in the network to fix a problem. We develop a machine-learning based solution which performs 2.3 times better than a business rule currently in place at the participating ISP.

1.2 Data anonymization

When an ISP wants to support an initiative for societal good, properly anonymized data is critical. Mobile phone data consists of multiple dimensions (geolocation, social network, application usage). All of these dimensions are considered intricately private. Due to the General Data Protection Regulation (GDPR) in Europe as well as special laws for the telecommunication sector, handling that kind of data is strictly regulated. Therefore, the data needs to be anonymized. Anonymized data cannot be traced back to any individual person. In particular, this can be hard for geospatial or very high-dimensional data. To anonymize mobile phone data, any identifiers (international mobile equpipment identifier (IMEI), international mobile subscriber identifier (IMSI)) are hashed by the

data providing entities with a randomized salt that changes every day and the salt is deleted after 24 hours. This prevents long-term analyses of any particular individual.

Data that is used for analysis for this research is aggregated additionally. During the aggregation process, k-anonymity of the individual records in each cohort is ensured. This means that in any group i.e. age, gender, or postal code, at least k individual devices need to be present. If fewer than k unique devices are available for a cohort, the whole cohort is deleted from the data and not available for analysis.

Additionally, we only use cell-id-based localization to enhance the subscribers' privacy due to its limited accuracy. Thereby the local regulations have been met and the recommendations of the GSMA, the alliance of mobile phone providers [GSM20], have been followed.

For complex attributes such as geospatial details, additional care should be taken in the anonymization process [HC22][3]. Traditional k-anonymization procedures ensure that at least k unique observations fall into each group to prevent the identification of any individual contained in the data by reducing the granularity of the data. However, imagine a geolocation such as a home location of a device which for each day points to a very similar location. Potentially, some identifying patterns could still be established. Trying to remediate this problem by reducing the granularity a lot more will yield data which is no longer valuable. Instead, by applying probabilistic k-anonymization a better anonymized dataset where a higher amount of detail can be retained can be used.

1.3 Publications and contributions

We list the publications that form a significant basis for the text in this dissertation and summarize the contribution of each publication.

1. **Comparing implementation variants of distributed spatial join on Spark**, [Georg Heiler, Allan Hanbury], IEEE BigData Conference 2019 [HH19]. We compare a distributed geospatial join for a data-locality-preserving and a non-data-locality-preserving variant and introduce a broadcast spatial join which is much faster for up to medium-sized data on the smaller side of the join. First, we compare various variants of a distributed spatial join. In particular, data-locality-preserving and non-data-locality-preserving methodologies were juxtaposed. Second, we introduce a broadcast (map-side) spatial join. It is well suited for enrichment of a large data set with small to medium sized metadata, as the small data set is copied to all the nodes. In a second step, a local join is performed for each partition.

2. **Country-wide mobility changes observed using mobile phone data during COVID-19 pandemic**, [Georg Heiler, Tobias Reich, Jan Hurt, Mohammad

[3] https://georgheiler.com/2021/03/08/can-you-tell-the-nuts-berries-apart-in-each-group, accessed 28th Nov 2021

Forghani, Aida Omani, Allan Hanbury, Farid Karimipour], IEEE BigData Conference 2020 [HRH+20]. We design an efficient processing pipeline for mobility analyses based on mobile phone data and demonstrate the influence of the NPI and confirm the usefulness of mobility data for modeling the disease spread. We design and implement an efficient processing pipeline that prepares mobile phone data daily for anonymized and aggregated mobility analyses. Then, we demonstrate the influence of the lock-down in Austria on the values of the following metrics: point of interest (POI) based counting, radius of gyration (ROG). We confirm the usefulness of mobile phone data for modeling disease spread by identifying a significant correlation with a time-shift of 8 days for the outflow of people from the highly infectious quarantined region Ischgl in Austria to other municipalities.

3. **The impact of COVID-19 on relative changes in aggregated mobility using mobile-phone data**, [Georg Heiler, Allan Hanbury, Peter Filzmoser], Austrian journal of Statistics 2022 [HHF20]. We contribute a compositional analysis of the movement data during the COVID-19 research and conclude that special groups (elderly and young cohorts during weekends) do not reduce their mobility.

4. **Behavioral gender differences are reinforced during the COVID-19 crisis**, [Tobias Reisch, Georg Heiler (equal contribution), Jan Hurt, Peter Klimek Allan Hanbury & Stefan Thurner], Nature Scientific Reports 2021 [RHH+21]. We find that for both genders, we observe an increase of total call duration. For women, the recovery time for total call time initially is as fast as for men, but later, it slows down. The decrease in mobility following the lock-down is more substantial for women. In addition, men recover their mobility behavior much more quickly after the measures are lifted.

5. **Monitoring supply networks from mobile phone data for estimating the systemic risk of an economy**, [Tobias Reisch, Georg Heiler (equal contribution), Christiane Diem, Stefan Thurner], Nature Scientific Reports 2022 [RHD+22]. Repurposing anonymized and aggregated mobile phone data for the first time to analyze and validate the correctness of the inferred supply chain network. The resulting networks allow us to quantify the systemic risk of individual companies reliably and, thus estimate a country's economic resilience. The method can be used for objectively monitoring change in production processes which might become essential in the green transition.We quantifiy the utility of mobile-phone usage data for supply chain analyses.

6. **Identifying the root cause of cable network problems with machine learning**, [Georg Heiler, Thassilo Gadermaier, Thomas Haider, Allan Hanbury, Peter Filzmoser], Preprint under review [HGHF22]. For hybrid fiber-coaxial (HFC) networks, searching for upstream *high noise* in the past was cumbersome and time-consuming. Even with machine learning, the task remains challenging due to the heterogeneity of the network and its topological structure. We contribute a label generation process and data pipeline to train machine learning models and

can advance 2.3 times over the baseline when applying machine learning models to the problem.

1.4 Research questions

The following research questions form the basis for the work presented in this thesis:

1. How large is the impact of the NPI on mobility and calls? To what extent can differences be observed in groups of the society formed by age and gender (RQ1)?
2. How well can supply networks be reconstructed from mobile-phone data (RQ2)?
3. How well can machine learning identify the root cause of a given problem in a cable network (RQ3)?

1.5 Structure of the thesis

We outline the high-level functioning of a mobile phone network (not with regards to its technical underpinnings, rather focused on mobility analytics) and describe the methods we use to describe mobility, as well as the scalable data processing pipeline and the dataset we have collected in *Chapter 2 Mobile-phone data analytics*. In *Chapter 3 Efficient mobility analysis*, we outline how to implement scalable geospatial operations and summarize the primitives which underpin the data processing pipelines with regards to the scalable geospatial operations. In *Chapter 4 COVID-19 mobility insights* we present the results of analyzing the mobile-phone usage data in the context of the pandemic (RQ1). The inference of supply networks from mobile-phone usage data is presented in *Chapter 5 Reconstructing supply networks from mobile phone data* (RQ2). We describe how the root cause identification process in cable networks can be improved by a factor of 2.3 over a naive business rule currently employed by the participating ISP in *Chapter 6 Identifying the root cause of cable network problems with machine learning* (RQ3).

The impact in other fields and an outlook is presented in *Chapter 7*.

CHAPTER 2

Mobile-phone data analytics

Mobile-phone usage data is applicable for research on a wide variety of topics: customer segmentation [Ahe11], identification of personality traits and lifestyle [CJGP11, HKPO20], the analysis of large social networks [AKU19, AMRD19, ALS18], hotspot detection [NIZ+16], prediction of movement [DLY19], mode of transport identification [ZBMR20], credit scoring [LMZZ18], disaster recovery [ALVC19, ML19], analysis of sleeping behavior of the population [MBG+17], migration [IFMFM18] and land usage classification [SLS+19, LPCR+15].

In the last two decades, it became possible to collect data on human behavior on a population-wide scale, see e.g. [Wat07]. Some of that data has been used to investigate human responses to crisis and emergencies [BWB11, LBH12, WT14, GR19]. Studying collective response to a crisis is essential for catastrophe planning and coordination [BPPC07, GRK+20] and policy makers in health and safety [OLS+20]. Response to crisis also reveals human qualities that only surface when facing different kinds of actual or perceived danger [CLL20, TKL+00, BZZ96, GR19].

Albeit strong legal regulation telecommunication data has been accessible to researchers since more than a decade. Mobile phone data in the form of call detail records (CDRs) that are collected by mobile phone operators for billing purposes have been used to study communication networks and the behavior of millions of people [BDK15], leading to spectacular insights into the structure of human communication and organization [OSH+07, EMC10], human behavior in emergency situations [BWB11], the spread of infectious diseases [BGĆC16, JLY+20b] and the principles of human mobility [GHB08, SQBB10, SDO+21]. CDRs allow for population-wide coverage, granular resolution of interactions on the person level, and the possibility to be combined with information, such as age and gender.

2.1 Mobile phone network

The topology of an ISP's network is organized hierarchically and consists of multiple Location area codes (LACs). Many Base transciever station (BTS) are placed within a LAC. Each BTS consists of several sectors. Usually these are oriented 60 degrees apart to cover a full circle. There are exceptions from this rule in case of special situations such as in-house cells, tunnel-cells or omnidirectional cells. Every sector contains multiple antennas, wich each send the signal on a variety of frequencies (handling base load or high capacity) and technologies (2G, 3G, 4G, 5G).

The network is separated in the radio- and core network. To route calls efficiently or deliver data packages, the radio network precisely knows the location of all the devices as calls should seamlessly continue even when crossing the borders of countries. The core network only observes events relevant for traditionally important use cases such as billing or perhaps legal interception. In particular, the localization information is not available as detailed as in the radio network.

The core network keeps track of each mobile phone by noting the attached cell-id. We as researchers get access to an anonymized dataset where these connections to specific cell-ids in the topology of the network are collected over time.

2.2 Mobility measures

Mobility information obtained from sources such as the Global System for Mobile Communication (GSM) network can be helpful to monitor mobility on a large scale [OLS+20]. We quantify the movement of the population by several measures ranging from simple counting to estimation of the mobility via radius of gyration to the evaluation of origin destination (OD)-flow matrix, which are described in the following subsections.

2.2.1 Counting devices

The mobile phone network consists of multiple topological layers. Some base stations cover a larger area. Indoor or underground cells as for example found in metro stations are well suited to monitor the number of daily commuters due to their small coverage area. For several POI, i.e. base station in the Viennese underground metro network or for an airport and region of quarantine (Ischgl) we count the number of unique subscribers per day.

2.2.2 Radius of gyration

We obtain mobility data as a stream of spatially localized network signaling events. It is transformed into a list of locations $\vec{x}_{i\mu} = (x_{i\mu}, y_{i\mu})$, with associated stay duration $t_{i\mu}$ for every individual $i = 1...N_{indiv}$ at location index $\mu = 1...N_{locations}$, where x and y represent longitude and latitude, respectively. Due to the anonymization procedure the location index μ is reset every day and the individual index i is reshuffled accordingly.

The radius of gyration R_G is calculated as the square root of the time-weighted mean of the squared distances d (Calculated as the Haversine distance which calculates a distance in meters from latitude and longitude coordinates given in degrees.) of the locations $\vec{x}_{i\mu}$ to the daily centroid $\overline{x}_i = \frac{\sum_\mu \vec{x}_{i\mu} t_{i\mu}}{\sum_\mu t_{i\mu}}$:

$$R_{\mathrm{G},i} = \sqrt{\frac{\sum_\tau d(\overline{x}_i, \vec{x}_{i\tau})^2}{\sum_\tau t_{i\tau}}} \tag{2.1}$$

It captures the amount of movement in a time-weighted manner and has the dimension of a length in meters.

2.2.3 Entropy

For our second mobility measure *entropy* the locations $\vec{x}_{i\mu}$ are binned into a hexagonal raster using Uber's H3 [IB18]. The chosen resolution level for the raster yields hexagons with an area of approximately $800m^2$ (This is H3's resolution level 8.). For each hexagon $\tilde{x}_\nu$ ($\nu = 1...N_{hex}$), the stay duration of the locations in each hexagon are aggregated to $\tilde{t}_{i\nu}$

$$\tilde{t}_{i\nu} = \sum_{\nu \forall \vec{x}_{i\nu} \in \tilde{x}_\nu} t_{i\nu} \tag{2.2}$$

The stay time distribution of an individual i is then defined as the share of its time spent in a given hexagon $\tilde{x}_\nu$

$$p(\tilde{x}_{i\nu}) = \frac{\tilde{t}_{i\nu}}{\sum_\nu \tilde{t}_{i\nu}} \tag{2.3}$$

The entropy of an individual's stay time distribution, S_i, is defined, using the standard formulation of Shannon Entropy, as:

$$S_i = -\sum_\nu p(\tilde{x}_{i\nu}) \log_2(p(\tilde{x}_{i\nu})) \tag{2.4}$$

2.2.4 Night location

[WYU+15] propose to use the most cell tower during night time to derive a home location. We only use night-time activity from 8 pm until 12 pm to obtain a spatial reference for each device i. The resulting spatial reference is mapped to a post-code for further analysis.

2.2.5 OD flow matrix construction

The mobility interaction network can be captured by extracting the origin–destination (OD) matrix, which specifies the amount of travel between regions throughout the study area.

It is calculated for multiple scales including macroscopic scales, e.g., at the inter-urban level, or at microscopic scales, e.g., at the intra-urban level. In recent years, OD matrices have often been constructed from mobile phone data [SK08, MBL+19, GZ18, Pur18, LLP+15, BKG+19, FKC20].

A *trajectory* can be modeled by sorting the localized events per user by time. To derive the OD matrix, the continuous stream of point localizations in the network is first rasterized to the desired resolution. We are analyzing various resolutions as defined in Section 2.2.5, e.g. municipalities and post-codes as well as mathematically well defined grids like Uber's H3 [IB18]. For each discrete location l a stay duration is computed, which is referred to as weight w.

We cluster these discretized point localizations by space and time in order to compute time-weighted stays for each user and raster cell. The most important points can be aggregated as a OD matrix, where *most important* refers to the points with a stay duration of at least s_k seconds. Each stay has an associated entry and exit time. We set a threshold of $s_k = 600$ seconds for our analyses based on an analysis of the distribution. Finally, we aggregate the matrix over all the devices i daily by counting the subscribers moving from one grid unit to the other.

The analyses were performed at a multitude of spatial resolutions. We use the following levels, which are increasing in the level of detail:

1. Austria as a whole
2. federal states
3. political areas [1]
4. municipalities and postal codes
5. specific points of interest

2.2.6 Points of interest (Shopping, Leisure)

Specific points of interest reflecting shopping and leisure zones in Vienna were analyzed in more detail. We first used H3 by Uber [IB18] to create a discrete raster for the whole country to speed up the analysis of specific locations afterwards. Then we count the number of unique subscribers in a set of manually defined hexagons. We limit our investigations to stays longer than 10 minutes and shorter than 4 hours. We assume this eliminates devices passing the shopping complex on the nearby highway, as well as persons working there, because these activities take much shorter or longer, respectively.

[1] https://www.statistik.at/web_de/klassifikationen/regionale_gliederungen/politische_bezirke/index.html

2.2.7 Graph-based movement analysis

The use of graph-based analyses in crowd-movement studies has been investigated, especially in the use of mobility data extracted from cellular networks [GZW20]. The OD matrix can be interpreted as a graph where pairs of nodes m and n represent origins and destinations which are connected by links with non-negative weights $A_{m,n}$ if one or more trips are made between the nodes. By modeling the crowd-movement in the structure of a graph, it is possible to characterize the architecture and dynamics of the population mobility and demonstrate relationships between people and places [FK18]. In graph theory, the topological criteria such as centrality, connectedness, path length, diameter, and degree play a vital role in the description of a graph where links are usually represented as binary states (i.e. adjacency matrix). For the mobility analysis, the difference in the strength of the interaction links between pairs of nodes is important [SMBH17].

2.3 Calling behavior

By analyzing calls, social interactions can be modeled. This part of the data consists of a list of outgoing (MO) and incoming (MT) calls, each associated with a source and destination. We filter to calls with a duration of at least 25 seconds to adjust for a shift in the distribution corresponding to calls that were not answered.

For each device c we find N_c^{MO} outgoing and N_c^{MT} incoming calls with k_c^{MO} and k_c^{MT} other individuals, respectively (in- and out-degree). The call duration is denoted by $\bar{t}$. Additionally, as described earlier for the mobility dimension, for each device, age group and gender are specified.

For all of these device-level metrics we report the median of the whole population, or for cohorts specified by age groups or gender. We will add superscripts g and h to indicate gender.

2.4 Gender and Age Group differences

To investigate gender differences we calculate the gender ratio r_x for the various aggregations x (calling, ROG) presented here. The ratio r_x is calculated as the quotient of the aggregate for the female cohort divided by the aggregate for the male cohort $r_x = x_{\mathrm{female}}/x_{\mathrm{male}}$ (x represents the aggregation, e.g. median R_G or median call duration $\bar{t}$). A gender ratio r_x close to 1 (or 100%) indicates that the quantity is of similar size for both genders, less (more) than 100% indicates smaller (larger) values for females.

2.5 Scalable data processing pipeline

Our data processing pipeline is depicted in Figure 2.1. Firstly, to improve the performance of our analyses and to ease the mental burden for the person conducting the evaluation, we

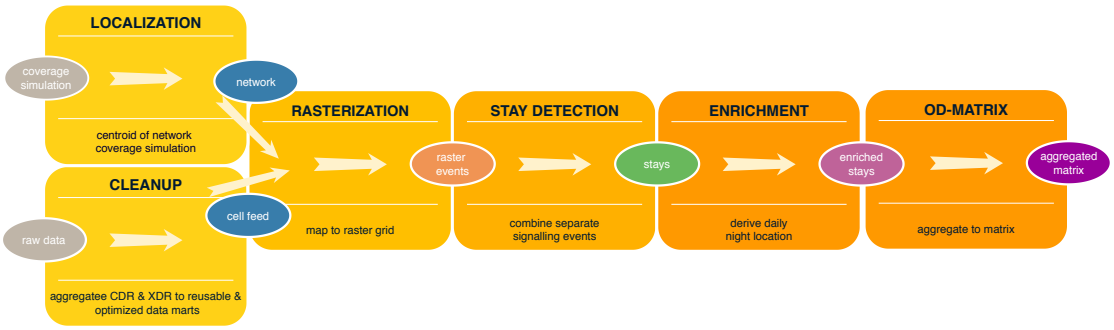


Figure 2.1: Daily aggregations are calculated as a first step to clean up and compress the data. Subsequent analyses can be implemented efficiently – as depicted here in the case of the computation for an OD matrix.

computed a daily aggregation of the raw data, removing domain-specific knowledge, prior to beginning the specific analyses. The GSM network and attached monitoring tools are a complex system that requires a lot of business knowledge (telecommunication-specific knowledge required to process the raw data, such as special types of events or structures of the topology). We abstracted this knowledge away and allow for effective and efficient analyses on top of our aggregations.

For each device, timestamp and cell-id were recorded and anonymized with a rotating key. Then, we enriched each cell-id with its location information which is provided by the ISP (described in 2.6.1). As a next step, stays within a raster cell, which are defined by the regional levels listed in Section 2.2.5, were detected by spatio-temporal clustering the signalling events. Thereafter, each stay was enriched with the daily night location, see Section 2.2.4 for details. Finally, the OD matrix was created as described in Section 2.2.5. After the daily cleanup, aggregation and compression, analyses were built on this solid and reusable foundation.

We needed to process an immense quantity of data as the raw events amount to more than one billion per day. Therefore, we rely on a cluster of computers to achieve good performance. Using Apache Spark [ZCDD12], a map-reduce style big-data framework, the burden of a distributed system is eased for the developers of the analyses as failures of compute nodes are handled automatically.

As Spark lacks support for geospatial primitives, GeoSpark was utilized for distributed spatial joins [YJM15]. Additional geospatial functionality is made available by using GeoMesa [HAE+15], which in its core is based on the java-topology-suite for providing the geospatial functionality. This is similar to GeoSpark, but still offers complementary functionality such as calculating spatial distances. To enrich spatial metadata such as political areas, we utilize a custom broadcast spatial join [HH19](Chapter 3), which is faster than its distributed equivalent.

Furthermore, columnar file formats with compression and run-length encoding are used, as these allow for significant compression of the data when sorted by the join keys.

2.6 Description of the datasets

The following Chapters take reference to a dataset describing COVID-19 infection rates in Austria as well as various aggregations obtained from the mobile phone usage data.

2.6.1 Mobile phone usage data

We partnered with a large Austrian ISP to get access to anonymized data from mobile phones. We use a combination of classical Call Data Records for the voice domain as well as a combination of generic data records (known as X Data Records) for the data domain. Thus we do not only register an event when a call is performed, but rather perceive additional events when data packages are transferred. Various network interfaces are connected via probes so we get data points from a multitude of network technologies for mobile data usage (2G, 3G, 4G), calls, text messages as well as Voice over LTE, from both user- as well as control plane.

The information on all exchanges made between a mobile phone network and its users are recorded as *events*. Any direct (user plane) as well as indirect (control plane) interaction with the network continuously generates events in the data set, which are aggregated daily. The data set is based on classical Call Data Record (CDR) and includes X Data Record (XDR) of the data domain, thereby providing anonymized metadata about voice and data usage. The events stem from various network interfaces covering all most widely used signalling technologies (2G, 3G, 4G, calls, text messages & Voice over LTE (VoLTE)).

The data set contains approximately 1 Billion events from 4.5 Million devices per day. For 80% of these, the subsequent event is received in 1.7 minutes, on average 4 minutes. This means that for some old, i.e. 2G devices, which are only rarely used, almost no data is transmitted whilst not actively in operation. Therefore, fewer events are generated and these devices are thus much harder to analyze when considering them for mobility use cases. Our analyses are based on filtering the data of approximately 1.2 Million devices registered with the partner ISP as mobile handsets excluding sensor devices from the Internet of Things as well as roamers[2] or events obtained from virtual network operators[3].

The GSM network registers events for each device i with a very accurate time information and a location with latitude (y_{lat}) and longitude (x_{long}). As the network continuously generates events, a near-real-time monitoring of aggregate population behavior is possible [XGM+20]. Our analyses were updated on a daily basis.

From this data we extract gender-specific features about communication patterns, such as the average interaction duration and the number of calls for all possible gender combinations of calling and being called. The data further allows us to characterize mobility. From location data we estimate the number of people shopping for food and the usage of recreational areas.

[2]Devices with a foreign SIM-card using the local network, i.e. mostly tourists.
[3]Virtual operators resell the existing network of the providers – often more cheaply.

Demographic information is not available for roamers or virtual mobile operators (MVNO) and thus they are excluded from this analysis.

Furthermore each day, only devices with a radius of gyration R_G (see eq. (2.1) below) larger than 0m and lower than 300km are considered. The lower bound aims to exclude internet of things-devices, which typically do not move, such as LTE-internet routers. The upper bound excludes a small number of devices which have a R_G larger than the theoretically maximal R_G inside Austria and are attributed to network artifacts.

Calls are filtered to a length of at least 25 seconds prior to aggregation to exclude calls that were not picked up, which form a distinct peak just below 25 seconds.

Our localization methodology is based on the topology of the network, namely the observed cell-id. This means that the accuracy is limited, and much less accurate than Global Positioning System (GPS)-based localization or the result of custom apps combining Bluetooth, WiFi and GPS. However, the data is available for a large quantity of devices. The ISP provides us with the localization information for each cell-id, which is based on the centroid of the network coverage simulation.

Sociodemographics: Using additional metadata, an individual can be assigned to a gender group (female, male), to an age group (here we consider the age groups in 15 year intervals: 15-29, 30-44, 45-59, 60-74, and 75+), and to an Austrian district of the daily night location to derive the groups. This metadata is provided in an anonymized format from the ISP. Since the distributions are generally very right-skewed, we work with the median per group and day in the following and also ensure k-anonymity for each one. The distribution of the genders: 454,000 women and 452,000 men is approximately equal.

The anonymized data covers the time period across the government interventions from February 1st to June 29th of 2020.

2.6.2 Infection data set

On behalf of the Austrian Ministry of Health, the company *Gesundheit Österreich GmbH* provides access to the electronic epidemiological reporting system (EMS)[4]. The data contains cumulative daily COVID-19 infection numbers from March 5th 2020 onwards.

[4] https://datenplattform-covid.goeg.at/

CHAPTER 3

Efficient mobility analysis

Many of the mobility analyses require a spatial data processing primitive: The spatial join. Having access to a fast implementation of it is important for scalable data pipelines.

3.1 The need for scalable spatial methods

The current mobile phone network already generates a vast amount of data. Also, 5G, as the new mobile phone network standard currently rolling out in various countries with many tiny cells (microcells), will generate even more data than previous versions. In both cases, processing a rapidly increasing amount of data is essential. Many use cases like urban planning, location-based advertising, recommendation of points of interest (POI), or socio-economic analyses require data in the spatio-temporal domain.

One of the most frequently used spatial operations is the spatial join. A naïve implementation is computationally expensive when performing a spatial enrichment on large quantities of data. Traditional geospatial information system (GIS) tools like PostGIS[1] offer such spatial processing capabilities, however their processing power is limited as they are usually bound to a single node. In the Hadoop ecosystem, it is possible to scale computation up to thousands of machines. The distributed architecture is only effective when network traffic is minimized. A naïve distributed implementation utilizing a cross-product would still be slow. Before filtering to the relevant results according to a spatial predicate (intersection, overlap, ...) the intermediate state which is required to be exchanged between compute nodes, is enormous as all tuples on the left side are paired with all tuples from the right side. Using frameworks like Spatial Hadoop [EM15], it is possible to achieve the desired level of scalability. However, based on classical map-reduce, queries are slow and also inherit the complexity from operationalizing Hadoop. Apache Spark is a popular, fast and scalable in-memory computation framework [ZCDD12].

[1] https://postgis.net/

It is sometimes used without Hadoop in the cloud to make Spark more accessible for newcomers. Spark - like many distributed computing frameworks - is still based on the map-reduce paradigm where computation is split into chunks and processed in parallel on multiple nodes. Spark is not only faster than classical map-reduce, but also offers a higher level API mimicking a local collection object with operations like *filter*, *join*, or *groupBy* as resilient distributed datasets (RDD). RDDs are immutable and state which is lost can be recomputed. Spark achieves speed by transferring data in memory and not writing to disk between each query step. With the addition of Spark-SQL, a graph of operations to be executed even allows for optimization to improve query performance.

However, no native support for spatial data types, queries, or most importantly spatial indices is built into Spark. Multiple frameworks are readily available to perform distributed spatial operations using Spark [TYM+15, YTA+17, XLY+16, YZS19, Sri14]. A detailed comparison of these systems can be found in [YZS19] and [GGCI+17]. With currently[2] 1.1k stars on gitHub[3], GeoSpark [YJM15] has a large community. Therefore, we chose it as the basis for our comparison. Its implementation is based on spatial partitioning, thus providing the possibility to join large spatial data sets. Exchanging data over the network (shuffling) is mandatory to colocate tuples which are close in the spatial domain and to enable fast local queries of a spatial index for each partition.

Accessing neighbours in the time and space domain is relevant for various trajectory-related computations like smoothing/noise reduction or clustering. These tasks are only efficient if local data, i.e. data which resides on the same node, is accessed when querying for neighbours. Not all use cases require two large data sets. When working with trajectories we propose a faster methodology for enrichment of spatial data which requires less network traffic and is thus faster.

3.2 Experiment description for comparison of join implementations

In distributed systems a join of two datasets commonly requires a shuffle operation to exchange data between the compute nodes. If the size of the data is large, this task can become very costly and slow due to large network IO. However, if one of the two datasets is small it could be broadcasted - and the other big one does not require to be shuffled. This is supported by a tool like Apache Spark out of the box. But as mentioned before, Spark does not include support for geospatial operations. We develop a broadcast spatial join wich preserves data locality and is efficient as no large amount of data needs to be exchanged. The small dataset needs to be indexed using a spatial indexing data structure (such as an R tree) and broadcasted.

We compare the computational speed of various distributed spatial join implementations utilizing three methodologies:

[2] 20th February 2022
[3] https://github.com/DataSystemsLab/GeoSpark

1. GeoSpark framework in a non-data locality-preserving way. Time to un-nest the data set was not counted, only a default inner join was performed, no re-aggregation to the original locality-preserving format (i.e., the least overhead when using out of the box tools to perform scalable spatial enrichment).

2. GeoSpark in a data locality-preserving way. The data initially fed to the spatial join was un-nested in a way that each observation from the array formed a new row. A left join was added manually as GeoSpark does not offer such an operation and we did not want to lose observations. Finally, the data was compacted again to allow for further processing in the data locality optimized representation. In more detail, data was aggregated for each user and period to contain the array of events with information about the joined POI.

3. Distributed systems like Apache Spark provide basic building blocks for generic data operations. However, as described before lack geospatial primitives. Usually they provide distributed joins and broadcast join primitives. A distributed join is very scalable, but also slow as the data needs to be exchanged over the network. The broadcast join can be much faster if one side of the join is small (which frequently is the case) and fits into the memory of a single worker node. In such a case, a full copy of the small dataset is broadcasted to all the worker nodes and then combined node-local with the other one. Our data locality-preserving method consisting of a spatial index (R-tree) created from the POI data which is using the broadcast approach to be distributed to all the worker nodes. Thus the join is performed without accessing the network locally on each partition of the data.

We investigated the enrichment of spatial trajectories with the nearest POI using a spatial join. The data was used to better understand recurrent patterns in trajectory data e.g., for classification of activity. The data was simulated and an exponentially increasing load of users was generated for multiple periods. For each user, time period (date) and a data locality-preserving array of events (time, latitude, longitude, accuracy (uncertainty of localization)) were stored partitioning the data per date as this allows effortless calculation of trajectory operations per group and easy addition of new data. All simulated locations were within Austria. Initially, the data resided in a locality-preserving format suitable for various trajectory analyses, but without the POIs.

We conducted our experiments on a Hadoop cluster using Spark version 2.2 on yarn with 37 containers using 4 cores each and 55GB of RAM per worker node totalling up to 145 cores. POIs were derived from the open street map (OSM) project as a subset with certain filter criteria.

The above mentioned methodologies were compared using the following configurations. Our code is available on gitHub[4].

(a) 200 events per period and user and 3 periods, 9.8k POI

[4] https://github.com/complexity-science-hub/distributed-POI-enrichment

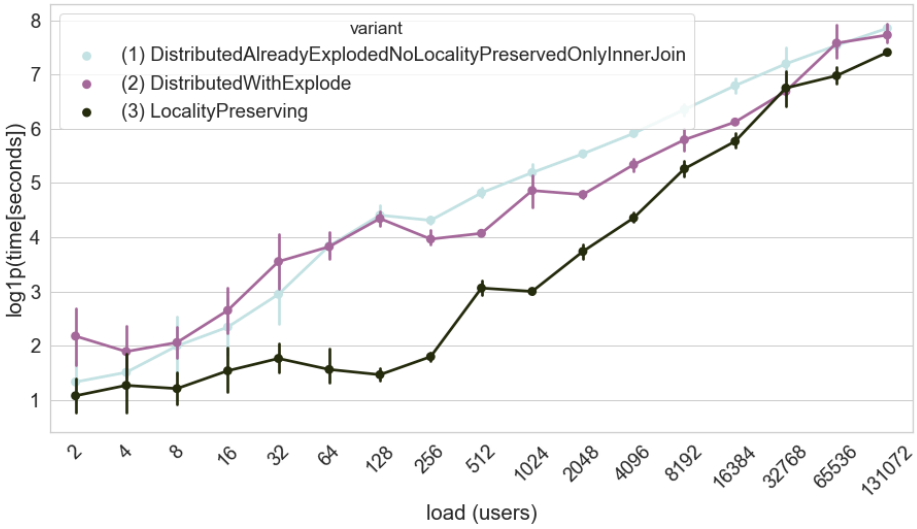


Figure 3.1: Configuration (a): 200 events per user per period for 3 periods. Load of users (x axis), processing time shown in logarithmic scale (y axis) for the 3 different implementations of a spatial join. Each one was run 5 times. The graph shows the mean and 95 confidence intervals as error bars.

(b) 2000 events per period and user and 3 periods, 9.8k POI

(c) 200 events per period and user and 300 periods, 9.8k POI

(d) 200 events per period and user and increased number of POI to 65.1 Million. Using all OSM POI for Austria.

3.3 Performance comparison of scalable spatial join implementations

In the following Section we describe the results of the experiment.

Configuration (a): As indicated in Figure 3.1, the locality-preserving GeoSpark join (2) is faster than the non-preserving approach (1) for large-enough quantities of data.

This is particularly surprising considering the larger amount of data being shuffled in the locality-preserving distributed GeoSpark join for: un-nesting, left join and aggregation. In almost all cases the custom implementation (3) using a map-side broadcast join was optimal, although in extreme cases the advantage of (3) diminished.

Configuration (b): When the number of events per user and period was increased, we obtained a more expected result concerning (1) and (2) where the latter was slower, as

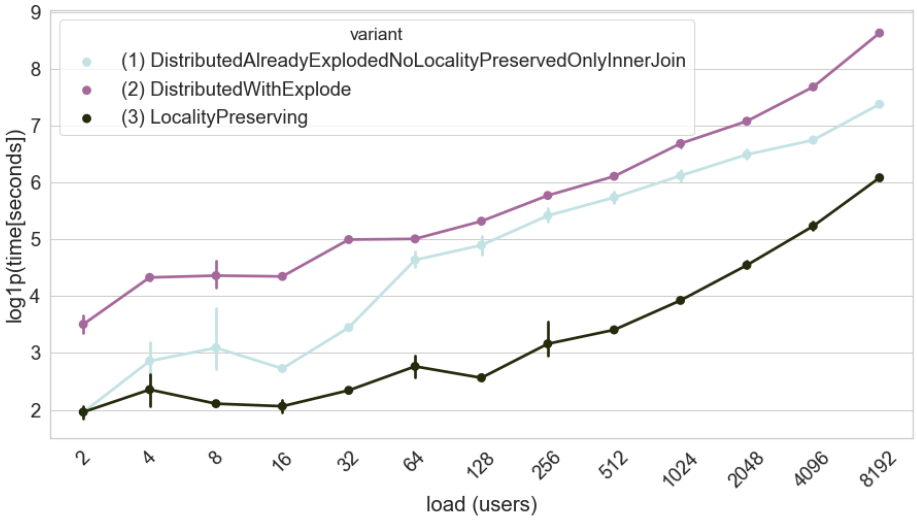


Figure 3.2: Configuration (2): increased number of events per user to 2000 events per period for 3 periods. Load of users (x axis), processing time shown in logarithmic scale (y axis). For each methodology 5 runs were computed. The graph shows the mean and 95 confidence intervals as error bars

indicated in Figure 3.2. We also noted that (2) generated a fairly large amount of shuffle IO when reconstructing the trajectory optimized format.

Configuration (c): As seen in Figure 3.3, (1) and (2) converge when increasing the load to 300 periods. This means that the overhead of shuffling for disaggregation and later re-aggregation is negligible from a time perspective, though it causes several 100GB of shuffle IO. Methodology (3) was the fastest variant. As an additional benefit no shuffle IO was caused.

Configuration (d): (d) was considered very specific as a high number of POI were within close proximity of each other. In this case more POI than trajectory points are present for small to medium sized workloads. Therefore, only in this case spatial partitioning was applied on the POI, not on the trajectory data set. A minimal workload already returned a large number of tuples. Methodology (3) was not suitable as it did not complete the computation. For this configuration, a distributed spatially partitioned join was the only option as each individual event already generated a large number of tuples and the parallelism was higher, resulting in smaller resource requirements compared to (3) and thus the completion of the queries.

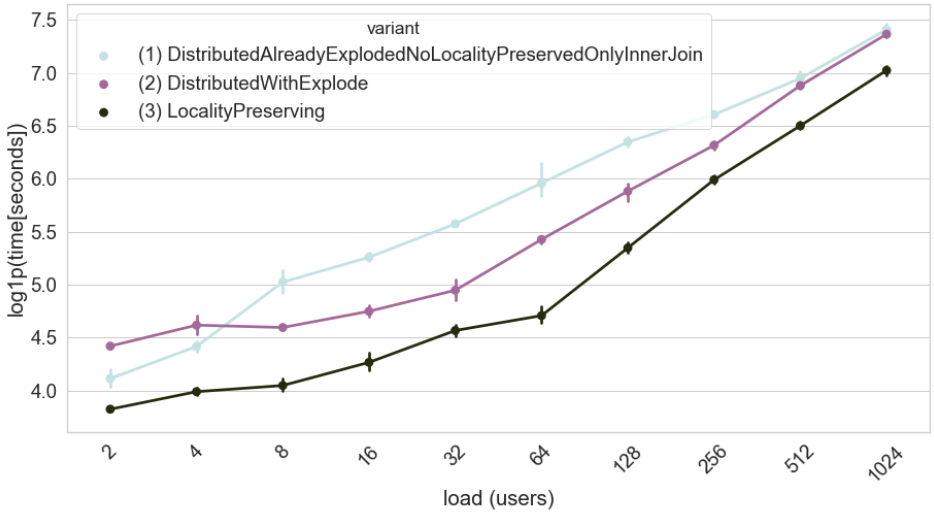


Figure 3.3: Configuration (3): 200 events per user per period. Increased number of periods to 300. Load of users (x axis), processing time as shown in logarithmic scale (y axis). For each methodology 5 runs were computed. The graph shows the mean and 95 confidence intervals as error bars

3.4 Discussion

Various use cases require different implementations for distributed spatial joins. A general-purpose framework like GeoSpark is useful, however, we observed that sometimes a more specific implementation, like methodology (3), the broadcast spatial join, proved more efficient. This property is particularly useful in scenarios where shuffle IO needs to be minimized, e.g., a real-time streaming computation or cases in which the spatial enrichment includes up to medium-sized (10k-100k) number of tuples. Fine-tuning Spark itself, i.e., setting the right level of parallelism might improve future approaches. In the future, a discretized spatial index like [IB18] could yield even more improvements, especially concerning event data as it would be possible to precompute an enrichment for all available raster cells of a specific resolution for a country and then applied very fast to new incoming data. Additionally, validation on a real-world data set like a mobile telecommunication data set should be performed.

Meanwhile geospark was renamed to Sedona and accepted as a top-level Apache foundation software project and has started to support the proposed methodology of a broadcast spatial join out of the box[5]. Furthermore other cloud databases such as BigQuery have added geospatial functions[6].

[5] https://sedona.apache.org/api/sql/Optimizer/
[6] https://cloud.google.com/bigquery/docs/reference/standard-sql/geography_

functions

CHAPTER 4

COVID-19 mobility insights

In March 2020, the Austrian government introduced a widespread lock-down in response to the COVID-19 pandemic. Based on subjective impressions and anecdotal evidence, Austrian public and private life came to a sudden halt. Here we assess the effect of the lock-down quantitatively for all regions in Austria and present an analysis of daily changes of human mobility throughout Austria using near-real-time anonymized mobile phone data.

Having access to properly anonymized data is important to perform mobility analytics to empower rapid response to combat COVID-19 and future pandemics. We analyze the absolute mobility of groups formed by age and gender and compare how the reduction induced by the NPI relates to the duration of calls. We show that a significant time-lagged relationship exists with the mobility outflow of COVID hotspots and the case numbers in the destination regions. Then, we apply compositional data analytics methods to identify how much groups (relative to each other) have changed their mobility behavior.

4.1 Mobility and the COVID pandemic

Extensive literature using mobility data during the COVID-19 pandemic has been published [PBG+20, JLY+20a, GRK+20, JWA+20, YTF+20, VMU+20, XGM+20, SFS+20, SSS+20, ISS+20, Heu]. International mobility reports of mobile-phone-based data are analyzed by the European commission to report on the effect of the COVID-19 lock-down including the comparison of the effect in different countries and estimation of cross-border effects [SFS+20, SSS+20, ISS+20]. Furthermore, [Heu] published analyses of mobility behavior subsequent to the lock-down using GPS-data evaluate the virus's spread from the highly infectious region in Ischgl to different countries. They used data from the private company Umlaut, which collected GPS measurements using tracking toolkits in apps. This data offers a higher positioning accuracy. However, the number of users and,

hence, the number of data records and the population's coverage are limited compared with GSM-based data.

Many studies, primarily employing data from China, analyze mobility to predict infection numbers. Jia et al. [JLY+20a] and Goa et al. [GRK+20] predict the number of infections in China with the outflow of people from Wuhan. Kraemer et al. [KYG+20], Jeffrey et al. [JWA+20] and Yabe et al. [YTF+20] report that the correlation between mobility and the infection rates dropped after implementation of the lock-down measures in China, the United Kingdom and Japan, respectively.

The two companies behind Android and iOS, Google[1] and Apple[2], both published aggregate mobility reports as well. These are available for many countries. Vollmer et al. [VMU+20] and Xu et al. [XGM+20] use mobility data to calibrate epidemiological models. We extend the literature by investigating Austrian mobility behavior and combining a wide variety of measures.

4.2 The pandemic in Austria

At the end of 2019 the SARS-CoV2 virus emerged in China, causing an ongoing, worldwide pandemic. In response to sharply rising numbers during the "first wave", on March 15th the Austrian government introduced a severe nationwide lock-down. The implemented non-pharmaceutical interventions (NPIs) included: school closures, restaurant closures, mandatory use of masks, incentives to use home-office, the complete prohibition of gatherings of any size, closure of all non-essential shops, and a general limitation of mobility. It was possible to leave the house for one of four reasons only: work that cannot be postponed, shopping for groceries, assisting others, and short recreational walks [DlDH+20].

The government's call seemed successful based on anecdotal evidence, such as reports of empty public spaces [orf20] or low traffic levels on highways [3]. However, to estimate the effect on epidemic spreading and plan further policy measures, a countrywide quantification of the impact of the actions was necessary. It is generally agreed upon that ensuring a minimum spatial distance between people and limited exchange between segregated communities are critical factors in preventing the spread of COVID-19.

These measures led to a massive reduction of mobility as measured for example with cellphone data [HRH+20], or traffic counts [Asf20]. The lock-down had severe consequences on public life: 58% of all Austrians who were in employment or self-employed reported that they were employed in a company that introduced home-office to at least some

[1] https://www.google.com/covid19/mobility/

[2] https://www.apple.com/covid19/mobility

[3] https://www.tt.com/artikel/16774378/zurueckversetzt-in-eine-andere-zeit-kaum-noch-verkehr-durch-tirol, accessed 18th of March 2020

extent[4], the number of people registered unemployed increased by 76% [5], more than 1,300,000 persons were temporarily laid off [Hag20], and public life, such as theaters, cinemas, restaurants, bars, shopping-malls and even large parks, came to a halt.

In our analyses we consider multiple phases indicated by Roman numerals, based on the data set of non-medical interventions published by Desvars et. al. [DlDH+20]:

I. pre-lock-down – before 11th of March. Pre-awareness phase. The population is practically not yet aware of the presence of the disease in Austria.

II. transition period from the announcement (March 12th) to the actual lock-down on March 16th.

III. lock-down – 17th of March until 1st of May

IV. easing – 2nd of May onwards.

For the Section on *Gendered impact analysis* we evaluate the easing process in more detail as it was conducted in a later publication and we decided to split the easing period into individual sub-phases there. III and IV are split into:

III. lock-down until first easing of NPIs (April 13th)

IV. period gatherings of more than 10 people are allowed, begins on May 1st

V. back to normal, restaurants and businesses re-open

VI. easing – 2nd of May onwards

4.3 Absolute changes of mobility and call duration

Mobility information obtained from sources such as the GSM network can be helpful to monitor the reduction in mobility on a large scale [OLS+20]. We monitored daily changes of mobility in Austria using anonymized mobile phone data, compared behavior before, during and after lock-down measures and published parts of our results online [6] due to the inherent relevance for the public. Here we present and extend the results and elaborate on the technological background of our efforts during the COVID-19 pandemic.

[4] https://www.market.at/market-aktuell/details/corona-definiert-arbeitswelten-von-morgen-neu.html, accessed 8th of October 2020.

[5] https://www.ams.at/arbeitsmarktdaten-und-medien/arbeitsmarkt-daten-und-arbeitsmarkt-forschung/berichte-und-auswertungen, accessed 8th of October 2020

[6] https://csh.ac.at/covid19, accessed 5th of February 2020

4.3.1 Overall mobility reduction

We found a reduction of commuters at Viennese metro stations of over 80% and the number of devices with a radius of gyration of fewer than 500 m almost doubled. The results of studying crowd-movement behavior highlight considerable changes in the structure of mobility networks, revealed by higher modularity and an increase from 12 to 20 detected communities.

Public transport usage

Figure 4.1 shows the reduction of passengers on the Viennese metro, which translates into the effectiveness of the far-reaching restrictions undertaken by the government of Austria. After a first press conference on the 10th of March (first black line), the measures were announced and activity was reduced until full implementation of the lock-down measures on the 15th of March. The frequency of metro usage was about 1/5 of a regular Monday in this state induced by full implementation of the lock-down measures. We still can observe the weekly trend that there is less usage of the metro during weekends. From Easter onwards metro usage starts to recover almost to previous levels.With the official end of the lock-down, mobility has recovered to 52.5% when comparing calendar week 22 with week 10 – i.e. with the levels of before the crisis. Even later until August a full recovery to previous levels is not reached.

POI analyses

For two selected locations (airport, quarantined region), see Figure 4.2, the dramatic reduction in devices present is depicted. Both can be seen as a proxy for long distance/international travel activities[7]. This also justifies that both locations have not recovered until the end of the analysis period.

ROG

Before the crisis, the median ROG for the whole population was 2 kilometers per day. After the announcement of the restrictions on the 15th of March, it reduced to 800 meters. The distribution of the ROG is heavily skewed. When creating discrete bins of the ROG the effect of very large ROG can be mitigated. We create three bins, $[0, 500\text{m}[$ for devices showing little to no movement, $[500, 5000\text{m}[$ for intermediate and $[5000\text{m}, max.]$ for large movements. Bin sizes were chosen based on qualitative experience with test devices. Figure 4.3A depicts how the lock-down measures increased the number of devices moving very little. Even after the official easing of the measures the population has not yet recovered to previous levels of movement until the end of our study period. Conversely, for medium distance movements in the range of 500–5000 m and large radii above 5000 m, Figures 4.3BC show the effect of the lock-down measures by depicting a dramatic reduction of movement.

[7] https://www.tirolwerbung.at/wp-content/uploads/2018/04/tiroler-tourismus-daten-und-fakten-2017.pdf

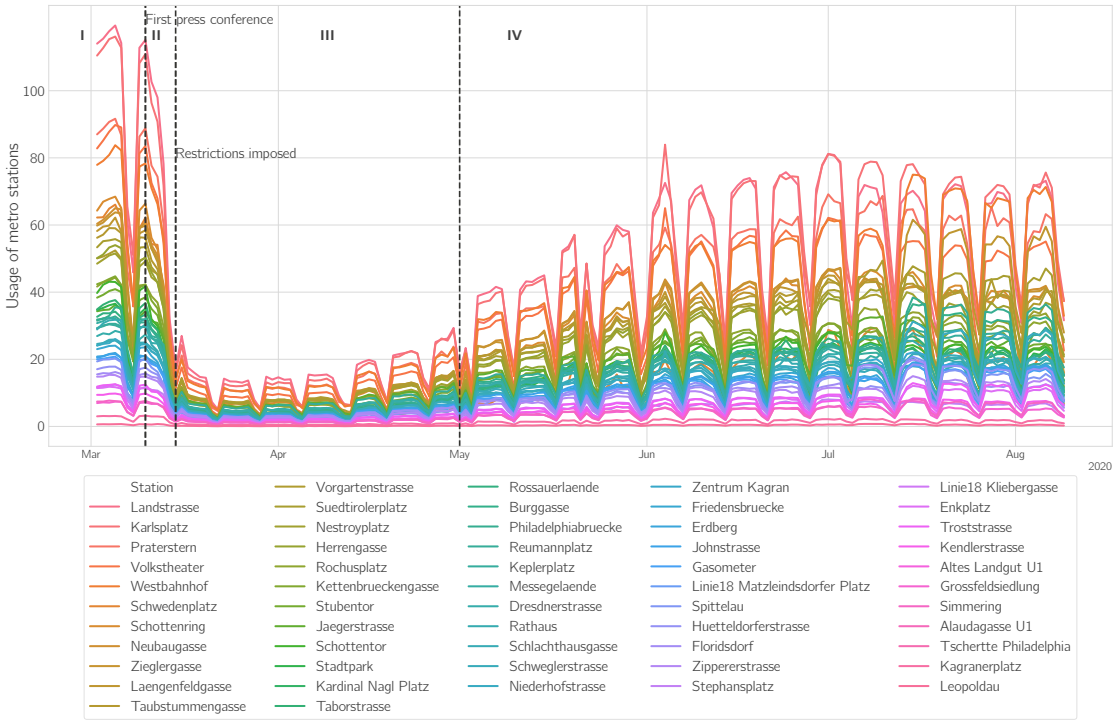


Figure 4.1: Reduction in public transport usage in Vienna during the COVID19 pandemic. Even after easing of the measures previous levels of metro usage are not reached again.

We additionally computed a daily night location as defined in Section 2.2.4 for each user. This location was assigned the matching post code allowing to produce a map of Austria, Figure 4.4, visualizing the spatial differences in the relative reduction of mobility, as measured by the ROG. The reduction is consistent throughout Austria, except for some small towns where the number of observations might be too small.

Figure 4.5 shows the effects of the lock-down. A reduction of mobility in the districts of Austria occurs from before the lock-down (panel A) to right after it (panel B). As a measure for mobility we use the median radius of gyration, R_G.R_G captures the time weighted, spatial extent of an individuals trajectory. We observe a decrease of R_G between 59% and 14%. Panel C shows the time evolution of R_G, averaged over all districts. After a sharp decline of almost 50% in phase III a rebound to almost pre-crisis levels is seen. In panel D we observe a more than 60% increase of call duration per call, $\bar{t}$. For a definition, see Methods. Panel E shows a brief increase of the number of calls per person, N_c, in the days just before the lock-down (phase II) followed by a 10% decrease. We now stratify these changes with respect to gender and age.

Moreover, we have broken down the analysis into hourly groups (Figure 4.6). Each line represents an hour. Before the lock-down, we observe a relatively consistent weekly trend. During the week there is a large spread between daily and nightly mobility, whereas on weekends this gap is reduced strongly due to less movement during the daytime and

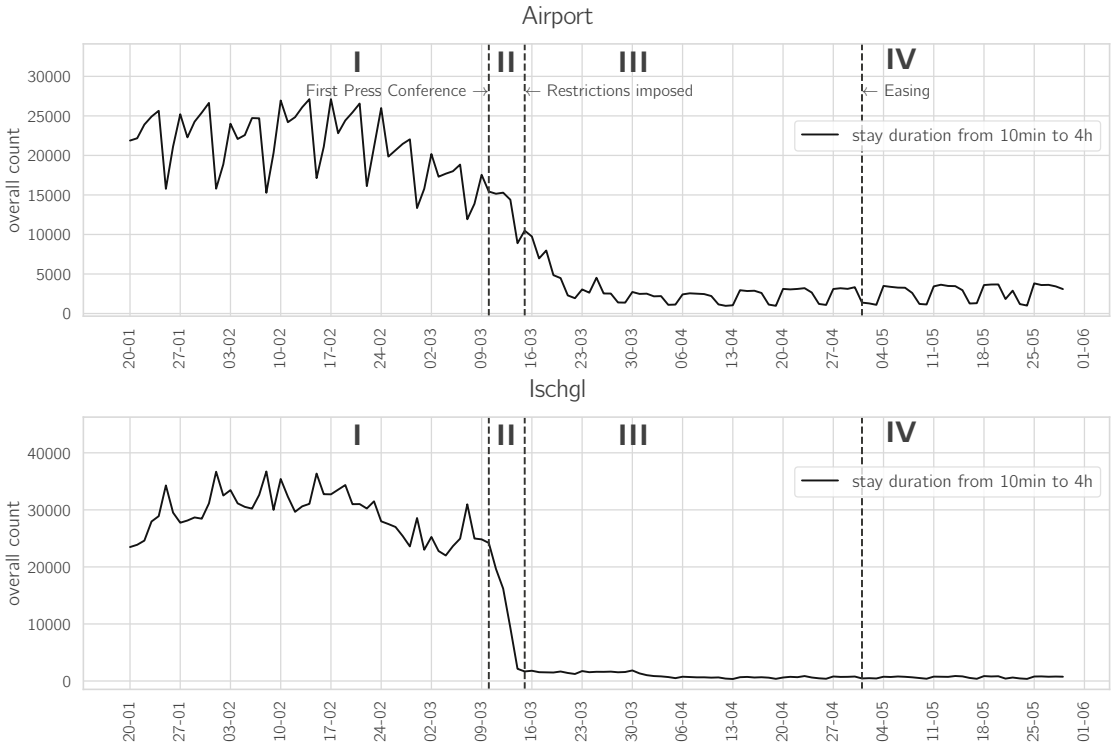


Figure 4.2: The count of mobile phones with a stay duration from 10 minutes to 4 hours for Ischgl and the Airport of Vienna. There is a clear difference between the Airport and Ischgl. While Ischgl went into quarantine, and all tourists were sent home on the 15th of March, the shutdown of the Airport happened the following week.

an increase in ROG at night. With the introduction of the lock-down measures ROG decreases for all times, but retains its weekly pattern, except for the characteristic nightly increase of activity on weekends, which is not recovering, even after reduction of the lock-down restrictions. Apart from weekend nights mobility at all times of day is slowly increasing towards pre-lock-down level.

4.3.2 Gendered impact analysis

Behavioral gender differences have been found for a wide range of human activities, including the way people communicate, move, provision themselves, or organize leisure activities. Using mobile phone data from 1.2 million devices in Austria across the first phase of the COVID-19 crisis, we quantify gender-specific patterns of communication intensity, mobility and circadian rhythms. We show the resilience of behavioral patterns concerning the shock imposed by a strict nationwide lock-down that Austria experienced at the beginning of the crisis with severe implications on public and private life. We find drastic differences in gender-specific responses during the different phases of the pandemic. After the lock-down, gender differences in mobility and communication patterns increased

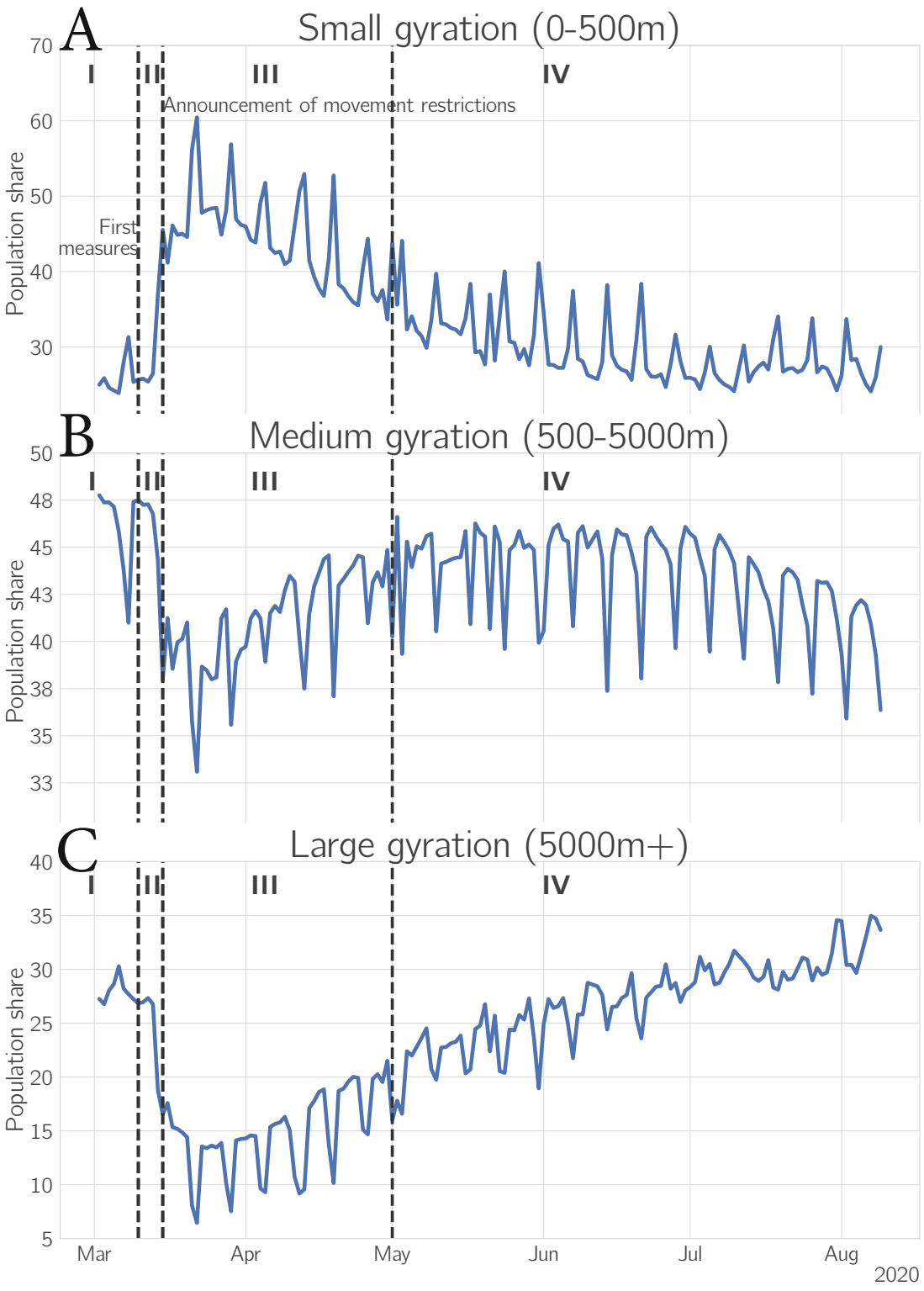


Figure 4.3: Bucketed ROG into small **A** (0 m - 500 m), medium **B** (500 m - 5000 m), large **C** (>5000 m) movement.

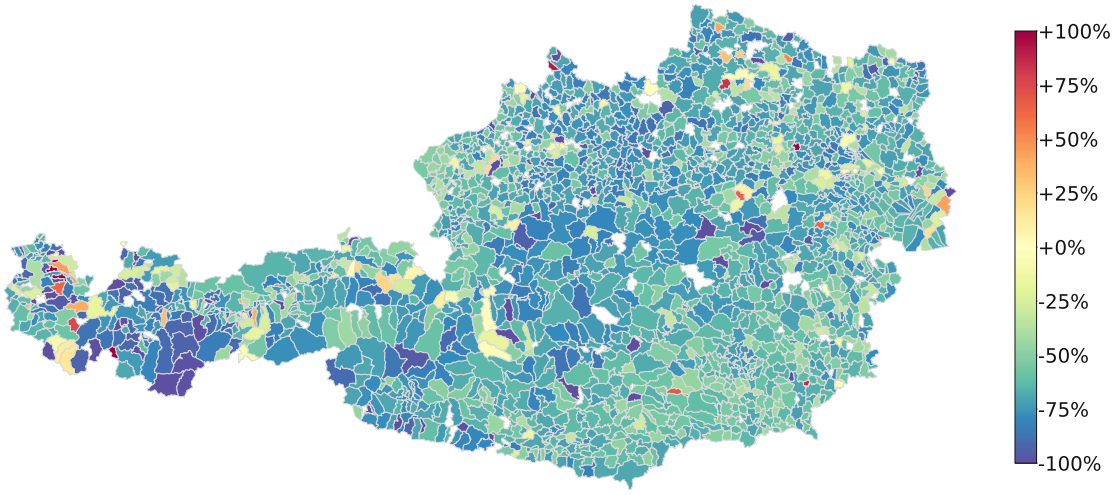


Figure 4.4: Relative change of mean ROG for week of March 2nd and week of March 23rd measured at postcode level.

massively. In particular, women had fewer but longer phone calls than men during the lock-down. Mobility declined massively for both genders. However, women tend to restrict their movement stronger than men. Women tended to avoid shopping centers and more men frequented recreational areas. After the lock-down, males returned to normal quicker than women; young age-cohorts returned faster. We interpret and discuss these findings as signals for underlying social, biological and psychological gender differences when coping with crisis and taking risks.

Empirical research has long been concerned with assessing whether women and men behave differently in their daily lives. Behavioral differences were reported in communication behavior, visible for example in the different investment in biological offspring across women and men's lifetimes [PKK+12]. Gender differences in mobility patterns do rise from a mix of cultural, infrastructure, resource, safety and socio-economic factors [GTP+20]. Psychological and cognitive and other non-reproductive differences have been studied for many decades, maybe even centuries, see e.g. [Hal13]. Also differences in stress perception and respective coping mechanisms have been known to exist for a long time [BZZ96, Mat04]. Non-reproductive biological differences include women having shorter circadian rhythms [DCC+11] and showing different co-morbidity patterns than men across their lifetimes [CKT14]. Even in virtual societies of online game players, strong behavioral gender differences were found. In particular, male and female players tend to behave differently in economic activities, their dealing with aggression and hostilities, and generally how they structure their social networks [ST13].

Times of stress may alter social norms, socio-economic constraints, and "typical" behavior. It is *a priori* not clear if and how these changes increase or decrease behavioral gender differences. On the one hand, one might speculate that stress leads to a more universal behavior, where gender differences become less critical and thus less pronounced [WXX+20].

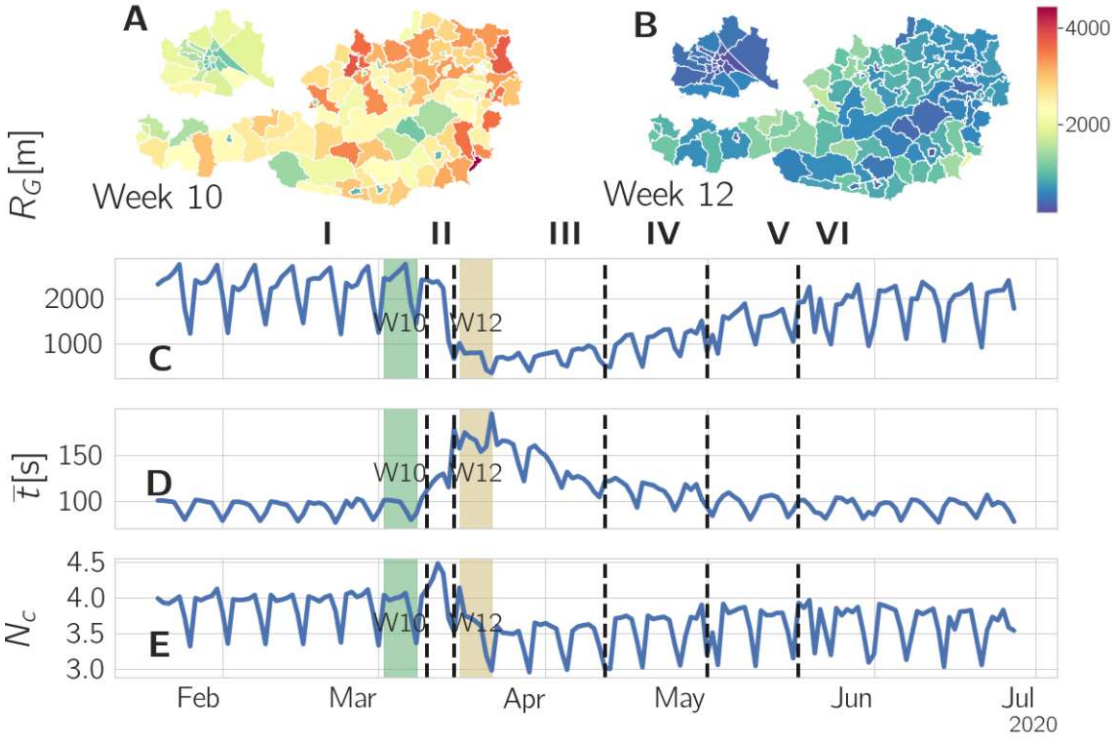


Figure 4.5: Population-wide response to the COVID-19 crisis. The maps show the mobility (radius of gyration, R_G) for calendar week (**A**) 10 and (**B**) 12 for Austria. The timeseries below outline the changes in (**C**) R_G, (**D**) the call duration per call $\bar{t}$, and (**E**) the number of calls per device N_c. During the lock-down mobility was drastically reduced throughout Austria. The call duration per call $\bar{t}$ increased dramatically and the number of calls, after a brief increase around the beginning of the lock-down, dropped below the pre-lock-down level.

On the other hand, psychological gender differences might become amplified when coping with crisis [BZZ96, GAR+02, JAD+20]. A crisis such as the COVID-19 pandemic is an exceptional shock to social systems and can be seen as a *natural experiment* that allows us to investigate the impact of population-wide stress and its consequences on gender-specific changes in behavior. Such a natural experiment can be used to estimate the *resilience* of behavioral changes, i.e., how long it takes after the onset of a well-defined shock to return to pre-crisis patterns of behavior. This characteristic time might also be necessary for a better objective understanding of temporal changes of psychological effects after emergencies, which are usually studied with self-reported data at a few points in time [BZZ96, GAR+02, Mat04, CLL20].

The uncertainty of the situation, especially the threat of job-loss or additional childcare duties caused stress and anxiety in the Austrian population [PLMG20]. Right from the start, it lead to the apprehension that women could be affected more by the lock-

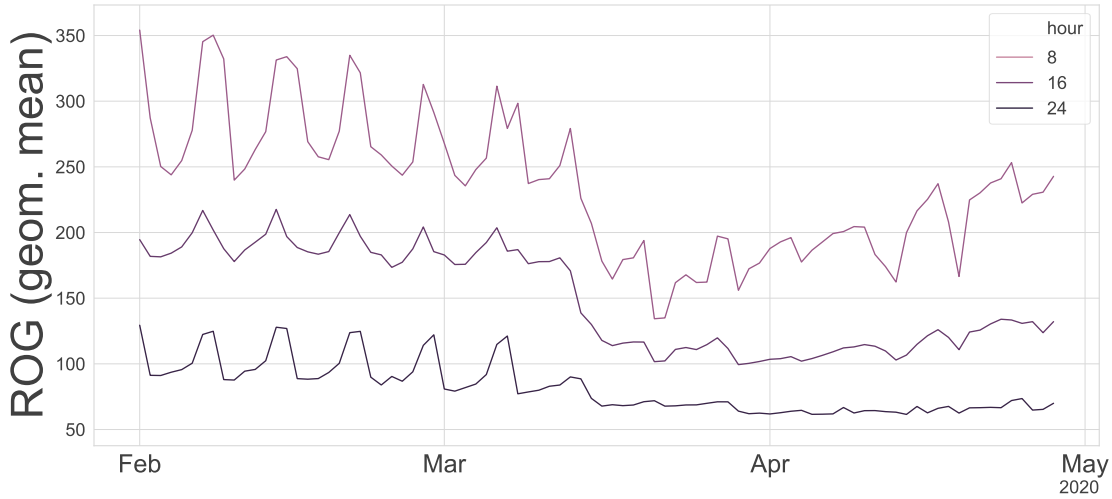


Figure 4.6: Hourly geometric mean of ROG for selected hours. The night activity on weekends is not recovering, even after reduction of the lock-down restrictions.

down due to additional childcare duties [IS20, Vig20, OEC20], domestic violence [BJI20], employment in high exposure jobs and simultaneously higher unemployment [OEC20]. Austrian women were more affected by unemployment and partial layoffs [Arb20], surveys registered an increase of domestic violence [SE20], and female scientists posted less preprints and started less projects [Vig20]. It has been argued that during the COVID-19 pandemic "disproportionately affected women and widened gender inequalities across the globe" [MRR21]. The fact that men and women react differently to stress and crises is not new. Women experience more stress [Zei06, BZZ96] and employ relatively more active and problem-focused coping strategies [BZZ96, Mat04], while men tend to emotion-focused coping, such as emotional avoidance [BZZ96].

Here we want to understand the effects of the COVID-19 crisis on behavioral gender differences in five directions: Changes in communication patterns, changes in mobility, changes in food supply, changes in spending leisure time and changes in circadian rhythms as seen in digital traces. We discuss gender as more than the distinction between biologically different sexes, but as a socially constructed categorization [Haw13].

To control for differences between our sample composition and the demographics of Austria, and to relate behavioral changes to different phases of life, we stratify our results with respect to age. Finally, we estimate the circadian activity of telecommunication and internet usage, from which we estimate e.g. gender differences in sleeping patterns.

The gender categories in our study are self-reported and are, for technical reasons, limited to female and male. We observe changes in the digital traces of humans in Austria that are shaped by the lived social experiences that are played out within specific contexts, constraints, and gendered opportunity structures. Many studies, including the present, empirically find behavioral and psychological gender differences. However, one should

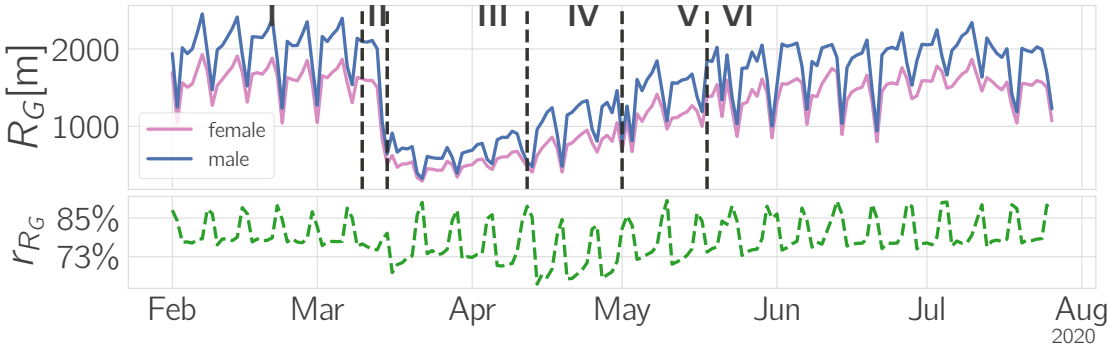


Figure 4.7: Mobility quantified by R_G. The upper panel shows R_G for men (blue) and women (pink). The lower panel depicts the gender ratio, r_{R_G}, over time. We observe a large drop in R_G for both genders in phase III and a drop in gender ratio in phases III (lock-down), IV, and V (lock-down eased).

not interpret these findings as a manifestation of an inherent difference between men and women, but as a starting point to discuss the roots of different experiences of the pandemic that are lived by women and men.

Telecommunication data has been used earlier to study the effect of crisis and emergencies. They were used to detect crisis [CGW+08], study communication patterns subsequent to different emergencies [BWB11], predict movement, e.g. subsequent to the Haiti earthquake 2010 [LBH12], and to help explain the spread of SARS-CoV-2 [GRK+20]. Gender differences in human mobility and communication were studied in [PKK+12, GTP+20]. In [PKK+12] changes in communication behavior across age and gender were reported and in particular, how reproductive investments and preferred relationships of both sexes shift over a lifespan. It is a known fact that males tend to have their workplaces further from home and thus generally move more, see e.g. [PSHS08]. Gender differences in mobility in Santiago de Chile are reported in [GTP+20]. There, significantly different movement behavior is found and is interpreted as a result of an interplay of socio-economic and urban factors. The gender specific behavioral response to seven terrorist attacks in six cities is investigated in [JAD+20]. They compare temporal mobile phone communication patterns in response to the attacks and report significant differences between the genders.

Mobility

Figure 4.7 A shows R_G for the two genders, R_G^f (red) and R_G^m (blue). The gender-ratio, defined as $r_{R_G} = R_G^f / R_G^m$ is depicted in panel B. The female population is moving less than males in pre-crisis times (phase I), as seen in the ratio r_{R_G} of 78% on weekdays and 88% on weekends. After a brief transition period II the Weekday ratio drops to around 73% during the lock-down phase III, while on weekends the ratio remains at initial levels. In phase IV, once restrictions were lifted, R_G for males returns back to

normal more quickly than for females, hence decreasing the gender ratio further down to 67%. The ratio starts to recover towards pre-crisis levels starting from phase V onward, once the main restrictions were lifted. When fitting the R_G curves as they converge to pre-crisis levels after the lock-down, we report a *half-life time* for men of $t^m_{1/2} = 34.8$d, and $t^f_{1/2} = 36.0$d for women.

The changes in gender ratios of R_G are significant between the phases. Especially the changes from phase I to the subsequent phases and from III to phase IV are indeed highly significantly. We find similar results if we replace the radius of gyration by an alternative measure for mobility that is inspired by entropy, $S_i^{f/m}$.

In Fig. 4.9 B we show the age-stratification of the gender-ratios. Before the crisis we observe very different gender ratios for different ages. Generally the ratio decreases with increasing age. For the young cohort of 15-29 years, the weekday-ratio is above 90%. For the two age cohorts above the average age of first childbirth (26.3 years for women and 28.7 for men [Aus19]), 30-44 and 45-59, the ratio is reduced to about 83%. For the age cohorts of retirement, 60-74 and 75+, gender disparity becomes even more biased towards men with a ratio of about 70%. In phase III, the three younger cohorts show an overall trend of increasing gender biases. For the age cohort 45-59, this trend is much less pronounced. Strikingly, the effect is reversed for the retirement cohorts where the gender ratio changes from around 70% to more than 80%, which again decreases towards pre-crisis levels in phase IV. The ratio for the old cohorts returns much more quickly to pre-crisis values than all the younger ones, which do not return to the previous values until the end of the observation period. We do not observe large differences in half-life times across gender, but $t_{1/2}$ is much smaller for older cohorts. For all cohorts we find values between $t_{1/2} = 38.8$d for 15-29 year old women to $t_{1/2} = 28.8$d for 75+ year old men.

The radius of gyration can be compared with corresponding data of the previous year (2019) in the same time period. We find that during the lock-down phase in 2020, there is less than 40% of the movement than in 2019.

Communication patterns

As proxies for the strength of social interactions we first analyze the call duration per pair of interaction partners, $\bar{t}^{gh}(t)$, the number of calls, $N_c^g(t)$, and the number of calling partners per user, $k^g(t)$, see Methods. The superscripts indicate gender, g represents the gender of the caller h is the gender of the called.

Figure 4.8 depicts the situation over time. In panel A we see a massive increase of calling times for the different gender combinations in phase II and the beginning of III. For the female-female calls we observe an increase of up to 140%, female-male and male-female rise by up to 81% and 97%, respectively, and male-male calls increase up to 66%. We find that calls involving women are generally longer than those involving men. Moreover, the call time increase is larger when women are involved.

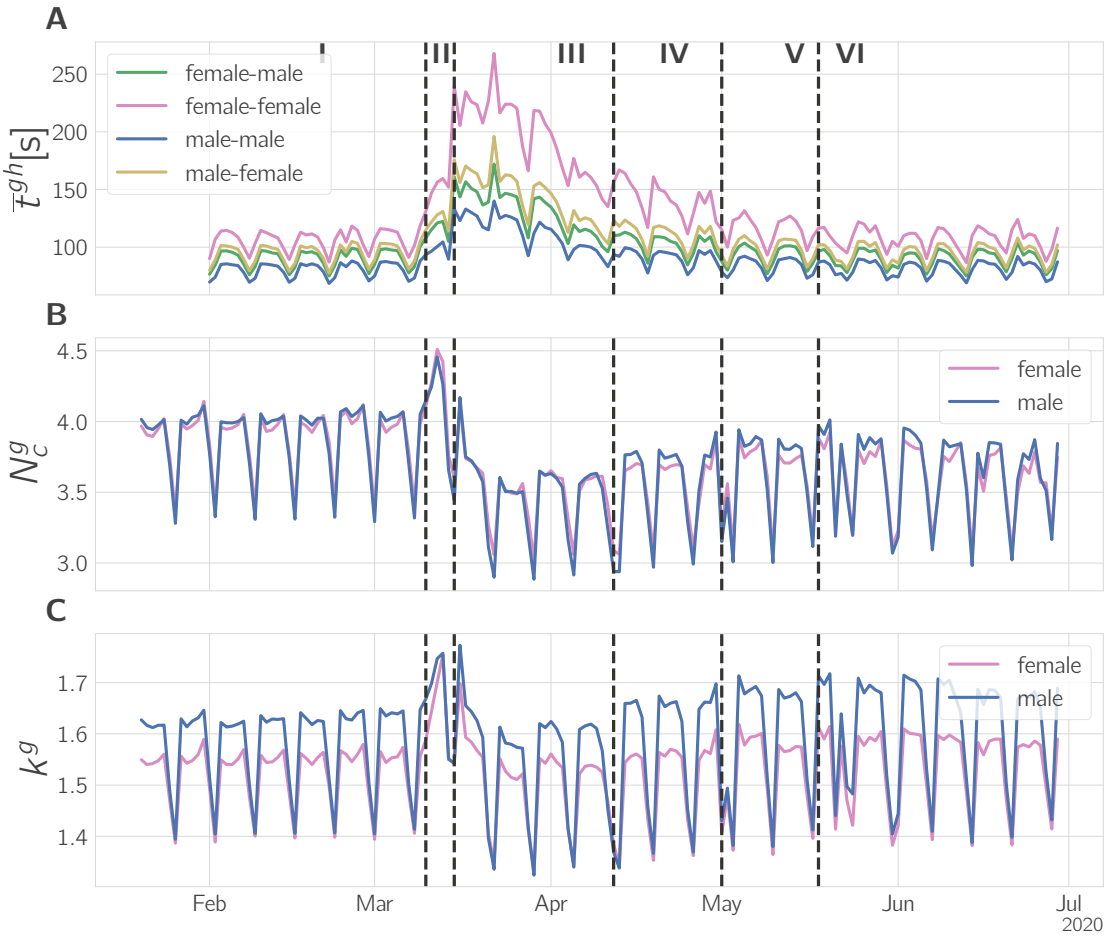


Figure 4.8: Gender-specific changes in communication behavior. (**A**) Median call duration of the four possible types of gender-specific calls, depending on who initiated the call and who received it. By mid-May pre-crisis levels are reached. Half-life times range from 17.3d in the female-female to 14.9 in the female-male case. (**B**) Number of calls originating from males (blue) and females (red). The median call duration peaks in phase III, particularly for female-female calls, whereas the number of calls assumes a minimum. Up to the end of the observation period, pre-crisis levels are not reached. (**C**) The number of communication partners, the degree $k^g(t)$, rises briefly and then drops below pre-crisis levels.

Calling times decrease gradually and reach pre-crisis levels in phase VI. This decay can be fitted with an exponential function. The exponents of the fits translate into corresponding "half-life" times, which are $t_{1/2,mm} = 15.9$d for male-male and $t_{1/2,ff} = 17.3$d for female-female interactions, the mixed interactions have half-life times of $t_{1/2,mf} = 15.5$d and $t_{1/2,fm} = 14.5$d for male-female and female-male interactions, respectively.Call times show a pronounced bias towards female initiated calls being longer. In phase I, female originated calls were 10% longer than male originated, and up to 30% longer on weekdays in phase III. From its maximum in phase III, the gender ratio continuously declines to normal levels in phase V.

The age profile for the median call duration is relatively flat for the adult and senior age cohorts and has very low values for the youngest cohort. The call duration increases slightly for the two youngest, but strongly for the two oldest cohorts. The gender ratio in call duration is biased towards women for all ages during the crisis, as seen in Fig. 4.9 A. Notably, the age cohort 15-29 is the only cohort having a more balanced call duration on weekends. For all other cohorts gender differences are increased on weekends. Around the beginning of phase III, the ratios for all except the 75+ cohort reach a maximum. The 75+ cohort reaches a maximum of the gender imbalance in phase IV.

In Fig. 4.8 B we show the number of calls, N_c^g, for male and female generated calls. Here we display the mean of N_c^g because the median due to its discrete nature in combination with the relatively small average N_c^g between 3.5 and 4.5, would make changes and gender differences hard to see. After a short increase in calls in phase II (female: +13%, male +6%) we see a significant drop in calls in phase III (both -9%), which never reaches pre-lock-down levels in the observation period. It stabilizes at a level of -5% and -4% of the previous level for women and men, respectively. There are only small gender difference in the number of calls.

In Fig. 4.8 C we show the timeseries for the number of different communication partners, k^g, i.e. the degree of men and women in their communication networks. For the same reason as for N_c^g, we show the average instead of the median for k^g. After a brief rise (up to 8% and 13% for men and women, respectively) in phase II, k^g falls below its pre-crisis level (-3% and -2%). In phases IV and V k^g rises to values higher than the initial values in phase I. In phase VI k_g is about 4% higher for men and 2.5% higher for women.

During normal times (phase I) we find that men have a slightly higher average degree (communication partners) on weekdays (f/m ratio 95%, men 1.6, women 1.55 unique contacts per day), while on weekends it is more or less balanced (women and men 1.4). In phase II, k_g is increased for both genders to a maximum around 1.73, with an increasingly smaller gender bias. In phase III the degree drops below pre-crisis levels, but men reduce k^g stronger, resulting in a smaller gender divide in phase III (96%). From phase IV onward, the degree slightly increases (even above pre-crisis levels: men 1.7 and women 1.6), even stronger for men, hence resulting in an increased gender divide (less than 94%).

Call duration increases much more than the number of calls decreases, regardless of gender. This is visible in Fig. 4.5 D and E. Just in phase II there is a drastic rise in both,

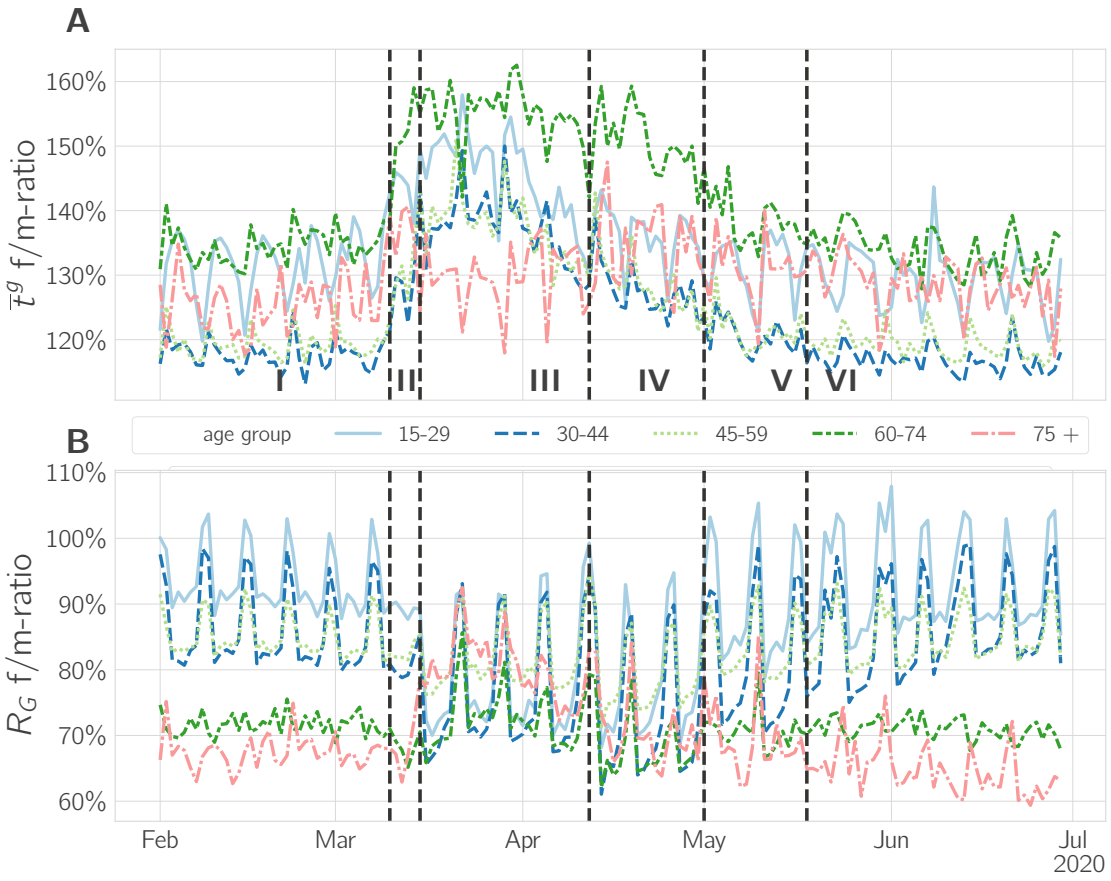


Figure 4.9: Gender ratios of communication and mobility for different age cohorts. The gender ratio of (**A**) the median call duration $\bar{t}$ and (**B**) the radius of gyration, R_G, is seen. In III the R_G gender ratio of young cohorts is shifted towards women moving significantly ($p < 0.001$) less, while for old cohorts it is shifted towards a more balanced value. In the same period, for all cohorts except 75+, the gender bias for the call duration increases towards women that have a higher call duration.

call time per call, and the number of calls. The concentration of communication partners is higher for females and increases during crisis. The bias is also shifted towards men having more communication partners in phase VI. All proxies indicate a strengthening of individual contacts and a focus on important contacts.

Gender ratios of different phases are considered to be distributed around different stationary values. Subsequently, we compare them with a two-sided Mann-Whitney-U test, and reject the null hypothesis that they are from the same distribution.

Basic provisioning

We show the number of unique devices as a proxy for the number of people at a shopping center across the lock-down. We count the number of unique subscribers in a specifically defined area.The shopping center is the largest of its kind in Austria and one of the largest in Europe. It is a cluster of 359 shops spread over an area of 670,000 m^2. Shops sell a wide range of products, including sports equipment, garments, furniture and electronics. It is visited by more than 20 million visitors each year from Vienna and its hinterland, especially in the south, as well as from Hungary and Slovakia. There are also 14 shops, including supermarkets, drug stores and pharmacies that were not affected by the lock-down.

The visiting patterns of the shopping center in phase I show a pronounced weekly periodicity with a maximum on Saturdays and very few visitors on Sundays, when all stores except cinemas and restaurants are closed. The gender ratio in phase I is close to one, indicating gender balance. In phase III the shopping complex was shut down to a large extent. No businesses other than stores for basic provisioning were allowed to open. Nevertheless we find a small number of visitors that we account mainly to persons shopping for food and drugs. The gender ratio in phases III and IV is clearly male-dominated In phase V, when shops were allowed to re-open, visitor numbers rose to pre-crisis levels at the beginning of the week, however without the strong peaks on Saturdays. The gender ratio returns to a balanced situation.

Leisure activities

In Fig. 4.10 A we count the numbers in a popular recreational area nearby Vienna, the Kahlenberg, frequented mainly for walks, and easy hikes. The number of visitors does not drop in phases II–V, but increases with the usual seasonal trend from March to June. We find more visitors on weekends and on days with good weather, explaining the high variance in numbers.

4.4 Relative changes of mobility and call duration

Evaluating relative changes leads to additional insights which would remain hidden when only considering absolute changes. We analyze a data set describing the mobility of mobile phones in Austria before, during COVID-19 lock-down measures. By applying

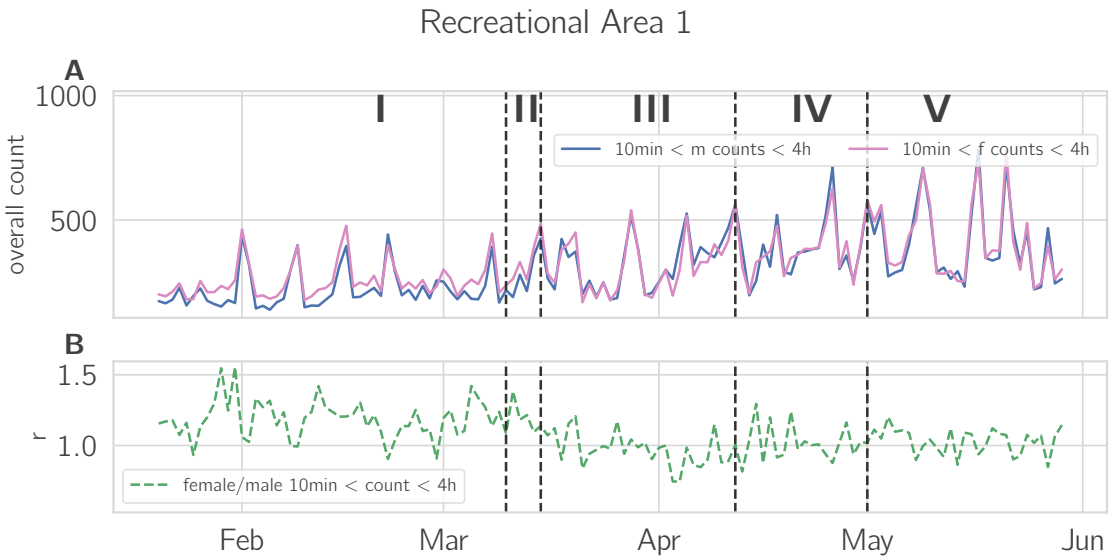


Figure 4.10: Visitors in a leisure area outside of Vienna, *Kahlenberg*, during the Covid-19 crisis. (**A**) The upper panel shows the counts of men and women present in the defined area. (**B**) The lower panel shows the gender ratio of the counts. The overall counts are unaffected from the lock-down, but the gender ratio changes from being from female-biased to equality.

compositional data analysis, we show that formerly hidden information becomes available: we see that the elderly population groups increase relative mobility. The younger groups, especially on weekends, do not decrease their mobility as much as the others.

Traditionally, a comparison is made in terms of *absolute information*, i.e., the ROG time series values of the different groups are analyzed in their unit of meters, such as [RHH+21] - see Section 4.3.

An alternative is to compare *relative information*, for example the ROG of the males with respect to females, or in terms of the ratio males to females. This leads to a dimensionless time series, and to a different aspect of data analysis which emphasizes the differences between the individual groups. A joint increase or decrease in both groups may not lead to a big change of the ratio. On the other hand, the ratio will change if the values of one group increase, and at the same time they decrease in the other group, or vice versa. Here again, the relative change rather than the absolute change is important. For example, if the ROG changes from 1000m to 2000m in one group, and from 2000m to 1000m in the other group, the ratio would change from 1/2 to 2. The same change could be observed if the absolute values in both groups would be bigger by a factor of 10. Thus, absolute values are no longer relevant in this consideration, because a multiplication by any positive constant leads to the same ratio. This is still trivial in case of comparing two groups, but it is no longer straightforward when relative information of several groups, such as age classes, should be compared. *Compositional data analysis* is devoted

to this problem of analyzing relative information [Ait86, PGETD15, FHT18]. In fact, compositional data analysis is frequently used in geosciences, but also more and more in other fields such as biology [ESS+20], bioinformatics [QERC18], economics [TMTAS19], marketing [JC18], medicine [DPPA+20], etc.

4.4.1 Compositional data analysis primer

From the point of view of compositional data analysis, a composition is defined as multivariate information, consisting of strictly positive values, where the absolute numbers as such are not of interest, and only relative information is relevant for the analysis [FHT18]. A composition can be given for example by the median ROG values of different age categories for a certain day, and every age category is denoted as a compositional part. We use the notation $x_1, \ldots, x_D$ for the compositional parts of D categories, and the composition is written as the (column) vector $\mathbf{x} = (x_1, \ldots, x_D)'$. For every day recorded in our data base we will observe such a composition, which in fact leads to a multivariate compositional time series. The interest is in relative information in terms of the ratios, and thus all pairs x_j/x_k, for $j, k = 1, \ldots, D$, should be considered in the analysis. Obviously, the pairs for $j = k$ are not relevant, and pairs of the reverse ratio x_k/x_j do not contain potentially new information. This motivates to consider the logarithm of the ratios, $\ln(x_j/x_k)$, so-called log-ratios. The reverse ratios have a different sign, and thus do not need to be considered, and their variance is the same as for the original ratio. Moreover, the distributions of log-ratios tend to be more symmetric than without a logarithm [PGETD15].

Still, the resulting $D(D-1)$ pairs $\ln(x_j/x_k)$, for $k > j$, only live in a subspace of dimension $\leq D-1$ [FHT18], and thus it is natural to aggregate this information. Consider an aggregation

$$y_1 = \frac{1}{D}\left(\ln\frac{x_1}{x_2} + \ldots + \ln\frac{x_1}{x_D}\right) = \ln\frac{x_1}{g(\mathbf{x})}\,, \tag{4.1}$$

where

$$g(\mathbf{x}) = \sqrt[D]{\prod_{j=1}^{D} x_j}$$

is the geometric mean of the composition $\mathbf{x}$. Then, y_1 represents all relative information about the part x_1 to the other parts in the composition in a form of an average of the log-ratios. This leads to the definition of so-called centered log-ratio (CLR) coefficients [Ait86]

$$\mathbf{y} = (y_1, \ldots, y_D)' \quad \text{with} \quad y_j = \ln\frac{x_j}{g(\mathbf{x})}\,. \tag{4.2}$$

The vector $\mathbf{y}$ contains all relative information about $\mathbf{x}$ in the above sense. It consists of D components y_j which are associated with the relative information about the corresponding part x_j. However, it turns out that $y_1 + \ldots + y_D = 0$, and thus a representation of data in terms of CLR coefficients leads to singularity [FHT18]. Although there are ways to circumvent this issue [FHT18], we will proceed with CLR coefficients for the following analysis for simplicity.

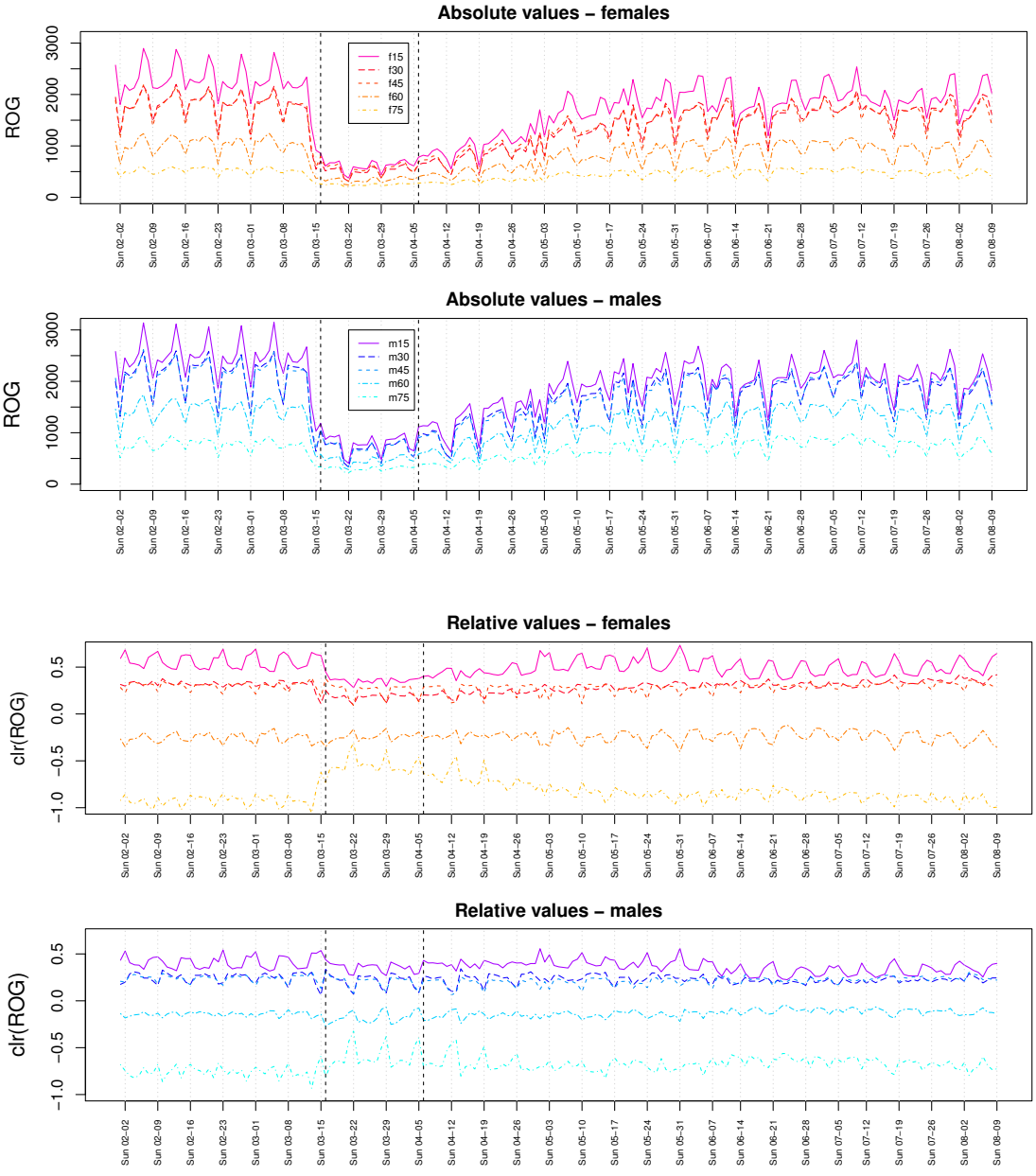
Consider now a multivariate compositional time series $\mathbf{x}_t = (x_{t1}, \ldots, x_{tD})'$, for the time points $t = 1, \ldots, T$, and the observations x_{tj} for each part $j \in \{1, \ldots, D\}$. The time series expressed in CLR coefficients is $\mathbf{y}_t = (y_{t1}, \ldots, y_{tD})'$, with $y_{tj} = \ln(x_{tj}/g(\mathbf{x}_t))$, with the geometric mean $g(\mathbf{x}_t) = (\prod_{j=1}^{D} x_{tj})^{1/D}$ per time point. Since this data representation only reflects relative information of the time series, an additional visualization of the absolute time series values can be interesting to get a more complete picture.

The CLR coefficients result in multivariate data that can be analyzed with the traditional multivariate statistical methods [FHT18]. A prominent way to represent the information in a lower-dimensional space is to use principal component analysis (PCA). Since PCA is sensitive to data outliers or inhomogeneous data, robust versions have been proposed, also in the compositional data analysis framework [FHT18]. The resulting loadings and scores are commonly represented in a biplot to get an overview of the multivariate data [AG02].

4.4.2 Compositional analysis of mobility

The results reported in this Section refer to the median values of the ROG per group. To begin with, Figure 4.11 A shows the absolute values for the females (top) and males (bottom) for different age groups as a reference to better understand the utility of the relative analyses. The legend indicates the considered age groups: 15 for age 15-29, 30 for age 30-44, 45 for age 45-59, 60 for age 60-74, and 75 for age elder than 75. For all of the following time series plots, the vertical dashed lines indicate the date March 16$^{\text{th}}$, 2020, when the restrictions came into action, and the date April 6$^{\text{th}}$, 2020, when they were relaxed. The data considered here are from the period February 1$^{\text{st}}$ until August 9$^{\text{th}}$, 2020. The plots clearly show the lock-down by an abrupt decay of the median ROG values in all age classes for both genders. After the lock-down, the order of the values remains the same, from the eldest group with the smallest values and the youngest group with the highest values, but it is on a much smaller level. The level then increased more or less systematically until the middle of June. Afterward, the level is not changing a lot, it is lower than at the beginning, and weekly patterns are clearly visible. Note that these weekly time series patterns that are very regular at the beginning are getting somehow distorted, partially also due to holidays (April 13$^{\text{th}}$, May 1$^{\text{st}}$, May 21$^{\text{th}}$, June 1$^{\text{st}}$, June 11$^{\text{th}}$), and they never get back to this regularity. Figure 4.11 B focuses on the relative information contained in the median ROG values. We consider the female age groups and the male age groups separately as two compositions. The plots show the corresponding CLR coefficients for females (top) and males (bottom). While in Figure 4.11 A we have essentially seen a decline of all values at the beginning of the lock-down phase, followed by an increase, we did not pay attention how differently the age groups declined and increased. This is the purpose of the relative view in Figure 4.11 B, where we mainly investigate the developments of the age groups to each other.

In both plots of Figure 4.11 B we can see roughly the same pattern after the lock-down: the biggest relative changes are visible for the youngest and the oldest age group, but they go into different directions. While group 15 had the biggest decline, group 75+



[A]

[B]

Figure 4.11: **A**: Median ROG values for different age groups over time for females (top) and males (bottom) in different age groups.; **B**: CLR coefficients of median ROG values for the female (top) and the male (bottom) composition.

increased the values relative to the other age groups. This seems to be counter-intuitive, but it can be explained by the fact that the geometric mean also went down significantly, and the ratio of the values of group 75+ to the geometric mean then even increased after the lock-down. Another interesting phenomenon is that the groups 60 and 75+ show the biggest increase in mobility (in a relative sense) during the weekends in this lock-down period. Although on a different level, the values from July show a similar structure to those from February. It is interesting to note that the youngest age group 15 shows a somehow mirrored weekly pattern compared to the elder age groups. This is not visible when looking at the absolute values in Figure 4.11 A. Relative information could also be understood in terms of data proportions. In particular, one could compute the proportion of a group on the total per time point, which in fact corresponds to normalizing the data per time point to a value of 1. Such a proportional presentation is shown in Figure 4.12 for the ROG values of the female age groups. Obviously, the information contained in this representation is different from CLR coefficients which focus on log-ratio information. One can hardly see any differences between the lock-down period and the remaining period, and thus this kind of "relative view" is not valuable for the analysis.

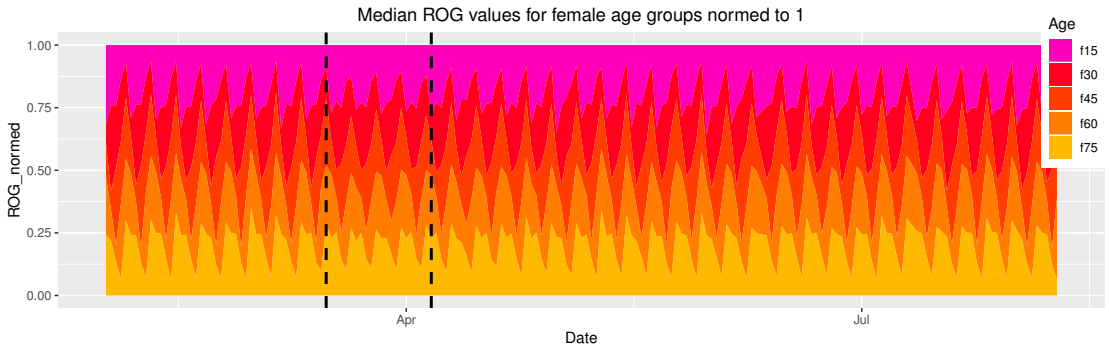


Figure 4.12: Proportional presentation of the median ROG values for the female age groups. For each time point, the data are normalized to a value of 1.

The median ROG values for the female and male age groups are analyzed in the following with PCA. Here, the method ROBPCA [HRV05] is taken, a robust version of PCA which downweights outlying observations. Figure 4.13 shows the biplot of the first two principal components (PCs) for the clr coefficients. The coloring is according to the time phases: green before the lock-down, pink during the lock-down period, purple after lock-down until mid of June, and light-blue after this period. The left biplot for the females identifies these four periods as clear clusters, while there is more overlap visible in the right biplot for the males. For the females, the direction of the first PC (71% explained variance) shows a transition of the relative ROG values from the young generation (f15, f30) before lock-down to the old (f75) one during lock-down, and then back to the center. Thus, younger and elder females show a contrasting behavior in this time period, which was already observed in Figure 4.11 B (top panel). The second PC (21% explained variance) shows also differences between the time periods, but it also reveals weekend effects.

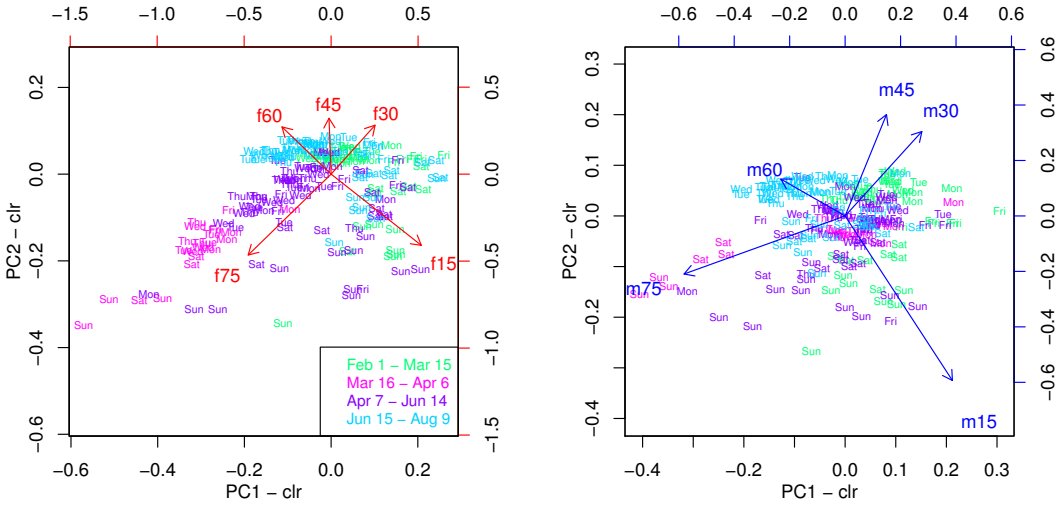


Figure 4.13: Biplots of the CLR coefficients of the median ROG values for females (left) and male (right) age groups. Green color for period before the lock-down, pink for lock-down period, purple after lock-down until mid of June, and light-blue after this period.

Especially on Sundays, the mobility for group f15 was bigger before and after lock-down, but it moved to group f75 during the lock-down phase.

The data structure in the biplot for the males (right plot) looks a bit different, but leads to similar conclusions. PC1 explains 69% and PC2 25% of the variance. Groups m75 and m15 have a similarly diverging behavior of Sunday mobility as observed for the females. The weekdays of the lock-down phase are in the center of the distribution, while for the females they were clearly moved towards group f75. On the other hand, the weekdays in the first time period (February 1st - March 15th) are better distinguishable from the weekdays of the last period (June 15 - August 9); a possible explanation is the fact that the working male population changed the mobility behavior more significantly than that of females due to home office.

A contrasting view is revealed in Figure 4.14, which shows the robust PCA results for the absolute values of ROG, for females (left) and males (right). In both analyses, PC1 explains 98% of the variability, and this direction essentially reflects the big change of the ROG over this time period. Otherwise, there is not much information left in these analyses, which reflects the limited usefulness of absolute information if the task is to compare age groups.

4.4.3 Compositional analysis of interaction

Figure 4.15 investigates the median call duration, reported in seconds, again for the two genders and the age groups. The absolute values are shown in the upper plot jointly for males and females. Here we observe the reverse ordering of the age groups compared

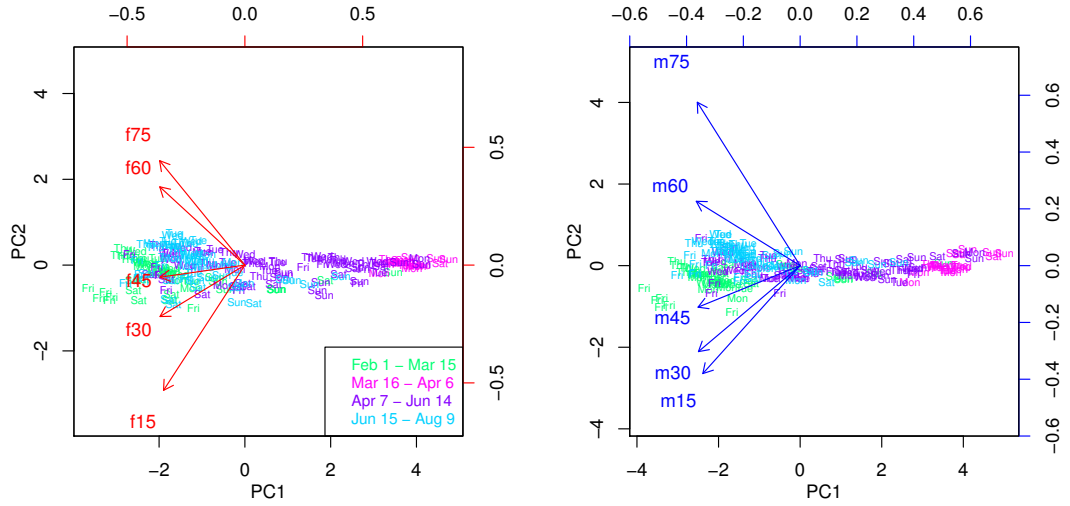


Figure 4.14: Biplots of the (absolute) median ROG values for females (left) and male (right) age groups. Green color for period before the lock-down, pink for lock-down period, purple after lock-down until mid of June, and light-blue after this period.

to the plots for the ROG values: the lowest values are for the youngest group, and the biggest for the oldest group. The values of the females are systematically higher than those of the males. It is interesting to see that the call durations already started to increase one week before the lock-down. While the ROG time series had their peaks during the weekend, we have the opposite here. This pattern, however, seems to change after the lock-down for group f75 (uppermost line), and it went back to *normality* only later on.

The bottom plot of Figure 4.15 presents the CLR coefficients, which are separately calculated for females and males, but presented here jointly for easy comparison. Although the absolute values of the youngest age group also increased with the lock-down, the increase was smaller compared to the other groups, which is reflected by decreasing CLR coefficients. The pattern of f15 and m15 has also an interesting structure: Before the lock-down, the groups had quite different behavior within their gender-group, but during the lock-down phase they became quite similar. From June on, they show again a similar behavior as at the beginning. Another interesting phenomenon can be seen after the lock-down: the two oldest groups show a contrary behavior to the other groups during the weekends. Their decline in call duration during the weekends was much smaller than that of the other age groups.

Figure 4.16 presents biplots of a robust PCA for the CLR coefficients for the female (left) and male (right) age groups. The coloring is taken as in the previous biplots, green before lock-down, pink during, purple after lock-down, and light-blue from June 15$^{\text{th}}$ onwards. PC1 explains 72% of the variability for the females, 54% for males, and PC1 and PC2 together explain about 98% variance in both cases. The different groups which are visible in the biplots are essentially weekend-effects or affects due to the lock-down.

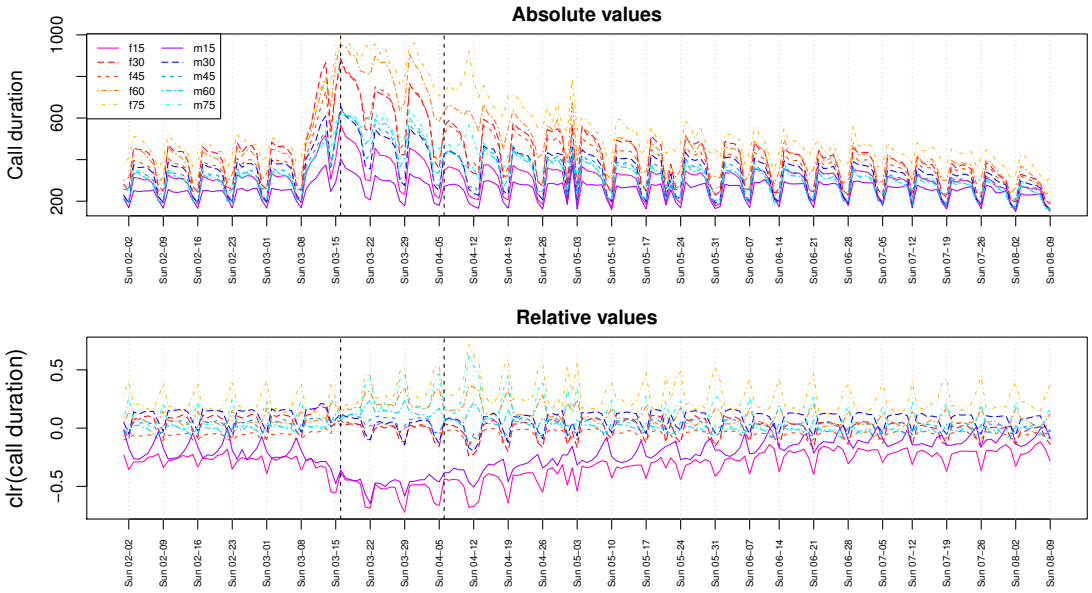


Figure 4.15: Median values of call durations per gender and age group over time (top), and clr representations separately for female and male age groups (bottom).

These grouping effects are essentially caused by the youngest and oldest age groups. When comparing the first observed time period with the last one, we can find quite clear differences in the corresponding PCA scores. These differences are essentially caused by the changing contrasting behavior between the youngest group and the elder groups; groups f75 and m75 (and also m30) do not seem to contribute to this difference. A possible explanation is the exploration of alternative methods for communication, especially for the elder groups.

Interactions between source and destination

It can be recorded who is actively calling a person, and who is receiving a call. The former person is called *source*, and the latter *destination*. Here we investigate the median ROG values for the different age groups of the females and males. However, the data set is more complex, because a person from a specific age group can be the source, while the destination can originate from a different age group. Moreover, both source and destination will have specific median ROG values.

Figure 4.17 illustrates these data for four specific cases: source f45 (f45_src) with destination f75 (f75_dst), and source f75 (f75_src) with destination f45 (f45_dst). In both cases, the median ROG values can be taken from the source group or from the destination group, see also figure legend. Throughout the whole period (here from February 1st - July 26th), the median ROG values from the source groups (solid lines) have slightly higher values than those of the destination groups (dashed lines) for the

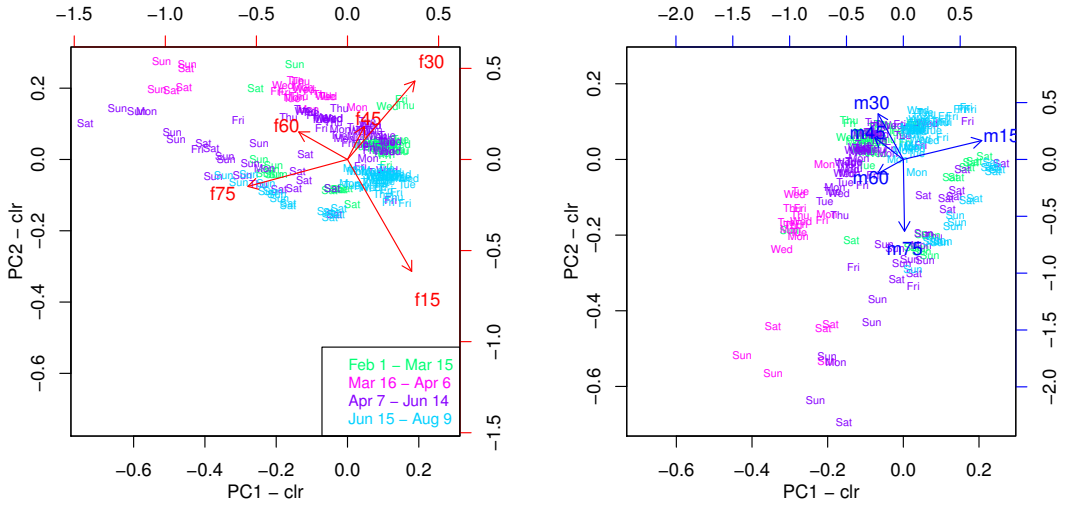


Figure 4.16: Biplots of the CLR coefficients of the median call duration values for females (left) and male (right) age groups. Green color for the period before the lock-down, pink for lock-down period, light-blue after this period.

same age classes, which can be expected because people from the source groups might call from a place outside their usual environment. While the lines are on a similar level at the beginning and at the end of the considered period, the weekly periodicity changes, probably caused by the summer holidays.

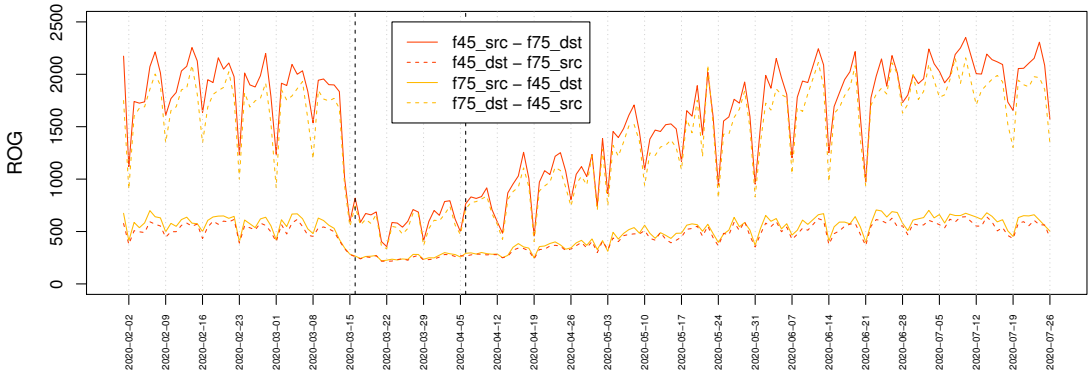


Figure 4.17: Median ROG values for the female age groups $f45$ and $f75$, depending whether they actively call (src) or they passively receive the call (dst). For example, line f75_src – f45_dst refers to the median ROG values for females in age group f75, actively calling females in age group f45.

In the following analyses we are interested in the similarity of the relative ROG values in terms of correlations, before lock-down (February 1st – March 15th) and after (March 16th – May 31th). In order to investigate relative information, the CLR coefficients are

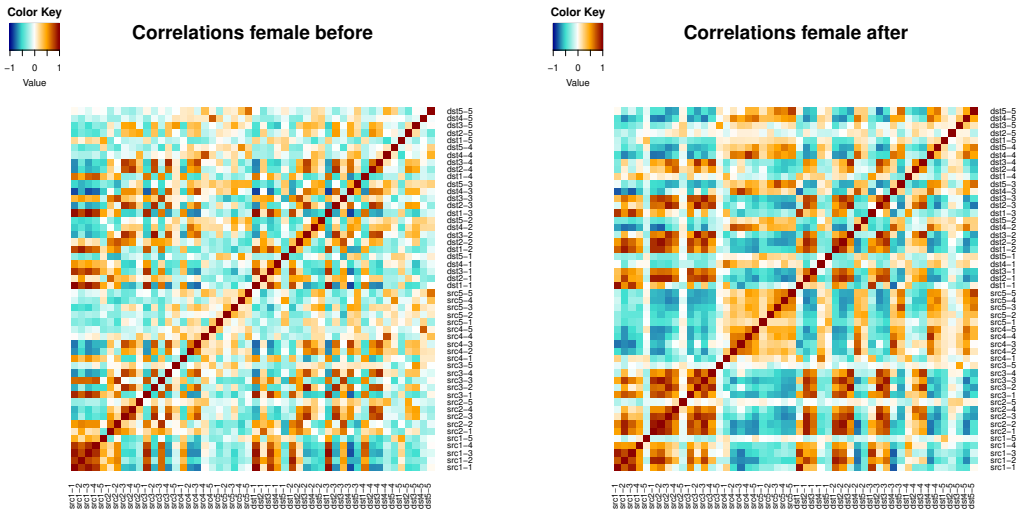


Figure 4.18: Correlations of the CLR coefficients for median ROG values for the female age groups (1 to 5, referring to 15 to 75+), when they are actively calling (src) or passively receiving a call (dst), recorded before March 16$^{\text{th}}$, 2020 (left), and afterwards (right).

computed for a composition with all 25 age combinations of the source-destination groups and all 25 age combinations of the destination-source groups, separately for females and males. Figure 4.18 shows the resulting correlation matrix for the females as a heat map, left for time points before the lock-down, and right after lock-down. The row and column labels are referring to the group numbers. For example, src1-3 refers to the time series f15_src – f30_dst, or dst5-1 is the series f75_dst – f15_src. The heatmaps show that the correlation structure before and after lock-down has clearly changed. Afterwards, there are more blocks with higher (absolute) correlations, and thus more similarity or dissimilarity between certain age groups. In general, there is a more pronounced difference after lock-down in the mobility behavior between the younger and the elder age groups.

4.4.4 Incorporating spatial location in compositional analyses

The mobile phone data also contain information about the location, in our case about the Austrian political district in which the phone has been used. The Austrian regions had different restrictions during the lock-down phase, and in particular people from all districts in Tirol had the strongest movement restrictions. Thus, in Figure 4.19 we compare the median ROG values for Kitzbühel, a district in Tirol, and Zell am See, which is also a rural district but located in Salzburg. The absolute values of the female age groups are shown in the upper plots, while the CLR coefficients are presented in the lower plots. Since the same scale is used along the vertical axes, one can clearly see the difference in mobility during the lock-down period in Kitzbühel and Zell am See, and this is also visible in the CLR coefficients. For Kitzbühel, there is much smaller variability of the values during lock-down, and also the relative differences between the age groups

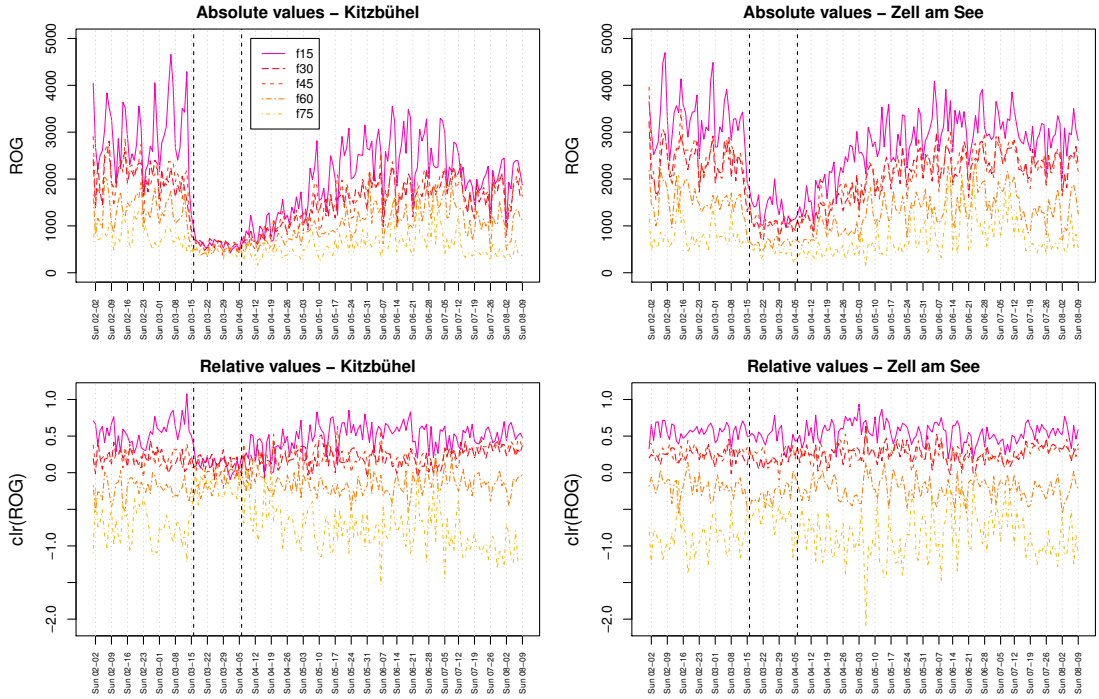


Figure 4.19: Median ROG values for Kitzbühel (left) and Zell am See (right) as absolute (top) and relative (bottom) information.

become much smaller. The change in the relative differences is not so pronounced for Zell am See. This means that also from a relative point of view, the data structure changes completely in Kitzbühel due to the restrictions.

Figure 4.20 focuses on the male age group m30, and compares the composition of all districts in Tirol with that of all districts in Salzburg. The dashed lines refer to the district capitals (Innsbruck and Salzburg, respectively). These districts behave differently compared to the other districts which are rural with many people commuting to their work place. The values of the districts in Tirol (except Innsbruck) get closer to each other after lock-down, and they start to diverge only in the middle of April. This may be explained by a similar mobility behavior of the m30 group within this period, probably caused by home-office or reduced working time. This seems different in districts of Salzburg, where the CLR coefficients show more variability after lock-down.

4.5 Summary

In this work, we described the changes in human mobility in Austria during the lock-down with regards to the SARS-CoV-2 pandemic using near-real-time, anonymized mobile phone data. We discussed mobility changes for very confined regions such as metro stations, airports or single villages, as well as regional and national changes. The results

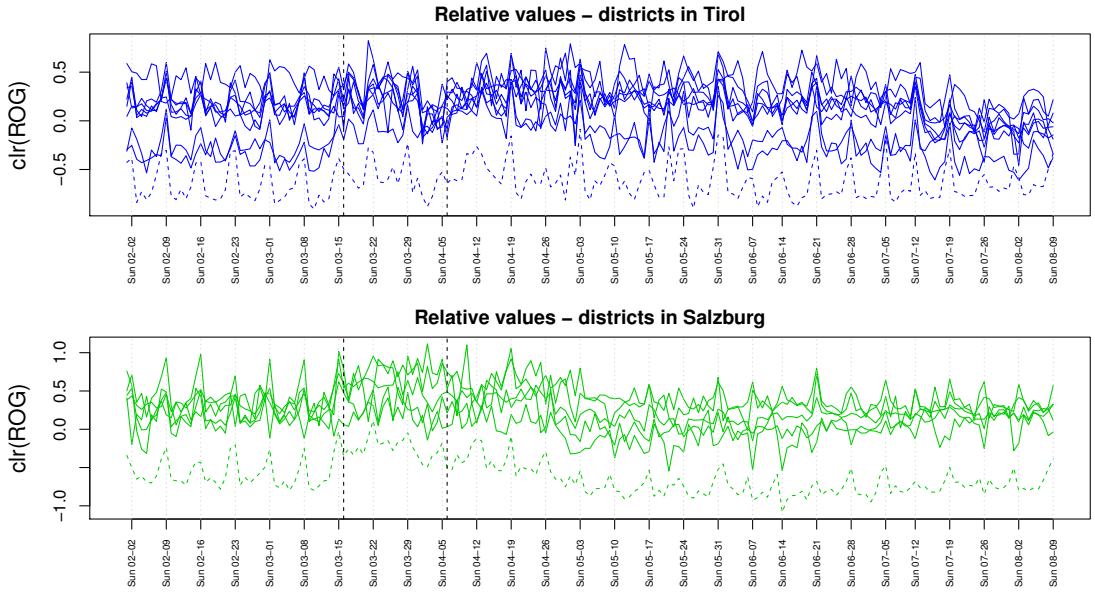


Figure 4.20: CLR coefficients of median ROG values for the male group m30 in all districts of Tirol (top) and Salzburg (bottom). The capitals are shown as dashed lines.

showed that the announcement of restrictions led to a dramatic reduction in human mobility in the whole country.

For all the movement metrics we consistently observed a reduction. This reduction is followed by an increase in modularity, as communities decompose into clusters. A similar observation can be made in the other direction as the mobility subsequently recovers. However, for the POI based indicators we cannot observe such a recovery.

Our analyses could be improved if data with a better accuracy were available, such as based on triangulation. Other limitations apply as well: we only analyze data from a single big ISP, its regional market share might vary. Furthermore, we localize the data only with the coarse cell-id. As a result, in rural areas the accuracy is in the range of a couple of kilometers, whereas for a city usually approximately 500 m. In certain rural areas there might be no cell tower coverage at all. Moreover, people from time to time leave their phone at home, which would not generate movement in our data set even though people are moving.

Furthermore, standardizing on a (sub-)set of the presented measures would allow comparison between network providers in the same country or between multiple countries. This would improve over the analyses of the European commission [SSS$^+$20, ISS$^+$20] where inconsistencies between different countries are mentioned. Additionally, limitations of this study such as the daily re-anonymization of the data could be lifted to perform a long-term evaluation of the impact of a pandemic on mobility.

The COVID-19 pandemic represents a unique natural experiment to understand individual

and collective coping mechanisms with respect to stress and crisis. Telecommunication data reveals almost real-time insights into many aspects of daily life without interfering with the subjects' actions and interactions. Using anonymized mobile phone data of a large fraction of the Austrian population, we find that gender differences that can be observed in communication patterns, mobility, and spending leisure time are amplified during the crisis, imposed by a severe lock-down in the first phase of the COVID-19 crisis. In the context of basic provisioning, we find indications that during the crisis there exists a bias toward men doing the shopping for food that is absent in normal times.

For both genders we observe an increase of total call duration, which is due to an increase of the call time per call and, interestingly, a decrease in the number of calls. This is a clear sign that communication becomes more focused and intense. This finding is in line with a general decline of the number of communication partners during the lock-down, suggesting a focus on a core of communication partners. The reduction of communication partners could result from the loss of conversation partners from work, however, we also observe a reduction on weekends, where one would not expect effects from professional contacts. The degree distribution before the crisis is in line with earlier work on mobile phone data [OSH+07]. While they find a mean degree 2.34 (averaged over 18 months), we get a smaller value of 1.53, presumably, because we average over 24 hours. However, we find the same power-law exponent, ~ -8, for the degree distribution. In these quantities we see a clear increase and amplification of the gender-biases.

Women show a smaller decrease in the number of calls and a stronger increase in call time per call. As a consequence, the gender ratios of the respective quantities shift towards females. Women have been reported to have more tightly knit (online) networks than men [ST13, ITY05]. We interpret our findings as a signal that this behavior intensifies during crisis. The tightening of the social network can also be attributed to social carework, such as calling lonely elderly, which was reportedly performed more often by women during the lock-down [Pra20]. In previous studies, women were reported to employ more active, problem-oriented coping strategies such as emotional and social support, while men show rational and detachment strategies in response to everyday stress [Mat04] and during a community crisis [BZZ96]. This, again, supports the expectation that women seem to tighten their social networks more than men.

We find that the recovery time to women's total call time initially is as fast as for men, but later, clearly slows down. The increase of demand for communication can be interpreted in the context with higher needs for communication as a coping strategy in an ongoing crisis [BZZ96, Mat04]. It also aligns well with the fact that women experience more stress than men [MAH+09], have higher levels of post traumatic stress disorder [SWF00], and have a higher prevalence to depression, partly due to "stress responsiveness" [PB10]. For the COVID-19 pandemic, similar results have been reported. For example, a study in Spain found that women showed more symptoms of depression, anxiety and PTSD, more feelings of loneliness, and less spiritual well-being when compared to men [AGSCM20]. Our result could be confounded by gender differences introduced by work environments. However, increasing gender-ratios in call times per call and the number

of calls on weekends are indicators that the confounder indeed weakens the effect on weekdays.

The age stratification of call times and the number of calls seemingly suggest that younger cohorts communicate less than older ones. We attribute this to a higher proportion of instant messaging services [TMJ+16] and other modern communication channels in the younger cohorts. Here a channel selection bias towards younger cohorts using web-based communication services more actively acts as a severe confounding factor.

The female population is moving less over the entire period, confirming earlier work in different countries and contexts [GTP+20, PSHS08]. The decrease in mobility, following the lock-down is stronger for women. In addition, men recover their mobility behavior much more quickly after the measures are lifted. This effect depends on age. For the young and adolescent population the existing gender-bias in mobility is enhanced, while for those above retirement age the bias reduces. We relate this to childcare duties during the reproductive age and gender specific differences in occupation. Unequal distribution of childcare work has been a large concern at the beginning of the pandemic [IS20, Vig20, OEC20]. Several studies identified it as a driver of gender inequality [MWNM20, Gra20]. Our data supports this hypothesis as the gender ratio is significantly (MWU $p<0.0001$) more equal after the school openings in phase VI. Occupational differences become apparent in the unemployment numbers at the beginning of phase III, where the increase for women was 8.7% larger than for men (women +67,5%, men 58,8%)[Arb20].

In addition to care-taking duties and occupational differences, the literature suggests an additional effect: Women have been shown to exhibit more ethical behavior, at least where it is socially desirable, while men often behave less community-aware [BOS89, DO11]. For women, it has been shown that they are 50% more likely to adopt non-pharmaceutical interventions in response to a respiratory epidemic [MDV16]. In this context, the reduction of mobility in women could be partly attributed to responsible behavior in staying at home to protect vulnerable parts of the population. This argument is supported by a qualitative panel survey, that reports women taking the COVID-19 pandemic more seriously in Austria [ELPK20, KKB+20].

Since it seems that men move more for work-related issues and are more often responsible for gathering basic provisions during the lock-down, they are more exposed to the perceived danger of catching SARS-CoV-2. One could speculate that this might be a sign of higher risk-taking behavior in men, in line with several previous arguments [Gus98, BMS99, ST13]. For a conclusive clarification of this matter, obviously, more research is needed.

Generally, gender differences in mobility decrease on weekends. We confirmed that the radius of gyration is larger for men because they commute more/farther [GTP+20]. This suggests that a main factor for our observed behavioral changes is indeed employment. Further evidence for this hypothesis is found in the fact that only for the 60+ age cohort the gender-ratio does not change between weekends and weekdays. Nevertheless, the

effects discussed above persist on weekends and our conclusions remain valid.

By analyzing relative changes using compositional data analysis methodologies formerly hidden insights can be identified. In this work of analyzing mobility during the COVID-19 lock-down measures we see that certain age-groups of the population (elderly, young during weekends) do restrict mobility less than other members of the population. Especially for the elderly which are at high risk of infections potentially additional reminders should be sent to adhere to the interventions. Similarly, for the young groups on weekends additional reminders to use mouth nose protection could be useful.

We find that that for both genders we observe an increase of total call duration, that the recovery time to women's total call time initially is as fast as for men, but later, clearly slows down. The decrease in mobility, following the lock-down is stronger for women. In addition, men recover their mobility behavior much more quickly after the measures are lifted.

CHAPTER 5

Reconstructing supply networks from mobile phone data

Remarkably little is known about the structure, formation, and dynamics of supply and production networks that are one foundation of society. Neither is the resilience of these networks understood, nor do we have ways to monitor their ongoing change systematically. Systemic risk contribution of companies was hitherto not quantifiable since supply networks on the firm-level did not exist except for a very few countries. We use readily available telecommunication data to reconstruct nationwide firm-level supply networks in almost real-time. We find the conditional probability of observing a supply-link, given a communication-link exists, to be about 90%. The resulting networks allow us to reliably quantify the systemic risk of individual companies and thus estimate a country's economic resilience. We identify 65 companies that could potentially cause massive damages. The method can be used for objectively monitoring change in production processes which might become essential in the green transition.

Even though possible, inter-firm or organisational networks have so far not been studied systematically with mobile phone data.

5.1 Estimating supply networks from mobile phone data

Bilateral interactions between the agents in an economy lead to networks that dominate practically all aspects of the economy, ranging from networks of production [FA10, DMR15], finance [BEST04], distribution [KKGB10], consumption [DGFP20], and recycling [SS97]. Networks are not only the basis of the efficient functioning of the economy. They are also the source of some of its implied risks and, in particular, systemic risk, or the risk that a large fraction of networks stop to function and do no-longer fulfil their function. Remarkably, the understanding of the economy in terms of its underlying networks has not arrived at mainstream economics [Art21].

For about two decades, systemic risk has been associated with network structures and ways to quantify it are nowadays available.The main idea behind the quantification of systemic risk is to estimate the economic or financial consequences of a defaulting node or link in a given network on the entire system. The fraction of the total system affected is typically associated with the systemic risk of a node or link. Knowing the systemic risk contributions of agents offers a way to quantify the resilience and robustness of a system. The first networks available to research were financial networks such as networks of inter-bank claims and liabilities [BEST04], or of overnight money markets [IDMP+08]. Systemic risk in these networks was first quantified with network measures like betweenness centrality [BST04], which were later improved by explicitly incorporating economic default mechanisms and the associated accounting procedures [BPK+12, TP13]. Further extensions involved multilayer networks [LBR14, PMBMJ+15], overlapping portfolios [PPT21, CS19], in the context of financial networks, as well as some applications in the real economy [FTFS16], and lately, also in production networks [IT19, DBR+21].

Systemic risk in mainstream economics has often been discussed not on the basis of networks [AB11, APPR17], but on financial time series data that obviously can't account correctly for cascading processes. It is precisely the cascading that leads to extraordinary large effects that are often associated with the fat tailed distributions of losses [MB19]. The default of Lehman Brothers in 2008 [Hal09, Lon10], the 2008-2010 global food crisis [dWK+16] and, more recently, world wide supply chain disruptions due to the COVID-19 pandemic [RL20, MT21] are only a few examples of severe events in financial markets, basic provision, or production networks, where cascading plays an essential role.

A network-based quantification of systemic risk makes it possible to identify the weak points in these systems and consecutively allows one to design mitigation strategies, for example an adaptive systemic risk tax to reduce the systemic risk in a banking system [PT16, LT17] or the computation of optimal networks that carry a minimum of systemic risk [DPT20, PPT21]. However, the computation of systemic risk requires the detailed understanding of the structure and dynamics of the underlying networks, which hitherto posed a major challenge [BL18].

This is particularly true for systemic risk in production networks. Only for very few countries buyer-supplier relations are known on a granular level of individual companies from which the supply-chain networks can be constructed. For Hungary value-added tax (VAT) data exists that specifies which company pays VAT to another. From this the *exact* national supply-chain network has been reconstructed [BS+20], containing more than 89,000 companies and 235,000 business relations (links). Using estimations for production functions for these companies makes it possible to obtain the national production network. Using this as an input, firm level systemic risk for all individual companies were computed by using an appropriately designed SR measure, the Economic Systemic Risk Index (ESRI) [DBR+21]. It is a network-based measure to estimate the fraction of the total production output (goods and services) of the economy that is affected by a firm's (short-term) failure.

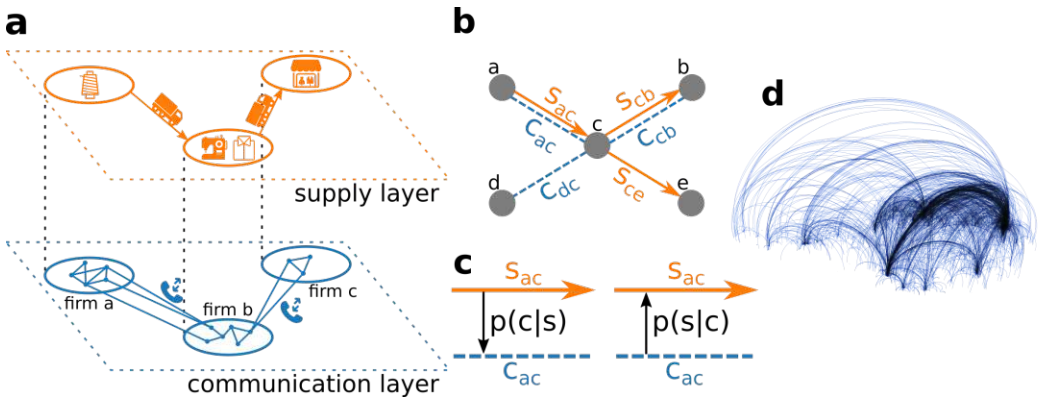


Figure 5.1: (a) Schematic view of the inter-firm multilayer network with a communication layer (blue) of phone calls between groups of devices that are associated to firms and the supply layer that captures the actual flow of goods (orange). (b) Section of the multilayer network where communication links, c_{ij}, exist if at least one phone call between firms i and j takes place and supply links, s_{ij}, exist if goods flow from i to j. (c) Conditional probabilities between supply links and communication links are defined as the probabilities to find a supply link, conditional on a communication link being present, $p(s|c)$. (d) The inter-firm communication network as provided by a mobile phone company. Arcs link firms that have an average call duration of more than 150s/d. Firms are slightly dislocated randomly, enough to ensure the anonymity of companies.

However, the Hungarian data is an exception. Granular and exhaustive data sets on the supply network of an entire nation are notoriously hard to obtain. Data exists only for a handful of countries, Japan [FA10], Belgium [DMR15], Brasil [SAT20], and Hungary [BS$^+$20]. Customer-supplier relations are inferred either from surveys and business intelligence [FA10], payment system data [SAT20], or VAT data [DMR15, BS$^+$20]. Survey data is typically very costly to collect and suffers from being outdated, highly incomplete, unweighted, and hard to verify [BL18]; on the other hand, payment system and tax data –in countries where it is collected– is sensitive and access is highly restrictive.

Here, we propose an alternative approach to reconstruct the supply-chain network by using the multilayer network structure of firm-to-firm relations. We assume that companies that communicate with each other also entertain customer-supplier relations. We thus focus on two network layers, the flow of goods and services that constitute supply relations and the mobile phone communication between companies. Figure 5.1a schematically depicts the two-layer network. The communication layer (blue) shows the mobile devices belonging to one firm, calling devices in other firms. The supply layer (orange) represents the flow of intermediate products (or services) between firms. In Fig. 5.1b we show the same situation by showing a communication link c_{ij} (blue) between firm i and j if they had at least one phone call within a certain time period and a supply link s_{ij} (orange) if goods or services flow from firm i to j. Note that communication links are undirected, supply links are directed.

The coordination of a customer-supplier relation, such as ordering, negotiating prices, or organizing shipping, requires communication between firms and has been studied intensively [HLC04, PLC08]. We thus expect the existence of strong link-correlations between the communication and supply layers. From the multilayer network in Fig. 5.1b we define the conditional probability, $p(s_{ij}|c_{ij})$, to find a supply link, s_{ij}, between firms i and j given that a communication link, c_{ij}, exists, and vice versa, the conditional probability, $p(c_{ij}|s_{ij})$, to observe a communication link given that a supply link exists, see Fig. 5.1c.

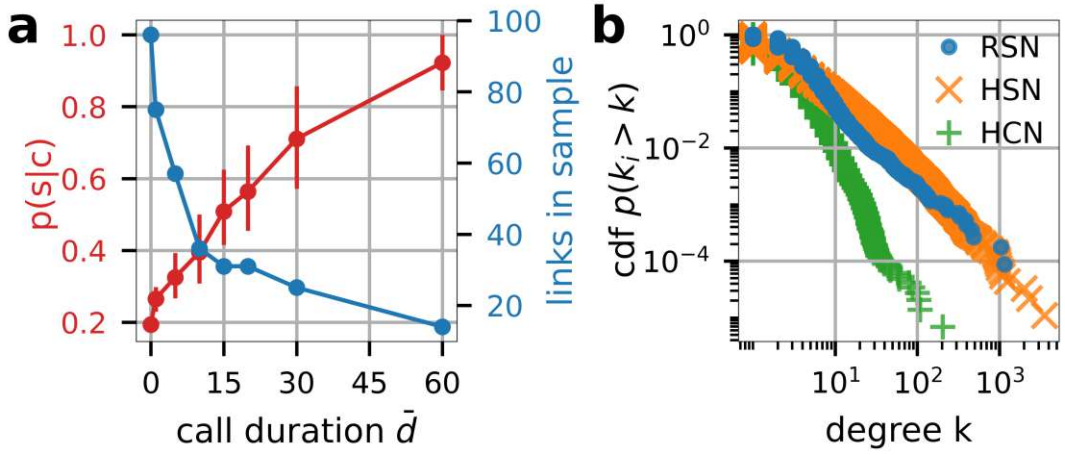


Figure 5.2: (a) Probability $p(s|c)$ to find a supply link, s_{ij}, given that there exists a communication link, c_{ij}, between firms i and j for communication links exceeding a given call duration, $\bar{d}_{ij}$. Error bars denote the quartiles of a bootstrap simulation.(b) Cumulative distribution function $p(k_i > k)$ for the degree k of the RSN (blue dots), HSN (orange x's) and HCN (green pluses). The degree distribution of the HSN is much more similar to the RSN than the HCN. Errorbars denote the quartiles of a bootstrap simulation.

Through a cooperation with a large mobile phone provider we have access to a dataset of CDRs in a medium-sized European country that allows us to identify groups of phones that are associated with a company through anonymized billing information, for details, see [RHD+22]. The dataset contains additional information on the firm's primary industry classification and balance sheet information. In Fig. 5.1d we show the corresponding firm-to-firm communication network (FCN) as obtained from our data. Firm locations are shifted by random distances (on average 30km) to ensure the anonymity of companies. Arcs in the figure represent communication links between firms. We find many short-range interactions within one city or economic region and few long-range interactions. We are intentionally vague with regards to information concerning the mobile phone provider because we are contractually bound to ensure its anonymity, as well as to protect sensitive business information such as the exact market

share in the business-to-business market.

Here we demonstrate that phone data can indeed be used to reasonably reconstruct supply networks that allow for a meaningful estimation of firm level economic systemic risk of an economy. The method is an efficient alternative to survey, tax, or bank transactions estimates. It uniquely allows us to study supply networks and monitor economic systemic risk in real time and provides a nearly complete overview of a nation's production network.

5.2 Description of the estimation process

The anonymized (but fine-grained, device-level) call detail record (CDR) data is mapped to an anonymized ID for each company. The observation period is approximately 125 days in autumn 2020 between two lockdowns. The obtained edge list is aggregated for the whole observation period, grouped by each source/destination, anonymized firm ID tuple and the call duration (in seconds) for each arc is summed up. Further, node-level statistics i.e. the number of devices is aggregated.

The firm communication dataset is merged with a commercially available business intelligence database that includes balance sheet information the industry classification in the NACE 2008 system [nac06]).For analysis, we drop NACE J61, J62, M70, and N82 to exclude businesses such as call-centers that have telephone activity at the center of their business and would confound the study with exceptionally high numbers of calls.

To compare the reconstructed supply network (RSN) with a real supply network we use a dataset based on granular VAT reporting in Hungary (HSN), described in [BS$^+$20, DBR$^+$21]. It contains a link between two firms only if at least two transactions occur in two different quarters. We use the data from 2017, where only transactions with a tax content larger than 1,000,000 Forint (approx. 3000€) are included. Hungarian VAT rates range from a 27% base rate to a 18% and 5% reduced tax rate for certain foods, pharmaceuticals, etc. and a 0% rate for public transport [Del15]. The calculations presented here are based on an unweighted version of the Hungarian production network.

We further compare the topology of the FCN with a human-to-human communication network (HCN). To this end we use a dataset provided by the same phone provider. It contains CDRs of calls between individual mobile phones which are anonymized with a new key every 24 hours. For this reason we can only analyze the HCN of one day. We choose September 17, 2020, a Thursday during the observation period outside of the holiday season and before the winter lock-downs. On that day we find 144,516 active devices and 154,557 calls.

We use input-output tables containing information on how many intermediate goods or services were used for the overall production of a certain good in a national economy in a given year. We use the input-output table of 2017, it is the latest available of the country studied. It contains 64 sectors in the CPA classification (*Classification of Products by Activity* [cpa08]), which is harmonized with NACE 2008 on level 2.

5.3 Conditional supply-link probability.

We determine the conditional supply-link probability $p(s|c)$ by comparing the firm communication network, shown in Fig. 5.1d, with ground-truth information on the real customer-supplier relations, obtained from a nationwide survey in April 2020. In the online survey more than 100,000 companies and businesses were asked to share their ten most critical suppliers and customers, respectively. More than 5,900 firms declared at least one supplier or customer with a total of more than 17,000 customer-supplier relations reported. For details on the survey see [RHD+22].We obtain the overall probability that a supply link exists between two companies, given that they had at least one conversation event in the observed period of approximately 150 days, is $p(s|c) = 0.19$. For the conditional communication probability we get $p(c|s) = 0.27$. For comparison, the respective marginal probability from the firm communication network directly is $p(c) = 0.002$. For the linking probability —using Hungarian data— we get $p(s) = 0.00005$. Since both values are orders of magnitude smaller than the conditional probabilities, highly significant link correlations between the supply and communication layers are indicated.

The conditional link probability increases with the intensity of the firm-firm communication. As a proxy for the latter we use the average daily call duration, $\bar{d}_{ij}$, in seconds per day. In Fig. 5.2a $p(s|c)$ is shown as a function of $\bar{d}_{ij}$ (red). The number of links used to calculate the overlap is shown in blue. $p(s|c)$ rises from 19% to values around 70% for $\bar{d}_{ij} = 30$s/d and around 90% for 60s/d. The number of links reduces from 75 to 14 links as $\bar{d}_{ij}$ increases. Note that errors do not increase, because a higher probability is associated with a smaller error. For details of the computation and errorbars see [RHD+22].

For the supply network ($p(c|s)$) the best proxy for tie strength would be the amount of traded goods. However, this information is not available, so we estimate the link weight as the product of the firm's sizes. Here, to stay consistent on the communication data, we proxy the firm size with the number of devices associated with a firm. We find an increase from 27% to around 60% for the network of firms with 4 or more devices. For thresholds larger than 4, the curve levels off and stabilizes around 70% for thresholds of 6 or more devices. The thresholds were chosen based on a distributional analysis of the data.

5.4 Reconstructing the supply network.

For obtaining an estimate of the supply network, based on the FCN, we chose $\bar{d}_{ij} = 30$s/d, with the aim to balance the loss of information due to ignored supply links and increasing link correlations due to the thresholds. This particular threshold is the result of a minimization of the Kullback-Leibler divergence for degree distributions of the HSN and thresholded FCNs.We arrive at an unweighted and undirected reconstructed supply network (RSN). To get an estimate for the link directions (firm i supplies j or vice

versa), we use classical input-output tables of the national statistical office. They contain information on the volume of trade between economic sectors in the economy. An element of the input-output table, G_{ab}, describes the flow of goods (in Euro) from sector a to sector b. We denote the number of links (firm-firm supply relations) from sector a to sector b by L_{ab} and assume that the ratio of links from one sector to the other is proportional to the ratio of goods flowing between these sectors, $L_{ab}/L_{ba} \approx G_{ab}/G_{ba}$. For example, the flow between the agricultural sector (a) and the food industry (f) is $G_{af} \approx 3,400$m€, while the food industry sold goods for $G_{fa} \approx 450$m€ to the agricultural sector. We now assume that it is $3,400/450 \approx 7.6$ times more likely that a supply link points from a firm a to one in f. We now consider every link from firm i in sector a to firm j in sector b in the RSN and assign it a direction according to the probability

$$p(i \to j) = \frac{G_{ab}}{G_{ab} + G_{ba}} \,. \tag{5.1}$$

Since we perform this assignment stochastically, we should think in ensembles of RSNs. Finally, we estimate a supply-link weight for every link in the RSN. We use the companies' total assets, calculated from the balance sheets, as size information, s_i; it is obtained from a commercially available business intelligence database, see Materials and Methods. Following the philosophy of "gravity models" in economics, we assume that large and small firms typically trade large and small volumes, respectively [And11]. Therefore we obtain a link weight estimate between firms i and j as the product of firm sizes, $W_{ij} = s_i s_j$. We will use only relative weights in the following.

5.5 Comparing network topologies of supply-chains, firm-firm communication, and human communication

It is enlightening to compare the network topology of the so-obtained RSN (blue) with the topologies of the Hungarian supply network (HSN) (orange) (for which exact topology is known [DBR+21]) and the private communication network between individual people (green) (i.e. not between companies). Figure 5.2b shows the degree distribution of the RSN (blue) in comparison to the exact Hungarian supply network (HSN) derived from VAT data [BS+20]. Both networks are similar and fat tailed, in contrast to the human communication network (HCN) that was obtained from the mobile phone data set. The RSN has an average degree of $\langle k^{RSN} \rangle = 4.79$. Its degree distribution has a maximum at $k^{RSN} = 2$ and its fat tail can asymptotically be approximated by a power law exponent $\alpha_k^{RSN} = 2.18(12)$ for $k^{RSN} > 30$. The HSN does not show an increase for small k but also exhibits a fat tail with $\alpha_k^{HSN} = 2.40(3)$, for $k^{HSN} > 30$. The average degree is $\langle k^{HSN} \rangle = 2.1$. For the HCN we find an average degree of $\langle k^{HCN} \rangle = 4.75$. There the decrease of $p(k)$ for high values is stronger, with an exponent of $\alpha_k^{HCN} = 4.89(26)$ for $k^{HCN} > 20$.

5.6 Economic Systemic Risk

With a reasonable reconstruction of the supply network, RSN, we turn to the quantification of economic systemic risk in the national production network in work with Tobias. For quantification we use the economic systemic risk index (ESRI) as developed in [DBR+21]; see the publication [RHD+22] for details.

We find 65 firms of high systemic risk to mainly belong to the manufacturing sector (NACE lvl. 1 category C, 77%), followed by companies in the electricity, gas stream and air conditioning supply (D, 8%) and financial and insurance activities (K, 6%) sectors.In contrast to the exact Hungarian production network [DBR+21], several companies from non-manufacturing sectors (NACE $\geq$ 45) are found in the plateau. This is somewhat unexpected since they are associated with linear production functions, which causes their shock spreading behavior to be less extreme than for Leontief producers.

5.7 Robustness of results

Our study is subject to several limitations, in particular (i) the imperfect overlap of the two communication and supply-link layers, limiting the possible accuracy, (ii) the limited market coverage of the phone provider (resulting in limited agreement even if $p(s|c) = 1$), and (iii) errors originating from the network reconstruction uncertainties in the estimations of directions and weights.

To estimate the biases and errors introduced by these weaknesses, we perform several simulation studies. First, we generate a synthetic communication network based on the HSN and the probabilities to find a communication link, where a supply-link is present $p(c|s)$, and where no supply-link is found $p(c|\neg s)$. From this synthetic communication network we then take a sample of nodes according to an estimated market share m of the data provider and calculate the induced subgraph comprised by links only between the sampled nodes.

Finally, following the procedure used on the empirical data, we reconstruct a supply network from this synthetic communication network and calculate the ESRI. We calculate Spearman's rank correlation coefficient, ρ, between the ESRI as calculated on the full, real HSN and on the reconstructed subgraph. After repeating these steps for 100 times with $m = 1/3$, $p(c|s) = 0.21$ and $p(c|\neg s) = 9.3 \times 10^{-5}$, we find an average Spearman correlation of $\langle \rho(ESRI_{HSN}, ESRI_{reconstr}) \rangle = 0.563(6)$. We find that the most relevant effect is caused by the limited market share with a drop of correlation of $\Delta\langle\rho\rangle = 0.31$, followed by the limited overlap, adding another, $\Delta\langle\rho\rangle = 0.13$. The effects from network reconstruction reduce the correlation by only $\Delta\langle\rho\rangle = 0.0004$, which is remarkably small. We calculate the probability that a node that is among the 0.1% riskiest nodes of the subsample is also among the riskiest 0.1% of *all* nodes and find 32.9(82)%. The probability that one of the top 0.1% of the subsample nodes is among the top 1% of the full network is 47.7(99)%.

5.8 Discussion

We show that mobile phone metadata can be used to reasonably reconstruct the flow of goods between firms in an economy, i.e., the supply network. The reconstruction is possible because of the similarity of the communication- and the supply layer of the inter-firm network. This method is one of the very few alternatives to obtain a comprehensive view on national supply network, when there is no VAT or payment system data.

Based on the supply network we calculate economic systemic risk and find that a small core of about 65 high systemic risk firms have the potential to affect large parts of the economic activity. Apart from these core firms systemic risk of companies is generally small. These results agree well with the previous results for Hungary, where a core of 32 high systemic risk firms was found to contribute to 45% of the overall systemic risk [DBR+21]. With a series of robustness checks we demonstrate the reliability of the results.

Using a large-scale survey on the actual customer-supplier relationships between companies, we find the probability of a supply link to exist, given an existing communication link as $p(s_{ij}|c_{ij}) \approx 0.19$. When thresholding for higher interaction strength of the communication relation $p(s_{ij}|c_{ij})$ the conditional probability increases strongly to 92%. Note that the survey asked for the firms' *most* critical suppliers. It is almost certain that in the FCN we observe connections to suppliers that are perhaps important but were not classified as essential in the survey, causing $p(s_{ij}|c_{ij})$ to be underestimated. Landline phones are still common practice in many firms; these communication links are not covered, thus further underestimating the overlap of communication and supply links.

We find that the degree exponents of the reconstructed supply network, $\alpha_k^{RSN} \approx 2.18$, and the exact Hungarian supply network, $\alpha_k^{HSN} \approx 2.40$, are similar; the degree exponent of the human-human communication network is much larger, $\alpha_k^{CN} \approx 4.89$. Also for the average nearest neighbor degree and the local clustering coefficient the topology of the RSN is more similar to the topology of the exact HSN than to the HCN.

We showed that the FCN and the HSN are most similar when thresholding communication strength to $d_{ij} > 30$s/d. We sample supplier directions using external information on companies' industry sectors and from input-output tables. Link weights are estimated by the product of firm sizes. Future improvement of the reconstruction method could be reached by using additional information contained in the FCN, such as asymmetries in the calling behavior, temporal patterns in the sequence of calls, as well as using dependencies of supply link weights on communication intensity.

The method has several limitations. We systematically investigate the error introduced by the imperfect overlap of the communication- and supply layers, the limited market share of the mobile phone provider, and the reconstruction of the link directions. In a simulation study we find an average rank correlation between the true ESRI in the HSN and the ESRI on a carefully simulated synthetic firm communication network of $\langle \rho(ESRI_{HSN}, ESRI_{reconstr}) \rangle = 0.563$. The limited market coverage and the imperfect

link overlaps contribute most of the effect. We expect $\langle \rho(ESRI_{HSN}, ESRI_{reconstr}) \rangle$ to be higher in reality since it is based on the estimate for $p(s|c)$ that is a lower bound. Further, despite the limited correlation, our method allows us to capture heterogeneity in shock spreading well and uncovers the localized effects of up- and downstream cascades on the firm level that traditional methods such as input-output models cannot describe.

There are also three limitations that could not be addressed explicitly. First, firms use many more communication channels than mobile phones such as landlines, e-mail or physical mail, and a growing number of new interaction channels, such as social media or online portals. Nevertheless, we assume that, if the supply relation is sufficiently strong, firms become more and more likely to use mobile phones to arrange spontaneous meetings, inform partners about delays, coordinate the quality, quantity and timing of deliveries, fix dates, provide support, etc.

Second, due to the anonymity of the telecommunication data it is not possible to perform targeted surveys on the customers of the phone provider. To reach significant overlap of the survey respondents and the customers of the phone provider, untargeted surveys need a response rate of a considerable fraction of firms within a country.

Third, another consequence of the anonymity of the data is that –by definition– firms cannot be identified and concrete policy statements can only be made on the level of the network. However, within the anonymity constraints, the effect of heterogeneous shocks in relation to economic sectors and geography can still be investigated. This is important since recent work has shown that heterogeneity in the initial economic shocks can cause dramatically different economic outcomes [IT19, DBR$^+$].

Since mobile phone data is easily available, the presented method to reconstruct a national production network is cheap, scalable, and easily implemented. It can be used for countries where no tax or survey data is available. The method also captures international links which allow us to identify economic exposures to specific countries. Maybe one of the most interesting features of the method is its temporal resolution, supply relations can be monitored in real-time. This offers the possibility to study how firm-ties form and rewire on the network-level. Monitoring the restructuring processes of the economy during natural disasters or economic crises are immediate areas of application and could become crucial for monitoring the progress of the green transition, where production networks have to transform such as to no longer produce greenhouse gases.

CHAPTER 6

Identifying the root cause of cable network problems with machine learning

Good quality network connectivity is ever more important. For hybrid fiber coaxial (HFC) networks, searching for upstream *high noise* in the past was cumbersome and time-consuming. Even with machine learning due to the heterogeneity of the network and its topological structure, the task remains challenging. We present the automation of a simple business rule (largest change of a specific value) and compare its performance with state-of-the-art machine-learning methods and conclude that the precision@1 can be improved by 2.3 times. As it is best when a fault does not occur in the first place, we secondly evaluate multiple approaches to forecast network faults, which would allow performing predictive maintenance on the network.

Hybrid fiber coaxial (HFC) networks deliver internet connectivity directly to end customers. Unfortunately, their reliability can be poor [BBF18, GPS+13]. The network contains separate channels for up (US) and downstream (DS) signals. The US signal of the HFC network refers to data that is transferred from the customer up to the central root node. In contrast, the DS part refers to the opposite direction of the signal, i.e., commonly used for downloads from the internet. In particular, for a problem related to the US channels, a fault usually affects only a single or limited group of customers. It relatively quickly spreads in the whole region of the network named fiber-node area. Therefore resolving such a problem fast and without disrupting connectivity further is essential. However, at the partnering internet service provider (ISP), the field technicians currently perform a binary search to identify the root cause of the problem by disconnecting certain amplifiers. This means that not only is a considerable amount of time spent searching for the device, which is the root cause of the incident, the search process itself temporarily disrupts the service for other customers.

The cable industry suggests using proactive network measurements (PNM) to diagnose problems. But the sheer volume of proactive alarms overwhelms the technicians as PNM data generically suggest areas of improvement and not the root cause of a specific incident.

Over time, implicit business knowledge has been built up to define a rule by the partnering ISP, but so far could not be executed automatically. We use it: Largest transmission power change before the incident – as a baseline when comparing our results. We demonstrate that by developing machine-learning enhanced models, precision can be improved over this baseline. This allows to 2.3 times better (measured by precision@1) direct the technicians and faster resolve high noise faults in the network. Such faults are sometimes referred to as common path distortion (CPD).

This problem is particularly interesting as normal behavior is different for each network region. The topological structure of the network as defined in Section 6.2.1should be included in the modeling approach.

In principle, it would be even better if a fault could be predicted before a field technician needs to be dispatched and customers observe degraded or unavailable service. Therefore, we develop a prediction pipeline for network faults to showcase the potential of predictive fault detection.

Our research question is (I) to evaluate whether machine learning enhanced models can steer technicians better to a given root cause of a high noise incident and (II) whether a future incident indicated by an overly high codeword error ratio can be predicted in advance.

6.1 State of the art

For a given high noise incident we use machine learning models to steer technicians to the root cause of the incident.

The scientific literature focuses on issues in the DS path of the signal [ZSR20, ZS17], identification of anomalies [ZSR20], prediction of hotline calls from incident tickets and telemetry [HZY+20, Eck21] spectral analysis of the telemetry data for fault detection [RW18, ZMLZ10], collection of better quality data [TH20] directly from the cable modems, generic network data analysis with neural networks [FHS+20]. Tool vendors in the industry offer software solutions for individual and manual spectral-analysis-based failure analysis for specific devices. However, too many warnings are created. Additionally, technicians are not guided to the root cause of an incident as these systems generate too much data to obtain detailed information for the whole network in real-time.

In addition to the US data used in [HZY+20] we furthermore utilize the DS PNM data in our study as features for the various models. The publications [Eck21, HZY+20] are trying to predict customer interactions on the hotline (based on generic faults), whereas we identify the actual root cause for any US high-noise-related incident automatically.

The authors of the tool CableMon [HZY+20] observe that they can predict approximately 80% of trouble tickets that would lead to a call. Eckert [Eck21] observes a similar result when using autoencoders.

However, here for the high noise root cause detection, we are in a different setting: Instead of only identifying an anomaly, we need to exactly pin-point the root cause of a given high-noise incident where often many cable modems start to act anomalously at almost the same time.

6.2 Problem description

In the following Section follows a description of the topological architecture of HFC networks as well as physical details of the problem.

6.2.1 HFC architecture

The HFC network resembles a tree-like hierarchy. An example is visualized in Figure 6.1. Often the network was built over a long period. Usually, some operators were bought and merged in this process. This contributes to further technical heterogeneity of the individual network segments (hubs). Hubs represent the physical structure of the network region. Commonly the devices in such a region were built together at the same time with the same technology and configuration. Interestingly, some regions in the country are worse than others. The root node of a hub named cable modem termination system (CMTS) contains several fiber-node areas which are connected using optic-fiber. Thus, these connections are highly reliable and in any case of failure, it is simple to identify the exact point of failure. The area of each fiber-node limits any signal interference. A fiber-node - typically using many line- and distribution amplifiers and potentially splitters - connects the *last mile* to the network. The last amplifier before a final consumer, i.e., the house, is named the last line amplifier. Based on coaxial copper cables, in particular, corrosion can badly influence the quality of the connections as parts of these networks are many decades old now [1].

PNM is recommended to improve fault resolution by the cable industry. Monitoring tools deployed in the industry can generate many proactive alarms. The sheer volume of proactive alarms can be overwhelming for the technicians. Therefore, even though included in the Data Over Cable Service Interface Specification (DOCSIS) standard since 2005 [Cab], dealing with PNM data remains a challenge as the recommendations for best practices and software deployed in the industry work with manually configured thresholds [Cab, WHTG]. These are often used statically and tailored to use cases such as general proactive network maintenance. Although the problems identified by PNM data indicate faulty network connections, they are not directly related to any specific customer disruption. As a result, these identified problem notifications might deliver too many findings to be handled for a specific incident. As there, the task is to identify the

[1] `https://calcable.org/learn/history-of-cable`, accessed 11th of September 2021

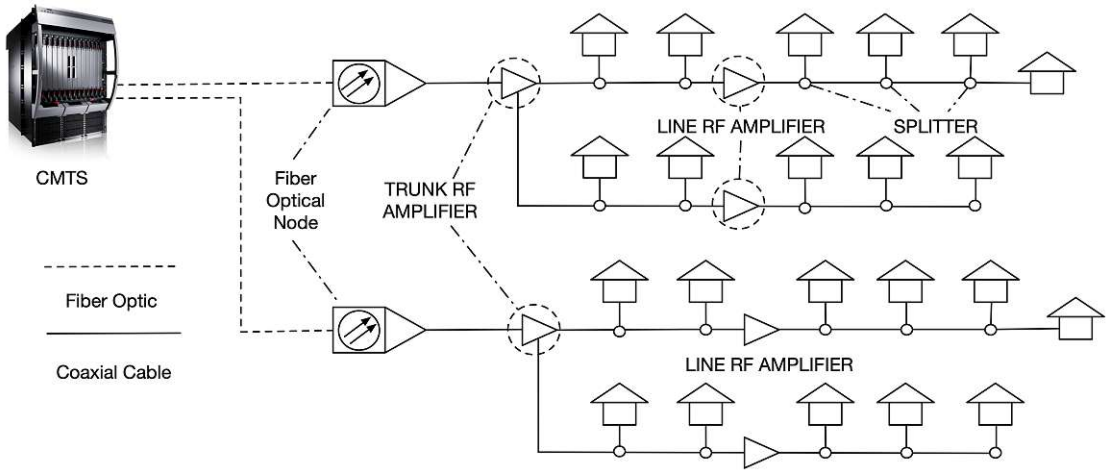


Figure 6.1: Overview of the architecture of an HFC network from [HZY+20].

root cause quickly, given the limited human resources of the technicians. Furthermore, the *proactive* PNM alarms of a HFC plant monitoring system do not resemble any kind of predictions for maintenance, rather only minor (=non-outage) faults, which could indicate the need for maintenance in the specific network elements if they frequently occur in a region of the network. In other words: PNM data can be helpful to proactively gradually improve the quality of the overall network, but this data does not offer the clue for a specific incident. In particular, PNM data does not outline which device is the root cause of any specific incident. However, this information would be needed to guide technicians when the fault resolution process should be improved. Accurate root cause indications have to be created with manual effort and this leads to the problem that this cannot be fulfilled with the available human resources of technicians.

6.2.2 Fault characteristics

High noise caused by CPD (Common Path Distortion) is an upstream distortion that is typically generated by corroded contact surfaces on a loosely tightened connector. An example is shown in Figure 6.2.x For the specific high noise problem characteristics, it is essential to understand that US channels (i.e. frequency bands) are shared. Therefore, a fault initially affecting only a single device on a specific frequency channel can quickly spread within the network region and in extreme cases destroy any connectivity in the whole fiber-node area. Unlike downstream faults, where tracing these to a common specific point for technicians to fix, the upstream channel becomes more complex in case of problems as many cable modems can depict anomalous behavior in such a scenario at almost the same time.

In a normally functioning US channel, each cable modem sends the signal to a common point at the top of the cable network (CMTS) without any disruption. The modems may not transmit on the same frequency at the same time. The CMTS uses the DOCSIS

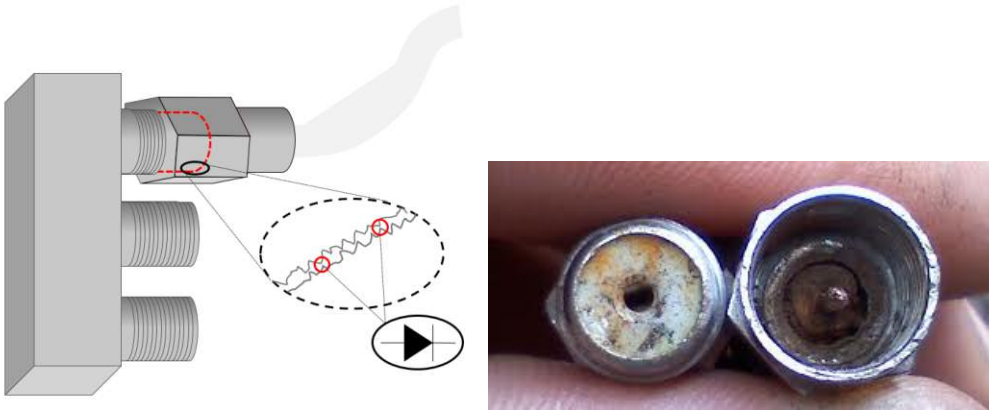


Figure 6.2: Damaged connector and loosely tied F-connector (partly due to corrosion): Upstream high noise is typically created by corroded contact surfaces or loosely tied connectors. Technicians of the ISP supplied these visualizations of corrosion.

protocol to control which modem is allowed to transmit at what time and frequency using the Time and Frequency Division Multiplexing (TaFDM) protocol. In case of disturbance on the US channel, any disturbance is transmitted onwards to the common point at the top of the network as coaxial cables are vulnerable to interference. Therefore, a single disturbance can negatively affect the DOCSIS signal for all modems in this fiber-node area or even make them unusable. The downstream signal contains many different frequency bands. These do mainly affect the US but also the downstream signal behind the corroded connector. In typical coaxial networks, this effect occurs at network points with a sufficiently high downstream signal where the US signal is relatively lower than the downstream. Thus, the disturbance affects the upstream more than the downstream. The term *High Noise* has become established as it forms a characteristic picture. An example is found in Figure 6.3. It materializes as a noise floor in the spectral domain due to the huge amount of frequency bands participating in the fault.

Technicians are faced with the challenge of not knowing which point in the network the disturbance is coming from. The current fault finding process is as follows: A technician has to go through the network and identify where the fault is coming from on the path through the network to the root cause by conducting a binary search. To make matters worse, the problem is often unstable and the technician cannot complete troubleshooting. Only when the source of the problem is found, the process of fixing the fault can be initiated. The main disadvantage is not only that technicians spend a lot of time troubleshooting, but that many customers are affected by the problem and that during the binary search procedure by the technicians to identify the root cause, additional customers might be affected.

In the following, we outline how the root-cause searching process can be improved by automating a simple rule-based classifier and utilizing machine learning enhanced methods. Secondly, we present a fault prediction method to prevent faults from happening in the

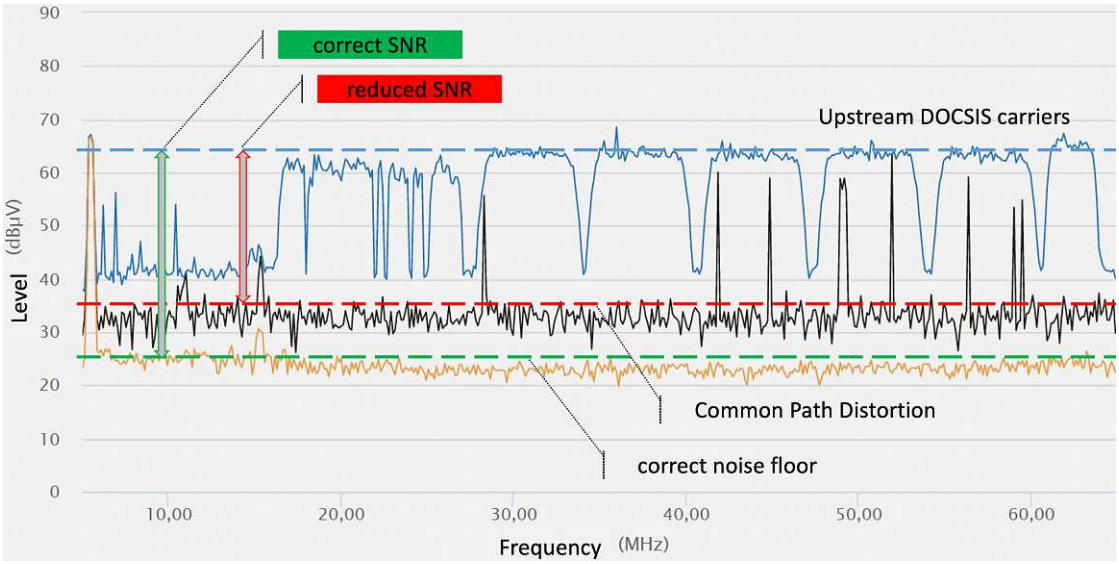


Figure 6.3: Only the upstream channel is visualized. A noise floor is created by the vast amount of frequency bands participating in the incident. The x-axis is the frequency of the signal and the y-axis is the signal level for each frequency. The yellow scenario (with the green marker) denotes a case with correct SNR, whereas the black scenario visualizes the noise floor (with the red marker) for reduced SNR. The blue line represents the basic DOCSIS user data frequencies (carriers).

first place ideally.

6.3 Dataset description

In this section the various data sources are detailed: From the raw telemetry data of the cable network monitoring tool to the generated alarms and the issue tickets that field force technicians are using when making any changes to the network or fixing a problem in the network.

6.3.1 Telemetry data

The telemetry information is collected on multiple levels: each cable modem reports data per each channel using simple network management protocol (SNMP) polling, but also the CMTS collects similar data. However, the cable-modem-based data might not be available in network outages for particular modems. The following are stored, in the raw form for each channel of each modem (MAC address), separately for both up- and downstream, with hourly resolution including a timestamp: signal to noise ratio (SNR), a cable modem transmission power (Tx power), the received signal power (Rx power), codeword error ratio (CER) and corrected CER (CCER). For the downstream, additional

micro reflections (m-reflection, impedance mismatch on the cable affecting the signal) are available.

6.3.2 Alarms

The network operating center (NOC) stores alarming events for each device in Elasticsearch. For each device, the start and end of the alarm are noted.

6.3.3 Truckroll-tickets

contain a free form text field for the notes of the technician, category of the incident, processing time and a free form text field for the amplifier causing the incident.

We developed a parsing logic here to extract the amplifier(s) which were identified as the root cause by the technician. Due to inconsistent naming of the amplifiers in different network regions and the process of parsing a free form text field, unfortunately, we are not able to utilize all tickets. The tickets are filtered to contain high-noise-relevant tickets already only.

6.3.4 Topology

Geospatial coordinates (location) for each amplifier as well as the path between the various amplifiers to the fiber-node.

The ground truth labels denote a root cause at a specific topology level. We decided to only accept accurate root cause identifications as valid labels, which denote an individual amplifier (on the lowest level) as the root cause. As the telemetry data is initially provided on the level of the fine-grained frequency bands where many belong to an individual amplifier, we decided to aggregate the data to the topological level of the last line amplifier. Due to the sheer size of telemetry data for the whole country of the ISP we choose to use Apache Spark (version 3.1.2) [ZCDD12] to perform the aggregation. During this aggregation process, the anonymity of the subscribers is ensured and we only ever receive anonymized data for our study. Here, after linearly interpolating missing data for each device, we compute descriptive statistics (mean, std, min, max), change ratio (current/previous) and relative changes ((current-previous)/previous) for each feature. Additionally, we consider a sliding window of 4 hours and calculate the change there as the difference between the largest and smallest value in each window instead of the difference between the current and previous observation. This data aggregation process is depicted in Figure 6.4.

We are evaluating a total of approximately five months of data (2021-02-25 – 2021-07-25). After the aggregation process, we consider 26069 last line amplifiers in the dataset, where some participate in multiple incidents.

An incident can become more severe (as more devices are affected). We need to aggregate the individual device-level alarms to the whole fiber-node as high-noise-related incidents

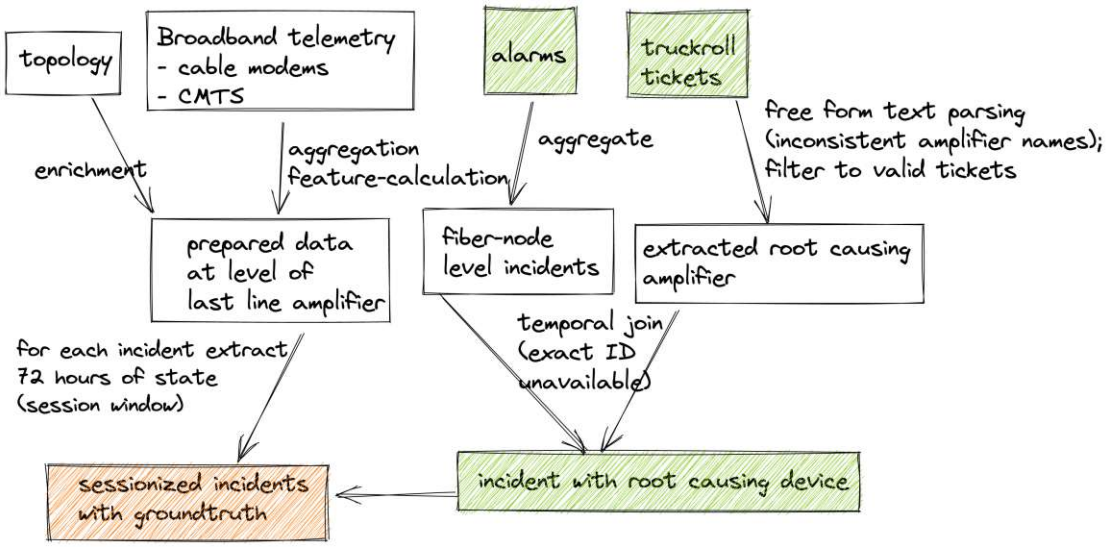


Figure 6.4: Data pipeline overview. For each labeled incident, a dynamic state (session window) of 72 hours before each incident is obtained.
Using the available data, alarms and truck-roll tickets are correlated using a temporal join.

often affect many devices. We need to have the global beginning and end of the incident. The global alarm time window is then used for a temporal join with the truck roll tickets as, unfortunately, no direct link between an incident, incident ticket, truck roll-ticket and the corresponding telemetry data was established before. Furthermore, only high-noise-related alarms are filtered for this specific use case.

With the alarms and parsed truck-roll tickets we can obtain ground truth labels for each incident. For each incident, we obtain a session window where one or more last line amplifier is marked with the label denoting a root cause for this particular incident. Figure 6.5 depicts an example case of the classical Tx spikes before a high noise incident that matches the positive class label.

As we need to identify the root cause for a specific incident, we can only keep incidents where a label is available in our dataset. 796 root cause amplifiers remain labeled from the ground truth data from 7 network regions for 457 unique fiber-node areas and 672 truckroll tickets. This means that for some tickets ≥ 1 offending (= root-causing amplifier) are suggested in the ground truth data. In total, we obtain 796 positively labeled amplifiers out of a total of 26069 for an amplifier identified as the root cause of a high noise incident. The remaining data are kept as negative examples. This makes the dataset highly unbalanced with regard to the target labels.

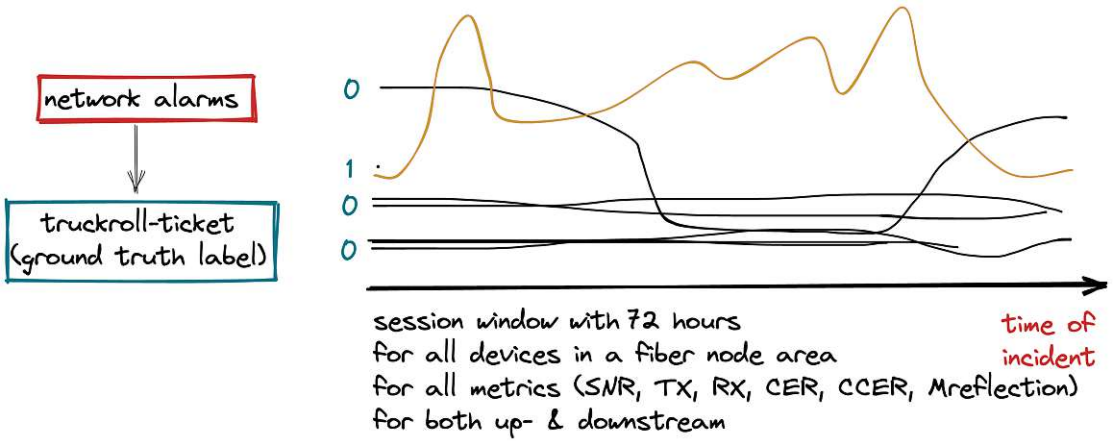


Figure 6.5: Given network alarms (in red) and truck-roll tickets we perform a multidimensional sessionization on the telemetry data. 72 hours before an incident are kept as training data. The root cause label (defined by the truck-roll ticket as ground truth in blue) is used to identify the offending amplifier. The y-axis contains the various amplifiers participating in the schematic incident session window. The line of the amplifier causing the incident is highlighted in brown.

6.4 Data Preprocessing

The network as described in Section 6.2.1 contains two crucial levels in the network's topology: hubs and fiber-nodes.

For most statistical models, numeric distances between the features are important. As the quality, if the network is different in each region differs, we must normalize the data in a way that we can learn from and compare all incidents taking both the physical effects and error boundaries mentioned above into account, as a feature that is considered high or anomalous in one region might be completely normal for another one. We propose to double normalize the data: As discussed in Section 6.2.1, the devices in a hub have similar physical properties, which can be handled using simple standardization (0 mean, unit variance). When introducing the HFC topology we already explained that any error is limited to the extent of each fiber-node area, Section 6.2.1. To make amplifiers comparable across fiber nodes, we need to standardize again, now taking time into account and do so for each of the 72 hours of session window for each feature, but standardize only within the amplifiers related to this particular incident.

6.5 Models

Starting with a simple business rule as our baseline we compare various state-of-the-art (SOTA) ML/AI-enhanced approaches.

6.5.1 Baseline: business rule

Decades of knowledge of the technicians define a very simple business rule as follows: Shortly before the incident, the largest upstream Tx change identifies the root causing amplifier.

This rule has a significant advantage: It is dynamically adapting to the specific situation of each incident due to choosing the largest change. Given two very different network regions with different physical properties or quality, the largest change is still a fairly reliable indicator for the amplifier causing the incident. Furthermore, this simple rule is well understood by the technicians. In case fine-tuning is required, they can easily adjust the cutoff parameters for themselves and instead of analyzing the top-1 (largest) change they could consider the top-n.

As we will see later, when evaluating statistical machine-learning models, such dynamics which is specific for each incident needs to be explicitly considered there as well during evaluation.

6.5.2 Subgroup discovery

Using explainable models can increase the trust of the non-tech business stakeholders as they can easily understand the inner workings. Singh et. al. [SNT+21] provide a package with implemented models that might be able to replace black-box models with simpler ones while improving efficiency and interpretability without sacrificing accuracy.

6.5.3 ML models

Logistic regression: A simple statistical baseline using a standard logistic regression procedure, it is implemented in scikit-learn [PVG+11]. *Lightgbm* [KMF+17] is one example for gradient boosted tree models which generally deliver good model fitting performance and is fast to train. Unlike neural networks, it does not require extensive fine-tuning. Compared to basic decision trees many trees are trained to improve the overall model. However, unlike random forests where random trees are used to stabilize and improve the results, gradient boosted trees are build that the next tree always best reduces the loss function (error) of the so far existing trees.

We compare various neural network-based approaches as well. These are based on tsai [Ogu20] as an implementation of various state-of-the-art time-series oriented architectures based on fast.ai [HG20]. We use the models of tsai for our dataset and in particular, adapt the data loaders for the sessionization and normalization as outlined above. For any of the neural network models, we use the learning rate finder[2] provided by the fast.ai library to balance the speed of training and accuracy of the models whilst still improving the performance of the models as there is a smaller chance being stuck in local optima. *LSTM* long-short-term-memory is a traditional neural network architecture for time series

[2] https://sgugger.github.io/how-do-you-find-a-good-learning-rate.html

handling [HS97] *InceptionTime*, is a recent SOTA architecture for time series [ILF+20]. *TST* BERT [DCLT18] and transformers revolutionized the field of sequence-based neural networks. Only recently first adaptations of these models for temporal tasks have been developed. The time series transformer (TST) [ZJP+21] is one such example. It is based on [VSP+17, HRKA21] the domain of information retrieval.

Both text- and image-based domains were revolutionized when pre-trained models could be used. This drastically decreased the required compute resources and datasets. For the time-series domain, the classical pre-trained models cannot be used as they stem from a completely different domain. Instead, we follow a self-supervised pre-training[3] approach by first training a BERT based model in unsupervised mode to create network embeddings for our LSTM core; secondly, we use this pre-trained model in three scenarios: *LSTM self supervised (fine-tuning)*, *LSTM self supervised (training)*, and *LSTM self supervised (train) + data augmentation* training with the CutMix [YHO+19] data augmentation strategy.

The hyperparameters of the models were optimized using Optuna [ASY+19] on a GPU-equipped server.

6.5.4 (Ranked) evaluation of results

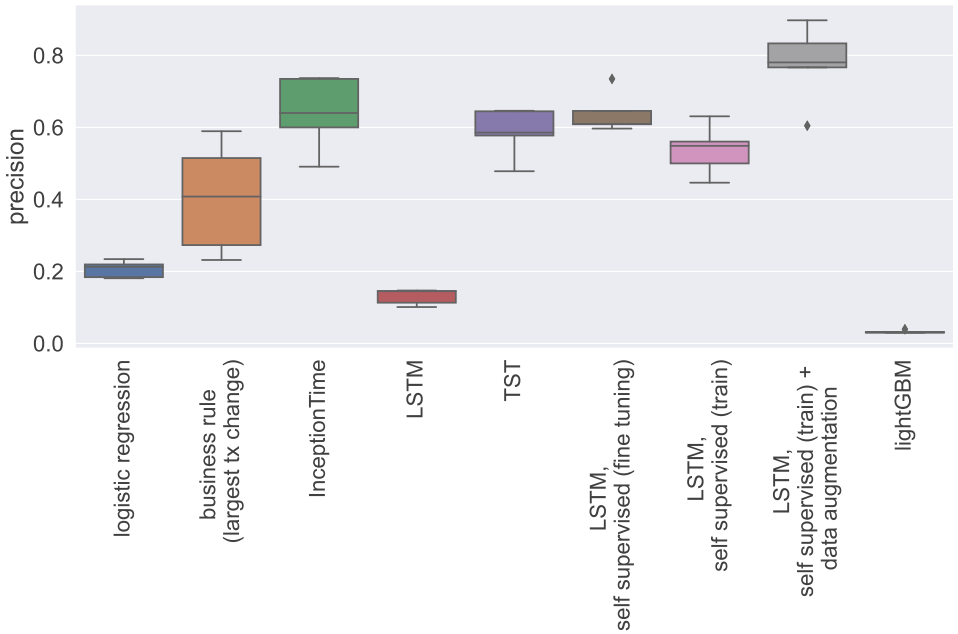
When evaluating the models we do *not only* perform a classical binary classification evaluation, where for one particular observation a probability is emitted. Rather, we classify a single incident session globally by obtaining the predictions of the model if any amplifier is a root cause for the incident and then rank these predictions. The ranked evaluation takes place in two stages: Firstly, the binary classification is performed by the various models. Secondly, the output probabilities are ranked and a top-k evaluation is performed. This is a deliberate decision as it enhances each of the models with the dynamics of the particular incident and network region as mentioned in Section 6.5.1 we could not account for otherwise. Empirically this proves to work well for all models as we are in a ranking a task where the most probable root cause for each incident needs to be identified when analyzing the precision@k, see Table 6.1.

6.6 Results

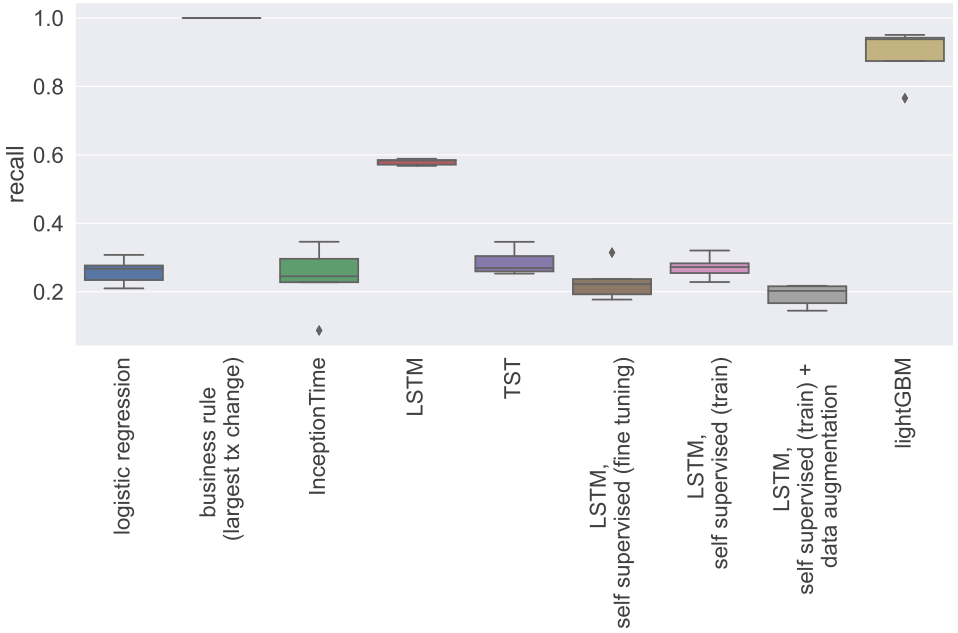
The *discovered subgroups* can be used to create rules which are easily understandable. Interestingly, these statistically discovered rules align well with the business practice of the field technicians. Showcasing the technicians that we can use these to derive their business rule increased trust in our other modeling activities.

The results of the first raw (binary classification) evaluation are depicted in Figure 6.6. The logistic regression is worse with regards to both precision and recall compared to

[3] https://github.com/timeseriesAI/tsai/blob/main/tutorial_nbs/08_Self_Supervised_TSBERT.ipynb [Ogu20]



(a) Precision



(b) Recall

Figure 6.6: Precision and Recall for the raw model outputs of the first binary classification stage for each cross-validation fold.

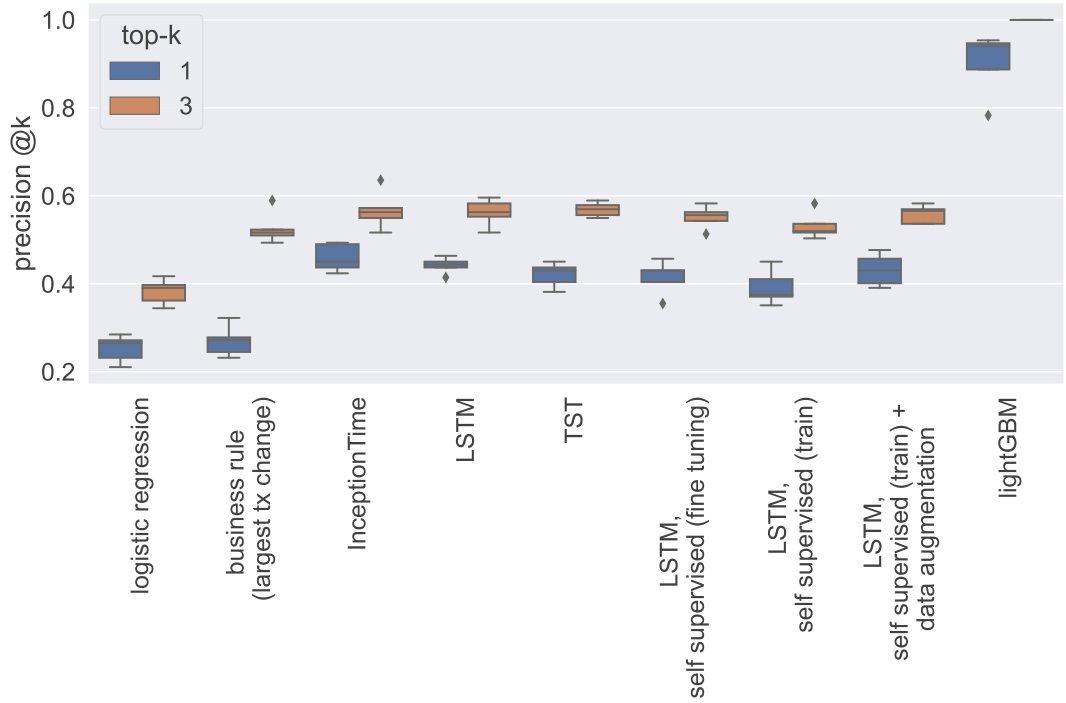


Figure 6.7: Comparison of models using precision@k (ranked). Any of the more complex models deliver better results than the simple business rule for both top-1 and top-3 evaluation. LightGBM in particular performs best with a wide margin. Only a naive logistic regression is worse than the business rule.

the business rule. Most of the other models (except LightGBM and LSTM) result in higher precision. With regards to recall, none of the other models is better than the business rule. However, the business rule baseline can achieve the high recall only with very limited precision. In any real-world scenario when deployed at an ISP the human resources of the technicians are limited to evaluate false alarms, therefore a high precision is more important than recall as the technicians otherwise might lose trust in the technical solution.

Furthermore, when reframing the task into a ranking task, where the most probable root cause is identified, the superiority of the ML enhanced models clearly becomes visible. The models are evaluated for a top-1 and any within the top-3 match. However, for sake of brevity and increased precision, we only discuss the top-1 match when comparing the results, as this is the variant that would most likely be used by an ISP to minimize the workload overhead of the technicians induced by faulty recommendations. Figure 6.7 depicts both cases for completeness. Detailed results for precision, recall and the precision@rank-k are listed in Table 6.1. The simple *business rule* (largest Tx change

top-k	model	precision step 1 mean	precision step 1 std	recall step 1 mean	recall step 1 std	precision @k mean	precision @k std	false positives@k mean	false positives@k std	true positives@k mean	true positives@k std
1	lightGBM	0.032	0.004	0.894	0.078	0.902	0.072	14.8	10.918	136.4	10.502
	InceptionTime	0.640	0.103	0.241	0.098	0.459	0.031	81.8	4.604	69.4	4.879
	LSTM	0.130	0.022	0.579	0.009	0.442	0.018	84.4	2.966	66.8	2.588
	LSTM, self supervised (train) + data augmentation	0.777	0.109	0.189	0.032	0.431	0.036	86.0	5.612	65.2	5.404
	TST	0.586	0.068	0.286	0.039	0.421	0.028	87.6	4.393	63.6	4.037
	LSTM, self supervised (fine tuning)	0.639	0.056	0.229	0.053	0.415	0.038	88.4	6.066	62.8	5.675
	LSTM, self supervised (train)	0.537	0.069	0.272	0.034	0.392	0.039	92.0	6.000	59.2	5.891
	business rule (largest tx change)	0.270	0.035	1.000	0.000	0.270	0.035	110.4	5.030	40.8	5.404
	logistic regression	0.206	0.023	0.259	0.038	0.253	0.031	113.0	4.899	38.2	4.550
3	lightGBM	0.032	0.004	0.894	0.078	1.000	0.000	0.0	0.000	151.2	0.447
	TST	0.586	0.068	0.286	0.039	0.569	0.016	65.2	2.387	86.0	2.550
	InceptionTime	0.640	0.103	0.241	0.098	0.567	0.044	65.4	6.580	85.8	6.611
	LSTM	0.130	0.022	0.579	0.009	0.562	0.031	66.2	4.658	85.0	4.583
	LSTM, self supervised (train) + data augmentation	0.777	0.109	0.189	0.032	0.558	0.021	66.8	3.114	84.4	3.209
	LSTM, self supervised (fine tuning)	0.639	0.056	0.229	0.053	0.552	0.026	67.8	4.087	83.4	3.715
	LSTM, self supervised (train)	0.537	0.069	0.272	0.034	0.532	0.031	70.8	4.712	80.4	4.615
	business rule (largest tx change)	0.526	0.037	1.000	0.000	0.526	0.037	71.6	5.683	79.6	5.459
	logistic regression	0.206	0.023	0.259	0.038	0.382	0.029	93.4	4.506	57.8	4.324

Table 6.1: Summary statistics (mean, std) for the results of the various root cause analysis models. Notice: The counts are aggregated high noise incidents. Each incident contains a varying but high number of underlying amplifiers.

before the incident) results in a top-1 precision (on average for the cross-validation folds) of 0.27. Any of the other more complex models deliver better results: In particular the *tree-based* model LightGBM results in a top-1 precision of 0.902. The various neural-network-based approaches differ in their precision only marginally (0.392 – 0.45) and are additionally worse and more complex to interpret and computationally expensive to train than LightGBM. Interestingly, LightGBM outperforms all the neural network-based approaches in our comparison. Most likely the reason for this is that the amount and quality of the training data is limited so far:

- *quantity*: we were only able to obtain ground truth labels for a limited area in the network see Section 6.3.4 due to free form text field parsing
- *quality*: due to a missing id field in the various data sources connecting the alarm and field-force ticket to an incident we need to perform a temporal correlation.

LightGBM has the advantage that the model training procedure is swift and allows for more experimentation with regards to hyperparamters. Especially in an industrial context where often AI is only an enhancing part of the overall process the optimization of the hyperparameters can be performed very fast.

6.7 Discussion

The machine learning aspects are only part of a bigger use case. Hence, it is important to understand the requirements of the ISP well in case it should be deployed in a scalable real-time setting with integration into an existing processing landscape. Indeed, the simple automation of the presented business rule will have the advantage of being most easy to

get started with and transparent to the HFC technicians maintaining and operating the network. However, as we have shown any of the other more complex models outperforms this simplistic business rule by a wide margin: The best one, LightGBM, is more than 2.3 times better than the baseline.

To benefit most from these ideas, the ISP should further consider creating a heat-map of the identified root cause devices over a more extended period of time. Thus if repeatedly problems are identified in an area, technicians can be dispatched there, perform maintenance and improve the overall quality of the network (not related to any specific incident). This can be especially useful in case of hard-to-reproduce (flaky) problems.

The suggested approach can be improved by making more and better quality data available:

- generating more training data: Creating a structured ticket reporting instead of free form text field parsing.
- collecting better quality labels: Currently, a temporal join is required to link the telemetry data with the alarms and incident tickets. Instead one stringent equi-join (id-based) identification of incident and label would further enhance the quality of the labels available to the ML pipelines

In the long run, upgrading the infrastructure as suggested by [TH20] to result in better measurements for individual cable modems leads to better data, but considerable market invest and time are required for such a change. Perhaps as an intermediate step upgrading the monitoring tool could be a more viable option: Other ISPs[4] use monitoring tools that intrinsically collect more data, which could further enhance the results if available.

Like in regular ethernet networks, [DS16] HFC could be inspired by some of the recent advances in fault localization methodologies in internet-protocol (IP) based networks.

Enhancing the fault-finding process with machine-learning enhanced models can improve the time to resolution as technicians do not need to follow a lengthy fault-finding process: The best model LightGBM, improves precision@k more than 2.3 times over the baseline for a $k$vof 1.

Furthermore, we have shown that predicting faults at future time steps of the network can be helpful to prevent failures in the network before they show customer impact.

[4] https://medium.com/tele2techblog/great-insights-in-hfc-networks-from-pre-equalisation-data-6b8cab2c1dab

CHAPTER 7

Conclusion

In this thesis large scale datasets with a graph structure are analyzed to better understand the real world and their abstract representation, such as behavioral changes in society or technology. We demonstrate how such datasets can be used to solve societal problems.

7.1 Research question answers

We are now able to answer the main research questions:

- RQ1: *How large is the impact of the NPI on mobility and calls? To what extent can differences be observed in groups of the society formed by age and gender?* We find that that for both genders we observe an increase of total call duration, that the recovery time to women's total call time initially is as fast as for men, but later, clearly slows down. The decrease in mobility following the lock-down is stronger for women. In addition, men recover their mobility behavior much more quickly after the measures are lifted. By analyzing relative changes using compositional data analysis methodologies formerly hidden insights can be identified. We see that certain age-groups of the population (elderly, young during weekends) do restrict mobility less than other members of the population.
- RQ2: *How well can supply networks be reconstructed from mobile-phone data?* We are able to construct supply networks from phone data. We find the conditional probability of observing a supply-link, given a communication-link exists, to be about 90% and hope to set the stage for a new area of research utilizing such datasets for this purpose and extending upon our work.
- RQ3: *How well can root cause analyses utilizing machine learning methods improve over traditional rule-based approaches?* We present the automation of a simple business rule (largest change of a specific value) and compare its performance with

state-of-the-art machine-learning methods and conclude that the precision@1 can be improved by 2.3 times. As it is best when a fault does not occur in the first place we secondly evaluate multiple approaches to forecast network faults, which would allow performing predictive maintenance on the network.

7.2 Impact

Due to the global interest in researching the pandemic we participated in several collaboration projects. Due to the inherent relevance for the public publications were featured on blog posts, press releases[1] and also national television.

7.2.1 Impact in other fields

Many contributions together with various collaboration partners in different fields were created. The most notable ones are listed below:

- *Varieties of mobility measures: Comparing survey and mobile phone data during the COVID-19 pandemic [KSH21]* Measures to reduce individual mobility are prime governmental non-pharmaceutical interventions to curb infection rates during a pandemic. To evaluate the effectiveness of these efforts, scientific research relies on a variety of mobility measures that commonly stem from two main data sources: survey-self-reports and behavioral mobility data from mobile phones. However, little is known about how mobility from survey self-reports relates to popular mobility estimates using GSM and GPS data. Spanning March 2020 until April 2021, this study compares self-reported mobility from a panel survey in Austria to aggregated mobility estimates utilizing (i) GSM data and (ii) Google's Community Mobility Reports. Our analyses show that correlations in mobility changes over time are high, both in general and when comparing different subgroups. Differences emerge if subgroup differences are compared between mobility estimates. Overall, our findings suggest that these mobility measures manage to capture similar latent variables, but researchers should be aware of the specific form of mobility different data sources measure.

- *Meteorological factors and non-pharmaceutical interventions explain local differences in the spread of SARS-CoV-2 in Austria [LKC+21]* The drivers behind regional differences of SARS-CoV-2 spread on finer spatio-temporal scales are yet to be fully understood. Here we develop a data-driven modelling approach based on an age-structured compartmental model that compares 116 Austrian regions to a suitably chosen control set of regions to explain variations in local transmission rates through a combination of meteorological factors, non-pharmaceutical interventions and mobility. We find that more than 60% of the observed regional variations can be explained by these factors. Decreasing temperature and humidity, increasing

[1] https://csh.ac.at/covid19

cloudiness, precipitation and the absence of mitigation measures for public events are the strongest drivers for increased virus transmission, leading in combination to a doubling of the transmission rates compared to regions with more favorable weather. We conjecture that regions with little mitigation measures for large events that experience shifts toward unfavorable weather conditions are particularly predisposed as nucleation points for the next seasonal SARS-CoV-2 waves.

- *National-scale surveillance of emerging SARS-CoV-2 variants in wastewater [AME+22]* SARS-CoV-2 surveillance is crucial to identify variants with altered epidemiological properties. Wastewater-based epidemiology (WBE) provides a complementary approach to sequencing individual cases. Yet, national WBE sequencing programs have not been widely implemented and data analyses remain challenging. The collaboration partners deep-sequenced 3.413 wastewater samples representing 94 municipal catchments, covering >59% of Austria's population, from December 2020 to February 2022. Our Variant Quantification in Sewage pipeline designed for Robustness (VaQuERo) enabled us to deduce abundance of predefined variants from complex wastewater samples and delineate the spatiotemporal dynamics of circulating variants, including the variants of concern Alpha, Beta, Delta and Omicron. These results were cross validated by epidemiological records of >311,000 individual cases, demonstrating the potential to apprehend the composition of circulating variants from WBE. Furthermore, the partners describe elevated viral diversity during the Delta variant dominated period, in contrast to the Alpha and Omicron variants. Finally, the partners provide a framework to predict emerging variants de novo and measure the reproductive advantage of variants of concern by calculating variant-specific reproduction numbers from wastewater. Together, this study demonstrates the power of national-scale WBE to support public health and promises particular value for countries without extensive individual monitoring.

- *Data Anonymization – The key to innovation [HC22]* To open up opportunities for innovation and to overcome the complex data protection hurdles that often prevail, entrepreneurs are increasingly relying on data anonymization. We answer legal and technical questions which arise about the effectiveness of the anonymization procedure and suggest a probabilistic k-anonymizaiton procedure for anonymizing high-dimensional data whilst retaining a high degree of detail.

7.2.2 Television and other media

In the introduction we stated that by analyzing data about society, research can have an impact in the real world. Our COVID mobility analyses were featured multiple times on national broadcast television as well as national and international news coverage. We hope and also are confident that such feedback loops provide value to society. In particular, the following press announcements and policy briefs were covered in the media:

- *Regionalized low incidence strategy*[2] By introducing exit tests for regions of high incidence we observe a reduction of the growth factor by 6% and for regions close by for 3%.

- *Do lockdowns wear off?*[3] During the first lockdown in Austria, beginning on March 15, 2020 streets were empty. On average, people moved 70% less than usual, according to anonymized mobile phone data. During the summer, mobility was still 20% down compared to 2019. The second nationwide lockdown, which started on Nov 17, 2020 still reduced mobility by around 60%. In the midst of the third hard lockdown—beginning at Christmas and reinforced by the mandatory use of FFP2 masks and two-meter distancing (replacing the 1 meter "baby elephant") people move around almost as much as in the weeks before lockdown 3 started.

- *How did the 4th Lockdown affect mobility in Austria?* [4] It becomes apparent that with the beginning of the 4th lockdown mobility decreased at a much steeper rate than during the weeks before. However, the mobility reduction has not yet reached a level as low as mobility during the 2nd lockdown in 2020. It remains to be seen how the situation will develop.

7.3 Future work

Future topics which could be investigated in more detail, were partly covered or could be extended upon:

- A potential future study could analyze co-movement patterns, identify human contacts and subsequently, determine the effect of (non) social distancing [HLW$^+$20]

- *Geo-spatially correlated supply chain shocks* Natural disasters, such as large floods, are predicted to become more likely in the next decades as a consequence of rising temperatures due to global climate change. Therefore, it becomes increasingly important to improve the assessment of the economic implications of such disasters. We propose approaching this challenge by combining a model for the localized effects of a flood, the spatial footprints of companies, and a model for shock spreading on the firms' supply network. The flood model will provide an area that would be flooded in the event of a given magnitude (e.g. 100-year event). If this area overlaps with the footprint of a company, we can assume that the company will receive an economic shock.

[2] https://www.csh.ac.at/wp-content/uploads/2021/04/2021-April-CSH-Policy-Brief-Ausreisetests.pdf

[3] https://www.csh.ac.at/lockdowns-and-mobility-in-austria/, https://www.csh.ac.at/wp-content/uploads/2021/01/2021-01-25-CSH-Policy-Brief-BewegungsradiusUpdate.pdf

[4] https://www.csh.ac.at/lockdown-for-unvaccinated-mobility-in-austria/, https://www.csh.ac.at/wp-content/uploads/2021/11/2021-11-26-CSH-Policy-Brief-Mobilitat-Herbst-2021-final.pdf

- With increasing complexity of the COVID-19 crisis in 2022 due to more complex containment policies, test concepts, vaccinations, virus variants and immunity waning, agent-based epidemic models slowly start to outperform their macroscopic compartmental counterparts. Although highly limited by long computation times, these models could have the potential of correctly depicting the current epidemiological situation. Nevertheless, their application is not without controversy since their validity strongly depends on the model for the geo-spatial contact process. Showing that the latter is correctly depicted is a highly challenging task. One could pursue a highly flexible approach for modelling contacts in epidemiological agent-based models based on contact locations and origin-destination matrices. Using the OD matrices we have calculated in this thesis a future publication could derive the speed of propagation of the virus and mutations as well as the calibration of an agent-based simulation of COVID like [BRU+20] by correlating the mobility flows with sequenced wastewater analyses.

z

List of Figures

List of Tables

Acronyms

BTS Base transciever station. 8

CDR Call Data Record. 13

CLR centered log-ratio. 40–45, 47–50, 87

EU European Union. 2

GPS Global Positioning System. 14, 23

GSM Global System for Mobile Communication. 8, 12, 13, 24, 25

IMEI international mobile equpipment identifier. 3

IMSI international mobile subscriber identifier. 3

ISP internet service provider. 2, 3, 6, 8, 12–14, 50

LAC Location area code. 8

NPI non pharmaceutical interventions. 1, 5, 6, 23, 81

OD origin destination. 8–12, 85

PCA principal component analysis. 41

POI point of interest. 5, 8, 10, 26, 50

ROG radius of gyration. 5, 11, 26–30, 32, 39–50, 86, 87

UK United Kingdom. 2

VoLTE Voice over LTE. 13

XDR X Data Record. 13

Bibliography

[AB11] Tobias Adrian and Markus K Brunnermeier. CoVaR. Technical report, National Bureau of Economic Research, 2011.

[AG02] John Aitchison and Michael Greenacre. Biplots for compositional data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 51(4):375–392, 2002.

[AGSCM20] Berta Ausín, Clara González-Sanguino, Miguel Ángel Castellanos, and Manuel Muñoz. Gender-related differences in the psychological impact of confinement as a consequence of covid-19 in spain. *Journal of Gender Studies*, pages 1–10, 2020.

[Ahe11] Shohin Aheleroff. Customer segmentation for a mobile telecommunications company based on service usage behavior. *Proceedings - 3rd International Conference on Data Mining and Intelligent Information Technology Applications, ICMIA 2011*, (March):308–313, 2011.

[Ait86] J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, London. (Reprinted in 2003 with additional material by The Blackburn Press), 1986.

[AKU19] Hidayet Aksu, Ibrahim Korpeoglu, and Özgür Ulusoy. An analysis of social networks based on tera-scale telecommunication datasets. *IEEE Transactions on Emerging Topics in Computing*, 7(2):349–360, 2019.

[ALS18] Talayeh Aledavood, Sune Lehmann, and Jari Saramäki. Social network differences of chronotypes identified from mobile phone data. *EPJ Data Science*, 7(1), 2018.

[ALVC19] Xavier Andrade, Fabricio Layedra, Carmen Vaca, and Eduardo Cruz. Risc: Quantifying change after natural disasters to estimate infrastructure damage with mobile phone data. In *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pages 3383–3391, 2019.

[AME+22] Fabian Amman, Rudolf Markt, Lukas Endler, Sebastian Hupfauf, Benedikt Agerer, Anna Schedl, Lukas Richter, Melanie Zechmeister, Martin Bicher, Georg Heiler, Petr Triska, Matthew Thornton, Thomas Penz, Martin Senekowitsch, Jan Laine, Zsofia Keszei, Beatrice Daleiden, Martin Steinlechner, Harald Niederstätter, Christoph Scheffknecht, Gunther Vogl, Günther Weichlinger, Andreas Wagner, Katarzyna Slipko, Amandine Masseron, Elena Radu, Franz Allerberger, Niki Popper, Christoph Bock, Daniela Schmid, Herbert Oberacher, Norbert Kreuzinger, Heribert Insam, and Andreas Bergthaler. National-scale surveillance of emerging sars-cov-2 variants in wastewater. *medRxiv*, 2022.

[AMRD19] Nour Raeef Al-Molhem, Yasser Rahal, and Mustapha Dakkak. Social network analysis in telecom data. *Journal of Big Data*, 6(1), dec 2019.

[And11] James E. Anderson. The Gravity Model. *Annual Review of Economics*, 3(1):133–160, 2011.

[APPR17] Viral V Acharya, Lasse H Pedersen, Thomas Philippon, and Matthew Richardson. Measuring Systemic Risk. *The Review of Financial Studies*, 30(1):2–47, 2017.

[Arb20] Arbeitsmarktservice Austria. Arbeitsmarktdaten - Berichte und Auswertungen. https://www.ams.at/arbeitsmarktdaten-und-medien/arbeitsmarkt-daten-und-arbeitsmarkt-forschung/berichte-und-auswertungen, 2020. Accessed 8 October 2020.

[Art21] W Brian Arthur. Foundations of complexity economics. *Nature Reviews Physics*, 3(2):136–145, 2021.

[Asf20] Asfinag. Asfinag verkehrszählung. https://www.asfinag.at/verkehr/verkehrszaehlung/. Online, sep 2020. Accessed 8 October 2020.

[ASY+19] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631, 2019.

[Aus19] Statistik Austria. Haushalte, familien und lebensformen. http://www.statistik.at/web_de/statistiken/menschen_und_gesellscha haushalte_familien_lebensformen/index.html, 2019. Accessed 8 October 2020.

[BBF18] Zachary Bischof, Fabian Bustamante, and Nick Feamster. Characterizing and improving the reliability of broadband internet access. *SSRN Electronic Journal*, 2018.

[BDK15] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):10, 2015.

[BEST04] Michael Boss, Helmut Elsinger, Martin Summer, and Stefan Thurner 4. Network topology of the interbank market. *Quantitative Finance*, 4(6):677–684, 2004.

[BGĆC16] Sanja Brdar, Katarina Gavrić, Dubravko Ćulibrk, and Vladimir Crnojević. Unveiling spatial epidemiology of hiv with mobile phone data. *Scientific Reports*, 6(1):1–13, 2016.

[BJI20] Caroline Bradbury-Jones and Louise Isham. The pandemic paradox: The consequences of covid-19 on domestic violence. *Journal of Clinical Nursing*, 29(13-14):2047–2049, 2020.

[BKG+19] Danya Bachir, Ghazaleh Khodabandelou, Vincent Gauthier, Mounim El Yacoubi, and Jakob Puchinger. Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies*, 101:254–275, 2019.

[BL18] Alexandra Brintrup and Anna Ledwoch. Supply network science: Emergence of a new perspective on a classical field. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(3):033120, 2018.

[BMS99] James P Byrnes, David C Miller, and William D Schafer. Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, 125(3):367, 1999.

[BOS89] Michael Betz, Lenahan O'Connell, and Jon M. Shepard. Gender differences in proclivity for unethical behavior. *Journal of Business Ethics*, 8(5):321–324, May 1989.

[BPK+12] Stefano Battiston, Michelangelo Puliga, Rahul Kaushik, Paolo Tasca, and Guido Caldarelli. DebtRank: Too Central to Fail? Financial Networks, the FED and Systemic Risk. *Scientific Reports*, 2(1):1–6, 2012.

[BPPC07] Carter T. Butts, Miruna Petrescu-Prahova, and B. Remy Cross. Responder communication networks in the world trade center disaster: Implications for modeling of communication within emergency settings. *The Journal of Mathematical Sociology*, 31(2):121–147, 2007.

[BRU+20] Martin Richard Bicher, Claire Rippinger, Christoph Urach, Dominik Brunmeir, Uwe Siebert, and Niki Popper. Agent-based simulation for evaluation of contact-tracing policies against the spread of sars-cov-2. *medRxiv*, 2020.

[BS+20] Andras Borsos, Martin Stancsics, et al. Unfolding the hidden structure of the hungarian multi-layer firm network. Technical report, Magyar Nemzeti Bank (Central Bank of Hungary), 2020.

[BST04] Michael Boss, Martin Summer, and Stefan Thurner. Contagion flow through banking networks. In *International Conference on Computational Science*, pages 1070–1077. Springer, 2004.

[BWB11] James P. Bagrow, Dashun Wang, and Albert-László Barabási. Collective response of human populations to large-scale emergencies. *PLOS ONE*, 6(3):1–8, 03 2011.

[BZZ96] Hasida Ben-Zur and Moshe Zeidner. Gender differences in coping reactions under community crisis and daily routine conditions. *Personality and Individual Differences*, 20(3):331–340, 1996.

[Cab] Cablelabs. Data over cable service interface specification proactive network maintenance pnm best practices primer : Hfc networks. pages 1–134.

[CGW+08] Julián Candia, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[CJGP11] Gokul Chittaranjan, Blom Jan, and Daniel Gatica-Perez. Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *Proceedings - International Symposium on Wearable Computers, ISWC*, pages 29–36, 2011.

[CKT14] Anna Chmiel, Peter Klimek, and Stefan Thurner. Spreading of diseases through comorbidity networks across life and gender. *New Journal of Physics*, 16(11):115013, 2014.

[CLL20] Keren Cohen-Louck and Inna Levy. Risk perception of a chronic threat of terrorism: Differences based on coping types, gender and exposure. *International Journal of Psychology*, 55(1):115–122, 2020.

[cpa08] Regulation (EC) No 451/2008 of the European Parliament and of the Council of 23 April 2008 establishing a new statistical classification of products by activity (CPA) and repealing Council Regulation (EEC) No 3696/93 (Text with EEA relevance), 2008. http://data.europa.eu/eli/reg/2008/451/oj, retrieved 20th august 2021.

[CS19] Rama Cont and Eric Schaanning. Monitoring indirect contagion. *Journal of Banking & Finance*, 104:85–102, 2019.

[DBR+] Christian Diem, András Borsos, Tobias Reisch, János Kertész, and Stefan Thurner. Heterogeneity in initial shocks causes heterogeneity in outcomes. *In preparation.*

[DBR+21] Christian Diem, András Borsos, Tobias Reisch, János Kertész, and Stefan Thurner. Quantifying firm-level economic systemic risk from nation-wide supply networks, 2021. *arXiv preprint arXiv:2104.07260*.

[DCC+11] Jeanne F. Duffy, Sean W. Cain, Anne-Marie Chang, Andrew J. K. Phillips, Mirjam Y. Münch, Claude Gronfier, James K. Wyatt, Derk-Jan Dijk, Kenneth P. Wright, and Charles A. Czeisler. Sex difference in the near-24-hour intrinsic period of the human circadian timing system. *Proceedings of the National Academy of Sciences*, 108(Supplement 3):15602–15608, 2011.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

[Del15] Deloitte. Taxation and Investment in Hungary (rates are updated to 2017) (PDF), 2015. https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Tax/dttl-tax-hungaryguide-2015.pdf, retrieved 30th august 2021.

[DGFP20] Giacomo De Giorgi, Anders Frederiksen, and Luigi Pistaferri. Consumption network effects. *The Review of Economic Studies*, 87(1):130–163, 2020.

[DlDH+20] Amelie Desvars-larrive, Elma Dervic, Nils Haug, Thomas Niederkrotenthaler, Alexandr Ten, Alija Dervic, Andrea Pacheco, David Cserjan, Marcia R Ferreira, Rainer Vierlinger, Samantha Holder, and Samuel Álvarez. A structured open dataset of government interventions in response to covid-19. (iDiv), 2020.

[DLY19] Thi-Nga Dao, Duc Le, and Seokhoon Yoon. Predicting human location using correlated movements. *Electronics*, 8(1):54, 2019.

[DMR15] Emmanuel Dhyne, Glenn Magerman, and Stela Rubínová. The Belgian production network 2002-2012. Technical report, NBB Working Paper, 2015.

[DO11] Derek Dalton and Marc Ortegren. Gender differences in ethics research: The importance of controlling for the social desirability response bias. *Journal of Business Ethics*, 103(1):73–93, Sep 2011.

[DPPA+20] Dorothea Dumuid, Željko Pedišić, Javier Palarea-Albaladejo, Josep Antoni Martín-Fernández, Karel Hron, and Timothy Olds. Compositional data analysis in time-use epidemiology: What, why, how. *International journal of environmental research and public health*, 17(7):2220, 2020.

[DPT20] Christian Diem, Anton Pichler, and Stefan Thurner. What is the minimal systemic risk in financial exposure networks? *Journal of Economic Dynamics and Control*, 116:103900, 2020.

[DS16] Ayush Dusia and Adarshpal S. Sethi. Recent advances in fault localization in computer networks. *IEEE Communications Surveys and Tutorials*, 18(4):3030–3051, 2016.

[dWK+16] Christopher Bren d'Amour, Leonie Wenz, Matthias Kalkuhl, Jan Christoph Steckel, and Felix Creutzig. Teleconnected food supply shocks. *Environmental Research Letters*, 11(3):035007, 2016.

[Eck21] Heinz Eckert. Detecting incidents in HFC broadband networks via LSTM classification and autoencoder Models. Master's thesis, University of Applied Sciences Wiener Neustadt, 2021.

[ELPK20] Jakob-Moritz Eberl, Noelle Lebernegg, Julia Partheymüller, and Sylvia Kritzinger. Die meisten nehmen die lage ernst. aber wer sind die corona-skeptiker? https://viecer.univie.ac.at/corona-blog/corona-blog-beitraege/blog12/, 2020. Accessed 8 October 2020.

[EM15] Ahmed Eldawy and Mohamed F. Mokbel. Spatialhadoop: A mapreduce framework for spatial data. In *Proceedings - International Conference on Data Engineering*, 2015.

[EMC10] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

[ESS+20] Josh L Espinoza, Naisha Shah, Suren Singh, Karen E Nelson, and Chris L Dupont. Applications of weighted association networks applied to compositional data in biology. *Environmental Microbiology*, 2020.

[FA10] Yoshi Fujiwara and Hideaki Aoyama. Large-scale structure of a nation-wide production network. *The European Physical Journal B*, 77(4):565–580, 2010.

[FHS+20] Jude Ferreira, Maher Harb, Karthik Subramanya, Bryan Santangelo, and Dan Rice. Convolutional neural networks for proactive network management developing machine learning models to detect and classify table of contents. 2020.

[FHT18] Peter Filzmoser, Karel Hron, and Matthias Templ. *Applied Compositional Data Analysis. With Worked Examples in R*. Springer Series in Statistics, Springer, Cham, Switzerland, 2018.

[FK18] Mohammad Forghani and Farid Karimipour. Interplay between urban communities and human-crowd mobility: A study using contributed geospatial data sources. *Transactions in GIS*, 2018.

[FKC20] Mohammad Forghani, Farid Karimipour, and Christophe Claramunt. From cellular positioning data to trajectories: Steps towards a more accurate mobility exploration. *Transportation Research Part C: Emerging Technologies*, 117:102666, 2020.

[FTFS16] Yoshi Fujiwara, Masaaki Terai, Yuji Fujita, and Wataru Souma. Debtrank analysis of financial distress propagation on a production network in Japan, 2016. *RIETI Discussion Paper Series* 16-E-046.

[GAR+02] Sandro Galea, Jennifer Ahern, Heidi Resnick, Dean Kilpatrick, Michael Bucuvalas, Joel Gold, and David Vlahov. Psychological sequelae of the september 11 terrorist attacks in new york city. *New England Journal of Medicine*, 346(13):982–987, 2002.

[GGCI+17] Francisco García-García, Antonio Corral, Luis Iribarne, George Mavrommatis, and Michael Vassilakopoulos. A comparison of distributed spatial data management systems for processing distance join queries. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.

[GHB08] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[GPS+13] Sarthak Grover, Mi Seon Park, Srikanth Sundaresan, Sam Burnett, Hyojoon Kim, and Nick Feamster. Peeking behind the nat: An empirical study of home networks. *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC*, pages 377–389, 2013.

[GR19] David Garcia and Bernard Rimé. Collective emotions and social resilience in the digital traces after a terrorist attack. *Psychological Science*, 30(4):617–628, 2019.

[Gra20] Lucia Graves. Women's domestic burden just got heavier with the coronavirus. *the Guardian*, 2020.

[GRK+20] Song Gao, Jinmeng Rao, Yuhao Kang, Yunlei Liang, Jake Kruse, Doerte Doepfer, Ajay K. Sethi, Juan Francisco Mandujano Reyes, Jonathan Patz, and Brian S. Yandell. Mobile phone location data reveal the effect and geographic variation of social distancing on the spread of the covid-19 epidemic. *arXiv:2004.11430*, 586, 2020. Accessed 8 October 2020.

[GSM20] GSMA. The gsma covid-19 privacy guidelines. Technical Report April, GSMA, 2020. Accessed 8 October 2020.

[GTP+20] Laetitia Gauvin, Michele Tizzoni, Simone Piaggesi, Andrew Young, Natalia Adler, Stefaan Verhulst, Leo Ferres, and Ciro Cattuto. Gender gaps in urban mobility. *Humanities and Social Sciences Communications*, 7(1):1–13, 2020.

[Gus98] Per E Gustafsod. Gender differences in risk perception: Theoretical and methodological perspectives. *Risk Analysis*, 18(6):805–811, 1998.

[GZ18] Zhichao Guo and Tongyu Zhu. Identifying route preferences over origin-destination using cellular network data. Technical report, 2018.

[GZW20] Mohammadhossein Ghahramani, Mengchu Zhou, and Gang Wang. Urban sensing based on mobile phone data: Approaches, applications, and challenges. *IEEE/CAA Journal of Automatica Sinica*, 2020.

[HAE+15] James N. Hughes, Andrew Annex, Christopher N. Eichelberger, Anthony Fox, Andrew Hulbert, and Michael Ronquest. Geomesa: a distributed architecture for spatio-temporal fusion. In *Geospatial Informatics, Fusion, and Motion Video Analytics V*, 2015.

[Hag20] Johanna Hager. Arbeitsmarkt: 1,3 millionen in österreich in kurzarbeit, 517.221 ohne job, https://kurier.at/politik/inland/live-die-aktuellen-zahlen-zu-arbeitslosigkeit-und-kurzarbeit/400928603. Kurier Online, jun 2020. Accessed 8 October 2020.

[Hal09] Andrew G Haldane. Rethinking the financial network, speech delivered at the Financial Student Association, Amsterdam, 2009. Bank of England.

[Hal13] Diane F Halpern. *Sex Differences in Cognitive Abilities*. Psychology press, 2013.

[Haw13] Mary Hawkesworth. Sex, gender, and sexuality: From naturalized presumption to analytical categories. In *The Oxford handbook of gender and politics*. 2013.

[HC22] Georg Heiler and Alexandra Ciarnau. Datenanonymisierung - Der Schlüssel zur Innovation. *ecolex*, pages 166–168, 2022.

[Heu] Thomas Heuzroth. Corona-pandemie: So hat ischgl das virus in die welt getragen.

[HG20] Jeremy Howard and Sylvain Gugger. Fastai: A layered api for deep learning. *Information (Switzerland)*, 11(2):1–27, 2020.

[HGHF22] Georg Heiler, Thassilo Gadermaier, Allan Hanbury, and Peter Filzmoser. Identifying the root cause of cable network problems with machine learning, march 2022.

[HH19] Georg Heiler and Allan Hanbury. Comparing implementation variants of distributed spatial join on spark. In *Proceedings - IEEE Big Data*, 2019.

[HHF20] Georg Heiler, Allan Hanbury, and Peter Filzmoser. The impact of covid-19 on relative changes in aggregated mobility using mobile-phone data, 2020.

[HKPO20] Martin Hillebrand, Imran Khan, Filipa Peleja, and Nuria Oliver. Mobisenseus : Inferring aggregate objective and subjective well-being from mobile data. 2020.

[HLC04] Paul K Humphreys, WL Li, and LY Chan. The impact of supplier development on buyer–supplier performance. *Omega*, 32(2):131–143, 2004.

[HLW+20] Huajun He, Ruiyuan Li, Rubin Wang, Jie Bao, Yu Zheng, and Tianrui Li. Efficient suspected infected crowds detection based on spatio-temporal trajectories. *arXiv*, 2020.

[HRH+20] Georg Heiler, Tobias Reisch, Jan Hurt, Mohammad Forghani, Aida Omani, Allan Hanbury, and Farid Karimipour. Country-wide mobility changes observed using mobile phone data during covid-19 pandemic. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3123–3132. IEEE, 2020.

[HRKA21] Ruining He, Anirudh Ravula, Bhargav Kanagal, and Joshua Ainslie. Realformer: Transformer likes residual attention. pages 929–943, 2021.

[HRV05] M. Hubert, P. J. Rousseeuw, and Karlien Vanden Branden. Robpca: A new approach to robust principal component analysis. *Technometrics*, 47:64–79, 2005.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[HZY+20] Jiyao Hu, Zhenyu Zhou, Xiaowei Yang, Jacob Malone, and Jonathan W Williams. Cablemon: Improving the reliability of cable broadband networks via proactive network maintenance. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*, pages 619–632, Santa Clara, CA, February 2020. USENIX Association.

[IB18] Joseph Gilley Nick Rabinowitz David Ellis Isaac Brodsky, Kevin Sahr. H3: Uber's hexagonal hierarchical spatial index. https://eng.uber.com/h3/, 2018. Accessed 8 October 2020.

[IDMP+08] Giulia Iori, Giulia De Masi, Ovidiu Vasile Precup, Giampaolo Gabbi, and Guido Caldarelli. A network analysis of the Italian overnight money market. *Journal of Economic Dynamics and Control*, 32(1):259–278, 2008.

[IFMFM18] Sibren Isaacman, Vanessa Frias-Martinez, and Enrique Frias-Martinez. Modeling human migration paerns during drought conditions in la guajira, colombia. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS 2018*, number 18, 2018.

[ILF+20] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

[IS20] IFES and SORA. Home-office: Positive resonanz, aber mehr stress. https://www.ifes.at/arbeitsklima-index-2020-home-office, 2020. Accessed 8 October 2020.

[ISS+20] Stefano Maria Iacus, Carlos Santamaria, Francesco Sermi, Spyridon Spyratos, Dario Tarchi, and Michele Vespe. *Mapping Mobility Functional Areas (MFA) by Using Mobile Positioning Data to Inform COVID-19 Policies: A European Regional Analysis*. Publications Office of the European Union, 2020.

[IT19] Hiroyasu Inoue and Yasuyuki Todo. Firm-level propagation of shocks through supply-chain networks. *Nature Sustainability*, 2(9):841–847, 2019.

[ITY05] Tasuku Igarashi, Jiro Takai, and Toshikazu Yoshida. Gender differences in social network development via mobile phone text messages: A longitudinal study. *Journal of Social and Personal Relationships*, 22(5):691–713, 2005.

[JAD+20] Jonas S. Juul, Laura Alessandretti, Jesper Dammeyer, Ingo Zettler, Sune Lehmann, and Joachim Mathiesen. Gender-specific behavior change following terror attacks. https://arxiv.org/abs/2004.02957, 2020. Accessed 8 October 2020.

[JC18] Abdennassar Joueid and Germà Coenders. Marketing innovation and new product portfolios. a compositional approach. *Journal of Open Innovation: Technology, Market, and Complexity*, 4(2):19, 2018.

[JLY+20a] Jayson S. Jia, Xin Lu, Yun Yuan, Ge Xu, Christakis A., Jianmin Jia, and Nicholas A. Population flow drives spatio-temporal distribution of covid-19 in china. *Nature*, 2020.

[JLY+20b] Jayson S Jia, Xin Lu, Yun Yuan, Ge Xu, Jianmin Jia, and Nicholas A Christakis. Population flow drives spatio-temporal distribution of covid-19 in china. *Nature*, 582(7812):1–5, 2020.

[JWA+20] Benjamin Jeffrey, Caroline E Walters, Kylie E C Ainslie, Oliver Eales, Constanze Ciavarella, and Sangeeta Bhatia. Report 24 : Anonymised and aggregated crowd level mobility data from mobile phones suggests that initial compliance with covid-19 social distancing interventions was high and geographically consistent across the uk. Technical Report May, 2020.

[KKB+20] Bernhard Kittel, Sylvia Kritzinger, Hajo Boomgaarden, Barbara Prainsack, Jakob-Moritz Eberl, Fabian Kalleitner, Noëlle Lebernegg, Julia Partheymueller, Carolina Plescia, David W Schiestl, et al. The austrian corona panel project: Monitoring individual and societal dynamics amidst the covid-19 crisis, 2020.

[KKGB10] Pablo Kaluza, Andrea Kölzsch, Michael T Gastner, and Bernd Blasius. The complex network of global cargo ship movements. *Journal of the Royal Society Interface*, 7(48):1093–1103, 2010.

[KMF+17] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 2017-December(Nips):3147–3155, 2017.

[KSH21] Fabian Kalleitner, David W Schiestl, and Georg Heiler. Varieties of mobility measures: Comparing survey and mobile phone data during the covid-19 pandemic, Oct 2021.

[KYG+20] Moritz U.G. Kraemer, Chia Hung Yang, Bernardo Gutierrez, Chieh Hsi Wu, Brennan Klein, David M. Pigott, Louis du Plessis, Nuno R. Faria, Ruoran Li, William P. Hanage, John S. Brownstein, Maylis Layan, Alessandro Vespignani, Huaiyu Tian, Christopher Dye, Oliver G. Pybus, and Samuel V. Scarpino. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science (New York, N.Y.)*, 497(May):493–497, 2020.

[LBH12] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, Jul 2012.

[LBR14] Carlos León, Ron Berndsen, and Luc Renneboog. Financial stability and interacting networks of financial institutions and market infrastructures, 2014. *European Banking Center Discussion Paper Series*, 2014-011.

[LKC+21] Katharina Ledebur, Michaela Kaleta, Jiaying Chen, Simon Lindner, Caspar Matzhold, Florian Weidle, Christoph Wittmann, Katharina Habimana, Linda Kerschbaumer, Sophie Stumpfl, Georg Heiler, Martin Bicher, Nikolas Popper, Florian Bachner, and Peter Klimek. Meteorological factors and non-pharmaceutical interventions explain local differences in the spread of sars-cov-2 in austria, 2021.

[LLP+15] Thomas Louail, Maxime Lenormand, Miguel Picornell, Oliva García Cantú, Ricardo Herranz, Enrique Frias-Martinez, José J. Ramasco, and Marc Barthelemy. Uncovering the spatial structure of mobility networks. *Nature Communications*, 6, 2015.

[LMZZ18] Huan Liu, Lin Ma, Xi Zhao, and Jianhua Zou. An effective model between mobile phone usage and p2p default behavior. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10861 LNCS, pages 462–475, 2018.

[Lon10] Francis A. Longstaff. The subprime credit crisis and contagion in financial markets. *Journal of Financial Economics*, 97(3):436–450, 2010.

[LPCR+15] Maxime Lenormand, Miguel Picornell, Oliva G. Cantú-Ros, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frías-Martínez, Maxi San Miguel, and José J. Ramasco. Comparing and modelling land use organization in cities. *Royal Society Open Science*, 2(12), 2015.

[LT17] Matt V Leduc and Stefan Thurner. Incentivizing resilience in financial networks. *Journal of Economic Dynamics and Control*, 82:44–66, 2017.

[MAH+09] Galanakis Michael, Stalikas Anastasios, Kallia Helen, Karagianni Catherine, and Karela Christine. Gender differences in experiencing occupational stress: the role of age, education and marital status. *Stress and Health: Journal of the International Society for the Investigation of Stress*, 25(5):397–404, 2009.

[Mat04] M Pilar Matud. Gender differences in stress and coping styles. *Personality and Individual Differences*, 37(7):1401–1415, 2004.

[MB19] José Moran and Jean-Philippe Bouchaud. May's instability in large economies. *Phys. Rev. E*, 100:032307, Sep 2019.

[MBG+17] Daniel Monsivais, Kunal Bhattacharya, Asim Ghosh, Robin I.M. Dunbar, and Kimmo Kaski. Seasonal and geographical impact on human resting periods. *Scientific Reports*, 7(1), 2017.

[MBL+19] Marco Mamei, Nicola Bicocchi, Marco Lippi, Stefano Mariani, and Franco Zambonelli. Evaluating origin–destination matrices obtained from cdr data. *Sensors (Switzerland)*, 19(20):4470, oct 2019.

[MDV16] Kelly R Moran and Sara Y Del Valle. A meta-analysis of the association between gender and protective behaviors in response to respiratory epidemics and pandemics. *PLOS ONE*, 11(10):e0164541, 2016.

[ML19] Aude Marzuoli and Fengmei Liu. A data-driven impact evaluation of hurricane harvey from mobile phone data. In *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pages 3442–3451, 2019.

[MRR21] Irma Mooi-Reci and Barbara J Risman. The gendered impacts of covid-19: Lessons and reflections, 2021.

[MT21] Kanika Mahajan and Shekhar Tomar. COVID-19 and Supply Chain Disruption: Evidence from Food Markets in India. *American Journal of Agricultural Economics*, 103(1):35–52, 2021.

[MWNM20] Helen Jaqueline McLaren, Karen Rosalind Wong, Kieu Nga Nguyen, and Komalee Nadeeka Damayanthi Mahamadachchi. Covid-19 and women's triple burden: Vignettes from sri lanka, malaysia, vietnam and australia. *Social Sciences*, 9(5):87, 2020.

[nac06] Regulation (EC) No 1893/2006 of the European Parliament and of the Council of 20 December 2006 establishing the statistical classification of economic activities NACE Revision 2 and amending Council Regulation (EEC) No 3037/90 as well as certain EC Regulations on specific statistical domains Text with EEA relevance, 2006. http://data.europa.eu/eli/reg/2006/1893/oj, retrieved 20th august 2021.

[NIZ+16] Ana Nika, Asad Ismail, Ben Y. Zhao, Sabrina Gaito, Gian Paolo Rossi, and Haitao Zheng. Understanding and predicting data hotspots in cellular networks. *Mobile Networks and Applications*, 21(3):402–413, 2016.

[OEC20] OECD. Women at the core of fight against covid-19 crisis. https://www.oecd.org/coronavirus/policy-responses/women-at-the-core-of-the-fight-against-covid-19-crisis-553a8269/, 2020. Accessed 8 October 2020.

[Ogu20] Ignacio Oguiza. tsai - a state-of-the-art deep learning library for time series and sequential data. Github, 2020.

[OLS+20] Nuria Oliver, Bruno Lepri, Harald Sterly, Renaud Lambiotte, Sébastien Deletaille, Marco De Nadai, Emmanuel Letouzé, Albert Ali Salah, Richard Benjamins, Ciro Cattuto, Vittoria Colizza, Nicolas de Cordes, Samuel P. Fraiberger, Till Koebe, Sune Lehmann, Juan Murillo, Alex Pentland, Phuong N Pham, Frédéric Pivetta, Jari Saramäki, Samuel V. Scarpino, Michele Tizzoni, Stefaan Verhulst, and Patrick Vinck. Mobile phone data for informing public health actions across the covid-19 pandemic life cycle. *Science Advances*, 6(23), 2020.

[orf20] orf.at. Leergefegte straßen und plätze, mar 2020.

[OSH+07] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[PB10] Gordon Parker and Heather Brotchie. Gender differences in depression. *International Review of Psychiatry*, 22(5):429–436, 2010.

[PBG+20] Emanuele Pepe, Paolo Bajardi, Laetitia Gauvin, Filippo Privitera, Brennan Lake, Ciro Cattuto, and Michele Tizzoni. Covid-19 outbreak response: a first assessment of mobility changes in italy following national lockdown. *medRxiv*, page 2020.03.22.20039933, 2020.

[PGETD15] V. Pawlowsky-Glahn, J.J. Egozcue, and R. Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. Wiley, Chichester, 2015.

[PKK+12] Vasyl Palchykov, Kimmo Kaski, Janos Kertész, Albert-László Barabási, and Robin IM Dunbar. Sex differences in intimate relationships. *Scientific Reports*, 2(1):1–5, 2012.

[PLC08] Antony Paulraj, Augustine A Lado, and Injazz J Chen. Inter-organizational communication as a relational competency: Antecedents and performance outcomes in collaborative buyer–supplier relationships. *Journal of Operations Management*, 26(1):45–64, 2008.

[PLMG20] Max Pellert, Jana Lasser, Hannah Metzler, and David Garcia. Dashboard of sentiment in austrian social media during covid-19. http://arxiv.org/abs/2006.11158, 2020. see also http://www.mpellert.at/covid19_monitor_austria/, accessed 8 October 2020.

[PMBMJ+15] Sebastian Poledna, José Luis Molina-Borboa, Serafín Martínez-Jaramillo, Marco Van Der Leij, and Stefan Thurner. The multi-layer network nature of systemic risk and its implications for the costs of financial crises. *Journal of Financial Stability*, 20:70–81, 2015.

[PPT21] Anton Pichler, Sebastian Poledna, and Stefan Thurner. Systemic risk-efficient asset allocations: Minimization of systemic risk as a network optimization problem. *Journal of Financial Stability*, 52:100809, 2021.

[Pra20] Barbara Prainsack. Solidarity in times of a pandemic: What do people do, and why? In preparation., 2020.

[PSHS08] Joseph Prashker, Yoram Shiftan, and Pazit Hershkovitch-Sarusi. Residential choice location, gender and the commute trip to work in tel aviv. *Journal of Transport Geography*, 16(5):332–341, 2008.

[PT16] Sebastian Poledna and Stefan Thurner. Elimination of systemic risk in financial networks by means of a systemic risk transaction tax. *Quantitative Finance*, 16(10):1599–1613, 2016.

[Pur18] Ida Bagus Irawan Purnama. Spatiotemporal mining of bss data for characterising seasonal urban mobility dynamics. *International Journal on Advanced Science, Engineering and Information Technology*, 8(4):1270–1276, 2018.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[QERC18] Thomas P Quinn, Ionas Erb, Mark F Richardson, and Tamsyn M Crowley. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16):2870–2878, 2018.

[RHD+22] Tobias Reisch, Georg Heiler, Christian Diem, Peter Klimek, and Stefan Thurner. Monitoring supply networks from mobile phone data for estimating the systemic risk of an economy. *Scientific reports*, 12(1):1–10, 2022.

[RHH+21] Tobias Reisch, Georg Heiler, Jan Hurt, Peter Klimek, Allan Hanbury, and Stefan Thurner. Behavioral gender differences are reinforced during the covid-19 crisis. *Scientific Reports*, 11(1):1–12, 2021.

[RL20] Neil J Rowan and John G Laffey. Challenges and solutions for addressing critical shortage of supply chain for personal and protective equipment (PPE) arising from Coronavirus disease (COVID19) pandemic–Case study from the Republic of Ireland. *Science of the Total Environment*, 725:138532, 2020.

[RW18] Dan Rice and Jon-en Wang. What gets measured gets done / what gets analyzed gets transformed analytics for a wider / deeper network view table of contents list of figures. Technical report, CableLabs, 2018.

[SAT20] Thiago C Silva, Diego R Amancio, and Benjamin M Tabak. Modeling Economic Networks with Firm-to-Firm Wire Transfers, 2020. *arXiv preprint arXiv:2001.06889*.

[SDO+21] Markus Schläpfer, Lei Dong, Kevin O'Keeffe, Paolo Santi, Michael Szell, Hadrien Salat, Samuel Anklesaria, Mohammad Vazifeh, Carlo Ratti, and Geoffrey B West. The universal visitation law of human mobility. *Nature*, 593(7860):522–527, 2021.

[SE20] Janina Steiner and Cara Ebert. The impact of covid-19 on violence against women and children in germany. preprint: https://www.hfp.tum.de/globalhealth/forschung/covid-19-and-domestic-violence/, june 2020. Accessed 8 October 2020.

[SFS+20] SANTAMARIA SERNA Carlos, SERMI Francesco, SPYRATOS Spyridon, IACUS Stefano, ANNUNZIATO Alessandro, TARCHI Dario, and VESPE Michele. Measuring the Impact of COVID-19 Confinement Measures on Human Mobility using Mobile Positioning Data (1), 2020.

[SK08] Keemin Sohn and Daehyun Kim. Dynamic origin-destination flow estimation using cellular communication system. *IEEE Transactions on Vehicular Technology*, 57(5):2703–2713, 2008.

[SLS+19] Xiaoying Shi, Fanshun Lv, Dewen Seng, Baixi Xing, and Jing Chen. Exploring the evolutionary patterns of urban activity areas based on origin-destination data. *IEEE Access*, 7:20416–20431, 2019.

[SMBH17] Meead Saberi, Hani S. Mahmassani, Dirk Brockmann, and Amir Hosseini. A complex network perspective for characterizing urban travel demand patterns: graph theoretical analysis of large-scale origin–destination demand networks. *Transportation*, 2017.

[SNT+21] Chandan Singh, Keyan Nasseri, Yan Shuo Tan, Tiffany Tang, and Bin Yu. imodels: a python package for fitting interpretable models, 2021.

[SQBB10] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[Sri14] Ram Sriharsha. Magellan, 2014.

[SS97] Erich J Schwarz and Karl W Steininger. Implementing nature's lesson: the industrial recycling network enhancing regional development. *Journal of Cleaner Production*, 5(1-2):47–56, 1997.

[SSS+20] Carlos Santamaria, Francesco Sermi, Spyridon Spyratos, Stefano Maria Iacus, Alessandro Annunziato, Dario Tarchi, and Michele Vespe. Measuring the impact of covid-19 confinement measures on human mobility using mobile positioning data. a european regional analysis. *Safety Science*, 132:104925, 2020.

[ST13] M. Szell and S. Thurner. How women organize social networks different from men: gender-specific behavior in large-scale social networks. *Scientific Reports*, 3:1214, 2013.

[SWF00] Murray B Stein, John R Walker, and David R Forde. Gender differences in susceptibility to posttraumatic stress disorder. *Behaviour Research and Therapy*, 38(6):619–628, 2000.

[TH20] Robert Thompson and Robert Howald. A proactive network management scheme for mid-split deployment. Technical report, CableLabs, 2020.

[TKL+00] Shelley E Taylor, Laura Cousino Klein, Brian P Lewis, Tara L Gruenewald, Regan AR Gurung, and John A Updegraff. Biobehavioral responses to stress in females: Tend-and-befriend, not fight-or-flight. *Psychological Review*, 107(3):411, 2000.

[TMJ+16] János Török, Yohsuke Murase, Hang-Hyun Jo, János Kertész, and Kimmo Kaski. What big data tells: sampling the social network by communication channels. *Physical Review E*, 94(5):052319, 2016.

[TMTAS19] Huong Thi Trinh, Joanna Morais, Christine Thomas-Agnan, and Michel Simioni. Relations between socio-economic factors and nutritional diet in vietnam from 2004 to 2014: New insights using compositional data analysis. *Statistical methods in medical research*, 28(8):2305–2325, 2019.

[TP13] Stefan Thurner and Sebastian Poledna. Debtrank-transparency: Controlling systemic risk in financial networks. *Scientific Reports*, 3(1):1–7, 2013.

[TYM+15] Mingjie Tangy, Yongyang Yuy, Qutaibah M. Malluhiz, Mourad Ouzzani, and Walid G. Arefy. Locationspark: A distributed in-memory data management system for big spatial data. In *Proceedings of the VLDB Endowment*, volume 9, pages 1565–1568, 2015.

[Vig20] Giuliana Viglione. Are women publishing less during the pandemic? here's what the data say. *Nature*, 581(7809):365–366, 2020.

[VMU+20] Michaela A C Vollmer, Swapnil Mishra, H Juliette T Unwin, Axel Gandy, Thomas A Mellan, Harrison Zhu, Helen Coupland, Iwona Hawryluk, Michael Hutchinson, Oliver Ratmann, Patrick Walker, Charlie Whittaker, Lorenzo Cattarino, Constance Ciavarella, Lucia Cilloni, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Nicholas Brazeau, Giovanni Charles, V Cooper, Zulma Cucunuba, Gina Cuomo-dannenburg, Amy Dighe, Bimandra Djaafara, Jeff Eaton, L Van Elsland, Richard Fitzjohn, Keith Fraser, Katy Gaythorpe, Will Green, Sarah Hayes, Natsuko Imai, Edward Knock, Daniel Laydon, John Lees, Tara Mangal, Andria Mousa, Gemma Nedjati-gilani, Pierre Nouvellet, Daniela Olivera, Kris V Parag, Michael Pickles, Hayley A Thompson, Robert Verity, Haowei Wang, Yuanrong Wang, Oliver J Watson, Lilith Whittles, Xiaoyue Xi, and Azra Ghani. Report 20 : Using mobility to estimate the transmission intensity of covid-19 in italy : A subnational analysis with future scenarios. Technical Report May, 2020.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[Wat07] Duncan J Watts. A twenty-first century science. *Nature*, 445(7127):489–489, 2007.

[WHTG] Larry Wolcott, John Heslip, Bryan Thomas, and Robert Gonsalves. A comprehensive case study of proactive network maintenance a technical paper prepared for scte/isbe. 80202(702).

[WT14] Qi Wang and John E. Taylor. Quantifying human mobility perturbation and resilience in hurricane sandy. *PLOS ONE*, 9(11):e112608, Nov 2014.

[WXX+20] Huiyao Wang, Qian Xia, Zhenzhen Xiong, Zhixiong Li, Weiyi Xiang, Yiwen Yuan, Yaya Liu, and Zhe Li. The psychological distress and coping styles in the early stages of the 2019 coronavirus disease (covid-19) epidemic in the general mainland chinese population: A web-based survey. *PLOS ONE*, 15(5):e0233410, 2020.

[WYU+15] Peter Widhalm, Yingxiang Yang, Michael Ulm, Shounak Athavale, and Marta C. González. Discovering urban activity patterns in cell phone data. *Transportation*, 42(4):597–623, 2015.

[XGM+20] Bo Xu, Bernardo Gutierrez, Sumiko Mekaru, Kara Sewalk, Lauren Goodwin, Alyssa Loskill, Emily L. Cohn, Yulin Hswen, Sarah C. Hill, Maria M. Cobo, Alexander E. Zarebski, Sabrina Li, Chieh Hsi Wu, Erin Hulland, Julia D. Morgan, Lin Wang, Katelynn O'Brien, Samuel V V. Scarpino, John S. Brownstein, Oliver G. Pybus, David M. Pigott, and Moritz U.G. Kraemer. Epidemiological data from the covid-19 outbreak, real-time case information. *Scientific Data*, 7(1):1–6, 2020.

[XLY+16] Dong Xie, Feifei Li, Bin Yao, Gefei Li, Liang Zhou, and Minyi Guo. Simba: Efficient in-memory spatial analytics. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2016.

[YHO+19] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *CoRR*, abs/1905.04899, 2019.

[YJM15] Jia Yu, Wu Jinxuan, and Sarwat Mohamed. Geospark: A cluster computing framework for processing large-scale spatial data. *SIGSPATIAL International Conference on Advances in Geographic Inormation Systems*, (3):4–7, 2015.

[YTA+17] Yongyang Yu, Mingjie Tang, Walid G. Aref, Qutaibah M. Malluhi, Mostafa M. Abbas, and Mourad Ouzzani. In-memory distributed matrix computation processing & optimization. In *Proceedings - International Conference on Data Engineering*, 2017.

[YTF+20] Takahiro Yabe, Kota Tsubouchi, Naoya Fujiwara, Takayuki Wada, Yoshihide Sekimoto, and Satish V. Ukkusuri. Non-compulsory measures sufficiently reduced human mobility in japan during the covid-19 epidemic. pages 1–9, 2020.

[YZS19] Jia Yu, Zongsi Zhang, and Mohamed Sarwat. Spatial data management in apache spark: the geospark perspective and beyond. *GeoInformatica*, 23(1):37–78, 2019.

[ZBMR20] Pengxiang Zhao, Dominik Bucher, Henry Martin, and Martin Raubal. A clustering-based framework for understanding individuals' travel mode choice behavior. In *Lecture Notes in Geoinformation and Cartography*, pages 77–94, 2020.

[ZCDD12] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, and Ankur Dave. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *NSDI'12 Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, 2012.

[Zei06] Moshe Zeidner. Gender group differences in coping with chronic terror: The israeli scene. *Sex Roles*, 54(3-4):297–310, 2006.

[ZJP+21] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. *A Transformer-Based Framework for Multivariate Time Series Representation Learning*, page 2114–2124. Association for Computing Machinery, New York, NY, USA, 2021.

[ZMLZ10] Li Zhang, Yong Tao Ma, Kai Hua Liu, and Yuan Zeng. Research of the noise characteristic on the upstream channel for hfc network. *ICSPS 2010 - Proceedings of the 2010 2nd International Conference on Signal Processing Systems*, 2:426–430, 2010.

[ZS17] Jay Zhu and Karthik Sundaresan. Access network data analytics (machine learning applied to cable access data). Technical report, CableLabs, 2017.

[ZSR20] Jingjie Zhu, Karthik Sundaresan, and Jason Rupe. Proactive network maintenance using fast, accurate anomaly localization and classification on 1-d data series. *Proceedings of the Annual Conference of the Prognostics and Health Management Society, PHM*, 2020-June:1–11, 2020.