Characterizing Large-Scale Mobile Traffic Measurements for Urban, Social and Networks Sciences

by

André Felipe Zanella

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

Telematic Engineering

Universidad Carlos III de Madrid

Tutor/Advisor: Marco Fiore

September 2024

Characterizing Large-Scale Mobile Traffic Measurements for Urban, Social and Networks Sciences

Prepared by:
André Felipe Zanella, IMDEA Networks Institute, Universidad Carlos III de Madrid
contact: andre.zanella@imdea.org

Under the advice of:
Marco Fiore, IMDEA Networks Institute

This work has been supported by:

Acknowledgements

I'd like to thank first Dr. Marco Fiore, who trusted me and invited me to work with his group. I grew a lot during this these years, and he was the principal source of inspiration, guidance and help throughout this journey. I learned through him not only how to do good research, but also how to be a better professional, colleague and a teacher to all around me. Thanks for all your patience and help during all our long deadline sessions, and for the counseling for my research and career.

A gigantic thanks for all my colleagues (who I'm lucky to call friends) at the Networks Data Science team, as they were the ones giving me great support from day one to help me learn the ways around our work and develop great research. I'd like to thank in special my PhD colleagues Orlando E. Martínez-Durive and Sachit Mishra for their help over late hours so we could always crunch through our terabytes of data and get the nicest plots possible, and for my Post-Doc colleagues Dr. Antonio Bazco-Nogueras and Dr. Diego Madariaga who also always gave great advice for my research.

I'd like to thank the full IMDEA Networks team, who were always there for (a few too many) coffees, ping pong matches, lunch breaks and great chats. I wish I could list all here, but I'd like to thank in special my good friends Michele, Akem, Sergi, Leonardo, Alan, Nikos, Andrea and Mauro. Having you all here always made things easier. I'd also like to thank the HR and Admin team of IMDEA Networks for giving great support to overcome all the bureaucracy I faced, specially for Alejandro Delgado and Elvira Conti.

I'd like to thank the great hosts and colleagues throughout my secondments for the BANYAN project. A special thanks for Dr. Zbigniew Smoreda and Dr. Stefania Rubrichi who were of great help during my stay at Orange, and for Dr. Ian Wassell and Dr. Kan Lin for their help during my stay at the University of Cambridge and Ranplan Wireless. Also, a special thanks for my BANYAN colleagues, specially Akem, Stefanos and Gabriel, who I had the pleasure of working together and having great discussions across quite a few different countries.

Last but definitely not least, I'd like to thank all my family and close ones for their support through these years. A gigantic thank you for my wife Thais, who joined me in this great international journey across many countries and cities and was always patiently there to support me in the good and bad times. Another gigantic thank you for my mother,

father and sister who were supporting me an ocean away. Also for all my friends back home who were always there over messages and video calls hearing me complain a bit too much, but always being supportive. And finally, a (necessary) thanks for my dog Tim who was usually by my side watching me work over long days and night hours in 3 different countries and was always a good boy.

Published and Submitted Content

This thesis is based on the following published papers (in reverse chronological order):

[1] Sachit Mishra, **André Felipe Zanella**, Orlando E. Martínez-Durive, Diego Madariaga, Cezary Ziemlicki, Marco Fiore. *Characterizing 5G Adoption and its Impact on Network Traffic and Mobile Service Consumption.* In: IEEE INFOCOM 2024 - IEEE Conference on Computer Communications. IEEE, 2024.

- The contents of this work are wholly included and are reported in Section 4.1.
- The author's role in this work was to perform the analysis on the experimental results, prepared and review the manuscript.
- The material from this source included in this thesis is not singled out with typographic means and references.

[2] **André Felipe Zanella**, Orlando E. Martínez-Durive, Sachit Mishra, Diego Madariaga, Marco Fiore. *Impact of Public Protests on Mobile Networks.* In: IEEE INFOCOM 2024 Poster - IEEE Conference on Computer Communications Workshop (IEEE WKSHPS). IEEE, 2024.

- The contents of this work are wholly included and are reported in Chapter 5.2.
- The author's role in this work was to conceptualize, perform experiments, write and review the manuscript.
- The material from this source included in this thesis is not singled out with typographic means and references.

[3] Gabriel O. Ferreira, **André F. Zanella**, Stefanos Bakirtzis, Chiara Ravazzi, Fabrizio Dabbene, Giuseppe C. Calafiore, Ian Wassel, Jie Zhang and Marco Fiore. *A Joint Optimization Approach for Power-Efficient Heterogeneous OFDMA Radio Access Networks.* In: IEEE Journal on Selected Areas in Communications (JSAC). IEEE, 2024.

- The contents of this work are is partially included and are reported in Chapter 7.5.
- The author's role in this work was to prepare data, conceptualize experiments, prepare and review the manuscript.

- The material from this source included in this thesis is not singled out with typographic means and references.

[4] Angelo Furno, **André Felipe Zanella**, Razvan Stanica, Marco Fiore. *Spatial and Temporal Exploratory Factor Analysis of Urban Mobile Data Traffic.* In: Data Science for Transportation 6.1 (2024): 4.

- The contents of this work are wholly included and are reported in Section 5.1.
- The author's role in this work was to perform the analysis on the experimental results, prepare and review the manuscript.
- The material from this source included in this thesis is not singled out with typographic means and references.

[5] **André Felipe Zanella**, Antonio Bazco-Nogueras, Cezary Ziemlicki, Marco Fiore. *Characterizing and Modeling Session-Level Mobile Traffic Demands from Large-Scale Measurements.* In: Proceedings of the 2023 ACM on Internet Measurement Conference. 2023.

- The contents of this work are wholly included and are reported in Chapter 7
- The author's role in this work was to conceptualize, perform the data collection and experiments, analyze results, prepare and review the manuscript.
- The material from this source included in this thesis is not singled out with typographic means and references.

[6] Stefanos Bakirtzis, **André Felipe Zanella**, Stefania Rubrichi, Cezary Ziemlicki, Zbigniew Smoreda, Ian Wassell, Jie Zhang, Marco Fiore. *Characterizing Mobile Service Demands at Indoor Cellular Networks.* In: Proceedings of the 2023 ACM on Internet Measurement Conference. 2023.

- The contents of this work are partially included and are reported in Section 4.2
- The author's role in this work was to conceptualize, perform experiments, write and review the manuscript.
- The material from this source included in this thesis is not singled out with typographic means and references.

[7] **André Felipe Zanella**, Orlando E. Martínez-Durive, Sachit Mishra, Zbigniew Smoreda, Marco Fiore. *Impact of later-stages COVID-19 response measures on spatiotemporal mobile service usage.* In: IEEE INFOCOM 2022-IEEE Conference on Computer Communications. IEEE, 2022.

- The contents of this work are wholly included and are reported in Section 6.1

- The author's role in this work was to conceptualize, perform experiments, write and review the manuscript.
- The material from this source included in this thesis is not singled out with typographic means and references.

The following papers are a part of this thesis and currently under review:

[8] **André Felipe Zanella**, Diego Madariaga, Sachit Mishra, Orlando E. Martínez-Durive, Zbigniew Smoreda, Marco Fiore. *Characterizing, modeling and exploiting the mobile demand footprint of large public protests.*

- The contents of this work are wholly included and are reported in Section 6.3
- The author's role in this work was to conceptualize, perform experiments, write and review the manuscript.
- The material from this source included in this thesis is not singled out with typographic means and references.

[9] **André Felipe Zanella**, Stefania Rubrichi, Zbigniew Smoreda, Marco Fiore. *Modeling and understanding the impact of COVID-19 containment policies on mobile service consumption in French cities.*

- The contents of this work are wholly included and are reported in Section 6.2
- The author's role in this work was to conceptualize, perform experiments, write and review the manuscript.
- The material from this source included in this thesis is not singled out with typographic means and references.

Abstract

Over the last few decades, it is difficult to pinpoint a technological advancement that shifted the daily life of the world's population in a more disrupting way than mobile phones and their applications. Their ubiquitousness has reshaped global behaviors and routines, transforming portable devices into the essential and on-the-go personal computer. Mobile phones enable communication, information access, and entertainment with little to no location constraints, thanks to their connectivity to the internet through a pervasive radio access network infrastructure. Of course, this seamless mobile access was not always a given, and decades of research, development and technology integration were needed to reach today's high-capacity support for broadband and low-latency mobile services.

From the first generations of mobile networks offering only on-the-go voice and text to the current fourth and fifth generations supporting high-resolution on-demand video streaming and low-latency cloud applications, every new release contributed to transforming mobile phones into essential items. With the rise in popularity of mobile applications in smartphones, any company or developer could release their own application, giving access to their product to consumers anywhere. Advancements in mobile networks meant an increase in data transfer capacities, leading users to be more comfortable utilizing their smartphones for tasks anywhere and at any time. The sucess of mobile technologies also signifies that the patterns of usage captured by mobile operators reflect in a rich and detailed way the endeavors of their vast user population.

Due to this reason, the data collected in operational mobile networks has today become a primary source of information for research in networking and beyond. Early research utilized analysis of mobile network traffic as feedback for the mobile operator itself, as a way to understand the spatiotemporal dynamics of the operational demands in the network, and how this could be leveraged to improve network deployments and operation according to consumption patterns. A second and broader direction lies in interdisciplinary research, seeking to explore how these measurements could be used to understand populations and urban environment dynamics.

This drives the need of research oriented towards networks data science: the study of tools and methodologies capable of handling large-scale measurements, asserting the quality and precision of the collected data in reflecting reality, as well as developing tools

capable of extracting insights and making the vastness of collected information useful for analysis. This thesis is a step in the direction of establishing said tools and methodologies, as well as showcasing several potential directions that mobile network measurement can support in interdisciplinary research domains.

The first part of this thesis presents a full contextualization of networks data science, expanding on the problem of the ever-growing scale of collected sets and presenting the many different fields that have been explored over the years, from classical network engineering applications to the study of populations, epidemics, socioeconomic and people's movement and transportation across cities. These studies are not possible without a well-established routine of data collection and processing, which is also discussed in the first part of this thesis.

The second part of the manuscript presents the original contributions provided by the thesis. Four chapters explore different directions where the collected data can be leveraged to derive new insights. First, an overview of the adoption of new technologies provided by the mobile network operator, with new findings into how changes in their traffic patterns may happen according to these new capabilities. Second, an exploration of how mobile consumption patterns and demands can be utilized to better understand the space within cities, with new methodologies presenting how both city-wide and location-specific insights can be gained just by looking at the traffic being consumed within base stations of the network. Third, a look into how special events may impact mobile networks, as such occurrences affect directly how users interact with their smartphones. It becomes important for the network operator to be able to extract insights and understand how these variations in traffic demand across time, space, and applications affect the functioning of the network, as well as these insights can be used by lawmakers to understand how these events affect populations. Lastly, a study characterizing session-level measurements to derive insights used to generate synthetic data sets containing new dynamics, generating simple models that can be utilized by anyone interested in research within mobile networks to test and validate their data-driven solutions.

In summary, the age of pervasive digital services leaves researchers with oceans of data and information to be explored in many different areas, with mobile networks shaping into a major source of rich information to guide innovation in both cutting edge research and technology development. This thesis guides the reader through the current state of affairs, showcasing the current opportunities opened by mobile network data and also presenting potential future directions that can be pursed in the following years.

Contents

List of Tables

List of Figures

List of Acronyms

BS Base Station
C-RAN Cloud Radio Access Networks
CDF Cumulative Distribution Function
CDR Call Detail Record
CI Confidence Interval
CN Core Network
DPI Deep Packet Inspection
DPO Data Protection Officer
EDGE Enhanced Data Rates for GSM Evolution
EFA Exploratory Factor Analysis
EMD Earth Mover Distance
GDPR General Data Protection Regulation
GGSN Gateway GPRS Support Node
GSM Global System for Mobile Communications
GPS Global Positioning System
GNSS Global Navigation Satellite System
HDFS Hadoop Distributed File System
IBN Intent-Based Networking
ICN Indoor Cellular Networks
IMEI International Mobile Equipment Identity
INSEE National Institute of Statistics and Economic Studies
IoT Internet of Things
IP Internet Protocol
IRIS Islets Regrouped for Statistical Information
IRLS Iteratively Reweighted Least Squares
ISP Internet Service Provider
KMO Kaiser-Meyer-Olkin
M2M Machine to Machine
MDAF Management Data Analytics Function
MEC Mobile Edge Computing

MINRES Minimum Residuals
MIGP Mixed-integer Geometric Program
MLE Maximum Likelihood Estimation
MMS Multimedia Messaging Service
MME Mobility Management Entity
MNO Mobile Network Operator
MSC Mobile Switching Center
MVNO Mobile Virtual Network Operator
NAS Network-attached Storage
NDA Non-Disclosure Agreements
NSA Non-stand Alone
NWDAF Network Data Analytics Function
OD Origin-Destination
PA Parallel Analysis
PC Personal Computer
PDF Probability Density Function
PDP Packet Data Protocol
PDN Public Data Networks
PGW Packet Data Network Gateway
QoS Quality of Service
RADIUS Remote Authentication Dial-In User Service
RAN Radio Access Network
RAT Radio Access Technology
RCA Revealed Comparative Advantage
RIC RAN Intelligent Controller
RNC Radio Network Controller
RNIB Radio Network Information Base
RRC Radio Resource Control
RSCA Revealed Symmetric Comparative Advantage
SA Stand Alone
SED Squared Euclidean Distance
SHAP SHapley Additive exPlanations
SLM Spatial Lag Model
SMS Short Message Service
TA Tracking Areas
TCP Transmission Control Protocol
UDP User Datagram Protocol
UE User Equipment
UGS Urban Green Spaces

VoIP Voice over IP
VoLTE Voice over LTE
XGBC XGBoost classifier
ZSM Zero-touch Network and Service Management

1 Introduction

From the list of disruptive technologies that appeared over the last few decades, it's undeniable that smartphones are one of (if not the) biggest and most impactful. Their pervasiveness has shaped behaviors and routines worldwide, becoming one of the most fundamental consumer electronics for many, taking the role of the essential Personal Computer (PC), allowing people to communicate and connect themselves to anyone in the world, organize their entire lives, access any information, content or entertainment, with no restrictions to where they are. Their widespread availability is enabled due to their internet connectivity, made possible by the wide spread of mobile networks, allowing seamless smartphone connections anywhere and anytime.

The pervasive nature of smartphones and mobile networks wasn't always guaranteed: decades of research and development by numerous researchers and organizations have brought these technologies to this point. Before becoming smart, mobile phones were limited to making phone calls and functioning as portable landlines. As their commercial spread and popularity grew through the 1980s and 1990s, more and more technologies were integrated into mobile phones. New sensors, functions, and communication methods were added, eventually giving rise to the concept of personal computers on the go.

1.1. The evolution of smartphones and mobile networks

Together with the evolution of mobile phones, Radio Access Technology (RAT) also grew in capabilities and complexity. In fact, the first generation (1G) of commercial mobile networks released by a Mobile Network Operator (MNO) was capable of basic communications, with analog audio signals being transmitted through a network of cells distributed across space, each utilizing its own transmitters and receivers. This was a differentiating factor to mobile radio systems, which would always connect to a centralized operator, instead of distributed cells. This simplicity did not last long, as soon new features were added to mobile networks leading to their second generation (2G), which saw a few different standards being used worldwide, perhaps more importantly the Global

System for Mobile Communications (GSM). Perhaps the most vital change was moving from analog to digital communications. By having voice communication transmitted as a digital signal, the benefit of having digitally encrypted conversations between the mobile phone and the Base Station (BS) emerged, raising the privacy of the network. This also led to better utilization of the limited radio frequency spectrum and enabled the first sort of text message communication in mobile networks, the Short Message Service (SMS), later joined by the Multimedia Messaging Service (MMS), which allowed for messages containing media other than text (such as pictures). After some time, small improvements in the standards were made enabling higher data rates and making it possible to perform packet switching operations, giving an early form to access the Internet through mobile phones compatible with Enhanced Data Rates for GSM Evolution (EDGE).

As operations in the mobile network became more data-intensive, soon increasing data transfer capacities of the network became the main goal, leading to the third generation (3G) of mobile networks. While the later releases of 2G (e.g., EDGE) mobile networks were already enabling early internet access through integrated browsers in certain phones, those connections were mainly associated with slowness and difficulty. The release of 3G networks aimed to establish themselves as a proper solution for convenient mobile internet access to users, with even the early adoption of 3G USB dongles that could be connected to computers to provide connection to the internet at homes without access to a fixed Internet Service Provider (ISP).

The chase for higher data capacities continued as the main trend through the fourth generation (4G) of mobile networks. As the previously mentioned relation between the evolution of mobile networks and phones, 4G showed itself as a solution for problems that appeared through the first generations of smartphones. As users' demand shifted from *on-the-go phones* to *on-the-go personal computers*, newer functionalities and applications raised the data demands of consumers, especially with the rise in popularity of video streaming services and social network applications based on the form of media sharing and consumption. While 3G networks allowed good connectivity to traditional web pages, the amount of media generated by the internet rose above its capabilities, so 4G rose to address these deficiencies. Through the 4G expansion, users saw a significant leap in data transfer capabilities between the mobile networks and their smartphones, allowing the on-the-go consumption of heavier types of media, such as video streaming and video calling. For the first time since their conception, mobile phones started to challenge not only personal mobile computers, but for many people they became the only computer needed (impacting even the market of PCs and laptops[1]).

[1]More details about how Smartphones impacted the PC market can be seen on: https://www.weforum.org/agenda/2016/04/4-charts-that-explain-the-decline-of-the-pc/

As an ever-lasting chase game, the evolution of smartphones together with the evolution of content consumed in applications leads to innovation and evolution in mobile networks. As of the time of this thesis, the world is experiencing the rollout and expansion of the fifth-generation (5G) mobile networks. While through the first few generations of mobile networks, the main headline had always been the increases in bandwidth capacity, allowing mobile networks to finally challenge the capacity of cable networks, one of the main driving forces for 5G was the idea of massification of devices connected to the network. Besides the widespread of smartphones, where the number of mobile subscribers is even surpassing the world's population [10], the rise in popularity of Internet of Things (IoT) and Machine to Machine (M2M) communication across space also occurred, especially in industry-related tasks and smart cities. This means more devices were connected to the network than ever[2], which exposed problems with massive access from the previous generations (as IoT and M2M also utilized the still available 2G and 3G networks), which should be addressed by newer 5G capabilities. Another new feature is the reduced latency experienced by users, which should mean less delay for many time-sensitive tasks (e.g., edge computing, cloud gaming).

The release of 5G is still an ongoing affair: as of the time of this thesis, most commercial MNO has released solely its Non-stand Alone (NSA) variant, which introduces a new access layer but repurposed the 4G core (to make its release more cost-effective for operators). But, while 4G is still the leader in subscribers, the adoption of 5G has steadily grown through every region of the world, with an expectation of 5G accounting for 25% of all mobile data traffic in 2023, and the number of 5G subscribers expected to exceed 5.3 billion by 2029 [10]. The capabilities of 5G will be further expanded through its Stand Alone (SA) release, which shall utilize an exclusive network core and higher frequency bands to maximize its capabilities. This shall enable higher traffic exchanges anywhere on the streets, with more modern computational tasks being performed outside of devices (e.g., offloading AI tasks and the graphical processing of games to external servers, instead of processing those in the smartphone), and more users and devices being connected and exchanging big amounts of traffic simultaneously.

1.2. Leveraging mobile network data for multi-domain research

Through the decades of advancements in the technology of mobile phones and networks, and their expanding role in daily lives through those years, the volumes of data generated on mobile networks grew steadily. Indeed, as observed in Figure 1.1,

[2]Intel anticipates over 50 billion IoT devices connected to 5G networks over the next few years: https://download.intel.com/newsroom/2021/archive/2017-01-04-editorials-intel-accelerates-the-future-with-first-global-5g-modem.pdf

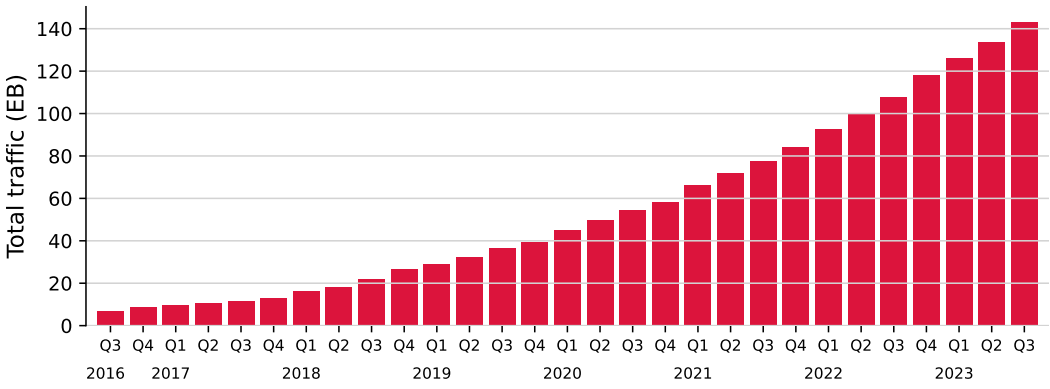


Figure 1.1. Quarterly mobile network traffic consumption, in exabytes (EB). Adapted from [10].

the estimated per-quarter data traffic consumed in mobile networks has grown at an exponential rate, from less than 10 EB in Q3 of 2016 to more than 140 EB in Q3 of 2023. This shall not be an unexpected trend, since as mentioned before, with every new generation of mobile networks, the most advertised new feature tends to be *faster data rates*, leading users to be every day more and more comfortable consuming any type of data in their mobile phones, even when they're outside of a Wi-Fi connection. Indeed, with the last releases of 4G and current releases of 5G NSA, most users may not even note differences in the quality of their connection versus the one they may have at home in their fiber network over Wi-Fi. With this gap ceasing to exist, it's only expected that traffic consumption levels over mobile networks will remain growing for the next few years, with Ericsson reporting a continuous growth in monthly consumption until 2027, with an expected stagnation only occurring after 2028[10].

Interestingly, through the years of growth in the data volumes flowing through mobile networks, MNOs have always been *passively collecting* information about data from their network, either for their own internal research, e.g., to dimension correctly their systems, or for billing purposes. Indeed, when calling and SMSs were the dominant operations occurring in the network, the operator always had to register the locations of the caller and receiver, as well as the duration in order to bill correctly the phone call (as it's been always done with fixed phone lines). When data-oriented products became the norm, the MNO also kept their interest in tracking how much data each customer was consuming since most mobile data plans had limits, and therefore this measure needed to be constantly done to be sure data contracts were being enforced correctly. The implication is that, for most MNOs, collecting and processing data was always certain.

It did not take long for researchers to see the potential of those rich and complex datasets that could be *relatively easily* be collected, as they could be used not only for network improvements and billing customers, but as a way to understand individual users and populations, either from the perspective of understanding the utilization of mobile

phones and their applications or by simply utilizing the measurements as a proxy to understand where people are located and how they're moving.

From this new source of large-scale datasets (which can easily be above the terabyte level) and the realized potential of its use for research, there was the birth of the need to understand how to properly collect and process it, as well as which tools could be used for analysis and proper explanations were needed, leading to the rise of the field of *networks data science*. Studies about the dynamics of computer networks are not a recent field[3] and have been an active topic in many communities and popular through many of the main journals and conferences in the area.

With the expansion of mobile phones and networks, it was unsurprising to see new studies on the dynamics of these growing networks. Researchers began to describe the dynamics of mobile traffic datasets collected from thousands to millions of users across various locations (sometimes spanning countries and continents) and over months or years. Other fields quickly recognized the potential of this data: as it's passively collected from large populations, it sparked an interest to understand if it could be an alternative to traditional census, survey, and interview data collected by sociologists, geographers, physicists, or epidemiologists. While deeply descriptive, the data from those traditional surveys usually required massive efforts[4] to be properly collected, with a continuous collection of such statistics being considered hard. Indeed, it did not take long to see research utilizing mobile network data expand rapidly: from the first works seen in 2006, the annual growth rate of new publications was 90% by 2016 [11].

Through those years, different datasets emerged. Due to the different products offered by MNOs over the years, a variety of events and statistics could be observed according to what was being utilized by customers. For example, due to the call and text messaging orientation 2G and 3G, the first collections involved Call Detail Record (CDR). Those detailed information and statistics about telephone calls and SMSs done by each customer across space and time. This simplicity in collection against traditional data used in the study of social and complex (human) networks was one of the positive arguments for mobile network data. As it was already being passively collected by the operator for billing purposes, little to no extra technical cost was needed for MNOs to pass down this data to interested researchers, with only the necessity for research agreements to be put into place. Also, by having independent researchers studying it, MNOs could easily benefit from insights gathered about their customers and even improve their services from the results of this cooperation.

As the network evolved, CDR data became less desirable due to the decline in

[3]Indeed, by searching through proceedings of either IEEE INFOCOM or ACM SIGCOMM during the 1990s and early 2000s, many works can be found, providing measurements and analytics about computer networks at many different points. Citing those here would steer away from the topic of this thesis, but an interesting reader can easily find many of those initial works still available in online repositories.

[4]Usually involving entire nations, such is the case of Census surveys.

traditional phone calls and SMSs, replaced by Internet Protocol (IP) equivalents like WhatsApp and Telegram for messaging. Additionally, standards like Voice over IP (VoIP) and Voice over LTE (VoLTE) began transmitting voice communication over the internet, similar to other online tasks. This increased interest in analyzing pure traffic flows, focusing on the overall data moving through the network (as opposed to the number of calls or events), using different IP standards like TCP, UDP, and QUIC. This data source is enriched by techniques like Deep Packet Inspection (DPI), which can label the mobile applications generating each traffic session. This allows for the study of the complex dynamics of different mobile applications, enabling custom solutions for mobile operators and researchers to explore spatiotemporal variations in app usage. This creates an unparalleled data source for sociologists and other modern humanities studies, as each person interacts differently with their smartphone, using different apps at various times and locations. This has sparked a new era of research.

In addition to traffic sessions and CDRs, another valuable data source from mobile networks comes from signaling events between mobile phones and base stations. These communications occur even when users aren't actively using services, helping determine the best base station for their location. This is done passively and preemptively: as a user expects his phone to always be ready to make a call to access the internet, those exchanges between mobile phones and the network are done automatically and, in many cases, happen with fairly low temporal granularity (in the order of seconds or even milliseconds)[5]. By adding an extra step of recording those signaling events, it's possible to derive location and mobility-related data sets, such as estimated home locations, origin-destination matrices, and trip distances, on a *per-user granularity*[6] or even aggregated over statistical regions. As mentioned earlier, with the right methodologies, this rich and complex data offers many possibilities for researchers in transportation, epidemics, and other fields interested in human mobility, as traditional mobility surveys were very resource-intensive to collect.

Another important detail is that, within the last few years, privacy regulators have become more aware of the impacts of collecting per-user data (not exclusive to mobile networks). This can be noted in Europe, where the measurements presented within this thesis are collected, with General Data Protection Regulation (GDPR) laws restricting many collections of per-user data. This leads to individual records becoming a rare sight, with researchers having to adapt their methodologies to work utilizing data with higher levels of aggregation. In mobile networks, this means that the data will be commonly aggregated on the antenna level, i.e., instead of knowing how much a certain user utilizes an app, it contains how much traffic of an app is seen on a certain BS. This increase in

[5] As the MNO is not only administrating its physical resources but also the radio-frequency resources and the demands across those can widely vary due to demand, user location and by simple nature events.

[6] This, of course, raises privacy-related issues, which are discussed during Chapter 3.

privacy regulations did not decrease the interest in mobile network data, as it can still be a more privacy-oriented alternative (e.g., against Global Positioning System (GPS)). Therefore, it's important to develop methods able to extract useful statistics about populations from mobile network measurements, considering the aggregation levels to be performed over time and space.

1.3. Contributions of the thesis

The work presented in this thesis aims to *advance the state of the art of networks data science*, by proposing methodologies and insights from recent large-scale measurements in MNOs. As those have the potential to be leveraged by many distinct fields, the contributions presented within this thesis navigate through a few different applications. Those shall all come together to help build a holistic perception of the field of network data science. The main contributions of this thesis are:

1. **Create new methodologies for the collection and processing of 5G per app traffic:** With the recent deployment of 5G NSA, a new challenge is how to properly identify the sessions per application corresponding to it, since the network core (where the majority of the traffic probes are located) is shared with 4G. This thesis proposes a way to overcome this problem, presenting guidelines and insights on how to properly handle traffic aggregated on a BS level. This contribution is described in [1], [5].

2. **Establishments of techniques for insights into application-level traffic demand:** With 5G and beyond mobile networks heavily dependent on IP networks, previous methodologies that focused on the analysis of CDR records need to be reviewed to fit the new collections that quantify data consumption. All works contributed throughout this thesis push the shift of a mentality oriented towards mobile application traffic. Among those works, it can be remarked:

    a) **Characterization of the adoption of new deployments of mobile networks:** As MNOs deploy new products over their network, an interesting aspect is to understand how customers adopt them, and more specifically if their usage patterns and app preferences are shifted. Two characterization studies about the adoption of new 5G NSA deployments and indoor mobile networks are presented. These contributions are described in [1], [6].

    b) **New methodologies and insights about land use characterization through mobile traffic:** As mobile networks become denser in urban environments, they provide higher-quality data that can be utilized for complex spatial and spatiotemporal analysis. Two techniques are presented

in characterizing land at different scales: firstly at general land uses on a city-wide scale, and secondly at a significantly smaller scale by characterizing green spaces within cities. These contributions are described in [4].

c) **Characterization of special events through mobile traffic:** Besides analyzing routine patterns, comprehending how special events may shift consumption patterns of network traffic is an important challenge, which can guide MNOs to have better preparation for such disturbances, or be utilized as a proxy for government agencies to better comprehend the impact of those events and their measures in the population. Two works are presented: one characterizing the impacts of the COVID-19 pandemic on mobile traffic, first at a nationwide scale and later at a city scale; the second one showcasing how the disturbances of traffic consumption can be a proxy to characterize massive manifestations within cities. These contributions are described in [2], [7]–[9].

3. **Characterization and modeling of transport-layer session dynamics and the generation of synthetic data:** The last main contribution of this thesis comes from the realization that while access to mobile network data can be a great enabler of research, not always this data will be available. The last work of this thesis presents a first-of-its-kind characterization of transport-layer traffic sessions, focusing on the statistical heterogeneity across mobile applications. This characterization leads to a few models that are openly shared with the community to generate synthetic traces of the measurement, with a few use cases presenting the potential of those models to help network research. These contributions are described in [3], [5].

1.4. Outline of the thesis

This thesis is composed of 8 chapters. Following this introductory chapter, the structure of the remaining chapters are organized as follows:

Chapter 2 presents the *Background works*, which contextualizes the field of *networks data science* for the last decade, through the different types of data that can be collected from mobile networks and the vast array of usages that these sets enabled researchers from many fields to expand their state of the art. Chapter 3 presents the *Materials and Methods*, which presents the fundamental methodologies and techniques used through this thesis for the measurement, processing, and analysis of mobile network measurements.

The remaining chapters contain the main contributions of this thesis. Chapter 4 explores *"How users adopt new mobile technologies"*. Two cases are presented: Section 4.1 presents one of the first large-scale collections of a production 5G NSA network deployed in France, looking at how the presence of 5G may have impacted the overall traffic demand.

Section 4.2 presents a first-of-its-kind characterization of mobile networks in indoor environments. Most works that characterized mobile traffic demands were indifferent about the BS being outdoors or indoors. But, as users expect a seamless connection, understanding how demand changes over indoor locations becomes an important task. This thesis showcases how the dynamics of mobile application usage vastly differ in these places when compared to outdoors.

Chapter 5 explores techniques that can help answer the question of *"How space and land use affect smartphone usage"*. Section 5.1 presents a work about the Exploratory Factor Analysis (EFA) technique, and how it can help uncover spatiotemporal patterns of mobile traffic consumption usage through two major urban centers in France. Section 5.2 reduces the spatial granularity and presents a new approach to characterize urban green spaces in cities through their influence on smartphone and application utilization.

Chapter 6 looks into *"How spatial events impact the network"*, and how this data can be leveraged to better comprehend populations and their relation to those events. A few works are explored: Section 6.1 explores how the COVID-19 pandemic restriction measures impacted the consumption of mobile traffic in France at a nationwide scale, both in the temporal and spatial dimensions. Section 6.2 poses similar questions, but zoom into the main urban cities in France to understand the impact within neighborhoods, and how those changes can be associated with the socioeconomic indicators within the cities. Finally, Section 6.3 loos into city-level, but proposing a characterization of the impact of large-scale manifestations in mobile networks, and how the insights gained from this analysis can be utilized to track the spatiotemporal progress of protesters, as well as their behaviors during the protest.

The final set of contributions of this thesis is seen in Chapter 7 about *"Leveraging data from smartphone usage for smarter networks"*, presenting a first-of-its-kind characterization and modeling of transport-layer session traffic datasets. Section 7.2 focuses on the characterization part, more specifically on how heterogeneous the traffic trends are amongst apps, even the ones within similar categories of data. Next, Section 7.4 presents a few mathematical models of those observed behaviors, which are shared with the community so anyone interested in replicating this data synthetically can easily do so. Finally, a few use cases are presented in Section 7.5 to highlight how the insight gained from this analysis can better guide network engineering research, by having models that are closer to reality than alternatives currently available.

This thesis wraps up in Chapter 8, which discusses the main takeaways and insights gained from the contributions presented through the chapters, as well as a few potential directions to be explored by future works.

2 Background

The pervasiveness of smartphones showcases their success as one of the most popular technologies of the XXI century. With an ever-growing number of users relying on them for both their personal and work lives, smartphones went from a luxury to a commodity, integrating themselves into the majority moment of moments of the daily routine.

As the differentiating factor of smartphones comes from being personal computers on the go, users expect that MNOs will provide ubiquitous connectivity, in any outdoor or indoor environment. With the advancements of RATs, the capacity and capabilities of mobile networks grew vastly, with users relying every day more on them for communication and access to internet services. These advancements, together with cheaper and accessible data plans, made users comfortable with using their smartphones attached to the mobile network anywhere.

As more users utilized more their devices connected to mobile networks throughout the years, the data passively collected MNOs grew in volume, complexity, and quality. This prompted researchers from multiple domains not exclusive to Network Engineering to become interested in testing the potential of those collected data sets in their field, spanning an impressive range of interesting studies that go way beyond traditional networks and telecommunications venues.

This expansion in multi-domain research relying on mobile network measurements has resulted in a few interesting surveys that characterize the fields and directions in which those sets are utilized [11]–[13]. Specifically, the work of [11] presents an interesting overview of the multiple disciplines outputting unique research from this source of data. According to them, mobile traffic measurements span 3 major fields and sub-fields of research, as seen in Figure 2.1: *Network analysis*, *Social analysis* and *Mobility analysis*. Those fields provide a heterogeneous background, due to the vast amount of disciplines interested in the insights that can be obtained from this type of data.

The aim of this Chapter is to present an overview of the state-of-the-art research utilizing mobile network measurements through those proposed fields. It's important to note that the contributions of this thesis will not span through all of the presented fields;

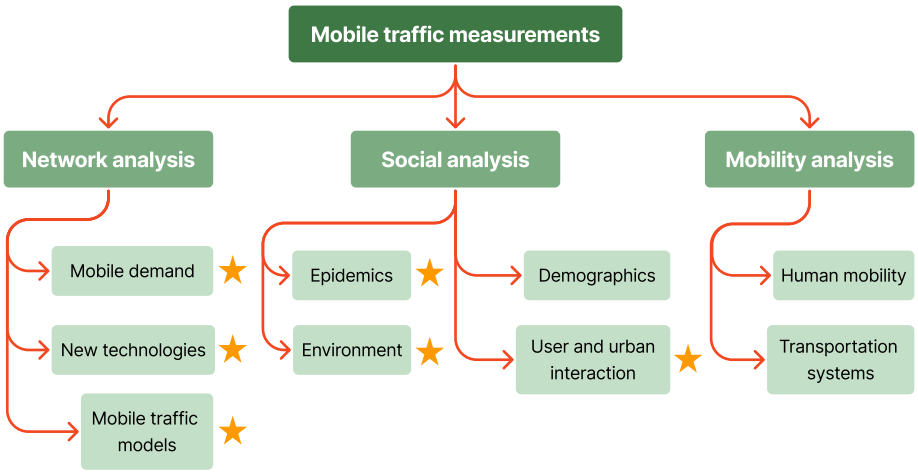


Figure 2.1. Classification of studies leveraging mobile traffic measurements in the literature. The sub-fields where this thesis contributes are marked. Adapted from [11].

the specific sub-fields where contributions are made are marked on Figure 2.1, as those will have more in-depth attention in order to present a coherent background to motivate the contributions to be presented through the later chapters.

The structure of this Chapter is as follows: Section 2.1 will go through papers that utilized measurements from MNOs for insights and improvements towards the mobile network itself; Section 2.2 will discuss the social aspect of mobile network data and how it can be utilized by many different areas who try to understand individuals; Finally, Section 2.3 will present the mobility side of this data, focusing on understanding how people move throughout the space by utilizing network measurements.

2.1. Mobile data analysis for network optimization

The most direct contribution of large-scale mobile network measurements is to understatement, analyze, and improve the network. Insights gained here come from understanding how production deployments are performing, the usage of mobile services through time and space, as well as producing real and synthetic data sets to test new algorithms to optimize the network. Therefore, works within this Section will go into the technical aspects of mobile networks.

Following the definition of [11], an adaptation of the sub-fields of *Network analysis* is seen on Figure 2.1. They will be related to the utilization of mobile network data for the improvements and understanding of networks, and encompass: 1) *Characterization of mobile demand*, where studies focus on characterizing the dynamics of traffic consumption, from both user and network perspectives; 2) *Impact of new technologies*, where the insights gathered by the data can be used to evaluate new technological solutions for mobile

networks; 3) *Modeling mobile traffic*, where measurements are used to characterize traffic patterns in order to generate synthetic models. Next, a few of the seminal works related to each of those sub-fields will be explored.

2.1.1. Characteization of mobile demand

Temporal patterns of consumption: When studying the overall demands of traffic consumption by users over the mobile network, the initial approach is to understand how traffic demands are characterized across the temporal dimension. Studies of temporal structures in mobile traffic can be as simple as straight comparisons of traffic volumes across seasons [14], or computing statistics about how many calls are happening over time [15]. These explored approaches in early mobile network literature allow appreciation of how planned and unplanned events induce significant variations in the typical temporal structure of the mobile traffic demand, but investigations at a finer granularity require more elaborated methodologies. A more composite method for temporal analysis of mobile traffic consists of building an expectation from historical data and labeling time slots deviating from the statistics as anomalous [16]. Those approaches are purpose-purpose, i.e., outlier detection, and have the limitation of not finding any hidden structures which would be desirable in more in-depth current studies.

Closer to the goal of revealing complex hidden structures are works that aim to create profiles of temporal network activity [17]. In this case, snapshots of the mobile traffic demand in a target region collected at different time slots are clustered via a dedicated distance measure. This allows the unveiling time periods with similar patterns in the spatial distribution of traffic. Recently, tools from spectral analysis have also been adopted to achieve a similar goal, on a per-service basis [18]. However, both approaches above rely on traditional clustering algorithms, hence forcing each time slot into a single specific category. This loses nuances in the data, e.g., during time periods where the spatial traffic is in fact at the boundary of two or more behaviors. Moreover, the approaches do not provide information about the root causes that lead to the resulting clusters, whose interpretation is thus left to the expert knowledge of the system.

Spatial patterns of consumption: A second direction when understanding mobile traffic demands is to understand where traffic is being consumed. To achieve this, a larger body of works addresses spatial structures in mobile traffic, focusing on urban settings due to their rich traffic patterns, and proposing multiple techniques to distinguish neighborhoods and spots of the city by their mobile traffic demands. Specifically, many studies start exploring spatial structures from the built work on temporal structures by collecting mobile traffic time series in different geographical zones, compressing them into normalized signatures that summarize the observed temporal patterns of traffic, and adopting clustering algorithms to group the zones based on their signatures [19]–[23]. Spectral methods have also been used to define signatures of routine activities and

deviations from them, with the same objective of using them for spatial clustering [24].

Information theory tools have been used to group large statistical zones across a whole country based on the local consumption of mobile services [25]; also, techniques from signal analysis relying on eigen-decomposition have been adopted to categorize buildings in a university campus based on the observed Wi-Fi traffic [26]. These methods focus on very large (i.e., nationwide) or localized (i.e., a single campus) geographical regions, and may miss the dynamics of exploration of the urban spaces occupied by modern cities; moreover, they still aim at generating rigid associations of zones to behaviors with little explainability, as in the legacy clustering strategies above.

Trying to explore more complex structures in both the *spatial and temporal dimensions*, some researchers have explored the use of factor analysis for the study of mobile traffic data and, utilizing both CDR data [27] and traffic flows [4], which showcase the co-dependence of space and time within mobile traffic consumption. Seeking to reveal complex spatiotemporal profiles of mobile traffic consumption, a gap will explored within Section 5.1, where a work will be presented contributing to a factor analysis technique that can showcase mobile traffic consumption patterns across time and space.

Patterns in mobile application consumption: A useful study for MNOs is understanding the adoption of mobile applications by users. This means finding insights about how those apps request their traffic demands to the network, as well as how users utilize those. This is especially useful for heavier loads and popular apps, which may necessitate special attention by the operator.

While the works shown so far focused on the dynamics of the total volume of traffic, a variety of works focus specifically on mobile services demands and their dynamics over space and time. For example, through the use of large-scale measurements in mobile networks, researchers were able to better comprehend how WhatsApp usage impacts MNOs [28], such as understanding that the hosting architecture of the app is exclusively located in the US (which could impact routing for non-US customers); that video sharing content is almost 40% of the demand generated by users of the app (which means that treating it as a simple instant messaging app may lead to incorrect resource allocation) and that the network flow characteristics depend on the smartphone operating system. A similar analysis of mobile network measurements was performed for WeChat [29], which also led to a better understanding of the impact of media-related messages (photos, videos) and uncovered different clusters of users that have significantly different demands when utilizing the app.

Another important class of studied applications was video streaming apps. Through their rise in popularity over the past decade, and with the higher capacity of 4G and beyond RATs, there was an increase in usage of those services over mobile networks and, due to their type of content, they quickly became one of the heaviest generators of mobile traffic. For example, an early work studied the user-side patterns of traffic of Netflix,

YouTube, and Hulu generated from both Android and iOS devices [30]. They observed that those services generated many redundant traffic flows, which negatively affected the mobile network resources. By also analyzing how the same set of applications have their traffic distributed through CDNs, researchers have also observed that assigned the CDN to deliver content to the user without considering the network conditions, which impacted the resource optimization for the network operator[31]; solutions for this involved an adaptive CDN selection considering the network resources to optimize the users' bandwidth. Such findings are important in the sense of understanding how modern applications actually operate and how both their mobile application and network infrastructure interact with the network operator.

Another category that saw surges in popularity leading to the study of their impact is Cloud services [32], since the system demands of those apps are quite unique and rely heavily on the uplink side of the chain, and have become more popular as alternatives to traditional in-device storage.

In essence, when studying the vast ecosystem of mobile applications and their traffic demands over the network, many works show how those apps act in unique ways [33]–[35], and current-age network resource optimization needs to take into account this uniqueness in their consumption patterns. Another set of interesting findings was a number of properties of service-level demands, such as the fact that they exhibit locality [36]–[38], temporal patterns that are diverse across applications [34] and possibly easy to predict [39], as well as consistent user bases [40], [41].

However, a gap in research appears due to the evolutionary nature of mobile networks: all these works employ demands collected in 3G or 4G networks and do not look into the adoption of one technology over another. This gap is addressed in this thesis in Section 4.1, which performs a study on the adoption of 5G on the application level. Also, whenever new RATs are deployed, a change in traffic consumption patterns could be expected, which will be the topic addressed next.

2.1.2. Impact of new technologies

Adoption of new RATs: Within the context of the deployment of newer RATs throughout the world, understanding their performance in the wild becomes a critical initial task for the operator, allowing it to understand if its clients are interacting differently due to the new RAT, as well as to diagnose the performance and quality of experience metrics of the network access. Also, comprehending those insights becomes critical in order to magnify gaps that need to be addressed.

For example, through the ongoing worldwide deployment of production-grade 5G networks, understanding their actual performances, as well as how users and their smartphones are dealing with the new technology became a critical task that can be explored through the study of network measurements. The majority of newer studies

have investigated the deployment and performance of 5G in deployments in two of the earliest adopters: the United States and China. As is the case with early studies, their focus is research built upon client-side measurements, targeting the evaluation of coverage, latency, energy consumption, or protocol operations observed at the level of individual 5G devices, usually providing results that employ 4G as a reference. Several of these measurement analyses target specific scenarios, such as Non-Standalone (NSA) and Standalone (SA) deployments [42], mmWave communications [43], multiple operators [44], [45], diverse urban environments [44], bus transit system [46], or high-speed trains [47]. It's important to note that those are not necessarily direct network measurements, but instead, client side, which means understanding what individual users are seeing, and not what the network is indeed observing.

These works generally acknowledge the much-improved throughput of 5G compared to 4G, yet also highlight a number of issues, including sub-optimal latency and high power consumption with respect to 4G [48], exceedingly aggressive strategies for the migration of radio resources from 4G to 5G [49], non-ideal support for high mobility scenarios [47], or inadequate resource management policies [45]. Some works also propose solutions to some of the problems above [45] or optimizations of applications to best utilize 5G features [42].

These early works in 5G measurements focused much on client-side studies, but few so far have been able to explore the network's perspective, analyzing traffic measurements collected by a major mobile operator in a nationwide production-grade infrastructure. This offers a different viewpoint and allows an analysis of the spatiotemporal demands generated by the whole user population, which is hard to obtain from smartphone measurements. A study of the impacts of 5G deployment on the network operator side was performed on an operational 5G NSA network in the UK [50], focusing on understanding the temporal evolution and characteristics of the local deployment, the diversity of the ecosystem of 5G-enabled devices, and the overall network performance.

To bridge this gap, the work presented in Section 4.1 contributes to the literature by providing one of the first large-scale country-wide characterizations of the adoption of a 5G production network over space and time, and as mentioned focusing on individual services. Most importantly, neither spatiotemporal nor application-level aspects were considered by previous network-oriented works.

Adoption of indoor mobile networks: Another newer trend in networks is the deployment of antennas and BSs within indoor environments. Users expect to use their devices no matter where they are. But, not all places may have available Wi-Fi connections, leading users only with their mobile network. Due to the absorbing nature of construction materials, the signal strength of outdoor antennas is severely reduced, limiting the connectivity in indoor environments. Therefore, networks deployed in indoor environments are becoming crucial for guaranteeing the quality of experience of users.

However, the majority of the research so far mentioned in this Chapter does not

distinguish where antennas are in relation to buildings. Even though some works may unveil spatial patterns of traffic consumption that could be related to indoor environments (e.g., peaks in commuting hours near metro stations [27]), those insights are gained by just analyzing which city features are close to the BS, with no distinction made about whether those deployments were indoors or outdoors. In fact, many other recent works do explore those distinct patterns of network usage through space [51]–[53], but neither discriminates the environment of the deployment.

This discrimination can in fact be made, as network operators have to account for the environment and record this in their data sets (as it will be later seen in Subsection 3.6.1). But, the literature in characterizing indoor environments is still quite limited. Few works compare the characteristics of outdoor BSs and fixed network demands [33], [54], revealing differences in terms of packet, flow, and session-level statistics, as well as in the temporal traffic patterns. However, the operation of fixed networks is fundamentally different from mobile networks, and thus the conclusions of those works cannot be straightforwardly extended to indoor mobile networks. The traffic from indoor mobile networks was considered in [55], exploring the mobile application utilization profiles in Santiago (Chile), through both distinguished indoor and outdoor BSs, but never fully characterizing the uniqueness of indoor mobile traffic consumption.

This gap in the characterization of indoor mobile network deployments will be further explored within Section 4.2, which will present how the temporal and application patterns change due to the deployment being indoor vs. traditional outdoor ones.

2.1.3. Modeling mobile traffic

The insights gathered from the study of mobile network data can be leveraged to model and evaluate solutions to optimize the network. For the validation of solutions within the network, data sets may carry technical specifics in relation to which measures may want to be optimized within the network (e.g., latency, packet drops). Therefore, a significant amount of work was put into characterizing measurements within the network, creating models describing those phenomena, and applying this data for optimization.

When modeling BS-level demands, understanding how to leverage temporal structures for traffic models becomes critical. BS-level statistics mainly describe aggregates of the traffic volume across all devices associated with the target antenna and are best characterized over timescales of minutes or hours. As such, they are different and coarser than the session-level dynamics, which instead occur at order-of-second granularity. Such examples of modeling and generating temporal traffic involve works employing α-stable distributions to model heavy-tailed samples of BS-level traffic [56]–[58], or recent generative neural networks that mimic BS-level dynamics over space [59], time [60] or both [61], and per service dynamics [62].

Another direction aims to model *statistical properties of mobile traffic within each*

session, i.e., at packet level. Such models typically operate at timescales of milliseconds or less, and provide analytical formulas for inter-arrival times between consecutive packets or requests from a device [63], sizes of individual files or number of packets per frame [64], intervals for deterministic reporting [65], or duration of activity and inactivity periods [66]. Different packet-level models are specified for broad classes of services like web browsing, video streaming, voice over IP, gaming, downloads via FTP, or machine-type communications, among others. But, there's still a lack of more application-specific models, which as mentioned before becomes a critical aspect for network operators to understand, as app dynamics even within the same category can be heterogeneous.

There's a vast amount of proposals for packet-level models for 5G systems, and an in-depth look at their proposals and limitations can be observed in a recent survey [13]. The key observation is that these models are primarily designed for the evaluation of low-layer technology in stationary environments, and do not capture, e.g., inter-session timings, how long a session generated by a given application persists in a BS, or how much traffic it generates there.

In the field of characterizing sessions at the transport-layer level, six models of individual mobile traffic consumption were developed [67] by classifying the demands generated by 6.8 million subscribers based on their temporal patterns and amount of data usage. In a similar spirit, six major temporal profiles were identified in the weekly demand generated by mobile devices [68], and predictors were proposed to anticipate the future load of each class of user. But, the granularity of those models could be considered coarser than desirable. First, they only consider overall user-level traffic. Second, they only consider the total traffic of each user, leaving a significant gap in relation to per-service models. Third, they're purely temporal and aggregate information over all BSs visited by each user, with no focus on behaviors recorded within a single BS. It can be argued that models with finer granularity, richness in application dynamics, and per-BS viewpoint could be more informative for the validation of mobile communication technologies and systems.

There's a significant gap in the characterization and modeling of transport-layer sessions, especially in relation the having models that take into account the heterogeneity of traffic generation across mobile applications. To address this, Chapter 7 will contribute to expanding the literature on transport-layer mobile traffic models.

2.2. Mobile data analysis for social science research

The collection of social data at a large scale has always been quite an extensive, difficult, and expensive task. Classically done with surveys through interviewers

questionnaires[1]. On smaller scales, those collections are done through population sampling, which while well studied is hard to guarantee to be free from biases.

Through the dissemination of the internet, new ways to collect social and demographic data originated. Their collection was significantly easier, especially when compared to surveys, as those new sources (such as mobile network data) can be passively collected through large populations (i.e., scale of millions). In the case of mobile traffic, not all the information usually required by traditional surveys is available. Because of this, it's not uncommon for it to be cross-checked and mixed with traditional surveys. This mix has enabled an interesting variety of new studies, which focus on understanding societies through their interaction with smartphones and mobile networks, how to derive statistics usually obtained from traditional surveys with mobile phone measurements, and mixing with classical survey data.

Based on [11] and as seen on Figure 2.1, four main sub-fields are present: 1) *Demographics*, where the ways of use of smartphones are related directly to sociodemographic surveys, in order to understand how parcels of the population may utilize their devices differently according to their societal context; 2) *Environment*, where the spatial and temporal structures of smartphone usage are related to the environmental aspect where users are; 3) *Epidemics*, where human movements and interactions can be leveraged to understand how diseases spread and affect populations; 4) *User and urban interactions*, where the interaction between users can be observed through mobile traffic consumption. Next, a few of the seminal works that utilize mobile network data within those sub-fields will be discussed.

2.2.1. Demographics

Measurement data obtained from mobile networks can be used for sociology research as a proxy to understand how the societal context of populations can affect smartphone usage, and on the opposite direction, how the study of smartphone usage can also give insights into the demographics of populations.

Perhaps the most direct approach is to utilize data to simply comprehend general statistics about the population. An early approach showcased that it was possible to note differences in the utilization of smartphones through genders in Belgium [69], and the possibility to determine the age and gender of users in Mexico through their mobile phone usage patterns [70]. Other works have also expanded the number of sociodemographic features possible to be determined through smartphone usage patterns, such as the income level and residential area [71] and even ethnicity [72].

Mobile network data was also utilized to predict the poverty index at fine spatial granularity in Senegal, together with environmental data, and obtained great results

[1]Perhaps the most widely known large-scale version are Census surveys, a country-wide collection trying to cover as close as possible to the absolute totality of its population.

when compared to traditional survey data obtained from the local Census [73]. Data from smartphone usage was also utilized to train models that can recognize patterns of poverty [74] or to create poverty maps [75], which can help organizations to better target humanitarian help. Another research approach was to utilize the diversity of income, educational attainment, and inequality as a proxy to infer the consumption of certain mobile applications, and the traffic of those apps could even be used to predict the socioeconomic status of areas in France [76].

Mobile network data can be also leveraged to understand the access to electricity and digital divide [77] and the electrification rates [78] in African countries, where the collection of classical survey data about this topic was previously considered difficult. It can also help understand the later stages digital divide in countries with higher penetration of smartphones, such as France [79]. Studies using mobility data collected from mobile networks also were able to showcase that diversity of mobility has a significant correlation with the socioeconomic status of users and it's the highest important feature in socioeconomic predictive models [80].

2.2.2. Environment

Dynamics and characteristics of cities: Besides understanding the demographic aspects of populations, the utilization of smartphone data can help understand the geographical and social environment, providing newer aspects of such spaces that were not previously observed by classical sources of data.

In an initial example, mobile network data helped unveil the geographical properties of individuals that are part of communities, helping researchers to understand the span of such communities through space [81]. The urbanization levels of spaces can also be observed, such as the differences in smartphone usage between urban and rural users, where users in urban spaces communicate more times, but users in rural spaces communicate with others for longer periods of time [82]. As expected, through different regions of the city, it's possible to determine high-activity places by understanding where mobile traffic is mostly consumed, with a correlation between traffic consumption and the nature of the environment [36]. Indeed, it's possible to determine different neighborhoods of cities just by studying their temporal profiles of mobile traffic consumption, where researchers were able to determine five categories: residential, commercial, industrial, parks, and others [20]. In general, the utilization of smartphones has been unlocking interesting insights that can help guide urban planners and geographers in better understanding the geographical limits of cities and urban environments, and how complex those can be besides pure census lines [83]

Large-scale events and manifestations: While many studies go through general environmental and land use identification, others may target more specific events across space. With the rise in popularity and dissemination of smartphones and mobile network

access through the population, their usage through urban environments had a significant impact on how humans interact and record such events.

Due to the essential role of smartphones in how decentralized civilians can organize, coordinate, and report in real-time large-scale protests, the study of their impact has become an important multi-disciplinary topic. Studies of geo-located tweets across 16 countries during the Arab Spring protests show that rises in activity related to protests on Twitter correlate with manifestations happening on the next day [84]. Indeed, the analysis of social media platforms shows that those act as a facilitator for information exchange and the organization of protests, even with the way information is disseminated through them influencing the success/fail rate of the organization efforts [85] and serving as a proxy do determine repression of the manifestation [86].

With this in mind, many works have explored the potential of observing and characterizing large-scale events in cities through the lenses of smartphones and social media, and how this data can be utilized to create models, and analytics and forecast those events. The data collected by mobile operators can help detect such urban anomalies and event attendance. Overall, it's possible to determine the type of event users are attending through measurements of mobile networks [87], or to utilize signaling data to determine the adherence of mass manifestations [88], or even to determine how much the attendance of family holidays is affected during conflicting times, such as during elections [89]. Other types of disturbances can also be detected through data, such as the prediction of transit congestion [90] and road accidents [91]. For more information about the different types of data utilized and models for urban anomaly detection, the following survey is recommended [92].

Still, the majority of works presented within the characterization of manifestations focus usually on single applications (especially Twitter), and may not be fully exploring the complex ecosystem of applications that can be unlocked through per-app mobile traffic. This will be addressed with a contribution presented in Section 6.3, which will give a full characterization of the spatiotemporal sensing of protests and the preferences of mobile applications during it.

Characterization of parks and green spaces within cities: The data generated by mobile networks can also help identify usages of very specific spaces within the cities. For example, it was possible to assess park accessibility and activity information on those spaces in Shanghai utilizing mobile phone data to assess geographic and activity-based inequalities [93]. Spatial disparities in park visit flow and duration were also observed, allowing parks to be classified based on environmental factors and facilities [94].

Furthermore, other works have employed mobile phone data to improve estimations of park catchment areas, revealing variations in park usage based on park size and visitor spatial patterns [95]. It was also possible to evaluate seasonal variations in park visitation and the correlation between visiting patterns and park attributes, highlighting

the influence of seasonality and park features on visitor volume and accessibility [96]. Lastly, park access for the elderly in Beijing was estimated by analyzing the impact of socioeconomic status on park accessibility [97].

These studies collectively enhance the understanding of park usage patterns, mobile internet consumption, and the socioeconomic dimensions influencing these behaviors. But still, the majority of them utilize either GPS data or signaling events from the network. A gap is addressed within Section 5.2 where a full characterization of parks will be made based on how they may influence the patterns of traffic consumption, going deep into how categories of applications may be consumed differently in those spaces and how this can be influenced by the type of features a park has.

2.2.3. Epidemics

General epidemics and extraordinary events: By extracting the mobility aspect of users through their connectivity within the network, it's possible to collect passively the movement and migration patterns of large populations, which is a great resource when studying epidemics and how diseases spread. For example, the use of mobile phone data in cooperation with traditional census surveys can help improve calibration and precision of general epidemic models [98], or in certain specific diseases models, such as Dengue [99]. It also enables better fine-grained simulations of the stochastic simulation of disease diffusion, such as Ebola [100], or understanding the hubs of Malaria [101].

Another possible usage of the signaling events in the network is for understanding the effectiveness of spatial-based mitigation strategies of diseases [102], as well as to describe recurrent mobility patterns which are useful for spatial epidemic models [103].

Besides epidemics, data collected from mobile networks can also help understand how disasters affect populations and how their recovery occurs [104], as well as to provide preemptive resources for government preparations for earthquakes [105]. More details about how data from mobile phones can be leveraged to understand disaster events, such as natural hazards and epidemics can be extensively seen in the following survey [106].

Impacts of COVID-19 from the perspective of mobile networks: In the past few years, one the most important usages of this type of data was in the studies that analyzed the impact of the COVID-19 pandemic on internet traffic, at different levels across the network, and also as a proxy to understand the impact and effectiveness of restriction measures imposed by governments.

At ISPs located in Central Europe, traffic increased by 15% to 20% during the initial 2020 lockdown, with much higher growth than in a typical year. Such dynamics can be ascribed to the restrictions mandated by governments, resulting also in dynamic changes in weekday patterns that started looking similar to those in weekends [107], [108]. Similar trends occurred in large ISPs in the United States, with an increase from 30% to 60% of peak traffic rates during the first quarter of 2020 [109]. Not only operators, but companies

that provide online services also noted significant traffic changes; for instance, Facebook initially observed short periods of sharp increase in their edge network traffic, with a subsequent steady increase of load; they also reported user behavior variations, such as an increase of interest in live streaming services [110]. The sharp traffic changes due to new user behaviors have been also seen at smaller scales: a 90% reduction in downlink traffic was recorded in the university network traffic, due to the classes becoming remote; at the same time, uploads grew due to the much more frequent usage of locally-hosted online teaching platforms [111]. The pandemic also impacted latency on the Internet, with delay values 3 to 4 times higher than in 2019 [112].

When looking at the specific context of mobile networks, the overall trend is different: restrictions in the UK resulted in a decrease of 24% in downlink mobile data traffic over the whole country. The changes were sharper across cosmopolitan areas, which experienced a 50% mobile data traffic decrease, while rural areas were more stable after the lockdown [113]. A lot of attention was in fact drawn by mobility measurements based on mobile network metadata: notably, network metrics showed a steep decrease in the population mobility over the whole UK during the first two weeks after the initial movement restrictions, followed by a slight uniform increase afterward; also, city-level analyses proved that people tended to move from the more dense metropolitan areas to the urban outskirts just before the measures took place [113]. At a national level, network mobility metrics showed a 65% decrease in displacements over France during the first nationwide lockdown, for both short- and long-range trips [114]. The phenomenon started one week before the lockdown was enforced, and, also in this case, fluxes could be observed leaving large conurbations towards more rural or touristic places. Restrictions during the first French lockdown also disrupted rush-hour commuting patterns. The study of trajectories through mobile network data also enabled tracking the effectiveness of reopening schools during the pandemic in France [115]. It also allowed the understatement of how socioeconomic status affected the mobility of populations during the pandemic, allowing governments to better understand how equality and even their actions through the population [116].

The vast majority of previous work analyses focused on the impact of the first lockdown on network usage, and none investigated the later stages of the pandemic response, where the enacted measures became more varied, and the population increasingly accustomed to those. Also, most research has considered aggregate traffic volumes, typically in ISPs or Internet Exchange Points (IXPs), possibly disaggregated into a few macroscopic service categories; very limited attention has been paid to individual mobile services.

In light of these considerations, new insights on how the COVID-19 pandemic affected the spatiotemporal consumption of hundreds of mobile apps in France were covered, presenting that now all apps across the country were affected the same, with a significant number of patterns appearing in relation to changes across both space and time [7]. Those

will be seen in better detail throughout Section 6.1 (focusing on the impact nationwide) and Section 6.2 (focusing on the impact inside major urban centers).

2.2.4. User and urban interactions

The study of mobile network data can also help leverage patterns in the interaction between users and across populations, enabling researchers to have a different view from such in-person interactions.

Early works have investigated the relation between where users make phone calls and their locations [117], with the discovery that more than 90% of the users who called each other share the same space (determined by the BS), even if their home locations are far apart. This suggested that it's possible to utilize telecommunications data as a way to study face-to-face interactions. Similar studies utilizing similar data have also shown that the number of contacts between users and the number of communication activities scale with the population size of cities, highlighting a phenomenon about the interaction-based spreading of diseases as cities get larger [118]. This leads to the conclusion that the mobile network can be used as a sensor for identifying spatial and temporal changes in everyday city activities [119] and how those individual activities can be clustered into patterns [120], which has great applicability as cities get smarter.

The potential for mobile networks could even be expanded, with the supplement of additional sources of data (e.g., bus and taxi GPS positioning) to create a real-time platform that evaluates urban dynamics, which could provide insights about traffic conditions and the movements of pedestrians through the city [121].

By estimating how users interact among themselves by observing their interactions with their smartphones, it's possible can be utilized to delineate the geographical regions and redefine maps, as long as the spatial resolution of data is high enough [122]. This becomes possible due to the possibility of utilizing smartphone data to determine profiles with the urban fabric of cities [123]. Those results are not exclusive to urban environments: the spatial distribution of mobile app consumption has the potential to be utilized to determine the urbanization level of regions, highlighting the heterogeneity of users' behavior across urban and rural spaces [25].

Finally, the understanding of users' patterns of smartphone utilization and the interaction amongst themselves can be even used to estimate the population of regions solely utilizing the data collected from mobile networks, not only statically but even the dynamic variations that urban centers see along the day from the migration patterns between neighborhoods [124].

2.3. Mobile data analysis for mobility research

As it was previously seen, data collected from mobile networks is an excellent proxy to measure and understand how individuals are moving through space. Due to the sheer size of the user base from MNOs, this can enable passively collecting insights into the mobility of populations in the range of millions, at virtually no extra cost for the operator, a feat which was hard (and expensive) to achieve before. This can be an interesting alternative to other new sources of mobility data, such as GPS (which requires more care in relation to users' privacy), or as an alternative to classical mobility and transportation surveys.

Following the definition of [11] and the observed fields of Figure 2.1, a few subcategories of studies using this source of data exist: 1) Human mobility studies, which focus on comprehending the mobility of users across time and space; 2) Transportation studies, which aim to comprehend the flow of roads and public transportation. This source of data has been greatly used by many fields outside of networks and telecommunications, such as urban planning, traffic engineering, sociology, and epidemiology. This subsection will present a short look into important works on the sub-fields to help create the full picture of the state-of-the-art, as mobility datasets are not the focus of this thesis.

2.3.1. Human mobility

These works set their goal of understanding how users move and travel within the space covered by mobile networks. This can be done by studying signaling events in the network (an essential and frequent exchange of information between devices and BSs) or by handover information (whenever a user is passed from one BS to another, usually due to their movement and to optimize the coverage), which can be utilized to geo-localize the trajectories of users, as well as estimate their home, work and other key travel locations, providing an important proxy to understand movement of users of the network [125].

For example, the mechanisms of urban mobility have been studied utilizing the temporal and spatial trajectories of the network, unraveling networks of mobility across populations [126]. Mobility datasets can be specifically used to generate Origin-Destination (OD) matrices [127], [128], one of the most important data sets for understanding human movements. Those are usually collected through traditional surveys but are usually very complex and time-intensive to obtain. Generating such sets from signaling events can help create larger and more up-to-date sets, and validations have shown good precision when compared to the traditional sets obtained from surveys [129]. As expected, in order to determine the origin of the OD matrices, an accurate representation of the home location of users needs to be present, with a significant number of algorithms being present in literature, with their evaluation being presented in [130].

The combination of mobile phone data with GPS data can also help uncover previously unseen behaviors in models of human mobility in relation to the distances traveled

by individuals [131] and to better detect land use [132]. It can also help understand intra-urban variations of vehicular and non-vehicular mobility [133], measure traffic congestion [134], and even identify accidents on the road [91]. Human mobility measured from smartphone activity can also be related to socioeconomic indicators, providing an interesting perspective for studies interested in diversities of mobility across the heterogeneity income levels of societies [135], [136].

As expected, the generation of those mobility datasets will have limitations [137], have their biases and may necessitate special techniques when the collection may be sparse or contain missing values [138]–[140], the needed attention about the temporal sampling frequency [141], differentiating stationary to moving users [142], as well as the necessity to preserve user privacy when performing the processing of those collections [143], [144]. When deciding whether mobility data from mobile networks in place of classical surveys, it's important that researchers take those limitations into account to be sure they're utilizing the data correctly without generating false assumptions. Overall, the usage of mobile network data can help train the future of models for human mobility, where more details about potentials and future directions can be seen on the survey [145]

2.3.2. Transportation systems

Besides the comprehension of how users and populations are moving throughout the geographical space, mobility datasets originating from MNOs can help validate and improve studies in the area of transportation.

Early examples come from the evaluation of this source of data versus GPS: for example, two separate works have concluded that the estimation of travel times using MNO data had an error within 5% to 15% of those collected from measurements with sensors on the streets [146], [147]. Another measurement possible with MNO data is to estimate travel times and traffic congestion on the roads [91], [148]. The data can also be useful to estimate the volumes of road traffic [149].

Besides traditional road traffic, mobile network data can also help to study data collected from multi-modal means of transportation (e.g., foot, bicycles, buses, cars). This could be done by estimating the velocity within users are switching BSs through the network [146], [148], [150], or by simply comparing against times measured by researchers [151] or modes which have their movement times well known [152].

Finally, those mobility datasets can also help urban planners in designing better transportation systems. For example, this data can be used to help tune the parameters of road traffic generators [153] or to better understand the performance of bus routes [154], [155], leading to adaptations that better suit its users.

3 Measuring and processing mobile network data

To ensure the quality of data used in research, it is important to select the best type of data available from mobile networks, perform accurate measurements and collection, and handle data processing and analysis effectively. This involves understanding and applying key concepts and techniques throughout the process.

Throughout this Chapter, the main materials and methodologies used throughout the contributions of this thesis will be presented. This Chapter is structured as follows: Section 3.1 will give a brief overview of the structure of the architecture of mobile networks; Section 3.2 will present the main types of data that can be collected over those networks by MNOs; Section 3.3 will explain how measurements initial processing are done, based on the deployment of the operator where the data utilized in this thesis was collected; Section 3.4 will go through the problem of distinguishing 5G NSA sessions from 4G sessions when measured at the network core; Section 3.5 will have an overview of the general tools utilized for data processing, especially focusing on the privacy side of the collection; finally, Section 3.6 will go over considerations needed when processing data in relation to spatial units.

3.1. An overview of mobile networks

Mobile networks are composed of two main structures, as observed in Figure 3.1: the Radio Access Network (RAN), which serves as the wireless access link between each user and the MNO, and the Core Network (CN), which will manage and process all tasks required by users, such as calls and IP tasks.

The RAN will be the access point of the network, composed of all BS available, which will cover all the geographical areas where the MNO provides its services. Devices will be communicating with said BSs in order to establish valid connections, according to the best BS that can both provide a good signal to the device and have enough space available (both in relation to computational and radio resources). Whenever a user exits the coverage area of a given BS, the network will perform a handover operation,

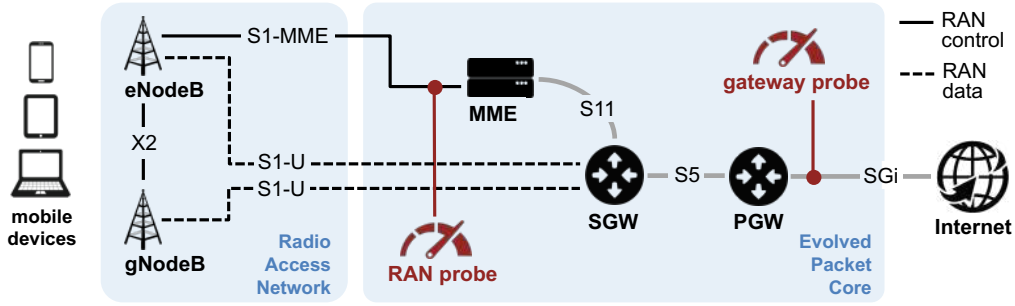


Figure 3.1. Simplified 4G and 5G NSA mobile network architecture illustrating the combined RAN and gateway data collection setup used by the MNO.

assigning a new BS that better covers the user according to their new spatial position. This task is handled by location updates that each device is required to send to the mobile network, even if no communications are being made at the time, in order to guarantee coverage at any instant the user may request an operation. Each BS may be composed of multiple antennas, which can provide access to different coverage areas (being directional antennas), or access through different radio frequencies available to the MNO (in order to optimize the allocation of radio resources).

The CN side of the mobile networks will be the part connecting the access layer to any external network. It will be performing the switching tasks, connecting the devices to their desired end-points, and handling all the packets exchanged between them. Throughout generations, the core of the network has seen significant changes, specifically in order to accommodate new services made available by MNOs.

As expected, inserting probes in different parts of the network may lead to different data being collected. For example, a probe at the Radio Network Controller (RNC) will be connected to the RAN side of the network, and able to collect any signaling events within the Radio Resource Control (RRC), at each BS with a fine granularity and collect statistics about them. Similarly, probes at the Mobile Switching Center (MSC) also collect signaling events, but on the CN side of the network. Mobility-related data sets are usually collected from one of those two probes, as the fine temporal granularity of signaling events offers a good proxy for the location and movement of users. In another direction, a probe at the Gateway GPRS Support Node (GGSN) or Packet Data Network Gateway (PGW) will be in the CN and used to collect statistics about IP-level traffic flowing through the network, being the common probe utilized to collect per-application session traffic. This is the type of probe used for the majority of the collections used for the contributions of this paper, and its details will be discussed in Section 3.3.1.

3.2. Data collected from mobile networks and devices

The term *Smartphone data* is loosely used for quite a few different types of sets that could be collected, according to where in the network this collection is done (or even outside of it), and which information is being collected. Due to these distinct origins, the information contained in each of those greatly varies, leading to their utilization for different fields and research questions. Since there's a usual misunderstanding on what can be smartphone data, this Section will address the possible types of sets that can be generated by smartphones, and what can be collected by MNOs.

3.2.1. Signaling data

A way to obtain the location and mobility of populations is to collect signaling events from the network. Those can be an interesting way to collect location while respecting individual users' privacy, as signaling events have a straightforward aggregation over spatial units (in contrast to traditional spatial noise techniques used with other location sets, such as GPS).

Data sets containing signaling events are collected through the RNC probes located at the RAN side of the network and will be recording passive events between the mobile phone and the network, especially the ones about assigning which BS the phone should utilize in case it requires services. They will have a lower granularity than GPS data, bound by the density of the MNO deployment, as the position of the user will be determined by the BS that they're currently attached to or being handed over through movements over space. Signaling data is mostly used to study a population's mobility, by processing it to create specific data sets such as OD matrices.

3.2.2. Call detail records

CDRs are one of the most common types collected through generations of RANs. This type of data is obtained through the recording of events that exchange information between any mobile phone and the network, e.g., telephone calls or text messages (SMS, MMS). It's possible to record statistics about those events, such as timestamp, call duration, type of operation, and other technical information that could be used by the operator for billing and study of users' patterns. They could also be utilized as a proxy for mobility, although due to their collection being on the CN side of the network, it's important to pay attention in relation to the biases they can have in calculating mobility statistics [138]. More importantly, those records do not capture the activity of mobile subscribers in terms of data traffic.

While interesting from a sociological perspective, calling and texting play an increasingly diminishing role in current and future generations of RATs; not only have

those sorts of communication services dropped in popularity, but even MNOs are shifting them to traffic-oriented services, such as VoIP or current instant messaging applications (WhatsApp, Telegram, Signal). Therefore, in the current age of network data mining, CDRs are dropping in popularity since they're not able to provide a view of hidden structures in users' dynamics, and more specifically, they do not allow observing patterns of mobile traffic (the main product demanded by users through the MNO).

CDR datasets are slightly more common to be encounter as open data and through challenges, but most of those sets are starting to become slightly *depreciated* through their age and the aforementioned drop in popularity of native (and not IP-based) call and text messaging services through MNOs. Examples are the Telecom Italia Big Data Challenge [156], which amongst many sources, contained CDR data from the city of Milan and the Province of Trentino, both in Italy; and the Orange Data for Development Challenges [157], [158], which contained CDR records for both Ivory Coast and Senegal.

3.2.3. Traffic flows

Current-day collections have been favoring IP traffic flows instead of CDRs. As mentioned, this is mainly due to the evolution of telecommunication technology standards placing the majority of services inside IP protocols. Therefore, most of the data and operations that an MNO sees are *IP packets*, leading the collection of traffic flows to be the most common collection.

Usually, whenever a mobile application requests content through the mobile network, the exchange of this content will be done in communication protocols, such as Transmission Control Protocol (TCP) and User Datagram Protocol (UDP). The network operator will be able to monitor those flows, extracting how much data was exchanged from both the uplink and downlink direction, and the duration of those flows (which could be used to estimate throughput). This sort of data will represent specifically *traffic consumption* through the BS of the network, and has the unique ability of differentiating traffic of specific mobile applications. In the current setting of networks, where operators optimize slices of the network according to the content of data flowing, being able to differentiate traffic flows becomes essential, and we can leverage this to also generate datasets that analyze these specifics.

Unfortunately, this type of data is significantly harder to obtain, since it holds quite sensitive information of the operators, making it harder to find openly released sets. An example of public releases of mobile traffic data was the NetMob23 Challenge Dataset [159], which contained the individual demand of 68 popular mobile services in 20 metropolitan areas in France, collected through the network of Orange.

3.2.4. GPS data

Data from GPS is the only type of data to be mentioned in this Section that although completely related to smartphones, cannot be collected by MNOs. Although GPS systems of smartphones can use BSs to help triangulate locations, whenever the satellite constellation is not fully available, its main collection points are the operators of each of the Global Navigation Satellite System (GNSS). Examples here would be the aforementioned GPS system[1] (operated by the US), GLONASS (Russia), BDS (China) and Galileo (EU). Therefore, although this data will represent the position of smartphones, this is indeed data from satellite networks and not mobile networks.

GPS is used specifically for *trajectory tracking* and has significantly better spatial granularity and precision than equivalent data obtained from MNOs. In order for it to be collected, it would be necessary to either have access to records of global positioning networks or have access to smartphones (through an installed application that can have system permissions to track GPS data). Its granularity will be related to the granularity in which the mobile phone updates its location (in relation to time) and can be as precise as the GPS system on smartphones allows. It's also not unusual to see GPS data being used to validate spatial data collected from mobile networks or used together in order to improve the measurements (as discussed in Section 2.3).

3.2.5. Other possible data sets

With mobile networks becoming smarter and operators collecting a higher volume of data from their premises, a new age of mobile network data has been emerging. A few examples of different datasets, not necessarily related to traffic and events are: *Network topology data*, by recording the deployment of new antennas; *Network coverage*, which observes the propagation of deployed antennas; and *User authentication and connection protocols*, such as the Packet Data Protocol (PDP) and Remote Authentication Dial-In User Service (RADIUS) authentication protocol. Other network metrics and statistics can also be calculated through traffic flow analysis, such as the delay experienced by the user and other packet-level statistics about the flows.

3.3. Mobile traffic measurements at MNOs

After understanding the usual data collected from mobile networks, the next step is to better comprehend how these sets can be collected by MNOs. The results presented in this thesis originate from the collection of traffic flows and their derivations. Also, matching

[1]This thesis will utilize GPS interchangeably with GNSS, as it's a far more common acronym in literature.

traffic flows with signaling data plays an important role in improving the spatial quality of these data sets.

Therefore, this next Section will give an overview of how the data used in this thesis was collected, with an overview of the probes used, their location within the network, as well as the implications according to the RAT they're measuring. All mobile phone data used in this thesis has been captured at the production network of Orange servicing the metropolitan France territory, spanning multiple years and a significant range of RATs (2G, 3G, 4G, and 5G NSA). The methodology to be described reflects the techniques used by the MNO; as expected, much of the technology here is tailored to the specificity of their infrastructure, so divergences could be expected in pipelines of other MNOs.

3.3.1. Mobile network measurement probes

Mobile network probes are the devices responsible for monitoring and registering the data flowing through the core of the network. They are considered passive devices, i.e., they are not sending signals through the network and measuring statistics about how this signal traversed the infrastructure (as it would be done in active measurements of the network, i.e., measuring latency and throughput on speed tests). Instead, they're simply *sensing* the data that is requested by users through the uplink and downlink direction, as well as any other metrics that are generated by the operator. Those passive probes are as simple as servers connected to access points, i.e., a daemon process running on a UNIX or Windows machine, attached to the network and observing flows.

The passive measurements probes employed by the operator essentially monitor the Gi, SGi, and Gn interfaces that connect GGSNs and PGWs to external Public Data Networks (PDN), gathering data about the traffic generated by mobile subscribers using 2G, 3G, and 4G connectivity. It's interesting to note that due to different network cores being utilized across 2G, 3G, and 4G, the probes utilized will also be *physically different*. This means that by solely identifying the probes by which the CN is located, it's possible to distinguish traffic flows across those 3 RATs. But, in the case of 5G NSA RAN technologies, as observed in Figure 3.1, the 5G gNodeBs coexist with 4G eNodeBs in the RAN. This means the measurement of 4G and 5G NSA traffic comes from the same probe with no distinction, leading to some slight complications to overcome when later processing the collected traffic from the probes, which will be further detailed in Section 3.4.

As observed in Figure 3.1, two complementary passive measurement systems are used by the MNO to compose the two different datasets utilized:

1. **RAN probes** deployed at the S1-MME interfaces of the Mobility Management Entity (MME) capture the *signaling data*. Due to the way the 5G NSA deployment operates, these probes can monitor the control planes of both eNodeBs and

gNodeBs. Signaling data is employed to geo-reference and time-stamp the session information. Specifically, the probes observe all signaling events generated by each User Equipment (UE), e.g., when it requests a service, exchanges data, performs handovers or moves across Tracking Areas (TA) and records the BS of attachment. By leveraging this information, it's possible to associate each UE (and the sessions it generates) with its serving BS at all times.

2. **Gateway probes** tapping at the SGi interface of the PGW monitor all IP traffic and extract information on each transport-layer session, extracting the aforementioned *traffic flows*. These probes record the total data traffic generated by the session, its start and end times, and the associated mobile service.

The measurement approach above allows for overcoming inherent limitations in the precision of the localization information available in the core network. Indeed, the UE location identifiers available at the PGW are updated infrequently, leading to stale positions and localization errors in the order of kilometers [160], [161]. Relying on the locations recorded by the gateway probes would thus jeopardize the capability to geo-reference session-level data at the granularity of the individual BS in a reliable manner. To overcome this problem, the UE and time information gathered by the gateway probes are crossed with the signaling data of the RAN probes so as to retrieve the BS(s) where each session occurs and assign the correct (fraction of) sessions to all BSs.

3.3.2. Identifying per application traffic flows

The app generating each IP session is identified by the MNO via a combination of DPI and proprietary traffic classifiers, within the GGSN/PGW probes. The DPI examines the headers at both network and transport layers to derive the per-flow mobile service information. While the algorithms used for traffic classification are confidential, the operator reports high accuracy in independent tests.

3.4. Identifying 5G NSA traffic flows

Although the collection methodology worked well for networks up until 4G (where both the RAN and the CN were different for each RAT), a problem emerges when the target network employs 4G and 5G NSA RAN. As depicted in Figure 3.1, in this configuration 5G gNodeBs coexist with 4G eNodeBs in the RAN and provide higher-capacity wireless communication to 5G-capable UEs. Yet, the lack of a dedicated 5G CN in the NSA deployment forces gNodeBs to depend on interactions with eNodeBs, via the X2 interface, for control operations towards the 4G MME. Also, gNodeBs connect to 4G gateways via a slightly modified S1-U interface for all data plane transmissions.

While the previously mentioned merge of signaling RAN probes and Gateway probes allowed a better localization of where the sessions were generated, it still does not allow for the differentiation between 4G and 5G NSA flows. Traffic flows are measured at the gateway probes, which is located at the CN and shared for both RANs for 5G NSA[2]. To overcome this problem, a different type of information collected by the gateway probes will have to be leveraged, which is the focus of this Section.

Consider a user connected to a mobile network during time interval Δt, utilizing multiple mobile applications in their smartphone. The traffic flow statistics of this user were collected by a Gateway probe in the format seen in Table 3.1, accounting for the number of bytes exchanged through the TCP and UDP communication protocols used by each mobile application in Δt at the BS the user was connected. Usually, to identify the RAT of this traffic flow, it's possible to match column `loc_start` with a secondary data set provided by the MNO which says the RAT of the CN where that probe was. In the case of 4G and 5G NSA, the CN is shared, therefore the traffic flow measurement is done by the same probes and the differentiation cannot be done.

In order for the user to have started the communication exchange with the BS, a previous event has occurred (and recorded by the gateway probes). This event is recorded as sessions called by the operator as CNX sessions, which record the start and end of the PDP context created after the RADIUS authentication. A simpler example of the type of data collected by the gateway probe for CNX sessions is seen in Table 3.2. It's important to note that both the samples of Tables 3.1 and 3.2 are simplifications, with both representing different network protocols occurring independently of each other. Those are recorded by the same gateway probe and have more information attached but, for the following example, only this information is necessary in order to understand how to flag either a TCP/UDP flow recorded at the core of the network as 4G or 5G NSA.

3.4.1. The PDP context and RADIUS authentication

Both the PDP context and RADIUS fit over the other types of data that could be collected from the RAN communication procedure in Section 3.2.5. Some extra clarification about them is needed to understand how they can be leveraged to differentiate 4G and 5G NSA flows.

The RADIUS is a user and server protocol that receives user connection requests, authenticates the user, and then returns the information about the configuration necessary for the client. Each entry corresponds to a user (identified by their `msisdn` connected to a BS) for one RAT and one known location.

The PDP context will be a data structure present at the gateway node created after the RADIUS authentication, with the so-called CNX session being the start and end of

[2]Which was the only 5G deployed by the operator in its production in France at the time of the collections.

Header	*Description*	*Example values*
`msisdn`	Customer ID	3367955511
`port_app`	Mobile application ID	66001
`start_value`	TCP/UDP session start timestamp	1667884574
`stop_value`	TCP/UDP session end timestamp	1667885115
`byte_up`	Session bytes in uplink direction	45638
`byte_dn`	Session bytes in downlink direction	163779
`loc_start`	Start location information	8102f81001733d03
`loc_end`	End location information	8102f81001733d03

Table 3.1. Example of a captured flow for either a TCP or UDP sessions at the gateway probes

the PDP context. This means, that whenever a user wants to exchange data with the network, it must first attach to a BS, run the RADIUS authentication triggering then the PDP context, where the network records all the necessary information about the user in order to determine which type of connection will be made.

Along those records gathered by the gateway probe during the user authentication (creating a CNX session), is a flag indicating whether both the user and the BS are 5G NSA enabled. In case both are, it will be possible to have TCP/UDP flows using 5G RAT. It's important to note that the CNX session will record the entire duration Δt_{CNX} where a user is connected to a certain BS. Inside this duration, multiple TCP/UDP sessions of duration Δt_{TCP} or Δt_{UDP} can be initiated by any mobile application to exchange data in any direction, as exemplified in Figure 3.2. Also important to note is that the CNX connection is CN dependent: if a user switches from 3G to 4G, a new session is opened. But, since 4G and 5G NSA share the same core, the CNX session is the same between 4G and 5G flows. Thankfully, the probe identifies 4G flows as *primary RAT* and 5G NSA flows as *secondary RAT*, which means it's possible to know if and exactly when a connection of the user exchanged 5G data. Next, this process will be shown.

3.4.2. Merging TCP/UDP flows with the PDP context

As mentioned, whenever a user (identified by `msisdn`) requests connection to a base station, the RADIUS authentication routine is run with the PDP context, and the following is recorded as a CNX session following the format of Table 3.2. The information about this CNX session will be recorded at the RAN probes when the connection between the user and the BS is finalized (either due to timeout or by being handed over to another BS). The information recorded, as seen in Table 3.2, is the customer identifier `msisdn` the time interval Δt_{CNX} where the user was connected (represented by `start_value` and `end_value`), the total bytes for each direction exchanged during the entirety of the

Header	*Description*	*Example values*
`msisdn`	Customer ID	3367955511
`start_value`	PDP context start timestamp	1667884563
`stop_value`	PDP context end timestamp	1667888204
`byte_up`	Overall connection bytes in uplink	3105082
`byte_dn`	Overall connection bytes in downlink	31887488
`loc_start`	Start location information	8102f81001733d03
`loc_end`	End location information	8102f81001733d0c
`sec_rat_type`	RAT type information	1
`sec_rat_start_ts`	In case 5G, time where 5G connection started	1667887716
`sec_rat_stop_ts`	In case 5G, time where 5G connection ended	1667888137
`sec_rat_byte_up`	In case 5G, bytes exchanges in 5G uplink	4049
`sec_rat_byte_dn`	In case 5G, bytes exchanges in 5G downlink	1386760

Table 3.2. Example of a captured CNX session at the RAN probes

connection (`byte_up` and `byte_dn`), the BS where the session was initiated (`loc_start`) and the next BS that the user was located (represented by `loc_end`). In the case where the user was not handed over to another BS (e.g., due to a timeout), the value is the same as `loc_start`. Most important is the sequence of columns representing statistics about the *secondary RAT* usage, which will contain any information about 5G flows. Column `sec_rat_type` will indicate whether both the device and the BS are 5G NSA enabled (1 if True, otherwise 0).

Looking solely at `sec_rat_type` is not enough to know if 5G flows happened. Indeed, when analyzing the collections, it was noted that a significant number of connections between users and BS could have been 5G, but the fields of 5G flow duration (`sec_rat_start_ts` and `sec_rat_stop_ts`) and bytes exchanged (`sec_rat_byte_up` and `sec_rat_byte_dn`) indicated 0. This means that even though 5G was available, for some reason outside of the knowledge presented in the data, both the user and network decided only 4G was needed for those flows. Therefore, in order to truly know whether a user utilized 5G, either `sec_rat_byte_up` or `sec_rat_byte_dn` has to be above 0 (and subsequently 5G duration $\Delta t_{CNX,5G}$ will also be recorded by `sec_rat_start_ts` and `sec_rat_stop_ts`).

Next, the process to merge the TCP/UDP sessions with the CNX sessions will be

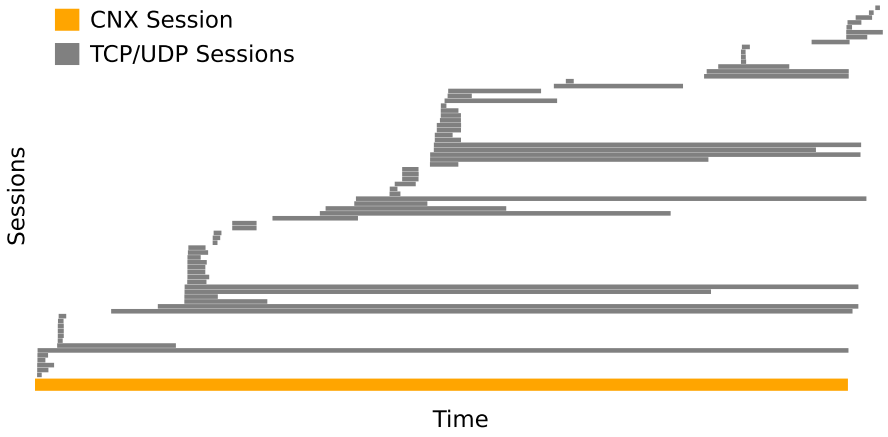


Figure 3.2. Example of a CNX session for a user in a BS, with all TCP/UDP that occurred inside its interval. In case this CNX session was flagged as 5G NSA, all TCP/UDP sessions that occur inside of it will be 5G.

described. A question may occur on why not use only the CNX session information, not having to rely on TCP/UDP flows and adding an additional merge to the collection pipeline. Unfortunately, the big downfall of the CNX session collection is that it contains only total traffic statistics, with no discrimination of the mobile applications used. Therefore for the type of analysis performed through this thesis, CNX sessions are not enough and need to be merged with TCP/UDP to get per-app data.

Considering the set of CNX and TCP/UDP sessions, in order to flag which flows of TCP/UDP were 5G enabled both datasets will be merged. Considering the sample data structures of Tables 3.1 and 3.2, a *left merge* between TCP/UDP with CNX files is done, with the following conditions all have to be respected in order to define a match:

1. The customer id `msisdn` is the same for both sets;
2. the `start_value` of TCP/UDP is higher or equal to the one of the CNX session;
3. the `start_value` of TCP/UDP has to be smaller than the `stop_value` of CNX;
4. the `loc_start` of both TCP/UDP and CNX have to be equal;

After the merge is done, the following set of rules has to be respected to declare that a TCP/UDP flow was 5G:

1. `sec_rat_type` $= 1$
2. `sec_rat_byte_up` > 0 `OR` `sec_rat_byte_dn` > 0

Figure 3.2 helps visualize the temporal conditions: considering the represented CNX session (orange) respects the previous conditions and represents a 5G NSA connection

between user and network, all the TCP/UDP flows (in gray) that lay inside its interval are considered to be 5G NSA. With this process done, it becomes possible to obtain a 5G per-app data set, utilized throughout the rest of the following chapters.

3.5. Processing mobile network data and ensuring privacy

Following the merge of the datasets of Tables 3.1 and 3.2, it's finally possible to obtain a set that describes 5G NSA (and below) flows per application. But, the way the current data is cannot be *allowed* to leave the premises of the operator. This next Section will describe the operations done in order to allow the data collection and processing to be approved by the Data Protection Officer (DPO) of the MNO, as well as authorized by the relevant national privacy-protection agency, so it was possible to be utilized for research, as well as tools and other processing steps done to allow the utilization of the large-scale network measurements.

3.5.1. Ensuring user privacy

Considering the measurement and processing pipeline described so far, it's important to remark that all data so far mentioned here is still individual (i.e., user-level) transport-layer level sessions that pass through the network core. After the collection by the probes, all data is moved directly to a temporary storage infrastructure within a secure platform at the operator's own premises. All file crossing (e.g., merges) and aggregation are done in memory without any intermediary results being permanently stored.

A set of filters is applied to the raw dataset before any aggregation is done. The first filter will be to remove all roaming and Mobile Virtual Network Operator (MVNO) users, as neither of those are officially clients of the MNO[3]. This will leave only the operator's customers who are informed of the possibility of such anonymous processing if their contract is signed with the operator. Also, every record that concerns less than 6 customers present on the same BS in a given time period in the later aggregated dataset is also filtered before being available to use. All operations are done complying with Article 89 of the GDPR guidelines [162].

3.5.2. Data aggregation

As it's not feasible to utilize user-level data, i.e., study Instagram traffic generated by a specific user, for the data to be released to researchers, an aggregation is needed in both the spatial and temporal dimensions. Following the set of rules and filters of Subsection 3.5.1, data will be aggregated at the BS level. This means that for a given time interval (e.g., an hour), all the per-app data generated by the set of users connected

[3]Instead, they're customers for other companies who pay to use the MNO structure.

Header	*Description*	*Example values*
`id_port`	Application ID	66001
`loc_start`	Antenna ID	8102f81001733d03
`sec_rat_type`	5G identifier	1
`timestamp`	UNIX timestamp bin	1667888204
`byte_up`	Bytes in downlink	45638
`byte_dn`	Bytes in uplink	163779

Table 3.3. Example of a TCP/UDP traffic flow dataset, with 5G traffic identified, which can leave the secure premises of the operator and be utilized for research guaranteeing user privacy.

at a given BS at the selected time interval will be aggregated (e.g., via a *Group By* function). Therefore, all will not refer anymore to *a user*, but to *an area*.

The user-level data described in Table 3.1 has its temporal granularity in *seconds*, represented as a UNIX timestamp. Such temporal granularity do not follow the specific guidelines of Subsection 3.5.1 and needs to be aggregated over time into a higher granularity. The easiest way to perform this operation is to perform a *floor division* of the UNIX timestamp by a given integer N and multiply the result again by N. For example, in the case of a UNIX timestamp 1667888204 of Table 3.3 (07:16:44 of 08/11/2022 in France), to change the granularity from 1-second to 1-hour, a $N = 3600$ can be used to obtain new timestamp 1667887200 (07:00:00 of 08/11/2022 in France). The same Group By operation done for spatial can be used here to aggregate over the higher coarse temporal granularity.

With both the spatial and temporal aggregations done, the final dataset obtained can be seen in Table 3.3. What this means is that for any given BS, it's possible to understand the temporal consumption of data generated by all users attached to it in the time interval, for any given app. This dataset shall give spatiotemporal insights about traffic consumption for all the use cases to be presented through this thesis.

3.5.3. Large-scale data processing

The collection of traffic consumption is a very resource-heavy operation. More specifically, the storage consumption of the resulting data sets can be quite intense: one day of per-app traffic through all antennas in France with 5-minute temporal granularity is around 12GB when stored as a Parquet file (which already includes quite efficient compression to reduce file size). Therefore, storing and processing months of data is not a trivial task, and is difficult be execute in *consumer-grade hardware*. To solve this, a combination of industry-oriented servers and tools needs to be used to process the data in order to perform research.

An internal cluster of servers is used for all big data tasks performed throughout this thesis. Two Dell EMC servers running Ubuntu 20.04 are utilized, interconnected by an

Ethernet connection, and accessed externally through a VLAN. Those servers are used for all data processing tasks (which are all CPU-oriented) and storage of data in use. All other sets not in use are stored in a Network-attached Storage (NAS), which is also connected to the servers through an Ethernet connection. In order to efficiently perform all big data processing tasks in a distributed manner, the servers are set up as an Apache Hadoop cluster. The idea of Hadoop is to facilitate the usage of a network of computers to solve large-scale computational problems in a distributed manner. It will consist of a storage module, known as Hadoop Distributed File System (HDFS), and a processing part which splits files into blocks which are then distributed across nodes for processing in parallel. This data split and processing parallelization allows the processing of large-scale network data to be significantly faster.

Alongside Hadoop, the other main tool utilized is Apache Spark, which is an analytic engine for the processing of large-scale data. It provided an interface for programming the processing to be performed on the Hadoop clusters, allowing for parallelism. The version of Spark used throughout this thesis is its Python variation PySpark. The idea behind PySpark routines is to operate MapReduce operations, where the raw collected data is filtered, processed, and therefore reduced in size in order to be later utilized for more complex analysis.

3.5.4. Feature scaling

Another important technique that's used throughout the contributions of this thesis is feature scaling, which will be done to achieve a few critical goals during the analysis and presentation:

First, the raw traffic values are very sensitive information for the network operator. Therefore, even though analysis can be done with pure values, their presentation like this is not possible to avoid disclosing private information. Therefore, feature scaling serves the purpose of *anonymize* traffic values so they can be later presented, changing the scale in which results are presented but keeping the overall volume patterns. Secondly, feature scaling can serve a variety of purposes related to making comparisons *fair*, i.e., two antennas may have different volumes due to the sheer number of people present in their coverage areas, but have the same usage patterns; feature scaling techniques a better comparison of those antennas, especially when the data will be utilized in models (e.g., linear models for feature explanation, cluster analysis). Next, a more detailed exposition will be done for a few techniques used.

Re-scaling - min-max normalization: Starting with the raw time series of the traffic of a certain antenna $n \in N$ in the mobile network, the full-time series is represented by T_n and the traffic at instant t will be $T_n(t)$. The min-max normalization will be:

$$T'_n(t) = \frac{T_n(t) - min(T_n)}{max(T_n) - min(T_n)} \tag{3.1}$$

The result of this scaling is a shift in the range of values, from the original $[min(T_n, max(T_n)]$ to $[0, 1]$. This will hide the original scale of traffic and the traffic distribution shape, but not interfere with the original patterns, being an ideal scaling whenever differences in absolute volume want to be preserved for the analysis This is valid for either per-app traffic or overall traffic aggregation dynamics.

Standardization - z-score normalization: This process can be done whenever the focus of the analysis will be in *patterns* of the time series, and not absolute volumes. Considering the same notation as before, the z-score of $T_n(t)$ will be:

$$T'_n(t) = \frac{T_n(t) - \mu_{T_n}}{\sigma_{T_n}} \tag{3.2}$$

where μ_{T_n} and σ_{T_n} are the mean and standard deviations of the full-time series T_n, respectively. The result will be a new distribution of traffic where the data will have $\mu_{T'_n} = 0$ and $\sigma_{T'_n} = 1$. This allows a better focus on patterns of consumption, i.e., to answer questions such as *"two services have the same peak times of usage, independently of their overall volumes of traffic?"*. This technique is also valid for either per-app traffic or overall traffic aggregation dynamics.

3.5.5. Quantifying the importance of mobile applications

In order to quantify which types of applications are utilized more in any specific space, a metric that's capable of representing the diversity of each app's mobile traffic consumption across space is necessary. The most straightforward approach would be to compare absolute volumes of traffic consumption of applications. But, this approach faces problems due to the sheer differences in the magnitude of traffic consumption different mobile applications have according to the type of content they're related to, i.e., video streaming applications intrinsically consume more traffic than instant messaging applications, no matter the locations where mobile traffic is being consumed. This not only introduces a bias of volume to the analysis but also complicates further clustering and modeling approaches. Also, there is volume bias according to where base stations are located, i.e., locations in busy downtown areas tend to consume more traffic than antennas in suburban and purely residential zones. This drives the need for a metric that is independent of volume, which can reflect the changes in the importance of applications across space.

An interesting solution is using the Revealed Comparative Advantage (RCA) [163] metric, originally proposed in the field of econometrics to compare the relative advantage/disadvantage of a sample inside one category, but can be successfully utilized

in the analysis of mobile traffic across space [6], [79]. In this case, the RCA will define the level of over or under-utilization of a certain mobile application at a specific base station, in relation to the entire set of applications and base stations. Considering one mobile application a that belongs to the set of applications A, and one base station v that belongs to the set of base stations V, the RCA $[\forall a \in A, \forall v \in V]$ is defined as:

$$RCA_{a,v} = \frac{T_{a,v} / \sum_{a' \in A} T_{a'v}}{\sum_{v' \in V} T_{av'} / \sum_{a' \in A, v' \in V} T_{a'v'}}, \tag{3.3}$$

, where $T_{a,v}$ is the traffic recorded for service a at antenna v, $\sum_{a' \in a} T'_a v$, refers to the traffic recorded by all services of set A at antenna v, $\sum_{v' \in V} T_a v'$ refers to the traffic of app a recorded across all antennas of set A and $\sum_{a' \in A, v' \in V} T_{a'v'}$ refers to the total traffic of the network, across all apps and antennas, during the studied period.

Values of RCA above/below 1 indicate that a base station has more/less consumption of a certain mobile application than other base stations of the network. A small problem is that while the lower boundary of RCA is 0 the upper boundary is ∞, which can highly unbalance the distribution of RCA values and result in problems in clustering algorithms [164], [165]. To overcome this, a solution is to utilize the Revealed Symmetric Comparative Advantage (RSCA) [164], defined as:

$$RSCA_{a,v} = \frac{RCA_{a,v} - 1}{RCA_{a,v} + 1}, \tag{3.4}$$

which results in bounded values in the interval $[-1, 1]$, with values above/below 0 meaning over/under utilization. This results in a distribution of RSCA values with no long tail.

3.6. Spatial resolution of mobile network measurements

After collecting the traffic flows over a period, a critical step becomes identifying the locations where the traffic demands were distributed across space. To ensure precision, especially in social and land use studies, the correct designation of traffic over space is a critical operation in the data processing pipeline. This Section will go through a few common techniques employed to ensure the correct assignment of spatial traffic demands.

3.6.1. Matching traffic flows with antenna deployment data

The matching of traffic flows with antennas can be done through internal data provided by the network about the topology of their deployment. A simplified example of such a set can be seen in Table 3.4, where the merge between this and the set presented in Table 3.3, through the `loc_start` field, allows the spatial localization of flows.

Besides the Antenna ID field, a few unique values are presented here, which can further assist in not only understanding the space where the network is but also a few technical

Header	*Description*	*Example values*
`loc_start`	Antenna ID	8102f81001733d03
`lon`	Longitude	0.42
`lat`	Latitude	43.1
`nom_site`	Internal name of the BS of	`PARIS_METRO`
`nom_cell`	Internal name of the antenna	`PARIS_METRO_120`
`type`	Technology	`ENODEB_CELL`
`azimuth`	Azimuth angle	120
`class`	Cell type	`MACRO`
`type_coverage`	Coverage type	`OUTDOOR`

Table 3.4. Example of the information about the antenna deployment, which can be used to geolocalize traffic demands across cities.

aspects of the deployment. Field `lon` and `lat` present the *Longitude* and *Latitude* of each BS, which `nom_site` representing the internal name of that BS and `nom_cell` the name of the antenna. As expected, each BS may have multiple antennas attached to it, so it's important to understand this differentiation.

Next, a few technical aspects of each antenna are presented. Field `type` is related to the RAT of each antenna. A BS may have antennas with multiple RATs connected. In the example given, `ENODEB_CELL` means that this antenna is attached to an eNodeB core (which could generate traffic from either 4G or 5G NSA, with the differentiation between both techs presented in Section 3.4. Other values encountered are `NODEB_CELL` and `BTS_CELL`, representing antennas attached to the 3G and 2G cores, respectively.

The next field is `azimuth`, which represents the direction of the antenna. Having only latitude and longitude gives information about a single point in space, but most deployments of the MNO are directional antennas, which radiate greater power to certain directions (thus extending their better coverage area to that direction). In case a deployed antenna is omnidirectional (i.e., a non-directional antenna), this value will be 0.

Another field present is `class`, which represents the type of access node of the deployment. For example, macro cells (as the example on the table) tend to be antennas with higher coverage and greater power, while micro cells (common in heterogeneous deployments) will be antennas specialized in covering only certain locations (e.g., due to shadows of macro cells coverage), with lower operational power.

The final field is `type_coverage`, which will tell the environment in which the antenna is deployed. The geodesic system only tells where in the globe each antenna is located, but this does not take into account built structures. Due to the mass popularization of mobile phones and the problem of transmission signals penetrating certain buildings, network operators are deploying more and more antennas inside *indoor environments*. Therefore, this field will differentiate if an antenna is deployed outdoors or indoors.

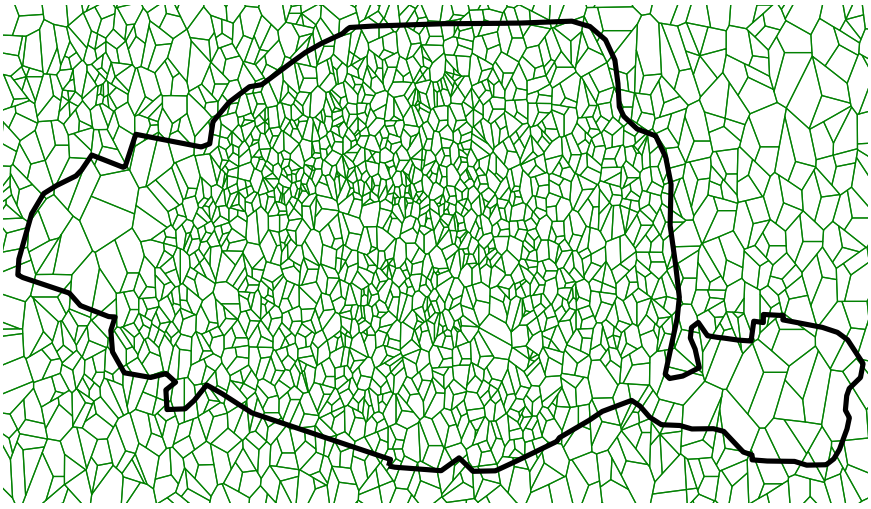


Figure 3.3. Example of Voronoi geometries (in green) generated by the 4G BS inside Paris (in black).

3.6.2. The Voronoi tessellation

The propagation path of an antenna is a significantly complex subject, which takes into account antenna type, power levels, antenna positioning, topography of the space, surface types, and reflections. Usually, the MNO has propagation maps for their deployments, but this information is not easily available and hard to utilize.

In order to have an estimation of a simplified propagation model of antennas, a common approach has been the Voronoi tessellation [166]. Its objective is to subdivide a 2-dimensional space into multiple sub-spaces, according to the nearest neighbors of a given set of points. In the case of mobile networks, the space will be the representative coordinate system of the area where the network is deployed (i.e., represented by latitude and longitude coordinates), with the set of points representing the base stations. Therefore, for any given user that is located inside the tessellation, the closest antenna to that user will be the antenna that generates the polygon in which the user is located. In denser deployments, such as the urban center of Paris seen in Figure 3.3, the size of Voronois will be significantly small, allowing for a fine spatial granularity. As expected, deployments in rural and less populated areas are significantly less dense, which indicates that the limits of the spatial grow in those scenarios, limiting *fine-grained analysis*.

It's actually not necessary to even leave urban centers to see the density of deployments reducing: in the example of Paris in Figure 3.3, it's possible to observe that the deployments in the two main parks of the city (Bois de Boulogne and Bois de Vincennes, located at the most western and eastern parts of the black contour, respectively) the area of the Voronois significantly grow, since the operator does not deploy as many antennas in those areas as it does inside the denser urban center.

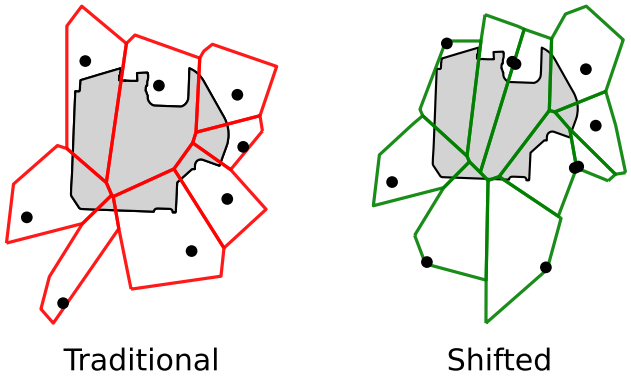


Figure 3.4. Example of traditional (in red) and shifted (in green) Voronoi geometries and their expected coverage in relation to the area of a park.

3.6.3. Leveraging the antenna azimuth to improve precision

As mentioned before, both the latitude and longitude values of Table 3.4 refer to the location of the BS. Therefore, all antennas located in this deployment will share the same coordinates, and result in the same Voronoi polygon for all. While this is not a problem for coarser spatial analysis, it can become a problem in finer-grained analysis, i.e., trying to determine the traffic produced in a highway or a park.

A trick that can be used to overcome this limitation and produce slightly more accurate representations is to utilize the azimuth angle to guide the generation of the polygons. This is possible because whenever an antenna is in a direction (i.e., has an azimuth angle), it's expected that its coverage will be better at a certain direction, which is not fully reflected if this antenna is being represented by the Voronoi of the BS. This can be done through the same process of generating Voronois, but with a slight change of the coordinates of the antennas based on the direction of their azimuth. The resulting differences can be observed in Figure 3.4, where the classical (left, in red) and shifted (right, in green) voronois are shown. Here, the objective was to find the antennas with the best *exclusive* coverage of the park area (in gray), i.e., the polygon overlaps as much as possible the park polygon. The importance of this technique will be further explored in Section 5.2, but the key point here is that those shifted Voronois result in a slightly improved idea of antenna coverage, which is extremely useful when trying to isolate traffic consumption in certain areas, especially in cases that require a fine spatial granularity.

3.6.4. Converting Voronoi geometries to other spatial units

The spatial granularity at the antenna level can be represented by Voronoi geometries that roughly represent the coverage area of each antenna, with traffic uniformly distributed over each geometry [167]. This may not be ideal for understanding intra-city

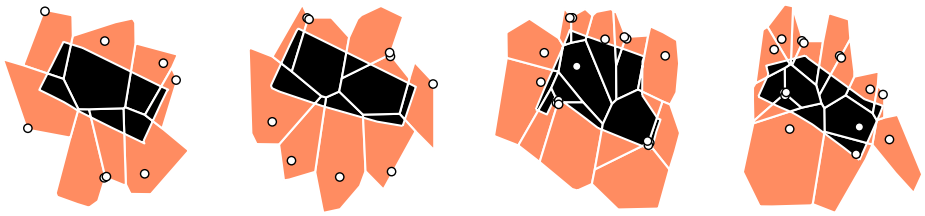


Figure 3.5. Examples of the conversion of spatial resolution, from Voronoi (in orange) to IRIS (in black), for different IRIS in central Paris. In this case, the % of the area of each Voronoi overlapped will represent its traffic inside the IRIS, with the remaining area in orange not counted as traffic for the selected IRIS.

dynamics, which are better represented by sub-municipal divisions called Islets Regrouped for Statistical Information (IRIS), defined by France's National Institute of Statistics and Economic Studies (INSEE) to better reflect changes in urbanization, geography, and demographics in sub-municipal statistics in France [168], acting similarly to US census tracts. Metropolitan France is represented by over 49,000 IRIS units.

Mobile traffic can be converted from Voronois to IRIS by calculating the intersection ratio of the Voronoi geometry that represents the space covered by each antenna with the set of IRIS geometries from each city. The traffic of IRIS i will be composed by the sum of the traffic ratio of every Voronoi $n \in N$ that intersects it, defined as:

$$T_x = \sum_{n}^{N} r_n T_n \tag{3.5}$$

where N is the total of Voronois that intersect IRIS i, r_n is the area ratio of the intersection of Voronoi n with IRIS i and T_v is the total traffic of the respective Voronoi.

This process is visualized on Figure 3.5 for 4 different IRIS located. As represented by Equation 3.5, the traffic of each IRIS (full area in black) will be represented by the sum of the traffic ratios of each Voronoi that overlaps with it, with the remaining traffic of the Voronois (represented in orange) not accounted for that specific IRIS.

It's interesting to note that this process is valid for any other geometry; the example utilizes IRIS as this is the geometry used throughout the rest of the thesis. For example, this same conversion could be done to calculate the traffic on a square grid (also seen in the literature). This process could also be done to convert to other bigger-sized geometries, such as city boundaries, in opposition to calculating the city traffic by just summing the traffic of all antennas located inside the city. This is helpful when for example calculating the traffic in cities in rural areas of countries, where the deployment is significantly less dense and Voronois can be quite big, even spanning through multiple cities. It's important to remark that there's a diminishing factor here: if the area to be converted is significantly larger than the Voronois that intercepts it, differences in calculation vs. just adding the traffic of antennas inside the geometry will reduce.

4 User adoption of new mobile technologies

Through the evolution of mobile networks, understanding how the users connected to the MNO's structure are utilizing newly deployed products is essential to assess the performance of these new technologies, the changes in preferences of service consumption, and any bottlenecks presented in the service that could be better addressed by the operator. With large-scale collections of network data, it's possible to obtain very early insights in a passive manner, without relying on surveys or expensive market research, which can aid researchers, engineers, and analysts in having a deeper understanding of the needs of their customers throughout the roll-out of newly deployed products.

This chapter will explore how it's possible to leverage measurements from production networks to assess the early adoption of new technologies. It will be structured as follows: Section 4.1 will contribute with a first-of-its-kind study about the adoption of 5G throughout mainland France, focusing on how mobile services may have their traffic consumption patterns affected, and how this can relate to space and socioeconomic indicators of areas in the country. Section 4.2 will contribute with another first-of-its-kind study on the adoption of indoor mobile networks, which are becoming a crucial point of network deployments, and how traffic consumption patterns may differentiate themselves from the ones usually observed in outdoor environments.

4.1. Characterizing 5G adoption and its impact on traffic

As the deployment of 5G networks is advancing globally, understanding the evolution, performance, and impact on users of this new generation of cellular technology is critical. However, due to industrial secrecy and competition in an aggressive market, operators tend to publicly disclose minimum information about their planning strategies for 5G or about the impact that the technology has on their users.

Market analyses by Ericsson [10] indicate that the adoption of 5G is well en route, with a declared coverage of the world population that has reached 35%. In Western Europe, which will be the focus of this study, that figure was already at 79% at the end

of 2022. The improving coverage is in turn uplifting the adoption of the technology by users. As a matter of fact, the total number of 5G subscriptions is projected to reach the 1.5 billion mark in 2023, with a growth of 500 million within the last 12 months, and a twofold increase in the past two years.

While these figures are interesting, they only provide a high-level picture of the adoption of the 5G technology. Many interesting questions related to whether and how the availability of a higher-performance wireless technology is affecting the way subscribers consume mobile traffic and services remain unanswered. For instance, there's little clarity on how the 5G technology is employed by users over space and time, and whether such patterns differ from those observed for 4G; if 5G is changing the utilization of specific mobile applications; or, whether there are specific portions of the population that are adopting the technology faster than others.

This section provides an in-depth analysis of the utilization of 5G by the users of a major mobile network operator, bringing forth a new understanding of the impact of 5G on the spatial and temporal dynamics of both total traffic and service-level demands.

4.1.1. Data processing for the analysis of 5G deployments

This study is built on top of a 3-month collection period, between March of May 2023, with a 5-min granularity. The collection follows the process described within Section 3.3, and the identification of 5G NSA flows within the measurement of TCP and UDP sessions follow the procedure described in Section 3.4.

The collected data is aggregated into BS-level, where each BS is expected to have a certain coverage area. However, the analysis of this section will take into account the area of cities and, within urban levels, the smallest urban unit present in France (IRIS). Therefore, it's necessary to perform an interpolation of data from BS-level to IRIS-level, as described in Section 3.6.4. Considering that the coverage of each BS can be associated with a Voronoi geometry, it's possible to assign BS-level traffic to IRIS level.

With this done, the next step is to understand in which IRIS 5G coverage is well established. More specifically, it's essential to understand if each IRIS is well covered by 5G antennas or if it may have a small number of antennas which are not covering a significant portion of space. This is well illustrated by Figure 4.1b and 4.1c, which presents how different IRIS can have significant diversity of 5G coverage, even though both have 5G antennas deployed in their area: the area of the IRIS in Figure 4.1c is almost entirely covered by BS with only 4G deployed, with the 5G deployment having an extremely short coverage; meanwhile, Figure 4.1b has most of its area well covered by antennas that are deploying both technologies. Therefore, for the remainder of the analysis, it will be important to focus only on the areas represented by the later example, while examples similar to the first one shall be filtered out to improve the quality of results. The results of this filter can be seen in a quantitative way in Figure 4.1a, which presents the number

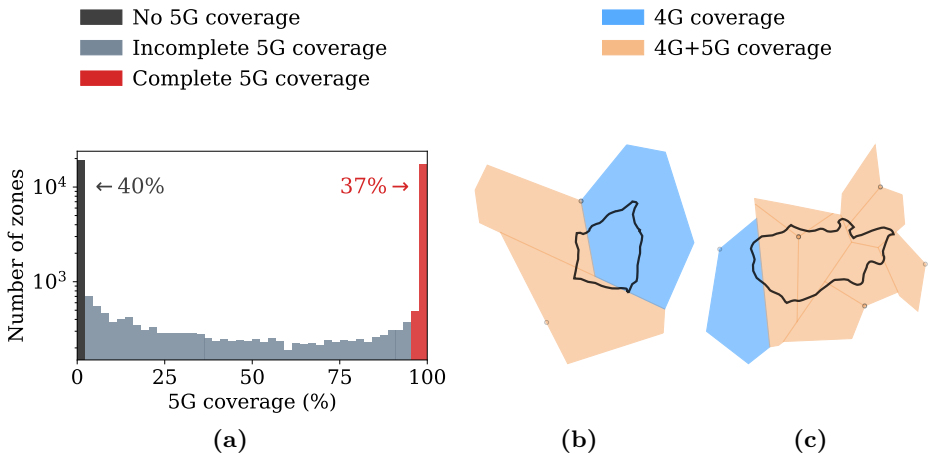


Figure 4.1. (a) Histogram of the 5G coverage across IRIS. Examples of IRIS that are covered by (b) 4G only and (c) 4G and 5G.

of IRIS with a percentage of their surface covered by 5G-enabled sites. This histogram has two clear peaks around 0% and 100%, meaning most areas in France are either not or fully covered, meaning 5G is still far from pervasive in the country. Indeed, only 36% of the IRIS have 5G covering more than 95% of its area (represented in red in Figure 4.1a).

Therefore, the rest of the study presented in this section will focus on the L=18,014 (out of 48,949) that have a complete 5G coverage, as those are expected to provide a significantly good 5G coverage to the user base, driving better utilization and adoption and leading to useful insights.

4.1.2. Overview of nationwide 5G adoption

This work starts by providing a high-level analysis of 5G adoption in France, exploring the fraction of traffic 5G users currently generate, and how it varies in space and time.

4.1.2.1. How much traffic does 5G generate?

In order to quantify the incidence of 5G on mobile data usage, it will be employed throughout the section the notion of *5G ratio*, i.e., the fraction of broadband traffic generated via the 5G technology. Formally, the 5G ratio measured at a given statistical zone ℓ and over a time period t is defined as:

$$R^{\ell}_{5G}(t) = \frac{v^{\ell}_{5G}(t)}{v^{\ell}_{4G}(t) + v^{\ell}_{5G}(t)}, \tag{4.1}$$

where $v^{\ell}_{\star}(t)$ is the volume of data traffic generated by technology $\star$. It's interesting to note that this analysis only considers 4G and 5G in the definition, since its interests are in high-speed service capable of supporting modern applications, for which 4G and 5G

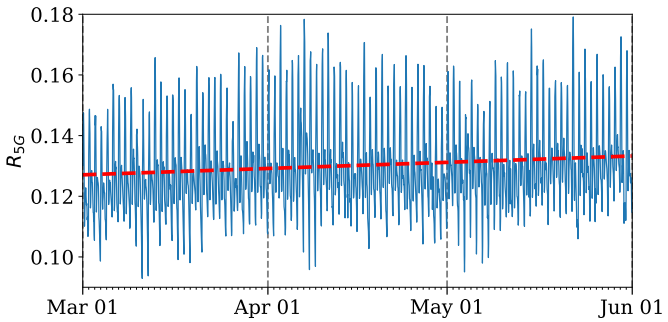


Figure 4.2. Nationwide 5G ratio, R_{5G}, computed on a hourly basis during the three-month observation period. Linear trend in red.

are the two major competitor cellular technologies. In fact, 2G and 3G demands are negligible in the context of the study, i.e., they contribute less than 0.5% of the overall traffic in the studied urban areas, hence they do not affect the value of $R^{\ell}_{5G}(t)$. Based on (4.1), $R^{\ell}_{5G}(t) = 0$ if 5G demands are absent and the mobile traffic is only generated by 4G devices, whereas $R^{\ell}_{5G}(t) = 1$ if 5G has already seized the whole user demand.

Figure 4.2 shows the evolution of the nationwide 5G ratio $R_{5G}(t) = 1/L \cdot \sum_{\ell} R^{\ell}_{5G}(t)$, at hourly time steps t and over the full data collection period. It's possible to observe that $R_{5G}(t)$ ranges typically between 0.10 and 0.18, i.e., 5G users are responsible for 10% to 18% of the overall mobile data traffic served by the operator across the whole country. The average value recorded over the three-month period is 0.1302, which can be used as a high-level figure for the penetration of 5G demands: in other words, in geographical regions where 5G service is available and well provisioned, 5G users typically generate 13% of the total mobile data traffic.

Interestingly, the average nationwide 5G adoption is not constant in time: it increases from 12.7% on March 1 to 13.3% on May 31. The dashed red line in Figure 4.2 reports the result of a linear fit on the data, which displays a clearly positive slope. Therefore, the incidence of 5G demands is steadily growing within areas where the new RAT is present, exposing how 5G is gaining momentum. While the trend is expected, from the fitting it's possible to quantify the effect and estimate a mean increment of 0.05 percent points per week at a national scale.

Key insights. *In areas of France where the 5G technology is pervasive, the 5G demand accounts for around* 13% *of the overall mobile broadband traffic. According to projections from the observed period, that percentage is growing at a pace of* 2.4 *points per year. These figures indicate that the adoption of 5G is still at early stages in France, and relatively slow in gaining 4G market shares.*

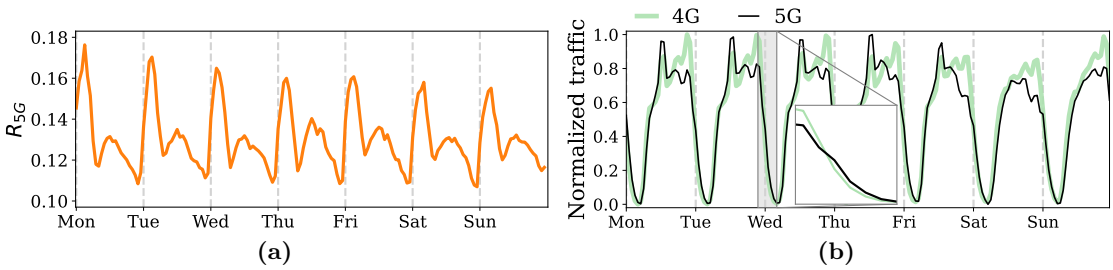


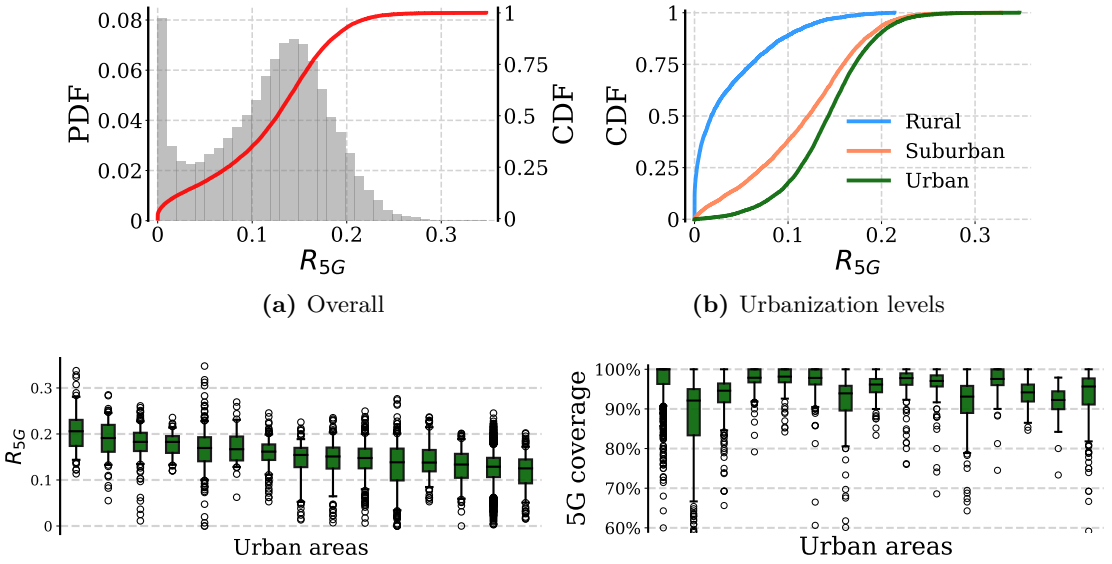
Figure 4.3. (a) Median week of 5G ratio in total traffic. (b) Median weeks of the traffic demands separated by technology.

4.1.2.2. How is 5G traffic consumed over time?

Figure 4.2 shows that the 5G ratio of traffic aggregated over the whole country, $R_{5G}(t)$, undergoes substantial fluctuations between different hours, which prompts a deeper investigation of temporal patterns in 5G adoption. To this end, it's possible to look at *median weeks*, i.e., one-week time series that summarize in a compact way the typical fluctuations of the target temporal phenomena, as the median value of each hour of the week across all weeks in the dataset. Figure 4.3a depicts the median week of $R_{5G}(t)$, and unveils a very clear daily periodicity: the fraction of mobile data traffic contributed by 5G in the whole country has (*i*) a high peak overnight, between 22:00 and 7:00 (8:00 during weekends), (*ii*) a smaller second peak in the first part of the day, and (*iii*) a drop in the afternoon until 22:00.

To ease the interpretation of the result, Figure 4.3b shows min-max normalized time series of the median weeks of nationwide 4G and 5G traffic. It's important to recall that these time series refer to statistical zones where both technologies are pervasive, as per Section 4.1.1, hence can be fairly compared. It's possible to observe that 4G and 5G traffic exhibit notably different time patterns. First, dynamics are similar throughout the morning, which results in $R_{5G}(t) \sim 0.13$ in Figure 4.3a during that period: that is, 4G and 5G users behave in a similar way in the first part of the day, and the 5G ratio is fairly stable around its typical value recorded in Section 4.1.2.1. Second, patterns diverge in the afternoon, as 5G demands drop while 4G traffic grows, which explains the decrease in $R_{5G}(t)$ in the second part of the day. Third, as seen in the inset plot, the trend switches from 22:00 through the night, as the 5G traffic curve has a less steep slope leading to higher (relative) values overnight, with $R_{5G}(t)$ peaks that last until early in the morning.

Key insights. *The incidence of 5G is not uniform over time, rather it follows a neat circadian pattern with fluctuations that make the 5G incidence almost double within each single day. The reason is that, quite surprisingly, 4G and 5G demands do not follow the same temporal dynamics throughout the day. The reasons for this phenomenon will be investigated in Subsection 4.1.4.2.*



(a) Overall

(b) Urbanization levels

(c) R_{5G} across urban cities

(d) 5G coverage across urban cities

Figure 4.4. (a) CDF of the overall 5G ratios over the total traffic across zones. (b) Breakdown of CDFs across urbanization levels. (c) Distribution of the ratio of 5G traffic across multiple cities. Boxes represent the range between the first and third quartiles, and encase the median line. Whiskers represent the 5$^{\text{th}}$ and 95$^{\text{th}}$ percentiles, and fliers are outside this range. (d) Percentage of 5G enabled antennas the considered urban areas.

4.1.2.3. Where is 5G traffic consumed?

All results presented before are aggregated over the whole country, yet 5G adoption is not necessarily heterogeneous over the territory. Figure 4.4a shows the Probability Density Function (PDF) and Cumulative Distribution Function (CDF) of the mean 5G ratio computed at statistical zones ℓ, i.e., $R^{\ell}_{5G} = 1/T \cdot \sum_t R^{\ell}_{5G}(t)$, where T is the number of hourly time steps t in the dataset. The distributions show how different zones experience very different 5G traffic incidence, ranging all the way from zero in some zones up to 35% in others. In particular, the PDF spotlights a clear peak of almost 10% of zones without 5G users. It's important to recall that all zones considered enjoy pervasive 5G coverage, hence RAT availability cannot be the root cause of the diverse adoption.

A great portion of the spatial diversity, including the peak at R^{ℓ}_{5G}=0, is in fact ascribed to the urbanization level of the region where the zone is located. Figure 4.4b breaks down the R^{ℓ}_{5G} CDF across urban, suburban and rural areas that are covered by 5G, and exposes how lower incidence almost exclusively emerges in rural 5G deployments, where half of the zones experience less than 5% 5G traffic; by contrast, no statistical zone in urban areas has such a low 5G adoption. Similarly, the percentage of zones where 5G users generate more than 10% of total traffic, i.e., R^{ℓ}_{5G}>0.1, is just 10% in rural regions but jumps to more than 80% of urban environments. It's interesting to note that in France, the distribution of IRIS per urbanization level is 64% rural, 21% suburban and 15% urban.

In light of these observations, the focus of the analysis will be on the major conurbations: the 15 largest metropolitan areas in France, which jointly include 35% of the urban and suburban zones where 5G incidence is significant as per Figure 4.4b. The 5G ratios R^{ℓ}_{5G} recorded in all statistical zones of one city are summarized into a single candlestick in Figure 4.4c: by juxtaposing the summaries of different cities, fairly comparable 5G ratios are observed, with medians that indicate a typical 5G adoption well above 10% in all cases, vary by 0.05 at most, and have a similar span of percentiles.

In fact, Figure 4.4c reveals how the spatial heterogeneity of 5G adoption is much larger *within* cities than *across* them: the intervals between the 5^{th} and 95^{th} percentiles tend to traverse 15 to 30 percent points in the 5G adoption across zones within a same city. In other words, almost all the variety observed in the CDFs in Figure 4.4b can be found within each single city. This adoption is also not related by a lack of coverage in those areas, as Figure 4.4d shows that the vast majority of IRIS of the studied urban areas are well covered by 5G-enabled BS.

Key insights. *5G adoption remains a prominently urban phenomenon to date in France, with all major cities experiencing similar levels of incidence of 5G demands on the total traffic. Yet, 5G usage becomes highly diverse among neighborhoods of each conurbation. These intra-city spatial diversities will be investigated in detail in Subsection 4.1.4.2.*

4.1.3. A service-level perspective on 5G adoption

The introduction of a new RAT is a unique opportunity to explore if users of specific mobile services are taking particular advantage of the more performing communication technology. The next portion of this section will focus in how 5G adoption breaks down across individual mobile services, and how it affects their behavior.

4.1.3.1. Is 5G adoption uniform across mobile services?

The first focus will be in whether the introduction of 5G has impacted the popularity of services among users. Figure 4.5a shows a ranking of the first 100 mobile services, based on the their 4G traffic volume (black line), with associated 5G volumes (green dots). Both 4G and 5G service-level traffic volumes are normalized by the demand of the first service in each technology. The quasi-linear curve over the logarithmic ordinate confirms the well-known power-law behavior of service-level demands [34], [39], which are highly skewed and encompass six orders of magnitude within the first 100 applications. More interestingly, the 5G traffic dots are mostly well aligned along the 4G curve, indicating that 5G does not bring any substantial change in the popularity ranking of services.

More interesting patterns emerge when considering the 5G ratio at a per services level. Figure 4.5b shows the PDF and CDF of the nationwide time-averaged 5G ratio,

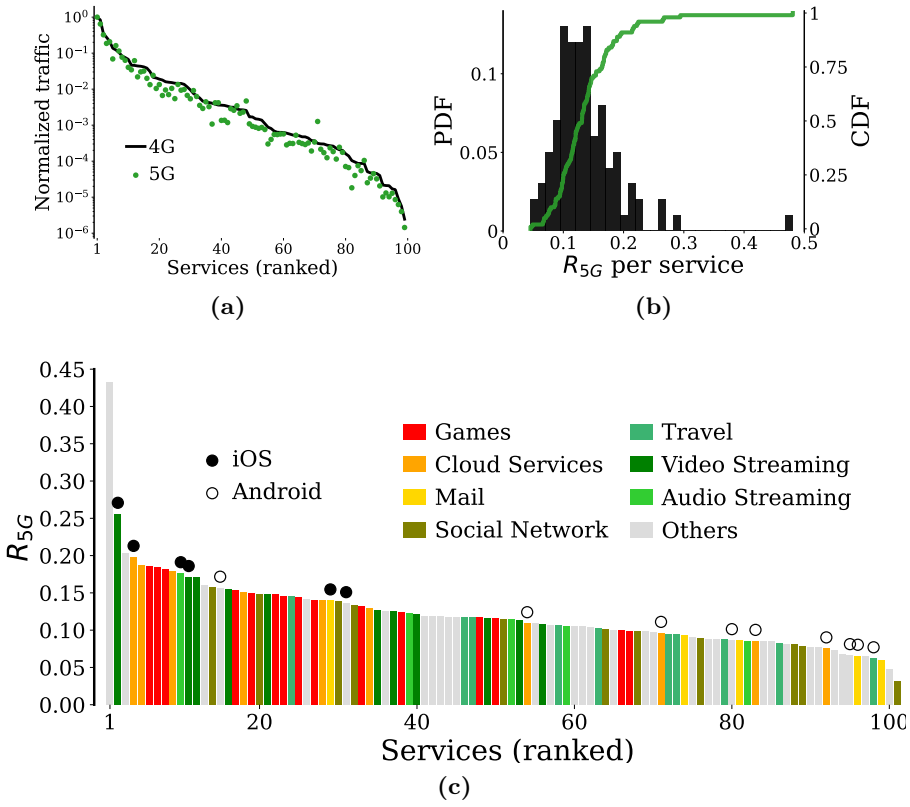


Figure 4.5. (a) Ranking for the top 100 services by their total traffic volume, for 4G and 5G, with values normalized per technology. (b) PDF and CDF of the 5G ratio across services. (c) Ranking of mobile applications based on their 5G ratio.

i.e., $R_{5G} = 1/(LT) \cdot \sum_\ell \sum_t R^\ell_{5G}(t)$, computed for each individual mobile service. The distributions show that the incidence of 5G is very heterogeneous across applications: the Gaussian-shaped PDF has an average around R_{5G}=0.13 that reflects the overall 5G ratio discussed in Section 4.1.2.1, but then spans all the way from services that have less than 5% 5G traffic to others than have almost 50% of their demand served by 5G already.

Figure 4.5c offers a more in-depth view on such a diversity, as a ranking of applications based on their R_{5G} value. Services are also associated to specific classes (i.e., colors), and those that are highly platform-specific (e.g., Apple Music for iOS, or Google Play Store for Android) are tagged with filled (iOS) or empty (Android) markers. Loose patterns can be noted at the level of service classes, such as gaming, video streaming, and Cloud services having a tendency to display higher 5G ratios, yet no strong conclusion. A more clear correlation emerges between R_{5G} and device types: iOS-specific applications consistently appear at the top of the ranking (with an average R_{5G}=0.18), while Android-based ones are gathered at the bottom (with an average R_{5G}=0.09). This suggests that Apple users are well ahead of the 5G adoption curve, which aligns well with the fact that they tend to be early technology adopters [169].

Key insights. *While the popularity of mobile services is not impacted by the introduction of 5G, individual services do show very a diverse incidence of 5G in their demands. This diversity is not strongly linked to the type of application, but the analysis suggests that the penetration of 5G among Apple users is around twice that observed in non-Apple users.*

4.1.3.2. Does 5G affect the way services are consumed?

The analysis of 5G so far relies on the demand of traffic, understanding the impact of 5G in mobile applications in relation to the *usage volume*. A different question could be elaborated by understanding if the availability of 5G changes the *way* mobile apps are being utilized by users.

The uplink-to-downlink (UL/DL) traffic ratio is proposed to represent the imbalance between traffic from and to the mobile device, being a good indicator of behaviors for how users interact with mobile applications [34]. Figure 4.6a presents the CDFs of the UL/DL ratio for the top 100 apps for 4G and 5G. While the CDFs may look relatively close between technologies, substantial differences emerge within a closer examination (represented by the inset plots): the CDFs cross each other, leading to the belief that 5G entails a higher probability of very low UL/DL ratios (i.e., applications with large download volumes, in the left inset) as well as a higher probability of high UL/DL ratios (i.e., services with substantial uploads, in the right inset). For the second behavior, it can be noted that only a single application has a UL/DL ratio higher than one in 4G, whereas around 7% of services break that barrier in 5G.

To expand this new relation between uplink and downlink seen in 5G traffic, the percent change of UL/DL ratio between 4G and 5G is computed. By denoting the UL/DL ratios for service s in 4G and 5G as ρ^s_{4G} and ρ^s_{5G}, respectively, then the percent change of s will be:

$$C^s = \begin{cases} 100\left((\rho^s_{5G}/\rho^s_{4G}) - 1\right) & \text{if } \rho^s_{5G} \geq \rho^s_{4G} \\ -100\left((\rho^s_{4G}/\rho^s_{5G}) - 1\right) & \text{otherwise.} \end{cases} \tag{4.2}$$

where positive values of C^s correspond to the increase of upload induced by 5G for s, with negative values meaning an increase of download in 5G. As in the case, C^s=40% means that the fraction of traffic of s due to upload grows by 40% in 5G with respect to $4G$ (leading to a reduction of the importance of download). C^s will be 0% if the UL/DL ratio does not vary across technologies.

Figure 4.6b presents the top 100 services ranked by C^s and helps clarify that the differences previously seen in Figure 4.6a between technologies are related to variations due to specific services. It's notable that a few mobile apps have an increase in upload traffic in 5G (positive changes, left part of the figure), while others have an increase in

(a)

(b)

Figure 4.6. (a) CDF of UL/DL ratio across services in 4G and 5G. (b) Percent change in the UL/DL ratio from 4G to 5G.

download traffic in 5G (negative changes, right part of the figure). Those changes can be quite drastic, with a quarter of the services having absolute changes above 50%, leading to the conclusion that the availability of 5G coverage has a significant impact on the way users consume mobile services. In some edge cases, services see a change above 100%, indicating upload/download traffic more than doubled due to the newer RAT.

Finally, it's interesting to see if those observed changes have any relation to the app categories (represented by colors in Figure 4.6b). It's possible to note that gamin, e-mail, and cloud apps are the main ones with an increase in upload prevalence due to 5G, while audio and video streaming lead to changes in download traffic. This could relate to the nature of those categories: the availability of better 5G coverage with better-expected performance than the previous 4G results in users who are more inclined to exploit the capacities of those apps when using mobile networks, for both uplink and downlink directions (a pattern that was usually reserved for Wi-Fi).

Key insights. *Access to 5G networks leads to users interacting with mobile applications differently; the improved capacity leads to users being more comfortable pushing the limits and intensifying the usage of applications within mobile networks, closing the gap with classical fixed Internet connections.*

(a) **(b)**

Figure 4.7. (a) Average $R^s_{5G}(t)$ median weeks and (b) breakdown of composition across service classes, for Clusters A and B.

4.1.4. Mobile services and spatiotemporal 5G usage

This study so far explored the diversity of 5G individually across time (i.e., over the week) and space (i.e., over urban zones). However, the observed heterogeneity in these dynamics can be traced to the spatiotemporal variations of individual services within 5G, leading the next part of the study to explore those.

4.1.4.1. How do services contribute to 5G time dynamics?

The next step of this work starts with clustering the median week of 5G ratio of each service s, in order to identify recurring patterns in the temporal relation of 5G consumption against 4G in mobile applications. A hierarchical clustering using correlation distance is proposed across $R^s_{5G}(t)$, with the following results: Cluster A has a set of applications with the 5G ratio extremely correlated among them; Cluster B is characterized by a set of apps that are basically not similar to Cluster A or even amongst themselves. Figure 4.7a presents the median $R^s_{5G}(t)$ of each of those clusters, where two remarks can be noted: 1) during day-light active hours (7am-10pm) the dynamics of both clusters are quite similar, aligning with the behavior previously mentioned in Figure 4.3a of lower 5G incidence in the afternoon; 2) the temporal patterns of $R^s_{5G}(t)$ diverge between both clusters during night/evening hours, where applications within Cluster A show a higher peak of 5G incidence.

Next, it will of interest to explore which applications are within each cluster. Figure 4.7b presents those results. Cluster A is characterized by video streaming and social network applications, which have a higher night incidence of 5G usage, while mail, cloud, and gaming apps are more present in Cluster B and do not have this peak. Together with the observations made previously on the temporal analysis, a few insights can be made: 1) the fluctuations in Figure 4.3a during the active hours are related to application categories, as the two predominant service temporal patterns in Figure 4.7a present the

same dynamic of a higher 5G incidence in the morning that progressively diminishes in the afternoon; 2) the peak during the evening in Figure 4.3a can be related to a subset of mobile services, by observing the overlap of categories in Cluster A and in the right of Figure 4.6b, which present applications with heavy download traffic which generates higher demands in the network at all times, hiding the pattern of other services in the total traffic plots.

Key insights. *During daylight hours, the contributions of specific services to the 5G traffic demand are uniform; this changes during the evening, when services are split between download heavy (which see a surge of 5G usage) and other work-related services (which do not experience this surge).*

4.1.4.2. Is there a spatial component to service-level 5G usage?

The final part of the analysis focuses on understanding if the service-level temporal patterns have a spatial component (i.e., do different areas of the cities experience 5G differently?). A choice is made to prioritize for this spatial analysis the services that are present in Cluster A, as those had significantly different and interesting patterns overnight in relation to the ratio of 5G traffic.

The median week of 5G ratio at each IRIS ℓ of the 15 major cities in France will be calculated, for services only in Cluster A (denoted as $\mathcal{S}_A$). The 5G ratio for the whole cluster A at each location ℓ and time step t will be computed, indicated by $R_{5G}^{\ell,\mathcal{S}_A}(t)$, using expression (4.1) on the total 4G and 5G traffic generated by all applications in $\mathcal{S}_A$. Then, the hourly median of such values is calculated at each location ℓ during a week.

Within each city, these median weeks of each ℓ are clustered with a similar technique of Section 4.1.4.1 (hierarchical clustering algorithm on the Euclidean distances between the median weeks of $R_{5G}^{\ell,\mathcal{S}_A}(t)$ for all locations ℓ of a same city). With those results, the final set of clusters determined (by analyzing the Silhouette score) is 2 (which will be from now on denoted as *red* and *blue* clusters, as represented by Figure 4.8). Observing together with Figure 4.8d, which represents the average median week of both clusters across all studied cities, it can be noted that the blue cluster has overall zones with lower adoption of 5G and that those zones do not show the overnight peaks previously observed in Figure 4.7a. Also, it can be noted that the red cluster has a higher adoption of 5G, with the peaks of $R_{5G}^{\ell,\mathcal{S}_A}(t)$ clearly visible. The maps also show that there are spatial patterns across IRIS of each cluster, which indicates that there's a spatial component to these trends observed in the 5G adoption in France.

This leads to the question of identifying which factors could explain this diverse adoption of 5G in urban spaces in France. For this, a number of features about the socioeconomic status (SES) and land use that characterize each statistical zone in the 15 cities in the dataset are collected. Specifically, it will be explored SES metrics that capture the educational level (e.g., fraction of the local population with a university degree),

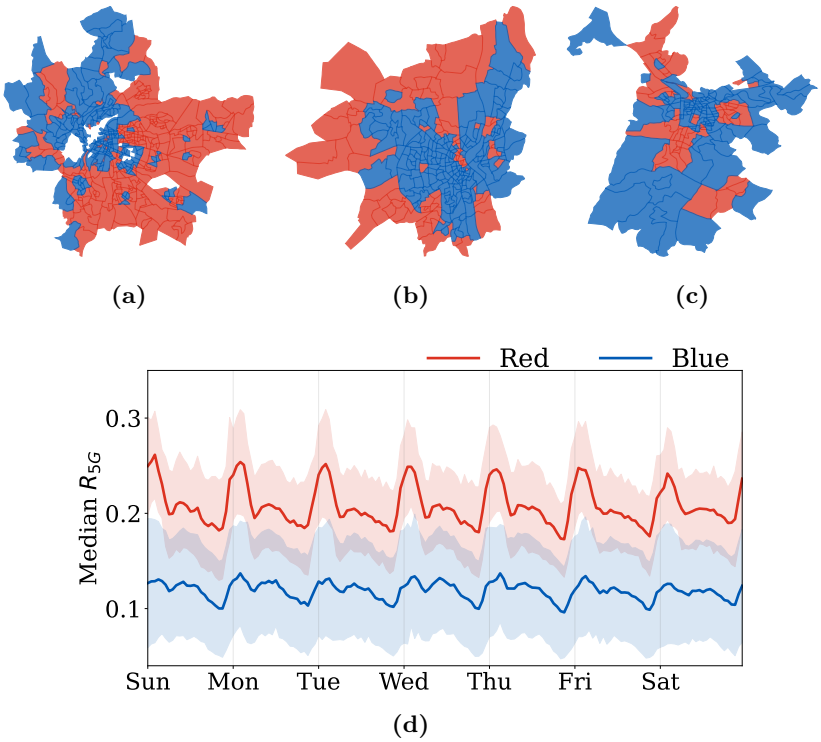


Figure 4.8. Maps of the blue and red clusters obtained from the clustering of $R_{5G}^{\ell,\mathcal{S}_A}(t)$ across city locations for (a) Lyon, (b) Bordeaux and (c) Grenoble; (d) Average median weeks (with standard deviation) of $R_{5G}^{\ell,\mathcal{S}_A}(t)$ across all cities, for the blue and red clusters.

economic status, (e.g., deciles of the distribution of income by the local population), employment status (e.g., fraction of the local population with executive or intellectual professions), and population status (e.g., fraction of the local population within multiple given age ranges). It will also be considered land use statistics, such as the fraction of the surface of each statistical zone that is covered by, e.g., residential or commercial buildings, industrial infrastructures, religious or sports facilities.

A Random Forest (RF) is trained on such features, with the aim of predicting whether a statistical zone belongs to the blue or red cluster based on SES and land use information. The RF consists of 100 decision trees, each with a maximum node density of 100. The Gini index is used to partition the nodes during construction, where each tree can consider 8 features at the time of partitioning. In addition, to avoid overfitting, a condition is imposed that at least 10 samples must be considered for partitioning and that each leaf node must contain at least 3 samples.

Overall, the provided features allow the RF model to determine with a mean 0.79 F1 score, i.e., a fairly high accuracy, the class of the statistical zone. In fact, when looking at feature importance, in Figure 4.9, it can be noted that land use features are given a negligible weight and essentially not used for classification. Therefore, it can be concluded

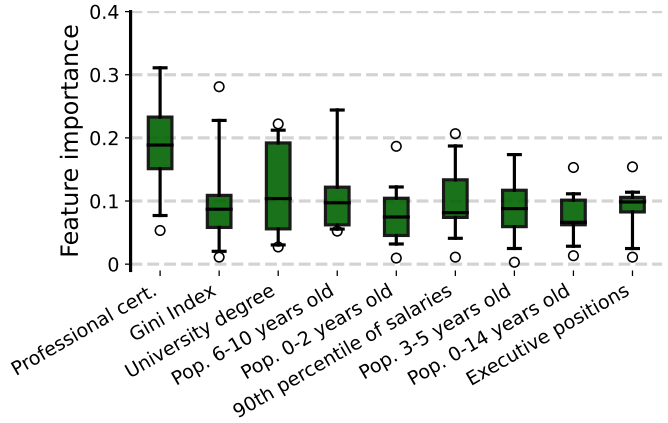


Figure 4.9. Importance of the main features used by the RF classifier of the statistical zones across blue and red clusters.

that SES is what primarily drives the temporal pattern in 5G adoption across the urban regions, and in particular, allows explaining where the surge of 5G incidence overnight that characterizes the statistical zones in the red cluster could be expected.

By analyzing the partial dependence of features used on the RF model, it can be seen that on an aggregate level, the RF model tends to consistently classify low-income and less-educated zones as red, i.e., to match them with the statistical zones marked by the typical nighttime peaks in the 5G ratio. This correlation with social status above appears to be consistent with the other features. A hypothesis could be laid that inhabitants of neighborhoods that are less affluent could be taking advantage of the availability of 5G 'dongles', i.e., small modems that allow creating a Wi-Fi network exploiting the 5G cellular connectivity, to get high-speed Internet access at home at a much lower cost than with regular fiber. This would explain the overnight peaks, due to, e.g., high-quality streaming or automated large file download activities that may characterize home usages during the evening and later into the night. While speculative, this account would be consistent with other effects, e.g., the fact that those same less rich areas record a higher incidence of 5G traffic at all times (see Figure 4.7a), or that services use intensifies in a home-like way with 5G (see Figure 4.6b).

Key insights. *While the daylight 5G demand is uniform across time and space, the higher demand of 5G during night hours can be linked to a set of services with higher download traffic, which are consumed in specific neighborhoods of cities. An explanation of those overnight peaks could be made with socioeconomic status indicators of those areas, where it can be determined that significant consumption of 5G is when users are at home and in less affluent urban areas in France.*

4.1.5. Main takeaways

This study was a first-of-its-kind analysis of 5G adoption and its impact on service consumption, using large-scale measurements collected in an operational nationwide mobile network. It takes an original network and service-oriented perspective on the utilization of the 5G technology, which sets it apart from previous works that looked into 5G performance, mainly from the end user viewpoint.

As such, this approach unveils a number of interesting and partially unexpected patterns in the incidence of 5G demands:

- The adoption of 5G is still at relatively early stages in the target country, with a low growth rate and consumption that is very strongly oriented in the main urban areas, although with large discrepancies within cities;

- The availability of high-speed 5G connectivity affects services and devices (e.g., iOS versus Android) in very diverse ways, and pushes users to consume applications much more intensively, seemingly towards closing the gap with typical usages of fixed Internet access;

- The fraction of overall mobile traffic contributed by 5G is surprisingly not constant in time, rather it shows a clear circadian rhythm with strong overnight peaks, which are in fact generated by a subset of download-intensive applications consumed in specific urban neighborhoods;

- The regions generating the aforementioned peak and higher 5G incidence, in general, are in fact characterized by lower education and income, suggesting that the local population makes significant use of 5G 'dongles' as a replacement for the more expensive fiber Internet access.

The main limitation of this study is that it focuses on one mobile network operator and a specific country: hence, the results are not expected to generalize globally, but are expected that they may hold for a substantial portion of Europe.

4.2. Characterizing the adoption of indoor mobile networks

Understanding the patterns of mobile network usage is a task been a task vastly explored, as observed in Section 2.1. The insights gathered by those studied drive both research within human activities, needs, and habits, but also inside mobile operators by giving a better comprehension of how the user base is interacting with the network. Interestingly, the majority of studies that exploit mobile traffic to characterize usage focus on measurements collected in outdoor environments. This could be explained by cellular networks historically being intended as a technology for mobile, outdoor users. However,

this use case is not exclusive anymore, and forecasts by industry indicate 80% of the future cellular data traffic will be generated in indoor environments [170]. As a result, 5G and beyond systems are anticipated to align with the emerging need to serve indoor users [171]: specifically, a number of major vendors and MNOs expect systems to transition from a legacy "outside-in" coverage approach, where indoor coverage is provided by antennas located outdoor, to the deployment of native Indoor Cellular Networks (ICN) [170], [172]–[174]. ICN deployments will allow improving substantially the quality of service for indoor UEs, and bring real competition to Wi-Fi technologies for the increasingly remunerative indoor market.

The emergence of pervasive ICNs makes it important to comprehend how such networks will be used. In this context, a considerable volume of research has focused on methods related to the modeling of radio propagation in indoor environments [175]–[177], the impact of the building layout on network performance [178]–[181] or the efficient planning of ICNs [182]–[185]. However, the dynamics and distinguishing features of the traffic data generated by in-building radio access network components have not been studied yet. Unlike outdoor BSs that tend to observe a general-purpose use, i.e., serve concurrently numerous subscribers engaged with diverse activities during different daily endeavors, ICNs are expected to target more specific use cases. For instance, ICNs are deployed in underground subway and train stations to compensate for the limited coverage of the outdoor wireless network. Likewise, corporate offices are equipped with indoor BSs to provide enhanced and reliable communications to support the work of their employees [186], whilst their installation is necessary in exposition centers and stadiums in order to accommodate the concentrated high-traffic demands intertwined with social events. Therefore, the ICN traffic is reasonably influenced by the context in which it is generated, which highly depends on the indoor environment type and, by extension, on the kind of activities in which users are involved. Eventually, the traffic dynamics of ICNs are expected to differ significantly from those of legacy communication system outdoor BSs due to their target to such specific scenarios.

In recent years, no research study thoroughly explored the characteristics of the traffic generated by ICNs. Unlike previous research in the field that focused on understanding and predicting the dynamics of the macro BSs traffic [23], [27], [51]–[53], the work on this section dwells upon the intrinsic particularities of indoor cellular traffic. In light of the proliferation of indoor communication systems and the establishment of private networks, the results of this analysis provide novel insights that may support an improved design and operation of these networks, e.g., via resource allocation, network slicing, caching, or energy adaptation schemes.

4.2.1. Data processing for the analysis of indoor networks

The collection of the data used in this section follows a similar procedure described in Section 3.3. After measurements were collected, the data was aggregated in one-hour intervals for the purpose of this study, with traffic from uplink and downlink being combined to utilize solely the total traffic consumed through both directions for each BS and mobile service.

This study contains a total of 4762 ICN antennas installed at more than 1000 base station sites, comprising different types of indoor environments through urban, suburban, or rural locations, with the recording period being around two months, from November 21st 2022 to January 24th 2023. Interestingly, the vast majority of those antennas are 4G, as apparently 5G was scarcely used for ICN at this stage of technology deployment in France. A total of 73 services are considered in this analysis, spanning a vast range of categories (e.g., video streaming, productivity, transportation, social media, etc...)

4.2.2. Classifying behaviors of indoor network usage

In this initial step of the analysis, the methodology of the data analysis will be presented, which should help uncover the dynamics of ICN traffic and unveil the distinct patterns that reside in the data.

4.2.2.1. Quantifying the importance of mobile apps in ICNs

Considering the total traffic generated by each indoor BS, the traffic consumed over the two-month studied period will be aggregated into a singular value, to build a $T^{N\times M}$ (of N indoor antennas by M mobile applications). The objective is to utilize this matrix to understand how services may act similarly in those environments. However, one problem to overcome when comparing the traffic of mobile services: differences in scale of mobile services traffic consumption are vast, i.e., video streaming services inherently consume more traffic than instant messaging. Indeed, clustering directly the aggregated traffic masks the impact of services and just groups them based on sheer *volume*.

The proposed solution involves using the RCA metric, introduced in Section 3.5.5. With it, it will be possible to quantify the degree of over and under-utilization of each mobile service in respect of the others, and in relation to the full set of studied BS, without having the bias induced by traffic volume. Going further, since this metric is intended to be used in a clustering approach, a transformation is needed to avoid bias due to the shape of the distribution given for the traditional RCA values (as mentioned in Section 3.5.5). Therefore, the symmetric RSCA version will be utilized for this analysis, since it results in a properly balanced distribution among entries. For the following clustering part of this study in the impeding clustering analysis, the RSCA of each mobile service will be

(a)

Figure 4.10. Silhouette score and Dunn index versus the number of clusters, serving as a stopping criterion to select the optimal number of clusters.

used to build matrix $T^{N\times M}$, separating the data based on the utilization profile of the M mobile services across the N distinct antennas.

4.2.2.2. Clustering patterns of mobile applications in ICNs

Since the initial goal of this analysis is to understand how mobile services are used differently in indoor environments, after establishing a metric capable of representing the behavior of each service M, the following step will be to cluster values of $T^{N\times M}$. Due to its simplicity and efficacy, a hierarchical clustering algorithm is chosen for this task. More precisely, it will be based on Ward's criterion [187], using the Euclidean distance between values. Due to the unsupervised nature of this algorithm, the ideal number of clusters is not known *a priory*. To help guide this decision, both the Dunn Index and Silhouette Score will be utilized.

After performing the clustering and the indexes analysis, as seen in Figure 4.10a, two values can be noted to have steps in both scores (which indicate an ideal number of clusters): $k = 6$ and $k = 9$. The second will be chosen, as it presents the highest step, meaning that this would be the highest number of partitions that could be made on the data before deteriorating cluster results.

To further understand the similarities between the 9 clusters, the dendrogram obtained from the hierarchical clustering process can be observed in Figure 4.11. From it, and considering that dissimilarity in terms of service usage is associated with distances along the y-axis, three large and previously unobserved groups of clusters can be seen, represented by the top blue connecting lines and color-coded deeper into the dendrogram in the sub-clusters colored orange, green, and red.

The first larger group comprises clusters 0, 7, and 4 (orange group), the second clusters 5, 6, and 8 (green group), whilst the third includes clusters 3, 1, and 2 (red group); the clusters found within the same group present a stronger similarity with each other. These

Figure 4.11. Dendrogram illustrating the iterative merging of antennas into clusters as returned by the hierarchical clustering algorithm run on SRCA features of individual antennas. Distance thresholds for $k = 6$ and $k = 9$ are highlighted. Colors tell apart the 9 clusters identified by the second threshold.

results reveal that if the choice was $k = 6$ instead of $k = 9$ for the number of clusters, results would only consolidate the orange group into a single cluster, instead of diving it into 3 sub-clusters, and merging clusters 6 and 8 at the second branch of the green group. For the remaining of the analysis, the 9 obtained clusters will be treated together in groups of 3, due to the resemblances seen. The goal will be to understand the differentiation of a larger group of 3, and within each larger group of 3 the uniqueness of each minor group inside it (that form the total 9).

Key insights. *RSCA enables acquiring an unbiased representation of the mobile service utilization at ICN antennas. Clustering the antennas based on their RSCA yields 9 distinct service usage clusters that can be aggregated into 3 larger groups. As expected, clusters within the same group clearly demonstrate more similarities compared to clusters in other groups.*

4.2.3. Spatial patterns of indoor mobile network usage

Although the previous analysis allowed identifying groups of clusters that are closer to each other, it does not delve into the feature importance of each cluster. Hence, an intriguing question arises: *which are the services that impact the clustering decision the most, and consequently characterize each cluster?* To answer this question, the spatial patterns of mobile services inside each cluster will be explored next.

4.2.3.1. Relation of clusters with indoor environments

By analyzing the location where each indoor antenna is located, it's possible to understand what are the spatial dynamics of the clusters previously obtained and search for patterns that could link them to those environments (i.e., do clusters gather specific types of places). Since each BS has an internal name given by the network operator

Environment	***Cases***	N_{env}
Metro	Underground railways in major cities	1794
Trains	National and regional railway stations	434
Airports	France's major airways	187
Work spaces	Corporate offices and Industrial Facilities	774
Commercial centers	Malls	469
Stadiums	Major sport event venus	451
Expo centers	Corporate, cultural and music event venues	230
Hotels	Accommodation units	28
Hospitals	Healthcare units	53
Tunnels	Highway and train tunnels	220

Table 4.1. Summary of Indoor environment types.

with may contain keywords (e.g., metro, tunnel, hospital), a text mining algorithm is used to quickly find patterns of each BS among the clusters. From this text mining, a group of indoor environments was created, as seen on Table 4.1, containing 11 categories of indoor locations (along with the number of antennas contained in each category): (i) metro stations, (ii) train stations, (iii) airports, (iv) workspaces, (v) commercial centers and shopping stores, (vi) stadiums, (vii) expo centers, (viii) hotels, (ix) hospitals, (x) tunnels, and (xi) public buildings.

The distribution of BS per indoor environment is unbalanced, as some locations are more common than others across, e.g., there are more offices than hospitals. Therefore, it's expected that network operators deploy more indoor BS in specific locations according to their presence inside cities. This imbalance in the data should not impact the analysis; as it does not use a supervised learning approach (which necessitates extra attention in the training of the model to account for the imbalanced classes), rather the problem is solved through unsupervised learning, which do not face the same issue.

By using this environment information, it's possible to associate classes of antennas with their specific surrounding conditions. A logical next step is then to quantify the type of indoor environments that reside within each cluster. A qualitative illustration of the correlation between the detected clusters and the indoor environment type is shown in the Sankey diagram of Figure 4.12, depicting how the samples of each cluster flow into the environments of Table 4.1. As expected, there are big portions of certain clusters flowing to the same environment types. For instance, metro and train stations are highly present in the orange group, while stadiums are within one of the green group clusters. Furthermore, it can be observed that the dominant flux towards workspaces originates from cluster 3, whilst clusters 1 and 2 populate the remaining environments.

This flux between clusters and indoor environmental type is quantified in Figure 4.13, which shows the percentages of each cluster in each environment type. Firstly, the orange

Figure 4.12. Sankey diagram depicting how the clusters flow into different environment types.

group clusters, seen in Figure 4.13a, comprise solely metro and train stations. Since this group represents antennas that serve users while commuting, the observed patterns in mobile services consumption through the RSCA show that this group of BS mainly utilizes music streaming services, navigation apps other entertainment-related services, which is an expected behavior while traveling. Notably, for clusters 0 and 4 more than 92% of the antennas are located in Paris and its suburbs, contrary to cluster 7 which consists solely of the Lille, Lyon, Rennes, and Toulouse metro antennas.

Another interesting pattern is within antennas of the green group, which are mainly located in stadiums, as seen in Figure 4.13b. That justifies the higher RSCA values seen for sports-related websites and the use of content-sharing applications, such as Twitter and Snapchat via which one can upload photos and information relevant to sports events. This behavior is more evident for the antennas in clusters 6 and 8, more than 75% of which are in stadiums. Remarkably, cluster 6 includes stadiums outside Paris, while approximately 60% of cluster 8 antennas are in Paris. On the other hand, stadiums make up only 35% of cluster 5, which also includes other diverse types of environments, such as expo centers, corporate offices, and commercial centers, equally distributed in Paris and other cities. Apart from the expo centers though, the remaining categories represent only a small fraction of their environment type, e.g., only approximately 5% of the commercial centers and working environments belong to cluster 5. Hence, given that cluster 5 is characterized by the under-utilization of most mobile services, it's considered to include antennas treating most of their Internet services equally, without demonstrating any advantages compared to other antennas in the data set.

For the red group, the most apparent connection is observed for cluster 3. More than 70% of cluster 3 antennas are workplaces, in particular corporate offices, which

(a) Orange group

(b) Green group

(c) Red group

Figure 4.13. Types of indoor environments per cluster.

clearly provokes a higher RSCA value for services such as Microsoft Teams, LinkedIn, and emailing applications. Notably, in contrast with the offices, the small number of industrial facilities included in the data set mostly occupies cluster 5. Expo centers also assume a strong presence, with more than 50% of them belonging to cluster 3, likely due to holding corporate events, conventions, and conferences.

While cluster 3 constitutes a single branch of the red group, clusters 1 and 2, which belong to the other branch, host many commercial centers. In particular, cluster 2 hosts 50% of the commercial centers, most of the hotels and public buildings, as well as almost all the hospitals. Notably, cluster 2 includes all the small retail shops of the network operator, which clarifies its advantage in Google Play Store usage to download apps.

Cluster 1 is more diverse, and along with commercial centers, contains almost all airport and tunnel antennas and a small percentage of all the available environment types. Given the large variety of environments encountered in cluster 1 and the absence of a dominant environment type, as well as due to the fact that it contains messaging, streaming, and music services, it can be speculated that this serves as a general-use cluster. It can be mentioned that while the antennas of cluster 1 are almost equally distributed between Paris and other cities, at around 92% of the antennas of cluster 2 are found outside Paris, and 70% of cluster 3 antennas are located in Paris and its suburbs.

Key insights. *It was possible to showcase that the clusters obtained from the analysis of mobile application usage within indoor environments were related to specific locations within the cities. Antennas in the Orange group were mainly located within metro and train stations, and had an over-utilization of music and navigation apps; meanwhile, the Red group was mainly present within commercial and office areas of city and was characterized by an over-utilization of work-related applications.*

4.2.3.2. Comparison with Outdoor Antennas

From the previous analysis, it's noted that ICN mobile traffic demands intrinsically exhibit distinct patterns that highly depend on the type of indoor environment. A natural question is whether these patterns are also present in the components of legacy communication system radio access networks. In order to answer that, the traffic generated by outdoor antennas found in proximity to the ICNs is probed from the data, in order to repeat the analysis and understand if the behaviors seen are indeed unique to indoor locations.

This portion of the study will assess whether the nature of the ICN strictly defines and differentiates its service demands from that of close-by outdoor antennas, e.g., if the service demands of the outdoor antennas nearby to corporate offices are drastically disparate from the in-building ones. Hence, for each indoor antenna, all the outdoor antennas found within a 1km radius will be considered, and a slight variation to the RCA metric, seen in Section 3.5.5, is proposed as:

Figure 4.14. Distribution among the identified clusters, for 22,000 outdoor antennas located in close proximity of the ICN antennas considered in this study.

$$RCA_{out_{i,j}} = \frac{T_{out_{i,j}}/T_{out_i}}{T_{in_j}/T_{tot,in}}, \tag{4.3}$$

where $T_{out_{i,j}}$ represents the traffic recorded for the j-th service at the i-th neighboring outdoor BS, T_{out_i} is the total traffic generated at the i-th outdoor BS for all the mobile services, while the ratio $T_{in_j}/T_{tot,in}$ expresses the level of utilization the of j-th service over the entire indoor traffic. Then, the $RSCA_{out_{i,j}}$ for the outdoor antennas can be computed following Equation 3.4.

It should be noticed that according to Equation 4.3, the RCA for the outdoor antennas measures the level of utilization for a certain service in an outdoor antenna compared to the level of the same service usage among all the indoor antennas. Indeed, rather than studying the traffic observed at outdoor antennas per se, this will explore whether this traffic is innately different compared to that generated in indoor environments.

With the $RSCA_{out_{i,j}}$ for all the neighboring outdoor antennas estimated, their cluster can be inferred by feeding these values through a trained random forest classifier to surrogate and generalize the unsupervised learning clustering results. The predicted cluster distribution for approximately $20,000$ neighboring outdoor antennas is presented in Figure 4.14. Evidently, the innate diversity encountered in the traffic demands of indoor antennas is absent for outdoor antennas. Indeed, almost 70% of the outdoor antennas appertains in cluster 1. That further corroborates the intuition that cluster 1 constitutes a general-use cluster. More importantly, though, it becomes clear that the distinct traffic behavior of workplaces, stadiums, metro, and train stations is now almost absent, as only a negligible percentage of the outdoor antennas lies within the respective clusters.

Key Insights. *The service demand observed in ICNs is absent from the neighboring outdoor BSs, despite their close proximity. This behavior shows that ICN traffic is highly environment-centric, whereas outdoor BSs accommodate general-purpose traffic.*

(a) Orange group, cluster 0. **(b)** Orange group, cluster 4. **(c)** Orange group, cluster 7.

(d) Green group, cluster 5. **(e)** Green group, cluster 6. **(f)** Green group, cluster 8.

(g) Red group, cluster 1. **(h)** Red group, cluster 2. **(i)** Red group, cluster 3.

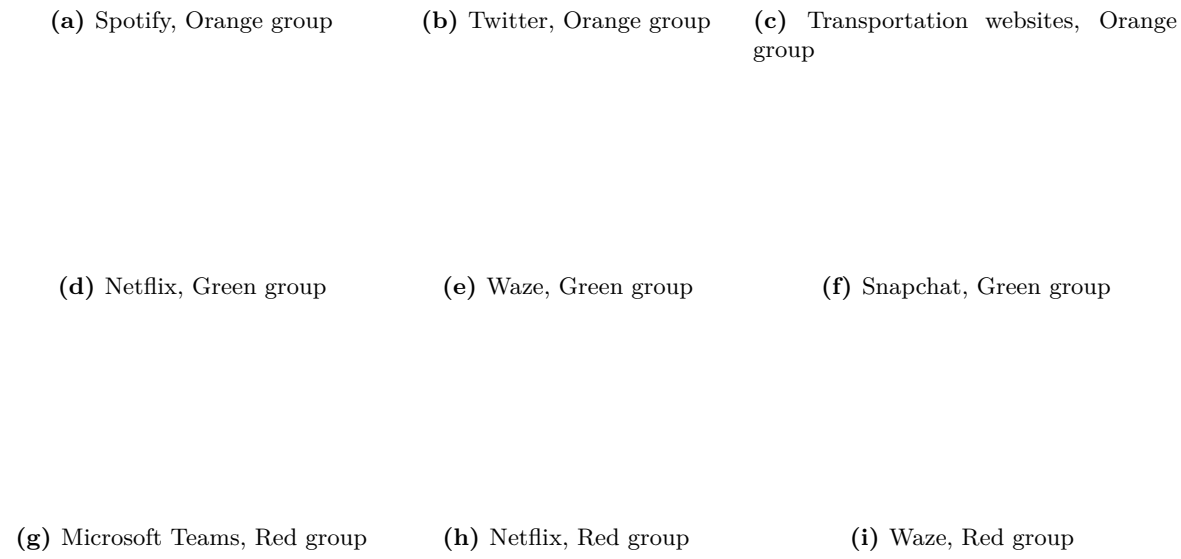
Figure 4.15. Normalized median traffic heatmaps per hour for a period between 04/01/2023 and 24/01/2023. Each heatmap represents the median traffic of all antennas that belong to the specified cluster at a specific hour and day, while the light gray dashed lines indicate weekends.

4.2.4. Temporal patterns of indoor mobile network usage

The next step in the analysis is to uncover the patterns that the ICN traffic exhibits over time. As expected, one should observe different patterns for each cluster, as it has been highlighted in previous research which focused though solely on the traffic recorded at outdoor antennas [23], [27].

4.2.4.1. Temporal patterns of ICN clusters

To this end, Figure 4.15 provides heatmaps showing the evolution of the normalized median traffic per hour across all the antennas belonging to the same cluster, between 04/01/2023 and 24/01/2023. It can be pointed out that there is a strong correlation between the temporal patterns and the indoor environment type. Indeed, as can be seen in Figure 4.15a to 4.15c, the orange group clusters, which are populated with metro and train stations, demonstrate a traffic peak during the common weekly commuting hours in France, i.e., 7.30 to 9.30 a.m. and 17.30 to 19.30 p.m. In the remaining hours of the day the traffic volume is considerably smaller, while the same holds throughout the weekends, e.g., the 7^{th} and 8^{th}, or the 14^{th} and 15^{th} of January. Remarkably, there is another day with negligible traffic in the period under consideration; the 19^{th} of January, which corresponds to a national general strike day. The strike's impact is not as severe for cluster 7, which includes metro antennas found in cities other than Paris, presumably



(a) Spotify, Orange group **(b)** Twitter, Orange group **(c)** Transportation websites, Orange group

(d) Netflix, Green group **(e)** Waze, Green group **(f)** Snapchat, Green group

(g) Microsoft Teams, Red group **(h)** Netflix, Red group **(i)** Waze, Red group

Figure 4.16. Heatmaps of per hour normalized median traffic between 04/01/2023 and 24/01/2023, for the antennas of each cluster and for a selected set of services based on their RSCA value. The light gray dashed lines indicate weekends.

due to the milder impact of the strike in these cities.

The green group clusters temporal patterns, presented in Figures 4.15d to 4.15f, are characterized by sporadic, non-canonical bursts of data usage. Again, this is an expected behavior for stadiums and convention centers, where unlike the other indoor environments discussed, surges of mobile subscribers appear on the premises and generate traffic only when events are taking place. For instance, in cluster 8 a traffic outbreak was observed only on the evening of January 19^{th} in a cross-Atlantic NBA special event conducted at Accor Arena in Paris, while the continuous burst in cluster 5 between the 19^{th} and the 24^{th} of January is a corollary of the 4-day Sirha Lyon event that took place at the Eurexpo Lyon convention center. Another difference between the green group clusters is that cluster 5 records a low volume of traffic throughout the entire day. Indeed, cluster 5 does not include only stadiums but other environments as well.

As depicted in Figures 4.15g to 4.15i, the red group clusters present a more vivid and diurnal pattern compared to the other two groups, with the traffic being almost evenly distributed from 10 a.m. to 8 p.m. Interestingly, clusters 1 and 2, which comprise numerous commercial antennas, record traffic throughout the weekdays and weekends, while cluster 3 consisting primarily of workspaces remains idle during weekends and after working hours, i.e., 5.30 pm. That is a unique behavior that clearly differentiates cluster

3 from the other two clusters. Delving further into the peculiarities of each cluster, cluster 2 yields a slight drop on Sundays, which can be attributed to the fact that it contains smaller stores, e.g., the MNO agencies, which are not as active on those dates compared to larger commercial stores and malls. A further difference between clusters 1 and 2, is that the latter presents higher traffic during nighttime, likely due to the larger number of hotels and hospitals which are more active during these hours.

4.2.4.2. Service-level temporal demands in indoor mobile networks

Finally, further insight can be gained by inspecting the temporal heatmaps for some mobile services that assumed a key role in the cluster decision-making process. Hence, for the orange group clusters, it selects Spotify, Twitter, and transport websites, depicting their normalized median traffic as heatmaps in Figures 4.16a-4.16c. As it can be seen, Spotify assumes a strong presence for the entire group, with alike motifs for all clusters, demonstrating its traffic peaks during the morning commuting hours. On the other hand, the usage of transport websites is scattered for cluster 7, while clusters 0 and 4 preserve a more lively commuting hour pattern. Moreover, Twitter usage is comparatively mitigated for cluster 4, since both clusters 0 and 7 present a persistent peak during morning or evening commuting hours.

The traffic patterns of the green group clusters are what is expected from event-oriented venues. For instance, Snapchat in Figure 4.16f presents traffic patterns significantly similar to the ones seen for total traffic in Figure 4.15, indicating the substantial usage of social media apps throughout these events. Likewise, in Figure 4.16e, Waze which is used for driving navigation and obtaining live traffic maps and road alerts, also shows an interesting traffic pattern. In particular, for these antennas, Waze assumes its peak values a couple of hours after the peaks of total traffic, as well as those of social media apps, which suggest the usage of vehicular navigation apps to guide the event attendants back to their home destination. In addition, video streaming apps, such as Netflix, fall into under-utilization in such venues, even on specific peak days and hours, as can be seen in Figure 4.16d.

For the red group, three very characteristic applications are selected to highlight their different utilization during day hours; Microsoft Teams, Netflix, and Waze. Microsoft Teams, which is a work-oriented service, attains very small peak values for clusters 1 and 2 which are composed mostly of non-working environments. On the contrary, cluster 3 records heavy traffic over working hours, or even during the lunch break, since it is populated with workspaces. Meanwhile, Netflix exhibits the opposite pattern, yielding a stronger usage in daytime and nighttime for clusters 1 and 2, respectively, but only being utilized during lunch hours for class 3. Again, cluster 2 includes most of the hotels where guests use streaming services during nighttime, while in working environments the use of streaming services is strictly limited during breaks. Finally, out of the red group clusters,

Waze attains the highest importance and recorded traffic for cluster 1 which possesses the biggest number of antennas located in tunnels. Furthermore, the Waze traffic peaks in Cluster 1 occur mostly on Saturdays, whereas for Cluster 3 the heaviest traffic is recorded on weekdays after office hours when the employees return to their residences.

Key insights. *The clusters designated by this analysis demonstrate diverse characteristics in the overall and per-application utilization patterns over time, which can be ascribed to the type of indoor environments associated with each cluster. This showcases interesting temporal patterns within the utilization of those indoor spaces, as characterized by the preferred times of peak consumption.*

4.2.5. Main takeaways

This analysis unveiled that ICN mobile service demands are site-specific and more specialized than outdoor ones; therefore, ICN resource orchestration should not target overall capacity, as in outdoor environments, but must take into account the most important application usage per indoor environment. In a sense, it fosters adopting a distinct network slicing dimension for indoor network resource planning, where the indoor slices will be tuned based on the characterizing applications for that specific indoor environment. Applications of such slicing could include adaptive power transmission control or content caching according to the insights provided by this analysis.

The proliferation of ICNs and the increasing mobile service demands generated by users in indoor environments call for an improved understanding of indoor traffic characteristics. The study presented in this section constitutes a primer in this direction, unveiling the unique behaviors that are intrinsic to the traffic generated by ICNs. Overall, the major contributed takeaways are:

- Hinging on a countrywide ICN Internet measurement traffic data set, an appropriate transformation of the traffic data is defined enabling probing the range of different Internet mobile service utilization profiles at indoor antennas. Then, employing an unsupervised learning approach, it's seen that distinct service utilization clusters are inherent in indoor communication systems. This rich and diverse behavior of Internet services has not been unveiled before, and as demonstrated, does not align with that of outdoor legacy communication systems.

- It's exposed that there is a strong connection between the clusters individuated by this analysis and the indoor environment type. In particular, the same mobile applications manifest very heterogeneous behaviors between ICNs and outdoor BSs, even for antennas in proximity, due to the determining influence of the environment type on indoor user activities. This phenomenon has not been highlighted or quantitatively analyzed before.

- It's revealed that the total Internet traffic data as well as the traffic generated by the individual applications exhibit different activity peaks and temporal patterns in the various identified clusters. That paves the way for the proactive management of ICN traffic by mobile traffic operators (MNOs).

5 Relationships between urban space and smartphone usage

The features of a city and how it spans the geographical space have great effects on mobile traffic consumption: In the same way, people may act differently when they are at home, at their office, at a shopping mall, having lunch in a park, or waiting at a metro station, it can be expected that the way they interact with their smartphones will also be greatly affected by those locations throughout the city. These differences in mobile traffic consumption over space can be leveraged as a proxy to also help differentiate and characterize the same locations, providing researchers from multiple fields with an interesting view of the spatial dynamics of current cities.

As expected, the possibilities given by mobile data drive the need for tools that are able to extract the patterns and profiles of utilization of mobile phones and be able to link them directly with different spaces, as well as techniques that are able to isolate specific sites within cities, so they can also be studied directly by the antennas that are providing coverage for their area, as well as any and all spatial limitations that data collected over BSs may have in regards to their precision.

This chapter will explore techniques capable of extracting information about smartphone usage over space, which can be later utilized for the characterization of said spaces based on both the temporal and spatial profiles of consumption. It will be structured as follows: Section 5.1 will present a technique capable of decomposing both temporal and spatial patterns of traffic consumption across two major cities in France, and how it can be utilized to present both long-term and short-term profiles of land utilization. Section 5.2 will significantly reduce the spatial granularity and focus its efforts on techniques that can localize mobile traffic consumption within green spaces within the city of Paris, and how the insights gained about the temporal and mobile application usage patterns are related to the location and features of said green spaces.

5.1. Spatiotemporal analysis of urban mobile data traffic

Mobile data traffic has been steadily surging over the past two decades, and forecasts from major players in the telecommunication ecosystem anticipate that this trend has not been exhausted yet. As a representative example, Ericsson indicates that global mobile network data traffic is rapidly approaching 100 Exabytes per month, with a year-on-year growth steady in the 40-50% range [188]. The phenomenon has spurred increased interest in the precise dynamics of mobile data traffic consumption. Indeed, even at relatively small scales, such as within single urban areas, the demand for mobile data is not homogeneous; rather, it undergoes substantial fluctuations in time and space due to the varying mobility and digital activities of the users. Characterizing these spatiotemporal changes and understanding their root causes has important practical applications, as it helps to establish new links between human endeavors, city fabrics, and the utilization of mobile services, and can support more informed network infrastructure management.

Early efforts in the analysis of mobile data traffic have revealed important features of its dynamics. For instance, the traffic generated by mobile subscribers is strongly periodic [39] and geographically localized [189], which enables its effective prediction. Also, social events can determine significant variations [190], with the consequent need for dedicated resource management policies in mobile networks. Similarly, the bandwidth consumed by individual subscribers is highly heterogeneous [167], yet it is captured by a limited number of typical profiles [40], [41], [191], [192], which enables, e.g., the informed tuning of traffic plans. Moreover, specific mobile services can generate traffic patterns that are highly heterogeneous in time [34], while differences over space depend on the considered geographical scale [25], [38], [193], hence calling for tailored resource management strategies in network slicing environments.

Analyses of mobile traffic demand in the literature can be divided into two broad categories [11], [12]. On the one hand, there are works that take a user perspective, and study the behavior of individual subscribers in terms of their mobility, the traffic they generate, and the mobile services they consume. On the other hand, there are studies that take the viewpoint of a mobile network operator and investigate properties of the demand aggregated over all users present in a given area, typically a cell sector or the coverage region of a base station. In this second category, the study of mobile phone traffic data has often supported transportation studies as a larger-scale and cost-efficient alternative to travel surveys for, e.g., the understanding and modeling of travel demand at a regional scale and its spatial classification [194], the identification of the most-traveled routes on a road network [128], the analysis of travel demand between transportation hubs in urban areas [195], and the reconstruction of travel mode for inter-city trips [196].

The work presented falls in the second category above. Despite previous efforts, dependable tools are still lacking to explore complex relationships in mobile data traffic.

In particular, while partial solutions have been proposed to detect either temporal or spatial structures in traffic demands, little attention has been paid to the more challenging *concurrent* inspection of both space and time dimensions. Development of a tool capable, at the same time, of automatically segmenting an urban territory into homogeneous areas and providing a temporal description of each identified area, could provide valuable information for travel demand estimation, as well as macroscopic traffic modeling.

In this Section, an original methodology is presented for the spatiotemporal classification of the mobile data traffic observed in operational networks, so as to fill the gap above. The proposed solution builds upon *Explanatory Factor Analysis (EFA)*, a well-established data analysis instrument in psychology research. EFA aims at identifying, in a fully automated way, latent factors that cause the dynamics observed in the data. This is achieved by identifying the variables of interest in the data and describing their covariance relationships in terms of the underlying and unobservable factors. EFA will be tailored to the specific problem of identifying recurrent behaviors in the mobile data traffic. The approach yields significant advantages over previous proposals:

- *Versatility in spatiotemporal analyses.* The methodology represents a unified approach to recognizing factors that are temporal or spatial in nature. Along the time dimension, EFA can detect temporal structures in the network-wide communication activity, revealing time periods that show a similar, stable spatial distribution of the mobile traffic demand. On the spatial dimension, EFA identifies hidden spatial structures, by automatically decomposing a target geographical area into zones where mobile data traffic follows homogeneous time dynamics.

- *Unsupervised and probabilistic nature of the approach.* EFA is a completely unsupervised tool that produces probabilistic structures. This allows for overcoming the limitations of methods previously employed for mobile traffic analysis, such as clustering, which only produces deterministic temporal or spatial categories, or supervised techniques that require labeled data.

- *Interpretability of results.* The proposed methodology eases the exploration of the root causes for the temporal and spatial structures above, by allowing an automated extrapolation of the structures hidden in the respective dual dimensions. In other words, EFA implicitly provides knowledge of the traffic geography that characterizes each temporal structure, and of the precise traffic time series that distinguish each spatial structure.

These advantages will be demonstrated throughout this Section with real-world mobile data traffic collected in production networks serving two major cities in France. The results highlight the vast range of unique temporal and spatial profiles that can be obtained from mobile traffic data through the use of EFA, as well as the ability to identify

short and long-term spatial structures and mixed land usage in cities. Ultimately, this work opens interesting perspectives on the use of EFA as a dependable tool for the analysis of complex hidden structures in mobile data traffic.

Relying on EFA allows for overcoming multiple limitations of previous methods. First, while the previously mentioned works explored the temporal and spatial structures of mobile traffic separately, EFA provides a unified framework that can be cast to explore both dimensions. Second, it surpasses the strictness of clustering, and allows for a probabilistic association of archetypal mobile traffic patterns to time periods or geographical areas, allowing to appreciate dynamics that are overlooked by traditional methods. Third, it outputs implicit information on the mobile traffic behaviors in space (respectively, time) that tell apart and characterizes diverse periods (respectively, areas), thus easing the explanation of results.

5.1.1. An introduction to Exploratory Factor Analysis

The approach proposed in this Section builds upon EFA techniques that root into the work of Spearman, over a century ago [197]. Since those early studies, EFA has emerged as one of the dominant classes of factor analysis, and has been widely employed in statistical psychology research [198]. Earlier works have introduced factor analysis for the study of mobile traffic data and, more generally, in the field of networking [27]. Only recently, a similar approach has been used to detect network anomalies problems in a campus Wi-Fi network [199], which is a different task than the proposed target, i.e., the inference of spatiotemporal structures in mobile traffic.

Relying on EFA allows for overcoming multiple limitations of previous methods. First, while all previous works explored the temporal and spatial structures of mobile traffic separately, EFA provides a unified framework that can be cast to explore both dimensions. Second, it surpasses the strictness of clustering, and allows for a probabilistic association of archetypal mobile traffic patterns to time periods or geographical areas, allowing to appreciate composite dynamics that are overlooked by traditional methods. Third, it outputs implicit information on the mobile traffic behaviors in space (respectively, time) that tell apart and characterizes diverse periods (respectively, areas), thus easing the explanation of results.

It is also worth noticing that previous works have analyzed CDR that encompass voice calls and text messages but do not capture the activity of mobile subscribers in terms of data traffic. While interesting from a sociological perspective, calling and texting play an increasingly diminishing role in network traffic, hence the results of many studies in the literature hardly apply to modern networks. Instead, this analysis fully focuses on data traffic collected in an operational 3G/4G network, and thus provides an up-to-date view on hidden structures in today's mobile traffic that can point towards a dynamic description of human presence and, therefore, travel demand. In order to lay the foundations of EFA,

Figure 5.1. EFA toy example: student grading across subjects. In this case, EFA can be used to identify a limited set of latent abilities of the students that may explain their grades.

Table 5.1 introduces the terminology used in the remainder of the section, using the toy example in Figure 5.1 to illustrate the semantics.

5.1.1.1. Fundamental model

Given a set of observed variables of interest, factor analysis is formally defined as *"a model of hypothetical component variables that explain the linear relationships existing between observed variables"* [200]. Such a hypothetical set of component variables can be derived mathematically from the observed variables, as follows.

Let $\mathbf{X}$ be a $N \times 1$ vector of observed *variables*, distributed with expectation $\mathbb{E}(\mathbf{X}) = 0$ and covariance $\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{X})$. Let also $\mathbf{F}$ be a $K \times 1$ vector of unknown normalized *common factors*, having mean $\mathbb{E}(\mathbf{F}) = 0$, covariance $\mathbf{\Phi} = \mathrm{Cov}(\mathbf{F})$ and order $K < N$. Next, let $\mathbf{\Lambda}$ be an unknown $N \times K$ matrix of common factor pattern coefficients (i.e., *factor loadings*). Let also $\mathbf{U}$ be a $N \times 1$ vector of independently distributed error terms (i.e., *unique factors*), with mean $\mathbb{E}(\mathbf{U}) = 0$ and finite covariance $\mathbf{\Psi} = \mathrm{Cov}(\mathbf{U})$. Since each unique factor is specific to one variable, the error terms are independent, and $\mathbf{\Psi}$ is a diagonal matrix. Finally, it's desirable to have common factors and unique factors to be uncorrelated, i.e., $\mathrm{Cov}(\mathbf{F}, \mathbf{U}) = 0$. It follows that:

$$\mathbf{X} = \mathbf{\Lambda F} + \mathbf{U} \tag{5.1}$$

is the *fundamental equation of factor analysis*, stating that the observed variables in $\mathbf{X}$ are weighted combinations of the common factors in $\mathbf{F}$ and the unique factors in $\mathbf{U}$. From (5.1), the covariance of the observed variables $\mathbf{X}$ can be written as:

$$\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{X}) = \mathrm{Cov}(\mathbf{\Lambda F} + \mathbf{U}) = \mathbf{\Lambda}\mathrm{Cov}(\mathbf{F})\mathbf{\Lambda}^{\intercal} + \mathrm{Cov}(\mathbf{U}) = \mathbf{\Lambda\Phi\Lambda}^{\intercal} + \mathbf{\Psi}, \tag{5.2}$$

which represents the *fundamental theorem of factor analysis*. In the case of EFA, no hypotheses concerning the factors[1] are made, and it is thus generally assumed that all

[1] Next, *common factor* and *factor* will be used interchangeably.

Term	*Description*
Variables	The phenomena of interest, related to a population, e.g., subjects taught to primary school students in Figure 5.1.
Samples	The set of individuals for which all phenomena of interest can be measured, e.g., students from the same class.
Observations	The realizations of all variables for each sample, e.g., the grades of examination tests in all subjects obtained by each student.
Common factors	Complex interrelationships among the observed phenomena that can be reasonably assumed to exist, e.g., inferring factors such as verbal and mathematical intelligence of the students, which may explain the aptitude to each subject.
Factor loadings	Numerical relationships that describe to what extent each common factor explains each variable, e.g., a factor with high loadings solely in algebra and geometry can reveal the existence of a common mathematical intelligence.
Factor scores	Estimated values that relate samples to common factors, e.g., scores indicate if the good/poor performance in scientific disciplines of students is well explained by their strong/weak mathematical intelligence.
Unique factors	Help explain the unique variance associated to each variable, pinpointing outlying behaviors in the data, e.g., unique factors could account for a rare talent of one student towards a specific discipline.

Table 5.1. EFA terminology and examples

factors are orthogonal, i.e., mutually uncorrelated and with unit variances. Therefore, $\mathbf{\Phi}$ can be replaced by the identity matrix in (5.2), and:

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^{\intercal} + \mathbf{\Psi}, \tag{5.3}$$

whose i-th diagonal element can be written as:

$$\sigma_{ii} = \mathrm{Var}(x_i) = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_{ii} = h_i + \psi_{ii}. \tag{5.4}$$

From (5.4), the variance of each observed variable, σ_{ii}, consists of two parts: the *communality* h_i, i.e., the portion of the variance shared with the other variables via the common factors, and the *unique variance* ψ_{ii}, i.e., the share specific to each variable, via the associated unique factor.

5.1.1.2. Factor extraction

Several methods have been developed to estimate the unknown variables $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ in (5.3). Two popular approaches utilized in this analysis will be presented next.

Maximum Likelihood Estimation (MLE) allows inferring the unknown variables $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ in (5.3) in a way that is efficient and robust [200]. MLE assumes $\mathbf{X}$ in (5.1) to have a multivariate normal distribution[2] with mean $\bar{\mathbf{X}} = \frac{1}{M}\sum_{a=1}^{M}\mathbf{X}_a$ and covariance $\mathbf{S} = \frac{1}{M-1}(\sum_{a=1}^{M}\mathbf{X}_a\mathbf{X}_a^\intercal - M\bar{\mathbf{X}}\bar{\mathbf{X}}^\intercal)$ computed from the M observations. The information provided by $\mathbf{S}$ may also be represented by a correlation matrix $\mathbf{R}$ and a set of standard deviations $s_1, s_2, \ldots, s_N$.

MLE maximizes the likelihood function:

$$\ln L = -\frac{1}{2}(M-1)[\ln|\mathbf{\Sigma}| + \mathrm{Tr}(\mathbf{S}\mathbf{\Sigma}^{-1})], \tag{5.5}$$

with Tr indicating the matrix trace operator. The $\mathbf{\Sigma}$ matrix maximizing (5.5) also minimizes the following fit function [201]:

$$F_K(\mathbf{\Sigma}) = \ln|\mathbf{\Sigma}| + \mathrm{Tr}(\mathbf{S}\mathbf{\Sigma}^{-1}) - \ln|\mathbf{S}| - N, \tag{5.6}$$

where K refers to the number of common factors considered. Using (5.3) in (5.6), the expression $F_K(\mathbf{\Sigma}) = F_K(\mathbf{\Lambda}, \mathbf{\Psi})$ can be used to compute the maximum likelihood estimates of the unknowns $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. The main steps are outlined below, while full details are found in [200].

Firstly, F_K is minimized with respect to $\mathbf{\Lambda}$, where the minimizer $\tilde{\mathbf{\Lambda}}$ is computed by imposing $\frac{\partial F_K}{\partial \mathbf{\Lambda}} = 0$. Denoting as $\Im$ the identity matrix, the above condition leads to

$$\tilde{\mathbf{\Lambda}} = \mathbf{\Psi}^{1/2}\mathbf{\Omega}_K[\gamma_i - 1]_K^{1/2}, \tag{5.7}$$

where the diagonal matrix $[\gamma_i - 1]_K$ contains the K largest eigenvalues of $\mathbf{\Psi}^{-1/2}\mathbf{S}\mathbf{\Psi}^{-1/2}$, and $\mathbf{\Omega}_K$ contains the corresponding eigenvectors. Replacing (5.6) in (5.7) one can derive the expression of the conditional minimum for a given $\mathbf{\Psi}$, as:

$$f_K(\mathbf{\Psi}) = -\sum_{j=K+1}^{N}\ln\gamma_j + \sum_{j=K+1}^{N}\gamma_j - (N-K), \tag{5.8}$$

where γ_j, with $j = K+1, \ldots, N$, are the residual eigenvalues of the matrix $\mathbf{\Psi}^{-1/2}\mathbf{S}\mathbf{\Psi}^{-1/2}$.

Secondly, the function f_K is minimized with respect to $\mathbf{\Psi}$, by imposing $\frac{\partial f_K}{\partial \mathbf{\Psi}} = 0$, which leads to the expression

[2] MLE is known to yield good estimations even when the actual distribution of $\mathbf{X}$ is not multivariate Gaussian [200]. This is proven in the context of mobile data traffic analysis by the minor discrepancy of the results attained with MLE and with MINRES, which does not require the Gaussian distribution assumption.

$$\mathsf{Diag}(\boldsymbol{\Psi}^{-1}(\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Lambda}}^{\intercal} + \boldsymbol{\Psi} - \mathbf{S})\boldsymbol{\Psi}^{-1}) = 0. \tag{5.9}$$

At this point, the maximum likelihood estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ can be computed by means of an iterative procedure based on the Fletcher-Powell method and applied to the function f_K and its partial derivatives in (5.8) and (5.9), respectively.

Minimum Residuals (MINRES) is an alternative approach to factor extraction. Unlike maximum likelihood estimation, MINRES does not rely on any assumption about the distribution of observed variables and can produce valid solutions even when applied to singular matrices [202]. The working principle of MINRES is to minimize the sum of off-diagonal squared residuals of the correlation matrices, i.e., differences between the observed ($\mathbf{R}$) and reproduced ($\boldsymbol{\Lambda}$) correlations, without requiring any estimation of the communalities, i.e., h_i in (5.4). In other terms, MINRES minimizes the fit function:

$$F_K(\boldsymbol{\Lambda}) = \|[\mathbf{R} - \Im] - [\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\intercal} - \mathsf{Diag}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\intercal})]\|\,, \tag{5.10}$$

where K refers to the number of common factors considered for the estimation of the $\boldsymbol{\Lambda}$ factor loadings matrix, and $\mathbf{H} = \mathsf{Diag}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\intercal})$ is the diagonal matrix of reproduced communalities. Equation (5.10) can be written in explicit form:

$$F_K(\boldsymbol{\Lambda}) = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(r_{ij} - \sum_{k=1}^{K}\lambda_{ik}\cdot\lambda_{jk})^2, \tag{5.11}$$

which is a function of the $N(N-1)/2$ off-diagonal residual correlations. MINRES minimizes $F_K(\boldsymbol{\Lambda})$ by successive approximations of the values of the factor loadings λ_{ik}. The communalities $h_i = \sum_{j=1}^{k}\lambda_{ij}^2$ are then obtained as a by-product of such method.

The minimization approach traditionally adopted with MINRES is based on a Gauss-Seidel procedure, originally proposed in [203]. The procedure starts with the choice of the number of factors K and the selection of an arbitrary loading matrix, for the given N variables and selected K factors. At each iteration, the arbitrary factor loading matrix is modified on a per-row (i.e., per-variable) basis, by considering an increment x_{ik} of the loading value of the considered variable i on each factor k, and by selecting the displacements vector X_i that minimizes the objective function. This optimal displacements can be iteratively determined by zeroing the partial derivatives of Equation (5.11) with respect to the considered variable i, considering the loading matrix modified at previous iteration, i.e.,:

$$\mathbf{X}_i = \mathbf{R}_i^0\boldsymbol{\Lambda}(\boldsymbol{\Lambda}_{)i(}^{\intercal}(\boldsymbol{\Lambda}_{)i(})^{-1}. \tag{5.12}$$

Here, $\mathbf{X}_i$ is the row vector of incremental changes of the factor loadings for variable i; $\boldsymbol{\Lambda}_{)i(}$ is the factor matrix obtained by replacing with zeros the elements in row i of the

current factor matrix; and, $\mathbf{R}^0$ is the vector of (observed) residual correlations of variable i with all other variables, with zeros for self-residuals.

Successive adjustments of the factor loadings matrix are performed multiple times on all variables, until a numerical convergence criterion, related to the rate of change of F_K, is satisfied, and the final loadings matrix is obtained. Specific details on the factor estimation procedure and suggested convergence criteria and parameters can be found in [203], while later optimizations are discussed in [202], [204], [205].

5.1.2. EFA for mobile traffic analysis

Next step will be to introduce the mobile data traffic measurement dataset employed throughout this study, and discuss how EFA can be tailored to the problems of inferring temporal and spatial structures in such type of data. By doing so, it's shown how the EFA framework can be cast to solve the two tasks, which are in fact each other's dual.

5.1.2.1. Measurement set

The mobile data traffic analyzed here was collected in the production infrastructure of Orange and was gathered during 12 consecutive weeks of 2016 (05/09 - 28/11), in two of the largest urban areas of France, i.e., Paris and Lyon, and covers the whole user base of the operator. The measurement data consists of total (i.e., aggregated in the uplink and downlink directions) volume of data traffic generated by the whole subscriber base of the operator in the considered regions, consisting in several millions of unique (resident or temporary) users. The traffic data is geo-localized at the level of the radio access antenna serving each user and is timestamped using a temporal granularity of 30 minutes.

5.1.2.2. Casting EFA for temporal and spatial analyses

As previously anticipated, EFA is a versatile tool that can be cast to identify both temporal and spatial structures hidden in the dynamics of mobile data traffic. The input to either problem is an aggregate representation of the communication activity of the mobile subscribers in the geographical region of interest. This definition of input is general and can accommodate any level of spatial and temporal aggregation, as well as any notion of mobile user activity (e.g., voice calls, text messages, generic data usage, or consumption of specific mobile services). In the specific case of this study, the input format is aligned to the measurement data: the activity maps to the volume of mobile data traffic, which is provided at a spatiotemporal resolution of the antenna location during every hour. It's important to remark that the spatial mapping of traffic is then performed by calculating the Voronoi tesselation (as discussed in Section 3.6.

In order to explain how to cast EFA to solve the problems of temporal and spatial analysis of mobile traffic, consider the example in Figure 5.2. Here, the traffic demand

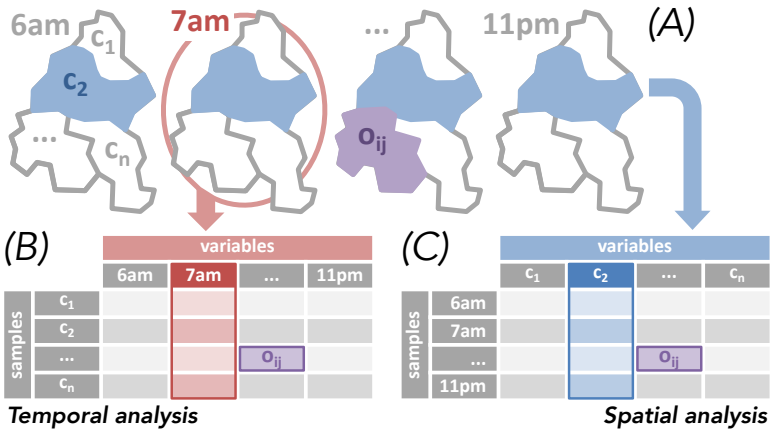


Figure 5.2. Mobile data traffic analysis with EFA in a toy scenario. (A) The one-week demand in the target region is aggregated on an hourly basis with respect to a spatial tessellation of n cells, each representing the coverage of one antenna. The resulting demand in the i-th cell during the j-th time slot is the EFA observation o_{ij}. (B) Temporal analysis: the hourly time slots are the EFA variables, each characterized by a set of observations over the cell samples. (C) Spatial analysis: the geographical cells are the EFA variables, each characterized by a set of observations over the hourly samples. Figure best viewed in colors.

in a given geographical region is aggregated on a hourly basis in n spatial cells, each corresponding to the coverage area of one antenna. One EFA observation o_{ij} matches the mobile data traffic volume recorded at the antenna cell i during the hourly time slot j. Then, the two problems are set apart by the mapping of variables $\mathbf{X}$ in (5.1), as follows.

Temporal analysis. *Time slots* will be modeled as the EFA variables. Each variable is described by the mobile traffic demand (i.e., the EFA observations o_{ij}) recorded over all spatial cells $c_1 \dots c_n$ during a given hourly time interval (e.g., 7:00 to 7:59 AM), as shown in plot (B) of Figure 5.2. In this EFA configuration, each variable (i.e., hour) is represented by a snapshot of the spatial distribution of traffic across cells. The common factors sought by EFA are then temporal structures that explain at what time instants the geographical distribution of the mobile demand is comparable.

An important remark is that, here, spatial cells map to EFA samples: hence, EFA scores relate cells to temporal profiles, revealing which geographical areas are especially important for a given temporal profile. This allows interpreting the temporal analysis results from a spatial dimension.

Spatial analsyis. EFA variables correspond to *geographical areas*. Each variable consists in the mobile traffic demand (i.e., the EFA observations) recorded in a specific cell through the complete monitoring period, as in plot (C) of Figure 5.2. In this EFA configuration, the EFA common factors represent structures in the geographical space that explain in what areas the mobile demand follows similar temporal dynamics.

Interestingly, time slots become the EFA samples in this configuration. Therefore, the EFA scores now explain what time periods are especially distinctive of the mobile

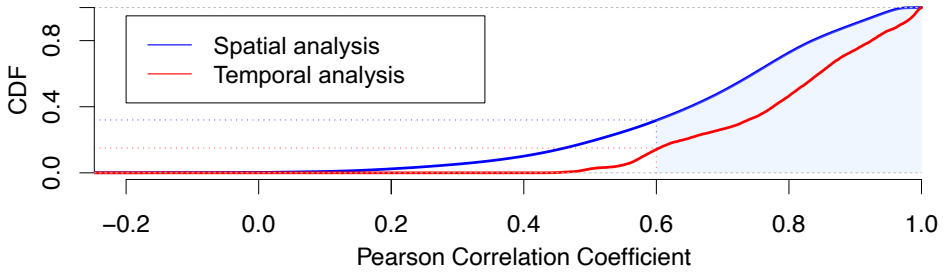


Figure 5.3. Distributions of the Pearson correlation coefficient computed between all pairs of EFA variables in the temporal and spatial mobile traffic analysis problems.

traffic usage in different areas, offering again a unique spatiotemporal interpretation of the results.

5.1.2.3. Tuning EFA for mobile data traffic analysis

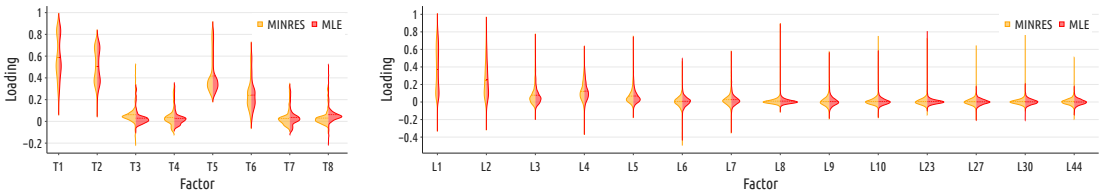
The two configurations of EFA previously outlined highlight how the temporal and spatial analyses of mobile data traffic are in fact dual problems. Beyond such a high-level mapping of variables and samples, several adjustments are needed in order to adapt the baseline EFA scheme to the specific problems at hand, including data verification and method parametrization. These aspects are discussed next.

Suitability of mobile traffic data for EFA. The definition of factor analysis builds on two major hypotheses on the input data: (a) the existence of a non-zero correlation among the observed variables, and (b) the linearity of the functional relationships among the observed variables and the unknown hidden factors. In practical cases, it is important to verify if these assumptions hold for the data to be analyzed. Thus, as a preliminary step in this study, the suitability of mobile traffic demand datasets for EFA is checked.

Tests exist that are dedicated to this purpose. Specifically, the *Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy* [206] is run on the reference datasets, which measures the proportion of the variance in the variables of the data that could be caused by the same common factors across all variables, and not just common correlations across pairs of variables. The test returns values in the range $[0, 1]$, where results close to 1 indicate a high suitability of the data to EFA. In both the classification problem formulations, and for all datasets, KMO returns values around 0.99.

As additional checks, it's verified: (i) the linearity of all pairwise relationships between EFA variables in the two mobile demand classification problems, finding strong correlation in 70–80% of cases, as shown in Figure 5.3; (ii) the sample-to-variable ratio, finding that it is always much larger than one, which is typically considered as a good rule of thumb for a meaningful factor analysis. In the light of all these results, mobile traffic data appears as an excellent candidate for EFA.

Choice of the number of common factors. An important design choice concerns



(a) Temporal analysis **(b)** Spatial analysis

Figure 5.4. Violin plots of the loading values on (a) temporal and (b) spatial factors, when solving the EFA problem with MINRES and MLE. For the spatial analysis, the distributions are shown for a selection of the factors returned by EFA.

the number of common factors that EFA should target. Many heuristics have been proposed to automatically determine such a number: most of them measure properties of the correlation matrix, search for the value that maximizes or minimizes the observed property, and suggest this value as the number of factors to retain [207].

Parallel Analysis (PA) [208] will be relied, which is recognized as one of the best methods for deciding how many factors to extract [209]. The PA method uses the eigenvalues of the data correlation matrix as rough estimates of the actual common factors. Specifically, PA compares such eigenvalues against those of uncorrelated normal variables that mimic the data variables (i.e., come in the same quantity, with identical sample size). The presence of common factors shall induce large eigenvalues: the number of factors is set to the lowest rank above which all data eigenvalues are larger than those from the uncorrelated variables.

Factor rotation. The common factors that satisfy (5.1) are subject to *rotational indeterminacy*, i.e., they are not mathematically unique, and linear transformations allow moving across the full space of solutions. A sensible rotation of common factors can maximize high and minimize low loadings. As explained in Section 5.1.1, loadings are the main instrument to link factors and variables: thus, the presence of stronger (i.e., closer to 1 or -1) loadings outlines more neat structures in the data and eases result interpretation.

VARIMAX rotation [210] is used to identify an appropriate rotation of factors. Given the unrotated $N \times K$ loading matrix $\mathbf{\Lambda}$, VARIMAX iteratively finds a $K \times K$ orthonormal transformation matrix $\mathbf{T}$ such that $\mathbf{\Lambda T}$ maximizes:

$$\sum_{j=1}^{K} \frac{N \sum_{i=1}^{N} (a_{ij}^2/h_i^2)^2 - \left(\sum_{i=1}^{N} a_{ij}^2/h_i^2\right)^2}{N^2}, \tag{5.13}$$

where a_{ij} are the elements of $\mathbf{\Lambda T}$ and h_i is the communality of the i-th variable defined as in (5.4) and computed from $\mathbf{\Lambda T}$.

Solving method. As previously detailed in Section 5.1.1, two different methods are considered to extract the common and unique factors via EFA, i.e., MLE and MINRES.

In order to measure the quality of the factors returned by each approach, the loading values are utilized. As already mentioned, higher loadings on an equivalent number of factors indicate a sharper and more explainable decomposition of the data, since variables are linked to hidden factors in a tidier way.

In fact, solving the EFA problems with MLE and MINRES returns the exact same factors, which already proves the robustness of the whole framework. It also simplifies the direct comparison of the loadings on such factors: specifically, Figure 5.4 illustrates the distributions of loadings on the returned factors for all variables, as violin plots. The main takeaway is that the MLE and MINRES yield almost identical distributions of the loadings, as discrepancies are minimal, and only affect a few factors. It's concluded that there is no operational difference in adopting one solving method or the other, and focus on results obtained by the computationally faster MINRES in the following.

5.1.3. Temporal structures in mobile traffic consumption

The results for the temporal analysis will be shown first. The focus will be on a compressed version of the temporal data by adopting a median week representation [17]: for each spatial cell, the median traffic volume observed at each hour of the week is computed (e.g., for all Mondays, 7:00 am to 8:00 am), and used as EFA variables. On the one hand, considering each hourly time slot in the 12-week period as an independent EFA variable would lead to a very large number of variables, higher than the number of cells, i.e., EFA samples: this is a condition that shall be avoided, as it makes factor inference less dependable [211]. On the other hand, hourly time variables recorded at the level of individual antennas are affected by substantial random noise; a median week compression is known to mitigate potential biases due to outlying behaviors, which, in this context, makes more representative traffic structures emerge.

5.1.3.1. Temporal structures across time

By solving the EFA problem in the temporal analysis configuration above, a total of 8 different common factors are obtained from this reference dataset, which will be labeled T1 through T8. An early insight provided by the methodology is therefore that *the weekly dynamics of mobile data traffic can be summarized into a very small number of archetypal profiles.* Indeed, the dynamics observed in the 24×7 time slots can be explained as a linear combination of just 8 patterns. These results are obtained by merging the measurement data for Paris and Lyon, i.e., using all cells in both cities as a single set of samples; this approach is preferred (rather than running for each city) as it allows the identification of more generalist temporal factors, which apply across cities and are not specific to a single urban area. Thus, both the above considerations and the remaining behaviors discussed to be discussed in this section will apply to both urban areas.

Figure 5.5. Temporal factors obtained from EFA, for the median week mobile traffic demand. Each plot refers to one common factor returned by EFA. In every plot, the hourly time slots (EFA variables) are arranged along 24-hour daily cycles (on the abscissa) for 7 days (Monday to Sunday, on the ordinate), and colors illustrate the loading of the time slot on the considered factor. Figure best viewed in colors.

The detailed loadings of all variables, i.e., hourly time slots in a week for each factor, are shown in Figure 5.5. Such loadings are color-coded according to the scale on the right side of the plot, with red representing higher values, and blue mapping to lower values.

The representation in Figure 5.5 allows for a preliminary interpretation of the factors, recapitulated in Table 5.2. Factor **T1** shows higher loading values for time slots from Monday through Friday, starting at 8 am and decreasing after 8 pm; these are easily mapped to the standard working hours. Factor **T2** presents a complementary behavior to T1, as its loadings are lower during work times, and increase after 8 pm until 2 am; moreover, loadings on T2 are also high during the whole weekend, which allows tagging T2 as characterizing relax hours of the week. Profile **T7** yields some similarity to T2, but with high loadings limited to weekends, specifically Saturday afternoon.

Three factors can then be related to behaviors occurring at the start or end of the weekdays. Two factors show high loadings in early hours: factor **T6** showing its peaks from 6 am to 8 am across all days of the week and interestingly a small 1h shift on the weekends (when peaks start appearing around 7 am), which can mean that this is a factor

Factor	*Activity*	*Hours*
T1	Working	Weekdays, 8 am – 8 pm
T2	Relax	Weekday evenings, weekends
T3, T5	Nightlife	Midnight through 8 am
T4	After-work	From 6 pm until 9 pm
T6, T8	Early	From 6 am until 10 am
T7	Weekends	Saturday & Sunday noon

Table 5.2. Temporal factors in mobile data traffic identified by EFA.

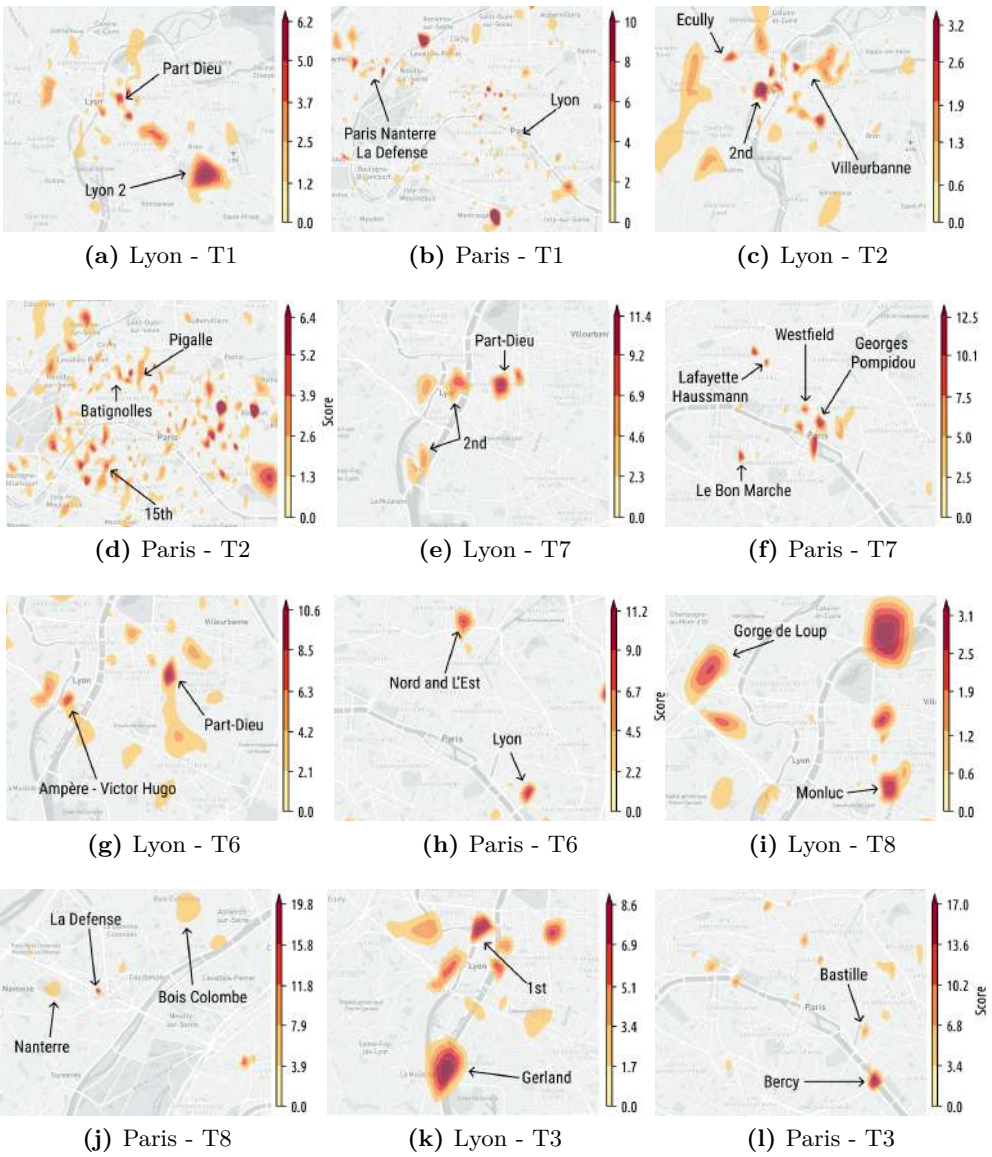
related to general early commuting; factor **T8** peaks are slightly shifted from the previous and go from 8 am until 10 am, but only for weekdays, which leads to the believe that this factor could relate to normal office commuting, as many workers in France start their work hours at the office between those times. Opposing the previous two, factor **T4** has its peaks at after work hours (5 pm to 8 pm) of weed days, which can relate to the general commuting related to leaving work and study spaces. It's interesting to note that there's no specific factor related to after work commuting on weekends, which can relate to the fact that this movement may be more dispersed throughout the weekend, with no clear pattern happening at the studied urban centers.

The two final profiles correspond to late night behaviors: factor **T3** focus in high activities exclusively during the weekends, with an initial growth in activity being seen Friday and peaking on Saturday and Sunday from midnight to 7am. Meanwhile, factor **T5** also has peaks of activity during late hours of the night, but with its peaks spread across all days of the week. Those differences may be due to factor **T3** being related to late night leisure while factor **T5** just as overall late night activity; further exploration of the geographical distribution of those factors during Section 5.1.3.2 can help confirm those hypothesis.

It is worth nothing that, unlike traditional clustering approaches [23], EFA yields a non-deterministic association of time intervals and factors. Indeed, a specific same hour of the week may be affected by multiple concurrent profiles with changes of intensity; this is the case of Saturday afternoons (composed by T2 and T7, plus T1 to a lesser extent) or 8 am during working days (T8 and T6, plus T2 to a lesser extent). This representation highlights the composite nature of mobile traffic patterns observed during the week, represented as a combination of multiple factors.

5.1.3.2. Geographical distribution of temporal structures

Looking at the loadings in Figure 5.5 may not be enough to disambiguate what determines the temporal structures identified by EFA. Only obvious root causes, such



(a) Lyon - T1 **(b)** Paris - T1 **(c)** Lyon - T2

(d) Paris - T2 **(e)** Lyon - T7 **(f)** Paris - T7

(g) Lyon - T6 **(h)** Paris - T6 **(i)** Lyon - T8

(j) Paris - T8 **(k)** Lyon - T3 **(l)** Paris - T3

Figure 5.6. Geographical distribution of temporal factors across Lyon and Paris.

as work activities of factor T1, can be pinpointed with some level of confidence.

As discussed in Section 5.1.1, EFA offers an automated way to better interpret the results via the factor scores. In the case of temporal profiles of mobile traffic, EFA scores allow quantifying the importance of each geographical location sample for every factor. In other words, it's possible to draw maps of the EFA score, where higher values in one area indicate that the local traffic has a higher influence for the targeted profile. Such visualization highlights locations and points of interest, which can explain the social phenomenon underlying each temporal factor. Those results are compiled for selected factors in both cites on Figure 5.6.

Factor T1 has higher score around work areas: universities (Jean Moulin Lyon 3, Lyon

2 and INSEEC in Lyon; Paris Nanterre in Paris), working zones (3th arrondissement of Lyon; downtown Paris and La Defense, one of the biggest business centers of Europe) hospitals and big stations (Part Dieu, Lyon's busiest station; Gare de Lyon and Montparnasse in Paris). This matches the temporal structures seen on Figure 5.5, meaning those regions have their main peaks of mobile traffic consumption during work hours. In contrast, the relaxing hours Profile T2 presents a different side of both cities. Regions with higher scores are either located close to shops and restaurants (such as 2th arrondissement in Lyon and Passy and Batignolles neighborhoods in Paris), which expect most of their traffic activity after work, as well as places with an active night live (Pigalle neighborhood in Paris). Also, this profile emerges in more residential zones of both cities (Villeurbane, Ecully and Tassin in Lyon; 12th and 15th arrondissement in Paris). This opposition between working and residential places between both profiles is further seen when comparing Figures 5.6a and 5.6c for Lyon, as well as Figures 5.6b and 5.6b for Paris: bigger clusters of one factor are mainly in empty places of the other.

Factor T7, seen on Figures 5.6e and 5.6f, presents some of the favored locations during the weekends: department stores. In Lyon, a cluster is seen around 2th arrondissement (which known for its large commercial centers) and around the La Part Dieu Shopping Center. The same can be seen in Paris, where clusters are close to Le Bon Marche Department Store, Westfield Forum des Halles Mall and Galeries Lafayette, as well as the cultural center Georges Pompidou. In France, many stores are closed on Sundays, which could explain an extra influx of mobile traffic around those locations on Saturdays.

The geographical distribution of factors also helps further differentiate the three profiles that appear to be associated to commuting behaviors. Factor T6 shows higher loading from 6am to 8 am, and its scores on Figure 5.6g and 5.6h help relate it to general peak-hour commuting behavior. Clusters are seen around bigger stations, such as Ampère Victor Hugo and Part Dieu in Lyon, as well as Paris' Gare du Nord, de L'Est and Lyon. Those are highly connected stations that attract many for their daily commute. Factor T8 (seen on Figures 5.6i and 5.6j) relates to early day commuting, with regions located mostly in residential and suburban areas, such as Monluc and Gorge de Loup in Lyon, or Bois Colombe in Paris, as well as some business zones like Nanterre and La Defense, both located in the capital. Finally, T4 represents specifically after work commuting, from 5pm to 8pm. This reflects in regions with higher scores closer to the city's downtown. A few exceptions can be made, like the La Defense region, which shows high values for both profiles (as well as T6). This is easily explained due to La Defense being one of the busiest work-only districts of Paris, attracting a strong influx and outflux of people who work in the area on a daily basis.

For the late night profiles, the locations for Factor T5 can be considered a subset of the locations for Factor T3 (seen on Figures 5.6k and 5.6l), as most zones with high scores on the first profile are also present in the latter. This includes regions in Paris that

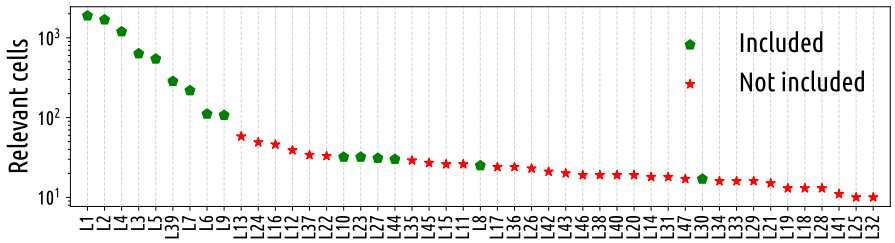


Figure 5.7. For each spatial profile, the number of cells that are considered relevant for it($loading \geq 0.1$), where green marks the profiles included in this analysis and red marks the non-included.

are known for having an active nightlife, such as Bastille and Bercy in Paris or the 1st arrondissement area in Lyon. Some regions known for relatively higher crime incidence also emerge, such as Gerland in Lyon.

5.1.4. Spatial structures in mobile traffic consumption

The different ways spaces in a city are related can be better understood by setting EFA variables as the spatial cells related to where the mobile traffic is, and the EFA samples as the 30-minutes aggregated time slots during the complete 12 weeks span of the dataset. This results in factors that provide insights on similar uses of smartphones through different regions of a city. Also, exploring the scores of each EFA factor can shed more knowledge about temporal behaviors. Patterns are summarized as *long-term behaviors* (such as commuting, working, relaxation) and *short-term events* (such as football matches, concerts).

A total of 47 factors resulted from the EFA analysis of spatial structures. This considerably big number is explained by the proportions of the dataset: a total of 2 287 cells from both Lyon and Paris were used as variables, with 4 080 30-minutes slots for mobile traffic used as samples, resulting in a long set with a significant variety of behaviors. It's important to note that not all profiles have the same importance; only a small portion capture large geographical areas with the majority being common to a small number of neighborhoods characterized by very specific human activity features. The number of cells for each profile that have $loading \geq 1$ is highlighted on Figure 5.7, where green marks the profiles of this analysis, including the 9 with over 100 relevant cells which represent long-term land use behaviors, as well as a few short-term land use behaviors that show interesting spots of the cities. A summary of all included profiles is presented in Table 5.3, and discussed in the following subsections.

5.1.4.1. Long-term land use behaviors

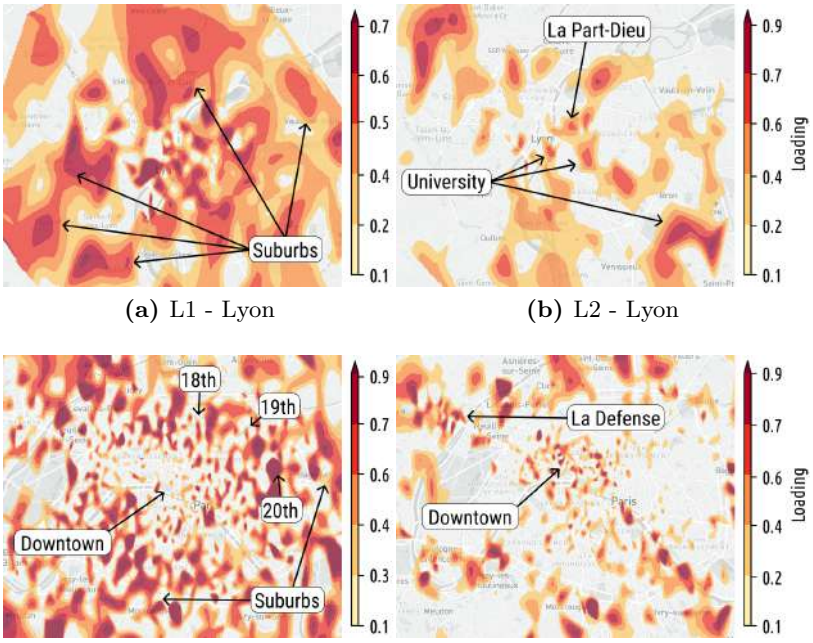
By observing constant behaviors across weeks and months, a number of profiles can be tied together by their long-term usage pattern. Those are determined by the types of

Factor	*Land Use*	*Examples*
L1	Residential, relaxation	Cities around Paris, Dense Residential such as 18th, 19th and 20th arrondissements.
L2	Working places	La Defense, Paris-Nanterre, La Part Dieu
L3	Long-range commuting (Train and RER)	Gare du Nord and L'Est, Rueil-Malmaison
L4	Short-range commuting (Metro)	Subway stations all around downtown Paris and Lyon.
L5, L7	Shopping places	La Confluence, Les Halles, Opera, Elysees.
L6	Cargo and maintenance sites	Gare du Nord, Saint Ouen, Noisy-le-Sec
L9	Clubs and night life	Bastille, Lyon 1st arrondissement.
L8, L30	Sport stadiums	Stade de France and Parc des Princes Stadium
L10, L23, L27	Expo centers	Expo Porte de Versailles, AccordHotels Arena, Euroexpo Center
L44	Catholic churches	Notre-Dame des Champs, Saint-Augustin, Sainte Trinité, Sainte Geneviève des Grandes Carrières

Table 5.3. Trivial land use cases for EFA spatial structures analysis, for both long and short-term behaviors.

land where loading values are higher as well as the scores obtained from the temporal samples. This results in a set of trivial land uses in cities, such as working places, residential neighborhoods and public transportation stations. The long-term land uses profiles present similar behaviors to the network profiling discussed in Section 5.1.3.1, but with more detailed spatial patterns than those in the network profiling maps.

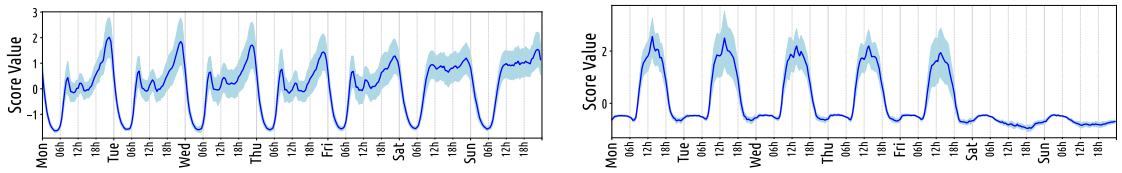
Land use profiles **L1** and **L2** exemplify well the extra details obtained from the spatial structures. The first one presents spaces characterized by **relaxation activities** (dense residential and suburban regions), while the second relates to **work/study locations**. The duality of these factors can be seen in Figure 5.8: active locations during the weekday, such as business districts (La Defense in Paris, La Part Dieu at the 3rd arrondissement of Lyon), universities and downtown areas are highlighted in L2, but disappear in L1. In the same way, suburbs and residential neighborhoods of Paris (such as the dense residential 18th, 19th and 20th arrondissements) are highlighted in L1 and not shown in L2. Scores seen on Figure 5.9 highlight the differences over time for the land uses: L1 has higher then average scores around night time, while L2 scores are higher around 8am till 8pm on working days and lower values around the weekend.



(a) L1 - Lyon **(b)** L2 - Lyon

(c) L1 - Paris **(d)** L2 - Paris

Figure 5.8. Land usage loading maps for long-term behavior profiles L1 and L2 in (a,b) Lyon and (c,d) Paris.



(a) Profile L1 **(b)** Profile L2

Figure 5.9. Land usage scores across time for long-term behavior profiles (a) L1 and (b) L2, for both Lyon and Paris.

Profiles L3, L4 and L6 present land uses related to the different usages of **public transportation**. Profile **L3** is associated to stations used for medium/long-range commuting, which offer high speed trains (TGV), standard regional trains (TER) and metropolitan trains (RER, exclusive to Paris); most of the major train stations of the cities emerge, with details seen on Figures 5.10a and 5.10b. Profile **L4** links to short-range commuting (metro stations), which explains the higher density of relevant spatial cells in the downtown areas, as it can be seen on Figure 5.10c with a few key metro stations highlighted. Profile **L6** is a profile that shows operational spaces for metro/train, such as a maintenance center for the SNCF railway company in Saint Ouen and Noisy-le-Sec, both located in the metropolitan area of Paris and seen on Figure 5.10f; the exit yard for trains of both Gare du Nord and de l'Est is also observed. Scores can assist the differentiation between all public transportation spaces. As seen on Figure 5.11a,

L3 sees peaks on the beginning and end of weekdays, which relates to the use of those long-range stations for the commuting to work (a common behavior for both cities due to their business and economic importance), and less during weekends. L4 sees more uniform scores across the day, showing that those short-range stations tend to have a more uniform usage across time, while the cargo and maintenance spaces of L6 see their activity rise late in the evening.

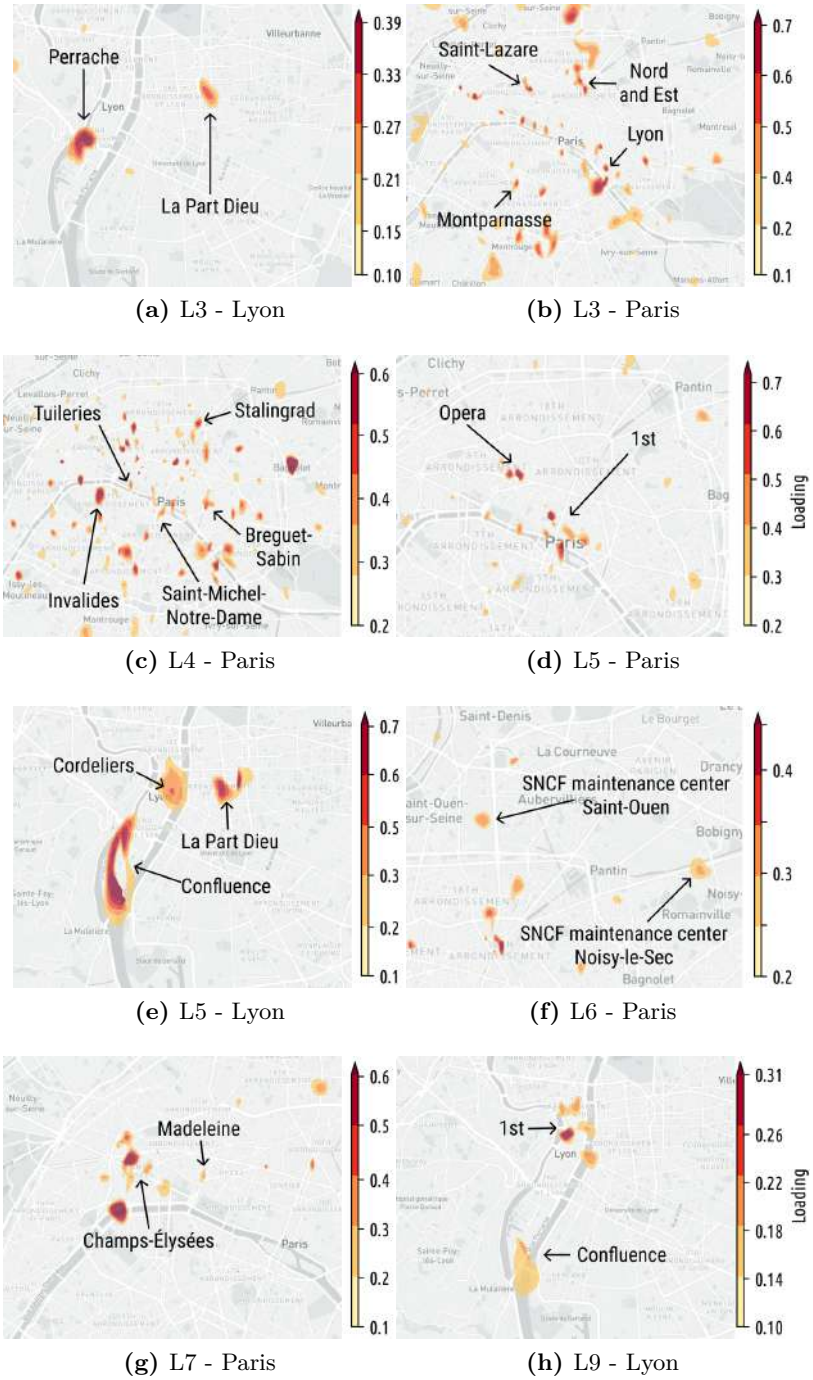
Profiles L5 and L7 relate to **leisure areas** of the cities, including entertainment spaces such as stores and restaurants. Profile **L5** presents locations known for their shops, bistros and galleries; this is seen on Figure 5.10e in Lyon for La Confluence, La Part Dieu and Cordeliers, as well as many shopping regions around Paris' 1st arrondissement (such as the Forum des Halles mall) and Opera regions (where many luxury galeries are present), seen on Figure 5.10d. Meanwhile, Profile **L7** presents commercial locations in Lyon's Ainay and Paris' Elysees and Madeleine, as seen on Figure 5.10g. L5 and L7 can be better differentiated when analyzing their scores over time. Profile L5 shows on Figure 5.11b a behavior previously seen in Section 5.1.3.1, with higher scores occurring on Saturdays (the preferred weekend shopping day in France); in the meantime L7 has more uniform scores from 8am to 8pm all days of the week. This might indicate that places around L5 are preferred by the population for shopping, while L7 shows a more mixed use that, although affected by a significant presence of commercial location, observes a more constant traffic due to the presence of offices.

The final long-term land use profiles explored is **L9**, seen for Lyon on Figure 5.10h highlighting locations known for their **late night attractions**. This is present on the 1st arrondissement of Lyon and the region of La Confluence, both having a good density of night clubs. Paris similarly has a good number of cells with high loading around Bastille, Pigalle and the 1st/2nd arrondissements, which are also known for clubs and restaurants opened until late night. When looking at the score for L9 on Figure 5.11c, the behavior in time of those locations match their usage, with values increasing on Fridays and Saturdays after 6pm, and being at the lowest from Sunday until Wednesday.

5.1.4.2. Short-term land use behaviors

The exploration of spatial structures also present regions of the cities used for short-term events that happen at punctual moments in time, e.g., during a few hours of one or multiple days. Such short-term patterns can be found when analyzing EFA scores for the time samples, which exhibit above average values for short intervals only.

A good example of short-term behaviors is provided by a couple profiles that emerge for stadiums in Paris. Profile **L8**, seen on Figure 5.12a, is related to Stade de France (the biggest sports stadium in France), while Profile **L30** shows the Parc des Princes (used by the Paris Saint Germain football team). The scores for each profile indicate in which moments those spaces saw mobile traffic that differentiate them from their neighborhood.



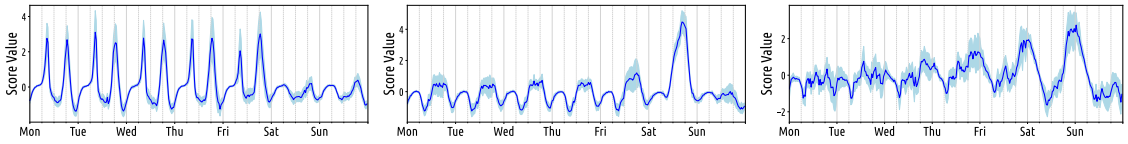
(a) L3 - Lyon **(b)** L3 - Paris

(c) L4 - Paris **(d)** L5 - Paris

(e) L5 - Lyon **(f)** L6 - Paris

(g) L7 - Paris **(h)** L9 - Lyon

Figure 5.10. Land usage loading maps for long-term behavior profiles: L3 in (a) Lyon and (b) Paris, (c) L4 in Paris, L5 in (d) Paris and (e) Lyon, (f) L6 in Paris, (g) L7 in Paris and (h) L9 in Lyon.



(a) Profile L3 **(b)** Profile L5 **(c)** Profile L9

Figure 5.11. Land usage scores across time for long-term behavior profiles (a) L3, (b) L5 and (c) L9, for both Lyon and Paris.

For State de France (L8), four dates during night emerged: Oct. 7th (football match between France and Bulgaria), Nov. 11th (football match between France and Sweden), Nov. 19th (rugby match between France and Australia) and Nov. 26th (rugby match between France and New Zealand). Those dates are seen on Figure 5.12b, which shows the scores for L8 across time. For Parc des Princes (L30), the scores show all of PSG home matches during the studied period: Sep. 9th (vs. Saint-Étienne), Sep. 13th (vs. Arsenal), Sep. 20th (vs. Dijon), Oct. 1st (vs. Girondins), Oct. 19th (vs. Basilea), Oct. 23th (vs. Basilea), Nov. 6th (vs. Stade Rennais) and Nov. 19th (vs. Nantes).

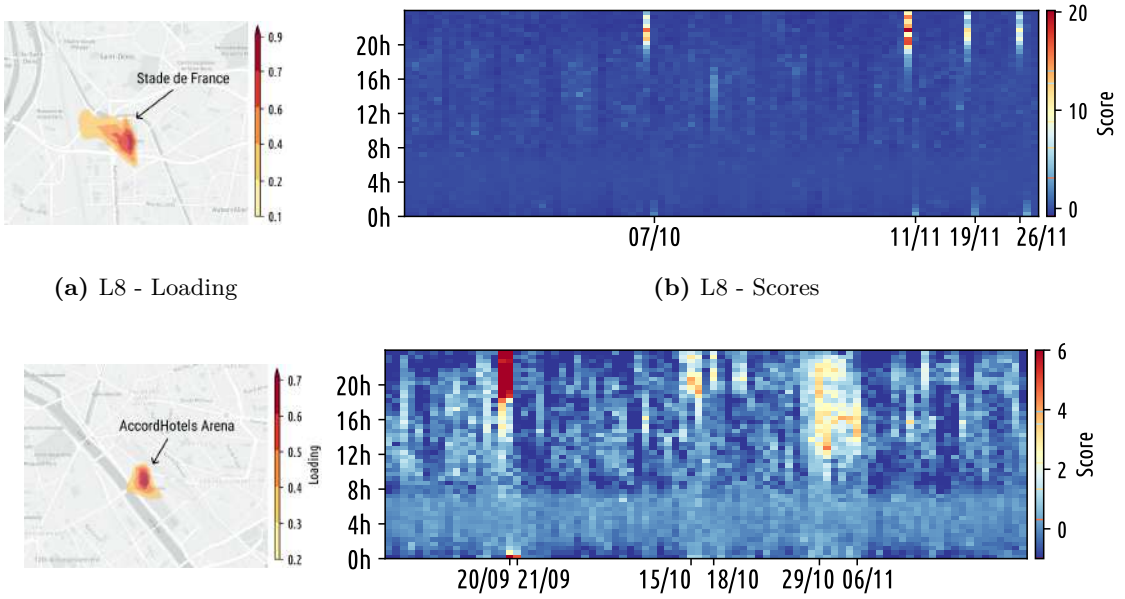
Three factors relate to exposition centers. Profile **L10** shows both the Euroexpo center in Lyon, as well as the Expo Porte de Versailles in Paris; Profile **L23** shows AccorHotels Arena, an indoor arena and concert hall located in the neighborhood of Bercy in Paris and seen on Figure 5.12c; Profile **L27** shows exclusively the Expo Porte de Versailles in Paris. Like was seen with stadiums, the scores for the profiles represent specific events that made those regions unique for their land usage. L23 had three different events happening at the arena during the studied period: two shows happening on the nights of Sep. 20th and 21st, three music concerts happening on Oct. 15th, 16th, and 18th; and a tennis tournament hosted between Oct 29th and Nov. 6th. The scores of the events of Profile L23 are seen in Figure 5.12d: it can be noted that the concerts have higher values at night time, while the tennis tournament starts around morning time and span throughout the entire week. It is also interesting to note the separation on scores for Oct. 17th, which presents lower values since no event happened at the arena on this date.

5.1.5. Mixed land use regions

A significant number of cells present high loading values for multiple spatial profiles, which reveals the presence of mixed land usage in the cities. To find occurrences of cells with mixed land use, the RCA metric is utilized (as discussed in depth in Section 3.5.5). For the problem presented in this Section, the metric will be described as:

$$RCA_{xf} = \frac{\lambda_{xf} / \sum_{x'=1}^{N} \lambda_{x'f}}{\sum_{f'=1}^{K} \Lambda_{xf'} / \sum_{x'=1}^{N} \sum_{f'=1}^{K} \lambda_{x'f'}} \tag{5.14}$$

where for this case $x \in \mathbf{X}$ is a cell that belongs to the $N \times 1$ vector of observed variables



(a) L8 - Loading **(b)** L8 - Scores

(c) L23 - Loading **(d)** L23 - Scores

Figure 5.12. Land usage loading maps in Paris for profiles (a) L8 and (b) L23, as well as the scores across time for (c) L8 and (d) L23.

$\mathbf{X}$, $f \in \mathbf{F}$ is a profile that belongs to the $K \times 1$ vector of common factors $\mathbf{F}$, and λ_{xf} is the loading value of a given profile f for cell x that belongs to the $N \times K$ matrix of common factor pattern coefficients $\mathbf{\Lambda}$.

Analyzing the histogram for the number of land use profiles per cell for Paris and Lyon in Figure 5.13a, a fairly small number of cells is noted to have a single land use; the majority of cells have instead between 3 and 4 significant profiles. It is important to highlight that regions with a higher density of uses are mostly seen in downtown regions, like the Paris' regions north of the Seine river, as well as 1st and 2nd arrondissements in Lyon. Similarly, areas on the suburban parts tend to have less mixed usage, such as the southern zones of Paris and suburban areas.

To exemplify the identification of regions with mixed land use through EFA, a few examples of cells that had $RCA_{xf} > 1$ for two land use profiles will be explored. Maps that showcase cells with mixed usage use of two profiles $[f_1, f_2]$ across cities are prepared by calculating for each significant cell $x \in \mathbf{X}$ that has $[RCA_{xf_1}, RCA_{xf_2}] > 1$ the value $(RCA_{xf_1} - 1) * (RCA_{xf_2} - 1)$, which will represent the *mixed usage RCA coefficient* for each set $[f_1, f_2]$. With this, it's possible to better observe the incidence and spatial pattern of mixed use cells over the cities. The first example encompasses cells with mix use of relaxing/residential (L1) and working (L2) hours profiles, a mixed land use that occurs on 209 of the total 1752 cells (11.93%) that are significant for either L1 or L2. Their distribution over space can be seen on Figure 5.13b, from which it can be highlighted

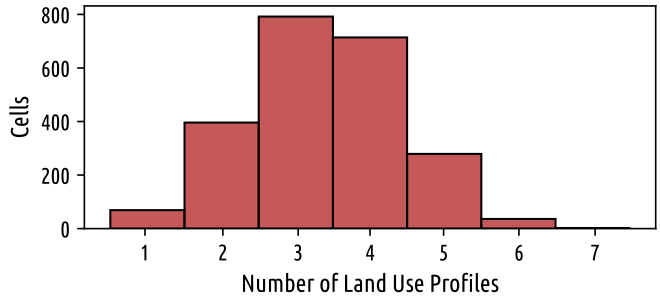
that the presence in Paris of many cells around the 1st and 7th arrondissement, which are highly touristic regions in the city, with many shopping stores, hotels and residential apartments. It's also observed that the 10th arrondissemnt, which although not a typical tourist spot of the city, it's a densely residential area with the presence of the two main train stations of Paris (Gare du Nord and de l'Est). Those are regions that are active the entire day due to their extreme diverse zoning.

Another combination of land uses that is worth exploring is that between relax/residential (L1) and leisure (L5) profiles, where 332 of 1579 cells (21.03%) significant for either L1 or L5 have this mixed usage. Those locations prove to be active both in after-work hours and during weekends, and, as seen in Figure 5.13c, encompass zones featuring good mixture of parks, houses, restaurants and cultural spots, notably Pentes de la Croix-Rousse, Fourvière and Les Cordeliers in Lyon, as well as the 9th arrondissement and Marais in Paris. Finally, one of the biggest intersections identified is that between the leisure (L5) and late night (L9) land use profiles, where 460 of 1457 cells (31.57%) can be associated to both behaviors. As observed in Figure 5.13d, this happens at regions where people tend to spend their free time, both during the night and on the weekends, with popular neighborhoods in Paris such as Bastille, 3rd and 4th arrondissements, as well as the 1st and 2nd arrondissements of Lyon.

5.1.6. Main takeaways

This section proposed an original approach to the spatiotemporal classification of mobile traffic data, which relies on EFA. Extensive tests with heterogeneous real-world datasets demonstrate the versatility of EFA, which provides a unifying framework to solve problems that have been studied in isolation in the literature, i.e., mobile traffic profiling and land use detection. In both cases, EFA attains results that improve those of state-of-the-art solutions (e.g., the richer information of network activity profiles), or match them while yielding greater consistency (e.g., the better abstracted land use classes, where loadings can be leveraged for intra-class analysis). In addition, EFA provides supplementary knowledge (i.e., the geographical perspective of profiles and the temporal view of land uses) that proves paramount to the interpretation of the results, and eases tasks that are otherwise complex to perform (e.g., the analysis of per-hour temporal data, or the detection of mixed land uses).

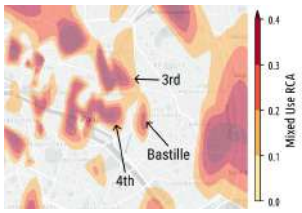
EFA-based classification can find multiple applications in data-driven studies, applicable across various levels and contexts. Firstly, this methodology offers fertile ground for travel-demand-related mobility studies. The temporal structures can provide precise insights into the key patterns of population presence across different days of the week. These patterns are linked to specific activities that people engage in within designated areas of the city. Factor loadings from the temporal description indicate times when individuals either currently require or will soon need access to transportation. For



(a) Distribution of cells with multiple land use profiles

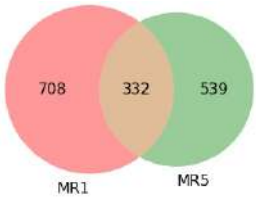
(b) L1 + L2 - Paris

(c) L1 + L5 - Lyon

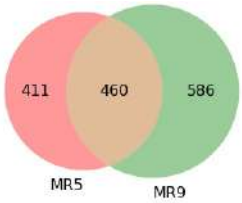


(d) L5 + L9 - Paris

(e) L1 + L2 - Venn



(f) L1 + L5 - Venn



(g) L5 + L9 - Venn

Figure 5.13. (a) Distribution of cells by the number of significant land use profiles; mixed use RCA coefficient and Venn diagrams for the mixed land use cases of (b-c) L1 + L2, (d-e) L1 + L5 and (f-g) L5 + L9.

example, in the context of commuting factors, the associated loadings clearly delineate the moments people are arriving at or departing from major transportation hubs, train stations and transit stops. Concerning work or residential factors, loading variations can precisely pinpoint population transition to different activities potentially located in other parts of the city, thus indicating potential peaks in the associated travel demand. On the other hand, the spatial analysis underscores the potential of this methodology for studying the attractiveness of specific urban areas over time. This is true for both recurring and episodic scenarios, such as special events, as observed through mobile phone data. This information is invaluable for traffic simulation and transportation planning, serving again as a precious proxy for travel demand data that can be otherwise challenging to acquire in a real-time like fashion. Specifically, the identification of areas occasionally or regularly experiencing high levels of activity can enable the easier identification of accessibility bottlenecks and support the optimization of transportation resources and transit schedules.

Moreover, the temporal structures identified by EFA expose non-trivial long-term dynamics in the mobile traffic demand that are relevant to the allocation of mobile communication resources in, e.g., Cloud Radio Access Networks (C-RAN) [212]. In addition, typical temporal profiles may serve as a basis for the detection of anomalous network usages, and for predicting the future demand in the context of transport resilience and anticipatory networking studies. In the spatial dimension, EFA classes neatly characterize the strong geographical locality of mobile demand. They can thus pave the way for cognitive network functions that aim at migrating network resources geographically, or at dynamically configuring the network topology; such functions are especially relevant to, e.g., Mobile Edge Computing (MEC) infrastructures [212]. Overall, EFA-based classification is a potential brick for future big data-driven 5G systems [213].

Future works using EFA together with mobile phone data can help in elucidating and modeling the specific links between the hidden structures and travel demand variations, building more robust models able to combine more traditional transportation data with information obtained from mobile network measurements.

5.2. Characterizing urban green spaces with mobile traffic

As populations across the globe become more centered around urban environments, cities face the challenge of creating gateway spaces, where individuals can be reconnected to nature and greenness, which can be an important factor in improving public health in those cities [214]. Urban Green Spaces (UGS)][3] are often described as places for physical, recreational, and relaxation activities inside cities, a runaway area from the stimulus encountered throughout major urban centers. During the age of information, where the internet (and all applications related to it) can be easily accessible from anywhere, due to the pervasiveness of smartphones and mobile networks, UGS could be considered a facilitator for a digital detox, where users can disconnect from their devices and applications.

Smartphones have been previously utilized as a proxy to understand the attraction factors of UGS (as discussed over Subsection 2.2.2). But, not much work has been done in the opposite way: how UGS affects the utilization of smartphones and their applications. This has limited a better comprehension of any healing effects parks and other green areas may have in relation to digital consumerism and smartphone activeness. More precisely, no other work has performed an in-depth analysis across mobile applications to understand how different categories of apps are affected by the UGS where users are connecting from. This raises a few questions: 1) are user overall utilizing less/more their smartphones in those spaces? 2) If such patterns are present, are they homogeneous across all urban green spaces? 3) Are the observed changes homogeneous across mobile applications,

[3]Throughout this Section, both UGS and Parks may be used interchangeably.

i.e., is the reduction observed across all apps or do certain classes of applications see an over/underutilization inside those studied areas?

To address those questions, this Section will present an in-depth characterization of smartphone and mobile applications consumption within urban green spaces, focusing specifically on a major European metropolis (Paris). This work encompasses a complete framework, first detailing the advantages, limitations, and all processing steps needed in order to perform such a fine-grained spatial analysis of mobile traffic, targeting specific sites and being able to filter out undesired traffic. Utilizing passively produced data also helps avoid biases that traditional surveys about smartphone usage within spaces could have. Next, a full characterization of spatiotemporal behaviors of smartphone consumption in UGS will be presented, laying an initial stone in fomenting mobile network traffic as a proxy to study fine-grained spaces within cities, and highlighting the heterogeneous impact of UGS over smartphone utilization.

5.2.1. Data processing for the study of smartphone utilization in UGS

This work utilizes mobile traffic time series collected from the production network of Orange, specifically within the city and metropolitan region of Paris, over the period of Jan/2024 to May/2024. This work makes no differentiation between RATs, as the main objective is to characterize the overall utilization of mobile traffic inside the targeted areas, independent of the network technology.

5.2.1.1. Data collection and preprocessing

Selecting applications and aggregating in categories: The collected data encompasses the time series of traffic generated at each antenna by each mobile application. Through the DPI of the MNO, 444 unique mobile applications and data protocols[4] are found. However, many of those applications may not be popular enough to generate temporal patterns that are desirable to analyze (i.e., measurement noise). It can also be noted that the distribution of traffic amongst applications is extremely imbalanced, with top 20 services generating almost 80% of all recorded sessions [5]. Also, whenever analyzing mobile traffic at the antenna level, the data will tend to become noisier (even on popular apps). Since the study of UGS involves is at almost antenna level (as the majority of parks is covered by less than 10), it will spark the necessity of filtering the applications utilized in the analysis and aggregating them into major categories to improve the data quality.

The set of apps within each selected key category for the study of UGS can be seen

[4]In some cases, the DPI is not capable of identifying the application due to the encrypted nature of mobile communications, leading to either a label related to the protocol (in case it was identified), or merely as *encrypted traffic*

Category	Mobile Application
Fitness	Garmin Connect
Games	Pokemon Go, Fortnite
Music	Apple Music, Spotify, Deezer, Soundcloud
News and Information	Wikipedia, Sports News, NewsPaper, NewsMag, Weather, Tripadvisor
Productivity	Skype, Microsoft Mail, Google Drive, Gmail, Finances, Dropbox
Shopping	Amazon
Social	Instagram, WhatsApp, Facebook, Snapchat, LinkedIn, Pinterest, Twitter, Facebook Messenger, TikTok
Travel	Uber, Waze, Airfrance, Transport
Video	Twitch, Periscope, Youtube, Netflix, Molotov TV, DailyMotion, Apple Video, Facebook Live

Table 5.4. Studied mobile applications for their usage within urban green spaces and their respective categories

on Table 5.4. It's important to note that those apps constitute 65.6%[5] of the total traffic generated in the studied network, meaning that this filter still encompasses the majority of traffic consumption patterns in the area.

A notable excluded app from the analysis is Strava, which would intuitively be in the fitness category. This was done due to how Strava generates traffic over its mobile app being closer to Social media than Fitness: during the time of the activity, Strava is not exchanging data with its servers; this is done solely when the application is over (which the BS that exchanged traffic about the activity may not relate with the places where the activity occurred). Therefore, the choice is made to remove it from the analysis.

5.2.1.2. Differentiating mobile traffic consumption within UGS

In order to assign traffic to parks, there's a need for a methodology that can confidently assign mobile traffic consumption to UGS. This is necessary in order to select BS that have their traffic mainly generated inside those spaces, excluding the ones that may have the demand shared within the neighboring regions. This methodology proposes a trade-off to maximize the data quality, resulting in a reduced number of studied spaces but minimizing the traffic generated from neighboring spaces falsely assigned to green areas (i.e., false positives).

This is done by creating the Voronoi geometries taking into account the azimuth angle (as discussed in Subsection 3.6.3). This allows selecting antennas that not only are near the UGS but also have their coverage focused in these spaces. This idea achieved previous success in detecting highway congestion through mobile network records, maximizing the avoidance of mobile traffic generated outside the desired spaces [91].

[5] 79.4% if the mentioned encrypted and protocol data are not considered, which would not be used in the analysis either way

The UGS geometries in France are obtained from OpenStreetMaps and are overlapped with Voronois (following the methodology discussed in Subsection 3.6.4). The objective is to choose only antennas mostly covering green areas, discarding any with significant coverage of spaces outside parks. To quantify the accuracy of the traffic correctly assigned to a park and its surroundings, a set of metrics is determined.

Considering the set of selected parks as $\forall p \in P$, and the set of voronois as $\forall v \in V$, the *illuminated park ratio* between a certain park p and antenna v is defined as:

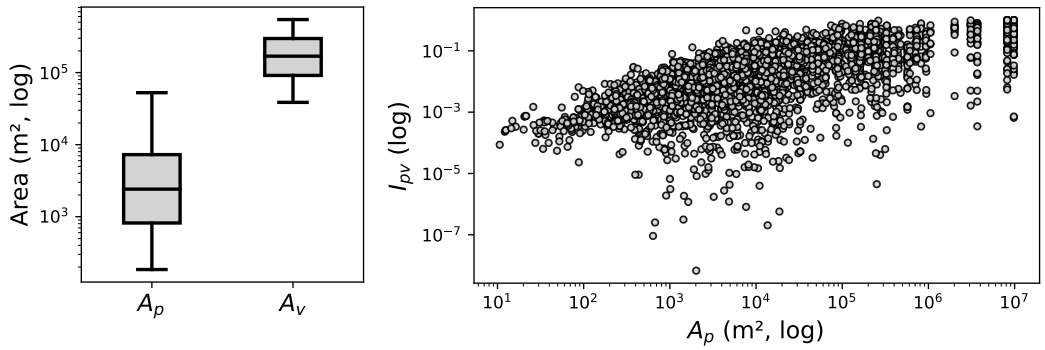
$$I_{pv} = \frac{A_v \cap A_p}{A_v} \tag{5.15}$$

where $[A_p, A_v]$ are the respective areas of park p and antenna v; a value equal to zero indicates that antenna v is not covering the park, while a value equal to 1 means the entire coverage area of the antenna is inside the park. The desire is to maximize $I_{pv} \forall [p, v]$ in for the set of V.

After calculating I_{pv} for $[P, V]$, a relation between the areas of parks and expected coverage is observed. As it can be seen in Figure 5.14a, the median area of Voronoi A_v inside the studied urban area is two orders of magnitude higher than the median area of parks A_p, with the 5^{th} percentile of A_v being on a similar scale as the 95^{th} of A_p. This means that a good portion of UGS may be *too small* to be correctly assigned traffic exclusively generated inside their area, as the covering Voronoi will be significantly bigger than it, making it hard to disassociate traffic in it versus from the streets. Therefore, it's a better choice to utilize only parks that have their area $A_p > \alpha$, where α will be the *minimum park area threshold* Figure 5.14b shows that the illuminated park ratio gets significantly better the bigger the area of the park. Therefore, it's important not only to select antennas that have their coverage mainly inside parks, but also where the filter will benefit from cutting parks that may be too small for the problem.

To select the final set of UGS with good coverage and the antennas covering them, two additional metrics are defined to evaluate I_{pv}, A_p, and A_v. Consider a park p with area A_p, covered by a set of $\tilde{V}$ antennas: a given antenna $v \in \tilde{V}$ with area A_v will have it's area illuminating A_p defined as the overlap $A_{i_v} = A_v \cap A_p$. The area of A_v outside A_p will be the relative complement $A_{n_v} = A_v \backslash A_p$, which is also equal to $A_n = A_v - A_i$. Consider β as the *minimum accept threshold* of illumination I_{pv}. Therefore, the set $\tilde{V}$ will be split in two: $\tilde{V}_{sel}$ is composed by $\{\forall v \in \tilde{V} \mid I_{pv} > \beta\}$; $\tilde{V}_{cut}$ is composed by $\{\forall v \in \tilde{V} \mid I_{pv} \leq \beta\}$. The precision in the coverage of antennas $v \in \tilde{V}$ covering park p is defined as:

$$precision_p = \frac{\sum_{v \in \tilde{V}_{sel}} A_{i_v}}{\sum_{v \in \tilde{V}_{sel}} A_{i_v} + \sum_{v \in \tilde{V}_{sel}} A_{n_v}} \tag{5.16}$$



(a) **(b)**

Figure 5.14. (a) Boxplot of the areas of parks A_p and Voronoi A_v, where whiskers indicate the 5^{th} and 95^{th} percentiles; (b) Relation between the illuminated park ratio I_{pv} and park area A_p, where the Pearson correlation is 0.61, indicating that for parks with bigger area, it's more likely to guarantee that the traffic generated is mostly exclusive by users within those spaces.

The recall can also be defined as:

$$recall_p = \frac{\sum_{v \in \tilde{V}_{sel}} A_{i_v}}{\sum_{v \in \tilde{V}_{sel}} A_{i_v} + \sum_{v \in \tilde{V}_{cut}} A_{i_v}} \tag{5.17}$$

Finally, the quality of coverage in park p will be:

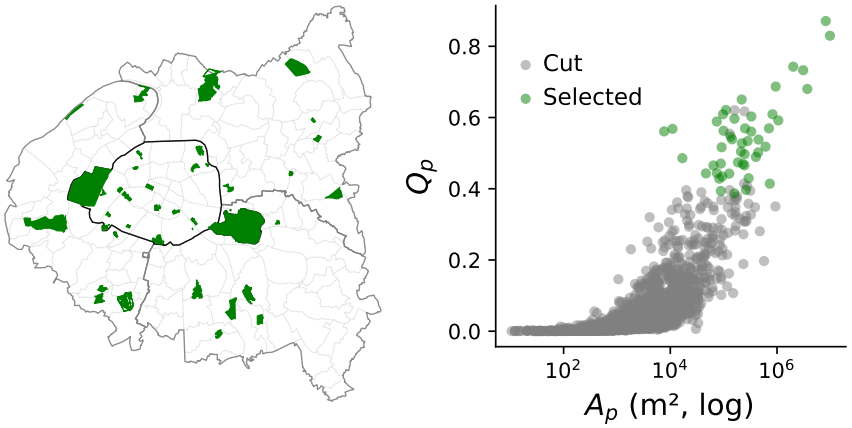
$$Q_p = 2 * \frac{precision_p * recall_p}{precision_p + recall_p} \tag{5.18}$$

Only UGS above a *minimum coverage threshold* $Q_p > \gamma$ will be selected for the analysis. The final goal is to select a set of values $[\alpha, \beta, \gamma]$ that results in the best set of UGS and antennas, respectively optimizing: 1) *minimum covered park area* α, 2) *minimum illuminated ratio* β, 3) *minimum quality of coverage* γ. From qualitative testings, the chosen values for $[\alpha, \beta, \gamma]$ were $[0.8, 0.1, 0.4]$.

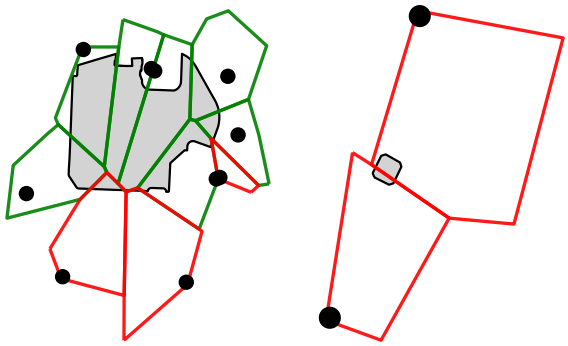
5.2.2. Results from isolating mobile traffic consumption inside UGS

As expected, not all UGS could have mobile traffic confidently assigned to them, with a considered certain that it was not generated in the neighboring areas. Indeed, considering the full set of parks P which contained 2406 parks, only 47 were selected for this analysis (seen over space in Figure 5.15a and detailed in Table 5.5). This was due to the significant influence the UGS area A_p had on the quality of coverage Q_p of each park: as observed in Figure 5.15b, the smallest UGS (with less than 10^3 m^2) had $Q_p \to 0$. As the area grew, the quality of coverage proportionally improved, with all selected parks having more than 10^4 m^2.

This relation between Q_p and A_p is greatly observed when looking at the Voronoi geometries. Figure 5.15c showcases a **selected UGS**: Jardin du Luxembourg, a medium-



(a) All selected UGS **(b)** Q_p x A_p



(c) Selected UGS **(d)** Cut UGS

Figure 5.15. (a) Distribution over space of the selected 47 UGS in Paris; (b) The relation between the UGS area and the quality of traffic coverage; Examples of a (c) selected and (d) cut UGS, showcasing the Voronoi covering them.

sized park in Paris. Two different sets of antennas are seen covering its space. First is the set of selected antennas (in green), which are considered sufficiently covering the park area, in relation to their total area A_v. All of those antennas had $I_{pv} > \beta$. However, a set of 3 antennas was removed (in red) due to having $I_{pv} < \beta$. Finally, considering the full set of antennas covering this park, the final confirmation came from its $Q_p > \gamma$, which means the overall quality of coverage of the park is sufficient for the analysis to be done, with also the total area of the park covered by the selected Voronois being above α. Not all parks however had sufficiently good coverage. Figure 5.15d presents the case for a **cut UGS**: Square Samuel Rousseau, a smaller park located in the 7th arrondissement of Paris. As it can be seen, this park is covered by two antennas, but unfortunately due to its smaller A_p, both the antennas had $I_{pv} < \beta$, which means it was not possible to confidently distinguish park and neighbor mobile traffic consumption, leading to the cut of this park from the analysis.

osm_id	Park name	A_p (km^2)	Cluster
4083189	Parc de Bercy	0.14	Touristic Parks
173204460	Parc des Buttes-Chaumont	0.25	Touristic Parks
53820452	Jardin des Tuileries	0.22	Touristic Parks
176234171	Square des Batignolles	0.02	Touristic Parks
305316290	Parc Pierre-Lagravère	0.29	Touristic Parks
43324060	Parc Georges Brassens	0.07	Touristic Parks
466176263	Cité Universitaire - Parc Ouest	0.07	Touristic Parks
466176883	Cité Universitaire - Parc Est	0.26	Touristic Parks
683921616	Jardin du Ranelagh	0.1	Touristic Parks
4050160	Parc Montsouris	0.16	Touristic Parks
6530383	Parc Jean-Moulin - Les Guilands	0.25	Touristic Parks
7574623	Parc de la Villette	0.33	Touristic Parks
7615434	Parc du Lycée Michelet	0.11	Touristic Parks
4430607	Bois de Boulogne	8.2	Touristic Parks
10885726	Domaine national de Saint-Cloud	3.69	Touristic Parks
4208595	Champ de Mars	0.28	Touristic Parks
142107768	Bois de Vincennes	9.8	Touristic Parks
2768926	Parc de Sceaux	1.73	Lunchbreak Parks
151567211	Parc André Citroën	0.09	Lunchbreak Parks
12273749	Parc Martin Luther King	0.1	Lunchbreak Parks
2826933	Parc Monceau	0.08	Lunchbreak Parks
154754263	Jardins Abbé Pierre - Grands Moulins	0.01	Lunchbreak Parks
13716106	Jardins du Trocadéro	0.14	Lunchbreak Parks
61366951	Parc Départemental des Chanteraines	0.71	Lunchbreak Parks
4221369	Jardin des Plantes	0.16	Lunchbreak Parks
14036411	Parc Suzanne Lenglen	0.19	Lunchbreak Parks
36902337	Parc Floral de Paris	0.33	Lunchbreak Parks
128206209	Jardin du Luxembourg	0.22	Lunchbreak Parks
10928376	Parc departemental Georges Valbon	3.09	Residential Parks
19578842	Parc de la Butte du Chapeau Rouge	0.05	Residential Parks
23411982	Parc des Sports du Grand Godet	0.21	Residential Parks
26374587	Parc Lefèvre	0.09	Residential Parks
29469692	Parc de la Fosse Maussoin	0.23	Residential Parks
34871331	Parc des Artistes	0.09	Residential Parks
91280888	Arboretum de Paris	0.13	Residential Parks
700409563	Parc des Saules	0.06	Residential Parks
39167900	Parc départemental du Sausset	2.01	Residential Parks
39239627	Parc des Sports - Plaine Sud	1.06	Residential Parks
448261881	Parc de la Vallée-aux-Loups	0.44	Residential Parks
44298483	Parc départemental des Lilas	0.82	Residential Parks
230599024	Parc départemental de la Haute-Île	0.73	Residential Parks
196945196	Parc Henri Sellier	0.25	Residential Parks
46172340	Parc départemental de la Plage Bleue	0.4	Residential Parks
1721301	Île de Loisirs de Créteil	0.8	Residential Parks
1060244887	Parc Nature du Plateau d'Avron	0.11	Residential Parks

Table 5.5. Selected parks, grouped by the cluster results.

Key insights. *It's possible to isolate quality mobile network traffic measurements that are generated within UGS, but there's a significant need for caution in relation to the area of those spaces. The methodology described can help better access this selection process, and could also be adapted for other urban spaces.*

5.2.3. The influence of UGS on smartphone usage

With the selected set of UGS, the characterization part of this study can start. The first open question is: *"Do green spaces lead users to interact with their smartphones in different ways than in other points of the city?"*. In case this hypothesis is true, the objective will be to differences the spatiotemporal traffic consumption in order to clarify how and why parks may lead to different interactions of users with their smartphones.

5.2.3.1. Parks experience different temporal patterns of smartphone usage

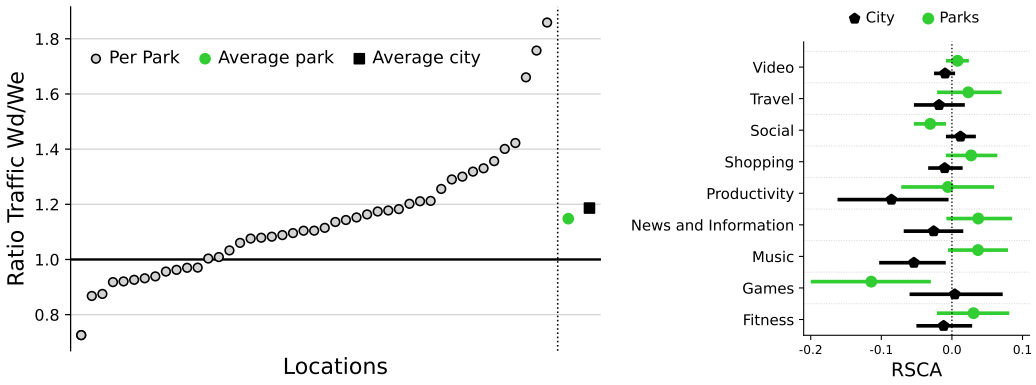
This first step involves understanding the temporal aspect of mobile traffic consumption within the studied spaces, i.e., is there any preference for traffic consumption within UGS in relation to weekdays or weekends?

Figure 5.16a presents for the selected parks the distribution of the ratio of traffic consumption across an average weekday versus an average weekend. Values above 1 mean the space has more traffic on an average weekday than an average weekend, while values below 1 represent that average weekends have see more traffic consumption than weekdays.

As it can be noted, there's a significant variation across the values ($[0.72, 1.86]$). By analyzing the edge values, an early insight into why those divergences happen appears: The bottom three parks[6] with the most weekend traffic are parks mostly oriented to nature activities. Two are located outside of Paris and have a good number of trees and dense green spaces, while the other (Parc Georges Brassens) is located on the outskirts of Paris but has a good number of nature appreciation and leisure locations. Either way, all 3 of those parks with higher weekend consumption are located in mostly residential areas. The top three parks[7] represent the ones where traffic is consumed more on weekdays. Those are all more centralized within the urban perimeter of Paris and are all also located in more mixed areas, where both offices (Monceau) and universities (Grands Moulins) are located. Parc Monceau has a good mix of benches (which could be used as an escape place during weekdays) and also sports activity features (such as a running track). The park with the highest ratio consumption on weekdays is Parc Suzanne Lenglen, which interestingly is the only dedicated sports center park included in the analysis.

Those initial results do not represent necessarily that nature and leisure parks lead

[6]Parc départemental de la Fosse Maussoin, Parc Georges Brassens and Parc de la Vallée-aux-Loups
[7]Parc Suzanne Lenglen, Jardins Abbé Pierre - Grands Moulins and Parc Monceau



(a) (b)

Figure 5.16. a) Ratio of traffic between an average weekday and an average weekend day, both for individual UGS (in gray), the average across all UGS (in green), and the average across the remaining of the city (in black). b) Distribution of RSCA values across UGS and the remainder of the city, representing how popular (RSCA>0) or unpopular (RSCA<0) mobile applications are across those areas. Markers represent the mean value, while lines represent the 95% confidence interval of the values across selected areas.

to a reduction in smartphone usage. Instead, they show that smartphones are used more on weekends, which could be hinted as a proxy for a bigger presence of users in those locations on the weekends. What this could represent is that those outskirt nature parks are preferred as weekend destinations.

Key insights. *The temporal aspect of mobile traffic consumption is heterogeneous across UGS, with preference varying between average use on weekdays and weekends. A brief look into edge cases shows that UGS with more smartphone use on weekdays tend to be more central and close to work and study places, while parks with more smartphone use on weekends are located in the residential outskirts and have a more natural presence.*

5.2.3.2. Preference for smartphone applications is different in parks than in the rest of the city

Besides temporal changes, understanding and characterizing if mobile applications are utilized differently inside UGS can help comprehend if those spaces lead to different behaviors than the remaining areas of cities. As mobile service usage is significantly heterogeneous across the users of the network, studying their variation over spaces can help quantify how their presence in UGS may affect their smartphone consumption.

For this analysis, the set of 43 mobile applications grouped into 9 categories is (as described in Subsection 5.2.1.1). The goal is to understand if certain groups of apps are utilized differently in UGS, in relation to the remaining spaces of Paris. An important aspect to note is that there's a significant variation in traffic volume generated across categories of mobile applications (i.e., video streaming apps inherently consume more

traffic than news and information applications), which can result in volume bias in the analysis of mobile services and lead to incorrect results in modeling algorithms. To overcome this, the RSCA metric is utilized (as described in Subsection 3.5.5) to compare the importance of applications across the studied areas, indicating how popular (RSCA>0) or unpopular (RSCA<0) an app is in a certain area, in relation to all other areas and apps. Those values are calculated across all categories of mobile applications for 2 sets of spaces: 1) all selected UGS and 2) the remaining spaces of the city. It's important to note that while the set representing the remaining of the city will contain green areas not selected, it can be assumed that those will not influence traffic since any park-related pattern is diluted through the other patterns present in the city (and is the reason those antennas were not selected).

The relative importance of applications across both groups is seen in Figure 5.16b, where the mean values and the 95% Confidence Interval (CI) obtained are presented. A few key differences in the usage of mobile apps within UGS can be highlighted: A set of application categories is characterized by almost complete overutilization in UGS while being underutilized in the rest of the city (with almost no overlap in their CI). These categories are Music and News and Information. This means there's a significant preference for usage of those categories across all parks, which does not occur across the rest of the city. Other categories also have this relation of over/underutilization in UGS/city, but with a higher overlap of the CIs, which would mean less homogeneous results across all areas that compose the distributions. This means there's a preference for usage of those categories in UGS, but some areas of the city may also present a preference. This includes Fitness, Video, Shopping, and Travel applications.

A few categories present the opposite: less importance in UGS, when compared to a higher preference of usage in the rest of the city. For example, Social Media apps show a very clear distinction, where all UGS present RSCA values below 0, while the majority of the city presents values above 0. This could mean that there's a significant underutilization of social media apps within UGS. This could have a few causes: at first, it can be said that green spaces would drive an *overall reduction* of smartphone usage, leading to less use of social media (which is one of the highest generators of mobile traffic). But, as previously noted, other categories had an overutilization in UGS, so this would not be the sole case. A more probable cause could be that users would have other stimuli, leading them not to crave the classical stimulus resulting from Social Media. As previously observed, users would be more willing to listen to music or read more *traditional* websites, instead of social media. Similarly, Gaming applications also have a clear underutilization in UGS, while their patterns are more heterogeneous across the city (some spaces with over, while others with under, due to the CI). The explanations for the underutilization of Gaming applications could be similar to Social Media, as those would also be less interesting for users in UGS as they'd tend to be less idle and have more outside stimulus.

Finally, it's interesting to note that Productivity apps have an underutilization across the city, but do see some overutilization in some UGS. While this sounds counterintuitive at first, this could represent the heterogeneity in types of parks and the locations they're placed within the city, as some users could go to central parks close to their work locations and still check their e-mails and have work calls. This will be explored later during Subsection 5.2.4.1.

It's important to note that the CI of RSCA values has a great variation in the importance (fluctuating between under and overutilization) of most categories among green and non-green spaces, which indicates that some UGS may have usage diverging from the overall trend, requiring an extra analysis in how different parks may lead to different patterns of smartphone usage.

Key insights. *The preference of which apps are utilized more in parks is different from the rest of the city, with apps related to Fitness, Music, News, Shopping and Travel being used more in parks, while Social Media and Gaming categories see a decrease in usage. The drops in usage of those two categories could be linked to a reduction in idleness in those spaces, leading users to crave less for those types of applications. Still, the CIs of RSCA are high, indicating a heterogeneity in the preference of applications within parks, signifying that not all parks influence smartphone consumption on the same way.*

5.2.4. The heterogeneous influence of UGS over smartphone use

After confirming that it's possible to differentiate UGS mobile traffic, noticing that they have significant temporal variations in their traffic consumption of mobile services and that this differs in UGS in relation to the city, the next step is to explore why those changes are happening, especially in order to understand the heterogeneity of behaviors in parks which could be noted by the wide CIs of the UGS RSCA. This next set of results presents a quantitative analysis of the differences in smartphone usage and mobile traffic consumption across the selected set of UGS.

5.2.4.1. Clustering parks according to preferences in applications

In order to identify and quantify what are the differences across UGS in relation to smartphone usage, a hierarchical clustering algorithm is performed, selecting as features the ratio of traffic on weekdays vs. weekends (seen in Figure 5.16a), as well as the RSCA across the app categories (seen in Figure 5.16b). The details are also seen in Table 5.5. After running the clustering, the Silhouette score is used to determine the optimal cut of the dendrogram resulting from the clustering algorithm, leading to a total of 3 groups of UGS in relation to those features.

The spatial distribution of the 3 classes can be seen in Figure 5.17a, with the median week of traffic for each cluster being presented in Figure 5.17b. By analyzing the regions

where each cluster element is located, as well as their temporal patterns, names are given for each: Residential (in blue), Touristic (in red), and Lunchbreak parks (in green).

Next, a further exploration of the resulting clusters will be made. Firstly, the Residential cluster (blue) is characterized by parks that are located in the suburban parts of metropolitan Paris (known as Banlieu and composed of cities in the departments of Hauts-de-Seine, Seine-Saint-Denis and Val-de-Marne). According to the 2017 Census, 68% of the inhabitants of the metropolis of Paris live in this area, which is considerably less dense than the center of Paris and more focused on residential areas. Interestingly, the UGS inside this cluster are the ones with the lowest consumption of traffic amongst the studied set: they consume on average 55% less traffic than the other parks studied.

Next, the Lunchbreak group (green) is present in areas denser with offices and universities. But, more than just the space, what truly brings those parks together is the temporal aspect of their traffic consumption: as observed in Figure 5.17b, their traffic is highly concentrated within weekdays (the median traffic on a weekday is 58% higher than the median traffic during weekends). As expected, UGS of this cluster dominates the right side of Figure 5.16a. These UGS are generally smaller (median area of UGS inside this cluster is 36% than other clusters) and less touristic, where regular users may not necessarily live in this area, but actually are working/studying in the surroundings and utilizing them during their breaks (i.e., having lunch).

Finally, the Touristic cluster (red) represents mainly parks either located in tourist areas of Paris (i.e., Jardin des Tuileries) or known by their own attractions (i.e., Champ de Mars and Eiffel Tower; Parc de la Villette and the Cité des Sciences et de l'Industrie museum and concert venues). Interestingly, these UGS have the most even distribution of traffic across weekdays and weekends, with working days having only a median of 7% more traffic than weekends. This could be explained due to these UGS being popular throughout all days, which could relate to the regions they are in and their purpose, especially in a town with such a tourist presence as Paris.

Further exploration of the relation of the temporal relation of traffic consumption inside UGS and their overall traffic, in relation to each obtained cluster, can be observed within Figure 5.17c. Here, the y-axis represents the ratio of weekday/weekend traffic, where initial separations between clusters can be seen: all parks within the Lunchbreak cluster are above the 1.2 value (median weekday traffic is 20% higher than the median weekend day), with solely one park of the two other clusters ever being above this threshold. An interesting example is the park with the highest ratio, Park Monceau, which is located in a mixed area with many offices and embassies. Around the value of 1 of the y-axis, there are the parks that have a uniform distribution of traffic through the week, and here there's a quite equal distribution between Residential and Touristic parks. Below 1, there are the parks that are more heavily centered towards weekend traffic consumption, which are major parks in the Residential group.

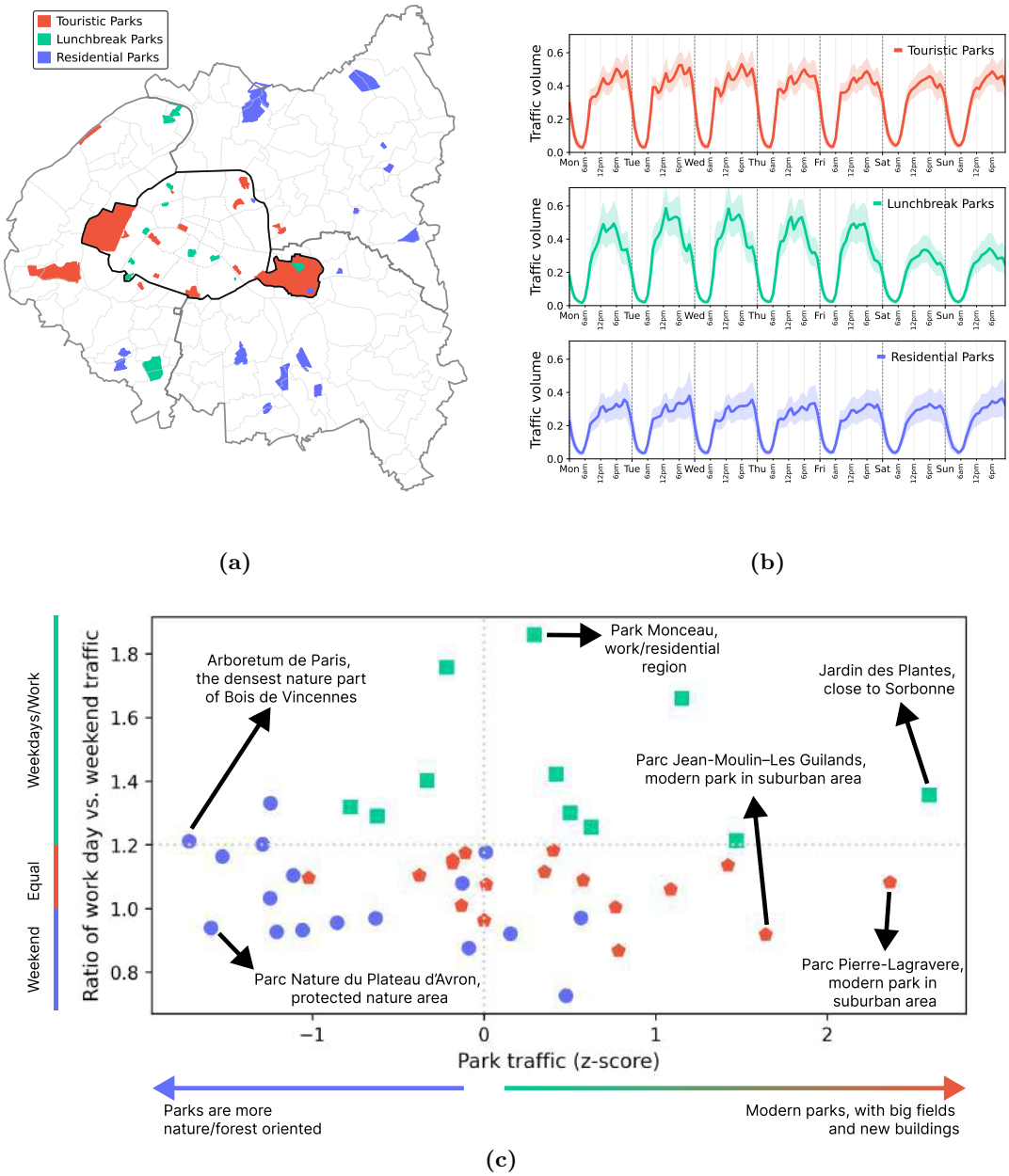


Figure 5.17. a) UGS Clusters over the space of Paris; b) Median week of total traffic per cluster, where the shade represents the 95% confidence interval of the distribution of parks inside each cluster. c) Distribution between the ratio of weekday and weekend traffic, versus the overall traffic of each park. Colors represent the clusters.

Meanwhile, the x-axis of Figure 5.17c represents the z-scored park traffic, where values above 0 mean that a park is above the median traffic of the set (i.e., a z-score value of 1 represents that a park has its traffic 1 standard deviation above the median), while values below 0 represents parks with traffic below the median. This gives an interesting insight into the traffic distribution of the UGSs per cluster: the parks below the median mobile traffic consumption are majorly within the Residential cluster, which also tend to be parks with a denser nature presence (such as the Arboretum de Paris and the Parc Nature du Plateau d'Avron). It could be noted that due to the features those parks have, their intended use drives people to utilize less than their smartphone than the remaining set of parks. On the opposite end of the axis, are the parks with a higher traffic consumption on the set, which is dominated by parks in the Touristic and Lunchbreak parks. The park with the highest traffic consumption is the Jardin des Plantes, which is located in a very central region of Paris very near Sorbonne University. Another interesting region is two other parks with very high traffic consumption amongst the UGSs, Parc Pierre-Lagravere and Parc Jean-Moulin-Les Guillands, both located in the suburban areas of Paris but not in the Residential cluster. Interestingly, both parks are extremely new, with modern buildings and big fields; this is a big contrast to other green spaces in the suburban part of metropolitan Paris.

Those results show that while geographical location plays a big role in how parks influence smartphone and mobile application utilization, there's an underlying factor related as well to the features of the park: parks within a city are not homogeneous and have significantly different intended usages, and this is a major factor driving how users interact with their devices within those spaces.

Key insights. *Parks can be clustered together solely using their preferences in mobile traffic consumption. Those clusters are related to where parks are (Touristic, Work, or Residential areas), which shows a relation between the context where the park is located and how they may affect the usage of smartphones.*

5.2.4.2. Clustered UGS present a difference in how mobile apps are used

The next step will be to explore how the mobile application categories are being used inside each cluster, as well as the relations that traffic being used on a weekday or weekend could have for each of those clusters in relation to app utilization. Following the same methodology of Section 5.2.3.2, the RSCA is calculated for each mobile application category within clusters on Figure 5.18. This result provides more context to the results previously seen in Figure 5.16b. It can be noted that the previously wide CI observed in UGS is broken down into clusters that may have, for the same application category, both an under and an overutilization of the same category. This proves the hinted heterogeneity observed in the full set of parks, and it can definitely be noted that the impact of park types in how applications may be consumed differently in urban green spaces.

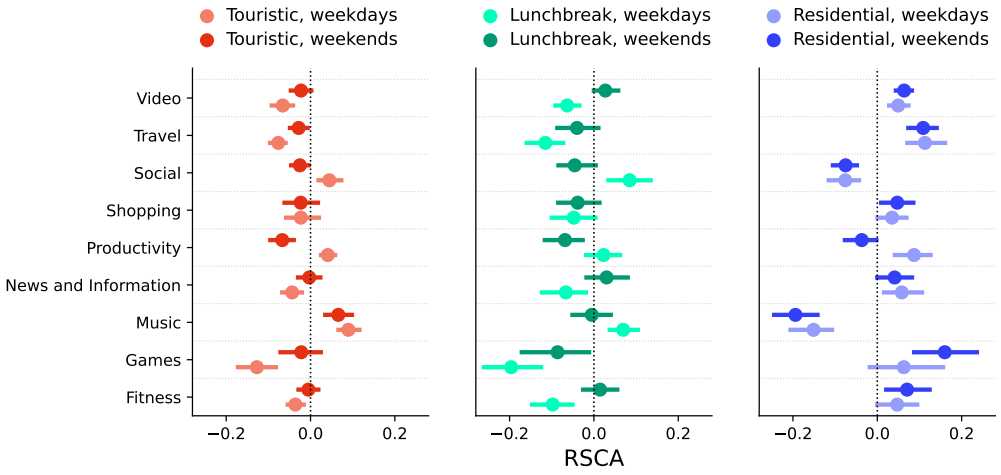


Figure 5.18. RSCA per application, discretizing the traffic on weekdays and weekends. Markers represent the mean values, while lines represent the 95% confidence interval of the values across parks on each cluster.

By analyzing each cluster, it's possible to better understand how different mobile applications are utilized within the categories of UGS. The RSCA values were also split between weekdays and weekends; this is motivated to understand if not only green spaces have different usages across the city, but also if the relation of weekdays and weekends affects the way parks and smartphones interact.

In the Residential cluster, app usage is consistent across weekdays and weekends. The exception is Productivity apps, which are used more on weekdays across all clusters, as expected. Social Media and Music apps are underutilized in Residential parks. This differs slightly from the overall results, where Social Media apps were also underutilized, but Music apps were important in the aggregated set of parks.

In the Lunchbreak category, which includes urban green spaces near workplaces and universities in central Paris, there is significant variation in app usage between weekdays and weekends. This change reflects the temporal dynamics of these areas: weekdays see a work-related population, while weekends see different users, shifting app preferences. All app categories except Shopping exhibit this pattern. For instance, Fitness apps are underutilized on weekdays but see a slight overutilization on weekends, suggesting that these parks become more like Residential parks on weekends. Conversely, Social Media apps are overutilized on weekdays and underutilized on weekends, indicating that workers use these parks to relax while staying connected during the week but engage in different activities on weekends. Additionally, video streaming apps gain importance on weekends, similar to their usage in Residential parks, further highlighting that Lunchbreak parks adopt a more residential-like usage pattern during weekends.

Finally, UGS from the Touristic cluster had a constant influx of traffic throughout the week and were mainly centrally located near famous city landmarks. These UGS

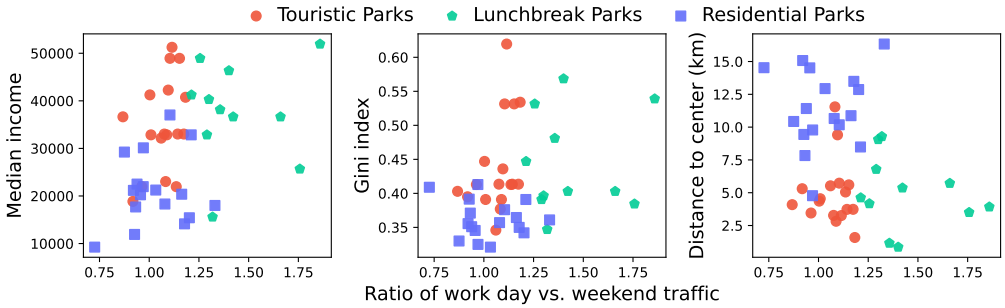


Figure 5.19. Relation between the ratio of weekday/weekend mobile traffic and socioeconomic indicators, with colors representing the obtained park clusters based on smartphone usage. Pearson correlation of the full set between traffic ratio and socioeconomic indicators are [0.4,0.31,-0.32], respectively.

saw heavy social media use during the week, dropping to underutilization on weekends, similar to the Lunchbreak cluster. Other categories showed similar patterns, reflecting the dynamic use of central parks by different populations on weekdays and weekends. Unlike Lunchbreak parks, the Touristic group did not see increased video streaming on weekends, indicating a consistent use pattern due to their central locations and Paris's status as a tourist hub.

Key insights. *Fitness apps are more popular in Residential parks, while social media apps dominate in Touristic and Lunchbreak parks. Social media and productivity apps are less popular on weekends in Touristic and Lunchbreak parks, a variation not seen in Residential parks. This suggests central city parks have multipurpose roles, while residential parks have more singular usage patterns.*

5.2.4.3. Link smartphone usage in parks with socioeconomic aspects

The 3 groups of UGS obtained from the clustering analysis reveal an additional pattern: spaces in each group can be put together due to the type of urban space they're in, in relation to their socioeconomic indicators.

Results are presented in Figure 5.19, representing the relation between the ratio of traffic on weekdays vs. weekends and three indicators about the socioeconomic status of areas where each green space is located: median income, Gini index, and the distance to the center of the city. It's interesting to note that through the 3 presented plots, a clear separation between the clusters is present, even though none of the socioeconomic features was utilized in the clustering process; this could indicate that the usage of smartphones in parks has a linkage to the socioeconomic status of areas.

The discussion begins with the unsurprising result that suburban UGS are the farthest from the city center. These parks have the least mobile traffic on weekdays, with higher usage on weekends. Contrary to expectations, people use their phones more during leisure

hours on weekends in suburban green spaces. The plot points, representing median traffic consumption, confirm this higher weekend usage. There's a slight negative correlation (Pearson $= -0.32$) between distance from the city center and weekend vs. weekday mobile traffic, indicating that the farther a park is from the center, the higher the weekend traffic, while closer parks see more traffic on weekdays.

Next, the relation of income and its inequality to smartphone usage in green spaces is examined. Both median income and Gini index show a positive correlation (Pearson $= 0.4$ and 0.31, respectively) with the ratio of weekday to weekend traffic. Higher-income and inequality areas see higher median traffic on weekdays compared to weekends. Suburban UGS, with the lowest income and greatest wealth equality in Paris, has the highest weekend smartphone traffic. In contrast, Lunchbreak and Touristic parks, located in wealthier central Paris, show higher income and inequality. Touristic parks have a more balanced traffic ratio between weekdays and weekends, while Lunchbreak parks experience higher weekday traffic.

This leads to an explanation of the heterogeneous patterns previously observed across parks, and as expected, the locations where they are within cities will drive different usages of the green spaces. Not all green spaces are equal, and the area they are located will drive different planning of their features, different people utilizing them for different reasons, leading to a vast array of interactions with users' mobile phones.

Key insights. *Residential parks, located farthest from the city center, have lower income and lower inequality, indicating a more homogeneous population and explaining their single-use behavior. In contrast, Touristic and Lunchbreak parks, closer to the city center, exhibit higher income and greater inequality. The diverse user base in these areas likely accounts for the varied smartphone usage patterns observed in these green spaces.*

5.2.5. Main takeaways

This section presents an analysis using mobile traffic data to study the behavior of specific urban locations, focusing on how green spaces relate to smartphone and mobile app usage patterns. The new methodology demonstrates that, despite the limitations of isolating mobile traffic in small city areas, it's possible to identify unique mobile traffic patterns within green spaces that differ significantly from other city areas.

The results indicate that UGSs are not homogeneous: their geographical locations and features are intrinsically related to user-smartphone interactions. This study fills a gap in the literature by providing a large-scale quantification of overall app usage in parks, highlighting the diversity of usage patterns based on park type and temporal variations.

6 Impact of special events on mobile traffic

The study of mobile network measurements can lead to important insights into the spatiotemporal patterns of traffic consumption. As previously observed in Chapter 5, the use of explanatory techniques can help to uncover long and short-term user behaviors. However, not all patterns observed within the network are routine: through the observation of the data, many anomalous events can be noted, some caused by network disturbances, others caused by unprecedented events, resulting in differences in the interaction between users, smartphones, and the mobile network.

With this in mind, this Chapter is interested in studying such anomalous events caused by users of the mobile network, in order to leverage insights about how such events may shift the overall and per-application traffic demands, in both the temporal and spatial dimensions. Such events can be sensed by either an over-utilization of certain applications (e.g., people recording an unexpected public event and posting in their social network) or an under-utilization (e.g., a closed area for construction works and a BS that would usually see many users has an unexpected drop in demand). Those disturbances in the daily routine can be passively captured by the mobile network measurements, allowing researchers to study such anomalies, how they may impact the lives of citizens, as well as how they may impact the functioning of the mobile network.

This chapter will explore the characterization of a group of significant anomalous events through the lenses of the MNO. Section 6.1 will explore the nationwide impact of the COVID-19 pandemic in France through 2020 and 2021 in mobile networks, uncovering how the restriction measures shifted traffic demands. Section 6.2 will extend the COVID-19 characterization, but focus on the impacts within the major urban cities in France, and how those observed shifts can be related to the socioeconomic status within the studied cities. Finally, Section 6.3 will characterize the Pension reform strikes that happened in 2023, focusing on understanding how large manifestations may affect the temporal traffic consumption which applications are mostly affected, and how those insights can help build a framework capable of identifying the protest throughout the city.

6.1. Impact of COVID-19 measures on mobile traffic

The COVID-19 pandemic has affected lives worldwide in a way that is unprecedented in modern times. The response measures that governments have adopted to contain the virus have changed the lifestyles of billions. Under severely restrained mobility regulations, the telecommunication infrastructure has played a key role, allowing people to communicate, work, entertain and even carry out physical activity in the most normal way possible. As proven by early studies, this has determined significant changes in the use of networks [108], [113].

This Section contributes to the body of knowledge about the impact of the COVID-19 emergence on network usage, by focusing on mobile services, or *apps*, and investigating how their consumption has evolved throughout periods characterized by different pandemic containment measures.

To this end, an analysis is made of the demands generated by hundreds of apps in the whole territory of France. The modification of such usages is observed across seven months from 2020 to 2021 and correlates with the heterogeneous restrictions enacted by the local government over that time frame. This study builds on mobile data traffic information collected in the nationwide infrastructure of Orange. The spatial scale and penetration level of the data allow for exploring also the geographical dimension of mobile service usage changes.

The perspectives taken in this work, combining individual mobile services, multiple response measures, and both temporal and spatial dimensions of the phenomenon, have not been explored by previous studies on the impact of COVID-19 on network traffic.

6.1.1. COVID-19 measures in France

During the first two years of the COVID-19 pandemic, France experienced 3 nationwide lockdowns, intertwined by periods with varied responses. A visual example of the evolution of the pandemic[1] is illustrated in Figure 6.1, along with the 7-day moving average of daily cases in France [215]. The first lockdown (March 17 – May 10, 2020) forced the majority of public places, including schools and restaurants, to close, social gatherings to be forbidden, and personal mobility to be limited to essential tasks, leaving only essential services open, with citizens asked to avoid gatherings and reducing as much as possible their mobility, which includes working from home when possible banning all travels except related to professional activities, buying supplies and health or family emergencies. As in many countries in Europe, the subsequent period (May 11, 2020 – October 29, 2020) was characterized by a progressive lift on the restrictions, with a re-opening of public spaces under the requirement to preserve social distancing. The

[1]Further details can be seen on: https://www.gouvernement.fr/info-coronavirus/les-actions-du-gouvernement.

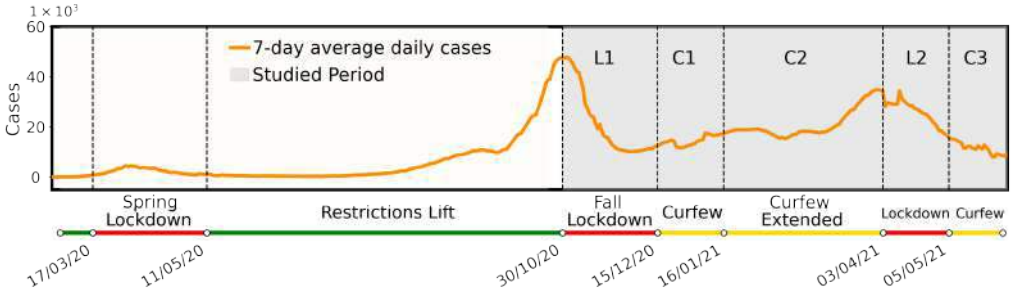


Figure 6.1. Timeline of COVID-19 cases and responses in France.

growth of COVID-19 cases later (October 17 – October 30, 2020) pushed authorities to enforce a 9 PM – 6 AM curfew, first in a few major cities and then in the majority of the country. A second nationwide lockdown followed (October 31 – December 14, 2020), with similar measures to the first one, except for primary and secondary schools staying open. Afterwards, non-essential services started to re-open, and travel restrictions were lifted, but an overnight curfew was maintained between 8 PM and 6 AM (December 15, 2020 – January 15, 2021). The curfew was later (January 16 – April 2, 2021) anticipated to 6 PM. A new increase in infections determined varied local measures, with an announcement by the government in March 31^{st} for the third national lockdown (April 3 – May 4, 2021); non-essential travel was again prohibited, and schools closed, but with lighter measures overall. This period was again followed (May 5 – May 30, 2021) by a progressive lift of restrictions, where nonessential travel was allowed again, with nonessential shops reopening on May 19^{th}, as well as shifting the daily curfew hours to 9pm - 6am. Restaurants were allowed to reopen on June 9^{th}, and the daily curfew was abolished on June 20^{th} of 2021. On a side note, vaccination started in France on December 27, 2020, but the incidence was not substantial during the observed period, with only 16.35% of the population fully vaccinated by May 30, 2021 [216].

This study will cover the period from October 2020 to May 2021, hence encompasses two lockdowns (L1 and L2 in the following), and three curfew periods (C1, C2, and C3). The rationale is that previous studies explored the impact of the first lockdown and subsequent period on mobile traffic; the aim is instead to investigate if and how diverse response measures have affected mobile service usage at later stages of the epidemics.

By adopting such an approach, new insights are provided into the spatial and temporal dynamics of both total traffic and demands for specific services, which stem from the succession of more and less restrictive pandemic response strategies.

6.1.2. The impact of COVID-19 on temporal patterns

In the first part of this analysis, a focus is made in the temporal dynamics of mobile service usage over the whole country of France, looking at both demand volumes and

Figure 6.2. Total traffic volume transiting in the Orange mobile network during the observed seven-month period, as a color scale (top), time series (middle), and linear interpolation over time periods with different responses (bottom).

typical weekly patterns throughout the studied period.

The following notation will be used. Let $d_c^s(t)$ be the demand for service $s \in \mathcal{S}$ recorded in commune $c \in \mathcal{C}$ at time t; $d_c(t) = \sum_s d_c^s(t)$ will be refereed as the total demand generated by all services in c at t, and $d^s(t) = \sum_c d_c^s(t)$ will be the demand for service s at time t over the whole country. Similarly, the global demand at t is $d(t) = \sum_s \sum_c d_c^s(t)$. Six macroscopic time periods will also be defined, each associated with different pandemic containment measures: T_{L1}, T_{C1}, T_{C2}, T_{L2}, T_{C3} denote the time span of the periods in the subscript. Also, T_{21} is the concatenation of all these periods.

6.1.2.1. Changes in mobile traffic volume

To analyze the volume, the temporal granularity will be reduced to days, as this will be an easier resolution to work with a 7-month-long time interval. Therefore, t indicates the day over which the demand is aggregated. Trends in total mobile traffic data will be explored, with a later dive into the dynamics of individual app usage.

Total traffic. The evolution across days for the volume of total traffic recorded between October 2020 and May 2021 is displayed in Figure 6.2. A standard score normalization is used in this and in all subsequent time series of the section to avoid disclosing the actual traffic volume of the operator, which is deemed sensitive information.

Formally, the standard score of daily total traffic in T_{21} is

$$z(t) = \frac{d(t) - \frac{1}{|T_{21}|} \sum_{t \in T_{21}} d(t)}{\frac{1}{|T_{21}|} \sqrt{\sum_{t \in T_{21}} \left(d(t) - \frac{1}{|T_{21}|} \sum_{t \in T_{21}} d(t) \right)^2}}. \tag{6.1}$$

It can be seen how the first lockdown L1 generated a significant drop in overall mobile traffic consumption, while the second lockdown L2 had a softer reduction effect. The linear interpolation at the bottom of the plot helps visualize those trends. After both L1 and L2, mobile traffic consumption recovered fairly slowly, as C1 and C3 are both

characterized by fairly constant loads in time. C2 is the only period where mobile traffic sees a more representative growth, which used to be the norm in pre-pandemic times [217]; the reduced growth at the end of C2 is explained by several densely populated departments of France already started local lockdowns a couple of weeks before the nationwide one.

Key insights. *The strict restrictions during lockdowns reduced the utilization of mobile networks, as was seen for the early stages of the pandemic [113], and this analysis shows that this reduction remained present in the later stages. Surprisingly, milder restriction measures such as curfews did not reduce mobile traffic, but allowed for a slow recovery of consumption levels after the stricter restrictions were over.*

Individual mobile services. The next question to be understood is if the total traffic dynamics was homogeneous across services, or if specific sets of apps behaved differently through the restriction periods. To understand this, the daily traffic of each application will be normalized over the seven months of data[2]

To answer the question, each service is described as its normalized daily time series $z^s(t)$ over the seven-month observation period[3] as:

$$z^s(t) = \frac{d^s(t) - \frac{1}{|T_{21}|}\sum_{t\in T_{21}} d^s(t)}{\frac{1}{|T_{21}|}\sqrt{\sum_{t\in T_{21}}\left(d^s(t) - \frac{1}{|T_{21}|}\sum_{t\in T_{21}} d^s(t)\right)^2}}. \tag{6.2}$$

It's worth mentioning that this normalization makes the temporal volume of different apps comparable, as it removes the bias of volume due to popularity and different demand levels due to content type. This makes it possible to directly compare time series by calculating the Euclidean distance across them, which can be used to construct a pair-wise distance matrix to cluster the behaviors. The methodology chosen here is hierarchical clustering using Ward algorithm [187], with both the Silhouette score [218] and Dunn index [219] being utilized to choose the ideal number of clusters. By observing both stopping criteria, a number of 18 clusters is chosen.

A comprehensive description of the 18 clusters is given in Table 6.1. This reveals the complex dynamic of the temporal evolution of apps under COVID-19 restrictions, which would be overlooked if the analysis solely focused on total traffic dynamics. Many of those clusters present trends akin to the COVID-19 containment measures enforced throughout the period of observation: some, like B, H or N, show steady patterns across all periods; others, like C, are characterized by a growing popularity; and others, like G, suffer from a fairly consistent loss of users. Instead, the interest is on dynamics that can be linked, even if only in certain periods, to response to the restriction measures.

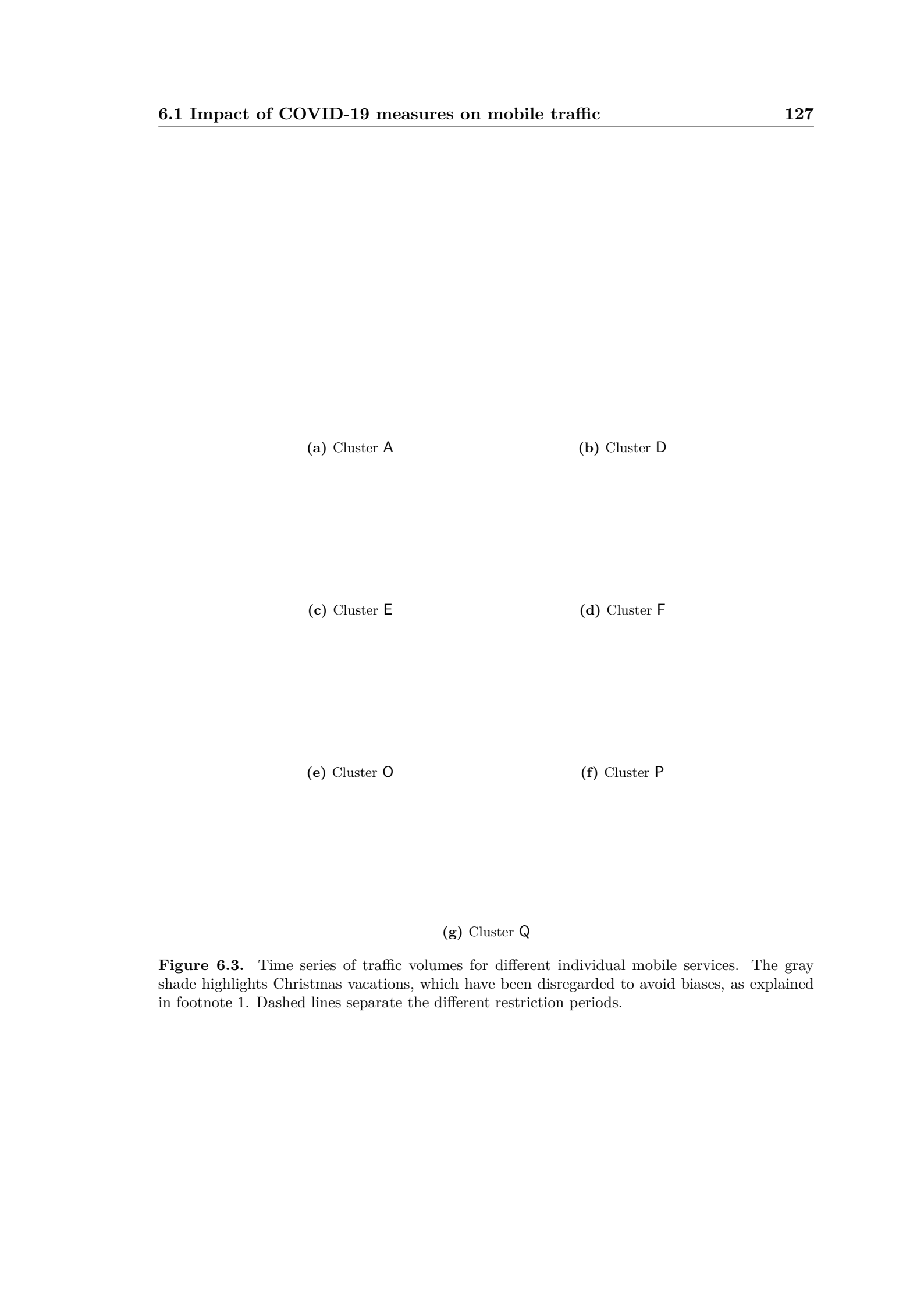
Focusing on clusters clearly affected by the pandemic, A, D, E and F all show a

[2]Filtering out vacation periods, as those have their own disturbances to mobile traffic usage which are not necessarily correlated to the pandemic.

[3]Here, data is filtered out from the data and subsequent analysis of $z^s(t)$ all vacation periods, which, as it will be later seen, can severely affect apps usage. By doing so, it's ensured that the results do not reflect (dis)similarities among services caused by the way they are consumed during holidays.

Cluster	Description	Samples
A	No weekly pattern, increasing in L1, C2 or L2; gaming and messaging apps mainly	WhatsApp, League of Legends
B	Higher usage in weekends, steady over time; video streaming and gaming apps mainly	Netflix, Youtube, Steam, PUGB
C	Higher usage in weekends, increasing over time; gaming apps mainly	MineCraft, Fornite
D	Higher usage in weekends, increasing around L1 and L2; video streaming and gaming apps mainly	Disney+, Apple Video, CounterStrike
E	No weekly pattern, increasing in proximity of L1 and L2	Pinterest
F	No weekly pattern, increasing in L1; gaming and conferencing apps mainly	Zoom, Clash of Kings, Angrywords
G	No weekly pattern, decreasing over time	Battle.net, Shazam
H	Slightly higher usage in working hours, steady over time	N26, Dropbox
I	Higher usage in working hours, decreasing in time; business apps mainly	Evernote, Twitter, Microsoft Office
J	No weekly pattern, increasing in C3	Prime Video, WeChat
K	No weekly pattern, noisy over time; OS update services mainly	MS Windows Update
L	Slightly higher usage in weekends, noisy in time; gaming apps mainly	World of Warcraft, Playstation
M	No weekly pattern, slightly increasing in C2	Psiphon, Coyote
N	Higher usage in working hours, steady over time; office applications mainly	Gmail, Skype, Google Docs
O	No weekly pattern, increasing in C2 and C3	Telegram, TikTok, Uber
P	No weekly pattern, increasing in C2 and substantially more in C3; location-based services mainly	TripAdvisor, Foursquare, Spotify
Q	Higher usage in working hours, increasing in C2 and C3	Google Maps, Waze, AirFrance
R	No weekly pattern, increasing over time	Twitch, Google Meet

Table 6.1. List of 18 clusters issued by the clustering algorithm.



(a) Cluster A

(b) Cluster D

(c) Cluster E

(d) Cluster F

(e) Cluster O

(f) Cluster P

(g) Cluster Q

Figure 6.3. Time series of traffic volumes for different individual mobile services. The gray shade highlights Christmas vacations, which have been disregarded to avoid biases, as explained in footnote 1. Dashed lines separate the different restriction periods.

significant increase in traffic volumes during the lockdowns. Sample time series of specific services in those clusters are in Figure 6.3. Although there are discrepancies that make the Ward algorithm cluster the apps separately, it's observed in all cases sustained higher traffic during L1 (in E, such as Pinterest in Figure 6.3c, and in F, such as Xiaomi Mi Home in Figure 6.3d), during L2 (in A, such as Houseparty in Figure 6.3a), or during both L1 and L2 (in D, such as Disney+ in Figure 6.3b). The reason for such dynamics is simple: videoconferencing, on-demand television, social media, or smart-home managers are all examples of mobile services that are more frequently consumed at home, where people spend a much larger portion of time during lockdowns. It's important to recall that this analysis looks at traffic generated by mobile devices connected to the cellular network, and not, e.g., to home Wi-Fi hubs. Therefore, the services in the clusters above all demonstrate that a non-negligible portion of the Orange user population is actually using RATs as a way to access the Internet from home.

Another notable behavior induced by COVID-19 is that of clusters that show a dual behavior to the one above. Namely, O, P, and Q are curbed by lockdowns, and record an increased usage during the relaxed measures in curfews. A closer look reveals that the vast majority of these services are directly or indirectly related to personal mobility: the limitations to movements determined by L1 and L2 clearly reduce their utility. Interestingly, the dynamics are slightly different in apps for general mobility and for more leisure-oriented mobility. The first case includes services with a marked working-hour pattern (in Q, such as Google Maps or Waze in Figure 6.3g) and with more regular usage (in O, such as Uber or Apple Maps in Figure 6.3e), and show moderate increase in usage in both C2 and C3. Instead, apps targeting mobility during free time (in P, such as Foursquare or TripAdvisor in Figure 6.3f) show a dramatic increase in usage in C3, which can be attributed to the combination of more relaxed measures and inviting weather conditions during that period.

To conclude the volume part of the analysis, it can be underscored how this approach of considering individual apps is critical to reveal the richness of behaviors above. For instance, Figure 6.4 shows the traffic time series of a number of popular video streaming services. These mobile services undergo very heterogeneous evolutions in the observed seven months, with volumes that are steady (e.g., YouTube, Netflix) and possibly higher during working hours (e.g., Skype), declining (e.g., Zoom), or heavily dependent on pandemic response measures (e.g., Amazon Prime Video). In fact, these services are classified in *different* clusters during the analysis. Had video streaming been treated as a single category, all this diversity would have been lost.

Key insights. *Individual mobile services' responses to COVID-19 restrictions form a complex ecosystem of usage patterns. Simply analyzing total traffic volumes or grouping services hides this complexity, potentially yielding inaccurate conclusions. These results uncover diverse consumption patterns among individual services in response to*

(a) Cluster B **(b)** Cluster F

(c) Cluster J **(d)** Cluster N

Figure 6.4. Time series of traffic volumes for different individual mobile services in the video streaming category, by micro-cluster.

restrictions. While some remain unaffected, others experience significant fluctuations depending on containment measures. Certain service types exhibit clear correlations between their nature and their reaction to restrictions.

6.1.2.2. Changes in temporal consumption patterns

An interesting question is whether COVID-19 measures not only impacted the volumes of traffic transiting in the mobile network but also the temporal distribution. For instance, previous works have shown that the typical difference between the hourly traffic pattern in working days and weekends tends to disappear during a lockdown [107]. Here, hourly traffic will be looked, i.e., the time index t denotes one specific hour, and weekly patterns will be examined in both total and per-app traffic. Also, equivalent weekly patterns computed from a three-month control period in 2019 will be used as a reference, so as to understand if and how the daily activity has changed due to COVID-19 responses. T_{19} will be denoted as the set of hours t in such a control period.

The focus will be first put on typical weekly dynamics that are known to capture most of the variance in the telecommunication activity of individuals [220], [221], and condense the seven-month traffic dynamics into a *median week signature* [23]. Formally, the median traffic in each hour of the week is computed as $w(t) = \mu_{0.5}\{d(t)|t \in \mathcal{M}(t)\}$, where $\mu_{0.5}$ denotes the median of the argument set, and $\mathcal{M}(t)$ is the set of same hours of the week as t (e.g., Mondays at 8 AM). Then, the median week signature is obtained by applying the standard score normalization in (6.1) to $w(t)$ instead of $d(t)$. It's important to note that disjoint $\mathcal{M}(t)$ is used for T_{21} and T_{19}, so as to obtain independent median weeks during and prior to the COVID-19 pandemic.

Figure 6.5. Total traffic median week in target and control periods.

Figure 6.6. Distances between the median week signatures of individual apps (columns), comparing pre-pandemic with COVID-19 periods (top rows), and different periods in 2020-21 characterized by varied response measures (bottom rows).

Figure 6.5 superposes the median week signatures for the considered COVID-19 response period and in the 2019 control period. While peak traffic hours stayed the same, minor changes can be accredited to the enacted restrictions. First, remote working sensibly reduced the need for daily commuting, which explains the disappearance of the early-morning traffic peak in 2020-21. Second, evening peaks during the pandemic are relatively higher than in 2019, which is likely caused by mobility limitations that forced people at home from late afternoon onwards, consistently through the observation period. Third, it can be confirmed that the reduced diversity between working and weekend days, which tend to be closer during COVID-19 than they were before.

The analysis is repeated on a per-app basis. First, it's computed $w^s(t) = \mu_{0.5}\{d^s(t)|t \in \mathcal{M}(t)\}$, for each mobile service s, and derived the app-specific median week signature applying in (6.2) to $w^s(t)$ instead of $d(t)$. In this case, separate median weeks can also be produced for T_{L1}, T_{C1}, T_{C2}, T_{L2}, and T_{C3}, so as to assess the impact of restrictions on the weekly patterns of app usage.

As median week signatures are also standardized, they can be directly compared. For each service independently it's computed the dynamic time warping between its signatures in different periods, i.e., L1, C1, C2, L2, C3, and in the 2019 control period. Figure 6.6 shows the result for all mobile services, along columns; the first five rows show the distances of the 2019 median week and those in L1, C1, C2, L2, C3, respectively. The following rows report the distances between different periods in the epidemics, i.e., L1–C1, L1–C2, L1–L2, L1–C3, C1–C2, C1–L2, C1–C3, C2–L2, C2–C3, and L2–C3.

(a) Skype **(b)** YouTube

(c) Netflix **(d)** Uber

Figure 6.7. Median week signatures of representative mobile services in the 2019 control period and in the 2020-21 target period.

Most apps do not show any significant change in their weekly pattern (i.e., have near-zero or negative distances in all rows), hence the way they are consumed is hardly affected by the pandemic. Among the mobile services that show some diversity (i.e., have positive distances), those on the left (group 1 in the plot) are less popular apps with inherently bursty dynamics that tend to vary all the time, even within weeks of 2019. More interesting is the group of services that show a clear distance between the control and studied period, but no differences in periods within the COVID-19 pandemic (group 2 in the plot). The weekly usage pattern of these apps clearly reacted –in a *uniform* way– to the restrictions.

The median weeks of representative mobile services in this group are illustrated in Figure 6.7. A video calling application such as Skype adjusted to a strongly work-oriented activity pattern, with high peaks in the morning and early afternoon, which also overflew to weekends. Major video streaming services are also in the group of interest. Both YouTube and Netflix saw their early morning peaks disappear, along with home-work commuters who created such demands; given the large volume of traffic of these apps, this also determines the same effect observed in the total traffic in Figure 6.5. In addition, the incidence of evening traffic grew dramatically during weekends for Netflix, due to the impossibility for people to enjoy nights out as in pre-pandemic times. Finally, with COVID-19, private cabs were hired with Uber in a new pattern, with reduced hours of operation in working days (owing to no commuting and early curfews), and no evening peaks on Fridays and Saturdays (due to almost absent nightlife).

Key insights. *Weekly patterns in the total mobile data traffic show changes during COVID-19, which are however mainly caused by a small subset of popular video streaming apps. In fact, when the vast majority of mobile services are consumed does not change in a significant way during the pandemic.*

6.1.3. The impact of COVID-19 on spatial patterns

The next step will look whether the temporal changes above occur homogeneously over the French territory, or are the result of geographically diverse effects of the epidemic responses. Disaggregated data is leveraged at the commune level to this end.

Total traffic. This analysis starts by considering the total mobile data traffic and computing the average traffic density in each commune during the 2019 control period T_{19} and in the target 2020-21 period. Formally, $\bar{d}_c(T_{19}) = (1/T_{19}) \cdot \sum_{t \in T_{19}} d_c(t)/a_c$, and $\bar{d}_c(T_{21}) = (1/|T_{21}) \cdot \sum_{t \in T_{21}} d_c(t)/a_c$, where a_c is the area of commune c. Since the interest lies in understanding if the *relative* geographical distribution of traffic has changed due to COVID-19, the traffic density is standardized over space in each period, as

$$z_c(T_{19}) = \frac{\bar{d}_c(T_{19}) - \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \bar{d}_c(T_{19})}{\frac{1}{|\mathcal{C}|} \sqrt{\sum_{c \in \mathcal{C}} \left(\bar{d}_c(T_{19}) - \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \bar{d}_c(T_{19}) \right)^2}}, \tag{6.3}$$

$$z_c(T_{21}) = \frac{\bar{d}_c(T_{21}) - \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \bar{d}_c(T_{21})}{\frac{1}{|\mathcal{C}|} \sqrt{\sum_{c \in \mathcal{C}} \left(\bar{d}_c(T_{21}) - \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \bar{d}_c(T_{21}) \right)^2}}. \tag{6.4}$$

In practice, z_c is a measure of how the traffic density of commune c compares to that of the average French commune.

The difference in the standardized traffic density between the 2019 control period and the observed pandemic time span in 2020-21 is illustrated in Figure 6.8. The manifest trend is a significant reduction of the relative importance of cities as the places where the overall mobile traffic demands are generated. Negative differences, in dark blue, pinpoint all large- and medium-sized urban areas in the country. Zoomed views are provided in the bottom part of the figure for the 10 most populated cities: they show even better how the phenomenon is strongly localized in urban centers, whereas the surrounding suburban areas possibly experience a positive difference, i.e., increased contribution to the overall traffic. Indeed, the higher incidence of countryside regions in also visible at a nationwide scale, and is especially strong at locations well known to attract metropolitan inhabitants during vacation periods[4].

The result very neatly demonstrates how COVID-19 measures not only forced inhabitants of major cities at home, so that they increasing relied on Wi-Fi access, and

[4]It's important to recall that vacation days were filtered out from the data, whose impact would be anyway diluted in seven months: the effect cannot be ascribed to holiday mobility.

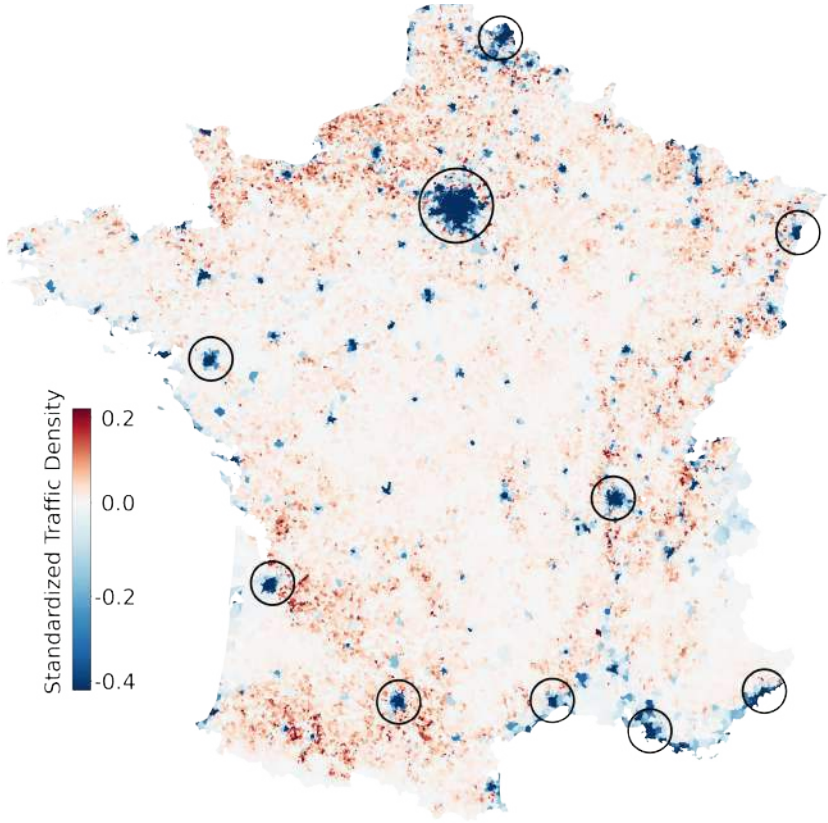
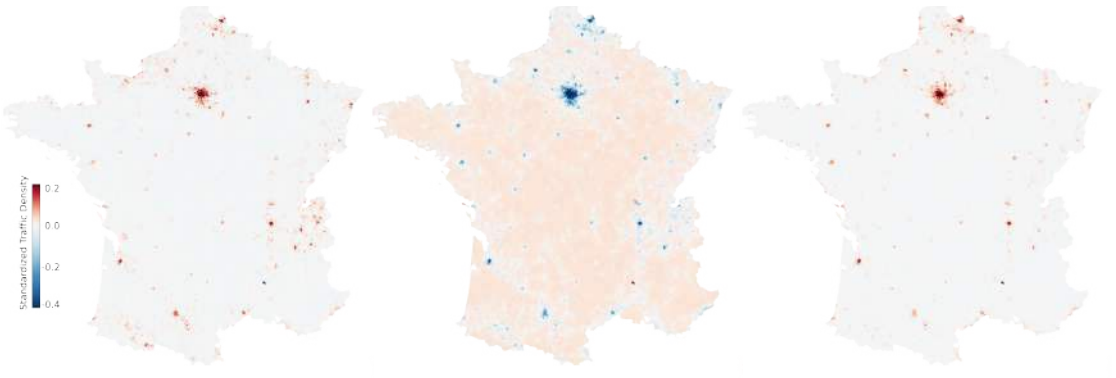


Figure 6.8. Difference in the standardized geographical distribution of traffic density between 2019 and 2020-21. Circles highlight the 10 most populated departments in France, for which detailed views are in the bottom part of the figure.



(a) L1–C2 **(b)** C2–L2 **(c)** L2–C3

Figure 6.9. Difference in the standardized geographical distribution of traffic density between periods during the pandemic.

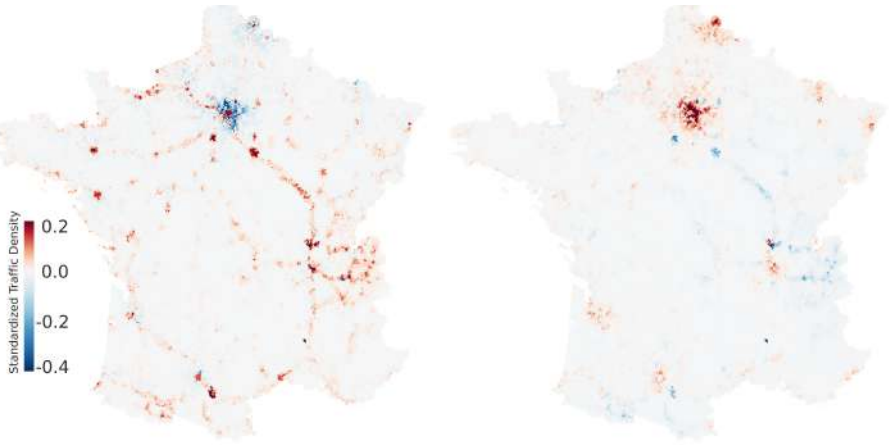
cut down their usual cellular network traffic; rather, it also pushed many people away from city centers, and towards second/vacation homes, or greener places. As most people kept working remotely, and relied on cellular access for Internet connectivity, such a mobility caused the de-urbanization of mobile traffic consumption in Figure 6.8. In fact, this phenomenon can be even broken down in time, across different periods during the epidemics. Figure 6.9 shows a similar difference map, but computed for the L1–C2, C2–L2, and L2–C3 period pairs. A striking effect emerge such that entering lockdowns, like C2–L2, determine the effect described above, whereas transitions into more relaxed measures, like L1–C2 and L2–C3 result in a return of traffic towards cities.

Key insights. *Restrictive COVID-19 measures exert a very consistent reduction of the contribution of all urban centers to the overall mobile traffic demand. Such de-urbanization of traffic is promptly reverted once restrictions are loosened.*

Individual mobile services. The dataset used for this study allows the exploration of if and how the spatial dynamics above are altered when considering independent mobile services, instead of their aggregated traffic. To this end, standardized traffic densities $z_c^s(T_\star)$ are computed on a per-app basis and for $\star \in \{\text{L1}, \text{C1}, \text{C2}, \text{L2}, \text{C3}\}$, by (i) computing the average traffic density of each service in every commune as $\bar{d}_c^s(T_\star) = (1/T_\star) \cdot \sum_{t \in T_\star} d_c^s(t)/a_c$, and (ii) applying similar equations to (6.4) where it's used $\bar{d}_c^s(T_\star)$ instead of $\bar{d}_c(T_{21})$. This allows producing difference maps like those in Figure 6.9, for each mobile service separately.

In order to discover macroscopic patterns, each combination of periods (e.g., C2–L2) is studied independently. For each case, the maps above are summarized as the probability distributions of the per-commune differences that compose them; then, the pairwise similarity is computed between the distributions of all services, using the Jensen-Shannon distance. This results in distance matrices like those depicted in Figure 6.10. While those are just two samples, all have a similar structure, highlighting how most

Figure 6.10. Sample matrices of pairwise distances between difference maps of each app. Left: L1–C2. Right: C2–L2.

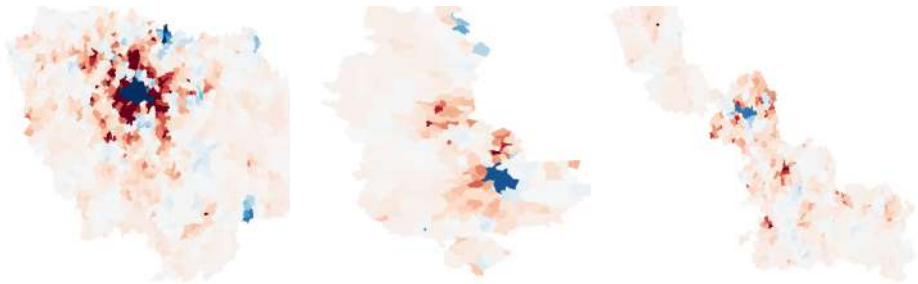


(a) L1–C2 **(b)** C2–L2

Figure 6.11. Difference in the standardized geographical distribution of Waze traffic density between example periods.

services have low distance among them, and spatial dynamics that are aligned to those observed for the total traffic. However, there exist apps for which the alternation of lockdowns and curfews entails fairly unique geographical variations –pointed by high/red values in the matrices.

Next, a few services with interesting spatial patterns of mobile traffic consumption due to COVID-19 restrictions will be explored. The first one is Waze, in Figure 6.11. Here, changes in the geography of the service traffic are bonded to the road infrastructure: when exiting from the lockdown in L1, an increased incidence of national transportation arteries is observed; this vanishes when entering again in the lockdown during L2, when instead it can be noted a higher activity around larger cities. This type of behavior is consistent with the fact that long-distance travel is cut out during lockdowns [114]; it also corroborates the previous reasoning on the causes of the de-urbanization of mobile traffic during lockdowns, which push people to move towards the close proximity of their



(a) Paris **(b)** Lyon **(c)** Lille

Figure 6.12. Difference in the standardized geographical distribution of TripAdvisor traffic density from C2 to L2, for 3 cities.

cities of residence –and use apps like Waze to find their way there. The second example is TripAdvisor. Here, *residential* urban areas located around the city centers consistently gain importance as mobile traffic sources when COVID-19 measures are tightened; city centers themselves have the opposite trend. This is shown in Figure 6.12 for three sample cities entering the L2 lockdown. It can be argued that these peculiar spatial dynamics can be the results of lockdowns keeping workers far from downtown offices, and forcing them to order delivery food from their homes.

Key insights. *While the majority of mobile services follow global spatial trends determined by COVID-19 control measures, some apps are affected in unique ways by the transitions between restrictions. As this is due to the distinctive nature of such apps, an in-depth analysis of their geographies across responses may have interesting applications beyond networking, e.g., to study how mobility, working, shopping, or eating habits evolve in different areas as a result of the pandemic.*

6.1.4. Main Takeaways

A first investigation was conducted on the impact of late COVID-19 response measures on the usage of individual mobile services, targeting a major European country, i.e., France, and a seven-month observation period that encompasses multiple lockdowns and curfews. The presented results reveal a number of previously unknown behaviors in both total traffic and specific service demands that result from the alternation of more or less restrictive control strategies enacted by the French government.

By doing so, new substantial knowledge is contributed to the literature on the effects of COVID-19 on telecommunication systems, by demonstrating the existence of unique patterns in the reaction to COVID-19 measures for specific apps with a quite narrow scope (e.g., targeting leisure mobility, e-commerce, or work activities) may pave the way to the use of mobile service consumption as a data source for research in other domains. Indeed, such data could offer unique insights to understand the behavior of individuals

in the presence of pandemic containment measures, across social dimensions like personal movements, shopping habits, or remote work schedules.

The key takeaways of this Section were:

- An extremely rich and varied range of reactions is unveiled, showing that individual mobile services have in front of restrictions of different severity; such diversity was hidden in previous analyses that have focused on traffic aggregates or service categories – and it's shown that apps on the same categories can have very heterogeneous behaviors.

- Some services are more affected than others by later-stage control measures in both time and space patterns, and detail representative cases.

- It revealed that the pre-COVID-19 weekly activity peaks of the total mobile data traffic have changed in a non-negligible manner one year into the pandemic, due to shifts in the consumption of a few high-volume services.

- It exposed the de-urbanization of mobile service usage during lockdowns, and showed how less restrictive measures such as curfews reverse that phenomenon.

6.2. Modeling the impact of COVID-19 on mobile networks

As was discussed in Section 6.1, prior studies on digital usages focus on the earlier stages of the pandemic, i.e., the so-called first wave and its consequent lockdowns, and investigate its effects at coarse spatial granularity (e.g., aggregating data at the level of countries or regions) and without separating individual digital applications (e.g., video streaming versus social media). As it was seen in Subsections 6.1.2 and 6.1.3, the dynamical changes of mobile app consumption are too heterogeneous to be aggregated inside major categories, ideally requiring specific per-service analysis. Following the results observed in a nationwide setting during Subsection 6.1.3, a question emerges on whether the dynamics inside cities are homogeneous or not, i.e., how mobile traffic consumption changes within major urban cities during the COVID-19 pandemic.

In this Section, the main goal is to close the above gap and explore the impact of policies enacted during the later stages of the COVID-19 pandemic (i.e., waves following the first one), across fine-grained neighborhoods in the largest cities of France at the level of single mobile services. This approach yields a high-resolution analysis of the consequences of governmental policies for COVID-19 containment on digital service consumption from mobile devices and carries substantial explanatory potential via the socioeconomic indicators typically associated with such geographical zoning.

6.2.1. Data preparation and model selection

This next section will go over the extensive preparation needed for the data in order to explore the inner city patterns of mobile traffic changes due to COVID restrictions, as well as the model selection process.

The timeline of the COVID-19 pandemic in France, together with government measures, was already discussed in Subsection 6.1.1. This study follows the same timeline, but with a focus on the impact within cities. A visualization of the timeline and how it fit into the data processing can be observed on Figure 6.13A.

6.2.1.1. Measuring changes in mobile service consumption

To explore the effects on mobile traffic consumption due to different restriction measures applied by the French government, this study explores three key transitions $T_\star$ of mobility restrictions at the later stages of the COVID-19 pandemic. Specifically, $\star \in \{L, D1, D2\}$, denoting the change of measures when (i) entering the third lockdown, (ii) leaving the third lockdown and transitioning into the following curfew, and (iii) further lifting of restrictions into final light curfew, respectively, as seen in Figure 6.13B. For every transition $T_\star$, traffic snapshots are obtained, representing the mobile demands before and after the transitions. To this end, the mobile service demands is aggregated in the period of 14 days immediately preceding the transition, as well as in the 14 days after the transitional moment has happened, as exemplified in Figure 6.13B.

The median traffic of each IRIS is obtained over the periods before $T_\star - 14$ and after $T_\star + 14$, followed by dividing the median traffic of each IRIS by its area in km^2 to remove the bias in traffic volume at large areas. Next, a standardization is applied on the before/after periods of each city by calculating the z-score of the traffic density, subtracting the value of each IRIS by the mean and dividing by the standard deviation of the selected set of IRIS of each city and period, so as to ensure that comparisons among different time periods are fair and not biased by confounding factors (e.g., changing traffic volumes). This results in values that highlight the importance of each IRIS for the traffic generated at the selected area and period. In addition, standardization allows for anonymizing the raw traffic values, which cannot be shared due to contractual obligations with the network operator that provided the data for this research. Finally, a traffic difference map is calculated, representing $T_\star$ by subtracting $T_\star + 14$ and $T_\star - 14$ and highlighting the effects caused by restrictions imposed on cities at the IRIS-level; as an example, Figure 6.13C shows the resulting traffic difference maps for Paris.

The analysis of transitions focuses on the evenings hours of workdays (Monday through Friday, from 7pm to 1am). The rationale is that such periods best capture differences induced by COVID-19 restrictions, for several reasons: (i) these hours represent a moment of the day when people conclude their workday and commute, are at their home or visit

leisure locations (if allowed by the restrictions); (*ii*) the time interval matches the peak time of mobile traffic consumption in France during the pandemic [7]; and, (*iii*) these hours help relate changes of traffic with geo-referenced socio-economic information that is related to the place of residency as well as with data about leisure areas in each city. This study considers the 10 cities with highest population in France: Paris, Marseille, Lyon, Toulouse, Nice, Nantes, Montpellier, Strasbourg, Bordeaux and Lille; in each city, mobile service demands are aggregated at the level of IRIS zones.

6.2.1.2. Handling spatial noise in traffic measurements

By observing the spatial distribution of traffic differences for each IRIS, the presence of spatial noise in the collected data was noted. This leads to spatial cluster patterns becoming less evident and an observed drop in model fitting performance. To avoid this, a spatial lag will be applied on the mobile traffic data [222], by calculating the neighborhood graph between IRIS using Rook's distance, and the resulting smoothed traffic difference will be the average between the value of the selected IRIS and its set of neighbors. This process will help highlight spatial patterns, as well as perform a filter on outliers (e.g., in a neighborhood all IRIS see an increase in traffic and there's a single one with a sharp decrease and no apparent reason for this behavior).

The collected data contains over 250 services with different degrees of traffic consumed. Especially when analyzing at IRIS level, noise can be significantly present in services that are less popular even with the aforementioned spatial lag, so not all services can be studied on such a small level of geographical granularity. To overcome this, all services are ranked by the total traffic generated. By observing their spatial distributions of traffic consumption, additional applications are removed due to noisy patterns at IRIS level.

6.2.1.3. Selection of socioeconomic features

To correlate changes of traffic at IRIS level, a set of socioeconomic features collected at the same granularity is used, as seen in Figure 6.13E. Those will be collected from three sources provided by INSEE. The first base comes from the 2017 Census in France [223], which is the most up-to-date Census that includes collection at IRIS level. From the complete set, the features chosen are the population and its density by diving the values by the area in km^2 of each IRIS; this results in a better geographical distribution of values, and also removes the bias that large areas may induce on the features.

The second base is related to revenue; as this was not covered in the 2017 Census a 2018 survey will be utilized, which studied revenue, poverty, and quality of life [224] to obtain the median income in Euros per IRIS.

Finally, the third set used is the SIRENE base of enterprises and establishments in France [225], which indicates the date of opening, if the place was still active at the

time of collection (or if not, the date of closure), as well as the location. From this base, the extracted features are the number of restaurants, bars, and nonessential stores (i.e. shopping centers, tech stores), and calculated as the total number per IRIS of those establishments and later divided by the IRIS area to obtain the density of leisure locations.

6.2.1.4. Modeling changes in traffic consumption concerning socioeconomic features

A single model will be used for each transitional period $T_\star$ in order to have the selected set of socioeconomic features predicting the mobile traffic differences due to mobility restriction for all cities (which is performed first for total traffic and later for each mobile service's traffic). It was initially evaluated using a simple linear regression model but quickly noted that R2 and Pearson correlation values were significantly low, which could be linked to spatial auto-correlation error on the residuals previously mentioned. It was then concluded the necessity of a linear model that takes into account the spatial characteristics of the data and selected a Spatial Lag Model (SLM) [226], which is a linear model and can be solved with traditional least squares methods and uses the spatial lagged version of the dependent variable as a regressor. It is described as:

$$y = \lambda W y + \beta_1 X + \epsilon \tag{6.5}$$

where y is the dependent variable, X is the matrix of features, Wy will be the spatial lagged feature of the dependent variable y and ϵ $N(0, \sigma^2)$ will be a Gaussian noise term. The goal is to estimate $[\lambda, \beta]$. In addition, two additional problems in the data have to be dealt with: heteroscedasticity and outliers. Both can be tackled together by treating the SLM as a robust regressor and solving it with an Iteratively Reweighted Least Squares (IRLS) [227], [228]. Those significantly reduced the spatial dependency of the model. Also, as the analysis is focused on coefficient values and not on the predictability power of the model, there's no necessity to split the set in train and validation.

6.2.2. Changes in total traffic at city level during the pandemic

Figure 6.13C shows maps of the changes in the overall mobile service demand in Paris during the transitional periods T_L, T_{D1}, T_{D2}, in terms of the increase or decrease of mobile traffic consumption generated by the transition with respect to the previous period. An evident heterogeneity of behaviors is noted across space (i.e., IRIS zones in a same city) and time (i.e., transitions T_L, T_{D1} and T_{D2}), which suggest diverse responses of the urban tissue to the COVID-19 restrictions in terms of digital usages. Such a response is in fact so different that, at each transition, both growth and reduction of the relative mobile traffic demands are concurrently observed in a same city depending on the neighborhood considered. This divergence was impossible to observe previously in

the nationwide analysis of Section 6.1, due to the granularity being city-level, i.e., only the median behavior of cities was observed, with the inner behaviors hidden due to it. The phenomenon also exhibits clear spatial dependencies, as areas experiencing similar increase or decrease of mobile digital consumption tend to be geographically clustered for all studied transitions.

For instance, in Paris, neighborhoods in the center and to the West of the urban region experience substantial reduction of mobile traffic activity induced by the lockdown in T_L, whereas areas to the East and South are those characterized by the highest growth of consumption during that same transition. West Paris recovers the lost digital service usage at the end of the lockdown in T_{D1}, whereas the city center does so only with when additional restrictions are lifted in T_{D2}. During both transitions to curfews T_{D1} and T_{D2}, the East and South regions display instead a progressive reduction of the demand for mobile services. As in the example of Paris above, all cities examined show cyclic patterns where neighborhoods are characterized by either (i) a reduction of demand during the lockdown, followed by an increased consumption in the subsequent curfews, or (ii) the opposite sequence of growth in lockdown and decrease with curfews.

Key insights. *Major urban cities in France present a heterogeneity of responses for the changes of mobile traffic consumption due to restriction periods, which can be linked to entire neighborhoods. It's noted that the changes due to lockdown are slowly reversed as soon as restrictions are lifted into the first curfew, and changes are almost fully reverted by the time of the looser second curfew.*

6.2.3. Socioeconomic explanation to mobile traffic usage changes

The clustered patterns of IRIS zones that display different responses to the same governmental restrictions suggest a connection to the similarly clumped socioeconomic characteristics of urban areas. To explore this relation, a spatial lag model is proposed, as described in Section 6.2.1.4, that uses socioeconomic indicators as regressors to predict mobile traffic variations in all of the 10 large cities considered in the study. All features input to the model are standardized by subtracting their mean and dividing by their standard deviation. The considered features are the population density, median income and leisure space presence (i.e.,, the spatial density of restaurants and non essential stores), whose geographical distribution is depicted in Figure 6.13E for Paris. These socioeconomic features are loosely correlated among them[5] and have low individual correlation with the mobile demand differences across transitions T_L, T_{D1} and T_{D2}[6].

An SLM fed by the three socioeconomic features above can explain the changes in mobile service demands induced by COVID-19 regulations on each IRIS of the 10 cities with good accuracy. As shown in Figure 6.14A, Pearson's correlation coefficients range

[5]Pearson's correlation coefficients of 0.21–0.49.

[6]Absolute value of Pearson's correlation coefficients below 0.2.

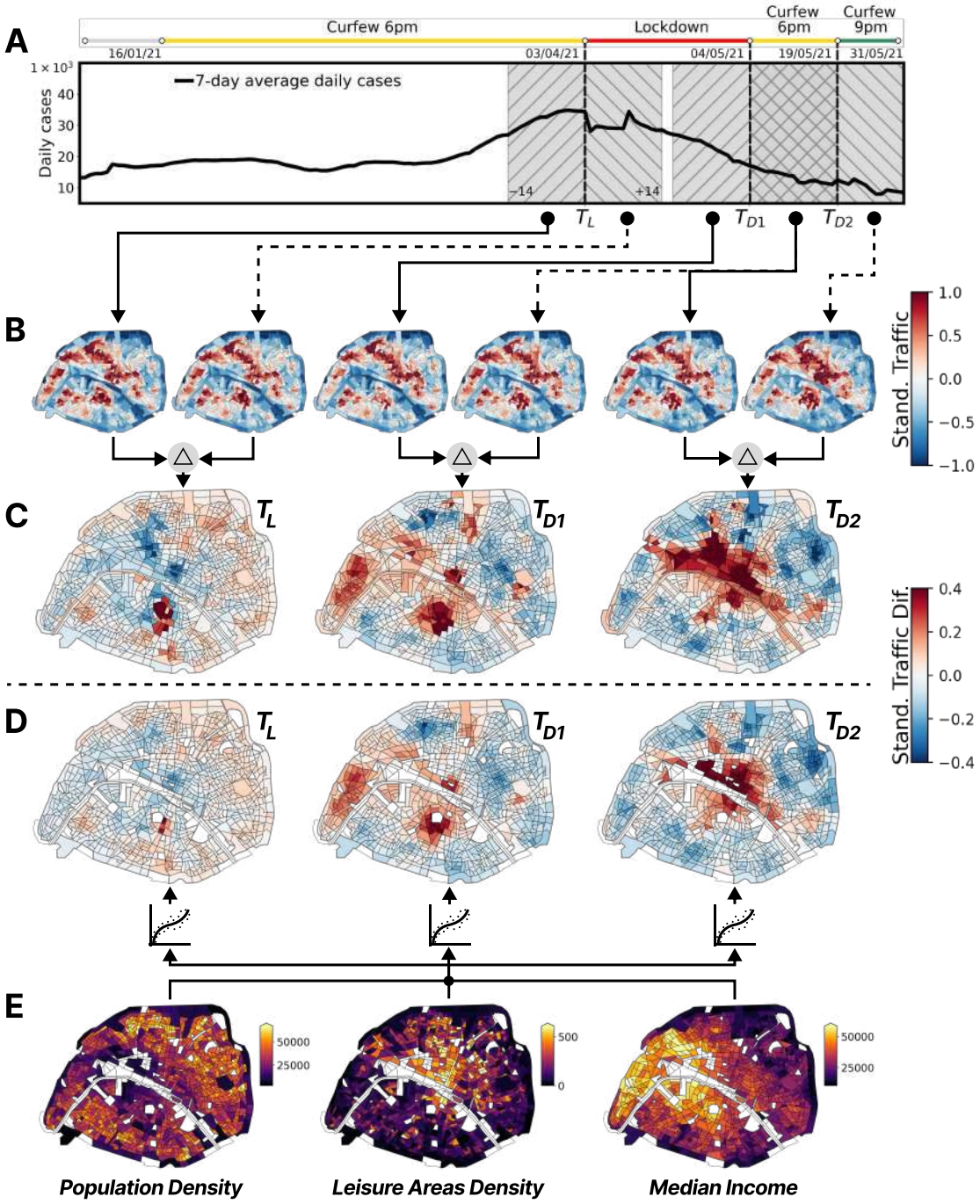


Figure 6.13. Arrangement of data. A) Timeline of COVID-19 cases in France, as well as the mobility restriction periods and studies periods; B) Mobile traffic per IRIS across each studied period; C) Traffic difference between periods; D) Predicted values by the proposed SLM model; E) Spatial distribution of the features utilized by the model for the mobile traffic difference prediction.

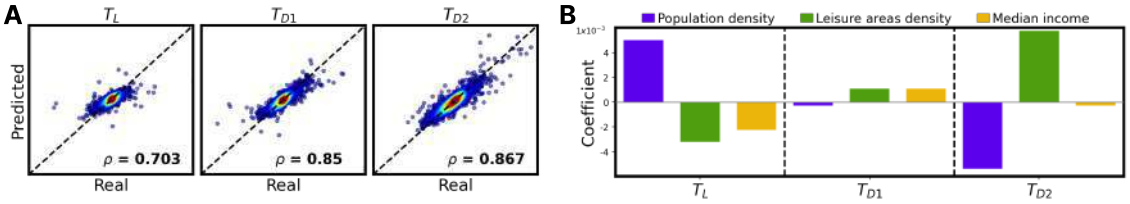


Figure 6.14. A) Pearson correlation between real and predicted changes in mobile service demands across transitions in COVID-19 regulations in France. B) SLM coefficients for the socioeconomic indicators used as features across the same transitional periods.

between 0.70 and 0.86, depending on the transitional period considered. Correlation values stay satisfactory when computed per city, with Pearson's coefficient that are above 0.6 in the vast majority of cases, which supports the generality of the model and of the predictive power of the selected features. An illustrative example of the quality of the model is provided for Paris in Figure 6.13D: the major spatial trends in mobile traffic changes are well reproduced across T_L, T_{D1} and T_{D2}.

The SLM coefficients, as shown in Figure 6.14B, display notable swaps of sign across periods that align with the cyclic effects previously seen on the traffic difference maps. Specifically, densely populated areas of residential destination (i.e., with low presence of leisure points of interest) and with lower income experienced traffic surges when the nationwide lockdown was put in place in T_L. In such neighborhoods, there is a tendency not to contract expensive fixed-line Internet access, and to use cellular connectivity as a cheaper replacement: being forced at home from the lockdown, local inhabitants consumed a substantially increased amount of mobile data traffic. This pattern is also consistent with multiple effects that include: higher-income households dropping cellular access in favor of high-speed fixed Internet at their home premises; richer families moving to second houses in the countryside during lockdowns, or leisure-dense areas suffering reduced visits under restrictions to non-essential mobility.

The transition to T_{D1} gives initial signs of reversed trends. With the end of the harsher lockdown limitations, the wealthier share of the population moved back to their urban residences, as indicated by the positive impact of median income on the growth of mobile data traffic usage. Similarly, partial resumption of leisure activities yielded an increase of mobile service demands in the associated neighborhoods.

Interestingly, the usage of cellular traffic in densely populated areas of the French cities remained unaffected by the transition from the lockdown to the first curfew period, i.e., during T_{D1}. The recovery in those areas only occurred with the removal of additional restrictions, as shown by the highly negative SLM coefficient of this feature during T_{D2}, which denotes a substantial drop in mobile data service consumption, i.e., a normalization of demands. This last transition is also linked to a full return to leisure activities in French urban areas, indicated by the positive SLM coefficient and increased traffic demands.

Key insights. *A linear model was proposed that relates the observed changes in mobile traffic consumption within cities due to COVID-19 restrictions with socioeconomic indicators. It's noted that areas with denser populations increase their mobile traffic consumption during curfews, while richer areas reduce (likely due to internal migration to secondary houses), as well as leisure areas. These changes are reverted when restrictions are lifted, with the population-dense areas seeing a reduction in traffic and leisure areas significantly increasing during the second curfew (when they were allowed to open again).*

6.2.4. Impact of containment policies on mobile applications

Mobile services are known to display a significant heterogeneity of spatiotemporal usages [34], and it can be verified that this diversity is also reflected on the impact of COVID-19 containment measures: the Pearson's correlation of the changes in the demands for total cellular capacity and for each service is as low as 0.08 on average across space and transitions T_L, T_{D1} and T_{D2}. As a consequence, the general effects of COVID-19 containment policies on the overall mobile traffic consumption that were observed above may not directly apply to the usage of individual mobile applications.

To investigate the existence and specificity of service-level effects, the same analysis described is Subsections 6.2.2 and 6.2.3 is repeated on the consumption of the individual mobile applications; the focus now is on the 38 services that are responsible for the generation of the highest traffic demands in France. Interestingly, the SLM approach based on the same three features of population density, median income and leisure space presence retains a high accuracy when applied to the demands for individual services[7].

Groups of services are identified whose consumption is similarly impacted by the different COVID-19 containment policies across T_L, T_{D1} and T_{D2}, by clustering applications based on their SLM coefficients. The agglomerative hierarchical clustering based on Euclidean distances among the SLM coefficients returns 5 clusters in each period, which are highlighted in the correlation matrices in Figures 6.15A1-A3. The result unveils how individual mobile applications display a relatively limited set of prototypical reactions to the imposed restrictions in major urban areas in France. Yet, when looking at the average SLM coefficients that characterize each of the identified clusters, it's found out that services in different clusters can exhibit highly diverse responses to COVID-19 policies, as depicted in Figures 6.15B1-B3.

Such heterogeneity of individual applications is rendered even more complex by the combined behavior of the same service across all transitions. The Sankey diagram in Figure 6.16 illustrates how applications move across the clusters of each transition: apart from some continuity in the services associated to the largest clusters, each mobile service reacts in a fairly unique way to the composition of the lockdown and curfew restrictions.

[7]Pearson correlation coefficients typically above 0.8 across all transitional periods and applications.

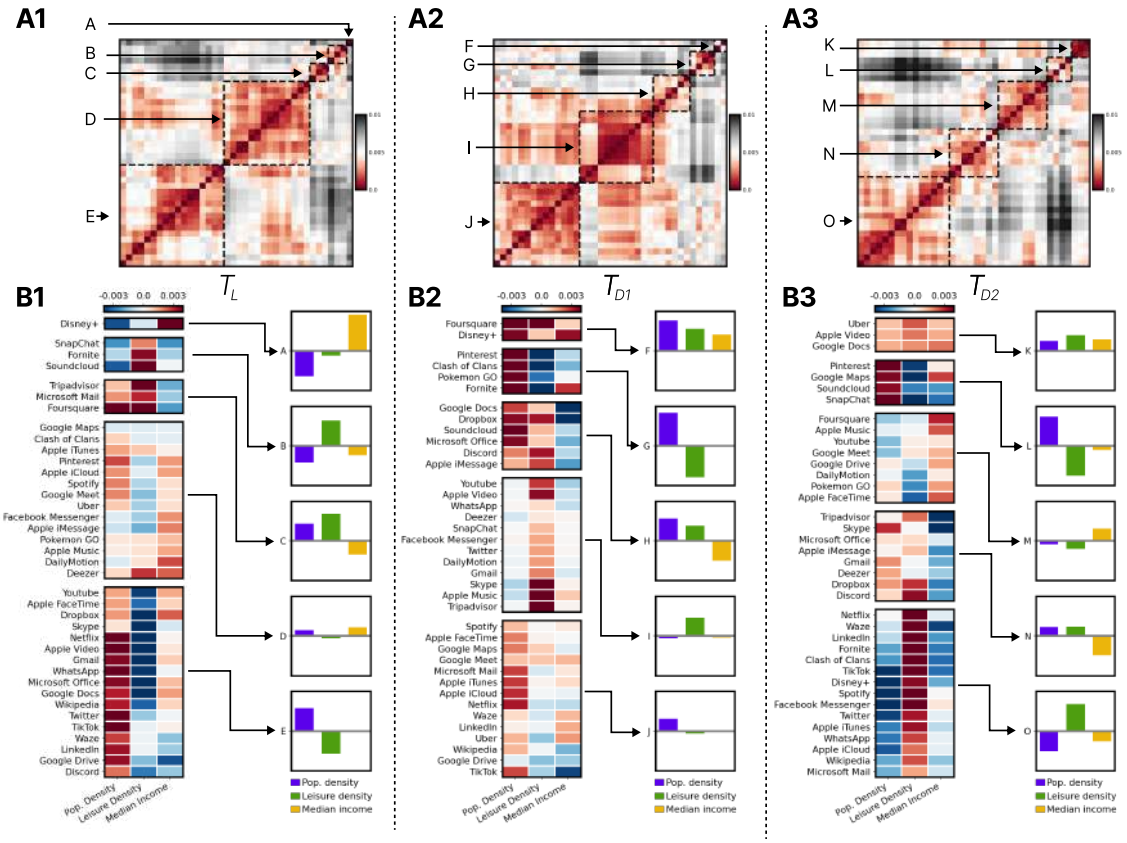


Figure 6.15. Results for the modeling of changes in mobile traffic due to COVID-19 restrictions for every service and period, ordered by a hierarchical clustering algorithm. A1-A3) Euclidean distance matrix showcasing how similar services in relation to their coefficients are grouped together. B1-B3) coefficient values for every service and restriction period are organized as clusters (left), as well as the average value of all coefficients inside each cluster (right).

Key insights. *Changes to the consumption of mobile services were heterogeneous, with not all applications following the same patterns observed in the overall traffic consumption. Those changes across periods could be clustered into 5 groups, where the major clusters of each period follow the same trends of overall traffic, but a significant number of apps still have diversities, which can be linked to the type of content consumed and target audience of each service.*

6.2.5. Implications of understanding the traffic changes in cities

This study unveils the impacts that policies enacted to contain the COVID-19 pandemic in France have generated on the consumption of individual mobile services. It shows that the spatial impact on urbanized areas was not uniform, with significant differences on usage across areas focused on residential or leisure activities, as well as a relation with the income level.

Moreover, complex and highly diverse dynamics are identified, showing that the

Figure 6.16. Sankey diagram of the shift of services across clusters in each period.

containment strategies have induced different changes of consumption behavior of each mobile application.

Specifically, it's noted that every transitional period has one main cluster (E in T_L, I in T_{D1}, and O in T_{D2}) with SLM coefficients similar to those observed for the total traffic; in some cases, a secondary large cluster (D in T_L, J in T_{D1}, and O in T_{D2}) also exists that also show slight variations with respect to the behavior of the total traffic. These clusters include very popular applications that generate high volumes of traffic, including Twitter, WhatsApp, TikTok, or Netflix. They drive the overall mobile traffic demand by creating increased demands during the lockdown in densely populated areas of residential destination and with lower income.

The results also reveal different and specific patterns, such as that of applications strongly liked to Apple devices (e.g., iMessage, iCloud, Apple Music) displaying an increased usage in higher income urban regions during the lockdown (cluster D), or a service of recent introduction at that time such as Disney+ being first adopted in richer and less populous areas of the French cities (cluster A).

Similarly, it is worth highlighting how information about food providers (e.g., Foursquare and Tripadvisor, in cluster C) was especially sought-after in both residential and leisure dense areas when the lockdown started, most likely due to local inhabitants looking for places that were open or accepting orders after the restrictions were enforced.

Also noteworthy are some of the effects induced by the end of the lockdown, such as younger shares of the population being allowed to stay outside but still barred from accessing leisure-dense places, so that they consumed mobile games (e.g., Fornite, Clash of Clans, Pokemon Go) closer to home than usual, as shown in cluster G.

6.2.6. Main Takeaways

The insights gathered by this analysis can help companies, including mobile network operators and mobile service providers, comprehend how their products might be affected by shifts in population mobility and allow a better policies to guarantee the quality of service requirements. Indeed, mobile application developers may employ spatiotemporal data about the impact of COVID-19 restrictions on the usage of their services to identify hidden adoption tends and discrepancies with respect to their competitors, leading a better offering on products inside their applications. Network operators can instead understand how to improve the scaling of their system for large-scale and persistent anomalous events such as epidemics.

Similarly, those results can help governments better understand the effectiveness of their actions across different indirect data sources, since the differences in smartphone usage can indicate flows of populations across spaces of the city. They can take advantage of mobile network data, together with models like the one proposed on this study, to comprehend how mobility restrictions and extraordinary events can impact different income levels inside urban centers.

6.3. Characterizing mobile demands in public protests

Among many other usages, smartphones and other mobile devices have become paramount instruments for the instant communication, advertising and reporting of significant social events. Organizers of gatherings of political, cultural or sports nature take advantage of a variety of mobile applications to announce, coordinate and disseminate the outcome of the happening, and journalists or online commentators use social media platforms as a means to report or opine about the ongoing events. Mobile services thus create a bridge connecting the physical world to the online ecosystem in the presence of large social events.

In the case of public demonstrations, recent years have witnessed a sparking interest in utilizing data from mobile phones and social networks to understand not only how protesters organize and communicate among themselves via these tools, but also how the digital sphere affects and is affected by large manifestations of social unrest. Existing works have explored in particular the relationship between protests and the X (formerly Twitter[8]) social media platform, by either analyzing geo-tagged tweets in the area of the rallies [229] or performing sentiment analysis of posted content related to the subjects of the demonstrations [230].

This Section proposes a different approach that embraces a more comprehensive view of the relationship between large protests and their impression on the digital world.

[8]Both names will be used interchangeably throughout this Section

Instead of focusing on an individual application or on content produced by its users, it will investigate how social manifestations impact the mobile traffic as a whole, by studying the relationship between the occurrence of public marches and the fluctuations in the overall demands for a variety of mobile services in the surroundings of the events. To this end, this Section will look at widespread social unrest episodes that happened in France in 2023 as a consequence of pension reforms proposed by the local government, and analyse traffic measurements performed in the affected regions and periods by Orange, the major network operator of the country.

This investigation unveils that large protests leave a very recognizable footprint on the overall mobile network traffic, especially when considering the consumption of individual services: indeed, a set of specific applications –mainly associated with social media and map navigation– experiences substantially increased demands, whereas others –generally related to entertainment– have significantly reduced relative usage. Such footprint is characterized in a more structured way by developing a learning model capable of accurately reconstructing the spatial and temporal dynamics of the target protests from mobile service demands only. The model informs about the significance of each application in pinpointing a protest in the historical measurements.

An additional step will be performed, linking the target demonstrations in France with mobile data traffic and showing how the total volume of traffic recorded by the network operator is a good indicator for the public participation in a large march. Combining this finding with the service-level model above allows building a framework for a-posteriori inspection of protests that discloses how the time-varying number of participants strode through the march area during each event. The framework is privacy-preserving as it only employs de-personalized traffic aggregates and does not offer opportunities for individual surveillance or re-identification of protesters. Instead, it offers the possibility of revealing, the precise progression of the marches, the alternate minor routes taken by participants or their dispersal at the end of the events, which can then inform both mobile network operators and local administrators towards better preparing the mobile connectivity infrastructure and coordinating safety measures for future similar situations.

6.3.1. The relationship of smartphones and public protests

Digital media consumption and dissemination is known to have a critical impact in how protests demonstrations occur online and break barriers offline [231]. Social media act as an alternative medium to traditional sources of news, and provide virtual stages where civil and political organizations can perform online recruitment of different segments of the population, and later move such masses of individuals to physical-world manifestations [232].

The role of smartphones in how a distributed population of protesters organizes, coordinates and reports in real time about the ongoing demonstration has thus become

a subject of multi-disciplinary research. Previous studies have proven that social media platforms in particular act as facilitators for information exchange *prior* to protests by shaping online relationships that then become a driving force for individuals to embrace the causes of dissent [233]. The effect is strong to the point that activity surges in geo-located tweets can be used to predict manifestation hours in advance [84] or to anticipate the rate of success or failure of the future demonstration [85]. Instant messaging applications, on the other hand, play a critical role *during* the social gatherings, with applications such as WhatsApp creating new ways for activists to meet, discuss and organize in real time [234]. Telegram is sometimes favored in that role as it mitigates the privacy concerns of the organizers [235], [236].

Mobile applications opened the door to unprecedented live coverage of public protests, as they offer an immediate way for any participant or local observer to report facts as they happen to global audiences, factually serving as a form of public journalism [237]. Protesters often share their visual expressions to help shape the narrative of the manifestation [238] and develop a community around their cause [239].

Real-time digital communication is also a double-edged sword when it comes to control enforcement of marches. While it can be leveraged by government agencies for monitoring purposes [240] or as a surveillance tool to counter-act against protests [241], social media coverage is also a means to announce violent repression of the manifestation by the authorities [86], [242] or even to coordinate counter-actions against police forces [243].

It is worth noting that not only protests but also many other categories of large social events have been studied through the lenses of the usage of mobile devices and applications. Across all different studies of the digital facets of public events, the most popular source of data is Twitter: the textual analysis of geo-located posts allows for thorough analysis [90], [230] and accurate forecasting [229], [244]–[247] of the target event. Images contained in posts can also be used for the same purposes [248].

Unlike prior work, the results showcased on this Section do not concentrate on a single mobile application nor solely examine the content of material shared therein. Instead, it shifts attention to the entirety of data traffic that encompasses a wide range of mobile services beyond single social media and messaging applications. While earlier studies have adopted a similar perspective to investigate different phenomena, none has ever applied it to public protests. For instance, network traffic measurements have been used to characterized cultural or sports events [87], understand the impact of holidays during elections [89] or monitor road traffic congestion [91].

Closer to this approach, a few studies have employed signaling data from mobile networks to study the attendance to mass protests [88] and to link it with the home location of participants [249]. Yet, such works depend on personal network localization information to investigate the dynamics of demonstrations in real-time, and do not look into mobile traffic demands or application consumption which are the focus of our study.

Also, this research is not targeted at live tracking of protesters which entails significant privacy risks, and raises ethical questions. Ultimately, this is the first investigation of the effects that large public demonstrations induce on the whole mobile data traffic and on the demands for a varied range of applications.

6.3.2. Data processing for the study of public protests

This study leverages network traffic measurements from the production network of Orange in France. The traffic data collection was performed in a continuous manner from January 31 to May 31, 2023 over 29,171 carriers covering five main urban areas of France, i.e., Paris, Lyon, Toulouse, Nantes, and Bordeaux. In these regions, all available RAN technologies are monitored, i.e., 2G, 3G, 4G and 5G: while 4G and 5G serve today the vast majority of the demand, it is not uncommon that 2G and 3G serve non-negligible portions of the data traffic in heavily loaded scenarios such as those produced by large manifestations.

By crossing the localization information with the service-level classification data (further details can be seen in Section 3.3), and aggregating the result over all mobile devices attached to every RAN carrier, the operator computed the demands for each application served by every carrier. For the purpose of this study, such information also accumulated over time into 5-minute intervals. The result are time series of the volume of traffic (in bytes) generated by over 400 mobile services at each of the target 29,171 carriers at every 5 minutes. These data measurements and processing abide by all applicable regulations, and the resulting aggregates are privacy preserving since they does not allow re-identifying individuals or retrieving personal information.

6.3.3. Baseline carrier-level service demands

During the observation period, a number of protests linked with the French pension reform took place in the considered cities. The full list can be found in Table 6.2.

In order to assess the impact of protests on mobile network traffic, it's necessary to establish a baseline period where the consumption of mobile services is deemed *normal*, i.e., not affected by similar demonstrations. As the majority of pension reform strikes in France occurred during work days, a set of 10 baseline work days are considered, being all non-holidays and not marked by manifestations according to the available records, and are not immediately adjacent to days characterized by protests: February 8, 9, 14, 15, 28, March 1, 2, and April 25, 26, 27, 2023.

For each carrier and mobile service, a reference daily demand is computed using the median volume of traffic at every 5 minutes in the 10 baseline days for the target carrier and service. This captures the typical expected traffic is normal conditions.

Protest		Attendance estimation	
City	Date	Organizers	Police
Paris	January 31	500,000	87,000
	February 7	400,000	57,000
	February 16	300,000	37,000
	March 7	700,000	81,000
	March 15	450,000	37,000
	March 23	800,000	119,000
	March 28	450,000	93,000
	April 6	400,000	57,000
	April 13	400,000	42,000
	May 1	550,000	112,000
Lyon	January 31	45,000	25,000
	March 7	50,000	25,000
	March 23	55,000	22,000
	March 28	30,000	12,500
	April 6	32,000	13,000
	May 1	45,000	17,000
Toulouse	January 31	80,000	34,000
	March 7	120,000	27,000
	March 23	150,000	30,000
	March 28	150,000	23,000
	April 6	90,000	15,000
	May 1	100,000	13,500
Nantes	January 31	60,000	28,000
	March 7	75,000	30,000
	March 23	80,000	25,000
	March 28	60,000	18,000
	April 6	50,000	15,000
	May 1	80,000	17,500
Bordeaux	January 31	75,000	16,500
	March 7	100,000	16,500
	March 23	110,000	18,200
	March 28	80,000	11,000
	April 6	60,000	10,000
	May 1	130,000	12,000

Table 6.2. City, date and estimated participation of the protest events investigated in our study.

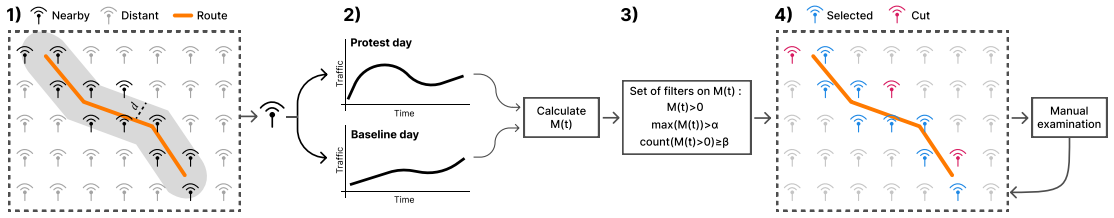


Figure 6.17. Methodology for establishing the true positive (blue) carriers for the ground truth, based on the variation of mobile traffic consumption during the days where a protest happened. The cut carriers (in red) were not selected due to not having a significantly clear change in their pattern.

6.3.4. Characterizing protests through mobile traffic analysis

In order to characterize the impact of large manifestations on mobile networks, it's first analyzed the changes in the dynamics of overall traffic consumption during the dates when protests occurred, with respect to the *normal* baseline days. More precisely, the goal is to gain insights from two perspectives, as follows.

1. Understand how protests impact the overall consumption of traffic in each carrier near the protest route, proposing a metric capable of identifying clear changes;

2. Understand how protests may affect the consumption of traffic of specific mobile applications; indeed, it is expected that not all apps are used in the same way, and some are especially impacted –in a positive or negative way– by the occurrence of a mass manifestation.

For this analysis, traffic from the city of Paris is used, taking into account 10 weekdays when public demonstrations against the French pension reform occurred, as per Table 6.2.

6.3.4.1. Pinpointing where and when protests affect overall mobile traffic

In the following, the methodology utilized to identify traffic pattern deviations caused by the presence of protesters and their disruption on the mobile network is described. This methodology is depicted in Figure 6.17, showing that its objective is to determine the clear true positives from protest days. Thus, establishing ground truth that can be later utilized for a more complex classification methodology.

For a given protest day, the first step involves an initial triage of carriers based on their proximity to the protest route authorized by the Paris Police Prefecture. This classification gives reasonable certainty that the observed patterns are happening near the expected protest location (and they are not uncontrolled outliers away from it). The proximity threshold is set as $d = 1$ km. Considering only this set of nearby carriers, the second step involves identifying carriers that had their traffic pattern clearly affected by protesters' presence. To achieve this, the deviation of the traffic consumption during

protest days with respect to the baseline days is computed, as defined in Section 6.3.3. It's important to note that at this stage this methodology is set to be conservative, i.e., it seeks to select only carriers greatly affected by the protest. As this is an intermediary step of the analysis, the objective is to find carriers to be used as true positives to better characterize the impacts of protests on mobile network traffic consumption.

In order to compute the traffic pattern deviation, the traffic on carriers is first normalized during both protest and baseline days. Here, a volume-oriented (e.g., min-max) standardization would not be ideal as it puts emphasis solely on volume variations, which may lead to incorrect interpretations. For instance, one carrier may yield a higher demand during the protest day than in the baseline and yet have aligned peaks of usage, meaning that the observed traffic consumption behavior is semantically the same and most likely not caused by the protest (which is limited to a few specific moments in time). Instead, a z-score standardization is preferred, which avoids volume bias and allows for a direct comparison of the traffic dynamics between different days.

For each 5-minute interval $t \in T$, $P(t)$ is denoted as the z-scored traffic observed during the protest day. The z-scored traffic at a given baseline day $n \in N$ will be defined as $B_n(t)$, with the set $\mathbf{B}(t) = \{B_1(t), ..., B_n(t)\}$ containing the time series of all baseline days, with the same range of values T as $P(t)$. The instantaneous mean of $\mathbf{B}(t)$ will be $\mu_b(t) = \frac{1}{N}\sum_n^N B_n(t)$ and the instantaneous standard deviation will be $\sigma_b(t) = \sqrt{\frac{1}{N}\sum_n^N (B_n(t) - \mu_b(t))^2}$.

The traffic pattern deviation is then defined as

$$M(t) = P(t) - (\mu_b(t) + 3\sigma_b(t)), \tag{6.6}$$

and time t is considered to be affected by a significant change in mobile activity due to protesters roaming within coverage of the carrier if $M(t) > 0$. In other words, significant deviations are identified as those exceeding $3\sigma_b(t)$, which, in case the distribution of baseline traffic over a given instant t on the set $\mathbf{B}(t)$ follows a normal distribution, corresponds to values outside the 99.7% probability range.

The third step of the methodology involves a set of filters done on top of $M(t)$ to remove any potential outliers. Besides having $M(t) > 0$, it's considered that whenever a protest affects a carrier, a significant number of marked time instants t will exist for the respective carrier. This sequence of instants will compose set $\tilde{T} \in T$, which are considered the potential interval where the presence of the manifestation was clearly affecting the usual traffic demand of that carrier. This means that two extra filters can be defined. The first will be α, which represents the maximum value of $M(t)$ for a given carrier. After examining results, it's determined that $\alpha = 0.5$, which means that for a carrier to be flagged only if $max(M(t)) > 0.5$, which will filter out cases of noisy outlier and keep only carriers where the disturbance was significant. The second is in relation to how many

instants are being flagged in each carrier. A carrier which may have a single $5-min$ interval flagged is considered to perhaps be a potential outlier, so the parameter $\beta = 10$ is chosen, which means that for a carrier to be marked as a true positive, it needs to have $count(M(t) > 0) \geq 10$.

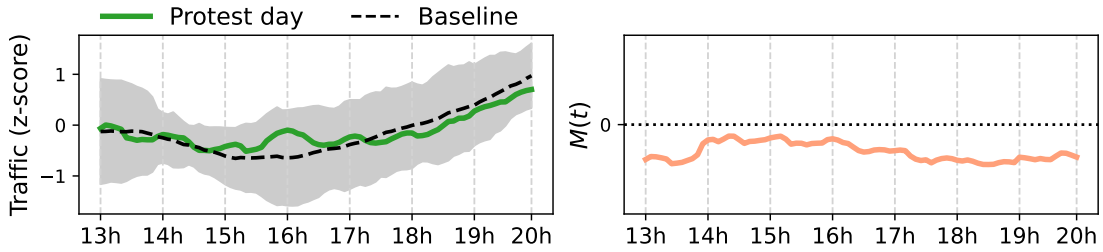
A set of clear true negative carriers is also calculated, which are denoted as the carriers (i) far from the protest route based on the initial triage, and (ii) presumably not affected by the protest, i.e., $P(t)$ is significantly similar to all values within $\mathbf{B}(t)$. After calculating $M(t)$ for all far away carriers, the true negatives are defined as any carrier where its 97^{th} percentile of $M(t)$ is smaller than -0.5. This means that at least 97% of the 5-min time instants for this carrier of $P(t)$ are within the interval estimate of $\mathbf{B}(t)$.

All the clear true positive and true negative carrier are then manually inspected, in order to obtain the final ground truth set. For the sake of clarity, two obtained examples will be discussed. Figure 6.18a presents a *true negative* carrier, where both $P(t)$ (in green) and $\mathbf{B}(t)$ (represented by $\mu_b(t)$ as the dashed line and the $\pm 3\sigma_b(t)$ interval around it) have a similar pattern for the z-scored traffic (left plot). This results in a traffic pattern deviation $M(t)$ (right plot, in orange) that is never greater than zero, implying that the temporal consumption pattern of this carrier was not affected by the presence of the protesters. On the opposite hand, Figure 6.18b shows a *true positive* carrier, with clear changes in its traffic consumption patterns. It can be specifically noted that this carrier experiences an uncommon surge of traffic demand $P(t)$ just after 15:00, which is only normalized after 17:00. Indeed, this surge of demand is completely outside the 99.7% interval estimate of $3\sigma_b(t)$, as observed over $M(t)$ on the right plot. The set of time intervals $\tilde{T}$ which were flagged as the potential hours that the protesters were within the coverage area of this carrier are marked in light gray.

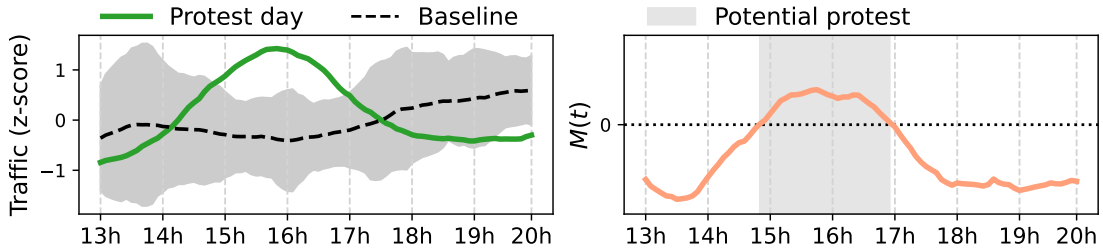
Key insights. *The proposed traffic pattern deviation can identify clear anomalies in the dynamics of traffic consumption of individual antennas of the network, by comparing traffic on the protest day with baseline values. This results can be utilized to establish a ground truth that will be utilized for a more robust classifier.*

6.3.4.2. Validation of the ground truth: Spatiotemporal tracking of the protest

After establishing a ground truth of carriers clearly affected by the presence of the protesters, the next step involves validating how the set of true positive carriers obtained for each day was affected over space and time, specifically in relation to the intervals $\tilde{T}$ that were marked as potential moments when the protest was passing by the carriers' coverage. The temporal evolution of $M(t)$ across the true positive carriers of each protest day can be observed on Figure 6.19, where red values indicate $M(t) > 0$ and black lines denote the interval $\tilde{T}$ where each carrier was affected. As expected, an evolution of $M(t)$ happened over time: some carriers were flagged only at the initial points hours of



(a) Carrier not affected by the protest



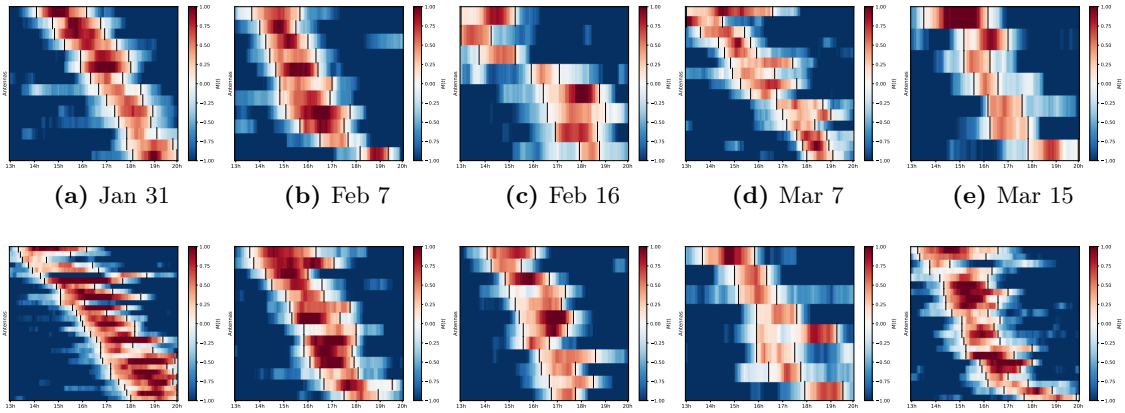
(b) Carrier affected by the protest

Figure 6.18. Examples of (a) true negative and (b) true positive carriers, in relation to how their traffic patterns deviated during the protest day against the baseline days. The proposed $\boldsymbol{M(t)}$ analyzes the difference of protest (in green) and baseline (in black) traffic, highlighting the potential hours when the protest affected the selected carriers whenever $\boldsymbol{M(t) > 0}$, as indicated by the gray range, together with additional filters.

the protest (14:00), some in the middle and others near the end (19:00). While this is an expected behavior, it's an initial validation for the ground truth set, as it shows that while protesters marched, their position over time changed and different carriers were affected at different instants. Indeed, no single carrier display the traffic pattern disturbance throughout the full 14:00 to 19:00 time interval of the protest, which confirms the natural handover of mobile network users as they progressed through the route.

The natural movements of protesters along the designed route, following the true positive carriers, can be observed over Figure 6.20, where lighter/darker colors indicate carriers that were affected earlier/later. As expected, carriers that were affected earlier are nearby the official starting points of each protest, while the ones affected later are nearby the official end points, with intermediary carriers being along the announced route by the Police of Paris. These results over both the temporal and spatial evolution of the carriers marked by having their traffic disturbed gives confidence that indeed mobile network measurements can be used to identify the presence of large-scale manifestations in both the spatial and temporal dimension, and that the obtained ground truth can be trusted for the next steps of this work.

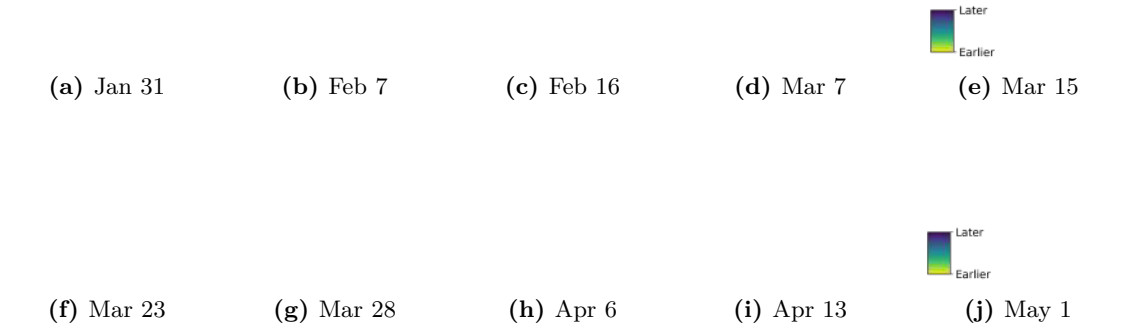
Key insights. *By observing the temporal and spatial evolution of all selected antennas with traffic disturbances, it can be noted that their consumption anomalies are within protest hours, where no single antenna is being disturbed; instead, the natural handover procedure of the network is observed, with protesters affecting at early hours the antennas*



(a) Jan 31 **(b)** Feb 7 **(c)** Feb 16 **(d)** Mar 7 **(e)** Mar 15

(f) Mar 23 **(g)** Mar 28 **(h)** Apr 6 **(i)** Apr 13 **(j)** May 1

Figure 6.19. Evolution of the protests, represented by the standardized traffic of the flagged carriers, ranked (from yellow to purple) by the average time when $\boldsymbol{M(t) > 0}$.



(a) Jan 31 **(b)** Feb 7 **(c)** Feb 16 **(d)** Mar 7 **(e)** Mar 15

(f) Mar 23 **(g)** Mar 28 **(h)** Apr 6 **(i)** Apr 13 **(j)** May 1

Figure 6.20. Evolution of the protests represented by the spatial distribution of the flagged carriers, color coded according to the time ranking.

nearby the official start of the route, and affecting the antennas nearby the end of the official route by the announced end hours of the protest. This gives confidence that the methodology for building ground truth can be trusted for the next steps of this work.

6.3.5. Modeling traffic changes due to mass protests

Following the procedure described in Section 6.3.4.1, a ground truth set is obtained, composed of more than 100,000 3-tuples with type (*carrier_id*, *time*, *label*), where *carrier_id* corresponds to a specific geographical location, *time* is a 5-min interval, and *label* classifies each data entry as a true positive or true negative sample. The label distribution is inherently imbalanced, with a ratio of 1 to 30 between positive (protest) and negative (non-protest) classes.

The ground truth, based on the traffic pattern deviation metric, can be used to

classify carriers affected by public protests at a 5-minute time resolution. Nevertheless, such a classifier model should not solely rely on overall traffic statistics, as deviations from baseline traffic patterns are not only attributable to public demonstrations, but can also be produced by other crowded events, such as sports games, concerts, and festivals. Therefore, mobile traffic during large public protests will be characterized and modeled based on the consumption patterns of individual mobile services.

6.3.5.1. Application consumption features

When it comes to per-application traffic, simply comparing volumes of applications across carriers is not a reliable approach, specially to later utilize those traffic differences as model features. Due to the natural diversity of traffic volumes, services like video streaming will generate much higher network traffic loads than low-traffic demanding services such as instant messaging applications. Moreover, carriers in general may expect different loads, according to the network optimization routines defined by the operator, i.e., a higher capacity macro cell will expect more users attached than a specific beaming micro cell, which will as well lead to different orders of magnitude volume between both. Consequently, those differences in magnitudes may just hide away anomalies and create volume biases for models. Therefore, an alternative method for quantifying the impact of manifestations on mobile application usage is required.

To overcome those challenges, a new metric is proposed for the characterization of changes in the volume of mobile applications' traffic due to massive events. Let $i \in I$ be a specific application from the set I of all mobile applications. For each carrier, let $T_{p,i}(t)$ be the traffic on the protest day at time t, and $T_{b,i}(t)$ be baseline traffic at time t. Then the proposed metric is defined as

$$F_{i,ns}(t) = \frac{T_{p,i}(t) / \sum_i^I T_{p,i}(t)}{T_{b,i}(t) / \sum_i^I T_{b,i}(t)} \tag{6.7}$$

$$F_i(t) = \frac{F_{i,ns}(t) - 1}{F_{i,ns}(t) + 1} \tag{6.8}$$

where $F_{i,ns}(t)$ is the non-symmetric metric and $F_i(t)$ is its the symmetric version.

$F_i(t)$ represents the percentage of traffic each mobile application generates in relation to the remaining set has changed on the day of the protest, versus the baseline. Values of $F_i(t) > 0$ would mean that the app i at instant t had a rise in importance against other apps in regards to the baseline, while $F_i(t) < 0$ would mean a loss of importance against other apps during the protest day, in relation to the normal period.

In order to check the validity of the proposed application metric, its values are compared for the set of true positives and the set of true negatives. As previously mentioned, the ground truth set was established from the overall traffic analysis only,

where no distinction between applications was considered. Therefore, this metric can help find applications whose consumption patterns changed because of the presence of protesters. Figure 6.21 represents the distribution of $F_i(t)$ for true positive carriers (in green) versus true negative carriers (in black). A few remarks can be made from this plot:

1. A first set of applications have a clear separation in their $F_i(t)$ values between true positive and true negative carriers, where there's a growth of the share during the protest day for the positives and a decrease for the negatives. This set includes instant messaging, localization, and news apps, such as WhatsApp, Twitter, Google Maps, and NewsPaper. Thus, suggesting that those applications had their consumption greatly affected by the protests.

2. A second set presents a clear separation between true positives and negatives, although not an increase of share. This set includes Telegram and Transport. Those apps are believed to be moderately affected by the protest since they had *a significantly lesser decline* in their share against the baseline for the true positives.

3. A third set of applications had a growth of share in the protest day against the baseline both for the true positives and negative sets. This includes popular social media apps, such as Instagram, Facebook and Snapchat. This can be related to not only people generating content within the march, but also users searching for this content throughout the city.

4. A final set is characterized by applications with a decrease of usage in the carriers affected by the protest, including specially applications related to audio (Spotify, Deezer) and video (Netflix) streaming, as well as work-related (LinkedIn).

Key insights. *A metric that can quantify variations in mobile application demand due to the presence of manifestations was proposed. After analyzing the distribution of the metric for the ground truth carriers against the remaining set, it's possible to find a set of apps with significant changes in demand during the protest day, in relation to the baseline. Those applications are mainly related to instant messaging, transportation, and news, which are classes expected to be highly utilized during mass manifestations.*

6.3.5.2. Model description and evaluation

After enriching the ground truth set with the application consumption features, a classification model is built to predict, at a 5-minute resolution, which carriers are affected by large public protests, against the baseline days. A XGBoost classifier (XGBC) is proposed, which relies on an ensemble of decision trees capable of handling class imbalance through a weighing class technique. After hyperparameter tuning, the resulting XGBC consisted of 1000 decision tree estimators, reinforcing the complexity of the model. Also,

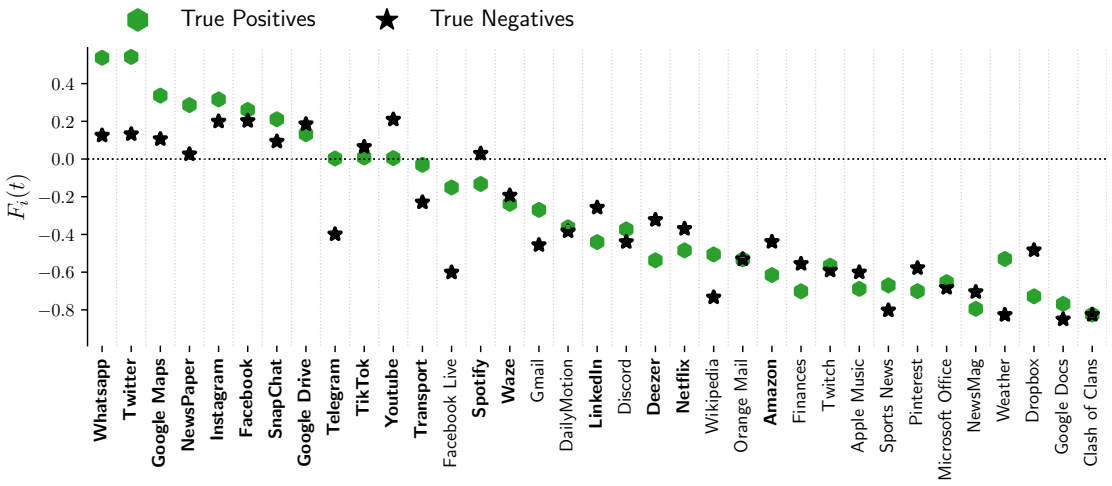


Figure 6.21. Distribution of $F_i(t)$ across all days in Paris. Apps that represent a significant change in usage are the ones where a clear separation between the distribution of protest and baseline happens, such as WhatsApp and Twitter.

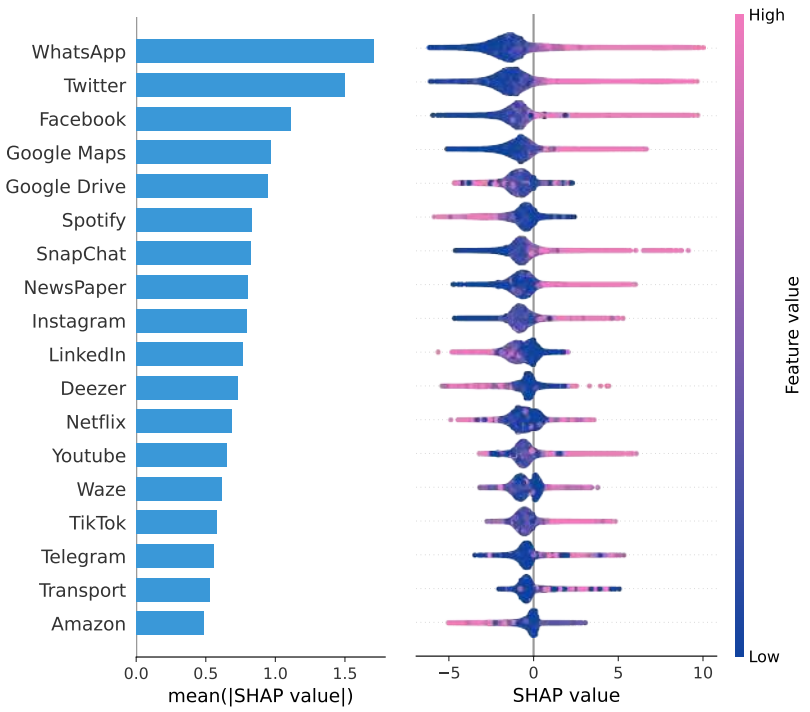
		Predicted	
		Protest	Non-protest
Actual	Protest	970	56
	Non-protest	10	31,116

Table 6.3. Confusion matrix of trained XGBC when applied to test dataset.

a logistic regression for binary classification is defined as a loss function, with a learning rate of 0.05. Moreover, the XGBC deals with class imbalance by weighing the balance of true positive examples relative to true negative examples, favoring better performance in the positive protest class.

To analyze the classifier's performance, the ground truth is randomly sampled into training and testing sets using a 70:30 splitting ratio. Then, the model is trained on the training dataset using a subset of 18 of the application consumption features, selected after a feature engineering process. The classification results using the XGBC are summarized in Table 6.3, where its high accuracy in distinguishing the carriers affected by public protests is evidenced. Accordingly, it's found that the provided set of features allows the XGBC model to determine with a 0.97 F1-score whether a large public protest has impacted a carrier during a specific 5-minute interval. The high F1-score value also implies the good performance of the model in terms of precision (0.99) and recall (0.95).

In order to provide insight into the explainability of the classifier model, the SHapley Additive exPlanations (SHAP) values are calculated for the trained model to analyze the contribution of individual features to the classification outcomes. As shown in



(a) Feature importance **(b)** Explanation summary

Figure 6.22. Contribution of application features to the XGBC model's prediction.

Figure 6.22a, the absolute SHAP values are employed to determine feature importance. From the figure, it can be seen that changes in the consumption of WhatsApp and Twitter contribute the most to the model's prediction. Additionally, Figure 6.22b provides insights into the associations learned by the classifier model, i.e.,, whether protest-affected carriers are related to increased or decreased mobile service consumption. The figure suggests that affected carriers relate to increased consumption of services such as WhatsApp, Twitter, Facebook, and Google Maps, as well as a decreased consumption of Google Drive, Spotify, LinkedIn, and Netflix, among others. It is important to note that these trends are well-aligned with the insights into service consumption changes presented in Section 6.3.5.1. Indeed, Figure 6.21 shows that, in general, the selected features (highlighted in bold) presented high differences when comparing true positives and true negatives data points.

6.3.6. Identifying protests through disturbances in the network

To further validate the proposed method, different experiments are performed by running the classification model over an entire target day. As a result, a list of potentially affected carriers by a protest during specific time intervals is obtained. Besides, given that public demonstrations are assumed to be continuous in space and time, a secondary step is incorporated to reduce the number of false positives and identify any potential

protest as a complex but consistent event. More precisely, a density-based spatiotemporal clustering algorithm is applied (ST-DBSCAN [250]) over the pairs of carriers and time intervals previously classified as protest by the XGBC. Therefore, detecting a potential protest lies in discovering a spatiotemporal cluster of carriers classified as affected. In all experiments, a specific parametrization is employed for the ST-DBSCAN algorithm to ensure consistency in detected protests. Accordingly, the neighborhood around a point is defined by a spatial radius of $1,2$ km and a temporal radius of 45 minutes. In addition, the minimum number of points to form a dense region is set to 30.

6.3.6.1. Intra-city testing

Firstly, an intra-city test is carried out, where the XGBC model is trained using a subset of the ground truth dataset to later classify all the carriers of Paris at a 5-minute resolution. The 2-step methodology (XGBC + ST-DBSCAN) is experimented to detect each of the ten large demonstrations in Paris, previously mentioned in Section 6.3.4.1. Given that the ground truth contains data from all ten protests under consideration, it will be first removed the target day from the training set. Thus, only the remaining nine protest days are used to train the XGBC in each case. Then, the 2-step protest detection approach is applied over the entire target day to corroborate the identification of the protest along the officially authorized route.

In order to illustrate the importance of employing a density-based clustering algorithm, Figure 6.23 depicts the spatiotemporal temporal patterns of the carriers classified as affected during a protest day (May 1). Those results are built on the output of the XGBC, which classified more than 1,500 carriers in Paris for each 5-minute interval between 8:00 and 22:00. According to the figure, less than 3% of all the carriers were labeled as affected at each time interval. Notably, a faction of the detected carriers emerges from the rest, exhibiting high density in both spatial and temporal scales. Therefore, the incorporation of ST-DBSCAN is essential, as it can help separate false positive data points from the actual group of carriers affected by the protest. Indeed, in all ten experiments, ST-DBSCAN distinguished a single dense cluster and labeled the rest of the points as noise.

Figure 6.24 shows the output of the protest detection approach, where the carriers labeled as affected are represented by their Voronoi cell, and the coloring depicts the average time of the day when each carrier was classified as affected. It can be seen that the methodology consistently identified the ten public protests in Paris, densely covering the official routes and exhibiting a clear time consistency referable to the protesters moving from the start to the end of the march. These results greatly enhance the rudimentary set of impacted carriers shown in Figure 6.20, not only by filling the gaps along the routes, but also by exhibiting some interesting phenomena, such as the dispersion of people at the end of the demonstration (e.g., Figures 6.25d, 6.24b, and 6.24i), and the existence of alternative less-pronounced routes (e.g., Figures 6.24d, 6.24h, and 6.24i). Those

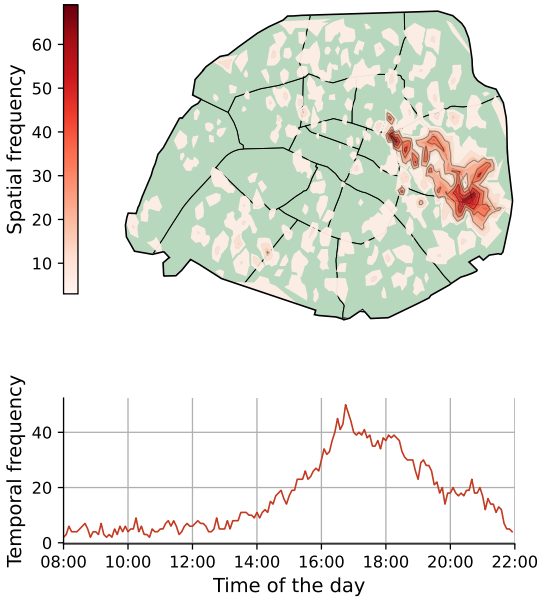


Figure 6.23. Carriers labeled as affected during May 1 in Paris. Frequency of labeled carriers in both spatial and temporal scales.

alternative itineraries were not part of the authorized routes, yet some were non-officially advertised as alternative routes to avoid high congestion levels during the marches.

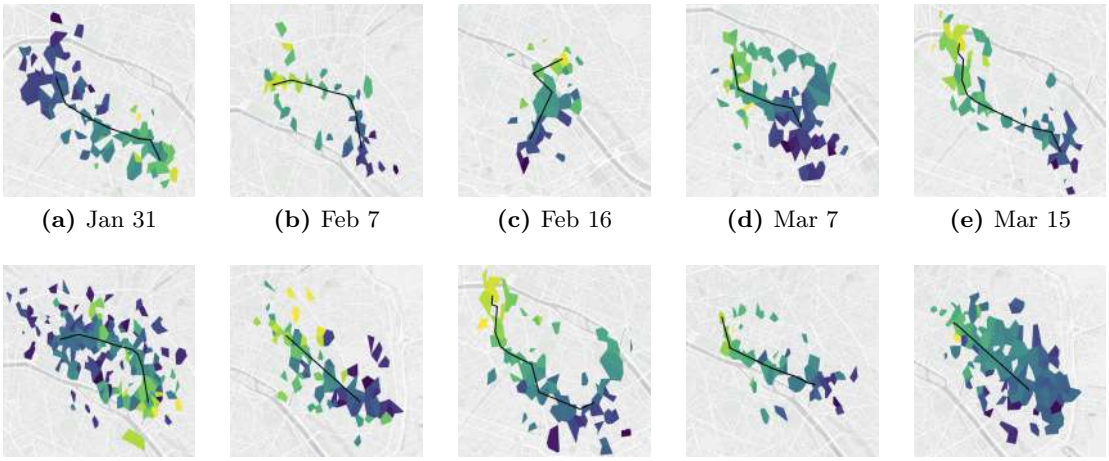
In addition, a secondary focus is put on inspecting the proposed detection method in terms of its precision in identifying public protests. Accordingly, the XGBC model is trained over the full ground truth set and apply the 2-step approach to four normal non-protest days in Paris: May 16, 17, 23, 24 and 25. The data for the selected days was generated following the procedure described in Section 6.3.5.1. As a result, no potential protests were found on the test days.

6.3.6.2. Inter-city testing

Secondly, an inter-city test is performed to discuss the feasibility of using an XGBC trained in a primary city to detect public demonstrations in secondary cities.

For this analysis, four French cities are considered (Lyon, Toulouse, Nantes, and Bordeaux) and their network traffic is inspected for six of the ten days of nationwide pension reform protests taken into account. Table 6.2 summarizes the selected protest days and cities, showing a great diversity among the magnitude of the events.

Thus, the XGBC is trained using the ground truth dataset containing information about the ten protests in Paris, and applying the proposed methodology to identify the selected 24 events, following the same model configuration as in the previous experiments. Figure 6.25 presents the obtained results according to the 24 test cases, where carriers are represented by Voronoi cells and the colors relate to the average time of the day each carrier was affected. For each target event, the methodology is able to identify a



(a) Jan 31 **(b)** Feb 7 **(c)** Feb 16 **(d)** Mar 7 **(e)** Mar 15

(f) Mar 23 **(g)** Mar 28 **(h)** Apr 6 **(i)** Apr 13 **(j)** May 1

Figure 6.24. Intra-city testing: Protest events identified in Paris based on the proposed detection methodology. Early affected carriers colored yellow, while later affected carriers colored purple.

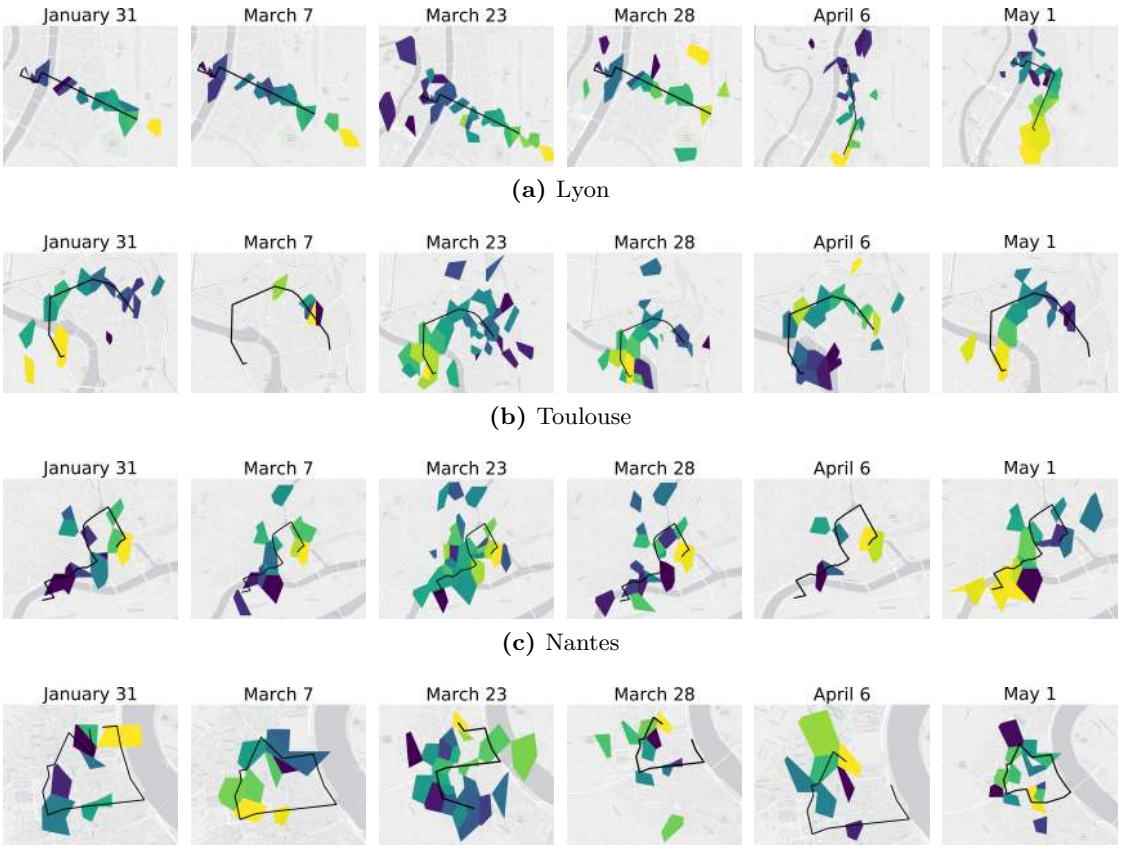
spatiotemporal dense set of affected carriers, separating the single discovered cluster from the noise data. A closer inspection of the figures exhibits consistency between the coverage area of affected carriers and the officially authorized route, which further supports the soundness of the proposed method.

Key insights. *Overall, these results show that the proposed methodology can identify public marches in various locations based on initial city traffic patterns. The successful detection of events of varying magnitudes demonstrates that this approach is a fundamental step toward developing a nationwide unified model.*

6.3.7. Dynamic estimation of protest attendance

Public demonstrations are mainly quantified in terms of the number of participants, which provides qualitative evidence of the magnitude of the event. Thus, organizers and local authorities employ different methods for people counting to announce their attendance estimations, typically during the course of the protest. Based on the outcomes of the proposed framework, the estimation of protest attendees from aggregated network traffic is explored next.

The ten identified protests in Paris are analyzed, and a protest-related traffic time series from the total traffic consumed across affected carriers every 5-minute interval is computed. Then, the maximum value of the traffic volume time series is associated with the organizers' and police's attendance estimations. It is well established from a variety of studies the existence of a power relationship between mobile network activity and population [251], [252]. Thus, the protest attendance is modeled as a power function of the peak traffic volume related to affected carriers.



(a) Lyon

(b) Toulouse

(c) Nantes

(d) Bordeaux

Figure 6.25. Inter-city testing: Protest events identified across different French cities based on the proposed detection methodology. Early affected carriers colored yellow, while later affected carriers colored purple.

As shown in Figure 6.26, a consistent relationship is found for both sources of estimations, having $R^2 \geq 0.69$, and thus, supporting the existence of such functional power relationships. However, a closer inspection of the data points in the Figure suggests that the relationship between attendance and peak total traffic could also be explained using a simpler linear regression model. Indeed, when adjusting a linear function to those points, R^2 values are obtained similar to the ones obtained by the power regression: 0.73 and 0.79 for the organizers' and police's estimations, respectively. However, the optimized intercept values of both linear regressions are overly high, associating zero network traffic with almost 230,000 (organizers) and 17,000 (police) attendees. Therefore, modeling the relationship between attendance and the peak of total traffic as a linear function, leaves no room for analyzing medium-sized events, such as most of the demonstrations outside Paris illustrated in Figure 6.25 and listed in Table 6.2. In contrast, it's shown next that a power regression adjusted to massive protests in Paris can be satisfactorily used to examine demonstrations of less magnitude in terms of attendance.

Accordingly, experiments are also done with estimating the number of participants

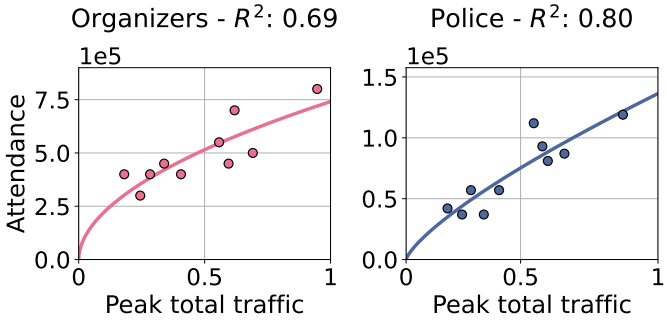
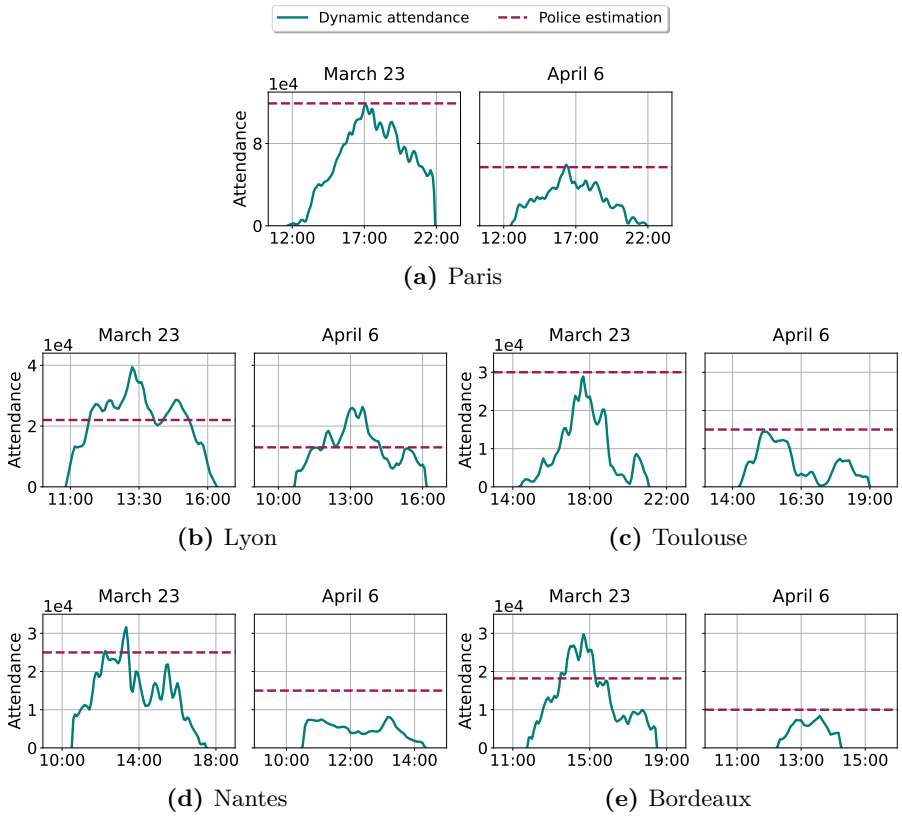


Figure 6.26. Power regressions of the attendance estimations against the peak total traffic related to a protest event. Real traffic volumes have been removed from the plot.

in the protests outside Paris identified in Section 6.3.6.2. Thus, the peak traffic volume is computed from the affected carriers and the power regression models are applied for organizers' and police's estimations previously adjusted to protests in Paris. Then, an analysis is made about how well the regression values approximate the actual estimations of both entities. On one hand, the regression model for the organizers' estimations predicted the announced protest attendance with a MAE of $57,045$. On the other hand, the regression model for the police's estimations obtained a lower MAE of $7,821$. These results suggest that the police's methodology for estimating participation during protests is more consistent across cities. Consequently, the following experiments devise a dynamic estimation of attendees, considering the police's power regression only.

Advantage is taken of the previously designed models to extend our a-posteriori analysis of public protests. Thus, the power regression models are extrapolated to the entire protest-related traffic time series, obtaining a dynamic estimation of protest attendance along the duration of the identified event. Then, two target days of nationwide demonstrations are selected to study the results of the proposed dynamic attendance estimation, March 23 and April 6, which cover a wide range of protest sizes. Indeed, for each city, the demonstration on March 23 is consistently listed among the protests with higher attendance, whereas the demonstration on April 6 is usually one of the less crowded events (see Table 6.2). Figure 6.27 shows the time-variant number of participants in both protest days across the five cities analyzed in this study. Intuitively, the dynamic number of participants is generally characterized by a bell-shaped curve, showing the increasing number of participants from the start of the demonstration achieving a maximum value around the median time of the event's duration. However, a surge in participants is also observed close to the end of some demonstrations (e.g., Figures 6.27b and 6.27d), which may be related to external factors such as the geographical area where the protests occurred. Additionally, as these attendance time series are based on the regression of police estimations, it's included in the Figures the official numbers (dashed line), which can be compared against the maximum value of the time-variant attendance. Thus, these



(a) Paris

(b) Lyon **(c)** Toulouse

(d) Nantes **(e)** Bordeaux

Figure 6.27. Dynamic attendance estimation in five French cities during the course of two nationwide protest events.

Figures provide further support for the suitability of a power regression to estimate the number of participants from aggregated volumes of mobile traffic.

Furthermore, a dynamic estimation allows the anonymization of the progression of a public march. For example, Figure 6.28 describes the evolution of the protest on March 15 in Paris, allowing to associate a time-varying number of participants with a well-defined area and time. Taken together, these results provide an important basis for the development of a privacy-preserving framework to investigate similar events in detail.

6.4. Main takeaways

The work presented in this Section is a first analysis of the effects of large public protests on the demands for mobile services, revealing the recognizable footprint that such events leave on the data traffic –in particular at the level of individual services. By leveraging these findings, it's shown that it is possible to use measurements collected by cellular network operators to perform an a-posteriori characterization of the spatiotemporal dynamics of the demonstrations that includes the identification of the paths taken by protesters as well as the estimation of their time-varying number.

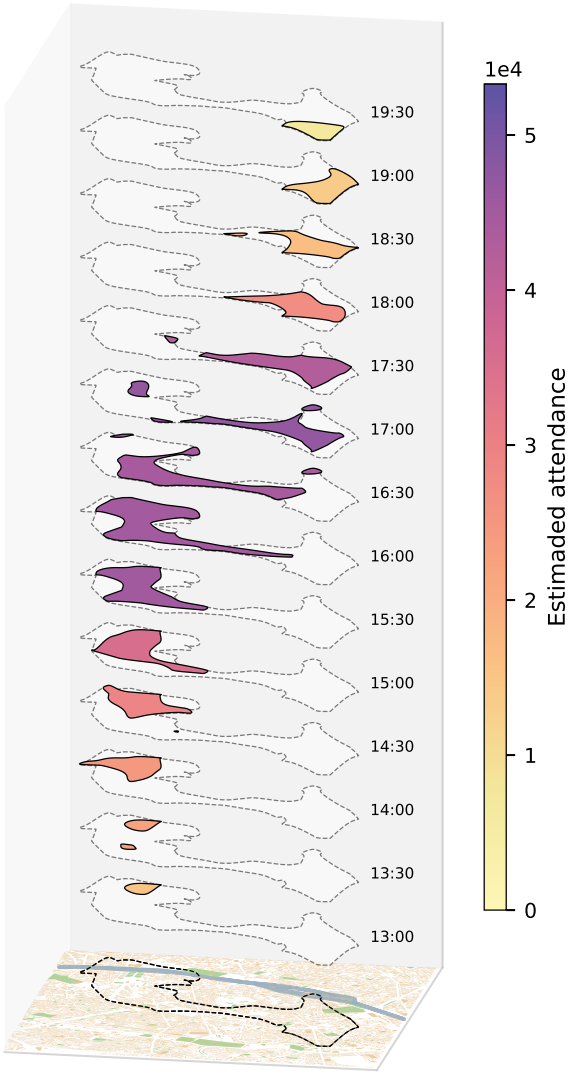


Figure 6.28. Comprehensive characterization of the public demonstration on March 15 in Paris, showing the spatiotemporal reconstruction of the event along with the dynamic estimation of participants.

On one hand, this provides insights that can be useful for operators to respond to these special circumstances, or to local administrations to better organize safety measures in similar future events. On the other hand, these result are preliminary to the development of a full-fledged framework to better understand protests or similar public events while preserving the privacy of the participants.

On that note, it is important to stress that the proposed study targets a-posteriori analysis of large events where tens of thousands of people gather, and (*i*) it does not target live monitoring or prediction and (*ii*) it does not apply to smaller populations that are not sufficient to generate detectable changes in the mobile network traffic demand aggregates. While these are limitations to the capabilities and resolution achievable by our methodology, they also provide inherent protection towards secondary uses of our study aimed at surveillance: building on top of our study to develop an on-line or predictive tool is not straightforward (e.g., the clustering stage cannot be performed live), and in all cases such a tool would not be usable to identify small groups (e.g., tens) of people and even less so individuals, whose impact on the global traffic demand is not detectable from the carrier-level aggregates.

7 Modeling modern mobile traffic for network optimization

Data-driven mobile networks have been an ongoing topic for the past few years and are expected to only grow throughout the deployment of 5G networks and the future deployment of 6G. As models and frameworks proposed to solve tasks, such as network resource optimization and assignment of new deployment, get more complex, their need for quality train and validation data also grows. Also, the network is becoming more and more centered around apps, where the volume of data that specific mobile applications request requires the operator to have special care with specific apps. However, the accessibility of mobile network measurement data sets is very scarce, especially in relation to the app level, which can harm the evaluation of proposed models by the research community. Therefore, a need for the characterization and open models in relation to per-app traffic becomes critical to guide state-of-the-art research.

This Chapter will present a contribution to this problem, proposing a full framework within per-app transport-layer traffic sessions. It will be structured as follows: Section 7.1 will provide an overview of the where models for mobile traffic data generation currently are. Section 7.2 will present the methodology utilized to capture the session-level data; Section 7.3 will present the full characterization of the collected data; Section 7.4 will present the modeling part, which takes a simplistic approach that could be utilized by anyone interested in generating synthetic data based on those models; finally, Section 7.5 will present a few use cases of the models contributed by this Chapter.

7.1. Context of open data traffic models

Data-driven solutions will play an increasingly important role in the 5G mobile network ecosystem during its evolution towards 6G. This trend is stimulated by the unprecedented access to traffic indicators and statistics enabled by a plethora of new network monitoring functions: prominent examples include the Network Data Analytics Function (NWDAF) [253] and Management Data Analytics Function (MDAF) [254] that appeared in 3GPP Release 16, or the database-like Radio Network Information Base

Figure 7.1. Graphic taxonomy of mobile network traffic models, with representative features and typical modeling timescales for models that operate at packet-level, (transport) session-level, and BS-level, respectively.

(RNIB) [255] and the consumer/producer Data Management and Exposure Services [256] for the near-real-time and non-real-time (respectively) RAN Intelligent Controller (RIC) in O-RAN. This abundance of data can feed innovative machine learning models that have been proven to yield promising performance in many complex network management tasks, including forecasting of mobile demand [52] and throughput [257], beam management in mmWave RAN [258], orchestration of network slices [259], classification of flow-level traffic [260] or control of virtualized RAN resources [261], just to cite a few representative examples. Overall, the combination of live data provisioning and learning-based inference is expected to pave the road for paradigms such as Zero-touch Network and Service Management (ZSM) [262] and Intent-Based Networking (IBN) [263].

In the emerging context above, the availability of vast and dependable mobile network data becomes even more critical to the development and evaluation of new network functions across all domains. Unfortunately, access to the large-scale real-world datasets that are needed to train and test original data-driven algorithms is today very limited. Broad measurements from actual production systems at city or national scales that capture the full diversity of mobile traffic demands are hard to come by and are typically protected by restrictive Non-Disclosure Agreements (NDA) that prevent their circulation. Public traffic data is scarce and outdated [156] or gathered via small-scale client-driven experiments whose representativeness is inherently circumstantial [264].

In this scenario, trustworthy models of mobile traffic become an indispensable asset to networking research: they allow generating realistic synthetic traces to remove the data access barrier, and implicitly enable verifiability and reproducibility of results. As illustrated in Figure 7.1 and previously detailed in Subsection 2.1.3, current models of mobile traffic target: (*i*) fine-grained packet-level statistics, e.g., about packet sizes or inter-arrival times [13]; or, (*ii*) aggregate dynamics at individual cellular base station (BS), e.g., describing the total mobile data traffic demand at a given BS over time [61].

In this paper, a different intermediate perspective is taken between those considered in the literature, and explores mobile traffic statistics at the level of *individual transport-*

layer sessions served by one BS. Transport-layer sessions, often also referred to as flows, are sequences of packets belonging to the same application-layer interaction[1] between a UE and a server, and are aimed at provisioning one specific (portion of) service to the UE. They are uniquely identified by a 5-tuple consisting of the transport-layer protocol, source/destination IP addresses, and source/destination ports. For instance, one session may be generated by a user launching the Netflix application on their smartphone to stream an episode of a show, or by a UE retrieving in background a software update for one of its installed applications.

As also portrayed in Figure 7.1, (transport) session-level models target previously overlooked features of mobile traffic: the arrival process of transport-layer data flows of a specific application at a given BS, the duration of such flows, their associated load, or the distribution of average throughput that the combinations of such duration and load statistics entail. Since transport-layer sessions are associated to the one application they serve, session-level models are inherently service-specific. The transport session-level models fill in fact a gap in the space of mobile network traffic modelling, and allow generating for the first time realistic demands aligned with those observed at the BSs of a modern 4G/5G RAN infrastructure. Specifically, they can complement studies on packet-level modeling so as to reproduce fine-grained mobile traffic loads at an individual BS that dependably mimic how the users attached to the target BS request specific services and what amount of traffic each such request entails.

As such, session-level models support the design of data-driven solutions and more credible performance evaluations for many networking tasks, including planning [265], dimensioning with respect to specific services [266], scheduling [267], or energy-efficient operation [268]; they can also inform new traffic generators for modern network simulators [269].

Overall, this Chapter yields the following contributions:

- It characterizes transport-layer sessions recorded at over $282,000$ BSs of a nationwide production mobile network covering continental France, investigating (*i*) the arrival process at individual cellular antennas of sessions associated to a wide range of applications, (*ii*) the distribution of the traffic volume generated by each such session, and (*iii*) the relationship between such load and the duration of the session. This analysis unveils statistical properties of session-level traffic that

[1]It's important to remark that a single application may establish multiple transport-layer sessions. This can happen over time (e.g., a messaging service initiating new sessions every time the user switches to a new chat with a different contact than the current one), or in parallel (e.g., a large file transfer application opening multiple FTP sessions). Multiple transport-layer sessions associated with the same application-layer session may have similar or different characteristics. However, in this chapter the focus is on individual transport-layer sessions only, leaving a thorough investigation of the relationships and interactions of such sessions at the higher layers as future work. Throughout the Chapter, transport-layer sessions will be refereed simply as *sessions*, hence all future references to session-level models implicitly refer to transport sessions.

have not been observed before, and that are heterogeneous across different mobile applications but persistent across space, time and radio access technology.

- It develops simple but accurate models of the statistical properties above for a variety of mobile services, which are released publicly[2] so as to contribute to removing the access barrier to dependable data needed to design and evaluate networking solutions.

- It shows the utility of the proposed models in two practical performance evaluation use cases, which prove how the proposed models substantially enhance the accuracy of the results compared to traffic models currently available for mobile network performance analysis that is not informed by session-level statistics.

7.2. Processing data into session-level statistics

This study builds upon massive measurement data collected in the operational nationwide mobile network of Orange. The target network employs 4G and 5G NSA RAN technologies, which are differentiated following the methodology of Section 3.4.

Consistently with the aim set forth in Section 7.1, it's recorded in the target network data about individual transport-layer sessions observed during 45 consecutive days at the 282,000 BSs that form the whole 4G/5G RAN of the operator. The session-level statistics are produced within secure compute premises of the network operator, and for the purpose of this work, there's only access to distributions and averages that do not contain personal or sensitive information. Next, the collection process and basic features of the data will be detailed.

7.2.1. Aggregation into session-level statistics

The gateway probes collect information about individual TCP and UDP sessions, which are uniquely identified by a 5-tuple consisting of the transport-layer protocol, source and destination IP addresses, and source and destination ports.

A TCP session is typically initiated by the three-way handshake and considered to be terminated shortly after a packet with the FIN or RST bits set is observed. Expiration timeouts that are service-specific are also employed to mitigate the effect of unorthodox TCP session terminations. In case UDP sessions, they start when a new 5-tuple is recorded, and ended once a timeout period without any transmitted packets elapses. Again, this timeout depends on the application that the traffic classification routines associate to the flow. It's worth remarking that, since this study is concerned with sessions served by a single BS, handovers are recorded in the measurement dataset as newly established or concluded transport-layer sessions, respectively.

[2] https://github.com/nds-group/MobileTrafficDists

Data about all sessions occurring at each BS for a given service are initially aggregated at one-minute granularity by the operator, before further processing; the additional transformations are performed to ensure the privacy of the data subjects as well as to strike a balance between a sufficient precision on the traffic representation and a dataset size viable for downstream analysis. Specifically, the data about all sessions occurring at each BS for a target service is aggregated on a daily basis, in the form of (*i*) the number of sessions arriving at the BS at every minute, (*ii*) a PDF of the total traffic volume generated by one session at the BS, (*iii*) value pairs composed of the duration of one session served by the BS and the traffic volume it generates. This is a compact, privacy-preserving representation that allows characterizing all major session-level properties, i.e., the arrival rate, duration, total load, and average throughput.

Formally, sessions occurring at each BS $c \in \mathcal{C}$ for service $s \in \mathcal{S}$ are aggregated over daily intervals $t \in \mathcal{T}$. For each tuple (s, c, t), the following statistics are stored.

- *Counts of sessions served by the BS*, denoted by $w_s^{c,m}$, capturing the total number of sessions received at BS c for service s each minute m of day $t \in \mathcal{T}$, which is further aggregated per day into a variable $w_s^{c,t}$.

- *PDFs of the traffic volume*, denoted by $F_s^{c,t}(x)$, describing the odds that a session of service s induces a total load x at BS c during day t.

- *Value pairs of discretized duration and traffic volume*, denoted by $v_s^{c,t}(d)$, capturing the mean load associated to sessions of duration d for service s at BS c in day t.

7.2.2. Statistics averaging

The dataset reports statistics per BS and day. For the analyses, there's a need to investigate behaviors averaged over multiple BSs and days. For duration-volume pairs, weighted average of each datapoint is computed; for instance, average pairs over all BSs and days for a service s are obtained as

$$v_s(d) = \frac{1}{\sum_{c\in\mathcal{C}} \sum_{t\in\mathcal{T}} w_s^{c,t}} \sum_{c\in\mathcal{C}} \sum_{t\in\mathcal{T}} w_s^{c,t} v_s^{c,t}(d), \ \forall d. \tag{7.1}$$

In the case of traffic volume PDFs, averaging is achieved via a finite-dimensional general mixture model. For an all-BS and all-day average PDF, this is expressed as

$$F_s(x) = \frac{1}{\sum_{c\in\mathcal{C}} \sum_{t\in\mathcal{T}} w_s^{c,t}} \sum_{c\in\mathcal{C}} \sum_{t\in\mathcal{T}} w_s^{c,t} F_s^{c,t}(x). \tag{7.2}$$

The expressions in (7.1) and (7.2) are straightforwardly extended to any subsets of $\mathcal{C}$ and $\mathcal{T}$, so as to merge statistics from any set of BSs and days. Illustrative samples of

nationwide traffic PDFs $F_s(x)$ and duration-volume pairs $v_s(d)$ averaged over all BSs and days are later reported in Figure 7.4.

7.3. Characterizing session-level demands at cellular BSs

Focus now will be in exploring the dataset and provide both qualitative and quantitative characterizations of session-level mobile traffic demands. The derived insights will inform the design of the proposed models to be introduced in Section 7.4.

7.3.1. Analysis of the arrival of sessions

This portion starts by analyzing the arrival process of sessions at a BS. Figure 7.2 reports the distribution of the number of new sessions established at every minute at different categories of BSs, i.e., the PDF of $w_s^{c,m}$ at all BSs $c \in \mathcal{C}_i$ of category i, and aggregated over all services $s \in \mathcal{S}$. The x-axis values are then normalized by the cardinality of set $\mathcal{C}_i$, so as to obtain the typical number of sessions arriving in one minute at a single BS of category i. Namely, categories $i \in \mathcal{I}$ tell apart BSs experiencing different loads: the distribution of total traffic served by each BS is computed during the whole measurement time, and separate BSs based on the decile they pertain to. Thus, each set $\mathcal{C}_i$ includes 10% of the BSs, with growing mobile traffic demands from the first decile to the last one. The rationale for this categorization is that it allows observing how the session arrival process is affected by the target BS load.

In fact, the plots in Figure 7.2 show that the behavior of the arrivals is semantically similar across all classes of BSs, or, equivalently, *the traffic volume served by the BS has no significant impact on the high-level statistics of the arrival process*. Indeed, the shape of the overall distribution is the same for all plots, apart from the obvious difference of scale in the abscissa induced by the growing demand across deciles. More precisely, *all PDFs of values $w_s^{c,m}$ show an evident bi-modal distribution*, which a close inspection reveals to be due to the well-known circadian rhythm of mobile network traffic, with low traffic (hence small number of sessions per minute) overnight and much increased demands (hence more frequent session arrivals) during daylight hours. Transitions between these two phases are very rapid, which leads to a negligible probability of having intermediate arrival rates.

Since sessions are naturally service-specific, a follow-up question is how such arrivals are distributed across different mobile services. Figure 7.3 offers a first result in that sense, ranking the top 100 services based on the fraction of total sessions they generate. The curve follows a negative exponential law (with a very high coefficient of determination R^2 of 0.97), implying that *the number of sessions generated by each service is very heterogeneous*: the top 20 services are responsible for over 78% of the sessions recorded overall. The imbalance is less dramatic than that in traffic, which is known to follow

Figure 7.2. Real: measurement PDFs of the per-minute session arrival rate for antennas serving different loads. Nonpeak and peak: fitted distributions modelling the bi-modal sessions arrivals (see in Section 7.4.1 for full details).

Figure 7.3. Services ranked by the fraction of sessions they generate, along with their normalized total traffic.

(a) Netflix **(b)** Twitch **(c)** Deezer

(d) Amazon **(e)** Pokemon Go **(f)** Waze

Figure 7.4. Probability density functions of the traffic volume $\boldsymbol{F_s(x)}$ (top plots in each subfigure), and value pairs of discretized duration $\boldsymbol{d}$ and traffic volume $\boldsymbol{v_s(d)}$ (bottom plots in each subfigure) for a selection of mobile services. PDFs and duration-traffic pairs are aggregated over working days (Monday through Friday) and weekends (Saturday and Sunday) separately.

a more skewed power law [34], [39]; nonetheless, it suggests that the probability that a newly established session belongs to a given application is far from uniform.

7.3.2. Qualitative analysis of session-level traffic

Figure 7.3 shows the total normalized traffic produced by each service. While some correlation with the number of sessions exists, the load dots are fairly scattered (on a logarithmic scale) for similarly ranked services: hence, *different applications entail a very varied traffic volume per session.* This motivates an investigation of such session-level traffic dynamics on a per-service basis, which is indeed the target of the next analysis.

Samples of session-level traffic volume PDF and duration-traffic pairs are portrayed in Figure 7.4 for six representative mobile services. All statistics are averaged over the

whole set of BSs and days, using the methodology described in Section 7.2.2, hence they capture archetypal behaviors of the demands of each application.

A first qualitative observation is that *total traffic volumes and duration values are highly heterogeneous, among sessions of a same service and even more so across different services*. Indeed, intra-service statistics show how sessions belonging to a same application can generate very diverse traffic volumes, spanning several orders of magnitude, over intervals that can range from seconds to hours; and, the shapes of PDFs and duration-traffic pairs are completely different at inter-service level. By looking at each subfigure, it can be noted that the traffic volume PDFs present multi-modal shapes with an overall smooth Gaussian-like trend (over the logarithmic abscissa) interrupted by abrupt and marked spikes of probability. Both the main statistics (such as the mean, standard deviation or skewness) and the probability peaks are not comparable across the selected services. Interestingly, *heterogeneous probability peaks also tell apart applications that ostensibly belong to the same class*, e.g., messaging services like Snapchat and Whatsapp, or video streaming services like Netflix and YouTube.

A closer look to the PDFs of each service reveals unique facets linked to the nature and usage of the mobile application. For instance, Netflix, a platform for movie streaming, has a clear mode around 40 MB, and a drop of probability just after the 200 MB mark. When a user is connected to a mobile network, Netflix adopts an automatic balancing of data usage and video quality, allowing 4 hours of playback per GB of data in typical cases. In this setting, the first peak occurs at around 10 minutes of streaming, and the drop after around 50 minutes: both values are consistent with intuition, as they match the duration of one short episode of a series and a full episode of a longer show.

The session-level traffic dynamics change substantially when looking at a different video streaming service, i.e., Twitch, which, unlike Netflix, focuses on live content. The main mode, around 20 MB, and the main knee, at 800 MB, are shifted to the right; also, the amount of traffic per minute is much higher. The data indicates that Twitch users engage in long sessions with a high bitrate, suggesting that live streams tend to be consumed in more stationary conditions than on-demand movies.

Another example is Deezer, a popular audio streaming service, which shows two main traffic modes that map to the highest probability values: one is located around 3.5 MB and the other at 7.6 MB. At the standard bit rate of 128 kbit/s [270], the two modes translate to 3:40 minutes and 8:00 minutes of listening time, respectively. These roughly match the duration of one and two songs, including advertisements: according to the data, Deezer users most often listen to a couple of tunes while connected to a same BS, and longer listening times, while possible, are less likely.

Applications that mainly rely on relatively short message exchanges, such as Amazon (an archetypal web browsing service), Pokemon Go (a popular location-based game) or Waze (a navigation service generating floating car data), show a completely different

behavior than streaming services. Loads per session are much lower, with traffic PDFs flattening to a zero value early on. Yet, the distributions and duration-traffic pairs are completely different also among these applications, highlighting once more the unique behavior exhibited by diverse services at the session level.

It is worth recalling that, in all PDFs, the duration of a session and the volume of traffic it generates are not only the result of the application or user's behavior, but also of the UE mobility. Indeed, *many sessions of mobile users occur only in part within a same BS, and generate a smaller-than-expected volume of traffic for a complete sessions of the same application*. This explains the presence of many very short sessions generating reduced traffic loads in the left part of the distributions of all services. Also, it allows interpreting the main mode of a streaming service like Netflix, which matches 3 MB and less than one minute of content: this is a reasonable mean dwell time in the BS for in-transit UEs running the application. Although frequent and thus important for a credible evaluation of mobile network performance, transient sessions have been ignored by traffic models proposed in the literature so far.

7.3.3. Quantitative analysis of session-level traffic

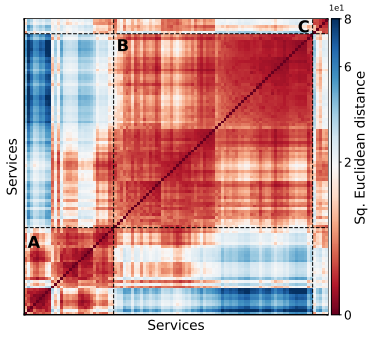
The qualitative analyses above unveil interesting aspects of session-level mobile traffic dynamics, which are however based on a close inspection of a few representative cases. To substantiate these observations, a quantitative study of the traffic volume distributions $F_s^{c,t}(x)$ is performed; it's considered for now data aggregated over all BSs $c \in \mathcal{C}$ and days $t \in \mathcal{T}$, and compared different services s, i.e., the PDFs $F_s(x)$ from (7.2), as per the following steps.

(i) $F_s(x)$ is normalized, for each service s, so that all PDFs have zero mean. This removes the impact of the sheer volume of traffic generated by each application, enabling a comparison of less obvious dynamics, such as the standard deviation or the modes of the PDFs.

(ii) Pairwise Earth Mover Distance (EMD) [271] is computed among normalized $F_s(x)$, compiling a similarity matrix.

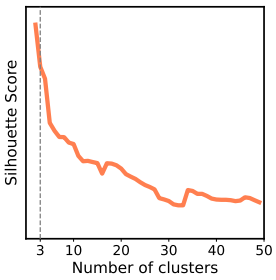
(iii) A centroid hierarchical clustering algorithm [272] is ran on the similarity matrix, so as to identify classes of services characterized by similar PDFs. This algorithm iteratively groups the two PDFs at minimum distance, computes their average via (7.2), adds it to the set of PDFs in place of the original pair, and recomputes distances from the aggregate to all other PDFs in the set. By doing so, it builds a hierarchy of PDFs based on their similarity.

The result of this process is summarized in Figure 7.5a. Three main clusters emerge;



(a) Similarity matrix: Traffic

(b) Similarity matrix: Duration



(c) Silhouette Score: Traffic

(d) Silhouette score: Duration

Figure 7.5. Similarity matrix of the normalized (a) PDFs $\boldsymbol{F_s(x)}$ and (b) pairs of duration d and traffic volume $v_s(d)$, for all services with the three major clusters highlighted. Distance values closer to zero (dark red) indicate more similar PDFs. Associated Silhouette scores for the (c) PDFs and (d) pairs of duration.

by looking at the services in each, it's possible to relate these groups to (A) streaming services, (B) low-duty-cycle services relying on short messages, and (C) outliers.

A similar hierarchical clustering process is performed on the set of duration-traffic pairs $v_s(d)$, swapping EMD for the Squared Euclidean Distance (SED) across services, which is more fitting for a *non-probability* function and avoids an additional square root needed for the traditional Euclidean distance, which becomes unnecessary when comparing distances. The results for this process are seen on Figure 7.5b, where once again 3 clusters emerge following the same pattern seen before on Fig. 7.5a: (A) Streaming services, (B) low-duty-cycle services and (C) outliers. Although differences across different $v_s(d)$ are not as visual as $F_s(x)$, it's not noted across applications that $v_s(d)$ has a behavior similar to a power law function, where streaming services that laid on cluster (A) tend to an exponent > 1, while HTML and messaging services on cluster (B) have their exponent $<< 1$. This behavior is further explored and exemplified when modeling $v_s(d)$ in subsection 7.4.3.

The emergence of two major behaviors in both clusterings is aligned with early observations in Section 7.3.2 about the difference between the dynamics of streaming applications like Netflix, Twitch and Deezer and those of less demanding services like

(a) Facebook Live **(b)** Facebook

Figure 7.6. Traffic volume PDFs $\boldsymbol{F_s(x)}$ (top) and duration-traffic pairs $\boldsymbol{v_s(d)}$ (bottom) for two applications with shared user base: (a) Facebook Live and (b) Facebook.

Amazon, Pokemon Go or Waze. It also confirms that this polarity does not depend on the user base but it is inherent to the nature of the service. Indeed, as shown in Figure 7.6, it affects services like Facebook Live (video streaming, cluster A) and Facebook (social media, cluster B), which have a largely common user population: the former has $F_s(x)$ and $v_s(d)$ aligned with that of the streaming applications in Figures 7.4a–7.4c, whereas the latter has flattened-out PDF and low-bitrate pairs as in Figures 7.4d–7.4f. It's concluded that *session-level traffic is marked by a main dichotomy between video and audio streaming services and applications that rely on short or lightweight message exchanges.*

However, clustering services beyond the two major groups above is not possible. Figures 7.5c and 7.5d show the evolution of the Silhouette score [218] over progressive splits of the services into a growing number of clusters. This index is widely used to identify meaningful clustering levels, where values closer to 1 indicates no overlaps and zero indicates overlapping clusters; the ideal cluster level is identified by a major drop of the score in the following level, as this indicates that breaking down the set into more classes generate significant overlap. Apart from the substantial change of value after the first 3 clusters, the Silhouette score stays nearly flat for all subsequent splits: finer-grained grouping of services is haphazard and does not reveal any informative pattern. Therefore, apart from a very macroscopic separation of streaming versus non-streaming traffic, *session-level statistics of mobile traffic demands cannot be characterized for whole classes of applications but must be studied for specific services independently.*

(a) Traffic across space/time **(b)** Traffic across RATs

(c) Duration across space/time **(d)** Duration across RATs

Figure 7.7. (a,c) Boxplots of differences in session-level traffic for (***i***) different services, and for each service across (***ii***) working days and weekends, (***iii***) urban, semi-urban and rural regions, and (***iv***) different cities. (b,d) Boxplots of differences in session-level traffic (***i***) for the same service across 4G and 5G RATs, and for difference services relying on (***ii***) 4G or (***iii***) 5G. Differences PDFs $\boldsymbol{F_s^{c,t}(x)}$ in (a,b) are computed via EMD, while those between pairs $\boldsymbol{v_s^{c,t}(d)}$ are computed using SED. Whiskers indicate the **5**-th and **95**-th percentiles, while the boxes outline the first, second and third quartiles.

7.3.4. Impact of space, time and technology

The analysis is now broken down over the temporal and spatial dimensions, by looking at PDFs $F_s^{c,t}(x)$ and pairs $v_s^{c,t}(d)$ that are not aggregated over all BSs $c \in \mathcal{C}$ and days $t \in \mathcal{T}$.

In the time dimension, mobile traffic workloads are known to differ primarily between working days and weekends [22], hence it will be explored if the same distinction exists in session-level dynamics. New aggregations of $F_s^{c,t}(x)$ and $v_s^{c,t}(d)$ are generated, over all BSs c but telling apart two sets of days: working days and weekends. Then, the EMD between the two traffic volume PDFs of a same service s for the two types of days is computed. EMD is symmetric and compares a pair of PDFs by calculating the minimum cost of displacing samples of one distribution to match the other, returning a value zero for identical PDFs. For duration-traffic pairs the SED of value vectors is used.

The distribution of these EMD and SED values is condensed in Figures 7.7a and 7.7c, under the 'Days' tag. As a reference, the distances between different services is reported, i.e., the values in the matrix of Figure 7.5a, under the 'Apps' tag. By comparing the two boxes, it is evident that *the dynamics observed for a same service yield negligible differences across working days and weekends*, whereas inter-service heterogeneity is much more pronounced. Visual examples of the lack of impact of the day type on session-level traffic are also in Figure 7.4 and Figure 7.6, where measurements collected in workdays and weekends does not show clear differences.

From a spatial perspective, $F_s^{c,t}(x)$ and $v_s^{c,t}(d)$ are aggregated for each service s, over all days t but separately over BSs that belong to different regions and cities. At a region level, PDFs and pairs for BSs will be computed from location into (i) dense urban, (ii) semi-urban and (iii) rural regions; urbanization level information is employed, from the data provided by the local national institute for statistics to tell apart the three types of regions. Concerning cities, statistics are derived for each of the 5 largest metropolitan areas in the country.

The test previously used for different days of the week is then repeated, by calculating for a same service the EMD of the traffic volume PDFs and the SED of the duration-traffic pairs among different regions as well as among diverse cities. The results are reported in Figures 7.7a and 7.7c, under the 'Regions' and 'Cities' tags. Again, distances are very small when confronted to those that affect diverse services under the 'Apps' tag. It can be concluded that *the geographical location of the BS has very limited impact on the session-level traffic statistics, hence a single model would generalize well across urbanization levels.*

Finally, the impact that different RAT have on the session-level statistics is investigated. For each service, it will be computed separately traffic volume PDFs and duration-traffic pairs for all sessions served by 4G eNodeBs and 5G gNodeBs. This allows studying if the statistics of a same application change when a user is connected via 4G or 5G. The result is reported in terms of EMD and SED in Figures 7.7b and 7.7d, under the 'RATs' tag, and shows that the diversity entailed by different RATs is negligible if compared to that determined by the service itself. The latter is reported in Figures 7.7a and 7.7c, under the 'Apps' tag, but is also broken down by technology in Figures 7.7b and 7.7d, under the 'Apps (4G)' and 'Apps (5G)' tags: there, it can be observed that the difference across applications remain stable no matter if those are served by 4G and 5G BSs. It can be concluded that *RATs do not impact in a significant manner the way users consume a same mobile service within a single transport-layer session.*

7.3.5. Key insights from the characterization of session-level traffic

The characterization of session-level traffic yields a number of takeaways relevant to modeling, as summarized below.

A) The arrival rates of newly established sessions at a given BS follow a bi-modal distribution, independently of the load served by the BS, hence a same modelling strategy can be applied to arrival processes of all BSs.

B) The fraction of total sessions generated by each service is not uniform, but follows a negative exponential law, calling for a suitable breakdown of arrivals on a per-service basis.

C) Services are characterized by unique multi-modal distributions of per-session traffic volume, which present varied probability peaks at specific load values. Apart from a broad distinction between streaming and best effort services, applications cannot be grouped on the basis of class or using statistical clustering methods: each service requires dedicated session-level modeling of the load and duration of the session they induce.

D) The statistics of session-level traffic and duration of a given service do not vary significantly across days, urbanization level, metropolitan areas or RATs; hence, a single model suffice to represent the dynamics of a service at a BS.

E) Transient, partial sessions generated by users crossing the BS coverage area for a short time period occur with significant frequency and should be properly modeled.

7.4. Obtaining models for session-level traffic

The insights above are built upon to develop original models of mobile network traffic at the session level. Insights A and B offer pointers on how to model arrivals of sessions $w_s^{c,m}$ at one BS. The remaining ones provide indications on the modeling of the traffic volume PDFs $F_s^{c,t}(x)$ and duration-traffic pairs $v_s^{c,t}(d)$. Specifically, insight C implies that dependable models need to target each service $s \in \mathcal{S}$ separately. However, following insight D, these per-service models do not need to be further specialized for individual BSs $c \in \mathcal{C}$ located in different regions and cities, for different days of the week $t \in \mathcal{T}$, or even across 4G and 5G NSA RATs. Ultimately, models of the aggregate $v_s(d)$ from (7.1) and $F_s(x)$ from (7.2) are enough to capture typical session-level mobile traffic reliably.

Based on these considerations,the following modeling approaches for $w_s^{c,m}$, $F_s(x)$, and $v_s(d)$ will be adopted, respectively.

- For session arrivals $w_s^{c,m}$, it will be used simple fittings of theoretical distributions on the bi-modal PDFs observed in Section 7.3.1, using a constant measurement-driven breakdown to associate each arrival to a specific service s.

- For the traffic volume PDFs $F_s(x)$, a novel algorithm is presented to decompose and approximate the distributions as log-normal mixture models. The proposed model achieves good estimation of the original $F_s(x)$ for a wide set of services s with a small set of components (hence parameters). The approach operates over the full PDF domain, thus including short-lived transient sessions and abides by insight E.

- For duration-traffic pairs $v_s(d)$, it will be shown that a regression using a power law model fits well all services s. Interestingly, these models enables comments on

how throughput varies non-linearly with the duration of a session, in ways that are unique to each service.

7.4.1. Fitting of session arrivals

Based on the analysis carried out in Section 7.3.1, the peak daylight arrival process of sessions at a BS and its off-peak nighttime counterpart separately will be modeled. This gives a degree of freedom in emulating either day or night traffic.

By looking at Figure 7.2, the mode during peak hours can be described by a simple Gaussian distribution. The mean $\mu^{c,w}$ of the Gaussian fitting is necessarily different across classes of BSs characterized by different loads, which observe diverse arrival rates: it ranges from 1.21 sessions/minute for the first decile class up to 71 sessions/minute for the busiest BS decile. For the standard deviation $\sigma^{c,w}$, a pattern is observed emerging across all classes of BSs, such that $\sigma^{c,w} \sim \mu^{c,w}/10$ in all cases: this allows automating the setting of $\sigma^{c,w}$ and simplify the models. The second mode, representing off-peak hours, is better modeled by a Pareto distribution, represented by:

$$b^{c,w} \cdot (s^{c,w})^{b^{c,w}} / x^{b^{c,w}+1}, \tag{7.3}$$

where $[b^{c,w}, s^{c,w}]$ are the shape and scale parameters, respectively. The measurement data is well fitted by fixing the shape to $b^{c,w} = 1.765$ and modify only the scale $s^{c,w}$ across antennas. In fact, the growth of $\mu^{c,w}$ and $s^{c,w}$ across BSs in increasing load decile classes is similar, i.e., exponential with akin rate. Examples of the resulting fittings are also shown in Figure 7.2.

According to the results of Sections 7.3.2 and 7.3.3, it is important to model arrivals associated to different services, which are not uniform. A simple yet effective way to break the aggregate arrival distributions above on a per-service basis is opted. This approach stems from the consideration that the share of sessions induced by each service is relatively constant across different BSs and over time. Specifically, Table 7.1 presents the expected fraction of sessions and traffic volume generated by 28 popular mobile applications. The table also report the corresponding Coefficient of Variation (CV), i.e., the ratio of standard deviation to the mean, across BSs and minutes. The CV thus represents the expected diversity of session and traffic shares yielded by each service. While the CV of the traffic share tends to fluctuate, that of the session share is fairly stable at around 1% across applications. In light of this observation, the session shares in Table 7.1 are used as probabilities to assign to a specific service a newly established session obtained from the fitted arrival rate PDFs.

Service	Sessions %	(CV)	Traffic %	(CV)
Facebook (FB)	36.52	±1.15	32.53	±1.68
Instagram	20.52	±1.27	31.48	±2.13
SnapChat	18.33	±1.17	9.52	±2.12
Youtube	4.94	±1.14	0.24	±1.39
Google Maps	2.76	±1.14	0.10	±2.82
Netflix	2.40	±1.29	11.10	±1.66
Waze	1.63	±1.39	0.62	±1.75
Twitter	1.46	±1.43	0.45	±1.49
Apple iCloud	±1.45	1.04	3.24	±4.20
FB Live	1.42	±1.17	1.80	±1.08
Spotify	1.12	±1.28	0.12	±2.54
Deezer	1.08	±1.91	1.59	±1.81
Amazon	0.96	±1.17	0.25	±1.11
Twitch	0.91	±1.22	3.67	±0.96
WhatsApp	0.85	±1.27	0.41	±2.91
Clothes	0.83	±1.23	0.85	±1.58
Gmail	0.54	±1.16	0.02	±1.17
LinkedIn	0.51	±1.23	0.54	±1.41
Telegram	0.44	±1.16	1.08	±3.27
Yahoo	0.32	±1.18	0.10	±2.40
FB Messenger	0.23	±1.25	0.01	±1.85
Google Meet	0.22	±1.11	0.14	±2.16
Clash of Clans	0.18	±1.25	0.09	±3.31
Microsoft Mail	0.11	±1.31	0.01	±4.48
Google Docs	0.09	±1.21	0.02	±3.58
Uber	0.07	±1.92	0.01	±1.55
Wikipedia	0.06	±1.30	0.01	±3.01
Pokemon GO	0.04	±1.21	0.01	±2.33

Table 7.1. Percent contribution to the total number of transport-layer sessions and to the total mobile traffic volume, for 28 applications and with associated CV.

7.4.2. Log-normal mixture models of traffic

The modeling approach for $F_s(x)$ is in three steps, which are illustrated in Figure 7.8 for one representative service, i.e., Netflix. In the first step, exemplified in Figure 7.8a, the experimental $F_s(x)$ is fitted using a log-normal distribution, i.e.,

$$LogN(x;\mu_s,\sigma_s^2) = \frac{1}{\sigma_s\sqrt{2\pi}} \cdot \exp\left(-\frac{(\log_{10} x - \mu_s)^2}{2\sigma_s^2}\right), \tag{7.4}$$

which allows representing the broad trend of session-level traffic volume for each service s, denoted by $f_s(x)$. The rationale behind the choice of a log-normal fit is that it is the single function best representing the whole $F_s(x)$ for the vast majority of services: indeed, it can be observed in all plots of Figure 7.4, Figure 7.6 and Figure 7.8a that the PDFs yield a resemblance to Gaussian-like shapes when the traffic is represented in a logarithmic scale. In this stage, the fitted PDF $f_s(x)$ is also subtracted from the measurement PDF $F_s(x)$, bounding the result to positive values and obtaining a residual probability.

The second step, depicted in Figure 7.8b, focuses on analyzing the residuals; these represent the unique peaks of session-level traffic of each service and are thus instrumental to a realistic modeling of $F_s(x)$. The process is automated as follows.

(a) Main component and residuals

(b) Residual selection

(c) Final model

Figure 7.8. modeling steps for the log-normal mixture model of the traffic volume PDF $\boldsymbol{F_s(x)}$, for a sample service, i.e., Netflix. (a) Decomposition of the measurement distribution (light blue) into a main log-normal component (dashed) and residual probability peaks (red). (b) Identification and characterization of the residuals to be modeled (light grey areas), using their first derivative (orange). (c) Final residual components used by the mixture model (red), and reconstructed PDF $\boldsymbol{\tilde{F}_s(x)}$ (black).

- The first derivative of the residual is computed, using a first-order Savitzky-Golay filter [273] that smooths the resulting curve and helps the subsequent steps.

- The derivative are checked against a threshold,[3] and record all continuous intervals of traffic values within which the derivative stays seamlessly above the threshold.

- The aforementioned intervals are ranked based on the residual probability they contain, simply computed as the integral of the residual curve within each interval.

This method employs the change rate of the derivative to single out the residual peaks of actual interest for the modeling process; these are characterized by a high rate of change over a short traffic interval, such as the (zoomed-in) light grey regions identified by the algorithm in Figure 7.8b.

In the third and final step, the retained residual peaks are modeled. As those resemble low-variance Gaussian PDFs in log scale, the n-th peak is represented as a log-normal

[3]Upon extensive tests, it was found that the algorithm to be robust to the choice of the derivative threshold, which avoids misinterpreting tiny oscillations as peaks. This allows using a same value, i.e., 10^{-5}, to model any service s.

function

$$f_{s,n}(x) = k_{s,n} \cdot LogN(x; \mu_{s,n}, \sigma^2_{s,n}), \tag{7.5}$$

where $LogN(\cdot)$ is defined in (7.4). $\mu_{s,n}$ is set to the traffic value with maximum probability in the associated interval, so as to properly center $f_{s,n}(x)$; $\sigma_{s,n}$ is then set to $(0.997 \cdot \ell_{s,n})/3$, where $\ell_{s,n}$ is the span if the n-th interval, so that 99.7% of the modeled probability lays inside the interval. Finally, $k_{s,n}$ is the residual probability used to rank the intervals, and allows scaling the log-normal distribution. Samples of modeled residuals are in Figure 7.8c for the case of Netflix.

The final mixture model for a service s, denoted by $\tilde{F}_s(x)$, is obtained by composing the main and residual functions:

$$\tilde{F}_s(x) = \frac{f_s(x) + \sum_{n=1}^{N} f_{s,n}(x)}{1 + \sum_{n=1}^{N} k_{s,n}}, \tag{7.6}$$

where N is the number of modelled residual peaks during the third step above, and the normalization factor at the denominator ensures that the expression in (7.6) is a distribution. Figure 7.8c provides an example of real $F_s(x)$ and its modeled counterpart $\tilde{F}_s(x)$, for the Netflix service.

To conclude, it can be noted that other approaches to derive $\tilde{F}_s(x)$ are possible, e.g., using traditional mixture models that automatically find the best decomposition of a PDF into multiple distributions of a given type. With respect to such alternative solutions, the algorithm not only produces models that are compact and accurate, but outputs components with a clear semantic (i.e., the main trend, and a set of characteristic peaks), easing results explainability.

In this regard, it is worth noting that, when applied to the measurement data, the procedure identifies and models at most 3 residual peaks for the majority of services; the rare additional peaks have negligible weight $k_{s,n}$ below 10^{-4}. Therefore, all models are aligned and avoid irrelevant components, by limiting the maximum number of residual contributions to 3.

7.4.3. Power-law fitted models between session duration and volume

Value pairs of duration d and traffic volume $v_s(d)$ tend to align into very consistent patterns, as shown by the examples in Figure 7.4 and Figure 7.6: therefore, the relationships between the duration of a session and the load it generates are clearly not random, but follow statistical trends. Longer sessions are largely associated to higher traffic volumes, which is reasonable. Yet, the exact growth patterns are quite different across applications, as also observed in the figures above.

In order to properly represent the expression of $v_s(d)$ for each mobile service, the data is fitted to varied functions from a range of families. Upon experimenting with polynomial,

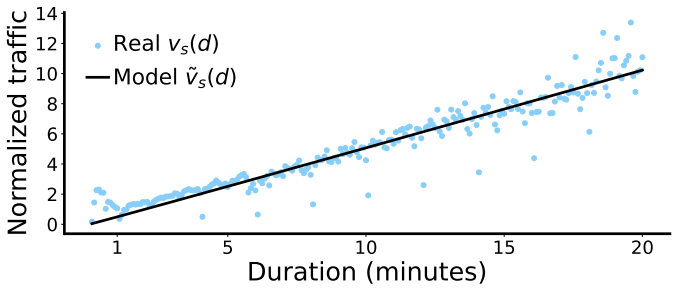


Figure 7.9. Relation between traffic consumption and duration of the session, for the service Netflix (linear scale on both axis). In blue, the fitted powerlaw function.

exponential, and power laws it was found that the latter yield the best quality of fitting across all services, while limiting the model complexity. Specifically, the following power-law model was obtained:

$$\tilde{v}_s(d) = \alpha_s \cdot d^{\beta_s}, \tag{7.7}$$

which will be unique for each application, by fitting $\{\alpha_s, \beta_s\}$ via the Levenberg-Marquardt non-linear least squares method. An example of the fit for the Netflix service can be seen on Figure 7.9, where the black lines represent the fitted $\tilde{v}_s(d)$ for the measured data.

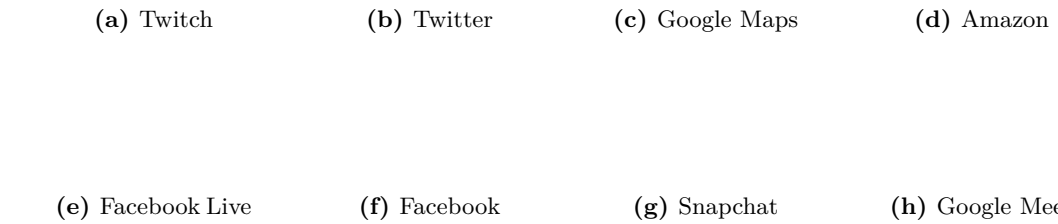
The low complexity of the power law model facilitates its explainability, and in particular it allows quantifying the diversity of behaviors in $v_s(d)$ discussed before. The fitted exponent β_s is especially revelatory in that sense. In a linear model where $\beta_s = 1$, then $\tilde{v}_s(d) = \alpha_s \cdot d$, and all sessions experience an average throughput α_s independently of their duration. A super-linear $\beta_s > 1$ denotes sessions whose mean throughput increases as they last longer, and a sub-linear $\beta_s < 1$ indicates that the instantaneous demand decreases for longer sessions.

Figure 7.10 shows the value of β_S for a representative subset of mobile services. The exponent spans a wide range of values, from 0.1 to 1.8, so each application has quite different scaling of the average throughput to the session duration. Interestingly, when looking at the super- or sub-linearity of the models, video streaming services dominate super-linear behaviors. It's speculated that this may be due to the fact that longer sessions within the same BS are generated by more stationary users, who also enjoy higher video bitrates thanks to a more stable and strong radio signal. Non-video applications have a sub-linear evolution of $\tilde{v}_s(d)$, as most require user interactions that tend to be less steady over longer periods.

7.4.4. Model quality and usage

Overall, session-level traffic models are generated as presented above for 31 mobile services, including all those listed in Table 7.1. The accuracy of the models for $\tilde{F}_s(x)$

Figure 7.10. Power law exponents of the fitted $\tilde{v}_s^{c,t}(d)$ for a subset of services. Coefficients R^2 are in bold.



(a) Twitch **(b)** Twitter **(c)** Google Maps **(d)** Amazon

(e) Facebook Live **(f)** Facebook **(g)** Snapchat **(h)** Google Meet

Figure 7.11. $\tilde{F}_s(x)$ and $\tilde{v}_s(d)$ (black solid lines) against measurement data (light blue) for a choice of services.

and $\tilde{v}_s(d)$ presented above is assessed by means of standard tests. For the traffic volume PDFs, the error of the models are computed by calculating its EMD with respect to the original measurement-based $F_s(x)$. Although the absolute value of EMD is not easily contextualized, in all cases results are obtained in the order of 10^{-5}, hence one order of magnitude lower than those recorded in the various tests on $F_s(x)$ in Figure 7.7a: this is considered a good indication of the fidelity of the models. In the case of duration-volume pairs, the coefficient of determination R^2 are computed as a measure of the quality of fit. The values are typically in the 0.7–0.9 range, which denotes a reliable fitting; in some cases, values as low as 0.5 were obtained, which are still reasonable and, upon close inspection, are mainly due to noisy measurement data that creates outliers. Examples of the R^2 values are on top of each bar in Figure 7.10. Finally, visual illustrations of the real data and models are provided in Figure 7.11 for a subset of services, and show the good resemblance of $\tilde{F}_s(x)$ and $\tilde{v}_s(d)$ with the measurements.

Each model is fully characterized by a tuple of parameters $[\mu_s,\sigma_s,\{k_{s,n},\mu_{s,n},\sigma_{s,n}\}_n,$ $\alpha_s, \beta_s]$, which are release publicly. This allows reproducing realistic session-level statistics for the traffic volume (extracted from $\tilde{F}_s(x)$), duration (obtained by applying the inverse

function $\tilde{v}_s^{-1}$ to the traffic volume) and average throughput (computed as the ratio of the volume to the duration). These open models can thus benefit the research community by empowering more dependable performance evaluations of mobile network systems and solutions, as demonstrated next in practical use cases.

7.5. Creating synthetic data from session-level models

This section describes three use cases that showcase the critical impact that accurate session-level per-service traffic modeling can have in network management. It can be remarked the goal is not to derive innovative solutions for these use cases, but rather providing examples that illustrate, through simple network management scenarios, the utility of the presented models with respect to more traditional, simpler and not data-informed modeling of traffic. The first use case highlights the importance of per-service traffic characterization; the second one shows the benefits of session-level modeling; finally, the third one provides a general use of replicating network topology and different throughput demands from both session-level and duration models.

7.5.1. Capacity allocation for network slicing

Correct characterization of mobile traffic demand is crucial for resource allocation in network slicing: An accurate knowledge of the expected traffic demand for each Service Provider (SP) requesting a network slice would allow the operator to adjust the reserved resources in a more efficient manner, considerably increasing its profit margin by reducing operating expenses for each slice while accommodating more slices.

This first example considers a scenario in which the operator signs a Service Level Agreement (SLA) with each one of the 28 SPs included in Table 7.1. Each SP acquires a network slice to guarantee its traffic demand during peak hours (i.e., except night from 10pm to 8am). The incoming sessions are sampled from the real data distribution, such that the share of traffic and number of sessions of each service follows the values indicated in Table 7.1, and the arrival time of the sessions is modeled so the number of arrived sessions per minute at each RU follows the distribution in Section 7.4.1. The terms of the SLAs are satisfied if the operator successfully delivers all the traffic demand from the SP's users at least 95% of the time. In this setting, the operator must decide how much capacity it allocates to each of the SPs at each of the antennas.

Algorithms: First are considered the derived models for the sessions' arrival time, the traffic per session, and the session's duration to determine the capacity allocation of each slice. Based on these models, it's possible to obtain the CDF of the traffic per service per antenna for different levels of demand. Considering this CDF and the average antenna load, the allocation can be done to each slice with the capacity that corresponds to its 95th percentile.

	Time with no dropped traffic (%)	Standard deviation
Model	95.15%	2.1%
BM A	89.8%	4.3%
BM B	87.25%	4.2%

Table 7.2. Performance results for capacity allocation for network slicing averaged over antenna and service.

This approach, which is only feasible with the derived session-level results, will be compared with two benchmarks. For that, consider the mobile traffic models available in the literature [274], [13] that provide shares of mobile traffic for 3 service categories (Interactive Web (IW), Casual Streaming (CS) and Movie Streaming (MS)). Through a literature search, no available models with higher levels of service specification are found. Thus, it can be considered as benchmarks BM A, which considers the three mentioned categories with the session shares derived from aggregating the corresponding values of Table 7.1 (IW: 49.30%, CS: 48.46%, MS: 2.24%), and BM B, which considers the three mentioned categories with the session shares from the literature (IW: 50%, CS: 42.11%, MS: 7.89%). For both, the capacity allocated to each service within a category is split uniformly, since no information w.r.t. the intra-category session shares is available.

Evaluation: The performance of the system is evaluated for a week in an area covered by 10 different antennas, for all the 28 services listed in Table 7.1. Table 7.2 shows the percentage of time for which the capacity allocated by the operator is sufficient to serve all the traffic demand, averaged over antennas and services. The solution based on the models proposed is the only one that achieves the SLA terms to guarantee the proposed Quality of Service. The other solutions suffer from the inaccuracy in estimating the share of demand of each service, and they also have bigger variability between services. An important aspect of this session-level per-service modeling is the robustness against outliers. Mobile traffic is very bursty, and dimensioning the slices based on traffic peaks may be very detrimental and lead to a waste of reserved resources. This can be seen in Figure 7.12, where the actual allocated capacity that satisfies the SLA terms is far below the traffic demand peaks.

7.5.2. Energy consumption in CU-DU

A standard virtualized Radio Access Networks (vRAN) scenario is considered, portrayed in Figure 7.13a, where Centralized Units (CU) located at a Telco Cloud Site (CS) serve traffic from a set of DUs at multiple Far Edge Sites (ESs), each associated to a group of Radio Units (RU). CUs run within physical servers (PS), whose energy consumption depends on the computing load. Therefore, the dynamic association of DUs to CUs within each PS, in accordance with the fluctuation of mobile data traffic at the RUs, determines the energy cost of the vRAN infrastructure for the operator. This is a

Figure 7.12. Normalized traffic demand and allocated capacity to Facebook network slice at one BS over time.

major operating expense that needs to be minimized [275], [276].

Energy optimization model: It can be assumed that all PSs at the CS are identical machines, whose capacity is limited by the maximum sum throughput of the mobile traffic they handle, up to 100 Mbps when working at full load [277]. The energy consumption is modeled at the each PS following real specifications of IBM servers [277, Table IV], such that a PS consumes a maximum power of 200 W when working on traffic at 100 Mbps; the power consumption is instead at 60 W when the PS is turned on but idle, and increases proportionally until the 200 W above at 100% load.

The operator employs then a dedicated algorithm for orchestrating the resources in the CU, executed at every time slot (TS) of one second. Due to the considered energy consumption model, minimizing the energy consumption of the system is equivalent to minimizing the number of active PSs. Thus, the algorithm is a bin-packing heuristic [278] that minimizes the number of PSs based on the current state of served sessions and the new session arrivals during the TS. While this model is relatively simple, it offers reasonable performance; more importantly, it provides a basis to assess the impact of traffic models on the compute resource management results, which is the goal.

Mobile traffic models: A vRAN system with one CS serving 20 different ESs is assumed, each handling 20 RUs. The arrival time of the sessions is modeled such that the number of arrived sessions per minute at each RU follows the modeled distribution in Section 7.3.1. The sessions are generated according to three different strategies: (*i*) using the measurement data presented in Section 7.3, by sampling $F_s(d)$ and matching the traffic volume values to $v_s(d)$ to derive duration and average throughput; (*ii*) using the proposed models as described in Section 7.4.4; (*iii*) from traditional mobile traffic models available in the literature [274, Table II], [13, Table XVII] that provide throughput and session size/duration for three service categories.

For all cases, the share of per service sessions are extracted from Table 7.1. For (*iii*), the 28 classes (services) are mapped into the 3 categories that their model considers, and

(a) System model

(b) Performance error of traffic models

(c) Power consumption over time

Figure 7.13. (a) vRAN system model considered in 7.5.2. (b) APE with respect to the measurements traffic in terms of active PSs and power consumption, for the model and the benchmarks. (c) Power consumption sample.

again generate sessions for each category according to Table 7.1. As the model in (*iii*) is used as a term of comparison, three different benchmarks are generated in fact from it: BM A fully adheres to the original models, BM B normalizes the generated data so that the total system throughput matches that observed in the measurement data, and BM C normalizes the throughput of each service class so that it matches that recorded in the measurement data. Clearly, BM B and BM C are not feasible with information from literature only, but they allow highlighting the advantages of the proposed models.

Performance evaluation results: Experiments for several emulated days are ran, orchestrating CS resources via the described strategy for all traffic models above. The same realization of class-level session arrivals is employed in all tests to avoid biases. Figure 7.13b summarizes the results, expressed as the distributions of the number of active PSs and of the power consumption. The absolute percentage error (APE) is reported with respect to the same figures obtained by feeding the optimization model with the measurement data. The proposed model tightly approximates the real scaling of the compute resources at the CS, with median APE well below 5% and very small deviation for both metrics. The difference is apparent with respect to the benchmarks, which incur into APE of 100%–1000%, hence leading to performance results that are completely off.

Clearly, the traffic generated by the benchmarks fails to capture real-world session-level statistics, which completely undermines the reliability of the performance evaluation. Figure 7.13c offers a close-up view of the temporal evolution of the power consumption with real data, the model and BM C: the result further highlights the quality of the contribution in mimicking real-world traffic.

7.5.3. Optimization of heterogeneous networks

The use of synthetic network data by the models described in Section 7.4 can help network optimization problems by providing realistic twins of the network, which could be beneficial for the optimization of the complex deployments that are becoming more common, such as heterogeneous networks. Works in this area usually utilize simplistic scenarios that are not based on real deployments [279]–[281], so the presented models can lead to a better assessment of the proposed algorithms.

The task of optimizing the many variables that characterize modern RANs is challenging for ultra-dense heterogeneous scenarios, where macro, micro, and femto cells are deployed together inside a relatively small area in order to provide higher capacity and enhanced Quality of Service (QoS) [282]. These dense and hierarchical deployments are present, especially in populated urban centers: Figure 7.14 shows an example of such a RAN configuration observed in a production-grade mobile network, which includes overlapping macro and micro cells providing coverage to the city center of Paris, as it can be seen on Figure 7.14. In these heterogeneous RAN infrastructures, a lack of optimization of transmission powers, resource allocations, and assignments between users and BSs might lead to high inefficiencies, such as high transmission powers and low QoS for users [283], [284]. The lack of realistic data that reflects this complexity could impact the assessment of proposed solutions

Algorithm: The proposed framework for optimizing denser heterogeneous RAN deployments is a joint optimization problem that allows minimizing the transmission power in OFDMA heterogeneous networks while meeting individual users' throughput requirements. The problem formulation allows considering multiple variables simultaneously, e.g., user association, resource allocation, and transmission power, instead of dividing the optimization problem into smaller ones, by deriving piece-wise concave approximations of the Shannon-Hartley formula, and consequently applying a variable transformation that enables posing the optimization problem into a Mixed-integer Geometric Program (MIGP) form. The effectiveness of the proposed optimization framework is demonstrated in a dependable scenario with up to 800 users generating demands modeled after mobile network traffic measurements and located in a real-world heterogeneous RAN, in the presence of wireless channel characteristics reproduced via a high-performance signal propagation solver.

Evaluation: The applicability of the proposed MIGP algorithm is evaluated in

Figure 7.14. Real-world topology of an operational heterogeneous network deployed in a neighborhood of a large European city.

realistic scenarios based on the models described on Section 7.4, where high-fidelity digital twins based on the collection of real deployment scenarios will enable the assessment of the proposed algorithm against other state of the art solutions. It's interesting to note that by utilizing the models proposed in this Chapter to generate synthetic network data, it's possible to create scenarios that are significantly bigger and more complex, in which most solutions previously mentioned were never tested. This enables a deeper exploration of the performance boundaries of network optimization frameworks.

For each scenario, S, a set of base stations BS_j is chosen respecting the operational network topology, with the number of sessions arriving on a 1-minute interval being estimated on proposed arrival models, considering the busiest scenario available. This interval represents a moment where the network operates at higher loads, with the number of connected users and their traffic demands reaching their peak values. The shares of sessions that each application is expected to generate will respect the values seen on Table 7.1. For each expected session arriving at BS_j, the traffic and duration (and thereby throughput) are estimated according to the demands of the corresponding mobile service based on the models described in Section 7.4. Each generated session is assigned to an individual user placed in a random location, lying within the covered area of s-th HetNet topology, determined by a Poisson random point process.

To evaluate the received signal distribution within the coverage area of each BS_j a ray tracing simulator is deployed which allows to carefully emulate the electromagnetic wave propagation in the urban fabric [285], [286].

In total, 5 scenarios generated from the proposed models were considered to evaluate the proposed framework, each assuming different spatial and network complexities. Scenario S_1 is a downtown location that contains 1 macro cell and 4 micro cells located in

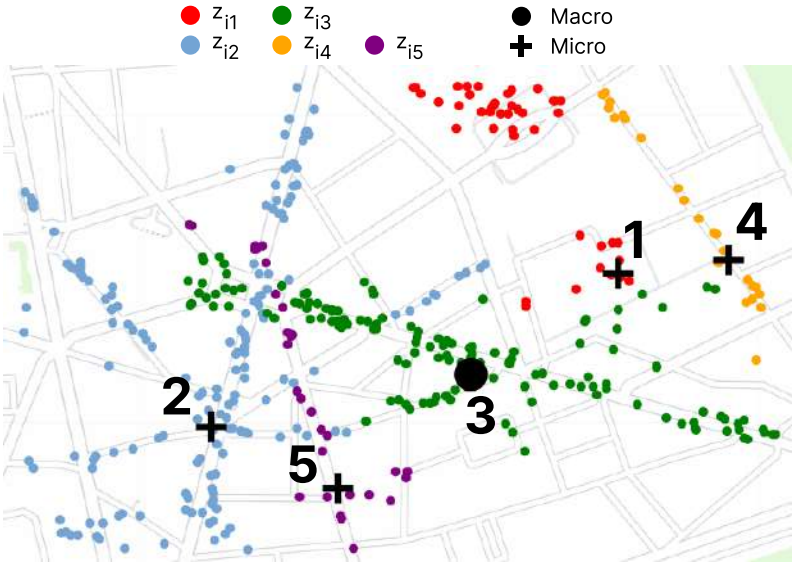


Figure 7.15. Users' assignment after MIGP optimization for S_1.

the historical center in the downtown area of the city under consideration. Scenarios S_2 and S_3 build on the HetNet operating at the same downtown region of S_1, but increasing the number of micro BS to 7 and 8, respectively, to evaluate the scalability of the proposed framework in denser network configurations. Scenario S_4 represents a less dense topology located by the river side of the city, consisting of 1 macro and 2 micro cells. Finally, scenario S_5 is on an area at the hill side of the city. All scenarios consider BSs that are 5G NSA-enabled.

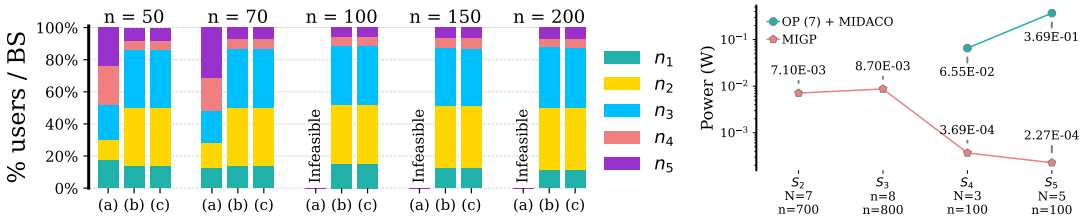
Initially, the applicability of the MIGP framework is tested on scenario S_1. This will consider a 5G RAN operating at 3.5 GHz with 500 resource blocks, utilizing an OFDMA technique with 100 MHz of bandwidth. As previously mentioned, S_1 consists of 1 macro and 4 micro BSs that accommodate the throughput requirements of 400 users, bounded between 1.35 Kbps and 18.72 Mbps; the noise power is set to -174 dBm/Hz. The goal is to minimize the total transmission power of the HetNet.

The resulting assignment is seen in Figure 7.15, which shows a clear tendency to connect the users to the BS providing the highest channel gain. It can be noted on the map a pattern of users being mostly around the outside of buildings, which is amplified by the propagation loss experienced by 3.5MHz communication inside buildings. Considering the 400 total users of this scenario, the number of users n_j connected to each BS j, the number of RB used in each BS, and the respective transmission powers are provided in Table 7.3. P_j^{Tot} stands for the total transmission power used by BS j (each resource block spends P_j Watts). This example guarantees that MIGP not only works but can solve complex scenarios with a significant number of users present.

The proposed models can be also leveraged to benchmark solutions against other state

BS j	n_j	RB_j	P_j (W)	P_j^{Tot} (W)
$j=1$	43	423	0.2894×10^{-3}	0.1224
$j=2$	161	495	0.7117×10^{-3}	0.3523
$j=3$	142	490	0.6535×10^{-3}	0.3202
$j=4$	24	413	0.0038×10^{-3}	1.5694×10^{-3}
$j=5$	30	435	0.2151×10^{-3}	0.0936
$\sum_{j=1}^{N}$	400	2256	0.0019	0.8901

Table 7.3. Number of users connected to each BS for S_1.



(a) **(b)**

Figure 7.16. (a) Number of users connected to each BS of S_1 for OP(7) + MIDACO, GP + fixed z_{ij} and MIGP.; (b) $\sum_{j=1}^{N} P_j$ for different networks.

of the art frameworks, as well as to text its scalability within more complex scenarios. A comparison of MIGP is performed with two other frameworks: *GP + fixed z_{ij})* (which is a simplified version of MIGP) and *OP (7)+MIDACO* (which is based in a commercially available solver). The number of users connected to each BS in S_1, considering a growth in users across the scenario, is seen on Figure 7.16a. With this approach, it's possible to note that MIGP gives a more homogeneous spread of users across BS than the two other frameworks, and that OP (7)+MIDACO was unable to solve problems with more than 100 users. It's possible also to assess how the frameworks perform across diverse scenarios, in relation to how they assign users to BS and how this affects the overall consumption of the network. Figure 7.16b presents those the results obtained from scenarios S_2 to S_5 for MIGP and OP (7)+MIDACO. This simulation not only vary in the number of BS deployed but also on the number of users present in each. With the data provided by the proposed models, it's possible to check that the assignment obtained from MIGP result always in lower consumption of energy, when compared to the commercially available solution of OP (7)+MIDACO.

7.6. Main takeaways

This Chapter presented first-of-its-kind exploration of mobile traffic at the transport-layer session level. This study builds on substantial measurement data and reveals

new facets of traffic, which are modeled accurately as a contribution to more reliable performance evaluations of mobile system. This work presents a few limitations: the granularity of the data does not allow for fine grained or intra-session simulations (i.e. packet level generation); since the models are at service level, they will require updates over the years to consider changes in popularity and new services that emerge; due to the aggregation of sessions at BS level to comply with user privacy regulations, the ability to study sequences of TCP/UDP flows a user may generate through the use of a mobile service is lost, which limits an expansion of this study to application-layer dynamics.

8 Discussions and perspectives

8.1. Discussion of the results from this thesis

The exponential growth of large-scale models, the rise in popularity of data-driven solutions within computer networks, and the rise in popularity of new data sources for multi-domain research (i.e., anthropology, epidemiology) are major drivers for the need for research in the domain of large-scale mobile network measurements. This means there's a significant rise in the demand for techniques for data collection and processing, in order to guarantee the veracity of collected data sets. Also, there's a need for research pushing the possibility boundaries of the data, in order to understand which types of insides can be gathered with confidence from the usage patterns of smartphones, so researchers from fields outside of mobile networks can have a better grasp of how to incorporate this data into their research interests.

The main motivation of this thesis was to present a significant overview of the field of Network Data Science and mobile network measurements, contributing to pushing the envelope of possibilities that the data can achieve. As it was originally mentioned in Chapter 2, works based on mobile network measurements tend to contribute within the fields of network, Social, and nobility analysis. The contributions seen throughout this thesis heavily favored the area of network analysis, with a set of insights also being new and significant for researchers in the field of social analysis, especially in the areas of environment and epidemics. These contributions are within the measurement and collection domain, the data processing domain, and presenting techniques and insights useful for the future of the field. Throughout the chapters, here are the highlights of each of their contributions and how their impact on the research community.

The initial contributions of this thesis involved detailing the data engineering and processing tasks involved in large-scale mobile network measurements. One of the most critical involved how it was possible to identify 5G NSA flows from measurements, as the network core (and therefore the probes) are shared with 4G, making it impossible for standard techniques used to distinguish between both. This process enables many of the

network analysis contributions throughout this thesis. Important and often overlooked concepts about the processing part needed before the analysis started were also detailed, especially in relation to the spatial unit of mobile network measurements.

A deep dive into how newer mobile network technologies affected the consumption of mobile traffic also showcased many important insights. A first-of-its-kind analysis of 5G adoption and its impact on service consumption was presented, using large-scale measurements collected in an operational nationwide mobile network. By taking a service-oriented perspective on the adoption and early utilization of 5G networks and focusing on the impacts over the network, it set itself apart from other works. Through it, it was possible to observe that the adoption of 5G in France is on a slow rise, with early adoption seemingly correlated with Apple devices, and that contrary to expectations, regions, where early adoption was stronger, had lower income and educational levels, which could be related to perhaps 5G substituting Wi-Fi in such residences. These user insights can prove significant importance for the business units responsible for the economics and marketing aspects of the deployment, as they offer a passive and privacy-preserving approach to measure adoption within the population. Also, the characterization study of indoor mobile networks presented insights on how the preference of mobile applications is tied to the indoor ambient where users were, independently of the outdoor environment, which can be an important tool for network optimization techniques, especially those focusing on network slices and optimization in the per-application level, e.g., optimizing music streaming applications in metro stations, as this category of apps had overutilization there. These insights can be critical in the development and evolution of future specifications of mobile networks.

Another set of explored techniques showcased the exploration of space and environments through the lenses of mobile networks. As multi-domain researchers become interested in utilizing mobile network measurements, there is a need for techniques and insights on how to fit smartphone measurements into environmental and spatial analysis, to which the works here presented intended to contribute. The use of EFA for the identification of land usage through mobile traffic utilization showcased a vast array in both temporal and spatial patterns, which can be deeply useful for urbanism and transportation researchers as alternatives to identify profiles of utilization within major cities. Also, new techniques for the study of urban green spaces in relation to smartphone utilization revealed the heterogeneity of green spaces in Paris, which could be related to the location and features of such spaces and how this affects the relationship between users and their smartphones.

Large-scale mobile network measurements can also be used to understand the effects of significant events. The impact of restriction measurements imposed by the government of France due to the COVID-19 pandemic was quantified by their impact on mobile network utilization, with a focus on the temporal, spatial, and per-application dynamic

changes. These highlighted population movements by studying the changes in hotspots of traffic consumption, specifically for certain key apps, and how their popularity changed as the movement was restricted to contain the spread of the disease. Also, a zoom into cities resulted in similar patterns within urban environments, especially how the socio-economical aspects of neighborhoods may have affected the reaction to the restrictions on smartphone usage. These insights show the potential of mobile network measurements as an alternative source of information for authorities to assess their measurements, as well for MNOs to understand how big events may affect their traffic demands. A different type of event was also explored: public protests. By characterizing the 2023 French pension reform strikes, it was observed that traffic demand patterns can be used as a proxy to track the spatiotemporal progress of manifestations across the city and that a few key applications are greatly affected by these events, which hints at their possibility of utilization for more complex protest detection algorithms.

Finally, a technical perspective on mobile network measurements was shown, providing a first-of-its-kind analysis of transport layer session-level traffic consumption of mobile services. By understanding the characteristics of services, it was possible to note that even services within similar categories (e.g., video streaming) had different statistical properties, and within the same service, whenever it was possible to identify the type of content being consumed (e.g., video or text), it was noted that the characteristics also shifted. More importantly, per-application models for the PDFs of traffic consumption were presented, as well as their relation with duration, throughput, and arrival rate. These models go a significant layer deeper than what was presented in the literature, and as showcased with 3 use cases, result in significant variations and improvement in the performance of many mobile network optimization use cases.

Overall, the field of network data science aims to achieve more than anything a deeper understanding of the data patterns, and to better enable data-driven solutions by providing critical insights to all interested parties, as well as techniques and data sets capable of shaping businesses. This thesis presented a significant number of contributions in this direction, which hopefully will help guide the next wave of works interested in large-scale mobile network measurements.

8.2. Perspectives for the area of networks data science

As expected, there are still many open questions in the topics touched on within this thesis, many of which can be tackled in the current context of mobile networks. Next, a few of these interesting future directions will be discussed.

Exploration of spatial representations of traffic: A significant number of studies that utilize mobile network measurements perform some sort of spatial analysis of the traffic. The vast majority of those will usually rely on two types of representation for the

traffic over space: Voronoi polygons or square/hexagon grids. As the expected spatial detail has been increasing, i.e., researchers want to be able to look more and more at specific features and locations within cities (such as parks, as discussed over Section 5.2), a potentially interesting direction for future research is to explore and generate a better understanding about the correct representations of traffic consumption over the expected coverage of antennas. Both Voronoi and grid representations greatly simplify and approximate the expected propagation of the antennas, which is a significantly more complex spatial entity. Two directions of research could be taken:

- Based on the coverage maps that MNOs generate for any new antenna deployment, a study on how to convert those into more simplistic representations would be of great interest to the research community, as those could substitute other spatial representations as more precise spatial localizations of traffic demands.

- An investigation of the limits for the Voronoi representation, how much uncertainty is added to the location users' traffic demand when Voronois are used, which could highlight how detailed over-space traffic demands can be estimated in the current state.

Study of smartphone model adoption and its relation to traffic demands: A currently unused field that's possible to obtain from the measurement data is the International Mobile Equipment Identity (IMEI), which is a numeric identifier which, while unique for each device, could be utilized to determine the maker and model of each smartphone. This means that it could be possible to study how different makers, models, and price ranges of smartphones could be related to mobile traffic consumption, such as:

- Understand if there's a correlation between smartphone maker, model, and price range in relation to the preferences of mobile app usage. A few studies have already explored how the relationship between socioeconomic indicators and traffic usage [79] or preferences over smartphone apps [76], so it could be interesting to understand if the value of smartphones could also play a relation with app preferences, or even to understand if the presence of certain smartphones were also related to socioeconomic indicators of certain areas.

- Adoption of devices with newer technologies. This for example could have a direct relation with the adoption of 5G, where not only the MNO could understand where the demands for 5G traffic are, but also the rate its consumer population is adopting 5G devices. This could also be expanded for any other new tech that can be assessed from the smartphone model, and even be cross-checked with app and phone makers in relation to certain smartphones and how their preferences for certain apps are.

Impact of non-smartphone devices in mobile networks: As the infrastructure of MNOs reaches deeper, and more and more devices are being connected to the internet, mobile networks play an essential role in connecting any kind of device deployed on the field to servers. The main impact here would be IoT and M2M device communication. Therefore, the study of the impact of those devices is an interesting future topic in Networks Data Science. For example, the work in [287] explored how connected cars are utilizing resources of newly deployed 5G networks. Other directions that could be utilized are to understand the RAT these devices are utilizing (i.e., to understand their usage of legacy infrastructure), as well as merging with a IMEI analysis to create a profile of which types of devices are currently leveraging the infrastructure of MNOs.

Standardization of datasets: The final, and perhaps one of the most crucial, future directions would be the standardization of mobile network datasets. For all studies contributed to this thesis and over previous works, a common problem becomes how the data collected from different MNOs may not follow the same standards, being it on how it is collected, pre-processed and its aggregation and privacy-related changes required. This means that it could be difficult to directly compare results from data sets obtained from different MNOs, as their full collection pipelines are different. While this perhaps would not be that impactful for mobile traffic volume demand analysis, this is significantly impactful for mobility analysis. Indeed, researchers interested in studying population movements and transportation using mobile network measurements as a proxy could be more likely to try to obtain bases from different sources, so guaranteeing that certain standards when calculating those sets would be followed would be of great gain for the research community. Certain companies who aggregate this type of data for later selling as a product may have this care when they process their sets, but this is usually a proprietary tool from their side. It would be critical for the expansion of mobile network measurements for multi-domain research that open standards about data processing are agreed, in order to allow better uniformity and validation across works.

Bibliography

[1] S. Mishra, A. Zanella, O. E. Martínez-Durive, D. Madariaga, C. Ziemlicki, M. Fiore, *et al.*, "Characterizing 5g adoption and its impact on network traffic and mobile service consumption", in *IEEE International Conference on Computer Communications*, 2024.

[2] A. Zanella, O. E. Martínez-Durive, S. Mishra, D. Madariaga, M. Fiore, *et al.*, "Impact of public protests on mobile networks", in *IEEE International Conference on Computer Communications*, 2024.

[3] G. O. Ferreira *et al.*, "A joint optimization approach for power-efficient heterogeneous ofdma radio access networks", *arXiv preprint arXiv:2403.14555*, 2024.

[4] A. Furno, A. F. Zanella, R. Stanica, and M. Fiore, "Spatial and temporal exploratory factor analysis of urban mobile data traffic", *Data Science for Transportation*, vol. 6, no. 1, p. 4, 2024.

[5] A. F. Zanella, A. Bazco-Nogueras, C. Ziemlicki, and M. Fiore, "Characterizing and modeling session-level mobile traffic demands from large-scale measurements", in *Proceedings of the 2023 ACM on Internet Measurement Conference*, 2023, pp. 696–709.

[6] S. Bakirtzis *et al.*, "Characterizing mobile service demands at indoor cellular networks", in *Proceedings of the 2023 ACM on Internet Measurement Conference*, 2023, pp. 645–659.

[7] A. F. Zanella, O. E. Martínez-Durive, S. Mishra, Z. Smoreda, and M. Fiore, "Impact of later-stages covid-19 response measures on spatiotemporal mobile service usage", in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, IEEE, 2022, pp. 970–979.

[8] A. F. Zanella, D. Madariaga, S. Mishra, O. E. Martínez-Durive, Z. Smoreda, and M. Fiore, *Characterizing, modeling and exploiting the mobile demand footprint of large public protests*, Submitted to ACM IMC 24, Under Review, 2024.

[9] A. F. Zanella, S. Rubrichi, Z. Smoreda, and M. Fiore, *Modeling and understanding the impact of covid-19 containment policies on mobile service consumption in french cities*, Submitted to EPJ Data Science, Under Review, 2024.

[10] Ericsson, *Ericsson Mobility Report*, Jun. 2023.

[11] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: A survey", *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 124–161, 2015.

[12] V. Blondel, A. Decuyper, and G. Krings, *A survey of results on mobile phone datasets analysis. epj data sci. 4 (1)*, 2015.

[13] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5g usage scenarios and traffic models", *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 905–929, 2020.

[14] J. C. Cardona, R. Stanojevic, and N. Laoutaris, "Collaborative consumption for mobile broadband: A quantitative study", in *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*, 2014, pp. 307–318.

[15] J. P. Bagrow, D. Wang, and A.-L. Barabasi, "Collective response of human populations to large-scale emergencies", *PloS one*, vol. 6, no. 3, e17680, 2011.

[16] D. Goergen, V. Mendiratta, R. State, and T. Engel, "Identifying abnormal patterns in cellular communication flows", in *Proceedings of Principles, Systems and Applications on IP Telecommunications*, 2013, pp. 1–6.

[17] A. Furno, D. Naboulsi, R. Stanica, and M. Fiore, "Mobile demand profiling for cellular cognitive networking", *IEEE Transactions on Mobile Computing*, vol. 16, no. 3, pp. 772–786, 2016.

[18] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and Z. Smoreda, "Identifying common periodicities in mobile service demands with spectral analysis", in *2020 Mediterranean Communication and Computer Networking Conference (MedComNet)*, IEEE, 2020, pp. 1–8.

[19] V. Soto and E. Frías-Martínez, "Automated land use identification using cell-phone records", in *Proceedings of the 3rd ACM international workshop on MobiArch*, 2011, pp. 17–22.

[20] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring land use from mobile phone activity", in *Proceedings of the ACM SIGKDD international workshop on urban computing*, 2012, pp. 1–8.

[21] M. Lenormand *et al.*, "Comparing and modelling land use organization in cities", *Royal Society open science*, vol. 2, no. 12, p. 150 449, 2015.

[22] S. Grauwin, S. Sobolevsky, S. Moritz, I. Gódor, and C. Ratti, "Towards a comparative science of cities: Using mobile traffic records in new york, london, and hong kong", in *Computational Approaches for Urban Environments*, M. Helbich, J. Jokar Arsanjani, and M. Leitner, Eds. Cham: Springer International Publishing, 2015, pp. 363–387. DOI: 10.1007/978-3-319-11469-9_15. [Online]. Available: https://doi.org/10.1007/978-3-319-11469-9_15.

[23] A. Furno, M. Fiore, R. Stanica, C. Ziemlicki, and Z. Smoreda, "A tale of ten cities: Characterizing signatures of mobile traffic in urban areas", *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, pp. 2682–2696, 2016.

[24] B. Cici, M. Gjoka, A. Markopoulou, and C. T. Butts, "On the decomposition of cell phone activity patterns and their connection with urban ecology", in *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2015, pp. 317–326.

[25] R. Singh, M. Fiore, M. Marina, A. Tarable, and A. Nordio, "Urban vibes and rural charms: Analysis of geographic diversity in mobile service usage at national scale", in *The World Wide Web Conference*, 2019, pp. 1724–1734.

[26] F. Calabrese, J. Reades, and C. Ratti, "Eigenplaces: Segmenting space through digital signatures", *IEEE Pervasive Computing*, vol. 9, no. 1, pp. 78–84, 2009.

[27] A. Furno, M. Fiore, and R. Stanica, "Joint spatial and temporal classification of mobile traffic demands", in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, IEEE, 2017, pp. 1–9.

[28] P. Fiadino, M. Schiavone, and P. Casas, "Vivisecting whatsapp through large-scale measurements in mobile networks", in *Proceedings of the 2014 ACM Conference on SIGCOMM*, ser. SIGCOMM '14, Chicago, Illinois, USA: Association for Computing Machinery, 2014, pp. 133–134. DOI: 10.1145/2619239.2631461. [Online]. Available: https://doi.org/10.1145/2619239.2631461.

[29] Q. Deng, Z. Li, Q. Wu, C. Xu, and G. Xie, "An empirical study of the wechat mobile instant messaging service", in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2017, pp. 390–395. DOI: 10.1109/INFCOMW.2017.8116408.

[30] A. Mansy, M. Ammar, J. Chandrashekar, and A. Sheth, "Characterizing client behavior of commercial mobile video streaming services", in *Proceedings of Workshop on Mobile Video Delivery*, 2014, pp. 1–6.

[31] V. K. Adhikari *et al.*, "Measurement study of netflix, hulu, and a tale of three cdns", *IEEE/ACM Transactions On Networking*, vol. 23, no. 6, pp. 1984–1997, 2014.

[32] Z. Li *et al.*, "An empirical analysis of a large-scale mobile cloud storage service", in *Proceedings of the 2016 Internet Measurement Conference*, 2016, pp. 287–301.

[33] Y. Zhang and A. Årvidsson, "Understanding the characteristics of cellular data traffic", in *Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*, 2012, pp. 13–18.

[34] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, C. Ziemlicki, and Z. Smoreda, "Not all apps are created equal: Analysis of spatiotemporal heterogeneity in nationwide mobile service usage", in *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*, 2017, pp. 180–186.

[35] A. Okic, A. E. Redondi, I. Galimberti, F. Foglia, and L. Venturini, "Analyzing different mobile applications in time and space: A city-wide scenario", in *2019 IEEE wireless communications and networking conference (WCNC)*, IEEE, 2019, pp. 1–6.

[36] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Measuring serendipity: Connecting people, locations and interests in a mobile 3g network", in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, 2009, pp. 267–279.

[37] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman, "Identifying diverse usage behaviors of smartphone apps", in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, 2011, pp. 329–344.

[38] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3g cellular data network", in *2012 Proceedings IEEE INFOCOM*, IEEE, 2012, pp. 1341–1349.

[39] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices", *ACM SIGMETRICS Performance Evaluation Review*, vol. 39, no. 1, pp. 265–276, 2011.

[40] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao, "Profiling users in a 3g network using hourglass co-clustering", in *Proceedings of the sixteenth annual international conference on Mobile computing and networking*, 2010, pp. 341–352.

[41] H. Li *et al.*, "Characterizing smartphone usage patterns from millions of android users", in *Proceedings of the 2015 Internet Measurement Conference*, 2015, pp. 459–472.

[42] A. Narayanan *et al.*, "A variegated look at 5g in the wild: Performance, power, and qoe implications", in *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, 2021, pp. 610–625.

[43] A. Narayanan *et al.*, "A comparative measurement study of commercial 5g mmwave deployments", in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, IEEE, 2022, pp. 800–809.

[44] A. Narayanan *et al.*, "A first look at commercial 5g performance on smartphones", in *Proceedings of The Web Conference 2020*, 2020, pp. 894–905.

[45] Y. Liu and C. Peng, "A close look at 5g in the wild: Unrealized potentials and implications", in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, IEEE, 2023.

[46] C. Fiandrino, D. Juárez Martínez-Villanueva, and J. Widmer, "Uncovering 5g performance on public transit systems with an app-based measurement study", in *Proceedings of the 25th International ACM Conference on Modeling Analysis and Simulation of Wireless and Mobile Systems*, 2022, pp. 65–73.

[47] Y. Pan, R. Li, and C. Xu, "The first 5g-lte comparative study in extreme mobility", *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 6, no. 1, pp. 1–22, 2022.

[48] D. Xu *et al.*, "Understanding operational 5g: A first measurement study on its coverage, performance and energy consumption", in *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, 2020, pp. 479–494.

[49] X. Yang *et al.*, "Mobile access bandwidth in practice: Measurement, analysis, and implications", in *Proceedings of the ACM SIGCOMM 2022 Conference*, 2022, pp. 114–128.

[50] P. Parastar, A. Lutu, G. Alay Ozguand Caso, and D. Perino, "Spotlight on 5g: Performance, device evolution and challenges from a mobile operator perspective", in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, IEEE, 2023.

[51] Y. Jin *et al.*, "Characterizing data usage patterns in a large cellular network", in *Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*, 2012, pp. 7–12.

[52] X. Wang *et al.*, "Spatio-temporal analysis and prediction of cellular traffic in metropolis", *IEEE Trans. Mobile Comput.*, vol. 18, no. 09, pp. 2190–2202, Sep. 2019.

[53] C. Zhang, H. Zhang, J. Qiao, D. Yuan, and M. Zhang, "Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data", *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1389–1401, 2019.

[54] J. Ridoux, A. Nucci, and D. Veitch, "Seeing the difference in ip traffic: Wireless versus wireline", in *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, IEEE, 2006, pp. 1–12.

[55] E. Graells-Garrido, D. Caro, O. Miranda, R. Schifanella, and O. F. Peredo, "The www (and an h) of mobile application usage in the city: The what, where, when, and how", in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 1221–1229.

[56] A. Karasaridis and D. Hatzinakos, "Network heavy traffic modeling using /spl alpha/-stable self-similar processes", *IEEE Transactions on Communications*, vol. 49, no. 7, 2001.

[57] R. Li, Z. Zhao, C. Qi, X. Zhou, Y. Zhou, and H. Zhang, "Understanding the traffic nature of mobile instantaneous messaging in cellular networks: A revisiting to α-stable models", *IEEE Access*, vol. 3, 2015.

[58] R. Li, Z. Zhao, J. Zheng, C. Mei, Y. Cai, and H. Zhang, "The learning and prediction of application-level traffic data in cellular networks", *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, 2017.

[59] K. Xu, R. Singh, H. Bilen, M. Fiore, M. K. Marina, and Y. Wang, "Cartagenie: Context-driven synthesis of city-scale mobile network traffic snapshots", in *IEEE PerCom '22*, Los Alamitos, CA, USA, Mar. 2022, pp. 119–129. DOI: 10.1109/PerCom53586.2022.9762395. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/PerCom53586.2022.9762395.

[60] Z. Lin, A. Jain, C. Wang, G. Fanti, and V. Sekar, "Using gans for sharing networked time series data: Challenges, initial promise, and open questions", in *ACM IMC '20*, Virtual Event, USA, 2020, pp. 464–483. DOI: 10.1145/3419394.3423643. [Online]. Available: https://doi.org/10.1145/3419394.3423643.

[61] K. Xu *et al.*, "Spectragan: Spectrum based generation of city scale spatiotemporal mobile network traffic data", in *ACM CoNEXT '21*, Virtual Event, Germany, 2021, pp. 243–258. DOI: 10.1145/3485983.3494844. [Online]. Available: https://doi.org/10.1145/3485983.3494844.

[62] C. Sun, K. Xu, M. Fiore, M. K. Marina, Y. Wang, and C. Ziemlicki, "Appshot: A conditional deep generative model for synthesizing service-level mobile traffic snapshots at city scale", *IEEE Transactions on Network and Service Management*, vol. 19, no. 4, pp. 4136–4150, 2022. DOI: 10.1109/TNSM.2022.3199458.

[63] 3GPP TR 36.814 V9.2.0, *3rd Generation Partnership Project; technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects (Release 9)*, Mar. 2017.

[64] 3GPP TSG-RAN1#48 R1-070674, *LTE physical layer framework for performance verification*, Feb. 2007.

[65] 3GPP TR 36.888 V12.0.0, *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE (Release 12)*, Jun. 2013.

[66] IEEE 802.16m-08/004r2, *IEEE 802.16m evaluation methodology document (EMD)*, Jul. 2008.

[67] E. Mucelli Rezende Oliveira, A. Carneiro Viana, K. Naveen, and C. Sarraute, "Mobile data traffic modeling: Revealing temporal facets", *Computer Networks*, vol. 112, pp. 176–193, 2017. DOI: https://doi.org/10.1016/j.comnet.2016.10.016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128616303644.

[68] J. Wu, M. Zeng, X. Chen, Y. Li, and D. Jin, "Characterizing and predicting individual traffic usage of mobile application in cellular network", in *ACM UbiComp '18*, Singapore, Singapore: Association for Computing Machinery, 2018, pp. 852–861. DOI: 10.1145/3267305.3274173. [Online]. Available: https://doi.org/10.1145/3267305.3274173.

[69] A. Stoica, Z. Smoreda, C. Prieur, and J.-L. Guillaume, "Age, gender and communication networks", *NetMob—Ananlysis of Mobile Phone Networks*, 2010.

[70] C. Sarraute, P. Blanc, and J. Burroni, "A study of age and gender seen through mobile phone usage patterns in mexico", in *2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2014)*, IEEE, 2014, pp. 836–843.

[71] Y. Wang, H. Zang, and M. Faloutsos, "Inferring cellular user demographic information using homophily on call graphs", in *2013 Proceedings IEEE INFOCOM*, IEEE, 2013, pp. 3363–3368.

[72] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks", *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, P10008, 2008.

[73] N. Pokhriyal and D. C. Jacques, "Combining disparate data sources for improved poverty prediction and mapping", *Proceedings of the National Academy of Sciences*, vol. 114, no. 46, E9783–E9792, 2017.

[74] E. Aiken, S. Bellue, D. Karlan, C. Udry, and J. E. Blumenstock, "Machine learning and phone data can improve targeting of humanitarian aid", *Nature*, vol. 603, no. 7903, pp. 864–870, 2022.

[75] G. Chi, H. Fang, S. Chatterjee, and J. E. Blumenstock, "Microestimates of wealth for all low-and middle-income countries", *Proceedings of the National Academy of Sciences*, vol. 119, no. 3, e2113658119, 2022.

[76] I. Ucar, M. Gramaglia, M. Fiore, Z. Smoreda, and E. Moro, "News or social media? socio-economic divide of mobile service consumption", *Journal of The Royal Society Interface*, vol. 18, no. 185, p. 20 210 350, 2021.

[77] G. V. Houngbonon, E. Le Quentrec, and S. Rubrichi, "Access to electricity and digital inclusion: Evidence from mobile call detail records", *Humanities and Social Sciences Communications*, vol. 8, no. 1, pp. 1–11, 2021.

[78] L. State, H. Salat, S. Rubrichi, and Z. Smoreda, "Explainability in practice: Estimating electrification rates from mobile phone data in senegal", in *World Conference on Explainable Artificial Intelligence*, Springer, 2023, pp. 110–125.

[79] S. Mishra, Z. Smoreda, and M. Fiore, "Second-level digital divide: A longitudinal study of mobile traffic consumption imbalance in france", in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 2532–2540.

[80] L. Pappalardo, M. Vanhoof, L. Gabrielli, Z. Smoreda, D. Pedreschi, and F. Giannotti, "An analytical framework to nowcast well-being using mobile phone data", *International Journal of Data Science and Analytics*, vol. 2, pp. 75–92, 2016.

[81] J.-P. Onnela, S. Arbesman, M. C. González, A.-L. Barabási, and N. A. Christakis, "Geographic constraints on social network groups", *PLoS one*, vol. 6, no. 4, e16939, 2011.

[82] N. Eagle, Y.-A. de Montjoye, and L. M. Bettencourt, "Community computing: Comparisons between rural and urban societies using mobile phone data", in *2009 international conference on computational science and engineering*, IEEE, vol. 4, 2009, pp. 144–150.

[83] L. Dong *et al.*, "Defining a city—delineating urban areas using cell-phone data", *Nature Cities*, vol. 1, no. 2, pp. 117–125, 2024.

[84] Z. C. Steinert-Threlkeld, D. Mocanu, A. Vespignani, and J. Fowler, "Online social networks and offline protest", *EPJ Data Science*, vol. 4, no. 1, pp. 1–9, 2015.

[85] J. T. Jost *et al.*, "How social media facilitates political protest: Information, motivation, and social networks", *Political psychology*, vol. 39, pp. 85–118, 2018.

[86] D. Christensen and F. Garfias, "Can you hear me now? how communication technology affects protest and repression", *Quarterly journal of political science*, vol. 13, no. 1, p. 89, 2018.

[87] F. Calabrese, F. C. Pereira, G. Di Lorenzo, L. Liu, and C. Ratti, "The geography of taste: Analyzing cell-phone mobility and social events", in *Pervasive Computing: 8th International Conference, Pervasive 2010, Helsinki, Finland, May 17-20, 2010. Proceedings 8*, Springer, 2010, pp. 22–37.

[88] A. Rotman and M. Shalev, "Using location data from mobile phones to study participation in mass protests", *Sociological Methods & Research*, vol. 51, no. 3, pp. 1357–1412, 2022.

[89] M. K. Chen and R. Rohla, "The effect of partisanship and political advertising on close family ties", *Science*, vol. 360, no. 6392, pp. 1020–1024, 2018.

[90] A. Salas, P. Georgakis, and Y. Petalas, "Incident detection using data from social media", in *2017 IEEE 20th International conference on intelligent transportation systems (ITSC)*, IEEE, 2017, pp. 751–755.

[91] A. Janecek, D. Valerio, K. A. Hummel, F. Ricciato, and H. Hlavacs, "The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring", *IEEE transactions on intelligent transportation systems*, vol. 16, no. 5, pp. 2551–2572, 2015.

[92] M. Zhang, T. Li, Y. Yu, Y. Li, P. Hui, and Y. Zheng, "Urban anomaly analytics: Description, detection, and prediction", *IEEE Transactions on Big Data*, vol. 8, no. 3, pp. 809–826, 2020.

[93] X. Ren and C. Guan, "Evaluating geographic and social inequity of urban parks in shanghai through mobile phone-derived human activities", *Urban Forestry & Urban Greening*, vol. 76, p. 127 709, 2022.

[94] Y. Liu, A. Lu, W. Yang, and Z. Tian, "Investigating factors influencing park visit flows and duration using mobile phone signaling data", *Urban Forestry & Urban Greening*, vol. 85, p. 127 952, 2023.

[95] C. Guan, J. Song, M. Keith, Y. Akiyama, R. Shibasaki, and T. Sato, "Delineating urban park catchment areas using mobile phone data: A case study of tokyo", *Computers, Environment and Urban Systems*, vol. 81, p. 101 474, 2020.

[96] P. Sun, P. Liu, and Y. Song, "Seasonal variations in urban park characteristics and visitation patterns in atlanta: A big data study using smartphone user mobility", *Urban Forestry & Urban Greening*, vol. 91, p. 128 166, 2024. DOI: https://doi.org/10.1016/j.ufug.2023.128166. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1618866723003370.

[97] S. Guo *et al.*, "Accessibility to urban parks for elderly residents: Perspectives from mobile phone data", *Landscape and urban planning*, vol. 191, p. 103 642, 2019.

[98] M. Tizzoni *et al.*, "On the use of human mobility proxies for modeling epidemics", *PLoS computational biology*, vol. 10, no. 7, e1003716, 2014.

[99] M. V. Kiang *et al.*, "Incorporating human mobility data improves forecasts of dengue fever in thailand", *Scientific reports*, vol. 11, no. 1, p. 923, 2021.

[100] N. Vogel, C. Theisen, J. P. Leidig, J. Scripps, D. H. Graham, and G. Wolffe, "Mining mobile datasets to enable the fine-grained stochastic simulation of ebola diffusion", *Procedia Computer Science*, vol. 51, pp. 765–774, 2015.

[101] F. A. Ihantamalala *et al.*, "Estimating sources and sinks of malaria parasites in madagascar", *Nature communications*, vol. 9, no. 1, p. 3897, 2018.

[102] S. Rubrichi, Z. Smoreda, and M. Musolesi, "A comparison of spatial-based targeted disease mitigation strategies using mobile phone data", *EPJ Data Science*, vol. 7, no. 1, pp. 1–15, 2018.

[103] C. Panigutti, M. Tizzoni, P. Bajardi, Z. Smoreda, and V. Colizza, "Assessing the use of mobile phone data to describe recurrent mobility patterns in spatial epidemic models", *Royal Society open science*, vol. 4, no. 5, p. 160 950, 2017.

[104] T. Yabe, P. S. C. Rao, and S. V. Ukkusuri, "Resilience of interdependent urban socio-physical systems using large-scale mobility data: Modeling recovery dynamics", *Sustainable Cities and Society*, vol. 75, p. 103 237, 2021.

[105] R. Wilson *et al.*, "Rapid and near real-time assessments of population displacement using mobile phone data following disasters: The 2015 nepal earthquake", *PLoS currents*, vol. 8, 2016.

[106] T. Yabe, N. K. Jones, P. S. C. Rao, M. C. Gonzalez, and S. V. Ukkusuri, "Mobile phone location data for disasters: A review from natural hazards and epidemics", *Computers, Environment and Urban Systems*, vol. 94, p. 101 777, 2022.

[107] A. Feldmann *et al.*, "The lockdown effect: Implications of the covid-19 pandemic on internet traffic", in *Proceedings of the ACM Internet Measurement Conference*, ser. IMC '20, Virtual Event, USA: Association for Computing Machinery, 2020, pp. 1–18. DOI: 10.1145/3419394.3423658. [Online]. Available: https://doi.org/10.1145/3419394.3423658.

[108] A. Feldmann *et al.*, "A year in lockdown: How the waves of covid-19 impact internet traffic", *Commun. ACM*, vol. 64, no. 7, pp. 101–108, Jun. 2021. DOI: 10.1145/3465212. [Online]. Available: https://doi.org/10.1145/3465212.

[109] S. Liu, P. Schmitt, F. Bronzino, and N. Feamster, "Characterizing service provider response to the covid-19 pandemic in the united states", in *Passive and Active Measurement: 22nd International Conference, PAM 2021, Virtual Event, March 29–April 1, 2021, Proceedings 22*, Springer, 2021, pp. 20–38.

[110] T. Böttger, G. Ibrahim, and B. Vallis, "How the internet reacted to covid-19: A perspective from facebook's edge network", in *Proceedings of the ACM Internet Measurement Conference*, 2020, pp. 34–41.

[111] T. Favale, F. Soro, M. Trevisan, I. Drago, and M. Mellia, "Campus traffic and e-learning during covid-19 pandemic", *Computer networks*, vol. 176, p. 107 290, 2020.

[112] M. Candela, V. Luconi, and A. Vecchio, "Impact of the covid-19 pandemic on the internet latency: A large-scale study", *Computer Networks*, vol. 182, p. 107 495, 2020. DOI: https://doi.org/10.1016/j.comnet.2020.107495. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128620311622.

[113] A. Lutu, D. Perino, M. Bagnulo, E. Frias-Martinez, and J. Khangosstar, "A characterization of the covid-19 pandemic impact on a mobile network operator traffic", in *Proceedings of the ACM Internet Measurement Conference*, ser. IMC '20, Virtual Event, USA: Association for Computing Machinery, 2020, pp. 19–33. DOI: 10.1145/3419394.3423655. [Online]. Available: https://doi.org/10.1145/3419394.3423655.

[114] G. Pullano, E. Valdano, N. Scarpa, S. Rubrichi, and V. Colizza, "Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the covid-19 epidemic in france under lockdown: A population-based study", *The Lancet Digital Health*, vol. 2, no. 12, e638–e649, 2020.

[115] L. Di Domenico, G. Pullano, C. E. Sabbatini, P.-Y. Boëlle, and V. Colizza, "Modelling safe protocols for reopening schools during the covid-19 pandemic in france", *Nature communications*, vol. 12, no. 1, p. 1073, 2021.

[116] E. Valdano, J. Lee, S. Bansal, S. Rubrichi, and V. Colizza, "Highlighting socio-economic constraints on mobility reductions during covid-19 restrictions in france can inform effective and equitable pandemic response", *Journal of travel medicine*, vol. 28, no. 4, taab045, 2021.

[117] F. Calabrese, Z. Smoreda, V. D. Blondel, and C. Ratti, "Interplay between telecommunications and face-to-face interactions: A study using mobile phone data", *PloS one*, vol. 6, no. 7, e20814, 2011.

[118] M. Schläpfer *et al.*, "The scaling of human interactions with city size", *Journal of the Royal Society Interface*, vol. 11, no. 98, p. 20 130 789, 2014.

[119] R. Ahas *et al.*, "Everyday space–time geographies: Using mobile phone-based sensor data to monitor urban activity in harbin, paris, and tallinn", *International Journal of Geographical Information Science*, vol. 29, no. 11, pp. 2017–2039, 2015.

[120] P. Widhalm, Y. Yang, M. Ulm, S. Athavale, and M. C. González, "Discovering urban activity patterns in cell phone data", *Transportation*, vol. 42, pp. 597–623, 2015.

[121] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in rome", *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 141–151, 2011. DOI: 10.1109/TITS.2010.2074196.

[122] S. Sobolevsky, M. Szell, R. Campari, T. Couronné, Z. Smoreda, and C. Ratti, "Delineating geographical regions with networks of human interactions in an extensive set of countries", *PloS one*, vol. 8, no. 12, e81707, 2013.

[123] A. Furno, R. Stanica, and M. Fiore, "A comparative evaluation of urban fabric detection techniques based on mobile traffic data", in *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015*, 2015, pp. 689–696.

[124] G. Khodabandelou, V. Gauthier, M. El-Yacoubi, and M. Fiore, "Population estimation from mobile network traffic metadata", in *2016 IEEE 17th international symposium on a world of wireless, mobile and multimedia networks (WoWMoM)*, IEEE, 2016, pp. 1–9.

[125] B. C. Csáji *et al.*, "Exploring the mobility of mobile phone users", *Physica A: statistical mechanics and its applications*, vol. 392, no. 6, pp. 1459–1473, 2013.

[126] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González, "Unravelling daily human mobility motifs", *Journal of The Royal Society Interface*, vol. 10, no. 84, p. 20 130 246, 2013.

[127] M. S. Iqbal, C. F. Choudhury, P. Wang, and M. C. González, "Development of origin–destination matrices using mobile phone call data", *Transportation Research Part C: Emerging Technologies*, vol. 40, pp. 63–74, 2014.

[128] J. L. Toole, S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González, "The path most traveled: Travel demand estimation using big data resources", *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 162–177, 2015.

[129] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin–destination trips by purpose and time of day inferred from mobile phone data", *Transportation research part c: emerging technologies*, vol. 58, pp. 240–250, 2015.

[130] L. Pappalardo, L. Ferres, M. Sacasa, C. Cattuto, and L. Bravo, "Evaluation of home detection algorithms on mobile phone data using individual-level ground truth", *EPJ data science*, vol. 10, no. 1, p. 29, 2021.

[131] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabási, "Returners and explorers dichotomy in human mobility", *Nature communications*, vol. 6, no. 1, p. 8166, 2015.

[132] A. Furno, N.-E. El Faouzi, M. Fiore, and R. Stanica, "Fusing gps probe and mobile phone data for enhanced land-use detection", in *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, IEEE, 2017, pp. 693–698.

[133] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti, "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example", *Transportation Research Part C: Emerging Technologies*, vol. 26, pp. 301–313, 2013. DOI: https://doi.org/10.1016/j.trc.2012.09.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X12001192.

[134] S. Çolak, A. Lima, and M. C. González, "Understanding congested travel in urban areas", *Nature communications*, vol. 7, no. 1, p. 10 793, 2016.

[135] S. Phithakkitnukoon, Z. Smoreda, and P. Olivier, "Socio-geography of human mobility: A study using longitudinal mobile phone data", *PloS one*, vol. 7, no. 6, e39253, 2012.

[136] L. Pappalardo, D. Pedreschi, Z. Smoreda, and F. Giannotti, "Using big data to study the link between human mobility and socio-economic development", in *2015 IEEE International Conference on Big Data (Big Data)*, IEEE, 2015, pp. 871–878.

[137] P. Bonnel, E. Hombourger, A.-M. Olteanu-Raimond, and Z. Smoreda, "Passive mobile phone dataset to construct origin-destination matrix: Potentials and limitations", *Transportation Research Procedia*, vol. 11, pp. 381–398, 2015.

[138] S. Hoteit, G. Chen, A. Viana, and M. Fiore, "Filling the gaps: On the completion of sparse call detail records for mobility analysis", in *Proceedings of the eleventh ACM workshop on challenged networks*, 2016, pp. 45–50.

[139] G. Chen, S. Hoteit, A. C. Viana, M. Fiore, and C. Sarraute, "Enriching sparse mobility information in call detail records", *Computer Communications*, vol. 122, pp. 44–58, 2018.

[140] G. Chen, A. C. Viana, M. Fiore, and C. Sarraute, "Complete trajectory reconstruction from sparse mobile phone data", *EPJ Data Science*, vol. 8, no. 1, pp. 1–24, 2019.

[141] P. Katsikouli, A. C. Viana, M. Fiore, and A. Tarable, "On the sampling frequency of human mobility", in *GLOBECOM 2017-2017 IEEE Global Communications Conference*, IEEE, 2017, pp. 1–6.

[142] L. Bonnetain *et al.*, "Transit: Fine-grained human mobility trajectory inference at scale with mobile network signaling data", *Transportation Research Part C: Emerging Technologies*, vol. 130, p. 103 257, 2021.

[143] M. Gramaglia, M. Fiore, A. Tarable, and A. Banchs, "Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories", in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, IEEE, 2017, pp. 1–9.

[144] M. Fiore *et al.*, "Privacy in trajectory micro-data publishing: A survey", *Transactions on Data Privacy*, vol. 13, no. 2, pp. 91–149, 2020.

[145] M. Luca, G. Barlacchi, B. Lepri, and L. Pappalardo, "A survey on deep learning for human mobility", *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–44, 2021.

[146] Z. Qiu, J. Jin, P. Cheng, and B. Ran, "State of the art and practice: Cellular probe technology applied in advanced traveler information systems", *Transportation Research Board 86th Annual MeetingTransportation Research Board*, no. 07-0223, 2007.

[147] H. Bar-Gera, "Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel", *Transportation Research Part C: Emerging Technologies*, vol. 15, no. 6, pp. 380–391, 2007.

[148] A. Janecek, K. A. Hummel, D. Valerio, F. Ricciato, and H. Hlavacs, "Cellular data meet vehicular traffic theory: Location area updates and cell transitions for travel time estimation", in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 2012, pp. 361–370.

[149] N. Caceres, L. M. Romero, F. G. Benitez, and J. M. del Castillo, "Traffic flow estimation models using cellular phone data", *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1430–1441, 2012.

[150] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in rome", *IEEE transactions on intelligent transportation systems*, vol. 12, no. 1, pp. 141–151, 2010.

[151] H. Wang, F. Calabrese, G. Di Lorenzo, and C. Ratti, "Transportation mode inference from anonymized and aggregated mobile phone call detail records", in *13th International IEEE Conference on Intelligent Transportation Systems*, IEEE, 2010, pp. 318–323.

[152] J. Doyle, P. Hung, D. Kelly, S. F. McLoone, and R. Farrell, "Utilising mobile phone billing records for travel mode discovery", 2011.

[153] M. Zilske and K. Nagel, *Building a minimal traffic model from mobile phone data.* Technische Universität Berlin, 2019.

[154] M. Berlingerio, F. Calabrese, G. Di Lorenzo, R. Nair, F. Pinelli, and M. L. Sbodio, "Allaboard: A system for exploring urban mobility and optimizing public transport using cellphone data", in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, Springer, 2013, pp. 663–666.

[155] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, "Exploring human mobility with multi-source data at extremely large metropolitan scales", in *Proceedings of the 20th annual international conference on Mobile computing and networking*, 2014, pp. 201–212.

[156] G. Barlacchi *et al.*, "A multi-source dataset of urban life in the city of milan and the province of trentino", *Scientific data*, vol. 2, no. 1, pp. 1–15, 2015.

[157] V. D. Blondel *et al.*, "Data for development: The d4d challenge on mobile phone data", *arXiv preprint arXiv:1210.0137*, 2012.

[158] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel, "D4d-senegal: The second mobile phone data for development challenge", *arXiv preprint arXiv:1407.4885*, 2014.

[159] O. E. Martínez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda, and M. Fiore, "The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography", *arXiv preprint arXiv:2305.06933*, 2023.

[160] Q. Xu, A. Gerber, Z. M. Mao, and J. Pang, "Acculoc: Practical localization of performance measurements in 3G networks", in *ACM MobiSys '11*, Bethesda, Maryland, USA, 2011, pp. 183–196. DOI: 10.1145/1999995.2000013. [Online]. Available: https://doi.org/10.1145/1999995.2000013.

[161] F. Metzger, A. Rafetseder, P. Romirer-Maierhofer, and K. Tutschku, "Exploratory analysis of a ggsn's pdp context signaling load", *Journal of Computer Networks and Communications*, no. 526231, 2014. DOI: https://doi.org/10.1155/2014/526231.

[162] European Union, *Eu general data protection regulation (gdpr): Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)*, Jun. 2016. [Online]. Available: https://gdpr-info.eu/.

[163] B. Balassa, "Trade liberalisation and "revealed" comparative advantage 1", *The manchester school*, vol. 33, no. 2, pp. 99–123, 1965.

[164] K. Laursen and C. Engedal, "The role of the technology factor in economic growth: A theoretical and empirical inquiry into new approaches to economic growth", *Unpublished MA dissertation*, 1995.

[165] K. Laursen, "Revealed comparative advantage and the alternatives as measures of international specialization", *Eurasian business review*, vol. 5, pp. 99–115, 2015.

[166] F. Aurenhammer, "Voronoi diagrams—a survey of a fundamental geometric data structure", *ACM Computing Surveys (CSUR)*, vol. 23, no. 3, pp. 345–405, 1991.

[167] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks", in *2011 Proceedings IEEE INFOCOM*, IEEE, 2011, pp. 882–890.

[168] INSEE, *Definition - IRIS / Insee — insee.fr*, https://www.insee.fr/en/metadonnees/definition/c1523, [Accessed 17-Nov-2022], 2016.

[169] C. K. Kim and S. Ian, "Examining applied multicultural industrial and organizational psychology", in IGI Global, 2023, ch. Why Do People Still Buy Apple Products?: Applying Psychological Modeling to Brand Image Management and Cultural Business Ecosystems. DOI: doi.org/10.4018/978-1-6684-7212-5.ch010.

[170] Cisco, "White paper: Cisco vision: 5G-thriving indoors", Cisco, Tech. Rep., 2017. [Online]. Available: https://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/ultra-services-platform/5g-ran-indoor.pdf.

[171] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey", *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 2016.

[172] Ericsson, "White paper: Bringing 5G networks indoors", Ericsson, Tech. Rep., 2017. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/white-papers/bringing-5g-networks-indoors.

[173] Huawei, "White paper: Indoor 5G networks", Huawei, Tech. Rep., 2018. [Online]. Available: https://carrier.huawei.com/minisite/Indoor-5G/pdf/Indoor-5G-Networks-White-Paper-V2.0-en.pdf.

[174] ZTE, "White paper: 5G indoor", ZTE, Tech. Rep., 2019. [Online]. Available: https://www.zte.com.cn/content/dam/zte-site/res-www-zte-com-cn/mediares/zte/files/newsolution/wireless/ran/white_paper/ZTE_5G_Indoor_White_Paper-EN.pdf.

[175] A. M. Al-Samman *et al.*, "Comparative study of indoor propagation model below and above 6 GHz for 5G wireless networks", *Electronics*, vol. 8, no. 1, p. 44, 2019.

[176] J. Medbo *et al.*, "Radio propagation modeling for 5G mobile and wireless communications", *IEEE communications magazine*, vol. 54, no. 6, pp. 144–151, 2016.

[177] S. Bakirtzis, J. Chen, K. Qiu, J. Zhang, and I. Wassell, "EM DeepRay: An expedient, generalizable and realistic data-driven indoor propagation model", *IEEE Trans. Antennas Propag.*, vol. 70, no. 6, pp. 4140–4154, 2022.

[178] Y. Huang, J. Zhang, and J. Zhang, "Wireless channel delay spread performance evaluation of a building layout", arXiv preprint arXiv:2212.05656, 2022.

[179] J. Zhang, A. A. Glazunov, and J. Zhang, "Wireless performance evaluation of building layouts: Closed-form computation of figures of merit", *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4890–4906, 2021.

[180] J. Zhang, A. A. Glazunov, W. Yang, and J. Zhang, "Fundamental wireless performance of a building", *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 186–193, 2021.

[181] S. Bakirtzis, I. Wassell, M. Fiore, and J. Zhang, "Stochastic evaluation of indoor wireless network performance with data-driven propagation models", in *GLOBECOM 2022 IEEE Global Communications Conference*, New York, NY, USA: IEEE, 2022, pp. 3587–3592.

[182] X. Ge, S. Tu, G. Mao, C.-X. Wang, and T. Han, "5G ultra-dense cellular networks", *IEEE Wireless Commun.*, vol. 23, no. 1, pp. 72–79, 2016.

[183] S. F. Yunas, M. Valkama, and J. Niemelä, "Spectral and energy efficiency of ultra-dense networks under different deployment strategies", *IEEE Com. Mag.*, vol. 53, no. 1, pp. 90–100, 2015.

[184] S. F. Yunas, A. Asp, J. Niemela, and M. Valkama, "Deployment strategies and performance analysis of macrocell and femtocell networks in suburban environment with modern buildings", in *39th Annual IEEE Conference on Local Computer Networks Workshops*, New York, NY, USA: IEEE, 2014, pp. 643–651.

[185] S. Bakirtzis, M. Fiore, I. Wassell, and J. Zhang, "Expedient ai-assisted indoor wireless network planning with data-driven propagation models", TechRxiv, 2023.

[186] A. Aijaz, "Private 5g: The future of industrial wireless", *IEEE Ind. Electron. Magazine*, vol. 14, no. 4, pp. 136–145, 2020.

[187] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function", *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[188] P. C. *et al.*, "Ericsson mobility report – june 2021", Ericsson, 2021.

[189] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary users in cellular networks: A large-scale measurement study", in *2008 3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks*, IEEE, 2008, pp. 1–11.

[190] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang, "A first look at cellular network performance during crowded events", *ACM SIGMETRICS performance evaluation review*, vol. 41, no. 1, pp. 17–28, 2013.

[191] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute, "Measurement-driven mobile data traffic modeling in a large metropolitan area", in *2015 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, IEEE, 2015, pp. 230–235.

[192] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng, "Characterizing user behavior in mobile internet", *IEEE transactions on emerging topics in computing*, vol. 3, no. 1, pp. 95–106, 2014.

[193] E. Peltonen *et al.*, "The hidden image of mobile apps: Geographic, demographic, and cultural factors in mobile usage", in *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2018, pp. 1–12.

[194] M. Fekih *et al.*, "Potential of cellular signaling data for time-of-day estimation and spatial classification of travel demand: A large-scale comparative study with travel survey and land use data", *Transportation Letters*, vol. 14, no. 7, pp. 787–805, 2022.

[195] Y. Chen, Z. Wang, H. Sun, Y. Zhang, and Z. He, "Analysis of travel demand between transportation hubs in urban agglomeration based on mobile phone call detail record data", *Journal of Transportation Engineering, Part A: Systems*, vol. 148, no. 7, p. 04 022 041, 2022.

[196] N. Breyer, C. Rydergren, and D. Gundlegård, "Semi-supervised mode classification of inter-city trips from cellular network data", *Journal of Big Data Analytics in Transportation*, vol. 4, no. 1, pp. 23–39, 2022.

[197] C. Spearman, """ general intelligence" objectively determined and measured.", 1961.

[198] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan, "Evaluating the use of exploratory factor analysis in psychological research.", *Psychological methods*, vol. 4, no. 3, p. 272, 1999.

[199] J. Camacho, R. Bro, and D. Kotz, "Automatic learning coupled with interpretability: Mbda in action", 2020.

[200] S. A. Mulaik, *Foundations of factor analysis*. CRC press, 2009.

[201] D. N. Lawley, "Vi.—the estimation of factor loadings by the method of maximum likelihood", *Proceedings of the Royal Society of Edinburgh*, vol. 60, no. 1, pp. 64–82, 1940.

[202] K. G. Jöreskog, "Structural analysis of covariance and correlation matrices", *Psychometrika*, vol. 43, no. 4, pp. 443–477, 1978.

[203] H. H. Harman and W. H. Jones, "Factor analysis by minimizing residuals (minres)", *Psychometrika*, vol. 31, no. 3, pp. 351–368, 1966.

[204] N. E. Briggs and R. C. MacCallum, "Recovery of weak common factors by maximum likelihood and ordinary least squares estimation", *Multivariate Behavioral Research*, vol. 38, no. 1, pp. 25–56, 2003.

[205] A. L. Comrey and H. B. Lee, *A first course in factor analysis*. Psychology press, 2013.

[206] H. F. Kaiser and J. Rice, "Little jiffy, mark iv", *Educational and psychological measurement*, vol. 34, no. 1, pp. 111–117, 1974.

[207] R. B. Cattell, "The scree test for the number of factors", *Multivariate behavioral research*, vol. 1, no. 2, pp. 245–276, 1966.

[208] J. L. Horn, "A rationale and test for the number of factors in factor analysis", *Psychometrika*, vol. 30, pp. 179–185, 1965.

[209] B. Thompson, "Exploratory and confirmatory factor analysis: Understanding concepts and applications", *Washington, DC*, vol. 10694, no. 000, p. 3, 2004.

[210] H. F. Kaiser, "The varimax criterion for analytic rotation in factor analysis", *Psychometrika*, vol. 23, no. 3, pp. 187–200, 1958.

[211] J. C. de Winter*, D. Dodou*, and P. A. Wieringa, "Exploratory factor analysis with small sample sizes", *Multivariate behavioral research*, vol. 44, no. 2, pp. 147–181, 2009.

[212] H. Assem, T. S. Buda, and L. Xu, "Initial use cases, scenarios and requirements", *H2020 5G-PPP CogNet, Deliverable D*, vol. 2, p. 1, 2015.

[213] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, "Big data-driven optimization for mobile networks toward 5g", *IEEE network*, vol. 30, no. 1, pp. 44–51, 2016.

[214] S. Wu, B. Chen, C. Webster, B. Xu, and P. Gong, "Improved human greenspace exposure equality during 21st century urbanization", *Nature Communications*, vol. 14, no. 1, Oct. 2023. DOI: 10.1038/s41467-023-41620-z.

[215] W. H. Organization, *Who coronavirus (covid-19) dashboard*. [Online]. Available: https://covid19.who.int/info/ (visited on 07/27/2021).

[216] E. Mathieu *et al.*, "A global database of COVID-19 vaccinations", *Nature Human Behaviour*, vol. 5, no. 7, pp. 947–953, May 2021. DOI: 10.1038/s41562-021-01122-8. [Online]. Available: https://doi.org/10.1038/s41562-021-01122-8.

[217] Ericsson, "Ericsson Mobility Report", Tech. Rep., Nov. 2019.

[218] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. DOI: https://doi.org/10.1016/0377-0427(87)90125-7.

[219] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters", *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973. DOI: 10.1080/01969727308546046. [Online]. Available: https://doi.org/10.1080/01969727308546046.

[220] H. Zang and J. C. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks", in *Proceedings of the 13th annual ACM international conference on Mobile computing and networking*, 2007, pp. 123–134.

[221] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, "Estimating origin-destination flows using mobile phone location data", *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 36–44, 2011. DOI: 10.1109/MPRV.2011.41.

[222] L. Anselin and O. Smirnov, "Efficient algorithms for constructing proper higher order spatial lag operators", *Journal of regional science*, vol. 36, no. 1, pp. 67–89, 1996.

[223] Insee, *Recensement 2017 : Résultats sur un territoire, bases de données et fichiers détail*, 2020. [Online]. Available: https://www.insee.fr/fr/information/4467366.

[224] Insee, *Revenus, pauvreté et niveau de vie en 2018 (iris)*, https://www.insee.fr/fr/statistiques/5055909, Accessed: 2022-10-05, 2021.

[225] *Base sirene des entreprises et de leurs établissements (siren, siret)*, https://www.data.gouv.fr/fr/datasets/base-sirene-des-entreprises-et-de-leurs-etablissements-siren-siret/, Accessed: 2022-10-05.

[226] L. Anselin and A. K. Bera, "Spatial dependence in linear regression models with an introduction to spatial econometrics", *Statistics textbooks and monographs*, vol. 155, pp. 237–290, 1998.

[227] P. J. Huber, *Robust statistics*, 2004.

[228] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares", *Communications in Statistics-theory and Methods*, vol. 6, no. 9, pp. 813–827, 1977.

[229] N. Alsaedi, P. Burnap, and O. Rana, "Can we predict a riot? disruptive event detection using twitter", *ACM Transactions on Internet Technology (TOIT)*, vol. 17, no. 2, pp. 1–26, 2017.

[230] C. Wu and M. S. Gerber, "Forecasting civil unrest using social media and protest participation theory", *IEEE Transactions on Computational Social Systems*, vol. 5, no. 1, pp. 82–94, 2017.

[231] J. Van Laer, "Activists online and offline: The internet as an information channel for protest demonstrations", *Mobilization: An International Quarterly*, vol. 15, no. 3, pp. 347–366, 2010.

[232] B. Enjolras, K. Steen-Johnsen, and D. Wollebaek, "Social media and mobilization to offline demonstrations: Transcending participatory divides?", *New media & society*, vol. 15, no. 6, pp. 890–908, 2013.

[233] J. Liu, "Communicating beyond information? mobile phones and mobilization to offline protests in china", *Television & New Media*, vol. 16, no. 6, pp. 503–520, 2015.

[234] S. Milan and S. Barbosa, "Enter the whatsapper: Reinventing digital activism at the time of chat apps", *First Monday*, 2020.

[235] A. Urman, J. C.-t. Ho, and S. Katz, "Analyzing protest mobilization on telegram: The case of 2019 anti-extradition bill movement in hong kong", *Plos one*, vol. 16, no. 10, e0256675, 2021.

[236] M. R. Albrecht, J. Blasco, R. B. Jensen, and L. Mareková, "Collective information security in {large-scale} urban protests: The case of hong kong", in *30th USENIX security symposium (USENIX Security 21)*, 2021, pp. 3363–3380.

[237] S. C. Pearce and J. Rodgers, "Social media as public journalism? protest reporting in the digital era", *Sociology Compass*, vol. 14, no. 12, pp. 1–14, 2020.

[238] C. Neumayer and L. Rossi, "Images of protest in social media: Struggle over visibility and visual narratives", *New Media & Society*, vol. 20, no. 11, pp. 4293–4310, 2018.

[239] O. Jenzen, I. Erhart, H. Eslen-Ziya, U. Korkut, and A. McGarry, "The symbol of social media in contemporary protest: Twitter and the gezi park movement", *Convergence*, vol. 27, no. 2, pp. 414–437, 2021.

[240] G. Grill, "Future protest made risky: Examining social media based civil unrest prediction research and products", *Computer Supported Cooperative Work (CSCW)*, vol. 30, no. 5, pp. 811–839, 2021.

[241] C. Neumayer and G. Stald, "The mobile phone in street protest: Texting, tweeting, tracking, and tracing", *Mobile Media & Communication*, vol. 2, no. 2, pp. 117–133, 2014.

[242] A. Hermida and V. Hernández-Santaolalla, "Twitter and video activism as tools for counter-surveillance: The case of social protests in spain", *Information, Communication & Society*, vol. 21, no. 3, pp. 416–433, 2018.

[243] T.-y. Ting, "From 'be water'to 'be fire': Nascent smart mob and networked protests in hong kong", *Social Movement Studies*, vol. 19, no. 3, pp. 362–368, 2020.

[244] P. Rozenshtein, A. Anagnostopoulos, A. Gionis, and N. Tatti, "Event detection in activity networks", in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1176–1185.

[245] R. Korolov *et al.*, "On predicting social unrest using social media", in *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, IEEE, 2016, pp. 89–95.

[246] S. L. Wilson, "Detecting mass protest through social media", *The Journal of Social Media in Society*, vol. 6, no. 2, pp. 5–25, 2017.

[247] A. M. Ertugrul, Y.-R. Lin, W.-T. Chung, M. Yan, and A. Li, "Activism via attention: Interpretable spatiotemporal learning to forecast protest activities", *EPJ Data Science*, vol. 8, no. 1, p. 5, 2019.

[248] D. Won, Z. C. Steinert-Threlkeld, and J. Joo, "Protest activity detection and perceived violence estimation from social media images", in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 786–794.

[249] V. A. Traag, R. Quax, and P. M. Sloot, "Modelling the distance impedance of protest attendance", *Physica A: Statistical Mechanics and its Applications*, vol. 468, pp. 171–182, 2017.

[250] D. Birant and A. Kut, "St-dbscan: An algorithm for clustering spatial–temporal data", *Data & knowledge engineering*, vol. 60, no. 1, pp. 208–221, 2007.

[251] R. W. Douglass, D. A. Meyer, M. Ram, D. Rideout, and D. Song, "High resolution population estimates from telecommunications data", *EPJ Data Science*, vol. 4, pp. 1–13, 2015.

[252] G. Khodabandelou, V. Gauthier, M. Fiore, and M. A. El-Yacoubi, "Estimation of static and dynamic urban populations with mobile network metadata", *IEEE Transactions on Mobile Computing*, vol. 18, no. 9, pp. 2034–2047, 2018.

[253] 3GPP TS 23.288 v16.1.0, *Architecture Enhancements for 5G System (5GS) to Support Network Data Analytics Services (Release 16)*, Jun. 2019.

[254] 3GPP TS 28.533 v16.0.0, *Management and Orchestration of Networks and Network Slicing; Management and Orchestration Architecture (Release 16)*, Jun. 2019.

[255] O-RAN.WG3.RICARCH-v02.01, *O-RAN Near-RT RAN Intelligent Controller Near-RT RIC Architecture 2.01*, Mar. 2022.

[256] O-RAN.WG2.Non-RT-RIC-ARCH-TS-v01.00, *O-RAN Non-RT RIC Architecture 1.0*, Oct. 2021.

[257] J. Lee *et al.*, "Perceive: Deep learning-based cellular uplink prediction using real-time scheduling patterns", in *ACM MobiSys '20*, 2020, pp. 377–390.

[258] M. Polese, F. Restuccia, and T. Melodia, "Deepbeam: Deep waveform learning for coordination-free beam management in mmwave networks", in *MobiHoc '21*, Shanghai, China: ACM MobiHoc '21, 2021, pp. 61–70.

[259] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "Aztec: Anticipatory capacity allocation for zero-touch network slicing", in *IEEE INFOCOM '20*, 2020, pp. 794–803.

[260] C. Zhang, M. Fiore, C. Ziemlicki, and P. Patras, "Microscope: Mobile service traffic decomposition for network slicing as a service", in ACM MobiCom '20, 2020.

[261] J. A. Ayala-Romero, A. Garcia-Saavedra, M. Gramaglia, X. Costa-Perez, A. Banchs, and J. J. Alcaraz, "Vrain: A deep learning approach tailoring computing and radio resources in virtualized RANs", in *ACM MobiCom '19*, 2019. [Online]. Available: https://doi.org/10.1145/3300061.3345431.

[262] ETSI, "GS ZSM 001 V1.1.1 – Zero-touch network and Service Management (ZSM); Requirements based on documented scenarios", 2019.

[263] 3GPP Technical Specification Group Services and System Aspects, "TR:28.812 – Study on scenarios for Intent driven management services for mobile networks, Telecommunication management", 2020.

[264] J. Huang and M. Xiao, "Mobile network traffic prediction based on seasonal adjacent windows sampling and conditional probability estimation", *IEEE Transactions on Big Data*, pp. 1–1, 2020. DOI: 10.1109/TBDATA.2020.3014049.

[265] D. Kim *et al.*, "Design and implementation of traffic generation model and spectrum requirement calculator for private 5g network", *IEEE Access*, vol. 10, pp. 15 978–15 993, 2022. DOI: 10.1109/ACCESS.2022.3149050.

[266] Y.-T. Lin, T. Bonald, and S. E. Elayoubi, "Flow-level traffic model for adaptive streaming services in mobile networks", *Computer Networks*, vol. 137, pp. 1–16, 2018. DOI: https://doi.org/10.1016/j.comnet.2018.01.027. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128618300318.

[267] H. Khedher *et al.*, "Real traffic-aware scheduling of computing resources in cloud-ran", in *ICNC '20*, 2020, pp. 422–427. DOI: 10.1109/ICNC47757.2020.9049679.

[268] S. Wu, Y. Wang, and L. Bai, "Deep convolutional neural network assisted reinforcement learning based mobile network power saving", *IEEE Access*, vol. 8, pp. 93 671–93 681, 2020. DOI: 10.1109/ACCESS.2020.2995057.

[269] B. Bojovic and S. Lagen, "Enabling NGMN mixed traffic models for Ns-3", in *Proc. Workshop on Ns-3*, Virtual Event, USA: ACM WNS3 '22, 2022, pp. 127–134. DOI: 10.1145/3532577.3532602. [Online]. Available: https://doi.org/10.1145/3532577.3532602.

[270] Deezer Support, *Deezer audio quality*, https://support.deezer.com/hc/en-gb/articles/115003865685-Deezer-Audio-Quality, Accessed: 2022-05-31, 2022.

[271] A. Ramdas, N. G. Trillos, and M. Cuturi, "On wasserstein two-sample testing and related families of nonparametric tests", *Entropy*, vol. 19, no. 2, 2017. DOI: 10.3390/e19020047. [Online]. Available: https://www.mdpi.com/1099-4300/19/2/47.

[272] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms", *arXiv preprint arXiv:1109.2378*, 2011.

[273] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures.", *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964. DOI: 10.1021/ac60214a047.

[274] I. Tsompanidis, A. H. Zahran, and C. J. Sreenan, "Mobile network traffic: A user behaviour model", in *2014 7th IFIP Wireless and Mobile Networking Conference (WMNC)*, 2014, pp. 1–8. DOI: 10.1109/WMNC.2014.6878862.

[275] H. Gupta, M. Sharma, A. Franklin A., and B. R. Tamma, "Apt-ran: A flexible split-based 5g ran to minimize energy consumption and handovers", *IEEE Transactions on Network and Service Management*, vol. 17, no. 1, 2020.

[276] R. Singh, C. Hasan, X. Foukas, M. Fiore, M. K. Marina, and Y. Wang, "Energy-efficient orchestration of metro-scale 5G radio access networks", in *IEEE INFOCOM '21*, 2021, pp. 1–10. DOI: 10.1109/INFOCOM42981.2021.9488786.

[277] S. Rawas, "Energy, network, and application-aware virtual machine placement model in SDN-enabled large scale cloud data centers", *Multimedia Tools and App.*, vol. 80, no. 10, 2021.

[278] D. Johnson, "Near-optimal bin packing algorithms", Ph.D. dissertation, 1973.

[279] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks", *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 248–257, 2012.

[280] W. Bao and B. Liang, "Structured spectrum allocation and user association in heterogeneous cellular networks", in *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, IEEE, 2014, pp. 1069–1077.

[281] F. Wang, W. Chen, H. Tang, and Q. Wu, "Joint optimization of user association, subchannel allocation, and power allocation in multi-cell multi-association ofdma heterogeneous networks", *IEEE Transactions on Communications*, vol. 65, no. 6, pp. 2672–2684, 2017.

[282] R. Borralho, A. Mohamed, A. U. Quddus, P. Vieira, and R. Tafazolli, "A survey on coverage enhancement in cellular networks: Challenges and solutions for future deployments", *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 1302–1341, 2021. DOI: 10.1109/COMST.2021.3053464.

[283] D. López-Pérez, A. De Domenico, N. Piovesan, G. Xinli, H. Bao, and S. Qitao, "A survey on 5G radio access network energy efficiency: Massive mimo, lean carrier design, sleep modes, and machine learning", *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 653–697, 2022. DOI: 10.1109/COMST.2022.3142532.

[284] A. De Domenico, E. Calvanese Strinati, and A. Capone, "Enabling green cellular networks: A survey and outlook", *Computer Communications*, vol. 37, pp. 5–24, 2014. DOI: https://doi.org/10.1016/j.comcom.2013.09.011.

[285] D. He, B. Ai, K. Guan, L. Wang, Z. Zhong, and T. Kürner, "The design and applications of high-performance ray-tracing simulation platform for 5G and beyond wireless communications: A tutorial", *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 10–27, 1st Quart., 2018.

[286] T. K. Sarkar, Z. Ji, K. Kim, A. Medouri, and M. Salazar-Palma, "A survey of various propagation models for mobile communication", *IEEE Antennas Propag. Magazine*, vol. 45, no. 3, pp. 51–82, Jun. 2003.

[287] P. Parastar, A. Lutu, Ö. Alay, G. Caso, and D. Perino, "Spotlight on 5g: Performance, device evolution and challenges from a mobile operator perspective", in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, IEEE, 2023, pp. 1–10.