



Thèse

2023

Open Access

This version of the publication is provided by the author(s) and made available in accordance with the copyright holder(s).

Context-aware Mobile Internet Quality Model: Quantifying and Facilitating Smartphone's Quality of Experience

De Masi, Alexandre

How to cite

DE MASI, Alexandre. Context-aware Mobile Internet Quality Model: Quantifying and Facilitating Smartphone's Quality of Experience. 2023. doi: 10.13097/archive-ouverte/unige:174205

This publication URL: <https://archive-ouverte.unige.ch//unige:174205>

Publication DOI: [10.13097/archive-ouverte/unige:174205](https://doi.org/10.13097/archive-ouverte/unige:174205)

© The author(s). This work is licensed under a Creative Commons Attribution (CC BY)

<https://creativecommons.org/licenses/by/4.0>

UNIVERSITÉ DE GENÈVE
Centre Universitaire d'Informatique
Programme Doctoral
Systèmes d'Information

FACULTÉ D'ECONOMIE ET MANAGEMENT
Prof. Dimitri Konstantas
Prof. Katarzyna Wac

Context-aware Mobile Internet Quality Model Quantifying and Facilitating Smartphone's Quality of Experience

THÈSE

présentée à la Faculté d'Économie et de Management de l'Université de Genève

Institut of Information Service Science

pour obtenir le grade de Docteur ès Systèmes d'Information

par

MSc Alexandre De Masi

de

Briey (France)

Members of the thesis committee:

Prof. Dimitri Konstantas (Jury President)

Prof. Katarzyna Wac (Thesis Director)

Prof. Markus Fiedler, Blekinge Institute of Technology

Dr. Niels Nijdam, University of Geneva

Dr. Jose Luis Fernandez Marquez, University of Geneva

Dr. Selim Ickin, Ericsson AB

Thèse N° 129

GENÈVE

Université de Genève

Août 2023

La Faculté d'économie et de management, sur préavis du jury, a autorisé l'impression de la présente thèse, sans entendre, par-là, émettre aucune opinion sur les propositions qui s'y trouvent énoncées et qui n'engagent que la responsabilité de leur auteur.

Genève, le 30 août 2023.

Dean

Markus MENZ



**UNIVERSITÉ
DE GENÈVE**

GENEVA SCHOOL OF ECONOMICS
AND MANAGEMENT

LE DOYEN

A Q U I D E D R O I T

I M P R I M A T U R

Je, soussigné, Professeur Markus MENZ, Doyen de la Faculté d'Economie et de Management, confirme que **Monsieur Alexandre DE MASI** obtient l'imprimatur pour sa thèse N°129, suite à sa soutenance publique du 24 août 2023 pour le grade de docteur en Systèmes d'Information.

Prof. Markus MENZ
Doyen

Genève, le 30 août 2023
MM/GK/kl

Context-aware Mobile Internet Quality Model

Quantifying and Facilitating Smartphone's Quality of Experience

Alexandre De Masi

Geneva School of Economics and Management
Institute of Information Service Science
University of Geneva
Switzerland

This dissertation is submitted for the degree of
Docteur ès Systèmes d'Information

August 2023

To Chen, your steadfast belief in me has anchored every step of this journey.

Acknowledgements

First, I would like to extend my deepest gratitude to my thesis supervisor, Prof Wac. Your patience and support have been invaluable throughout this journey. I am also deeply thankful to Prof Konstantas for his insightful feedback and continuous encouragement. My appreciation extends to the jury members, who have contributed significantly with their rigorous evaluations and constructive critiques.

To my beloved wife, Chen, your support and understanding have been foundational during this endeavor. I owe a great deal to my parents for instilling in me the values of hard work and perseverance. My brother, your constant encouragement and belief in my abilities have been a driving force.

A special mention to my brewer friends, Mattia and Marios, for their camaraderie and shared experiences throughout this process. To the friends I made along the way and who were consistently present at our Friday support meetings at Rue de l'École de Médecine: Allan, Jody, Igor, Dejan, Emmanuel, and Ashley, your collective wisdom, encouragement, and the weekly light-hearted moments made the process more enjoyable.

I cannot express my gratitude sufficiently for the unwavering support I received from Yann, Julien, Nicolas, Romain, Jonathan and Jérôme during the course of my thesis. Hailing from the beautiful country of Dolbistan, not only did they provide encouragement throughout my academic endeavors, but our bonds of friendship, which have endured and grown for more than two decades for some, have been a source of strength and inspiration. Their unyielding faith, shared memories, and the countless moments of fellowship have been instrumental in shaping this journey, and I am profoundly grateful for having them by my side.

Prof Moccozet, your teachings have transformed my perspective on education and have enabled me to evolve as a better teacher. I am eternally grateful for the wisdom you imparted. Lastly, to all those I have had the privilege to meet at the University and within the lab, your contributions, discussions, and shared experiences have enriched my academic journey.

Geneva, August 2023

Abstract

On average, people spend 20% of their daily time on smartphone applications in 2021. The smartphone has become a ubiquitous ally in decision-making, planning, entertainment, and communication when user is mobile. However, such activities' quality is deeply linked to multiple factors, such as the smartphone user's context and networking state. To satisfy the networking needs of applications, Quality of Service (QoS) was introduced in 1994 to study the links between user experience and network hardware metrics such as jitter, round-trip time, packet loss, throughput, and availability. These QoS attributes could be guaranteed following a service-level agreement between the operator and their clients. However, QoS is not enough due to the complexity and content of modern smartphone applications, and the fast-changing environment of smartphone users. Hence, Quality of Experience (QoE) was introduced to explain and improve the relationships between the actors who make the experience (e.g., network infrastructure, user, and smartphone hardware maker and software developer).

QoE integrates QoS's technical aspects and requirements, but puts the human at the system's center. That includes the application context, the user's intentions, and expectations to understand the experience. Current work in the QoE domain focuses mainly on improving the network protocol to deliver the required application content or adapting the content to fit diverse network scenarios. Most studies are in a laboratory setting and only collect data from a few participants over a short period (i.e., a few hours maximum). However, no QoE studies systematically take human factors into account conducted over long period studies in-the-wild, i.e., natural daily life environments of the user, to the best of our knowledge.

Many efforts have been made to improve QoE, such as upgrading smartphone hardware, smartphone systems, and mobile network hardware. A large group of QoE studies focus on improving QoS attributes. However, there is a gap between QoS and smartphone applications' QoE. For example, a large available bandwidth cannot guarantee a good experience of a family video call with multiple users (e.g., poor app navigation and interaction design). To fill the gap, and address the human factors, we first studied QoE in-the-wild. We studied the connectivity levels of three cohorts of smartphone users in the same geographical area over different periods to identify the features impacting the QoE. To facilitate the improvement of the QoE, we proposed a method with a forecasting model based on crowdsourced application usage history to predict application launches. This model could help to understand the application's usage pattern found in the users' interaction with their smartphone. Furthermore, toward quantified and enhancing smartphone application QoE, we adopted a mix-method approach to collect quantitative and qualitative information in situ. As such, we ran two studies in-

the-wild. The first study enables us to determine the contextual factors which should be considered to create an accurate and context-aware QoE user-centric quantifying model for smartphone applications. Moreover, we investigated the creation of a contextual model capable of accurately predicting the expected QoE of smartphone application users. In the second study, we deployed a QoE notification system (expectQoE) built from the first study's results. We inquired about the system's impact on reducing the smartphone user's self-reported burden.

In this thesis, we proposed a method to quantify and facilitate the smartphone application end-user QoE by introducing a focus on the context. Our system can be easily transplanted into a device, and our results show the possibility of positively influencing the perceived smartphone users' QoE, which could be used as a ground for future QoE-enabled smartphone services. The results showed that a QoE notification system effectively reduces annoying experiences and time spent on smartphones.

Résumé

En moyenne, en 2021, les gens passaient 20% de leur temps quotidien sur les applications de leur smartphones. Le smartphone est devenu un allié omniprésent dans la prise de décision, la planification, le divertissement et la communication lorsque l'utilisateur est en déplacement. Cependant, la qualité de ces activités est profondément liée à plusieurs facteurs, tels que le contexte de l'utilisateur du smartphone et l'état du réseau. Pour répondre aux besoins de réseau des applications, la Qualité de Service (QoS) a été introduite en 1994 pour étudier les liens entre l'expérience utilisateur et les métriques matérielles du réseau comme le jitter, le temps aller-retour, la perte de paquets, le débit et la disponibilité. Ces attributs QoS pourraient être garantis selon un accord de niveau de service entre l'opérateur et leurs clients. Cependant, la QoS ne suffit pas en raison de la complexité et du contenu des applications modernes de smartphones, et de l'environnement en constante évolution des utilisateurs de smartphones. Par conséquent, la Qualité d'Expérience (QoE) a été introduite pour expliquer et améliorer les relations entre les acteurs qui créent l'expérience (par exemple, infrastructure réseau, utilisateur, fabricant de matériel de smartphone et développeur de logiciel).

La QoE intègre les aspects techniques et les exigences de la QoS, mais place l'humain au centre du système. Cela inclut le contexte de l'application, les intentions et les attentes de l'utilisateur pour comprendre l'expérience. Les travaux actuels dans le domaine de la QoE se concentrent principalement sur l'amélioration du protocole réseau pour livrer le contenu d'application requis ou adapter le contenu pour s'adapter à divers scénarios de réseau. La plupart des études sont menées en laboratoire et ne collectent des données que sur quelques participants pendant une courte période (c'est-à-dire quelques heures maximum). Cependant, à notre connaissance, aucune étude QoE ne prend systématiquement en compte les facteurs humains menée sur de longues périodes dans des environnements de vie quotidienne naturels de l'utilisateur.

De nombreux efforts ont été faits pour améliorer la QoE, tels que la mise à jour du matériel des smartphones, des systèmes de smartphones et du matériel de réseau mobile. Un grand nombre d'études sur la QoE se concentrent sur l'amélioration des attributs de la QoS. Cependant, il existe un fossé entre la QoS et la QoE des applications de smartphones. Par exemple, une large bande passante disponible ne peut garantir une bonne expérience d'un appel vidéo familial avec plusieurs utilisateurs (par exemple, une mauvaise navigation et conception d'interaction de l'application). Pour combler cette lacune, et aborder les facteurs humains, nous avons d'abord étudié la QoE en situation réelle. Nous avons étudié les niveaux de connectivité de trois cohortes d'utilisateurs de smartphones dans la même zone géographique sur différentes périodes pour identifier les caractéristiques impactant la

QoE. Pour faciliter l'amélioration de la QoE, nous avons proposé une méthode avec un modèle de prévision basé sur l'historique d'utilisation des applications crowdsourcées pour prédire les lancements d'applications. Ce modèle pourrait aider à comprendre le modèle d'utilisation de l'application trouvé dans l'interaction des utilisateurs avec leur smartphone. De plus, pour quantifier et améliorer la QoE des applications de smartphones, nous avons adopté une approche mixte pour collecter des informations quantitatives et qualitatives in situ. Ainsi, nous avons mené deux études en situation réelle. La première étude nous a permis de déterminer les facteurs contextuels qui devraient être pris en compte pour créer un modèle de QoE centré sur l'utilisateur précis et conscient du contexte pour les applications de smartphones. De plus, nous avons étudié la création d'un modèle contextuel capable de prédire avec précision la QoE attendue des utilisateurs d'applications de smartphones. Dans la deuxième étude, nous avons déployé un système de notification de QoE (expectQoE) construit à partir des résultats de la première étude. Nous avons enquêté sur l'impact du système sur la réduction de la charge auto-déclarée de l'utilisateur de smartphone.

Dans cette thèse, nous avons proposé une méthode pour quantifier et faciliter la QoE de l'utilisateur final des applications de smartphones en introduisant un accent sur le contexte. Notre système peut être facilement transplanté dans un appareil, et nos résultats montrent la possibilité d'influencer positivement la QoE perçue des utilisateurs de smartphones, qui pourrait être utilisée comme base pour les futurs services de smartphones habilités par la QoE. Les résultats ont montré qu'un système de notification de QoE réduit efficacement les expériences agaçantes et le temps passé sur les smartphones.

Table of contents

List of figures	xv
List of tables	xvii
1 Introduction	1
1.1 Context and Motivation	2
1.2 Research Questions	5
1.3 Research Methods	6
1.4 User Study Design (S1,S2)	7
1.5 Thesis Contributions	11
1.6 Thesis Outline	14
2 Related Work	19
2.1 Introduction	20
2.2 Overview: Quality of Experience	20
2.3 Quantification and Modeling of Quality of Experience	21
2.4 Smartphone Quality of Experience Assessment	26
2.5 Mobile User’s QoE Management and Services	28
3 Article I: mQoL smart Lab: Quality of Life Living Lab For Interdisciplinary Experiments	31
3.1 Introduction	32
3.2 Current Platform: mQoL Living Lab 1.0 (mQoL-Lab)	33
3.3 New mQoL Platform: mQoL Smart Lab 2.0	35
3.4 Discussion and Conclusive Remarks	37
3.5 Revisions to the Published article	38
4 Article II: The Importance of Smartphone Connectivity in Quality of Life	39
4.1 Introduction	40
4.2 Related Work	41
4.3 Mobile Network Connectivity Study: Methods	45

4.4	Mobile Network Connectivity: Results	53
4.5	Discussion	60
4.6	Conclusion	64
5	Article III: Forecasting Smartphone Application Chains: an App-Rank Based Approach	67
5.1	Introduction	68
5.2	Related Work	71
5.3	Method and Implementation	73
5.4	Results	82
5.5	Discussion	83
5.6	Limitations and Conclusions	86
6	Article IV: You're using this app for what ? A mQoL Living Lab Study	89
6.1	Introduction	90
6.2	Methodology	92
6.3	Results	96
6.4	Discussion	97
6.5	Conclusive Remarks and Future Work Areas	98
7	Article V: Predicting Quality of Experience of Popular Mobile Applications from a Living Lab Study	101
7.1	Introduction	102
7.2	Related Work	103
7.3	User Study	104
7.4	QoE Prediction Model	110
7.5	Discussion	112
7.6	Limitations	113
7.7	Conclusions and Future Work	114
8	Article VI: Towards Accurate Models for Predicting Smartphone Applications' QoE with Data from a Living Lab Study	115
8.1	Introduction	117
8.2	Related work	119
8.3	The Approach: User Study	122
8.4	Building QoE Prediction Models	128
8.5	Results	137
8.6	Discussion	139
8.7	Study Limitations	142
8.8	Conclusions and Future Work Areas	143

9 Article VII: Less Annoying: Quality of Experience of Commonly Used Mobile Applications	145
9.1 Introduction	146
9.2 Related Work	148
9.3 Methodology	149
9.4 ExpectQoE Evaluation and User Study Results (S2)	152
9.5 Discussion and Limitations	159
9.6 Conclusions	161
10 Discussion and Conclusion	163
10.1 Discussion of the Results	164
10.2 Limitations	175
10.3 Future Work Areas	179
10.4 Conclusion	181
References	183
Appendices	211
A Informed Consent Form: Study 1	213
B Informed Consent Form: Study 2	219

List of figures

1.1	Studies' S1 and S2 Timeline and Research Methods	9
1.2	Research Methods (Ickin, 2015)	9
1.3	QoE Level Query (EMA Question)	10
1.4	User Expectation Query (EMA Question)	10
1.5	expectQoE Output Notification with Buttons	11
1.6	Thesis Achievements	11
1.7	Schematic Representation of the Thesis	13
3.1	Data Flow in the mQoL-Lab Platform	34
4.1	RAT distribution of participants in P1 ($N = 50$)	55
4.2	RAT distribution of participants in P2 ($N = 55$)	55
4.3	RAT distribution of participants in P3 ($N = 5$)	55
4.4	Overall average RAT distribution over P1, P2, and P3	55
4.5	Mean Signal Strength per Data Collection Period	56
4.6	Overall Signal Strength Distribution per Data Collection Period	57
4.7	Pearson Correlation Between Signal Strength and Network Access Technology Type	57
4.8	Data Consumption CDF	58
4.9	Normalized Weekly Mean Amount of Data Received per Data Collection per Period	58
4.10	Total Downloaded and Uploaded Bytes per Period Normalized by Number of Data Collection Days	59
4.11	Mean Cumulative Cell Tower and Wi-Fi AP ID Changes per Data Collection per Period	60
5.1	Application Chain Example	69
5.2	Forecasting Model Pipeline	73
5.3	Mean F1 Score Over Number of Weeks (K)	83
6.1	Study List in mQoL Lab	91
6.2	Running Study Notification	91
6.3	New EMA Notification	92
6.4	Distribution of Selected 7 Apps Used per User	94

6.5	What action were you trying to accomplish ?	95
6.6	What action were you trying to accomplish in the application ?	96
6.7	Distribution of top 10 apps per user	99
6.8	Time spend in each app per user [min]	100
7.1	User selecting QoE=5 on the scale after using (Google) Maps	105
7.2	Application usage QoE/MOS rating distribution.	109
7.3	Application use distribution.	109
7.4	Network connectivity distribution.	109
7.5	Model AUC for each user as test datasets.	113
8.1	User Selecting QoE=5 on the Scale After Using (Google) Maps	123
8.2	Application Usage QoE/MOS Rating Distribution Per Participant	127
8.3	Network Connectivity Distribution Per Participant	127
8.4	Physical Activity Distribution Per Participant	127
8.5	Applications Distribution Per Participant	128
8.6	QoE Modeling Process Pipeline	129
9.1	Study S1 and S2 Timeline and Research Methods	149
9.2	S2: Effect size and Amount of Valid Analysis for Causality Between Application Usage Duration and ExpectQoE Notifications During T1	157

List of tables

1.1	Synoptic Table of Dissertation Articles	16
2.1	Selected QoE Studies*	27
4.1	Data Collection Periods	45
4.2	Participation Statistics for the Filtered Dataset in Each Data Collection Period	46
4.3	Number of Minutes of Data Collected for the Three Periods	47
4.4	Data Collected by mQoL-Log	48
4.5	mQoL-Log Network Data	49
4.6	Average Measurement Minutes Collected Post Resampling, Per Participant in a Period	50
4.7	Generation of Cellular Network Access Technologies	52
4.8	Overall Average RAT Distribution (%) per Data Collection Period	54
4.9	Percentage of Connectivity to Internet Distribution per Data Collection Period	56
5.1	Literature Review on Forecasting Based On Application Usage Records	71
5.2	Smartphone Screen State Combinations	75
5.3	Dataset States Pre and Post Filtering	76
5.4	F1 Score Performances Over the Both Datasets	82
6.1	Application Metadata from Google Play Store	93
6.2	Time spend to reply to EMA in each app [s]	97
7.1	Study Participation Metrics	106
7.2	mQoL-Log: Background Logger Data Collection	108
7.3	Network Features	111
7.4	AUC score on the validation dataset (all users)	112
8.1	Study EMAs Questions	121
8.2	mQoL-Log: Background Logger Data Collection	124
8.3	Study Participation Raw Metrics	125
8.4	Network features as Collected via mQoL-Log During Application Usage	131
8.5	QoE Prediction: Metrics on the Validation Dataset for Multiple Common Classifiers	132

8.6	Times to Reply (TR) to EMAs Per User	134
8.7	Aggregation Features (FA) Threshold in Minutes for Each mQoL-Log service	135
8.8	Aggregated Features Importance from TR , FA and ULT	136
8.9	Model's Features Per Scenario	138
8.10	QoE Models Performance on Test Dataset for Each Scenario	139
9.1	Information Collected per Study (mQoL-Lab)	150
9.2	S2: Participants Ratings to ExpectQoE Predictions	153
9.3	S2: Causality Between Application Usage Duration and expectQoE Notifications	157

Chapter 1

Introduction

Chapter Contents

- 1.1 Context and Motivation 2**
- 1.2 Research Questions 5**
- 1.3 Research Methods 6**
- 1.4 User Study Design (S1,S2) 7**
 - 1.4.1 S1: Study 1 (2018) 8
 - 1.4.2 S2: Study 2 (2021) 10
- 1.5 Thesis Contributions 11**
- 1.6 Thesis Outline 14**

1.1 Context and Motivation

A recent study from [Deboitte \(2018\)](#) reports that 92% of all people in Switzerland own at least one smartphone, with 97% of those respondents using them every day. In advanced economies (i.e., developed world), the median smartphone ownership was reported to be 76% compared to 45% in emerging economies ([Silver, 2019](#)). A study by [Dey et al. \(2011\)](#) shows that smartphones remain near their owners for at least 88% of the day (24-hour period). Throughout the day we carry our smartphone everywhere. Society has become accustomed to using smartphone applications in numerous situations, such as repetitively checking social media applications for new content, on moving public transportation (e.g., a bus, train, or tram), crossing the city on a bicycle following the directions of a navigation application, and watching videos and listening to audio content from a streaming platform via their application. Smartphone applications may also impact people's mobility choices, as explained by [Khan et al. \(2020\)](#). The applications provide real-time information about public transportation traffic and notify their users about other available transportation modalities (e.g., taxis, walking, and bicycles). They provide information on decision making and help to accomplish daily tasks through the mobile Internet.

Smartphone applications can satisfy user needs and enhance their Quality of Life (QoL ([WHO-QOL, 2012](#))). The plurality of individual needs and tasks has caused the smartphone application market to grow in the last 10 years. For example, the Google Play Store contains 2.7 million applications from 28 different categories as of November 2022, compared to 0.4 million applications on December 2011 ([Statista, 2022](#)). However, smartphone applications do not always function as the end-user expects. For example, a user may expect its application to be ready to satisfy their communication intent (e.g., video call). However, the video may fail to initialize due to an external constraint. For example, connectivity can be impacted in mobile contexts when the end-user's smartphone is roaming between telecommunication antennas.

Mobile network operators are mainly concerned with providing access to wireless networks and improving their Quality of Service (QoS; [Telecommunication Union](#)). QoS focuses on quantitative information (e.g., jitter and round-trip time) to evaluate the quality of a service (e.g., voice call) over the Internet Protocol. QoS can have an impact on the experience of mobile applications for users ([Casas et al., 2015b](#)), but it is not enough to ensure the quality of a smartphone application use only. The smartphone operating system (OS) cannot provide a high-level user experience during all application usages due to the multiple possible sources of annoyance. The challenge arises from the broad range of applications, contexts, hardware, and users' needs. Therefore, the term Quality of Experience (QoE) was coined to complement QoS; the QoE measure is an expansion of QoS and includes qualitative information related to the experience itself, and therefore prioritizes the end-user, as presented in the Qualinet white paper ([Le Callet et al., 2012](#)). Consequently, the authors of the white paper defined QoE as:

QoE: The degree of delight or annoyance of the user of an application or service. It results from fulfilling his or her expectations with respect to the utility and/or enjoyment of the application or service in light of the user's personality and current state. (Le Callet et al., 2012)

Smartphone systems are build with contribution from UX designer (User eXperience) and developers. The designers focus on the user interface and their interactions with the system. The developers build and optimize the system. However, the appreciation of UX and the performance of the system are insufficient to guarantee a delightful application experience. To provide a better overall experience, it is needed to quantify the QoE of smartphone applications and maximizing the enjoyment of the end-user, potentially at every context. In the end, obtaining a delightful experience would benefit smartphone application developers and end-users alike.

Research about smartphone applications and the corresponding user experience so far has been limited to focusing on the application interface and networking aspects and metrics (Casas et al., 2017a, 2022; Chen et al., 2014; Gao et al., 2020; Mitra et al., 2015). In the field of telecommunication networks, studies focus on modeling and forecasting the network patterns of smartphone applications and proposing a model to map the state of the network to the QoE level of the end-user (Huet et al., 2021; Mozetić et al., 2021; Pibiri et al., 2012). However, such a method does not integrate the users themselves or their context.

New network protocols and enhanced hardware are insufficient to facilitate end-user QoE (Narayanan et al., 2021; Schwind et al., 2020; Szabó et al., 2016). Although the authors propose QoE management methods based on network condition status; they fail to take into account the user, its perception, and the content (Seufert et al., 2021). The authors only focus on the medium on which the content is delivered. Important aspects such as context and habits are outside the scope.

The pattern in which applications are used as been previously studied by Davison and Hirsh (1998) on Linux based command line. Their work indicated a certain predictability in how we use applications. In today context, the smartphone user application habits were studied and forecast by Liao et al. (2013); Zhao et al. (2016). The authors have shown the impact of application recommendations on the end-user experience, in particular for launching an application faster (via precaching application data). The current focus for forecasting application usage is directed toward digital wellbeing intervention (Roffarello and De Russis, 2021) and toward bandwidth optimization on core network operator (Celenk et al., 2021; Deljac and Randić, 2022; Shen et al., 2019; Xiang et al., 2017).

New methods and models considering the user's expectations, from their own mental models built from past experiences, from external and contextual stimuli (e.g., physical activity) could be employed to maximize the degree of delight and minimize the degree of annoyance of an experience. Mental models include beliefs and knowledge structures that describe what, how, and why we know something and also what we think in respect to a particular matter (Cannon-Bowers and Salas, 2001). Thus, those mental models are used to assess the level of one's experiences. Furthermore, the models embody the perceived impact of the factors influencing the user QoE on smartphone.

As reported by [Ickin et al. \(2012\)](#), the most influential factors for QoE of smartphone application users were:

- Application interface design: position and location of UI elements (e.g., layout, buttons, text fields)
- Application performance: speed of in-application task execution
- Battery: the efficiency limits the user smartphone usage
- Phone features: availability of applications and hardware
- Applications and data connectivity cost: monetary cost of application and due to data connectivity cap (e.g., limited bandwidth or amount of data)
- Routine: recurring user actions on the smartphone (e.g., morning and evening smartphone notification check)
- Lifestyle: support a user's lifestyle choices (e.g., mHealth ([Khan et al., 2020](#); [Wattanapisit et al., 2020a](#); [Wulfovich et al., 2019](#)), mobility and digital wellbeing ([Monge Roffarello and De Russis, 2019](#)) applications)

However, these factors were defined ten years ago, they may be not valid today ([Dery et al., 2014](#); [Shaheen et al., 2017](#)). The operating system and smartphone evolution (e.g., application interface and battery capacity) and the rise of new application usages (e.g., mobile streaming ([Lall et al., 2020](#); [Wamser et al., 2015](#); [Wassermann et al., 2019](#))) have affected the factors. Furthermore, the factor named "phone feature" is directly linked to the connectivity of the smartphone user. The new telecommunication technology advancement had an impact on smartphone applications ([Huang et al., 2013](#)). From a 3G network with limited bandwidth, we observed the upgrade to a 4G/5G network (e.g., higher speed and bandwidth) that enabled new habits, such as video streaming ([Hegde et al., 2022](#); [Jiang et al., 2021](#)). Video streaming became mainstream on smartphones. In 2021, [Statista \(2021b\)](#) reported that more than 61% of the worldwide viewing time of YouTube was done on a smartphone. Therefore, it is necessary to understand the features that affect smartphone user connectivity and experience over long periods.

This thesis explores how individuals use their smartphones and applications based on their collected perceived QoE assessment. Subsequently, we propose an on-device system supported by smartphone sensor data and user actions, intent, and behavior (e.g., application usage) to maximize the end-user's experience. Consider this figurative case of Bob trying to use his messaging application to call Alice during his train commute back home in the evening. This commute path is known to create high annoyance for smartphone users riding the train. The network operators managing the cell towers bordering the train track hardly check the QoE of their users, nor plan or could invest in new infrastructure (e.g., blocked by regulation or political and public opinion). Moreover, Bob's smartphone operating system (OS) and application only indicate the network signal quality, which

Bob perceives as sufficient for his call. However, because the system can quantify the current QoE level of Bob's applications, due to knowing Bob's habits, his current context, and past experience; and consequently it notifies him of the imminent annoying experience. It makes Bob change his behavior, hang up the call with Alice and switch to texting. In this case, Bob's intention of communicating is still satisfied and the annoying experience does not occur.

Our work is the basis of future QoE-enabled service on smartphone. By leveraging context, crowd-sourced inputs, and machine learning algorithm, such service can be proposed by external providers or directly implemented in the operating system to provide higher QoE of smartphone applications, limiting meaningless and annoying application usage.

1.2 Research Questions

We explore the following research questions (RQs, in-box) based on the context and motivation of this thesis :

RQ1a - What are the factors that affect smartphone users' connectivity in-the-wild, over time?

RQ1b - How to accurately forecast the usage of smartphone users' applications with application usage record?

Connectivity is central to understand the QoE of smartphone applications due to the influence of QoS on QoE. In addition, the evolution of the smartphone network and services has impacted its users. As such, we propose, via statistical analysis of past smartphone datasets and the current knowledge in the domain state-of-the-art, to identify the factors which impact one's connectivity over different periods. Furthermore, the pattern of application use is a factor influencing QoE. Therefore, we propose to use application usage records to forecast the application usage sequence toward its integration in a system that facilitate QoE enhancement. In the context of this thesis, the term "accurate" refers to the machine learning model performances obtained (i.e., metrics).

RQ2a - What contextual factors should be considered to create an accurate and context-aware QoE user-centric quantifying model for popular mobile applications?

RQ2b - How to accurately predict the smartphone application user's QoE with a context-aware model?

The state-of-the-art models to quantify QoE of popular mobile applications have been trained with a focused on mapping QoS to QoE. As such, the user's context was often left out, and the user's intent was never part of the model. Hence, we explore the creation of models trained with contextual and intent based information. This user-centric approach is in opposition to the system-centric approach found in QoE applied to smartphone applications.

RQ3 - What are the challenges in quantifying QoE on-device with the users' context factors in-the-wild?

Inference on-device of QoE is problematic due to the amount of external factors influencing the user's QoE. Furthermore, some modeling features that were obtained to create an in-the-wild model may be unavailable due to the nature of the feature. For example, the time spent using an application has been reported to be a feature of great importance. However, it is problematic to know preemptively how much a user will use an application. Hence, we study the challenges to quantifying QoE on-device.

RQ4 - What is the evolution of the most influential factors for the user's QoE on smartphones since their last assessment?

The factors influencing the use of smartphone applications were first studied by [Ickin et al. \(2012\)](#) more than ten years ago. Therefore, a new assessment is needed due to changes in the smartphone operating system (Android new versions), in the smartphone user population (worldwide adoption), in the applications themselves (material design guidelines), and its ecosystem and services (Google Play store and services' new features). Hence, we explore the evolution of the factors and its trends.

RQ5 - What human-computer interaction method can contribute to manage smartphone users' expected QoE in-the-wild?

QoE management methods often involve the use of an adaptive algorithm to change the consumed content format or the use of another networking path to deliver the content (e.g., multihoming and path optimization). These methods are managed and operationalized by the content provider, without any control or transparency on the user side. As well, the methods are based on QoS metric without including the user in the loop (i.e., no human-centric design). Hence, we propose an informational approach based on human-computer interaction to facilitate the smartphone user experience.

1.3 Research Methods

We employed a mixed methods approach that combined both qualitative and quantitative methods to achieve our goal of quantifying and managing smartphone applications' QoE in-the-wild. The adoption of the *in-the-wild* approach enables the investigation of participants within their naturalistic settings, wherein everyday data produced by individuals in ordinary circumstances is recorded. The term mixed methods refers to a research methodology that integrates quantitative and qualitative data within a single investigation or study ([Creswell and Tashakkori, 2007](#)). The essential premise of this methodology is that such a combination allows for a more comprehensive use of the data than

separate quantitative and qualitative data acquisitions and analyses. Considering the research tools available from previous studies on smartphone-based phenomena within our lab, we decided to employ a logger application (mQoL-Log) to collect the quantitative data. mQoL-Log was developed by De Masi et al. (2016); Gustarini et al. (2016a) to passively gather sensor information from Android smartphones within the mQoL-Lab application stack. Smartphone-based data collection has previously been used to study smartphone addiction (Lowe-Calverley and Pontes, 2020), sleep (Ciman and Wac, 2019), and other human factors. We built a data collection platform (mQoL-Lab) to support our research (De Masi et al., 2016), later improved by (Berrocal et al., 2020b), and facilitated its execution and reusability for QoE studies. Through our application, we also used surveys to implement an Ecological Momentary Assessment (EMA; Stone and Shiffman (1994)). EMAs are used to collect a person's momentary self-assessment of a particular outcome of interest (e.g., satisfaction, QoE level, or expectation). Studies in the QoE domain have traditionally collected their participants' assessment during or after (controlled) testing a service or application. However, this approach limits the habit-forming event and could consequently impact the data collected, particularly the participants' expectations (Sackl et al., 2017). We used the Day Reconstruction Method (DRM; Kahneman et al. (2004)) according to the best practices for in-the-wild studies (Gustarini et al., 2013; Hoßfeld et al., 2014; Ickin et al., 2012). This method consists of a semi-structured interview during which the user describes their last 24 hours on a particular topic. Researchers learn new information in relation with user habits (e.g., commuting time, popular application usage, and lifestyle), which impact the research context and data collected.

Furthermore, to answer our RQs, we designed two studies. The studies provided data which enable us to answer the research questions. Due to the interactive nature of a user and its smartphone (i.e., deeply couple to the current context), the studies were conducted *in-the-wild*. The *in-the-wild* approach, as presented by Rogers and Marshall (2017), in human-computer interaction research, enables the collection of user experience phenomena in normal environment that differ from those derived from other laboratory-based methods. Most of the existing research on the QoE of smartphone applications uses a classic laboratory setting. As such, these studies often miss important external factors that influence smartphone users' daily activities. Therefore, we chose to develop two studies *in-the-wild* for our research.

Finally, we also employed the mobile phone use (MPU) dataset (Pielot et al., 2017) to model the forecasting of the usage of smartphone applications. The output information of such model could be used to preemptively stop annoying or disturbing application usage (e.g., limiting access due to triggering content).

1.4 User Study Design (S1,S2)

We designed two human subject studies with different objectives to answer our RQs. We also utilized previously collected passive data with the mQoL-Lab application (De Masi et al., 2016), for longitudinal studies in the Greater Geneva Area Switzerland (mQoL Living Lab). Figure 1.1 presents the protocols

of Study 1 (S1) and Study 2 (S2), and their respective timelines (days). The informed consent forms for both studies are included, respectively, in the Appendix of this thesis (S1: Appendix A, S2: Appendix B).

S1 (2018) focused on quantifying smartphone applications' QoE levels in-the-wild. S2 (2021) focused on the factors influencing the QoE of smartphone applications, and tested both the QoE model derived from the data collected in S1, and the influence of *expectQoE*, a QoE alert system, on participants' application usage habits. Both studies had a duration of four weeks. Contrary to S1, in which we collected data unobtrusively and passively, the dynamics during S2 was different. S2 was a partial intervention study. After an initial baseline data collection period (T0), the expected QoE level outputted by the QoE model was shown during the T2 period. As we wished to understand the impact of *expectQoE*, we included in the timeline a special period T3. During T3, participants were shown random QoE levels. Participants were recruited on the campus of the University of Geneva through brochures, mailing lists, and social networks. We focussed on smartphone users who lived in the Greater Geneva region and at least 5 years of Android usage. The geographical requirement is due to the study protocol (i.e., ethics requirement) and the study design (i.e., meeting the participants face to face to introduce the study). The Android usage requirement is due to the bias, and habits which another smartphone operating system could introduce in the experiments. As well, the data collection application mQoL-Lab was only available on the Android platform due to the software limitation to enable data collection on iOS.

The requirements for S1 and S2 were identical. Potential participants were required to use a few or all target applications as usual, and have the most recent version of the operating system on their smartphones. The target applications were Spotify, WhatsApp, Facebook, Facebook Messenger, Google Maps, Chrome, and Instagram. The applications were selected due to their popularity in the Google Play Store in 2018 (Statista, 2021a). Candidates completed an online survey from which we estimated their interest in our studies and verified whether they met our requirements. Each participant gave consent before enrolling in the study and downloading our mQoL-Lab application from the Google Play Store.

Figure 1.2 presents the research methods used in this thesis, classified by their use for the collection of objective or subjective data and the cyclicity (i.e., timeliness) of the data.

1.4.1 S1: Study 1 (2018)

For S1, we focused on quantifying the QoE of popular smartphone applications. We employed an EMA to query the study participants to rate their QoE level after an application usage (Figure 1.3), as employed in past QoE studies (Ickin et al., 2012). EMAs were limited in time and amount. Only 12 EMAs per day could be triggered between 7:00 and 21:00. In addition, to limit the participants' burden, the interval between two consecutive EMA was set up at least to 20 minutes. Within each EMA, we also asked participants about their intent; what were they trying to do or what action to accomplish

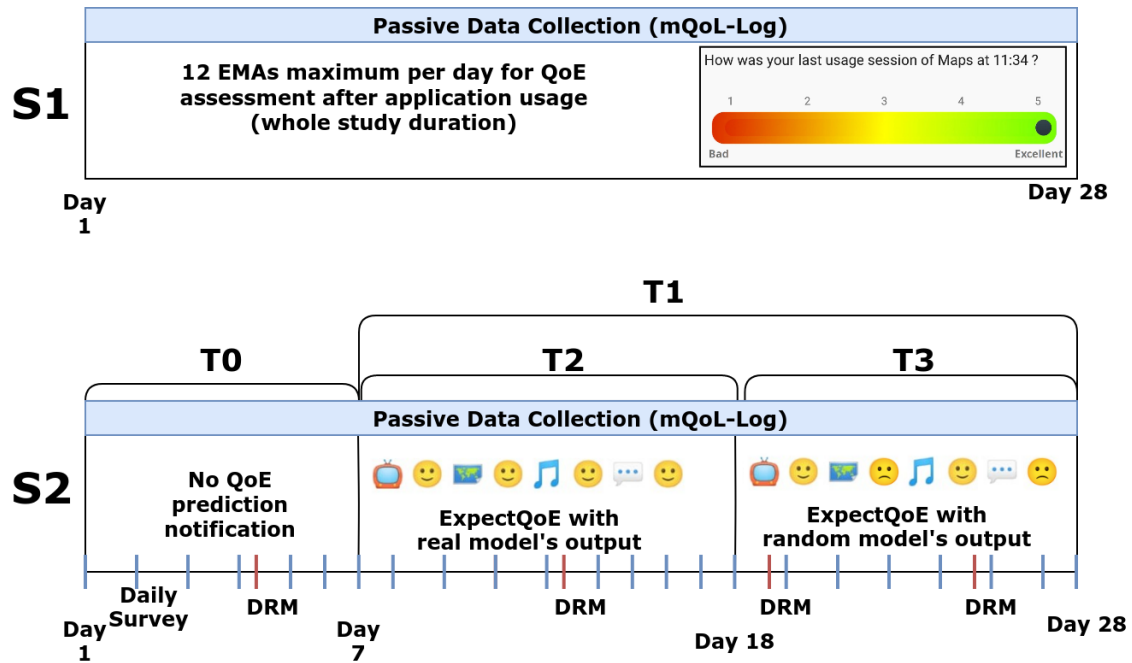


Figure 1.1 Studies' S1 and S2 Timeline and Research Methods

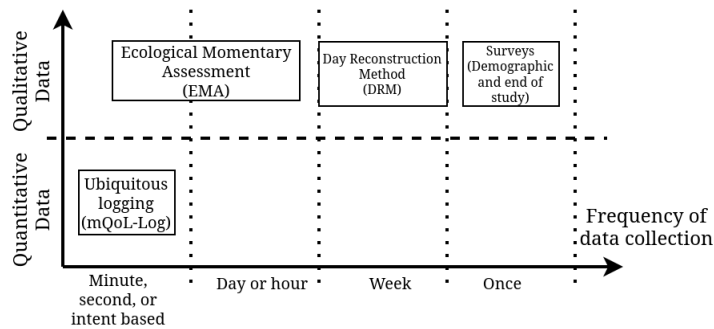


Figure 1.2 Research Methods (Ickin, 2015)

through application usage (Figure 1.4). The EMA presented itself throughout a notification on the smartphone.

Once the participant touched the notification, they could answer the questions. Simultaneously, our smartphone logger mQoL-Log passively and unobtrusively collected data from the onboard sensors (Berrocal et al., 2020b). These data comprised the QoS information of the network (e.g., signal strength and wireless technology), the participant’s physical activity (e.g., walking, running, or in a vehicle), battery state, the onscreen application name, and the duration of the participant’s interaction. The data obtained from S1 provided insights into smartphone application usage habits (EMA and passive sensing). This data enables us to build a QoE level classification model and ascertain the most important features impacting the user QoE on smartphone.

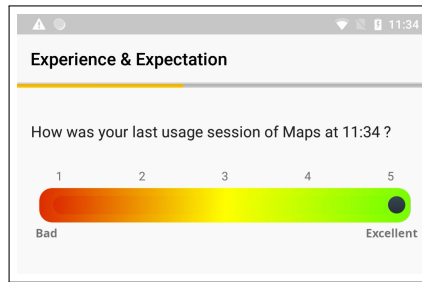


Figure 1.3 QoE Level Query (EMA Question)

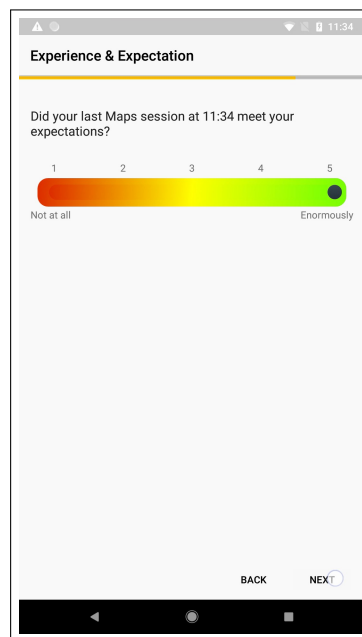


Figure 1.4 User Expectation Query (EMA Question)

1.4.2 S2: Study 2 (2021)

For S2, we focused on employing the QoE model developed after S1. The first seven days of the study (T0) were used to passively collect the application usage habits of the participants. We wanted to analyze the impact of such a model on the participants' smartphone usage habit. Does alerting users of an expected QoE limit their annoyance? We integrated the model in a system named *expectQoE*. Furthermore, the *expectQoE* output was only available from Day 7 of the study until the end. We implemented *expectQoE* in mQoL-Lab through Android notifications. The system showed the expected QoE level (i.e., high or low) of four different application categories: communication and social (WhatsApp and Instagram), audio (Spotify), video (YouTube and Netflix), and tools (Google Maps).

The participants were able to rate *expectQoE*'s outputs directly from the notification via buttons, presented in Figure 1.5. Emojis replace plain text (Lu et al., 2016; Wu et al., 2022) and have been heavily used to generate notification-to-application interactions (Esteves et al., 2022; Tauch and Kanjo, 2016).

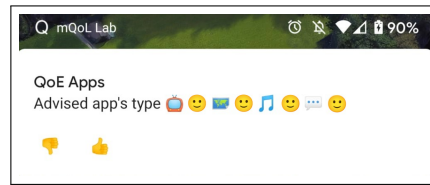


Figure 1.5 expectQoE Output Notification with Buttons

The category of application and their respective expected QoE level were represented by emojis. We mapped each application category to a specific intuitive emoji: communication and social 🗣️, music and audio 🎵, video 📺, and travel and local 📍. In-notification buttons were used to provide feedback with the thumbs-up 👍 and thumbs-down 👎 emoji. The QoE level was indicated via the slightly-smiling-face emoji 😊 for a predicted high-level QoE and the worried emoji 😟 for a predicted low-level QoE. After Day 18 of the study, the expected QoE levels presented were randomized (T3). Thus, enabling us to test the impact of our model on the smartphone's user habits, versus a random output. We hypothesized that the participants would avoid launching low QoE applications due to *expectQoE* indication. The participants received the output of the real QoE model for 11 days and the output of the random QoE model for 10 days. At the end of each day, the participants had to answer a survey about their daily QoE, their expectations, and the usefulness of *expectQoE*. Following the DRM method, we conducted weekly remote interviews with participants for all four weeks. We asked about the factors influencing their QoE and their last 24 hours of application usage. They described their smartphone usage routine and annoying and delightful experiences they had with their smartphone.

1.5 Thesis Contributions

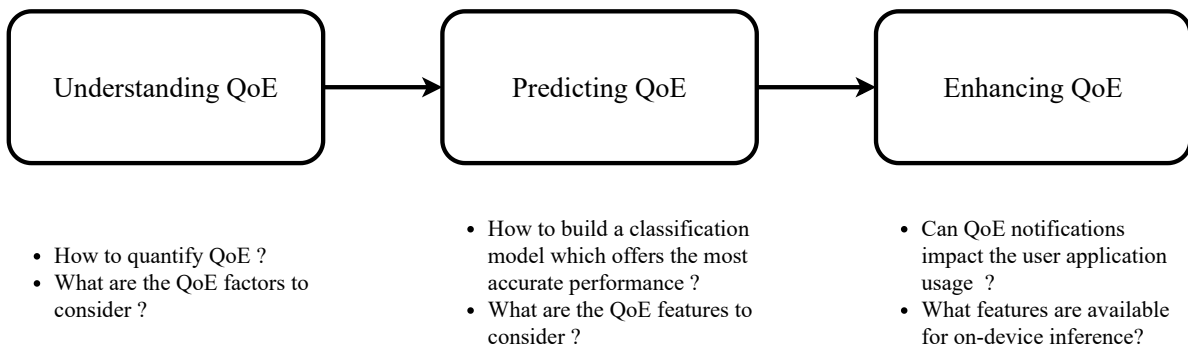


Figure 1.6 Thesis Achievements

The achievement of this work, throughout this thesis (Figure 1.6), can be summarized around three main scientific contributions (SC) with our results:

1. **SC1:** Building models to quantify QoE

- We identified smartphone connectivity patterns through time and three different user cohorts (RQ1a).
 - We conducted an in-the-wild study to collect and analyze the contextual factors that affect the QoE of the smartphone user (RQ2, RQ3).
 - We investigated several approaches to build models to quantify the QoE level of smartphones application. The models obtained were validated following the best practices in machine learning modeling (RQ3).
2. **SC2**: Developing a novel method for forecasting smartphone application usage
- We presented a novel method and algorithm to forecast smartphone application usage (RQ1b).
3. **SC3**: Designing and implementing a system to facilitate smartphone application QoE
- We carried out the first in-the-wild study that implemented a system to facilitate QoE on smartphone (RQ3, RQ5).
 - We identified the evolution of the user reported factors influencing QoE since 2012 (RQ4).
 - We quantified the impact of QoE level notifications on smartphone application session duration (RQ5).

Figure 1.6 does not show the non-linearity of this thesis research process.

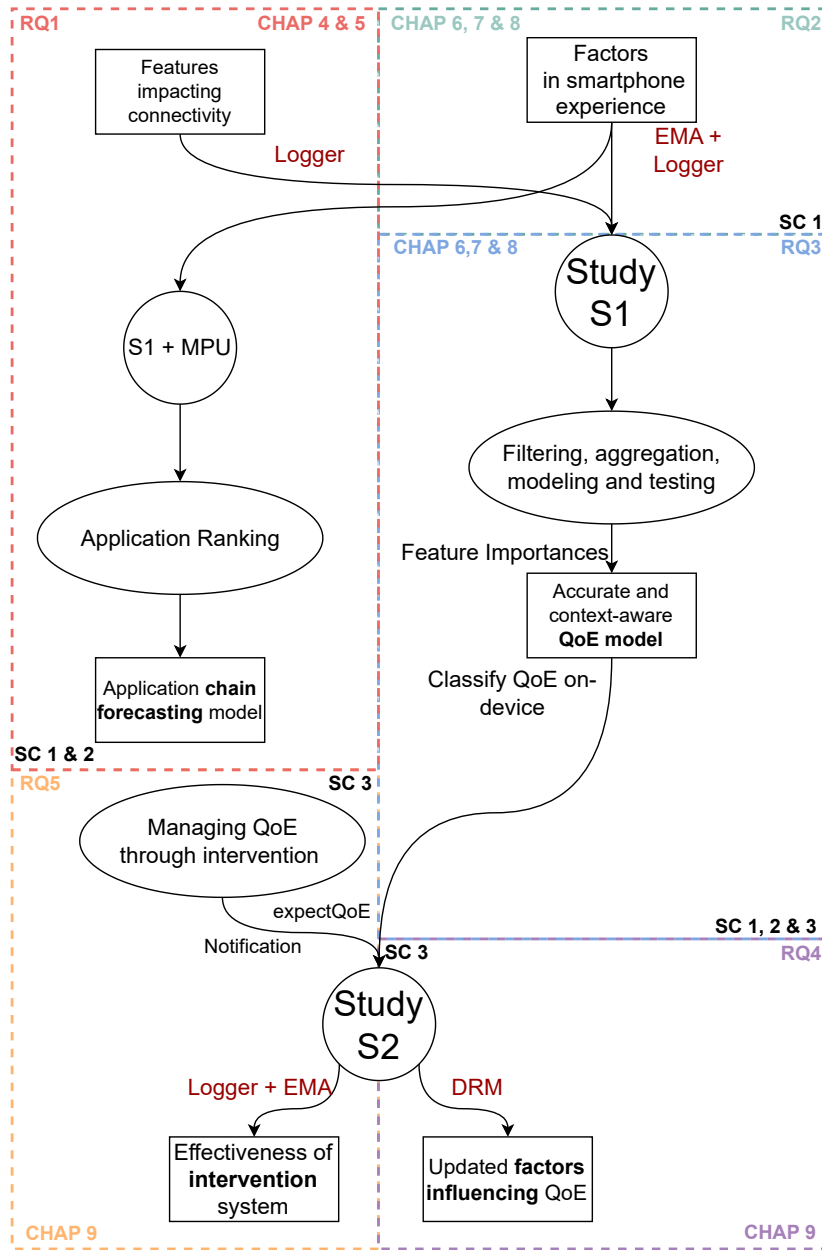


Figure 1.7 Schematic Representation of the Thesis

In Figure 1.7 we show the method and achievement of this work. Semicircles are the actions needed to fulfill the related goals, and rectangles are representing their outcomes. The dashed lines refers to the research questions.

1.6 Thesis Outline

Table 1.1 presents an overview of the research articles (workshops, conferences, and journals) written during the thesis, their connection to the different chapters, and their description of relevant scientific contributions.

Chapter 2 presents the related work and background information. The limitations of existing research are discussed. Chapter 3 describes the development of the underlying technical infrastructure (mQoL-Lab) that was used to conduct the experimental work of this thesis (S1 and S2), thus enabling our work. This chapter was published in the Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (De Masi et al., 2016).

Chapter 4 focus on **RQ1a**. We present the evolution of smartphone connectivity through three cohorts in the Great Geneva region from 2015 to 2020. This chapter was published as a book chapter in *Quantifying Quality of Life: Incorporating Daily Life into Medicine*, Springer, Cham, 2022.: (De Masi and Wac, 2022)

Chapter 5 presents a novel approach to forecasting smartphone application usage. It supports the exploration of **RQ1b**. In this chapter, we investigate the behavior of smartphone application users. We build a prediction model to forecast a ranked list of application to be used. The model is based on the most recent application used and the user history of the past application to forecast the next set of applications that a user will launch. In the future, this information could be used for enhancing preemptively the user QoE. For example, it could trigger a smart caching service for high bandwidth application, reducing the user's annoyance when watching video. The chapter was published in the *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia (MUM '23)*.

Chapters 6, 7, and 8 examine empirically **RQ2** and **RQ3**. These chapters focus on understanding smartphone application usage habits and modeling smartphone applications' QoE level with data collected during S1 (Section 1.4.1).

Chapter 6 addresses smartphone application habits. We investigate the actions that are commonly performed in popular applications. This chapter was published in *Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (De Masi and Wac, 2018).

Chapter 7 presents our findings regarding building a model using the data collected in S1 to predict the QoE of popular mobile applications. We created a classification model to predict the perceived level of QoE ("high" or "low") of the application usage. This chapter was presented and published at *11th International Conference on Quality of Multimedia Experience* (De Masi and Wac, 2019) where it received the **Best Student Paper Award**. Chapter 8 is an extension of Chapter 7. It was published in the journal *Quality and User Experience*, Oct. 2020, Springer (De Masi and Wac, 2020). Following the rules of the journal, 40% new content was added. In this chapter, we investigate the integration of other features (i.e., reported expectations) in the building of QoE models. We also present multiple QoE prediction models constructed using only data available to quantify QoE on the smartphone in its current context.

Chapter 9 examines and **RQ4** and **RQ5**. We describe our system to manage smartphone users' expectations based on a QoE classification model, and we explore the evolution of the factors influencing smartphone applications' QoE. We present our implementation of the QoE prediction model in mQoL-Lab and the QoE notification system, *expectQoE*. This chapter was published and presented in *13th ACM Multimedia Systems Conference (MMSys2022)*. Chapter 10 summarizes the answers to the RQs examined in the thesis, the limitations of the thesis, and a description of future research and concludes the thesis.

This thesis is based on already published papers which are included as chapters. However, the text included from the papers was further adapted to adjust the writing style throughout the document.

	Article 1	Article 2	Article 3	Article 4	Article 5	Article 6	Article 7
Chapter	3	4	5	6	7	8	9
Research Focus	Data Collection Platform	Quality of Life and Connectivity	Application Launch Forecasting	Application Usage Modelling	Modeling Quality of Experience	Modeling Quality of Experience	Managing Quality of Experience
Paper Title	mQoL Smart Lab: Quality of Life Living Lab for Interdisciplinary Experiments	The Importance of Smartphone Connectivity in Quality of Life	Forecasting Smartphone Application Chains: a App-Rank Based Approach	You're Using This App For What? A mQoL Living Lab Study	Predicting Quality of Experience of Popular Mobile Applications from a Living Lab Study	Towards accurate models for predicting smartphone applications' QoE with data from a living lab study	Less Annoying: Managing end user QoE expectation's of Smartphone's Apps
Authors	A. De Masi, M. Ciman, M. Gustarini, K. Wac	A. De Masi, K. Wac	A. De Masi	A. De Masi, K. Wac	A. De Masi, K. Wac	A. De Masi, K. Wac	A. De Masi, K. Wac
Methodology	Empirical research	Empirical research	Scoping Review Design and Empirical research	Empirical research	Empirical research	Empirical research	Empirical research
Scientific Contribution	Produces an ambitious system	Thoroughly explores an area	Develops new method	Thoroughly explores an area	Derives superior algorithms	Derives superior algorithms	Opens up new area
Published in peer-reviewed conference/ Journal	Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Workshop UbiComp	Quantifying Quality of Life: Incorporating Daily Life into Medicine, Cham. 2021	2023 Proceedings of the 22th ACM International Conference on Mobile and Ubiquitous Multimedia	Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Workshop MHC	2019 11th International Conference on Quality of Multimedia Experience, QoMEX 2019 Acceptance rate: 43% Best Student Paper Award	Quality and User Experience, vol. 5, no. 1, p. 10, Oct. 2020, Springer Tracked for Impact Factor	2022 Proceedings of the 13th ACM Multimedia Systems Conference, MMSys 2022 Acceptance rate: 34%
Status	Published	Published	Submitted	Published	Published	Published	Published
Author Contributions	A.D., M.C. & M.G. wrote the paper K.W. supervised the project & provided feedbacks on the paper	A.D. analyzed the data & wrote the paper K.W. supervised & wrote the paper	A.D. analyzed the data & wrote the paper K.W. supervised & participate in previous version of the paper	A.D. designed & conducted the study, analyzed the data & wrote the paper K.W. supervised & wrote the paper	A.D. designed & conducted the study, analyzed the data & wrote the paper K.W. supervised & wrote the paper	A.D. designed & conducted the study, analyzed the data & wrote the paper K.W. supervised & wrote the paper	A.D. designed & conducted the study, analyzed the data & wrote the paper K.W. supervised & wrote the paper

Table 1.1 Synoptic Table of Dissertation Articles

Chapter 2

Related Work

Chapter Contents

- 2.1 Introduction 20**
- 2.2 Overview: Quality of Experience 20**
 - 2.2.1 From Quality of Service to Quality of Experience 20
- 2.3 Quantification and Modeling of Quality of Experience 21**
 - 2.3.1 QoE Influencing Factors 22
 - 2.3.2 QoE Assessment and Modeling Approaches 25
- 2.4 Smartphone Quality of Experience Assessment 26**
 - 2.4.1 Landscape 26
 - 2.4.2 Code Driven vs Human Driven Application Experience 27
 - 2.4.3 Device influences on QoE 27
- 2.5 Mobile User’s QoE Management and Services 28**
 - 2.5.1 QoE Management 28
 - 2.5.2 QoE Dependent Mobile Services 29

2.1 Introduction

This chapter presents the state of the art in the Quality of Experience (QoE) domain, from its origins to current approaches to modeling QoE, based on the type of service in the context of mobile services. How QoE could be quantified and enhanced via traditional and novel approaches (e.g., content adaptation), is included in this chapter as well.

Firstly in Section 2.2, the definition of QoE is presented as its application and adaptation concerning mobile services. Then, Section 2.3 illustrates how QoE is quantified and modeled, including the widely applied QoE modeling approaches and their application scenarios. We focus on the related multidisciplinary work (e.g., combining cognitive science and machine learning), due to the characteristic of QoE. For instance, QoE could be studied in cognitive science on how one forms and constructs an experience (i.e., from mental model (Carroll and Olson, 1988; Olaverri-Monreal and Gonçalves, 2014; Rouse and Morris, 1986) influenced by past events) and illustrating the impact of past experiences in creating new ones. In this section, we also explain the metrics and factors used in related QoE studies. Unlike quality of service (QoS) in the network and telecommunication domains, QoE is human-centric. The QoS quantitative metrics are succinctly discussed in this chapter. In addition to QoS metrics, other factors from the literature that influence user experience are presented and described as context. Section 2.4 targets QoE studies specified in smartphone usage. We summarize and reflect on the past studies focused on smartphone's application QoE modeling and its application to predict future QoE levels. Lastly, in Section 2.5, we provide an overview of the existing tools and frameworks for QoE studies, including a method for managing the perceived QoE of the end-user and its expectations. We also summarize and discuss the state of the art at the end of this chapter.

2.2 Overview: Quality of Experience

2.2.1 From Quality of Service to Quality of Experience

As aforementioned in Chapter 1, to clarify QoS and QoE, we start with their definitions and the brief study evolving history. The International Telecommunication Union (ITU) defines QoS as :

QoS: The totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service (ITU-T Recommendation P.800.1, 2019).

QoS typically targets telecommunications services with quantitative metrics (e.g., loss, delay, jitter, throughput) obtained from the devices on the network (e.g., switches, routers and cell towers) or on its edges. This technological approach is based on analytical approaches and empirical or simulation measurements. QoS addresses the performance aspects of physical systems (wired and wireless), while QoE covers QoS and includes the user experience broader domain (i.e., expectation and context). Finally QoS is rooted in the network domain. However, some experiences sometimes does not involve telecommunications or networks (for example, a video played locally on the smartphone).

QoS cannot provide quality metrics on end-user service satisfaction (e.g., video glitches, excessive waiting time in applications, and audio artifacts in music). Thus, missing the user aspects to understand the factors that influence an experience and the perceived quality of a communication service that provides an experience.

The Qualinet COST Action (IC1003) (i.e., a task force focusing on defining QoE) in their white paper (Le Callet et al., 2012) explain that QoE requires a multidisciplinary and multi-methodological approach for its understanding (i.e., mixed-methods). Also, they have analyzed the overall implications of QoE. They showed that QoE incorporates the user's assessment of system performance. The performances are influenced by context, culture and users' expectations. Also, the user's experiences are influenced by the system or service and ' their achievements, their socioeconomic status and issues, and their psychological traits.

QoE: The degree of delight or annoyance of the user of an application or service. It results from fulfilling his or her expectations with respect to the utility and/or enjoyment of the application or service in light of the user's personality and current state. (Le Callet et al., 2012)

Focusing on the user aspect, QoE was introduced to fill this gap and propose a codification of the path to quantify the quality of an experience of a service. Furthermore, user-centered QoE has taken over the role of network-focused QoS by increasing the importance of user-focused quality rather than the technical performance of the entire service. Although there are many differences between QoS and QoE, it is important to notice that QoE is often dependent on QoS (Hossfeld et al., 2020). The technical aspects of a system's performance can have a significant impact on some dimensions of QoE (Fiedler et al., 2010), especially in the case of multimedia systems.

2.3 Quantification and Modeling of Quality of Experience

QoE is quantified and modeled through multidisciplinary approaches, as it is influenced by many aspects. The technical aspect (i.e., hardware and software to provide the experience and consume the content) is an essential component of the entire experience, as described by Le Callet et al. (2012). The economic incentive also plays a role (Bilal and Erbad, 2017; Najjar et al., 2021) in QoE. For example, a smartphone user who uses network intensive application (e.g., 4k virtual reality video) may be willing to pay more for a higher download speed to satisfy their QoE. Meanwhile there are users may would like to reduce their mobile bills by limiting the download speed. Thence, these users would consciously reduce their expectations on application usage, which also impact their future QoE. Before quantifying and modeling QoE, we firstly introduce the four main components presented in existing studies that influence QoE. Then the widely used QoE quantification and modeling methods are presented as well as its application scenarios.

Additionally, we present the state of the art on smartphone application QoE. Finally, we explore the current QoE management methods in the domain.

2.3.1 QoE Influencing Factors

Based on existing studies, such as (Gao et al., 2020; Le Callet et al., 2012; Mazhar and Shafiq, 2018; Nam et al., 2016; Tiotsop et al., 2019), QoE is quantified and modeled within these following aspects: user perception, context, usage content and end-user device. We explore each aspect in the subsequent sections.

2.3.1.1 Human Sensory Perception

The human component is central in the QoE domain. Human sensory organs are biological factors that are directly linked with experience (e.g., eyes, skin, and ears). With the information obtained by the sensory organs, the brain cognitive functions process and manage the information, and quantify an experience as annoying or delightful with different levels/intensities. The Human Visual System (HVS), which plays an important role in user perception, has been studied via the Visible Difference Predictor (Daly, 1992) and provided information for the assessment of QoE.

Human visual perception of differences are based on the current perceived degree. Hence, some quality changes cannot be perceived by humans unless they are above a certain threshold. As such, Bouguer-Weber (Ross, 1997) presented a formula based on “just-noticeable difference” (JND) linking human perception to relative changes in stimuli, presented as the formula below. ΔI is the differential perception and is relative to the changes I (original stimuli intensity), with the characteristic constant k of the sensory modality.

$$k = \frac{\Delta I}{I} \quad (2.1)$$

The formula can be illustrated as such: If a light source is notably dim, people might notice a smaller fluctuation in intensity than they would if those same changes were made to brighter light. Based on JND, Seow (2008) proposed a set of time duration thresholds in a web browsing scenario to find the maximum tolerance (100 ms) that the user could define the system reacting instantly. Even though human beings can recognize disturbances of less than a second, the user may still feel uninterrupted.

In addition to HVS, cognitive and general psychophysiology is also important for QoE. Psychophysiology is defined *as the scientific study of the relationship between physiological and cognitive processes* (Browne, 2016). Engelke et al. (2017) provided a classification method, which is relevant to QoE assessment. The authors’ approaches to QoE assessment are similar to those used in affective computing (Picard, 2003). In which sensors are often used to collect biological information to assess valence and arousal (e.g. for stress detection or emotion recognition). For example, Arnau-Gonzalez et al. (2017) used electroencephalography (EEG) and peripheral electrocardiography (ECG) and electromyography (EMG) from physiological signals to evaluate the perceptual quality of video from these signals. Their results showed the potential of their proposed methods for an accurate evaluation of perceptual video quality.

Moreover, the review from Bañuelos-Lozoya et al. (2021) presents the cognitive states that are being investigated in the context of QoE evaluation. The authors reported a low number of studies

compared to cognitive state research in other contexts, such as driving or other critical activities. From the 29 studies ran between 2014 to 2019, the recognition methods of cognitive state were based partially on the acquisition of EEG, ECG, galvanic skin response (GSR) and eye tracking (ET) signals. However, this provides a starting point to analyze and interpret states such as mental workload, confusion, and mental stress from various human signals and propose more robust QoE models for multimedia content.

Furthermore, [Sackl and Schatz \(2014b\)](#); [Sackl et al. \(2012\)](#) have successfully linked unexpected quality perception ratings with the social-psychological theory of cognitive dissonance and provided a method to evaluate real-world economic scenarios in which subjects actually represent active decision-making agents. As such, [Sackl and Schatz \(2014b\)](#) presented an experimental setup that includes users as active and decision-making subjects in the context of subjective perception of video quality and the influence of cost. However, no defined thresholds could be obtained because the experimental setup was not representative of real video watching situations.

In addition, past memories influence the experience as shown by [Aru et al. \(2016\)](#); [Kahneman and Riis \(2005\)](#); [Strijbosch et al. \(2019\)](#). Thus, [Hosfeld et al. \(2011\)](#) were able to demonstrate the implications of the memory effect as a key influence factor for Web QoE modeling and derived extensions of their basic models. Finally, expectations are part of quantifying a user's QoE. [De Letter et al. \(2021\)](#) have shown the importance of expectation in that aspect in the domain of virtual reality (VR) QoE. The authors used the repertory grid technique ([Marsden and Littler, 2000](#)) which emphasized the holistic attributes of an experience combined with the precision of qualitative analysis. The attributes embody the meaning structures and values of the participant related to their experiences. Their results have shown that a QoE model for VR consists of attributed from three groups: users, content, and system. Their study participants ranked the content higher than the other attributes that influence their QoE. The QoE research on smartphone applications suffers from a lack of interest in the holistic aspects.

2.3.1.2 Content

The consumed content of an experience could influence QoE in different ways.

- It has certain signal properties that may be affected by processing, such as capture, delivery, or presentation ([Le Callet et al., 2012](#)).
- It is related to a message in the content ([Amour et al., 2018](#)).

Artists or content producers create experiences and may try to deliberately achieve predetermined user experiences (i.e., generating emotion in one consuming the content).

Content is subjectively processed from the point of view of the user experiencing it (i.e., the end-user) [Muscarì \(1985\)](#). The content creator assigns a certain form to the content (i.e., layout or format). It is related to signal properties (e.g. bitrate), but also symbolic properties (e.g., chosen colors and image ratio). The symbolic properties are linked to the artistic decisions which were made to construct the experience (i.e., aesthetic). Wherever the signal properties are associated with the

support on which the symbolic properties are presented. On the other hand, the content can be processed by the carrier (e.g., before transmission on the network), possibly changing its form, which will be perceived by the end-user [Barsalou \(1999\)](#).

Due to the perspective difference, the meaning assigned by the recipient will likely differ from the one intended by the creator ([Jekosch, 2005](#)). However, this is an ongoing topic under QoE research, how content and QoE are interrelated. As defined by [Le Callet et al. \(2012\)](#) QoE does not address the degree of success achieved to transmit the intended message, but rather how a technical system, technical processing, or a context change may have positively or negatively influenced the end-user perception of the experience. This is the scope within this thesis.

2.3.1.3 Device

The device is the object from which the end-user will interact and perceive their experience. The hardware performance of the device plays an essential role in the interaction and leads to the resulting QoE ([Dasari et al., 2018](#)). Even in the same content (e.g., text communication), the interaction experience is different from using smartphones, tablets and laptops, varying from using a software keyboard or a hardware keyboard. Thus, the expectations are also different interacting with different devices.

In communication applications, methods for text input influence directly the QoE. For example, applications often handle and enhance user text input by providing possible end words, phrases, or short sentences, or translation via autocompletion and the suggestion bar ([Pandey and Arif, 2020](#)). That is, a user will not expect the same input method when writing a text message on a laptop.

In addition to the input method, other hardware components also affect the user with direct interaction ([Ruiz et al., 2011](#); [Teng et al., 2006](#)). Taking the device screen as an example, its size ([Dunaway and Soroka, 2021](#); [Kim and Sundar, 2014](#)), luminosity (lux), color level, and technology (e.g., OLED, AMOLED, e-Ink) are the factors that have an impact on QoE. For less direct interaction, the data processing components (CPU, GPU, RAM, storage, and network) ([Narayanan et al., 2021](#)) which support the operating system and the application determine whether the user's intent could be satisfied. For a mobile device providing experience, battery level, related to its effect on application usage, and other hardware components (e.g., modem, memory, processor) can also be the influential factors of QoE in smartphones shown by [Ickin et al. \(2012\)](#).

2.3.1.4 Context

[Dey \(2001\)](#) exemplified the context as the implicit information on the situation exchanged by humans during communication to increase the information bandwidth. However, in human-computer interaction, this ability does not transfer well. Hence, by improving computer's access to context, it is possible to build a more useful computational service. As such, [Dey \(2001\)](#) defined context as :

Context: Any information that can be used to characterize the situation of an entity. An entity is a person, place or object that is considered relevant to the interaction between a user and an application, including the user and the applications themselves. Dey (2001)

Le Callet et al. (2012) characterized the context factors that influenced QoE. That is, the context embraces any situational property to describe the user's environment in terms of physical, temporal, social, economic, task, and technical characteristics cf. (Jumisko-Pyykkö and Vainio, 2010). These factors can occur at different levels of magnitude, dynamism, and patterns of occurrence, either separately or as typical combinations of all three levels. The physical context describes the characteristics of location and space, including movements within and transitions between locations. Temporal aspects of the experience, e.g., time of day, duration, and frequency of use (of the service/system), are covered by the temporal context. Costs, subscription type, or brand of a service/system are part of the economic context (Schatz et al., 2011). The experience can be perceived as focused or in a multitasking situation (i.e., task context), alone or with other people present or even involved in the experience (i.e., social context).

2.3.2 QoE Assessment and Modeling Approaches

In the literature, quantifying QoE has two meanings, one is to quantify the content to deliver (i.e., video) or the intent to satisfy (i.e., communication need). Several approaches (Chen et al., 2014; Wamser et al., 2015) have been developed to assess user perceptions both in a laboratory setting (controlled environment) and in-the-wild (uncontrolled environment). Some methods are originally from other domains and employed in QoE, such as Hossfeld et al. (2018). Studies reported by the QoE community, to a large extent, relied on the use of a standardized 5-point Absolute Category Rating (ACR) scale to calculate Mean Opinion Score (MOS) values (ITU-T Recommendation P.800.1, 2019). The ACR scale uses the following five-level ratings: 5 Excellent, 4 Good, 3 Fair, 2 Poor, and 1 Bad. The numbers may be displayed on the scale. One issue with the scales used in quantitative QoE evaluation is their limited bound at both ends (Hossfeld et al., 2017). Therefore, the individual rating scores of a participant are limited. The participant should be provided with a free text entry field to add any comment on the experience just rated. As such, Hossfeld et al. (2014) included the need for a feedback channel in QoE experiments, enabling a deeper annotation of the QoE events on hand. A classic channel is email to contact the study principal investigator (PI) with reports and questions that could impact the study results. In addition, the authors present their best practices and recommendations for crowdsourced QoE methods. For example, they indicated that the study protocol should contain reliability checks in the test design, during the test, and after the test. These checks increase the quality of the collected outcomes. Due to the nature of crowdsourced studies, the validity of the subjective ratings obtained could be problematic (e.g., a participant always rating 5), as well technical difficulties can arise (e.g., software update). Although, crowdsourced or in-the-wild studies have their shortcomings (e.g., participants dropping out, data collection issues and outside factors influencing

experiment), observed events in this context enable a closer understanding of QoE factors in daily settings (Capponi et al., 2019).

Once the experience has been reported by the study participants, this assessment is used to link QoS and other factors to a QoE level. As such, Sackl et al. (2015) proposed a logarithmic relationship between the QoS and QoE are revealed in the context of web browsing and file downloading (Sackl et al., 2016). The IQX hypothesis model (Fiedler et al., 2010) is used to describe QoE as an exponential function of QoS factors (i.e., describing the MOS-QoS relation). IQX relates the impact of absolute stimulus change to the current perception level. In other words, the change of QoE caused by a change in QoS depends on the actual current level of QoE reflecting the actual level of expectation. Fiedler et al. (2010) have shown that QoS metrics such as the packet reordering rate, due to delay jitter and packet loss, are easily matched to the QoE (MOS assessment score). Furthermore, (Menkovski et al., 2011) presented a nonlinear relationship between the video bitrate and the QoE of perceived users. Additionally, (Hofsfeld et al., 2019) showed that to derive QoE metrics in a system, it is necessary to use the corresponding QoS-to-QoE metric mapping functions derived from the user rating distribution in subjective research.

2.4 Smartphone Quality of Experience Assessment

Different from the section above, which discusses QoE in a relatively general context, in this section, we specify QoE in smartphone usage.

2.4.1 Landscape

The landscape in the smartphone industry has remained stable in the past ten years (Statista, 2022). New hardware makers have arisen to propose new incentives to use their platform/devices instead of the two most popular on iOS from Apple Inc and Android from Google/Alphabet such as Fairphone (2021). Fairphone proposes a repairable smartphone with a higher life duration. Samsung, Xiaomi and Huawei have made foldable smartphones, trying to enact new interactions based on this form factor. Apple only allows for the installation of an application on their smartphone from the Apple Store. However, any application can be downloaded online from any website and installed on the Android smartphone (after enabling one setting). This paradigm plays an important role in the population that uses these platforms. The expected interactions with a iOS device (e.g., iPhones) are different from an Android one (Barea et al., 2013; Furini et al., 2019; Novac et al., 2017). Apple has greater control over the application running on their device and provides a seamless and integrated experience on the applications on their platform (McAran and Shaw, 2020). Contrary to Android, in which the application not developed by Google teams has UIs suffering from a lack of uniformity. As such, the different UI styles directly affect the smartphone user (Liu et al., 2021).

Table 2.1 Selected QoE Studies*

Paper	Duration	# Participants	Device	Goal	Application Focused	In the Wild	Context		Scale
							Location	Physical Activity	
Hosfeld et al. (2011)	X	127	Laptop Desktop	Memory Effect on QoE	No	Partially	No	No	ACR-5 MOS
Schatz and Egger (2011)	X	17	Laptop	Web Browsing File Download	No	Yes	No	No	ACR-9 MOS
Aggarwal et al. (2014)	X	X	Smartphone	Passive QoE Estimation from QoS	Yes	No	No	No	X
Chen et al. (2014)	9 Months	20	Smartphone	Map UI State and QoS to QoE	Yes	Yes	No	NO	X
Hosek et al. (2014)	X	108	Smartphone Mobile phone	Web Browsing File Download File Upload	No	No	No	No	ACR-5 MOS
Casas et al. (2016)	14 Days	42	Smartphone	Map QoS to QoE	Yes	Yes	Yes	No	ACR-5 MOS
Lee and Cha (2017)	3 Days	15	Smartphone	QoE from UI and QoS	Yes	Yes	No	No	X
Boz et al. (2019)	2 - 4 Weeks	292	Smartphone	QoE from QoS, App Popularity and User Profile	Yes	Yes	No	No	X

x: not reported in the paper

*: excluding our own studies

2.4.2 Code Driven vs Human Driven Application Experience

Modern application development methods are based on Test Driven Development (Astels, 2003) of the source code. The engineering aspects, as well as the UI, are tested repetitively by software programs, A/B testing (Kaufmann et al., 2014) and released the beta applications to a predefined user group. The collected metrics (e.g., number of crashes, application engagement, frame rate) are used to find issues, support decision-making, and enhance the application for a future release. However, these methods often exclude the context of the application users. The systematic mapping study on the smartphone application from Luo et al. (2020) showed the need to test an application in-the-wild, as it is often ignored by application developers. Therefore, it is important to investigate the QoE application of the smartphone in the context of the experience.

2.4.3 Device influences on QoE

This section focuses on QoE assessment studies on different devices (i.e., smartphone, TV, laptop) with various content types. In short, there is one study from Casas et al. (2016), which tried to achieve the same goals as S1 with different methods and protocols. Their study focused on data collection with network QoS to map QoE levels. Meanwhile, their participants were instructed to accomplish a specific task in an application and then rated the experience. This method could introduce a bias (Hornbæk, 2013) since the intent of using the application did not originate from the participant.

Table 2.1 presents a selection of QoE studies focusing on experiences of laptop, desktop, and smartphone application. In general, most studies were performed with the ACR MOS scale to report QoE levels. However, none of them integrated the physical activity of their participant as a context factor in their QoE assessment, which was shown by Chen et al. (2016); Verdejo et al. (2010) to influence users' QoE.

Smartphone application QoE studies mainly focus on the assessment of an experience directed by the study investigators. These studies were conducted in laboratory setting or in-the-wild, with

an unrepresentative participants' population, due to difficulties in participant recruitment (Koo and Skinner, 2005; Lazar et al., 2017; Tahaei and Vaniea, 2022) or bias in selecting (Linxen et al., 2021; Offenwanger et al., 2021). The ITU recommends a minimum of 24 participants (in the laboratory experiment setting) or 35 participants (in the public environment setting) for subjective evaluation of audiovisual quality (ITU-T, 2021) (e.g., films).

The in-laboratory studies enable more control of the experiment protocol and design (Chamberlain et al., 2012), confounding factors are limited in these studies. However, it does not capture important contextual events that may influence participants' QoE, which is possible with in-the-wild study (Kjeldskov and Skov, 2014). Moreover, in-the-wild studies also introduce other challenges as technical feasibility (data logging), privacy issues, and reproducibility of the derived results (Ballou et al., 2021). Protocols for QoE studies are not standardized or registered (Garcia and Casas, 2020). However, a common pattern in QoE assessment looks as follows: researchers observe the action of their participants, collect passive data (interaction information), trigger a task to accomplish (stimuli), or wait for a specific intent from the participants. Once the task is finished, the participant annotates their QoE level (Hossfeld et al., 2014).

Another study from Boz et al. (2019) have shown a method to model QoE from data collected in-the-wild. The models' features were networking metrics (i.e., Wi-Fi and cell signal strength), time context information (day of the week and time of day) and user profile (age). However, they did not use a validated scale to collect the ratings of their participants, which increase the measurement error. Contrary to the best usage and recommendations (Hossfeld et al., 2014) employed in the QoE domain. Furthermore, the features presented for their models did not integrate any application usage duration data.

Overall, the previous QoE assessment studies often had a low number of participants and were conducted in a laboratory setting, missing valuable contextual information and employing nonvalidated scales to measure QoE.

2.5 Mobile User's QoE Management and Services

Based on the literature, we classified the methods to manage or enhance the user's QoE into two groups. The first group is content adaption, which modifies the experience's content accordingly to satisfy the user's expectation in their current context. The second is intervention, based on proposing alternative content and experience to limit exposure to low QoE in their current context.

2.5.1 QoE Management

2.5.1.1 Transport Medium and Content Adaptation

Current research in the QoE domain (Gao et al., 2020; Kiani Mehr et al., 2021; Li et al., 2018) mainly focuses on adapting the content delivery method and the transport medium (i.e., via an adjustment of the network protocols parameters). For example, Szabó et al. (2016) leveraged the QUIC protocol

(not an acronym according to [Iyengar and Thomson \(2021\)](#)) to enhance multimedia content QoE. The QUIC protocol defines an object priority assignment, which the authors employed to enhance the video initial buffering time with success (i.e., 6-49% faster depending on the video properties and network condition).

Another method to maximize the end-user QoE focuses on transforming the content (i.e., [Bilal and Erbad, 2017](#); [Moldovan and Muntean, 2017](#); [Petrangeli et al., 2019](#)). These works are based on reducing the size of the files to serve to the user; dynamically changing their base properties (e.g., bitrate) on context changes (e.g., better or worst signal strength), without impacting the perceived QoE. However, to the best of our knowledge, none of the works proposes to manage the user expectation and so its experience. For video or audio streaming application an indicator of the current resolution and bitrate is often present. All the methods are transparent to the user. As such, no QoE level indicators are shown ([Wamser et al., 2015](#)).

2.5.1.2 Intervention

Interventions on smartphones are widely present in health-based research and marketing (e.g., nudging ([de Ridder et al., 2022](#))). One example, a smartphone-based intervention showed success in motivating physical activity in a student population ([Muntaner-Mas et al., 2021](#)). The end goal of the intervention was to change the behavior of the participants through interactions with an application or notifications ([Morrison et al., 2017](#)). Although the choice of application and the content of notification are often set up to direct the users to a certain action ([Visuri et al., 2019](#)), dynamic systems exist to serve a wide range of purposes. For example, determining the best time to interrupt a smartphone user with notification ([Mehrotra et al., 2015](#)) to forecasting application usage. In such context [Cao and Lin \(2017\)](#) proposed a system predicting the next-used application and recommend it to the user, influencing the launch of the next application. However, the aforementioned systems did not apply models or methods to manage the users' expectations. Thus, intervention is one tool to influence a smartphone user behaviour.

2.5.2 QoE Dependent Mobile Services

This section goal is to present the mobile services which depend on QoE for a better experience.

There are several companies (e.g., Netflix, Youtube, Hulu) that employ QoE metrics to quantify the quality of their services ([Bampis and Bovik, 2018](#)). Specifically, in the domain of multimedia services, providers monitor the consumption of their content on multiple levels. For video services, which consume a large amount of bandwidth due to the nature of the content, they continuously adapt the video quality based on the information collected by the end-user devices (e.g., amount of skipped frames and technology of network connection) ([De Pessemier et al., 2013](#); [Gao et al., 2020](#); [Juluri et al., 2015](#); [Mazhar and Shafiq, 2018](#); [Nam et al., 2016](#); [Tiotsop et al., 2019](#)). Meanwhile, other multimedia content, such as audio content (podcast, music, and phone calls) and text content, are also QoE enabled

(Sackl et al., 2015). Experiences in software applications (smartphone applications) and web browsing often include different types of multimedia content (Baraković and Skorin-Kapov, 2017).

Often, services are using QoE work as a control loop (Du et al., 2017; Rehman Laghari and Connelly, 2012). An experience is observed (e.g., engagement with an application, game, or video session) and its quality is continuously quantified. An expected QoE score is computed and updated; if scores reach a certain threshold, the content flow is optimized to go below an acceptability threshold (Moldovan and Muntean, 2017; Song et al., 2019). Such systems are dominant in QoE-aware video (Abar et al., 2020; Mozetić et al., 2021) and edge computing services (Xia et al., 2020a) (e.g., for accelerated video and image processing). In this way, service providers can monitor and forecast the amount of interruption (e.g., server disconnect, low bandwidth events, and short usage session) in the user interaction with the service (Seufert et al., 2019).

Many researchers have contributed in QoE or QoE related studies, proposed theories, experiment protocols, quantification, and modeling approaches. However, to the best of our knowledge, there are no long-term QoE studies in smartphone application usage, taking the user context and QoE management into consideration. Based on the state of the art, we proposed our research questions (Chapter 1) and designed the experiments S1 and S2 to answer them. Detailed information will be presented in the following chapters.

Chapter 3

Article I: mQoL smart Lab: Quality of Life Living Lab For Interdisciplinary Experiments

Published in Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, New York, NY, USA, Sep. 2016, doi: 10.1145/2968219.2971593.

Chapter Contents

3.1	Introduction	32
3.2	Current Platform: mQoL Living Lab 1.0 (mQoL-Lab)	33
3.2.1	mQoL 1.0 platform requirements (2012)	33
3.2.2	mQoL 1.0 platform limitations	35
3.3	New mQoL Platform: mQoL Smart Lab 2.0	35
3.3.1	Parse: Open source cloud solution (scalability)	35
3.3.2	MongoDB: Database (scalability and flexibility)	36
3.3.3	Visualization tools (flexibility)	36
3.3.4	Machine Learning (scalability and flexibility)	36
3.3.5	Experiments Builders (flexibility)	36
3.3.6	Targeting Potential Users (scalability and flexibility)	37
3.4	Discussion and Conclusive Remarks	37
3.5	Revisions to the Published article	38

Abstract

As a base for hypothesis formulation and testing, accurate, timely and reproducible data collection is a challenge for all researchers. Data collection is especially challenging in uncontrolled environments, outside of the lab and when it involves many collaborating disciplines, where the data must serve quality research in all of them. In this paper, we present our “mQoL Smart Lab” for interdisciplinary research efforts on individuals’ “Quality of Life” improvement. We present an evolution of our current in-house living lab platform enabling continuous, pervasive data collection from individuals’ smartphones. We discuss opportunities for mQoL stemming from developments in machine learning and big data for advanced data analytics in different disciplines, better meeting the requirements put on the platform.

3.1 Introduction

In the last years, smartphone penetration in individuals’ lives enabled acquiring data about their behaviors and context for extended periods of time and in their natural environments, i.e., “in-situ”. Social and behavioral, as well as medical, scientists, are now leveraging it –performing experiments in which it is necessary to assess individual’s state repeatedly “in-situ”. They often require collecting data throughout the individual’s day in a non-invasive manner, i.e., without influencing their lifestyle and habits. To design and develop new experiments and to be able to understand and model the variability of human behavior “in-situ”, a simple computer science approach to user studies, in which we (computer scientists) collect data, derive algorithms, and then experiment with them, providing final answers to social and behavioral scientists, is not sufficient. It becomes necessary to introduce an interdisciplinary approach and to build an adequate technological infrastructure to let social and behavioral scientists acquire accurate, timely and reproducible data. We motivate this statement given our research experience, as follows. In March 2012 we have launched the mQoL Mobile Computing and Communications Living Lab ([Gustarini et al., 2013](#); [Wac et al., 2015a](#)) in our Quality of Life group at the University of Geneva. It enables researchers to run studies and collect data from smartphone users in the Geneva region in Switzerland. The Living Lab has more than 40 individuals participating in different studies; some of them are participating in the lab since 2012 to date. We have published several results about mobile computing technologies, Human Computer Interaction and behavior assessment “in-situ” - from Quality of Service (QoS) ([Wac et al., 2015b](#)), Quality of Experience (QoE) ([Ickin et al., 2013](#)), to intimacy ([Gustarini et al., 2016b](#)), stress ([Ciman et al., 2015](#)) and context awareness ([Gustarini and Wac, 2012](#)). Even if the mQoL platform performed well for our experiments and data collection, along our research efforts we understood that to fulfill the new requirements of the interdisciplinary approach in data collection and experimentation a full redesign of the current platform is necessary. The ultimate goal of the platform is to enable experimentation that facilitates the improvement of Quality of Life of its users. In this paper, we present the current mQoL Living Lab platform, its current shortcomings and the redesign and prototype of the new mQoL Smart Lab. The

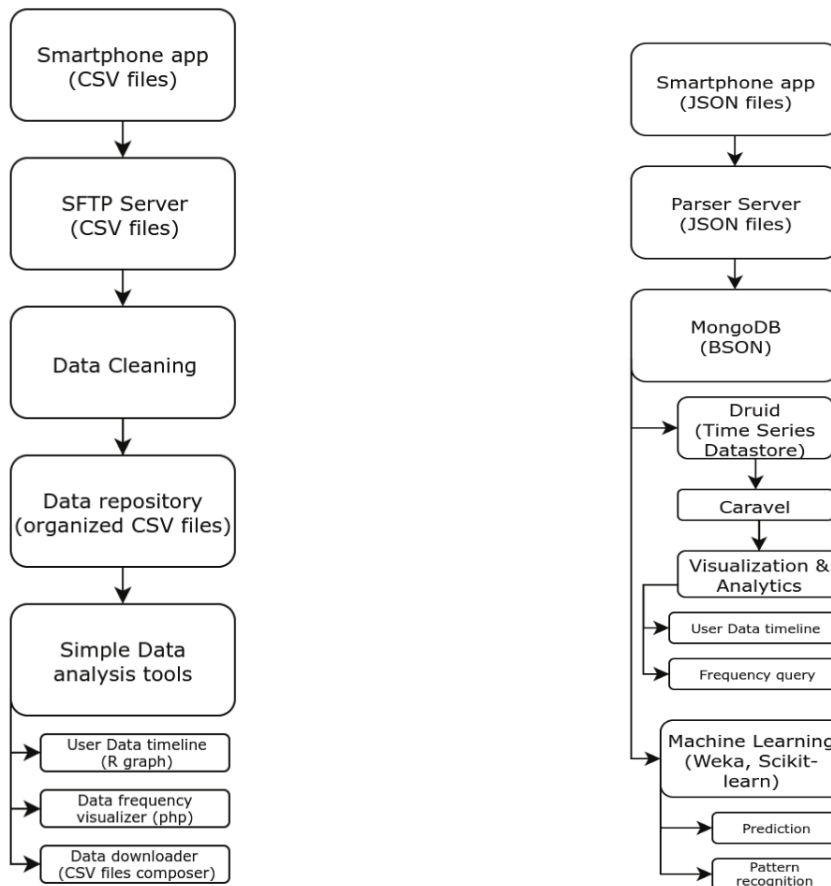
redesign is specifically aiming at improving the platform scalability and flexibility, which will facilitate interdisciplinary research efforts. Some other platforms for “in-situ” experiments and data collection exist. For example, the University of Buffalo “Phone-Lab.org” platform (Nandugudi et al., 2013), or the “AWARE” framework of the University of Oulu (Ferreira et al., 2015) are frameworks and infrastructures for data collection. They mostly focus on research aspects in computer science like scalability, accuracy and privacy-preservation for collecting vast amounts of raw data. OpenmHealth.org is a longitudinal framework aiming at collecting raw data related to the individual health and wellbeing, but it is intended only for computer scientists collaborating in diverse mhealth projects. The MITs Human Dynamic Lab (MediaLab, 2016) focuses on computational social sciences; however, it does not release its framework for data collection or the data itself. Finally, the Kalvi’s HUMAN (Kavli, 2016) project focuses on socio-demographics and urban factors influencing individual’s life, involving 2,500 families for 20 years. The biggest shortcoming of all these platforms is that they do not embed in their design an interdisciplinary approach, but focus only on data acquisition and the specific data analysis.

3.2 Current Platform: mQoL Living Lab 1.0 (mQoL-Lab)

The platform comprises two main components: the mQoL logger and the mQoL server. The mQoL logger is an Android app installed in our living lab participants’ phones. The app collects the data we are interested in, including, but not limited to screen touches, screen ON- OFF events, apps use and duration, wireless access network type and delays. The mQoL logger synchronizes with the server, which then processes the data and allows us to create simple visualizations and download the data for further analysis (the simple data flow is in Figure 3.1a). We built this platform at the beginning 2012 given the requirements of research projects at that time.

3.2.1 mQoL 1.0 platform requirements (2012)

The main requirements for the current platform were: fast to build, own hosting, security and privacy, reliability, and being non-intrusive for study participants. It had to be fast to build because of an urgency of own research. We avoided any complex server logic and used SFTP (security requirement) to send the data stored as simple CSV text files from the mQoL logger to our server (Figure 3.1a). To comply with privacy laws, we hosted the data in our university servers. To further preserve the privacy of our participants, we do not store any personal identifiers along with the data on the server. The data collection and transmission had to be reliable. Therefore, we decided to store data on users’ devices first and transfer it only over Wi-Fi. We do not remove the data from the users’ devices until the uploading operation is successful. Otherwise, we retry the whole upload. The QoL server has a nightly process that cleans the data from duplicates and broken records (Figure 3.1a). The cleaning operation prepares the data for the simple visualization engine and easy download (e.g., we can select time periods, data kind, and users’ identifier to download CSV ready for the analysis).



(a) mQoL 1.0: The flow of the data from participants' smartphones and an automatic process of cleaning it, enabling its further use.

(b) The flow of the data in the mQoL Smart Lab platform. Data comes from participants' smartphones and the Parse server manages the data cleaning, storage, visualization

Figure 3.1 Data Flow in the mQoL-Lab Platform

3.2.2 mQoL 1.0 platform limitations

Having this platform for the experiments was precious for our simple computer-science driven research along the past years; we also understood its limitations. Right now, we have the goal of supporting an interdisciplinary approach, including use of the platform for experiments run by psychologists, sociologists or economists, and at the current stage, we have two main challenges: scalability and flexibility. The current architecture does not allow us to scale to more than 80-100 users. We cannot distribute our mQoL logger out of the controlled environment of our living lab. Server resources are limited (e.g., memory and CPU), and the data collection through SFTP in CSV files does not allow us easily to scale our experimental and data analysis needs. The platform is not flexible enough: neither we (nor our interdisciplinary collaborators) cannot easily plug in new components such as event-based collection triggers, new experiments (with open to the world onboarding campaigns), machine learning tools (ML), more complex visualizations engines, and so on. We decided to evolve our current platform to a next level

3.3 New mQoL Platform: mQoL Smart Lab 2.0

The interdisciplinary approach goal is to allow any type of researcher to experiment with datasets and run different algorithms, e.g., a sleep analysis from smartwatches data could be process in a few clicks. We identified new components able to fulfill the new requirements, as follows: server (Parse – see sidebar for links), database (MongoDB), data visualization tool (Caravel), ML tools (TensorFlow, Scikit-learn, and Weka), experiment builders (our own modular Android library, Gradle and Google Play API), and marketing tools (Facebook Ads and Google AdSense). All of these components enable scalability and flexibility, as follows. The updated data flow is presented in Figure 3.1b.

3.3.1 Parse: Open source cloud solution (scalability)

Parse is an open source backend solution for infrastructures requiring scalability. It is based on Node.js and requires a MongoDB database. Parse supports user authentication, enabling creation of an anonymous login for study participants. It also features a key based client verification to verify the genuineness of the data coming from participants. It allows sending notifications to smartphones, to communicate with study participants (e.g., notify them about a new experiment). It also features live queries to subscribe to database updates and trigger events in the server or client apps. This feature allows us to automate future data analysis tasks, or to respond to specific events with actions. Parse offers clients API (e.g., Android, iOS, PHP, Javascript, and more) to easily implement the data transmission between our participants' smartphones (or any other device we may use) to our Parse instance.

3.3.2 MongoDB: Database (scalability and flexibility)

The new architecture will use MongoDB, a NoSQL database. It offers features that our previous data management scheme does not have: high scalability, load balancing, fault tolerance, replication, aggregation, and querying. Multiple programming languages support access to MongoDB database and allow for fast processing of the data. For example, our collaborators in e.g., psychology and sociology would get access to data via their regular interfaces (e.g., R) enabling them to run data analytics and conduct statistical significance tests directly on the data. Since most of the collected data are time series, we are also exploring the area of Time Series Data Store (TSDS), a special class of NoSQL database highly optimized for handling time series data. Overall, having different programming languages supporting the access to the MongoDB database, it is easier for researchers with diverse backgrounds and data analytics skills to have direct access to the data source and the database.

3.3.3 Visualization tools (flexibility)

The interpretation of data is often the next step after the collection procedure. We may need to visualize complex data to detect a particular trend, or to allow the researcher to validate or refute their hypothesis. There exist new powerful open-source technologies to deal with big data visualization coming from millions of users. For example, AirBnB Inc. designed [Caravel](#) as a data exploration platform to be visual, intuitive and interactive. It allows the researcher to write multiple, easy, and fast query to a database and output the data in a powerful visualization form, e.g., distribution of mobile applications usage of participants aged between 18 and 23.

3.3.4 Machine Learning (scalability and flexibility)

Machine Learning (ML) enables the design of complex algorithm simplifying manual data analytics task (which we have followed so far). We plan to integrate ML into our data analytics platform, enabling the platform to provide predictive services for the researchers running the experiments with participants. For example, one of the future goals of the platform will be to predict the application use experience of nomad smartphone user via ML, or to assess intimacy or stress levels of individuals “in-situ”.

3.3.5 Experiments Builders (flexibility)

We envision a special module that allows researchers of any discipline without coding skills to create their apps to collect data and perform “in-situ” studies. The tool, e.g., a web platform, will use our modular Android library to package the sensors in the app, Gradle to automatically create an Android apk file and Google Play API to deploy the app for distribution. For example, an economist could create an app to perform self-reports based on the Experience Sampling Method (ESM) ([Hektner et al., 2007](#)), about situational marketing, launch it on the Google Play store and distribute it to participants.

3.3.6 Targeting Potential Users (scalability and flexibility)

One of the difficulties to run studies is to attract a representative group of participants. In the past, we used the mQoL website to promote the living laboratory and the benefits of participating in our studies. Our new approach consists of using the powerful online ad platforms, e.g., Facebook Ads, Google AdSense, to display ads for specific studies to a representative part of the target population. For example, a psychology researcher could need twenty new participants between 20 to 45 years old who live in the Geneva canton and are likely buying luxury goods. With Facebook Ads, we can send the ad about the new study directly on Facebook feeds of people fulfilling these criteria. Discussion and Conclusive Remarks Based on the previous experience and the new requirements to support an interdisciplinary approach, especially including social and behavioral research, it became apparent that we require a new platform. Therefore main goal of the mQoL Smart Lab approach is to redesign the existing smartphone-based research platform (MQoL Living Lab) to be able to support the collaboration between scientists from different research fields, e.g., computer science and social and behavioral sciences. In particular, the platform will take into account the fact that different research areas, in the future, could require the integration of various modules, visualization tools, and machine learning algorithms. It will support the ad-hoc add-on of new instruments for data collection analysis, to improve the collected data quality and timeliness, and reproducibility of the results acquired based on this analysis. Thanks to the interdisciplinary approach and the collaboration with researchers from different research areas, the mQoL Smart Lab will boost behavioral and computer science research. We will develop interactive, smartphone-based behavior assessment technologies, tools, and algorithms that are unobtrusive, that we can deploy “in-situ” and longitudinally, and allow to combine self-reports, experimental tasks, and data from the smartphone sensors. Using this platform, we will perform large-scale innovative research in different areas, by bringing experimental paradigms into real-world situations. All these insights will provide an essential basis for research towards enhancing people’s life quality following the mission of our lab.

3.4 Discussion and Conclusive Remarks

Based on the previous experience and the new requirements to support an interdisciplinary approach, especially including social and behavioral research, it became apparent that we require a new platform. Therefore main goal of the mQoL Smart Lab approach is to redesign the existing smartphone-based research platform (MQoL Living Lab) to be able to support the collaboration between scientists from different research fields, e.g., computer science and social and behavioral sciences. In particular, the platform will take into account the fact that different research areas, in the future, could require the integration of various modules, visualization tools, and machine learning algorithms. It will support the ad-hoc add-on of new instruments for data collection analysis, to improve the collected data quality and timeliness, and reproducibility of the results acquired based on this analysis. Thanks to the interdisciplinary approach and the collaboration with researchers from different research areas,

the mQoL Smart Lab will boost behavioral and computer science research. We will develop interactive, smartphone-based behavior assessment technologies, tools, and algorithms that are unobtrusive, that we can deploy “in-situ” and longitudinally, and allow to combine self-reports, experimental tasks, and data from the smartphone sensors. Using this platform, we will perform large-scale innovative research in different areas, by bringing experimental paradigms into real- world situations. All these insights will provide an essential basis for research towards enhancing people’s life quality following the mission of our lab. The mQoL Smart Lab platform we research aspires to become a base upon which a whole interdisciplinary ecosystem can be build.

3.5 Revisions to the Published article

This articles was original published in 2016, it was supersede by [Berrocal, Manea, De Masi, and Wac \(2020b\)](#) “*mQoL Lab: Step-by-Step Creation of a Flexible Platform to Conduct Studies Using Interactive, Mobile, Wearable and Ubiquitous Devices*”. This article describes the data collection platform that we developed to support S1 and S2.

Chapter 4

Article II: The Importance of Smartphone Connectivity in Quality of Life

Published in *Quantifying Quality of Life: Incorporating Daily Life into Medicine*, Eds. Cham: Springer International Publishing, 2022, doi: 10.1007/978-3-030-94212-0_23.

Chapter Contents

4.1	Introduction	40
4.2	Related Work	41
4.2.1	Mobile Network Connectivity and QoL	41
4.2.2	Smartphone Apps and Their Impacts on QoL	42
4.2.3	Smartphones as Sensors of Daily Life	44
4.3	Mobile Network Connectivity Study: Methods	45
4.3.1	Data Collection Periods and Overall Summary of the Collected Data	45
4.3.2	Measurement Framework: mQoL-Log	47
4.3.3	Final Dataset	47
4.3.4	Features Derived from Mobile Network Connectivity	49
4.4	Mobile Network Connectivity: Results	53
4.4.1	Network Access Technology	53
4.4.2	Signal Strength	56
4.4.3	Data Consumption	56
4.4.4	Users' Physical Mobility	59
4.5	Discussion	60
4.5.1	Discussion of Overall Results	60
4.5.2	Study Limitations	62
4.5.3	Quantified Self Movement	62
4.5.4	QoL Technologies	63
4.6	Conclusion	64

Abstract

Mobile network connectivity enables individuals to use various Internet-based applications and is nowadays an integral part of the physical environment. More specifically, this connectivity shapes individuals' modes of gathering information and their communication capabilities. In turn, this impacts the individual's decision-making and, in the long term, may influence their health and quality of life (QoL). This chapter focuses on longitudinal modeling of the availability of mobile connectivity such as Wi-Fi and 3G or 4G for individuals living in the Geneva area (Switzerland). We analyze connectivity over five years (2015–2020) based on data collected from 110 mQoL (mobile QoL) Living Lab participants. The participants are from three different cohorts corresponding to distinct data collection periods (2015–2017, 2018–2019, 2020). We derive four features that quantify an individual's connectivity level: the network access technology (Wi-Fi or cellular), signal strength, the overall data consumption (upload and download), and the participants' mobility patterns while connected. We also compare the connectivity levels of the three cohorts over time. Our findings reflect the relations between mobile connectivity and the smartphone network activity of the mQoL study cohorts during their daily activities, which may impact their QoL. We summarize the results and conclude this chapter by exploring the different QoL technologies and services enabled by mobile connectivity. However, the effects of connectivity on specific QoL domains, such as psychological aspects (i.e., positive/negative feelings) or social relationships, should be investigated further.

4.1 Introduction

The World Health Organization (WHO) has defined Quality of Life (QoL) as an “individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards, and concerns.” The WHO expands this definition across several domains, namely physical and psychological health, social relationships, and the environment. In this chapter, we focus on one facet of the environmental domain: the physical environment. We explore the availability of mobile network connectivity in one's environment without considering other variables that contribute to this environment, such as noise, pollution, climate, and the general aesthetic. Determining the impacts of connectivity on an individual's QoL is important for considering improvements or adverse effects on their day-to-day life.

Wireless networks have been present in our physical environment since the invention of over-the-air transmission of information (ALOHAnet ([Abramson, 1977](#))) in 1970. Recent developments in communication technology have now made it affordable to own a powerful, ubiquitous, network-enabled device. Today, wireless networks are present throughout the shared physical environment, especially in the developed world and in areas with high population density. Indeed, the accelerated digitalization of the population can be attributed to the global adoption of smartphones. The number of smartphone users reached 3.2 billion worldwide in 2019 and will continue to grow ([Katz and Calhorda, 2018](#)). Likewise, the networks that support them have been deployed at a similar pace and are

continuously updated to cover larger areas and upgraded to utilize new technologies (e.g., from 3G to 5G).

The majority of mobile applications require an Internet connection, and in this study we focus on connectivity to mobile networks, whereby human-to-human interaction is enabled by computer-based networks. Networks support instant information transfer in various formats, including text, image, and video, and enable the necessary interaction between nodes (i.e., people or machines). Furthermore, they provide access to a number of services that can be used to improve an individual's decision-making capabilities and ultimately their QoL. A 2018 study by [Chan \(2018\)](#) found that smartphone use predicts relationship quality and subjective well-being, while [Kim \(2018\)](#) suggested that the use of information and communication technology, such as smartphones, in old age generally plays a positive role in enhancing the psychological, mental, and social aspects of one's QoL.

This chapter presents features of mobile network connectivity derived from smartphone use data collected from different cohorts in the Geneva area (Switzerland) between 2015 and 2020. We explore four connectivity features and examine the evolution of connectivity during the last five years as derived from data gathered unobtrusively from the consented mQoL (mobile QoL) Living Lab participants.

This chapter is structured as follows. We present the literature review in Section 4.2. In Section 4.3, we provide the study parameters, describe the collected data, and outline the studied connectivity features. In Section 4.4, we report the results obtained from the analysis of the features. In Section 4.5, we discuss the limitations of the study and different approaches to connectivity quantification. Finally, in Section 4.6, we describe the lessons learned and provide recommendations for future areas of work, especially the quantification of the impact of mobile connectivity on QoL.

4.2 Related Work

4.2.1 Mobile Network Connectivity and QoL

Previous work has shown the benefits of deploying mobile networks in rural and developing areas (e.g., Ghana, Nigeria, Kenya, and Tanzania) ([Ericsson, 2010](#)). Researchers have found that it facilitates improved communication between the local population and distant services such as health and governance. The same authors have documented income growth in the Southeast Asia region in the last 10 years due to the rising usage of mobile applications and voice calls as the population gained access to new services and information relating to the weather, agriculture, finance, and music, for example. The income growth has only been reported in low-income countries, but surprisingly, in 2018, the GSM Association ([Association, 2019](#)) found that the top reason to use mobile instant messaging was the same for low, middle, and high-income countries. This indicates that the benefits of messaging applications are not the prerogative of high-income countries. In recent years, messaging applications have created new markets and services that are available on their platforms. For instance, WeChat (est. 2011), Facebook Messenger (est. 2011), and WhatsApp (est. 2009) have all integrated payment

functions into their applications in selected countries including China (WeChat Pay), Brazil, and the USA. Before the prevalent use of smartphones, the development of mobile payment solutions using a fast and reliable network was stagnant. Today, mobile networks are a critical gateway to the digital economy, as these solutions have been widely adopted to simplify the exchange of money and goods. Overall, 90% of Chinese tourists claim that they would use WeChat Pay overseas if given the opportunity (Association, 2019).

The direct impact of broadband network access on Gross Domestic Product (GDP) per capita has also been studied; one investigation found that a 10% increase in broadband penetration has a notable impact on GDP per capita, increasing it from 0.9 to 1.5 percentage points on average for the Organisation for Economic Co-operation and Development (OECD) economies. Furthermore, the authors explained that if digital services are established alongside a reliable infrastructure, new services will be created (Katz and Callorda, 2018).

In recent years, the Asia-Pacific region has been improving its environmental QoL through connectivity and will continue to do so particularly by way of smart city initiatives (Insights, 2017). Such initiatives are described as cross-sector endeavors that link people to public and private infrastructures. Connectivity is crucial for smart city services, from the use of Internet of Things devices and a cloud-based platform to monitor and analyze air quality at street level, to the publishing of open data by public authorities to enable faster development of online-based services. In summary, a link between mobile connectivity and QoL around the world has been proven to exist—to such an extent that connectivity has a direct impact on a country's GDP.

4.2.2 Smartphone Apps and Their Impacts on QoL

The revolution in mobile devices, which have evolved from basic cell phones to smartphones, has created a new market for mobile applications. New application types were created for those devices, and as of November 2020, the Google Play Store hosted 2.56 million different applications across 32 application categories and 17 game categories (Google, 2020). Two application categories that may have a direct impact on users' health are (i) health and fitness, including personal fitness, workout tracking, dieting and nutritional tips, health, and safety applications, and (ii) medical, including drug and clinical references, calculators, medical journals, news, and handbooks for healthcare providers.

Health and fitness applications such as food diaries allow users to track their food intake for multiple purposes. These applications connect to a central database that contains nutritional information about various foods (e.g., calories, carbohydrates, fat, and vitamin content). Users have to scan the barcode on a food item or use the search box to find and manually add the specific food and item weight, and the application computes its total nutritional value. Chen et al. (2017) reported that users' quality of experience is much higher with smartphone application diaries than with pen and paper diaries. They also found that diabetic patients using application diaries reported a better food intake control than those using pen and paper diaries. Furthermore, a recent study by Bracken and Waite (2020) demonstrated that non-patient users wishing to lose weight (e.g., managing pre-

obesity) and others wishing to gain weight (e.g., building muscle) utilize diary applications to attain their nutritional goals.

Medical applications are oriented towards health workers and healthcare practitioners. These professionals can use these applications as a productivity tool in their work, which enables them to automate necessary tasks (White et al., 2016). Recent work (Ventola, 2014) has indicated the advantages of medical applications: they increase access to point-of-care tools, thus improving patient outcomes that stem from better clinical decision-making. Wattanapisit et al. (2020a) investigated whether a medical smartphone-based application can replace a general practitioner. They praised the use of an application for tasks such as recording medical history, making diagnoses, promoting health, performing some physical examinations, and assisting in urgent, long-term, and disease-specific care. However, the application was unable to support clinicians in performing medical procedures, appropriately utilizing other professionals, or coordinating a team-based approach. A recent literature review by Wattanapisit et al. (2020b) focused on medical counseling for physical activity and returned mixed findings regarding the usability and utility of medical applications. The review suggested that technical issues and the complexity of programs were barriers to usability, thereby implying the possibility of unfavorable patient outcomes such as inaccurate advice and diagnoses.

Mobile network connectivity plays a significant role in always-online smartphone applications. These applications may help to enhance an individual's decision-making and thus result in an improved QoL through connectivity to the Internet. However, such applications can also lead to the reverse effects. One example is smartphone addiction. According to the observations of Kwon et al. (2013b), "the overuse of smartphones can be easily seen in today's society." The examples provided in the study include physical impacts (e.g., car accidents caused by smartphone use) and mental impacts that create issues for smartphone-addicted children (e.g., a loss of concentration in class). The authors proposed the Smartphone Addiction Scale (SAS) to quantify this addiction. The SAS consists of 48 items relating to smartphone usage in distinct contexts, such as taking the smartphone to the toilet or feeling stressed when the smartphone is not connected to a network. Also derived from this scale is the Smartphone Addiction Scale for Adolescents (SAS-SV) (Kwon et al., 2013a), evaluated by the same authors. The SAS-SV was used by Haug et al. (2015) in a study on young people in Switzerland, which found that social networking applications were the applications most closely associated with smartphone addiction.

Smartphone addiction has also been attributed as a source of loneliness, poor bonding, and lack of integration, as shown by Bian and Leung (2015). Samaha and Hawi (2016) observed the relationships between smartphone addiction, stress, academic performance, and satisfaction with life. Through the use of multiple surveys, the SAS-SV, the Perceived Stress Scale, and the Satisfaction with Life Scale, they found addiction risk to be positively related to perceived stress. Finally, a large study by Carbonell et al. (2018) demonstrated a substantial overlap between smartphone use, Internet addiction, and social media use in a student population. Smartphone addiction also has physical effects. For instance, Akodu et al. (2018) described higher scapular dysfunction found in a population of students who are addicted to their devices.

A considerable amount of literature has been published on the influence of smartphones and has found that smartphone applications may influence users' QoL. Applications can contribute to users' well-being both positively and negatively, depending on the applications used and the user profile.

4.2.3 Smartphones as Sensors of Daily Life

Research by [Dey et al. \(2011\)](#) established that smartphones are within arm's reach of their users an average of 88% of the time. Therefore, they are a beacon of one's presence. Indeed, smartphones have been used during the COVID-19 pandemic as a proximity sensor for contact tracing ([Cencetti et al., 2020](#)). In recent years, smartphones have become a critical tool for researchers in all fields, as one of the greatest challenges to conducting a study is collecting participants' data. To solve this problem, a set of applications and software libraries have been developed to collect raw sensor data from smartphones as proxies for their users. These libraries collect similar data in different ways, although iOS devices are more restricted than Android devices.

Smartphone data can be collected from the following onboard sensors: accelerometer, location, proximity, barometer, gravity, light, magnetometer, audio, and temperature. Communication data can also be recorded from Bluetooth, SMS, telephony, and social applications. Tools such as AWARE exist to simplify the data collection process ([Ferreira et al., 2015](#)). However, AWARE is often unable to integrate with other software platforms, while other tools such as Sensus ([Lathia et al., 2013](#)) have customization issues. Meanwhile, libraries such as SensingKit ([Katevas et al., 2016](#)) cannot support data collection alone. Furthermore, other software platforms like the CARP Mobile Sensing framework ([Bardram, 2020](#)) propose a multi-platform approach (Android and iOS) with a reusable UI (Flutter) and support sensing for numerous features, but they lack low-level, hardware-based, detailed information.

Smartphone data collected with such tools have been successfully used in human studies ([Opoku Asare et al., 2019](#)). For example, [Ciman and Wac \(2018\)](#) and [De Ridder et al. \(2018\)](#) leveraged data collected from smartphone sensors to propose a stress assessment method. The first study used the data generated by finger swipes on the screen to detect stress, while the second paper showed through a meta-analysis that a tailored smartphone application can directly extract the heart rate variability (HRV), which is a stress indicator, from images of the subject's finger as it touches the smartphone's camera under illumination from the smartphone's flashlight. This process is called photoplethysmography. Smartphones are also used as sleep duration sensors, which was explored by [Ciman and Wac \(2019\)](#), and can predict users' intimacy, as claimed by [Gustarini et al. \(2016a\)](#).

In summary, smartphones are a proven source of daily-life data in multiple research domains, and their output has been validated experimentally.

4.3 Mobile Network Connectivity Study: Methods

QoL Lab was established in 2010, and since 2011, our research group has collected smartphone-based datasets for various human-based research studies and has used its own logging software for research into human activity recognition (Hausmann and Wac, 2011), mobility (Wac et al., 2015b), and intimacy (Gustarini et al., 2016a), among the other areas of study. The goal of this prior research was to quantify those aspects of human behavior with the use of smartphone sensors (i.e., gathering data using accelerometers, gyroscopes, and networking information, for example) and participants’ self-reported inputs. We now focus on human subject studies “in-the-wild” and the practical aspects of smartphone data collection (Gustarini et al., 2013) through various research topics such as Quality of Service (QoS) (Wac et al., 2006), Quality of Experience (QoE) (De Masi and Wac, 2020), and behaviors such as sleep (Ciman and Wac, 2019) or stress assessment (Berrocal et al., 2020a). Smartphone data is collected in these different studies using the same framework (mQoL-Log), and it is tailored for each study. The mQoL-Lab application (Berrocal et al., 2020b) enables background data collection through the mQoL-Log framework and implements surveys and remote notification to support human and smartphone-based research studies. Updates are necessary as the target system (Android OS) is always evolving. This section presents the tools used to acquire the data as well as their characteristics and discusses the selection of the derived features that are important for modeling individuals’ day-to-day mobile network connectivity. Furthermore, we detail the processes used for feature engineering and data filtering.

4.3.1 Data Collection Periods and Overall Summary of the Collected Data

We investigated participant connectivity with the use of mQoL-Log data records. We focused on the networking data collected through different studies conducted in Geneva over three time periods, which is presented in Table 4.1. Each participant was only present during their period. P1 was aggregated from a mQoL-Lab Living (mQoL LLab) observational study that focused on people’s smartphone usage. P2 studies were also “observational” and focused on quantifying the QoE of smartphone applications. They also focused on stress assessment via peers (PeerMA (Berrocal et al., 2020a)). The P3 study was the first “interventional study of smartphone application category recommendations made based on the QoE model”, where the intervention aimed to maximize user QoE in any context.

Table 4.1 Data Collection Periods

Period ID	Period Years	Study Focus	References	Number of Participants (N)	
				pre-filtering	post-filtering
P1	2015–2017	QoS, mQoL LLab	De Masi et al. (2016)	53	50
P2	2018–2019	QoE, PeerMA	Berrocal et al. (2020a); De Masi and Wac (2018, 2020)	63	55
P3	2020	QoE	[soon]	5	5
Total				121	110

The presented meta-study focuses on participants’ mobile connectivity throughout their days. The 121 participants collected a total of 69,761,823 samples. A sample is a piece of timestamped

network-related information that was collected automatically via the mQoL-Log either when mQoL-Log requested information (i.e. by pulling the network state every 60 seconds for P1) or when an event occurred, such as a handover between different network connection types (e.g. 4G to Wi-Fi network connection or disconnection for P2 and P3) being pushed to the logger. The different ways of collecting the networking data (push/pull) were dictated by the Google API changes over the years. A “day of the collection” is a calendar day (midnight-to-midnight) for which at least one sample exists. On average, each participant collected data for 85 days (\pm std.err 9), 21 days for the 25th percentile (Q1), 31 days for the 50th percentile (Q2), and 128 days for the 75th percentile (Q3). We observed outliers in the aggregated dataset: one participant recorded 322 days of collection (max), while another only submitted one day of collection (min). Filtering was applied to the dataset following two exclusion criteria: (i) a participant collected less than ten samples, or (ii) a participant collected less than three consecutive days of recording. The filtered dataset contained 110 participants; the filter removed 11 participants and 18,550,170 randomly distributed samples. The remaining 51,211,653 samples were retained for further analysis. Table 4.2 presents the participation statistics for the filtered datasets collected in each period. A “day of measurement” is defined as any sample collected in a 24 hour period during the collection period; this is valid for P1, P2, and P3. Contrary to a “day of the collection”, this new metric is not based on a calendar day but on the availability of samples in a 24 hour period (defined as a moving window or 24h from a previous sample).

For example, for the P1 participant who recorded 322 days of collection (max), we have defined 322 days of measurement, meaning that the time difference between any two samples was less than 24 hours and that at least one sample per calendar day (Monday, Tuesday,...) was available. On average, each participant collected data for 93.3 days of measurements (\pm std.err 9.66), 27 days for the 25th percentile (Q1), 32.5 days for the 50th percentile (Q2), and 170 days for the 75th percentile (Q3).

Table 4.2 Participation Statistics for the Filtered Dataset in Each Data Collection Period

Period ID	Avg number of measurement days/ participant	Standard error	Min	Q1 days (25%)	Q2 days (50%)	Q3 days (75%)	Max	Missing days avg \pm std.err
P1	168.6	15.4	6	79	187	270	322	59 \pm 10
P2	30.4	2.2	4	24	29	32	98	22.8 \pm 1
P3	32.6	1.6	30	31	31	32	39	0 \pm 0
Total	93.3	9.66	13.33	27	32.5	170	153	27 \pm 3

Given that a *sample* is a piece of timestamped network-related information, if $n \geq 1$ samples are generated at a specific minute (hh:mm), we classified this as *one minute of data collected*. Table 4.3 details the total number of minutes of data collected per period. We computed the mean rate of minutes acquired to understand how much data was collected per collection period overall. This rate differs from the days of data collection and the days of measurement, as it is minute based. We compared each sample acquired at a minute level to the possible number of data collection minutes during the collection period, assuming zero data loss, i.e., with data for all the minutes available. The

last column of the table shows the overall acquired minute rate over the three data collection periods. Compared to P2 and P3, as explained above, the data collected in P1 was acquired more frequently.

4.3.2 Measurement Framework: mQoL-Log

In 2011, within the context of the mQoL Living Lab, we developed the first version of a smartphone logger for the Android operating system, and we implemented a cloud-based infrastructure to collect smartphone data. The smartphone application was composed of two modules: the data logger (mQoL-Log) and the user interface (mQoL-Lab). The user interface contained the participant's communication medium to complete the study and provide the possibility to contact the study's principal investigator. A cloud-based (our university-hosted) component was able to trigger surveys remotely and control the quality of the data collected on the smartphone, for integrity purposes.

mQoL-Log collected the data from the smartphone as mentioned previously (see Section 4.3). Table 4.4 presents data collected from the smartphone's sensors through mQoL-Log. The logger included an energy policy to preserve the participant's smartphone battery life by stopping all data collection at a threshold of 30% battery capacity. Collection resumed once the smartphone was charging or when the battery capacity was above the threshold.

4.3.3 Final Dataset

As we wished to compare the connectivity of participants for the given data collection periods, we resampled the acquired P1, P2, and P3 datasets to one sample per minute, and completed the missing data points with the last known connectivity value. This method interpolates the missing points between two samples (upsampling), thus enabling a minute-based analysis of the smartphone's connectivity. The following assumption was made to validate this dimension change (i.e., to discretize it to one minute frequency): if no data is present between two samples, this means that no event occurred. With this, we propagate the last known value to the next minute until a different event-generated sample is found. However, we are fully aware that this process does not allow us to make generalizations about a representative sample of the population.

Table 4.3 Number of Minutes of Data Collected for the Three Periods

Period ID	Avg [min]	Standard error	Min	Q1 (25%)	Q2 (50%)	Q3 (75%)	Max	Mean acquired minute rate \pm std.err [%]
P1	86,148	9,387.8	855	15,769	89,008	138,676	224,684	34 \pm 2
P2	1,499	151	189	634	1,202	2,021	5,559	4 \pm 0.2
P3	1,370	286	546	1,017	1,232	1,992	2,045	3 \pm 0.5
Total	39,970	5,858	536	1,088	2,471	75,339	77,429	17 \pm 1.8

Table 4.4 Data Collected by mQoL-Log

Variable name	Definition	Study period	Trigger and frequency of collection
Screen activity	The status of the smartphone screen and the user interaction.	P1, P2, P3	Changes in screen events (on, off, user presence, rotation) (push)
Touches	Number and duration of user touches on the screen during a usage session.	P1, P2, P3	Screen event-based: each smartphone session (push)
Active app name	Application name on the user screen	P1, P2, P3	Changes in the application on-screen (push)
Background app	Application services running in the background (list)	P1	Every 60 seconds (pull)
Connectivity and network	Wi-Fi Wi-Fi level, Wi-Fi BSSID, Wi-Fi SSID, Wi-Fi interface speed, cell ID, cell operator, cell strength, cell radio access technology (RAT), cell network code,	P1	Every 60 seconds (pull)
	Internet connection status, cell bandwidth up and down stream, number of packets and bytes sent and received on wireless interfaces	P2, P3	Changes in network connection state and during user app usage (push)
Round Trip-Time [ms]	The RTT is the time needed for a ping to be sent by a smartphone to a server, plus the amount of time taken for an acknowledgment to be received.	P1 (always unige.ch server)	Every 60 seconds (pull)
	A ping is an active probing connection to a specific server via its address. A ping is executed six times; the first is discarded to remove any noise from DNS resolution time. We derived statistics (mean, stdev, and variance) from five executions.	P2 (app server)	When the app usage session starts (pull)
Battery	Battery status (e.g., charging, full, discharging), battery level, battery temperature	P1, P2, P3	Changes in battery state (push)
Physical activity	Physical activity of the user from Google Play Services activity (still, tilting:between two states, in-vehicle, on a bicycle, on foot, running).	P1, P2, P3	Changes in the user activity (push)

Table 4.5 mQoL-Log Network Data

Name	Description
Network type	Type of cellular or Wi-Fi network (RAT).
Signal strength	The signal strength is defined as the received power present in the Wi-Fi and cellular radio in dBm (RSSI). dBms were transformed into the representation used in the Android OS, i.e. bars, as the participant would see this information on-screen.
Operator	The name of the cellular network operator.
Unique identifier (ID)	Cellular network tower ID (cell ID) or Wi-Fi basic service set identifier (BSSID).
Network name	Wi-Fi network name.
Handover	Flag indicating a change in network type, cell ID, or BSSID.
Total downloaded data	Cumulative sum in bytes of downloaded data since the last smartphone reboot.
Total uploaded data	Cumulative sum in bytes of uploaded data since the last smartphone reboot.

Theoretically, P1 should have been sampled at one-minute frequency, since the pull method was leveraged for collecting the data every minute. However, we observed a skew in the pulling time, due to the Android OS giving lower priority to the collection process; the mean acquired pull rate was not 100% at 1 minute period. Following the resampling process, the P1 was hence resampled to a one-minute frequency. As for P2 and P3, the resampling process generated a time series from the discrete events collected by the push method. The total size of the resampled dataset is 234 million samples as presented in Table 4.6.

4.3.4 Features Derived from Mobile Network Connectivity

In this subsection, we describe the four features derived from the raw dataset: (i) network access technology, (ii) signal strength, (iii) data consumption, and (iv) user's physical mobility.

Network access technology or radio access technology (RAT) is defined as the physical connection system for a radio-based communication network. Smartphones support several RATs, such as Wi-Fi Bluetooth, GSM, UMTS, LTE, or 5G NR (New Radio). The focus of this analysis lies on RATs that enable Internet connection, so the Bluetooth standard is out of scope.

The signal strength is defined as the received power present in the Wi-Fi and cellular radio signal. The signal strength feature directly impacts a user's network context and provides an insight into the

Table 4.6 Average Measurement Minutes Collected Post Resampling, Per Participant in a Period

Period ID	Avg [min]	Standard error	Min [min]	Q1 (25%)	Q2 (50%)	Q3 (75%)	Max [min]	Mean acquired minute rate \pm std.err [%]
P1	241,832	22,261	7,245	113,091	267,415	388,762	462,466	100 \pm 0
P2	42,171	3,144	3,931	33,402	39,637	45,089	139,726	100 \pm 0
P3	45,122	2,473	41,448	42,447	42,930	43,902	54,885	100 \pm 0
Total	109,708	9,292	17,541	62,980	116,660	159,251	219,025	100 \pm 0

connectivity level at that moment to the current Internet provider (i.e., a cell tower or Wi-Fi access point).

Data consumption is defined as the amount of data (bytes) transferred from and to the smartphone through upload and download. The amount of data transferred during a specific time window provides information about the immediate network bandwidth. Some types of smartphone applications consume more data than others; for example, a video call application sends and receives more bytes than a text-based chat application.

The fourth feature is the user's physical mobility. Smartphones are used on the move, and their small size allows users to keep them in their pockets. In this way, they are a proxy for the user's mobility. Mobile connectivity is dependent on the physical network infrastructure around the user. Therefore, we analyzed the mobility aspect registered in the dataset. Mobility is defined as the number of cell towers and/or Wi-Fi access points with unique identifiers that a participant passes through during a specific time window.

4.3.4.1 Network Access Technology

Wireless network access technology on a smartphone consists of two Internet-enabled subtypes: Wi-Fi and cellular. Wi-Fi allows smartphones to connect to a wireless local area network (WLAN). Often these local networks are also routed to provide Internet access. A smartphone's Wi-Fi interface connects to an access point (AP) to provide an Internet connection, which has a network name and a unique identifier. In contrast to a cellular connection, Wi-Fi enables a smaller coverage range depending on the generation used (on the scale of meters rather than the kilometers of a cellular connection). For this reason, Wi-Fi is primarily used to connect to the Internet from home, work, or university. Various generations of cellular networks have been developed (e.g., 3G, 4G, 5G) with the evolution of access technology (see Table 4.7).

A cell tower offering Internet connectivity also has a unique identifier, but the main differences between cell-based technologies generation are the speed of the connection and their coverage range

from the antenna. A smartphone's baseband processor is the chip on its motherboard, which manages all radio functions. This processor is separated from the main smartphone processor for three reasons: (i) radio performance: the main processor is too slow to handle the type of work done by the baseband processor, such as encoding and modulation; (ii) legal: authorities require the software that manages radio transmission to be certified; and (iii) reliability: the OS or new application versions should not interfere with the baseband processor functions. The baseband processor is the component that manages the handover between network access technologies. When a tower is located too far from a smartphone, the signal may drop and end the user's connectivity. The baseband processor then automatically connects to a closer antenna to provide network access. If an antenna is not available in the same RAT, the baseband processor, selects a lower technology RAT, as older RAT often provide a larger range of coverage. For instance, if a 4G signal is unavailable because the user is on the move, and no other 4G link can be established, the smartphone will attempt to connect to a 3G antenna.

The type of network access technology is important because it is directly linked to the quality of connectivity. As Table 4.7 shows, an EDGE-based connection theoretically has a maximum download speed of 0.0375 Mbit/s, which is not enough to watch a YouTube video (Wamser et al., 2015). Wi-Fi technologies have also undergone several stages of evolution with different maximum download speeds, i.e., Wi-Fi type (e.g., a, b, g, n, ac). However, this information was not available during dataset collection, so information regarding Wi-Fi speed is not included in this analysis. Connection to a Wi-Fi network is not automatic, as the user must enter credentials to connect to the network. These credentials ensure the encryption of the communication between the smartphone and the wireless AP. The credentials exchange is transparent on a cellular connection, in which case the baseband processor communicates with the Security Information Management (SIM) card and the operator network to authenticate the smartphone on the network.

4.3.4.2 Signal Strength

We examined the overall network connectivity signal strength over the collection periods. Signal strength is always presented on the smartphone screen and is located in the upper right-hand corner on Android and iOS. Icons represent the signal strength sensed by the onboard antennas for both Wi-Fi and cellular networks in a human-readable format. The mQoL-Log application was able to collect that information in decibel-milliwatts (dBm). To utilize this information, we determined how the Android OS presented this data to the end-user, and mapped the dBm to the number of bars (0 to 4) shown on-screen. The signal strength represents the power present in the received radio signal.

For smartphones, this directly impacts the QoE of smartphone services such as video streaming and online games. The minimum signal strength needed to achieve a "good" experience when watching an online video on the move depends on the network access technology and the video format (e.g. HD or 4K). The signal strength plays a significant role during handovers. The baseband processors collect the signal strength continuously and choose whether to switch between RATs (i.e. conduct a vertical handover for the same RAT or a horizontal handover if RATs change) or between cell anten-

Table 4.7 Generation of Cellular Network Access Technologies

Generation	Acronym	Full name	Max download speed	Estimated download time for a 3 minute 1080p YouTube video (75 MB)
2G	GPRS	General Packet Radio Service	0.0125 Mbit/s	800 minutes
	EDGE	Enhanced Data Rates for GSM Evolution	0.0375 Mbit/s	27 minutes
3G	UMTS	Universal Mobile Telecommunications System	0.0375 Mbit/s	27 minutes
	HSPA	High Speed Packet Access	0.9 Mbit/s	11 minutes
	HSDPA	High Speed Downlink Packet Access	14 Mbit/s	1.1 minutes
	HSUPA	High Speed Uplink Packet Access	14 Mbit/s	1.1 minutes
	HSPA+	Evolved High Speed Packet Access	42 Mbit/s	13.8 seconds
4G	LTE (Cat4)	Long-Term Evolution	150 Mbit/s	1.5 seconds
5G	NR	New Radio	400 Mbit/s (sub-6Ghz) 1.8 Gbits/s (mmWave)	0.001 seconds

nas. Connectivity-wise, the smartphone user sees the signal strength as an overall health indicator of the network connection. Thus, a user may decide not to start a video call if the smartphone reports low signal strength, instead preferring to communicate via audio call only.

4.3.4.3 Data Consumption

Data consumption is a significant feature in the context of connectivity. The RAT limits the amount of data that can be transmitted, measured in seconds. Accordingly, the amount of data consumed is bound to the current network access technology. The data consumption depends on the type of services utilized by the smartphone user. Video applications consume a large amount of data (e.g. by downloading video, while a video calling application simultaneously generates and consumes a large amount of data by uploading and downloading a video). The overall data consumption also provides insight into the network traffic state. If the network encounters a large amount of traffic, this impacts the bandwidth available for use in a live video or other application by the user, and the user connectivity is affected. The amount of data downloaded and uploaded also indicates the user profile type, as some users consume less data than others. This may be due to the nature of their subscription to their operator (financial), the services used on their smartphones (behavior), and the quality of the link connecting them to the Internet (structural) over time (Fiedler et al., 2010).

4.3.4.4 User's Physical Mobility

User mobility is essential, as discussed previously. Indeed, connectivity and mobility are crucial to understanding participants' smartphone usage and connectivity changes. We explored participants' mobility per hour and the number of times each participant connected to the same tower or the same AP for multiple periods (days to weeks). A large number of unique identifiers (ID) is an indication of high mobility for a participant. One cell tower covers a few kilometers of land in a densely populated area (e.g. a 4G tower has a 16 km range), while a Wi-Fi AP covers only a few meters (e.g. a Wi-Fi ac reaches 12–35 m inside and up to 300 m outside).

4.4 Mobile Network Connectivity: Results

We analyzed results for the four features that quantify the connectivity level of an individual relying on the connection and usage of their smartphone network: the network access technology (subsection 4.4.1), its signal strength (subsection 4.4.2), overall data consumption (subsection 4.4.3), and mobility (subsection 4.4.4). For each feature, we present the overall statistics (post-filtering) of the 110 participants organized by their respective collection period.

4.4.1 Network Access Technology

Table 4.8 presents the overall average of RAT distribution per measurement period. Figures 4.1, 4.2, and 4.3 illustrate the distribution of network access technology for P1, P2, and P3 participants, re-

Table 4.8 Overall Average RAT Distribution (%) per Data Collection Period

RAT/period [%]	P1	P2	P3
Wi-Fi	52.1±3.7	51.0±3.2	47.7±7.1
LTE	29.8±4.1	28.7±2.9	49.8±0.5
HSPA+	3.1±0.8	3.3±1.2	0.8±0.5
HSUPA	0.1±0.1	0.0±0	0.0±0
HSDPA	2.2±1.2	0.0±0	0.0±0
HSPA	3.9±1.1	0.4±0.1	0.1±0.1
UMTS	3.1±1	0.7±0.3	0.3±0.2
EDGE	2.0±0.8	0.6±0.2	0.2±0.1
GPRS	0.0±0	0.0±0	0.0±0
GSM	0.0±0	0.0±0	0.0±0
UNKNOWN	2.6±0.9	0.5±0.3	0.3±2
NOCO	1.0±0.2	14.7±2.4	0.7±3
Download speed on cell network in Mbit/s Avg± std.err	6.9±3.3	6.4±3	9.4±4.6

spectively. The figures clearly show the adoption of LTE (4G). In the P1 distribution, we observe a high presence of HSPA, while the P2 distribution suggests that some participants (particularly P2S98 and P2S64) were not connected (NOCO) for the majority of the study. Overall, we see lower access to the Internet in P2 than in P1 and P3. The most recent data demonstrate the rise of LTE and Wi-Fi over the RAT. Furthermore, during P3 the participants had the most stable connection to the Internet (low NOCO), as presented in Table 4.8.

Figure 4.4 presents the overall average distribution over the three periods. We observe that LTE is more present than Wi-Fi in P3.

The data imply that overall, on average for all periods, any connection to the Internet is present 93±0.8% of the time (averaging 104,540±64,36 minutes across all periods). This information is computed from the RAT distribution. Table 4.1 presents the distribution of the connectivity and the average minutes of connection for each period and reveals that P2 connectivity is lower than that of P1 and P3.

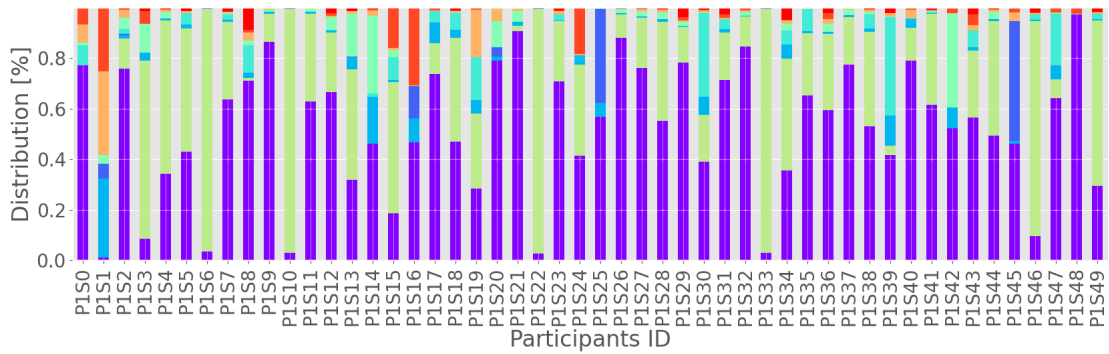


Figure 4.1 RAT distribution of participants in P1 ($N = 50$)

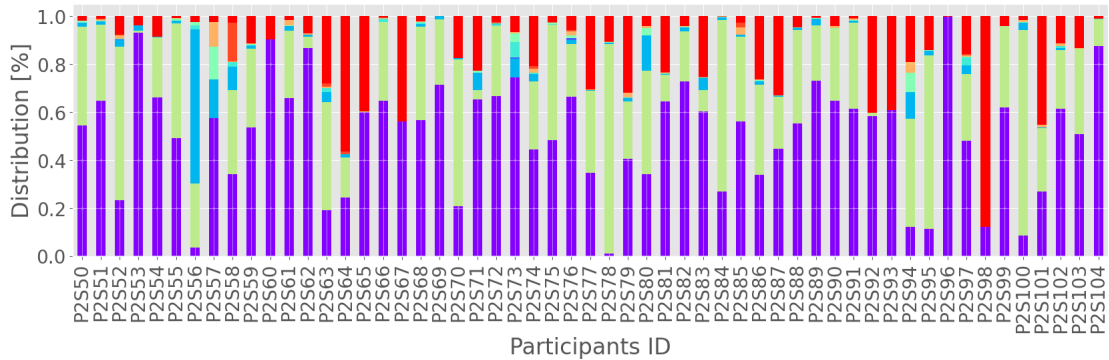


Figure 4.2 RAT distribution of participants in P2 ($N = 55$)

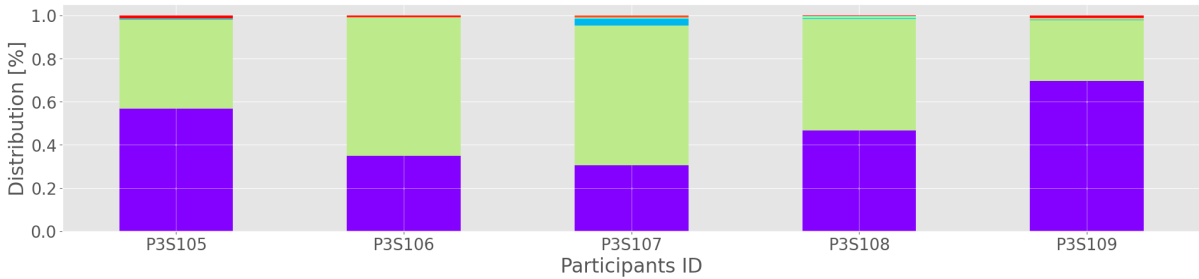


Figure 4.3 RAT distribution of participants in P3 ($N = 5$)

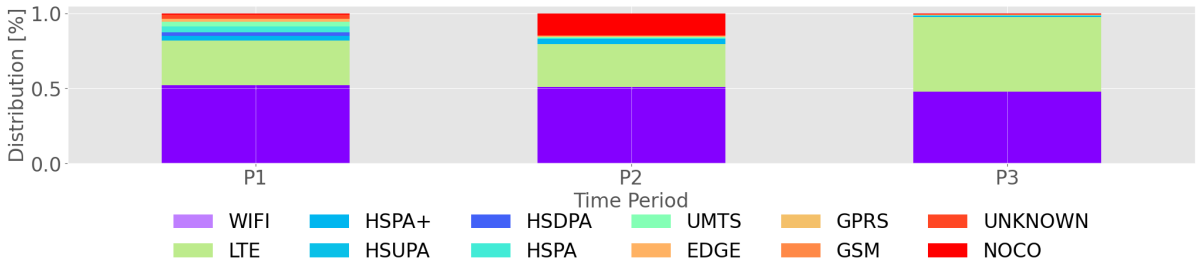


Figure 4.4 Overall average RAT distribution over P1, P2, and P3

Table 4.9 Percentage of Connectivity to Internet Distribution per Data Collection Period

Connectivity (%) per Period	P1	P2	P3
mean	0.96	0.85	0.99
std	0.06	0.18	0.00
min	0.69	0.12	0.99
25%	0.97	0.77	0.99
50%	0.99	0.92	0.99
75%	0.99	0.97	0.99
Mean in Minutes/total time/ in period \pm std.err	233,162.51 \pm 21,371	35,775.89 \pm 2,672	44,682.18 \pm 2,448

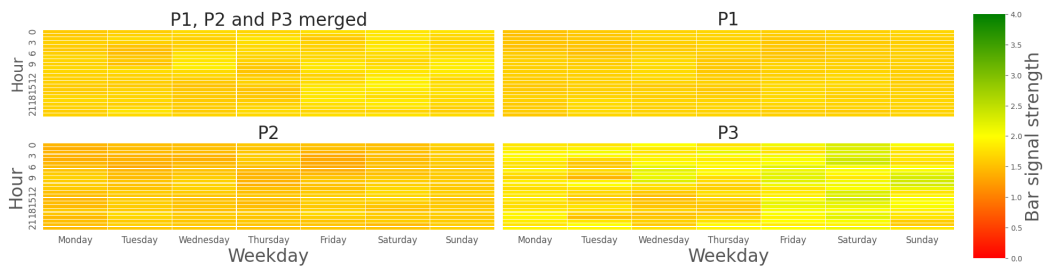


Figure 4.5 Mean Signal Strength per Data Collection Period

4.4.2 Signal Strength

The temporality of signal strength for each group is presented in Figure 4.5. Signal strength increased with time for each group. P1 and P2 feature homogeneous signal strength, in contrast to P3, which exhibits a higher signal strength at weekends and during mornings.

Figure 4.6 presents the overall signal strength distribution per period. The resampling process explains the high prevalence of the 0 bar.

Figure 4.7 presents the correlation between the signal quality and the connection type over all three periods. We note a high degree of correlation between Wi-Fi and signal strengths of 1 and 2 bars, while LTE network technology and signal strengths of 3 and 4 bars display a moderate correlation.

4.4.3 Data Consumption

During the analysis, we observed high data consumption by particular participants, as depicted in Figure 8 with the cumulative distribution function (CDF) for monthly data usage and in Figure 9 with the daily data usage for each participant (each data point on one of the lines corresponds to a

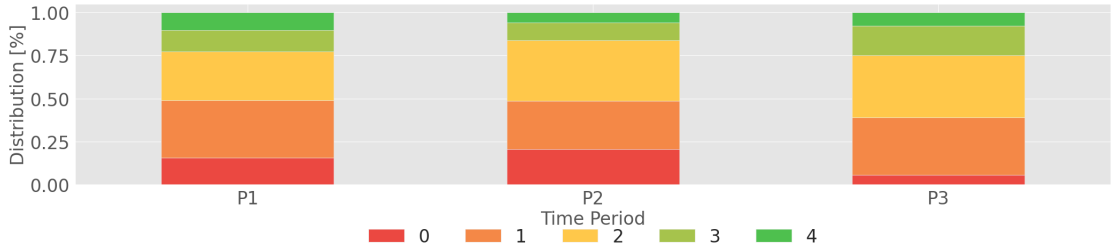


Figure 4.6 Overall Signal Strength Distribution per Data Collection Period



Figure 4.7 Pearson Correlation Between Signal Strength and Network Access Technology Type

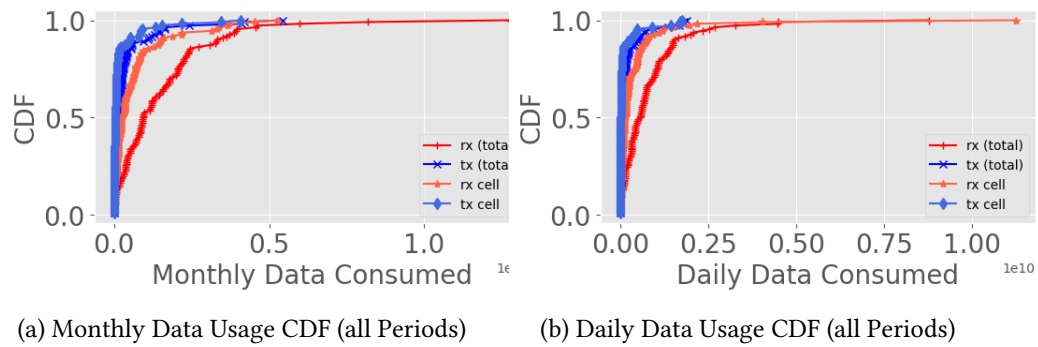


Figure 4.8 Data Consumption CDF

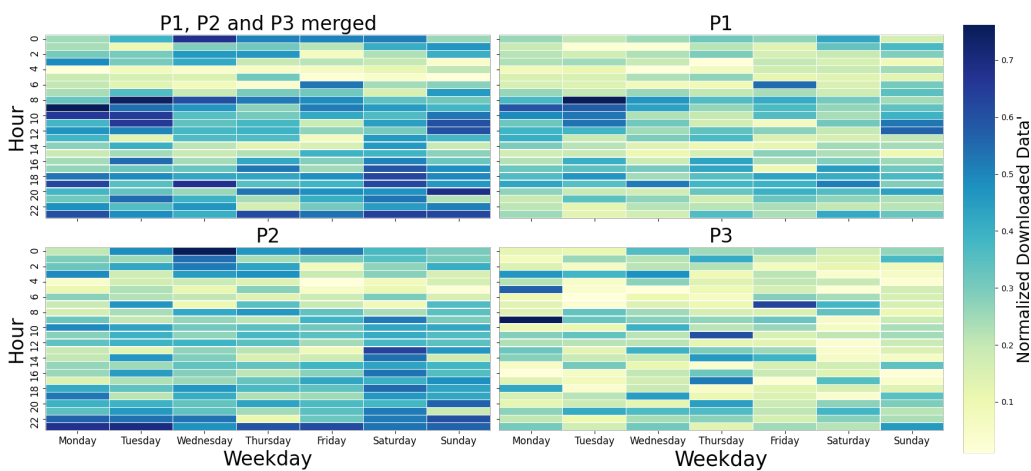


Figure 4.9 Normalized Weekly Mean Amount of Data Received per Data Collection per Period

participant). In both figures, each data point represents the average monthly data consumed by one study participant in terms of (i) rx (received, downlink) and (ii) tx (transmitted, uplink), overall and for cell-based networking. The majority of the participants display similar data-consuming behavior, regarding both data receiving and transmitting. In both temporal modalities, the amount of data transmitted from the smartphone to the cellular network is lower than the amount of data received. The monthly and daily data usage follows the same pattern (Figure 4.8a), while we observed faster consumption in the daily data usage (Figure 4.8b), in both figures each sign represents a participant.

Figure 4.9 presents the min-max-normalized weekly mean data received from all participants over the three periods. A larger amount of downloaded data can be observed during the weekends compared to the rest of the week. Participants consumed more data during mornings and evenings, and downloaded more data on weekends. We observed clusters of spikes during afternoons and evenings. The P3 participants received fewer data during the weekend. As expected, a low volume of data was received by smartphones during the night.

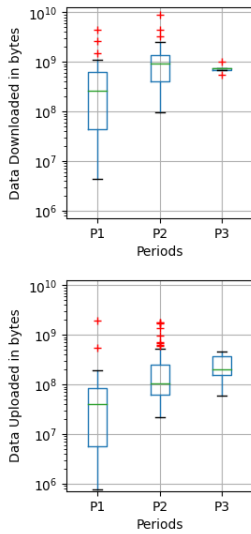


Figure 4.10 Total Downloaded and Uploaded Bytes per Period Normalized by Number of Data Collection Days

We found that participants in P2 consumed more data than the other cohorts, as shown in Figure 4.10. P3 data consumption is less sparse, likely due to the number of participants in this cohort. In all three periods, we observed some outliers that consumed more data than other participants.

4.4.4 Users' Physical Mobility

We focused the analysis on the number of individual cells and AP IDs. The full dataset contains 59,602 unique cell IDs and AP IDs combined. It is important to note that the same Wi-Fi network can be accessed via different APs, in which case the ID is different, but the network is the same. This enables roaming between the different APs in the same domains. This type of configuration is often found in large networks, for example in companies, universities, and large houses. In such cases, a Wi-Fi repeater is installed to obtain better signal quality over the entire area. The repeater has the same Wi-Fi network name as the main AP, but it has a different ID. Like smartphones, these devices reconnect to another AP when they lose a connection, such as when the user is on the move.

The vertical handover process is seamless, and the device automatically reconnects to a Wi-Fi network that shares the same name as the previous network. In this case, the device already knows the security configuration to obtain a secure connection, namely a previously established authentication. Figure 12 shows the mean cumulative cell tower and Wi-Fi ID changes per hour and per day of the week, normalized from 0 to 1. In Figure 4.11, we observe a lower number of unique IDs on Sundays for all periods. Other patterns are present; for instance, on Friday evenings participants were highly mobile, and the reverse is found during the night. P3 demonstrates lower mobility on Saturday evenings than P1 and P2. One possible explanation is the data collection time; P3 was recorded after the end of

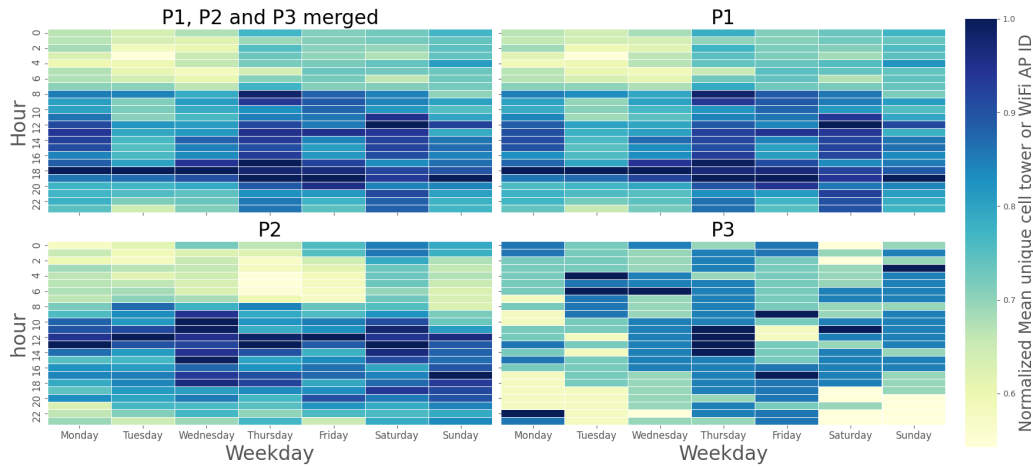


Figure 4.11 Mean Cumulative Cell Tower and Wi-Fi AP ID Changes per Data Collection per Period

the first partial-lockdown in Switzerland during the COVID-19 pandemic. At this time, participants would have been less inclined to participate in external social gatherings on two consecutive nights.

4.5 Discussion

The results confirm our hypothesis that network connectivity and consequently the mobile Internet is widely available in today's developed world. The results indicate that the participants' smartphones were connected to the Internet for 93% of their day ($\pm 0.8\%$) on average. Their devices were always either connected or searching for new network access via Wi-Fi APs and cellular towers. The quality of the connection was high overall, and we found a strong correlation between LTE and high signal strength. Furthermore, as data quantity is directly connected to the services used on the smartphone and the available network bandwidth, we observed multiple data consumption patterns that could be used to profile the users. Taken together, these findings provide important insights into the four features that impact users' connectivity and may influence an individual's decision-making and consequently their QoL. In this section, we discuss the results and their limitations before recommending other data sources for modeling environmental QoL via connected services.

4.5.1 Discussion of Overall Results

Over the data collection periods (2015–2020), the adoption of 4G (LTE) network access technology was close to complete in the Geneva area. The low presence of network access technologies other than LTE and Wi-Fi in P3 can be attributed to the continuous efforts of the mobile operators in updating network infrastructure (i.e., new antenna deployment), an update in performance of the smartphones' baseband processor (i.e., which leads to a faster handover), and the low mobility of the participants. A participant would have a higher number of connections if they were more mobile. Furthermore,

contrary to the data in P1 and P2, data from P3 was acquired during a shorter period of time from a smaller sample size.

We found a strong prevalence of Wi-Fi usage during all three periods. As Wi-Fi is commonly used at home and at work, we made the assumption that Wi-Fi usage occurs mostly indoors, where participants are located. Furthermore, while Wi-Fi connection costs are not linked to the amount of data consumed, this is not the case for cell-based connections. Wi-Fi is generally provided by a broadband Internet connection. As noted in subsection 4.1, some participants use Wi-Fi less than others, possibly for cost and quality reasons. The cost of broadband is high in Switzerland, and it is cheaper to obtain an unlimited 4G connection than to have both a (Wi-Fi) broadband connection at home and a 4G subscription. The broadband connection quality also plays a role; if an area has a low population density, broadband operators will not invest in high-throughput infrastructures. As a result, using a smartphone's 4G connection to provide home Internet may become convenient and financially attractive.

Our results introduce an additional reflection with respect to the cellular and Wi-Fi connectivity, and especially the handover between the two. Autonomous handover between cell-based networks and Wi-Fi has not always been possible in smartphones. However, smartphone OSs have evolved and can now automatically switch between a Wi-Fi and a cell-based connection. In fact, the switching between the two types of connection is common in everyday smartphone usage. For example, after entering a home, a smartphone will automatically connect to the home's Wi-Fi router. A smartphone will switch to cell-based connectivity if the Wi-Fi connection is of low quality. This so-called smart assist feature is totally transparent to the user and does not require interaction with the smartphone. However, this process only operates in one direction (i.e., Wi-Fi to 4G); the smartphone does not subsequently test the Wi-Fi network to attempt to revert to the Internet's connection source. Given our results, we would recommend that the smart assist feature operate both ways.

Additionally, connection and disconnection events to a cell tower are important data for a network operator. Notably, operators ultimately use the data collected from their core network, particularly the number of smartphones connected to an antenna, to generate connectivity maps and understand how to improve their services. Indeed, the services can also be improved by the network operator by enabling better connection during times of increased demand for connectivity in a given area. Conversely, the network operator could also enable a low-power mode of their system during low data consumption hours, decreasing their standby energy consumption. As shown in subsection 4.3, we found the same pattern as [Walelgne et al. \(2020\)](#): low data consumption during the night and higher consumption during the evening and early morning. These patterns reflect how people use their devices and connectivity. The observed higher throughput could originate from video consumption (leisure) or video conferencing with loved ones. This information can be used by a network operator to rent more bandwidth from its network provider, thus enabling a high-quality video conferencing experience at a specific time.

4.5.2 Study Limitations

This study has several limitations. The populations of participants in the three periods are not identical, so we are unable to comment on the evolution of the individual populations. Additionally, the two main OSs for smartphones are Android (Google) and iOS (Apple), but data was only collected from Android users in this study. As a result, information and insights about the population of iOS users is missing from this study. Additionally, the number of participants and the duration of P3 is lower than that of P1 and P2, so the generalization of the results between the cohorts is limited. We encountered another limitation during data logging due to the shortcomings of the OS de-prioritizing our logger. With the P1 dataset, we found that it was impossible to collect at least 50% of minute-based samples, even by sampling with the minute-based pulling method. Future studies shall be designed such that they address these limitations.

4.5.3 Quantified Self Movement

The Quantified Self (QS) movement brings together individuals from different backgrounds who wish to learn about themselves. The QS practitioners use tools, principles, and methods that are mostly enabled by smartphone applications and services and allow them to measure, analyze, and share their data (Bode and Kristensen, 2016). The QS tools can include medical test results or well-being-oriented connected objects (e.g., fitness trackers, smartwatches), mobile applications, and web applications. Those sources of information can also contribute to collecting a high-dimensional connectivity dataset and data to quantify individuals' behaviors, health, and QoL (Wac, 2018). Currently, the QS practitioners are mostly interested in their habits and health. They collect large amounts of data that they usually openly share on online platforms (e.g., quantifiedself.com, openhumans.org) for others to experiment with. In doing so, they expect to learn about themselves through their own analyses and through others'.

In the QS movement, smartphones are the main collection devices. For instance, diary and reminder applications are often deployed to collect one's day-to-day emotions, mental states, social interactions, and other aspects of human life currently unquantifiable via autonomous, connected devices. Those devices and applications depend on mobile network connectivity to function. However, the collection of network connectivity by the followers of the QS movement is often neglected. At the same time, smartphone data loggers that collect smartphone user habits, such as mQoL-Log, are uncommon in QS. In the future, it will be important to explore the potential use of additional data sources in the QS movement such as smartphone connectivity levels and their influence on the daily life of the individuals. Knowledge and anecdotal data (i.e., a study with one participant) obtained by QS's followers could prompt further investigation by researchers into the links between mobile network connectivity, physical health, social iterations, individuals' overall decision-making, and QoL, for example.

4.5.4 QoL Technologies

The evolution of QoL technologies (QoLT), defined as technologies that enable assessment and assurance of life quality for individuals (Wac, 2019), is deeply linked to the development of individuals' connectivity. The possibility to improve one's life with QoLT would likely involve a component of communication to the Internet (e.g., a cloud) or edge network devices. The large amounts of data produced by personal wearable health sensors and smartphones, for example, would be processed for immediate use (in emergency situations) or for later use. The degree of QoS offered by QoLT would depend on the supported mobile network connectivity level. Therefore, the four features described in this chapter are important, as they are fundamental aspects that define the individual's connectivity. Without connectivity, there may be no QoLT. To elaborate on this point, we discuss QoL aspects defined according to the WHO and connectivity-dependent services.

The domain of physical health includes many important facets, including daily living activities. Some of these activities rely on indoor connectivity being provided in the home or at work, school, or other frequent locations. The activities may require a low-latency, high-throughput network connection to operate. For instance, smartphone applications can provide medication schedule reminders and notifications to a patient and their family. Energy, fatigue, and mobility are factors that can be quantified by smartphone and wearable data, and adequate real-time personalized care services can be provided to the individual, depending on their needs. The applications can also help a population with substance dependence issues; for example, some applications can put at-risk individuals in real-time communication with medical professionals. In the case of assisted living, connectivity can enable support services like remote healthcare and, in the future, robot care. Overall, many day-to-day physical health services provided to an individual in a given context can be supported by connectivity.

The psychological health domain of QoL may be influenced positively or negatively by smartphone applications. Connectivity to services through smartphone applications can contribute to improving this domain. Services that influence this field include entertainment (e.g., watching a video), which can influence feelings, and information services (e.g., reading news on social media), which can influence thinking processes.

In the social relationships domain, services enabled through an Internet connection can range from simple text-based messaging to smartphone-based video conferencing. More generally, opportunities for social relationships provided by connected services are extensive and are evolving. These services may range from interactive entertainment services (e.g., joint use of online games, which influences feelings) and social networks (i.e., communication and exchange of information, thus influencing the quality of the relationship). The sex industry understood this potential market and created multiple devices for remote sexual interaction through the Internet, providing intimacy for long-distance couples (Liberati, 2017). In the social relationships domain, the specific challenge is to ensure sufficient mobile network connectivity for both receivers to enable content exchange with sufficient user experience during the interaction.

The features of the environmental domain of QoL may be difficult to quantify, as it contains the most facets of any QoL domain and is influenced by contextual variables that may not yet be understood. For example, opportunities for leisure or education may involve the possession of interactive entertainment (and a joint use of devices such as smart TVs, for example, thus influencing feelings), the use of social networks, or the use of online education services (e.g., services for peer communication and the exchange of information). Because of the high interactivity of these examples of online leisure and education opportunities rely on connectivity to succeed. Overall, there are many services in the users' environments that may enable a better QoL and rely on mobile connectivity to be provided. However, the challenge is that a unified, well-understood model of these services and their connectivity does not exist yet.

In conclusion, QoLT may impact all the QoL domains in beneficial and detrimental manners, all depending on the implementation of the services it supports.

4.6 Conclusion

This chapter quantifies the mobile network connectivity of individuals in the Geneva area during three data collection periods between 2015 and 2020. Our results demonstrate that connectivity is ubiquitous in the day-to-day life of the participants of this study, as they could access their online services anytime and from any location. We also observed a time-based evolution of the participants' Internet connection throughout the day. Overall, our results suggest that connectivity in the same geographic location improves over time. The explored features (signal access technology, signal strength, data consumption, and users' physical mobility) offer some insights into the participants' connectivity.

We observed a high correlation between signal strength and several network access technologies. According to our data, on average, a better signal strength is available on LTE than on Wi-Fi. Furthermore, knowing the individual data consumption patterns (amount of data received and transmitted) permits the profiling of study participants. Users who consume more data during a short period (spike) may use services that other users may not access because of their low connectivity. Additionally, we considered the amount of data received and transmitted by the smartphones at different times of the day. Although we found peaks during the evenings for P1 and P2, P3 did not exhibit this pattern. It is possible that a large amount of data consumption was taking place on other devices for a better experience during the evening (e.g., watching YouTube videos on a television screen instead of a smartphone screen). In addition, we also observed less mobility on Sundays across all periods. We compared the overall mobility in all periods and noticed a lower mobility in P3, which was possibly due to the COVID-19 situation in Switzerland at the time of the study.

We discuss the results in the context of emerging QoLT, which, embedded in personal devices including wearables and smartphones, enable the collection of health information, which may support an individual's progress towards better health behaviors and, consequently, a better QoL. Overall, an increase in the use of QoLT may contribute to a better life. The range of services provided by QoLT rely

on network connectivity, so future research work is needed to ensure that this connectivity matches the requirements of the technologies anywhere the user may be at any time.

Acknowledgments

This work was supported by AAL Guardian (AAL-2019-6-120-CP, 2019-2022), SNSF MIQModel (157003, 2015-2019), H2020 WellCo (769765, 2018-2020), AGE-INT (2021-2024), QoL@hip2neck, and QoL@GVA.

Chapter 5

Article III: Forecasting Smartphone Application Chains: an App-Rank Based Approach

Published in the Proceedings of the “22nd International Conference on Mobile and Ubiquitous Multimedia” (MUM 2023), December 2023. doi: 10.1145/3626705.3627802.

Chapter Contents

- 5.1 Introduction 68**
- 5.2 Related Work 71**
 - 5.2.1 Profiling Smartphone Application Users 71
 - 5.2.2 Forecasting Application Usage 71
- 5.3 Method and Implementation 73**
 - 5.3.1 Model Requirements 73
 - 5.3.2 Forecasting Model Overview 73
 - 5.3.3 Datasets 74
 - 5.3.4 Data Wrangling 74
 - 5.3.5 Modeling and Evaluation 79
 - 5.3.6 Statistical Analysis 81
- 5.4 Results 82**
 - 5.4.1 Model Performances 82
 - 5.4.2 Application Usage Habits 83
- 5.5 Discussion 83**
 - 5.5.1 Previous research 83
 - 5.5.2 Models’ Performances 84
 - 5.5.3 Ranking Importance 84
 - 5.5.4 Implication for Quality of Experience and Digital Wellbeing 85
- 5.6 Limitations and Conclusions 86**

Abstract

Research indicates that smartphone users reuse the same applications throughout the day. This study aimed to forecast a list of probable applications to be launched on a smartphone based on prior usage patterns, without the use of contextual information. We proposed a ranked-based algorithm that considers the sequential behavior of application usage history and presents usage sessions as “application chains”. We evaluated the algorithm using datasets from 397 users, comprising 433,663 application chains with a minimum of three applications and 174 applications for the longest chains, averaging 40.91 ± 18.76 applications per chains, recorded over varying time periods and across multiple countries. Our results indicate that the proposed algorithm outperforms alternative approaches, achieving a significantly higher F1 score of $62 \pm 6\%$ without the use of contextual information. The ability to predict application launch can enable the provision of additional services such as digital wellness and improved application’s Quality of Experience.

5.1 Introduction

A study reports that a vast majority (92%) of individuals in Switzerland own at least one smartphone and use it daily (97%) (Deboitte, 2018). As applications are increasingly integral to daily decision-making, understanding usage patterns and habits have become important. We employ the habit definition by Oulasvirta et al. (2012) which refers to the repetitive inspection of dynamic content on a smartphone device through an application. Smartphone applications can both positively and negatively affect an individual’s life, by either preempting needs or causing addiction (Haug et al., 2015). Forecasting application usage could lead to a better understanding of habit formation and inform the development of systems to enhance or reduce these habits.

Previous research has shown that smartphones are frequently accessed in various contexts and participate in everyday routines (Dey et al., 2011). However, the origins of these routines have been only partially explained and modeled on a limited scale, with fewer than 50 participants and study durations of two weeks (Lukoff et al., 2018; Tran et al., 2019). It is known that applications are used in specific patterns and that users revisit the same applications depending on context (Jones et al., 2015). AURs generated with each launch of an application are used to profile smartphone users (Zhao et al., 2019a) and to derive behavioral markers (Qin et al., 2014). Such as the application-use routines in the morning and evening. AURs are of interest in interaction research, particularly in the context of forecasting and profiling. A recent extensive review from Li et al. (2022) highlights emerging technologies and key trends in smartphone application usage behavior, impacting academia and industry.

Much of the current literature on the usage of smartphone applications focus on profiling. The extensive survey by Zhao et al. (2019a) presents an overview of the use of information from smartphone applications for user profiling. The authors found that AURs are particularly useful in profiling five attributes: demography, personality traits, psychological status, personal interests, and lifestyle. Profiling can detect a class, it is unable to predict whether a habit based on a smartphone application

is recurrent. The application sequence order and its repetitiveness are indicators of habit (Jones et al., 2015).

This knowledge could lead to recommendations and interventions in digital wellbeing (Vanden Abeele, 2021). Digital wellbeing focuses on one's interaction with smartphones. Particularly, the negative impact such as reduced mental performance, higher stress, compulsive behavior, unhealthy sleep pattern, and lower cognitive capacity (Zimmermann, 2021) due to smartphone usage. our method can be applied in the context of a personalized digital wellbeing assistant that utilizes application sequence forecasting to optimize the user's app usage and promote healthier habits. The assistant can provide timely reminders and suggestions based on the predicted sequence of applications, helping individuals manage their time and attention more effectively. By analyzing the user's app usage patterns and leveraging the forecasting capability, the assistant can identify potential areas of improvement and offer tailored recommendations for reducing excessive screen time, encouraging breaks, or promoting usage of specific apps that promote relaxation, mindfulness, or productivity.

Furthermore, applications forecasting can also be utilized in parental control applications to foster a healthier digital environment for children. By understanding the likely sequence of applications that children may engage with, parents can set appropriate usage limits, designate focused study periods, and ensure a balanced and age-appropriate application selection. This promotes responsible and mindful use of digital devices while safeguarding the well-being and development of young users. Understanding and forecasting probable smartphone use is the first step in digital interventions (van Velthoven et al., 2018).

As well, forecasting those habits can address solutions in the Quality of Experience (QoE)(Le Callet et al., 2012) domain for smartphone applications (De Masi and Wac, 2020), for instance by recommending a list of the next application launches to use in a specific context (Baeza-Yates et al., 2015), thus enhancing the user's experience (i.e., by preemptively catching content and pre-processing data). Overall, forecasting application launch could have a direct impact on one's wellbeing and smartphone's QoE.

Nevertheless, few works have used a methodical approach to capture and model smartphone usage. Previous studies that forecast with AURs have not been replicated due to the nature of the datasets collected for the studies (Li et al., 2022). The methods employed to process the data (i.e., aggregation and filtering) and build the models, moreover, are often superficially described and unavailable as a digital appendix (e.g., raw data and processing scripts). Hence, a replicable method for working with AURs is needed.



Figure 5.1 Application Chain Example

Attempts have been made in this context to employ previously defined analysis approaches from other computer domains. [Jones et al. \(2015\)](#) define a revisitation chain as the chain of applications used across different sessions by a smartphone user. Their definition is based on the revisitation analysis in the context of web browsing ([Tossell et al., 2012](#)). Expanding on their definition, we define an *application chain* as the sequence of applications used during a screen session. A screen session begins when the smartphone screen is on and ends when the screen is off. The example application chain presented in [Figure 5.1](#) depicts WhatsApp, Spotify, and Facebook Messenger usage. The launch frequency is important for observing revisitation patterns (i.e., revisiting an application already present in the chain).

In this work, we propose a rank-based algorithm that utilizes an analysis of smartphone application usage habits. This algorithm is based on similarities between consecutive application sessions, and it is implemented in a model to forecast the probable launch of applications in the next session, using historical usage data without the use of contextual information. Additionally, we present a method for building and evaluating forecasting models based on AURs. While several models have been proposed to predict the next application launch based on the previous one ([Baeza-Yates et al., 2015](#); [Liao et al., 2013](#); [Lu et al., 2014](#); [Shin et al., 2012](#)), none of them focus on forecasting a full list of applications within a chain based on ranking. Unlike the previous methods in the application forecasting domain, which can only predict a class or a value (i.e., regression), the rank-based algorithm we propose can forecast the relevance ranking of the set of applications. Additionally, these cited works employed sensitive information such as location as input for their models, which may negatively impact the privacy of the smartphone user.

In this paper, we present three contributions. First, we investigate the relationship between consecutive application chains and whether they contain indicators of subsequent chains. We hypothesize that previous chains can predict future ones, and consecutive chains could be useful for forecasting and ranking applications in future usage sessions. Second, we propose a method for building a forecasting model based on habit-forming patterns found in AURs ([Oulasvirta et al., 2012](#)). We also provide an open-source codebase to facilitate comparisons with other works that share our research goals. Third, we implement our algorithm in a model to forecast the next application chain and test its performance against three existing models. We use user-dependent models and evaluate them using two independent datasets collected from Android smartphone users in different countries. Our results indicate that ranking applications based on usage history is more accurate than the existing models. The paper is organized as follows: [Section 5.2](#) provides an overview of related research on forecasting with AURs, and [Section 5.3](#) presents our proposed method and describes the implementation of our algorithm for application chain forecasting. [Section 5.4](#) presents the performance of our models, and [Section 5.5](#) discusses the challenges. Finally, [Section 5.6](#) presents the limitations and concludes the paper.

predicting a high probability of WhatsApp usage between 6 and 9 pm (Do and Gatica-Perez, 2014; Liao et al., 2013; Xia et al., 2020b; Yu et al., 2018, 2020). The final goals are divergent and focused only on the screen events to predict the state of the next screen. Both Shin et al. (2012) and Baeza-Yates et al. (2015) aimed to provide smartphone end-users with a list of probable applications to open based on their context via a smart launcher application. These works implemented their models in real-world applications with some success, as demonstrated by high user adoption. However, the data collection methods differed enormously between the works. The datasets employed by Xia et al. (2020b); Yu et al. (2018, 2020); Zhao et al. (2019b) were collected by an Internet service provider (ISP). The network traffic traces collected contained data only from Internet-enabled applications; this collection method did not include applications that do not generate traffic (e.g., camera application). Hence, the resulting models are particularly problematic as they do not reflect the interactions and in-situ smartphone application usage generated by the smartphone user. The ISP-based studies fail to acknowledge the shortcoming of such a method. However, other studies exist in which the authors collected their datasets directly from smartphones by developing their data loggers (Baeza-Yates et al., 2015; Jones et al., 2015; Kostakos et al., 2016; Liao et al., 2013; Natarajan et al., 2013; Shen et al., 2019; Shin et al., 2012; Xu et al., 2013). This crowdsourced data collection method was unobtrusive for the studies' participants and conveyed signification information about habit formation by not interfering with the users' interactions or devices. On the contrary, Stanik et al. (2020) used Amazon Mechanical Turk, which is an online crowdsourcing platform, to collect their dataset. The study participants had to install an application that collected data directly on their smartphones and annotate their usage when the researchers' application required it. One major drawback of this approach is that the annotation process can influence the participants, changing the way they use their smartphones and modifying their routines. Although a subset of the works (Do and Gatica-Perez, 2014; Huang et al., 2012; Lu et al., 2014; Moreira et al., 2020; Xiang et al., 2017; Zou et al., 2013) used a popular publicly available dataset (i.e., the Mobile Data Challenge dataset by Laurila et al. (2013)), collected for over a year in-the-wild, the methodology for pre-processing the dataset is not described in any of the studies. Finally, Roffarello and De Russis (2021, 2022) proposed a method to obtain an explanation for the habits of smartphone application usage (Roffarello and De Russis, 2019). The method is based on bagging, clustering, and association rule to extract habits from smartphone collected data (i.e., context and smartphone application usage). The rules take the form of a *if-then* statement: "if connected to Wi-Fi and at home, then WhatsApp".

In summary, previous studies have focused on the contextual, temporal, and sequential nature of application usage habits, particularly on historical application and previous application usage to forecast the next application launch (within a session). In addition, the number of participants, application launches, model performance metrics, and data collection duration varies between the studies. Hence, the results are difficult to compare. We propose a pipeline method to standardize the approach and facilitate replication and comparison in forecasting smartphone application chains. Our method focuses on forecasting the full next application chains, providing a more comprehensive prediction be-

yond just a ranked list of probable next applications to launch. This distinction highlights the unique contribution of our approach.

5.3 Method and Implementation

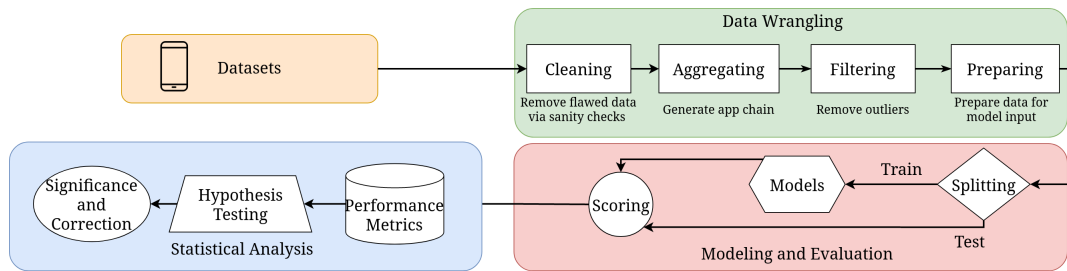


Figure 5.2 Forecasting Model Pipeline

5.3.1 Model Requirements

The proposed method should facilitate the construction of a forecasting model that predicts application ranked chains based on past application usage. In this model, the rank signifies the relative probability of a user accessing a specific application in a session, with a higher rank indicating a greater likelihood of interaction with the application. The model should be created using human-smartphone interaction data collected in-situ, providing a timestamped dataset (time series) that reflects application usage behavior. Validation of the model should be carried out through a time series split for cross-validation, allowing an examination of model performance in accordance with the timelines of application usage. It's also crucial to evaluate the significance of these results. Lastly, the method should encourage replicability by providing comprehensive instructions and sharing code and tools.

5.3.2 Forecasting Model Overview

Figure 5.2 visually represents our method, broken down into four primary components. (i) Datasets: This pertains to the identification and selection of relevant datasets required to verify our method and hypotheses. (ii) Data Wrangling: This encompasses the steps taken to prepare the dataset from the raw data, involving tasks such as cleaning (sanity checking of data), aggregation (assembling the application chains), filtering (removing outliers), and preparation (which may include feature extraction and data organization for model intake). (iii) Modeling and Evaluation: This stage involves defining the model and developing a cross-validation strategy to assess the model's performance. (iv) Statistical Analysis: This involves establishing the significance of our results based on hypothesis testing and the evaluation of performance metrics derived from model testing.

5.3.3 Datasets

The choice of App Usage Record (AUR) datasets is contingent on the model's ultimate objective. Depending on the requirements, a dataset can either be collected or an existing open dataset could be utilized. An extensive exploration of open data repositories, such as [Crowdad.org](https://www.crowd.com/), is a prerequisite before collecting AUR in real-world conditions. The model's specific needs may restrict the selection of an appropriate open dataset. In cases where there's no public data suitable for the task, the development of a study protocol and selection or creation of a data logger becomes necessary ([Bardram, 2020](#); [Ferreira et al., 2015](#); [Kumar et al., 2021](#); [Li et al., 2022](#)).

Our focus was on readily available datasets, collected in real-world scenarios using smartphone loggers. We chose not to include ISP-based datasets as they emphasize application network behavior rather than user-device interaction. We used two datasets: the Mobile Quality of Life (mQoL) dataset ([Berrocal et al., 2020a](#); [De Masi and Wac, 2020](#)) and the Mobile Phone User (MPU) dataset from Telefonica ([Pielot et al., 2017](#)). Although the Carat dataset ([Oliner et al., 2013](#)) offers valuable insights into mobile device energy diagnosis, its lack of clear information regarding application user session sequences, a vital aspect of our work, led us to exclude it from our study.

Both datasets were collected on Android smartphones, with the participants' consent. The mQoL dataset was based on two real-world studies conducted in 2018, assessing smartphone users' Quality of Experience and the Peer-ceived Momentary Assessment method (PeerMA, , N=55, [Berrocal et al. \(2020a\)](#)). The MPU dataset, collected from 342 participants, was used by [Pielot et al. \(2017\)](#) to predict opportune moments to engage smartphone users. Both datasets provide similar information, including application session data and user-smartphone screen interactions. The mQoL dataset covers an average of 33.89 ± 14.79 days per participant, while the MPU dataset covers an average of 25.10 ± 6.93 days per participant, with an overall average of 29.28 ± 10.88 days of participation across both datasets. On average, participants contributed data for around 3.81 ± 1.31 weeks (ranging from a minimum of 1.3 weeks to a maximum of 13 weeks).

5.3.4 Data Wrangling

5.3.4.1 Cleaning

It's imperative to perform integrity checks on the AUR data to ensure its quality and reliability, as issues are common in data collected in real-world settings, as highlighted by [Ferreira et al. \(2015\)](#); [Gustarini et al. \(2013\)](#). Due to potential glitches and limitations in the data collection software, the resulting dataset could be flawed, making data cleansing a crucial preliminary step before any data transformations or aggregations occur. For instance, scrutinizing the range of key input parameters (e.g., timestamps) can provide valuable insights into the data quality. Any data points falling outside of the minimum and maximum range can be discarded. The start and end dates of the data collection period usually serve as the minimum and maximum values for timestamps. Application usage collection loggers, such as Aware ([Ferreira et al., 2015](#)) and mQoL-Log ([Berrocal et al., 2020b](#)), are

designed with a battery threshold to conserve battery life and maintain a positive user experience. When the battery level hits this threshold, the logger enters a sleep mode, potentially causing gaps in the recorded events and resulting in incomplete application interactions in the data, which need to be eliminated to prevent any skewing of study results.

Given the assumption that the datasets may contain artifacts due to limitations in the logger application, we implemented filters to remove all AURs with inaccurate timestamps (i.e., those falling outside the study period) and erroneous application names (e.g., “NULL” values).

5.3.4.2 Aggregation

The next step in data wrangling involves aggregating individual AURs into an application chain (a many-to-one mapping). Past practices involved merging application launches occurring within a 7-minute interval (Zhao et al., 2019b), but this approach doesn’t accurately reflect application usage behaviors and is merely based on the overall dataset distribution. Jones et al. (2015) implemented a 30-second window for merging repeated application usages to account for notification-induced behavior changes. An appearing notification can alter the application chain as users may launch a new application via the notification, a common scenario for communication applications. However, the arbitrary choice of this window duration could introduce bias into the analysis.

The selection of a suitable aggregation operation depends on the model’s needs, the AUR’s origin, and its characteristics. The operation applies to one dimension only, either (i) time, where AURs are grouped based on a literature-derived time window, or (ii) a specific dataset feature that demarcates the start and end points of the aggregation (e.g., the screen turning on and off). Once the application launches are processed through the chosen window, a reduction operation must be applied to discard unnecessary applications and simplify the application chain. The rationale for identifying unwanted applications must be supported by literature and fit the specific task at hand.

States	Triggers
On → Off	A notification turns on the screen with no user interaction.
Present → Off	The user unlocks the screen and interacts with it (until shutoff or timeout).
Off → On	Time between subsequent screen interactions (i.e., no interaction).
On → Present	A notification turns on the screen and user interacts (e.g., unlocking).

Table 5.2 Smartphone Screen State Combinations

Our aggregation methodology drew inspiration from Kostakos et al. (2016), who demonstrated modeling of a smartphone’s screen state using Markov chains, considering user input and potential screen state combinations. Table 5.2 enumerates these critical combinations. The screen state was invariably logged using an application. Our attention was particularly on the On → Off and Present → Off sequences, which offered the beginning and ending timestamps for each chain. Following that, we linked the applications utilized in the interim between these two events with AURs. AURs were

inclusive of events occurring between On \rightarrow Off and Present \rightarrow Off triggers. The outcome of this step was a record of applications the user activated from the time their screen was on until it was turned off.

We then executed a reduction operation, condensing multiple launches of the same application within a chain into a single launch, tagged with the corresponding start and end timestamps. System or background applications that do not appear on the screen, such as the keyboard, were removed. This helped to address instances where a user opens an application, like WhatsApp, types a message, sends it, and the logger records a sequence like: WhatsApp, Keyboard, WhatsApp, Facebook Messenger. Removing the keyboard launch from this sequence leaves two identical application launches which are in fact the same application from the user’s perspective. Consequently, these two launches were merged into one, resulting in a chain: WhatsApp, Facebook Messenger.

5.3.4.3 Filtering

Filtering involves selecting a subset of the aggregated application chain for the purpose of modeling and forecasting. AURs may include elements such as software keyboards, launchers, installers, permission control managers, settings, User-Interface (UI) system processes, and data loggers. However, these elements need to be filtered out, following the precedent in this type of analysis (Zou et al., 2013), as they do not denote a user launching an application to fulfil a specific need. Consequently, filter parameters should be chosen based on established literature, for instance, the minimum interaction time with a smartphone application that signifies a new user need. Detailed statistics regarding the remaining data and its distribution must also be provided.

Given our focus on sequential patterns, we restricted our analysis to users who had logged activity for over 10 days, in line with Jones et al. (2015). This threshold is generally applicable for analyzing habitual smartphone application usage. We proceeded to filter out “micro-usage” instances of applications (i.e., application usage of less than 3 seconds (Ferreira et al., 2014)) to prevent them from affecting the analysis. Additionally, we computed a z-score for the duration of each participant’s application chain, calculated from the timestamp of the initial application in the chain through to the end of the final application. We retained the chains falling within the $z > 3$ and $z < -3$ range, thereby preserving 99.9% of the cumulative percentage. Lastly, we eliminated nonconsecutive chains—those without a sequential identification number—due to the filters applied earlier.

Dataset	Pre-filtering		Post-filtering	
	#Participants	#Chains	#Participants	#Chains
MPU	342	368,273	334	272,404
mQoL	55	65,390	55	50,828
All	397	433,663	389	323,232
Avg		216,831.5		161,616

Table 5.3 Dataset States Pre and Post Filtering

Table 5.3 provides a breakdown of the data distribution before and after filtering for both datasets in unison. Eight participants from the MPU were excluded from the study due to less than 10 days of active participation, while 323,232 chains remained following the filtering process. The filtering steps implemented earlier, such as system applications and keyboard usage removal, resulted in the discarding of only 25.5% of the total collected chains. These discarded chains comprised 73.3% of a single application, owing to the micro-usage application filtering. The statistical details furnished in the remainder of the paper concentrate on the datasets post-filtering. On average, each participant had 830 ± 574 chains, ranging from a high of 4,724 chains to a low of two consecutive chains.

While there was a broad variance in smartphone usage and the lengths of unique chains per participant, the distribution of the amount of unique chains seemed to follow a certain pattern, hinting at consistent reuse of applications by users. To validate this, we performed a Mann-Whitney U Test on the distribution of the number of unique chains and the total number of chains. This test does not predicate on a specific distribution for a dataset. Its null hypothesis asserts that the distributions of two datasets are identical. There was a statistical difference between the distributions ($p < .05$, rejecting the null hypothesis). However, upon testing their distribution with a normal test, both were found to follow a normal distribution (unique chains: $p < .001$, the total number of chains: $p < .001$). Thus, we looked into the ratio between the count of unique chains and total chains to assess the variability in overall application usage habits and determine if it was correlated with the number of chains. We found the average ratio to be 0.34 ± 0.14 . We divided our dataset into two groups based on the median number of collected chains (726 chains). Those with more than 726 collected chains were classified as extensive application users, and those with 726 or fewer collected chains were classified as low application users. We applied a one-way ANOVA test to the ratio of the two groups ($p < .001$). As a result, we deduced significant differences between the two groups due to the number of collected chains and the propensity of smartphone users to repetitively use the same application in a similar pattern. Furthermore, we examined the variability in the lengths of the chains per participant; we found an average of 40.91 ± 18.76 different chain lengths across both datasets, with a maximum of 174 different chain lengths and a minimum of three. However, the length of the chains (the number of applications within a chain) reveals that on average less than 2.62 ± 1.1 applications are used in a chain (mQoL: 3.15 ± 1.63 , MPU: 2.53 ± 1.07), with a minimum of 1.04 applications, a maximum of 10.86 applications and a mode of 4.49 applications. The top 10 most frequently used applications among all participants were WhatsApp, Chrome, Contact, Facebook, Gmail, Instagram, Twitter, Phone, Photo Gallery, and Email Client, which aligns with the top 10 Android applications listed by [Jones et al. \(2015\)](#).

5.3.4.4 Preparation: Derived Features

The preparation step depends on the approach chosen to build the model. This transformation step includes feature engineering (deriving new variables from available data, e.g., amount of unique application used, the time duration between two usages of the same application ([Jones et al., 2015](#))),

enhancing the model's forecasting performance. Additionally, the data format (i.e., size, dimension, distribution) has to be compatible with the machine learning algorithms employed. Finally, an assessment of the obtained feature distribution must be done over the participant's data to provide insight.

We have focused on the chain distance as the derived feature. The results of [Martinez et al. \(2014\)](#) suggest that the study participants' ratings should be transformed into ranked representations to obtain more reliable and generalizable models. Hence, we hypothesized that an application chain could be mapped to a ranked chain without duplicate items. Duplicate applications indicate a higher frequency of usage, and they obtain a higher rank. Rankings allows for a natural way to capture and quantify the user's preferences and habits. The assumption here is that a user's interaction with different applications is not random, but instead follows a pattern that reflects their personal preferences, tasks at hand, and habitual behaviors. By ranking applications based on their usage frequency, we can assign a meaningful order to applications that abstracts away the specificities of individual sessions, but still retains the essential information about the user's preferences and habits.

Moreover, the ranked representation provides a form of data normalization and is less sensitive to variations in the absolute frequencies of application usage. Instead of focusing on the raw frequency of app usage, which can be noisy and subject to various external influences, ranking emphasizes the relative importance of different applications.

Furthermore, the ranked chain prediction model inherently takes into account the sequential nature of user interactions. It doesn't just predict which apps will be used, but also in what order they will be likely engaged with. This adds another level of depth to the predictions, providing a more realistic and useful forecast for user behavior.

In addition, by transforming data into ranked representations, we inherently introduce a level of noise reduction, since the fine details of app usage (exact timestamps, duration of use, etc.) are not taken into account in the ranking. This can help the model focus on the most salient patterns in the data and avoid overfitting to the training set. Each use of an application is a deliberate interaction made by an individual over the fixed application set available on their device, comparable to the selection of a rating on a scale. Accordingly, the rank order is based on the popularity of the application in the current session. However, it is impossible to compare ranked chains of different lengths ([Li et al., 2019](#)). The precondition is that all possible items are ranked. In our context, this condition would require a user to open all their applications one by one during one session, which action does not represent a common behavior. To mitigate this problem, we implemented a random algorithm to fill the missing ranks based on the application set available (i.e., the total number of applications) per participant, a common method to fill an incomplete ranked list ([Marlin and Zemel, 2009](#)). Because of the time series nature of the data, other methods for filling (e.g., frequency-based) are impossible. These methods leak data from the future (e.g., application choice) in the current chain. Therefore, the model forecasts X applications in the chains, with X corresponding to the amount of applications a user used during their longest chain in the past. For this analysis, we wanted to observe the difference, also named distance, between two consecutive chains. We, therefore, applied the normalized Kendall Tau distance to count the number of pairwise disagreements between two chains. A distance close to

zero indicates a low disagreement, and hence a high similarity between chains. Due to the nature of the random filler, we repeated the filling and the computation of the distance 10 times. We observed a normal distribution of the average distance for all participants ($N=389$), validated with a normal test ($p < .001$). Our analysis found a high similarity between consecutive chains for all participants ($\tau_{mean} = 0.06 \pm 0.03$, $\tau_{min} = 0.02$, $\tau_{q25} = 0.04$, $\tau_{q50} = 0.06$, $\tau_{q75} = 0.08$, $\tau_{max} = 0.39$). However, we found a $\tau = 0$ for only 12.6 ± 13.6 % of the chains. Hence, the majority of consecutive chains are not identical. This result contradicts the assumption stated in Section 5.1 that the majority of concurrent chains are identical.

5.3.5 Modeling and Evaluation

5.3.5.1 Model Selection

The model selection is built upon previous work in the domain. The model must be compatible with the desired end goal. For example, a regression model is unable to forecast a list of applications. Moreover, the literature should provide candid models to test against allowing model comparison. For example, one standard candid model for forecasting the next application is forward filling by repeating the last known application (Shin, 2012). Finally, the features' selection must be based on the output data from the data wrangling block.

The chain forecasting task is closely related to how an individual uses their smartphone. The ranked chains contain information about application habits, and can potentially make it possible to forecast the next chain by only using the participant's history without relying on sensitive context information such as location. Unlike past research, which has focused on either a single model for all participants or individual models trained using contextual information, our approach takes a different approach by focusing on one participant's application habits at a time. Therefore, our models are trained on the data of one participant only, resulting in individual models. While a general model addresses the cold start problem and limits overfitting, it also requires fine-tuning to correspond to a specific user's habits. Our proposed algorithm model is based on the ranked chain, which can predict the application rank within a chain based on the previous chain. However, the length of the chain may fluctuate. To address this limitation, chains are mapped to ranked chains and padded for comparison and input to the model. We used a tree-boosted algorithm (XGBoost (Chen and Guestrin, 2016)) with multiple output classifiers to predict the rank of each application (XBGRank). This algorithm was selected due to its high performance in learning-to-rank tasks (Lucchese et al., 2020). Each classifier was trained to predict the rank of a specific application within a chain, and the previously ranked chain was used as feature. We also assessed the distribution of ranked chains and adjusted the model to limit overfitting with the most common ranked chains in the participant's dataset.

Specific attributes of the XBGRank model are presented below:

- Input: The input of the model is the previous application usage session. In other words, the algorithm takes as input the ranked list of applications that the user interacted with in the previous session.

- **Feature:** The feature is the ranked list of the previous application session.
- **Machine Learning Problem:** The machine learning problem is formulated as a classification task. Given the previous application usage session, the goal is to classify which applications are more likely to be started in the next session based on their past in-session usage frequency. The algorithm learns from historical patterns to predict the most relevant applications for the subsequent session.
- **Output:** The output of the algorithm is a ranked list of applications sorted by their predicted usage frequency in the predictive chain. The algorithm assigns a score to each application, indicating the likelihood of it being started in the next session, prioritizing applications that are expected to be more frequently used.
- **Granularity and Sequential Nature:** The granularity of the problem is at the session level. The algorithm predicts the applications that are likely to be started in the next session based on the usage patterns in previous sessions. It operates on a sequential basis, considering the historical sequence of application usage sessions to make predictions for the subsequent session.

From the literature on identifying the next application launch (Figure 5.1), we identified three candid models to compare against our model. The most-frequently-used (MFU) application and the most-recently-used (MRU) application models (Shin, 2012) were transposed to forecast the application chain. The MFU model always predicts the most frequent chain from the training dataset. The MRU model returned the last application chain present in the training dataset. The only input feature used to train these models was the application chains. Furthermore, we also framed our forecasting task as a sequence-to-sequence (Seq2Seq) task. Indeed, the Seq2Seq models have partly resolved sequence prediction tasks, as shown by Sutskever et al. (2014). For instance, they are popular in natural language processing tasks, especially for text translation, image annotation, conversation modeling (e.g., chatbots), and text summarization. Seq2Seq models use Long Short-Term Memory (LSTM). We implemented a Seq2Seq model using the classical encoder and decoder architecture employed for text translation tasks. Each application chain was tokenized to be fed to an encoder LSTM gate connected to a decoder gate. The decoder outputs a possible application chain. This vanilla LSTM network is used for comparison, without specific hyperparameter optimization for this particular problem.

5.3.5.2 Evaluation Metrics

In the literature, it is often difficult to compare forecasting models for application usage due to the various metrics employed. Table 5.1 notes the use of four different metrics to determine the performance of application forecasting models. In this paper, we propose contextualizing these metrics for the specific task of application chain forecasting: (i) Accuracy: the proportion of correct ranked applications within the full forecasted chain; (ii) Precision: the proportion of correct ranked applications within the full forecasted chain, over the total number of ranked applications, both correct and

incorrect; (iii) Recall: the proportion of correct ranked applications within the full forecasted chain, over the total number of relevant ranked applications; (iv) Root-mean-square error: a measurement of the differences between predicted and observed values which does not apply to this forecasting approach.

Precision and recall are commonly used to evaluate the performance of forecasting models built with AURs using cross-validation, as they are widely reported in the literature. Cross-validation helps to avoid overfitting by withholding a portion of the data as a testing set to validate the model while using the remaining data to train it. The validation score is computed using a metric function such as F1. The method of splitting and ordering the training and testing sets is crucial to the evaluation of the model and must be compatible with the forecasting goal. When working with time series data like AURs, it is important to consider the order of the data, as a shuffling strategy for cross-validation would disrupt the continuity of time and lead to data leaks. An approach for ordered data is to use a time series split for cross-validation, which protects from data leaks and allows for observation of the model's performance over time (Bergmeir and Benítez, 2012).

We adopted the F1 score as the performance metric for our model, as it is commonly used in the literature on application forecasting and offers a more balanced assessment than using precision or recall alone. To evaluate our model's ability to forecast ranked application chains, we trained and tested it using time series data from both datasets. To avoid issues of data leakage and overfitting, we employed a time series split, which is a variation of K-fold cross-validation. Specifically, we trained the model on data collected for one week and tested it on the following week's data. We repeated this process until the testing set contained less than one day of data. This approach allows us to observe the model's performance over time and to capture weekly patterns in application usage habits. Additionally, it should be noted that a month of data collected is sufficient to explore patterns in application usage habits, as highlighted in previous studies (Jones et al., 2015). The final K-fold split comprised an average of $80 \pm 5\%$ training data and $20 \pm 5\%$ testing data. Additionally, it should be noted that validation data was not provided manually to train the model. For the algorithms which use validation data (i.e., NN and XBG), the validation set is 20% of the training set, which is the default value of the program used to implement the models (i.e., Keras and XGBoost).

5.3.6 Statistical Analysis

The validation of model performance should be done via statistical analysis. Statistical analysis is a process in which trends, patterns, and relationships are investigated using quantitative data. The goal is to assess the superiority of the results from multiple models trained on the same data in different ways. Statistical significance tests are designed to address this problem, and they quantify the likelihood of observed metrics (i.e., F1) given the assumption that they were drawn from the same distribution. If this assumption is rejected, it suggests that the difference in scores is statistically significant; however, the formulation of the assumption (null hypothesis) is error-prone. The appropriate statistical tests are applied to test the null hypothesis against the alternative hypothesis, which is that

the performances are not random. The test depends on the distribution of the data (normality) and its attributes. If the test is correct and the p-value is lower than a significance level (often selected at $\alpha = .05$), the null hypothesis is rejected and the alternative hypothesis is accepted. This means that the models' performances are different. However, a post-hoc Bonferroni (Armstrong, 2014) correction has to be applied for solving the multiple comparisons' problem on the obtained p-values.

We selected the one-way ANOVA test as the statistical significance test to validate our assumption that one model performs better than the other. To mitigate the issue from multiple hypothesis testing (i.e., reduce the probability of getting Type I error), we selected the Bonferroni correction (Armstrong, 2014) as a p-value corrections method.

5.4 Results

5.4.1 Model Performances

Model	MFU			MRU			Seq2Seq			XBGRank		
Dataset	MPU	mQoL	ALL	MPU	mQoL	ALL	MPU	mQoL	ALL	MPU	mQoL	ALL
Mean	0.01	0.01	0.01	0.01	0.00	0.00	0.25	0.18	0.23	0.61	0.58	0.62
Std	0.01	0.01	0.01	0.02	0.01	0.01	0.16	0.12	0.08	0.14	0.13	0.06
Minimum	0.00	0.00		0.00	0.00		0.00	0.00		0.30	0.28	
Maximum	0.28	0.10		0.23	0.03		0.95	0.55		0.98	0.77	

Table 5.4 F1 Score Performances Over the Both Datasets

We aggregated the results of each cross-validation step for each participant and calculated a mean score per model. Overall, our ranking-based model performed better than the candid models. Table 5.4 presents the aggregated F1 values per model and dataset. Then, we applied a one-way ANOVA test on the performance of each model per participant, to investigate the significance of our findings. Overall, after a Bonferroni correction, we found $p < 0.01$ for all participants. Consequently, the ranking approach performed statistically better than the candid models. We next observed the models' performances over time separately for each participant. Each model was trained $K - 1$ times with data from a cumulative period, based on our selected cross-validation method. K corresponded to the number of weeks of data available for each participant. Across both datasets, 183 participants (of 389) collected data over more than four weeks. Hence, their results highly influenced the mean F1 score for $K > 4$. The models were trained with different $K = \{0, 1, 2, 3, \dots, 11\}$ values, representing each of the test sets. At $K = 11$, the participant with the most data recorded weeks (13 weeks), the training set contains the last 12 weeks of data ($K = 0, \dots, 11$), and the test set contains the 12th week. Overall, the ranking method demonstrated a higher performance for all folds (Figure 5.3).

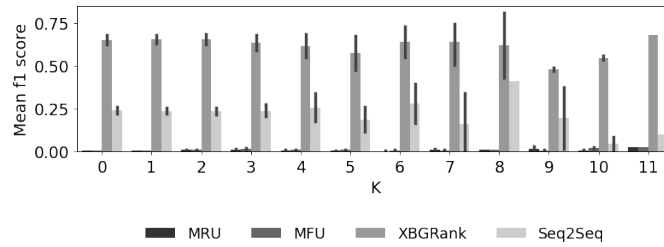


Figure 5.3 Mean F1 Score Over Number of Weeks (K)

5.4.2 Application Usage Habits

In this study, we aimed to understand the predictability of participants' application usage habits using the F1 score as an indicator of overall predictability. Building on previous research, which has focused on descriptive statistics and association rules based on contextual factors and the probability of launching a group of applications (Jones et al., 2015; Roffarello and De Russis, 2021), we investigated whether the predictability scores per participant are correlated with the average time spent on their smartphones. This time-based habit has been previously employed as a feature to predict the next application launched (Moreira et al., 2020). To test the statistical significance between the predictability score and the average smartphone usage time per day, we used a one-way ANOVA test. Since the dataset includes participants with varying usage profiles, we used the daily average instead of the average duration across the entire dataset. Results show that there is a statistically significant correlation between these two variables, with a $p < .001$ after a Bonferroni correction for all participants. This suggests that the habit of application usage duration can influence forecasting performance.

5.5 Discussion

5.5.1 Previous research

There is limited existing work proposing a defined pipeline (Roffarello and De Russis, 2021) and using ranking-based methods to forecast application launches or application chains. Thus, we can only compare our model performance with the decision-tree-based methods. The main difference between these two methods is how the participant's smartphone usage habits are encoded as the model input. The ranking approach uses fewer features than previously used by Lu et al. (2014). However, the performance metrics reported in the works presented in Table 5.1 make it difficult to compare our approach against the state of the art. Only two studies reported their precision and recall results, which are used to compute F1 scores. Moreover, the previous work only focuses on the next application to be launched. The next application may not be enough in a long-term forecasting context (e.g., scheduling intervention) as the trigger application chain is longer than average. Contrary to the task of forecasting the application ranked chains, the goals are different. Additionally, the models presented in Section 5.2 were often trained on a dataset collected by network operators, the effects of the

data collection method were not considered by the authors as a limitation. Furthermore, the models with the best performance were trained on a dataset collected over multiple months and with more participants than the mQoL and MPU studies. Finally, the methods employed to build the models are insufficiently described. On the other hand, the method presented in this paper offers a path to build application forecasting models.

5.5.2 Models' Performances

The simplistic MFU and MRU models obtained the lowest performance on both datasets from the different models implemented, followed by the Seq2Seq model (without hyperparameter tuning) and our ranking-based model XBGRank. Compared to the literature review, model for forecasting the next application launch, the ranking approach ($F1 : 62 \pm 6\%$) performed better than the decision tree-based model ($F1 : 36\%$). The habits of smartphone use encoded through the ranking methods lead to better performance among all participants. The effectiveness of our ranking-based method is also verified on both independent datasets. Moreover, the ranking approach for application chains consumes fewer resources than the deep-learning method, which often requires high-energy-consuming hardware to be trained. In addition, the comparison with models trained with ISP-based application usage is also hazardous due to the nature of the original data. The performances of ISP-based models are the results of the network activity, in contrast to the performance of the XBGRank model built with real application usage data. Also, we found that the dataset origin (mQoL or MPU) does not significantly impact XBGRank performance. As such, this ranking approach can be generalized and applied to other application usage datasets.

5.5.3 Ranking Importance

In this analysis, we investigated the similarity between consecutive application chains as a potential forecasting feature. Our rank-based approach applied to two independent datasets showed a low Kendall-Tau distance between consecutive chains for a given participant, indicating a high correlation. These findings provide evidence that ranked application usage patterns can be a highly predictive feature for forecasting application chains. Our approach is a simplification as it does not assume or predict the length of the chains, a property that is inherent to the ranked list approach. We observed that on average, the length of chains is less than 2.62 ± 1.1 applications, but they do fluctuate. In contrast, past studies have been able to only forecast the use of more than one application. With our approach, all the applications in a chain can be forecast with one inference, regardless of its length. Our study also found that $96.23 \pm 2.03\%$ of 389 participants primarily used communication and social apps, consistent with previous research (Böhmer et al., 2011). This highlights that smartphones continue to primarily serve communication purposes, with different apps within the same category providing various types of mediums, such as text messaging and video conferencing.

Although the random ranks may introduce some noise, we believe it is a reasonable trade-off considering the benefits it brings to the experimentation process. By including all applications in the

ranked list, even if they were not used, we ensure that the algorithm has the opportunity to learn from the entire application set and capture potential user preferences that may emerge over time.

5.5.4 Implication for Quality of Experience and Digital Wellbeing

Previous research has integrated forecasting models into “smart” application launchers for Android to aid in application selection, but this information was not leveraged to enhance the QoE of specific applications. If we successfully forecast the entire application chains, it could allow for system optimizations such as better battery management, processor operation scheduling, and network utilization based on smartphone usage patterns, thereby positively impacting smartphone users’ QoE. Our method could contribute to these optimizations by fostering a more in-depth understanding of application chains, potentially facilitating dynamic resource allocation, battery conservation, and smoother smartphone performance. If it can anticipate the next application chain, it could enable context-aware application management. This includes preloading necessary resources, optimizing application startup times, and enhancing responsiveness, potentially improving user experience. The method could also bolster personalized content delivery by predicting the full application chain. If successful, content providers could prefetch and optimize the delivery of relevant content, resulting in a seamless, personalized user experience. Furthermore, it could enhance application recommendations and discovery. By analyzing usage patterns, our method could provide tailored application recommendations that align with the user’s interests and preferences of the moment, thus refining application discovery systems. This amalgamation of functionalities illustrates the potential comprehensive benefit our method could deliver, spanning from system performance to user experience.

Existing digital wellbeing applications primarily focus on reducing certain application usage by monitoring the time spent using an application. Studies have shown that such interventions can be effective in reducing meaningless application usage ([Roffarello and De Russis, 2021](#)) (20 participants, 36.80 ± 20.59 days on average). Our method could enhance personalized digital wellbeing if it accurately forecasts the user’s sequence of application interactions. This would enable proactive recommendations for healthier screen time management and could potentially improve overall digital wellness. As such, knowing which applications will be launched through the use of a chain-based model, could trigger an intervention and help users avoid starting harmful application usage patterns such as continuous checking of social networks or excessive gaming which leads to mental health problems ([Samaha and Hawi, 2016](#)) and negatively impact social interactions.

Regarding the overhead of such system on-device, we propose a computationally efficient method, suitable for on-device implementation, that strategically utilizes idle, charging, and night-time periods for training and updating the smartphone usage forecasting model. These periods are typically marked by reduced user activity and increased device resource availability. Leveraging these times allows for optimal allocation of computational resources, reducing the impact on device performance and user experience. It allows for more intensive operations during charging periods while conserving battery life, and it utilizes night-time when devices are less likely to be in use. This balanced

approach helps maintain system optimization and QoE improvements, while being cognizant of resource consumption, battery life, and user experience.

5.6 Limitations and Conclusions

One limitation of this study is its confined focus on Android users. These results should be replicated on the iOS platform. Moreover, this work focused on forecasting application usage through a ranking approach. We assumed that the participants' behavior in selecting one application over another derives from the intent they want to accomplish within the application. This assumption may have limited the performance of the models because notifications (Mehrotra et al., 2016) also cause application engagement. As well, not hyper parameter tuning was done on the LSTM network. The focus of our research was primarily on introducing and evaluating our proposed method for application sequence forecasting. While hyperparameter optimization is an important aspect of optimizing deep learning models, given the scope and goals of our study, we did not extensively explore hyperparameter tuning for the LSTM network. Another limitation, which is standard for wild studies, is that the software logger may have failed to collect all relevant data due to an operating system update or other factors. Thus, influencing the results. Additionally, the sensitive information needed to create a more accurate model (i.e. better performance) could be used via federated learning or local training to create the model. However, both approaches come with technical difficulties (i.e., common hardware for training) and optimization issues.

In this paper, we found that consecutive application chains are significantly relevant to each other. Smartphone users often launch the same applications consecutively. We presented and implemented an algorithm, via our replicable method to build a forecasting model based on habit-forming patterns, motivated by previous research on smartphone application forecasting. Our paper extended this concept to application chains by employing ranking on application frequency. In the end, we build an application chain forecasting model which performs significantly better than the candid approaches previously employed in the application forecasting domain. Further investigation is needed to understand whether application and notification content can influence application behavior, thus mitigating its effect and building more accurate forecasting models.

Data and Code Availability

The pre-processing scripts are available on GitLab¹ to foster replicability of the pipeline. The MPU dataset is available on the site of the authors.

¹Link to come

Acknowledgement

SNSF MIQmodel (157003) (2015-2019) & AAL CoME (2014-7-127). The University of Geneva ethics committee approved those studies under *CUREG.201803.02* and *CUREG,201909.12*.

Chapter 6

Article IV: You're using this app for what ? A mQoL Living Lab Study

Published in Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, New York, NY, USA, Oct. 2018. doi: 10.1145/3267305.3267544.

Chapter Contents

6.1 Introduction	90
6.2 Methodology	92
6.2.1 Recruitment	95
6.3 Results	96
6.3.1 Use of Selected Seven Apps	96
6.3.2 What you Say vs. What You Do: Real Use of Mobile Apps	96
6.3.3 EMA Context Survey	97
6.3.4 Limitations	97
6.4 Discussion	97
6.5 Conclusive Remarks and Future Work Areas	98

Abstract

Smartphones are personal ubiquitous devices that provide an immense source of information via diverse applications (apps) that contribute to our decision-making process throughout the day and improve our quality of life in the long term. In the past, an app only had one or a few specific functions, while nowadays, given the same interface, an app provides multiple interactive services to their users. However, we still have a weak understanding of user expectations and experiences with these apps. Towards this end, we extended our previous smartphone logging app, to the new mQoL Lab for **m**obile **Q**uality **o**f **L**ife, to strategically trigger user surveys and to achieve a better understanding of the user's actions in popular Android apps, like: Spotify, WhatsApp, Instagram, Maps, Chrome, Facebook and its Messenger. We present and discuss the results and their implications acquired during our first pilot study conducted with five users for four weeks in our Living Lab settings.

6.1 Introduction

In the last 15 years, mobile phones became “smart” and widely available, and its usage evolved into a crucial skill in supporting our day-to-day needs for information and communication “on the go”. There exists a copious amount of different applications (apps) helping us in these needs, especially via a particular app developed for a specific goal, e.g. video-chat, reading the news or playing a game. The smartphone becomes an extension of our body in our daily life, being at least 50% of the time within our arms reach (Dey et al., 2011). It is highly probable that a smartphone (and wearables connected to it) is becoming a tool that allows us to improve our everyday activities and health, considerably contributing to our Quality of Life (QoL).

We started researching the app experience in 2010 indicating the major factors influencing the popular apps (Ickin et al., 2012), through challenges in human subject studies “in-the-wild” (Gustarini et al., 2013) and connectivity patterns of smartphone users (Wac et al., 2015c), to living-lab approach to data collection in mobile studies (De Masi et al., 2016), explicitly focusing on mobile users in trains or other moving vehicles. In our research, we have leveraged hybrid research methods - with different sources (the participant and his/her smartphone), different data granularity and timeliness of the data acquired from the participants.

Our smartphone-based logger app, previously named mQoL-Log (Wac et al., 2015b) had several limitations, e.g. it was challenging to configure, was predefined for a specific study goal and did not include context-based triggers for user surveys launched on a participant smartphone. The current, updated version, mQoL Lab, has a better study management system. It is able to run multiple human subject studies at once and engage the participants via the same interface (Figure 6.1).

For this study, we selected a set of the most popular interactive apps (Table 6.1) and asked the user to rate their experience after using these apps. Our primary goal was to test mQoL Lab on a small scale study. Our secondary goal was to understand and model the end-user's Quality of Experience (QoE).

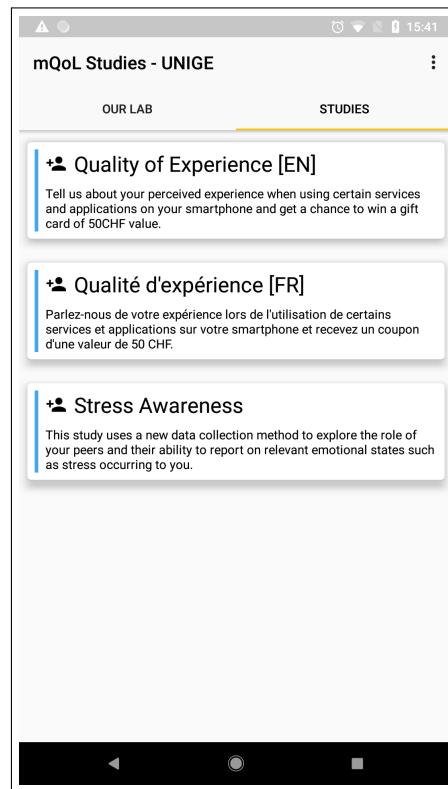


Figure 6.1 Study List in mQoL Lab

Le Callet et al. (2012) define QoE as *the degree of delight or annoyance of the user of an app or service*. QoE findings are just indicated and not discussed in details in this paper due to space limitation.

The currently available and open source research platform for human subject studies, AWARE framework (Ferreira et al., 2015), does not propose fine-grained tuning of context-based event triggers, specifically user survey after a specific app utilisation. Ickin et al. (2012) used AWARE for their QoE study with this limitation. The survey's trigger was random during the day without a specific set of apps to rate for the user to provide their QoE rating on. Casas et al. (2015a) developed an app for collecting user experience in the field, where they ask their participants to execute a specific task on a

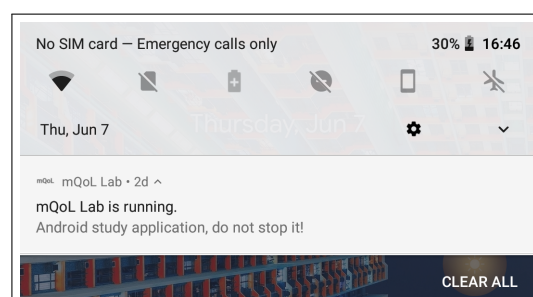


Figure 6.2 Running Study Notification

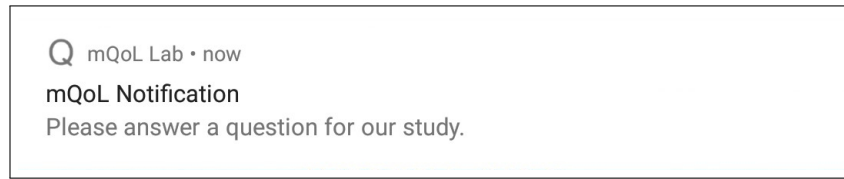


Figure 6.3 New EMA Notification

set of app after which (user's intent), a preprogrammed user survey appears. Most of their datasets are composed of ratings of Youtube video visualisation. We decided to discard the video apps for our work as the QoE community is heavily focused on this subject (Ickin et al., 2015; Juluri et al., 2015; Ketykó et al., 2010; Nam et al., 2016). De Moor et al. (2010) proposed a detailed yet theoretical framework for evaluating QoE in a living lab setting in 2010. Since this framework was created, the mobile operating system landscape changed; hence it can not be entirely used today. Their solution needs profound system information not publicly available via the Android API, without installing a custom Android OS on the participant's device. Finally, Tossell et al. (2015) collected smartphone-based sensing data to explore smartphone addiction, yet their solution does not make use of context-based event triggers, as we do.

6.2 Methodology

In this study, we used a qualitative and quantitative hybrid method. The study entry survey (demographics information request and app habits) and user surveys deployed via Ecological Momentary Assessment (EMA) Stone and Shiffman (1994) are qualitative. EMAs allows the users to report their momentary experience with their smartphone, as the intent, they are trying to satisfy. The mQoL Lab app integrates two components, the EMA-survey manager and the mQoL Lab data logger.

Research Methodology

Qualitative: (a) Entry surveys enabling us to understand the individual’s socio-demographics and current experience with the smartphone/apps (b) Ecological Momentary Assessment (EMA) (Stone and Shiffman, 1994) enabling us to understand the momentary, just in context attitude, needs and behaviours of study participants using their smartphone/apps.

Quantitative: Raw and analysed datasets obtained from smartphone built-in sensors via the mQoL Lab app.

Hybrid: The simultaneous applications of qualitative and quantitative methods, allowing for better accuracy of datasets collected towards our end goal.

We have limited the scope of this study to seven popular Android apps (Spotify, WhatsApp, Instagram, Google Maps, Chrome, Facebook and Messenger (Table 6.1)) due to the constraint of the maximum number of EMAs to be triggered per day to minimise the participant’s burden. The study has been approved by the University of Geneva’s ethics commission. Overall, our study focused on expectations and the resulting QoE of mobile smartphone users. In this paper we only present results for one question that users had to reply during each EMA: *What action were you trying to accomplish?*. The user can reply to categorise his/her intent along seven labels: “CONSUME content”, “SHARE or create content”, “READ text message”, “WRITE text message”, “CONTROL an app (start/stop music)”, “VIDEO call” or “AUDIO call”. The same seven labels are always presented for any apps; it is not dynamic. It allows us to discard false replies, e.g. using Maps to make a video call.

The remaining EMA questions and possible responses are as follows:

- Did your usage of app name at use start time go as expected? Yes/No/I’m not sure.
- How was your last usage session of app name at use start time? Slider (1 to 5) with a Mean Opinion Score (Union (1993), MOS scale).
- Did your last app name at use start time meet your expectations? Slider (1 to 5).

Table 6.1 Application Metadata from Google Play Store

	Category
WhatsApp	Communication
Chrome	Communication
Messenger	Communication
Facebook	Social
Instagram	Social
Spotify	Music and Audio
Maps	Travel and Local

- If something went wrong, please tell us more about it. Free text entry.

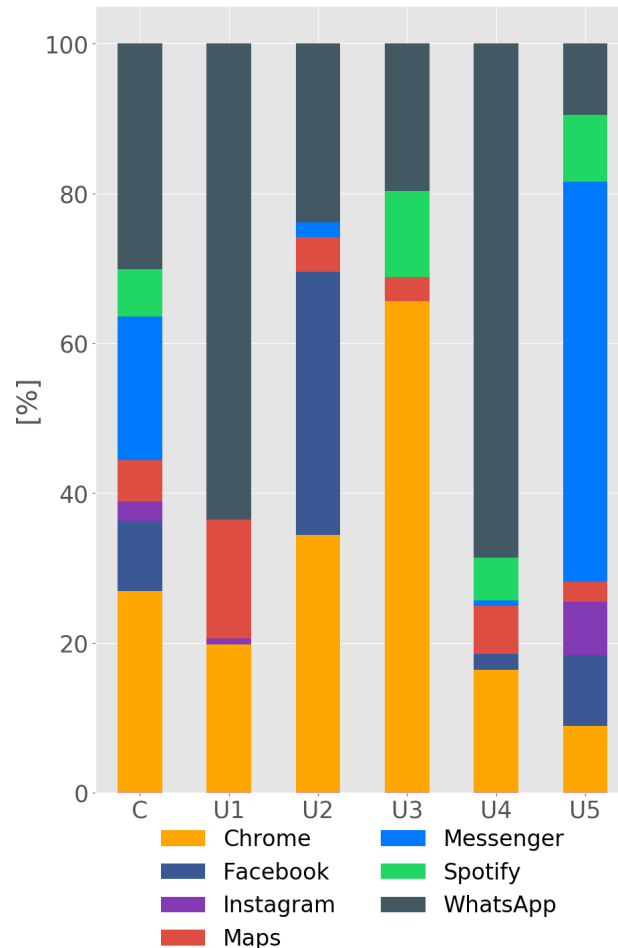


Figure 6.4 Distribution of Selected 7 Apps Used per User

Once the study starts, a notification (Figure 6.2) is always present on the screen remembering the user that data are being collected. An EMA notification is triggered after the use of a specific app as in Figure 6.3, give access to a survey containing 5 questions. To further limit the user annoyance of our notifications, we set up a policy, that the mQoL Lab only trigger 12 surveys per day between the hours of 7AM to 9PM, and the minimal time between two consecutive surveys is set to 20 minutes.

Additionally to EMA, the mQoL Lab logger collects quantitative data as follows, enabling the gathering of the following information:

- Application: package name of app on the user's screen.
- Activity: user physical activity from the Google Play Services (still, tilting: between two states, in vehicle, on bicycle, on foot, running).

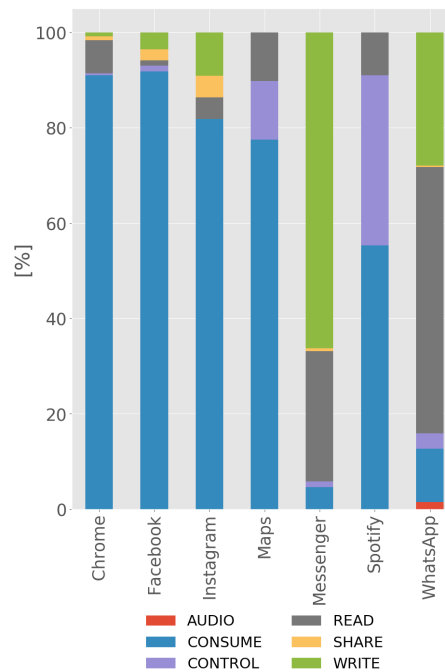


Figure 6.5 What action were you trying to accomplish ?

- Network: signal strength, basic service set identifier and service set ID (network name) of Wifi Access Point, routing tables, IP address, domain name server, ping to app server, packet and kilobytes send and received per an interface.
- Cell: signal strength, cell ID, operator name, network code and network time of the connected and neighbour cells.
- Touch: number and distribution of user touch while interacting with the screen, per a user session
- Battery: level, temperature, health and status of the battery.

6.2.1 Recruitment

Adult participants were recruited inside the University of Geneva (Switzerland) for the duration of the study (28 days). They needed to be active users of the Android OS smartphone and users of the set of apps which we have focused on in this study; they have self-reported the top 5 apps used at the study entry time.

6.3 Results

6.3.1 Use of Selected Seven Apps

We have collected data for five users (one female, aged 26-35, avg 32 y.o., all employed, four with MSc, one with a PhD), denoted as U1 to U5 for over 28 days. The average participation was 26 days, with an average of seven EMAs filled per user per day. We collected 40,320 minutes of cumulative data. They spend an average of 1503.6 minutes in all app session combined; that corresponds to around five hours of total app time per each participant (11 minutes/day).

From the time stamp of the app usage, we aggregate all the different mQoL Lab data sources. The cumulative distribution of the selected seven apps usage is available in Figure 6.4 (C) as the distribution of apps per each user (U). One-third of our dataset is composed of WhatsApp, followed by Chrome and Messenger. The main activity during an app used is “still”. Participants, when mobile, used their app “on foot”, followed by “in vehicle” (tram, bus, train or car). The other activity as “tilting” and “on bicycle” represents a small part of our samples, presented in Figure 6.6.

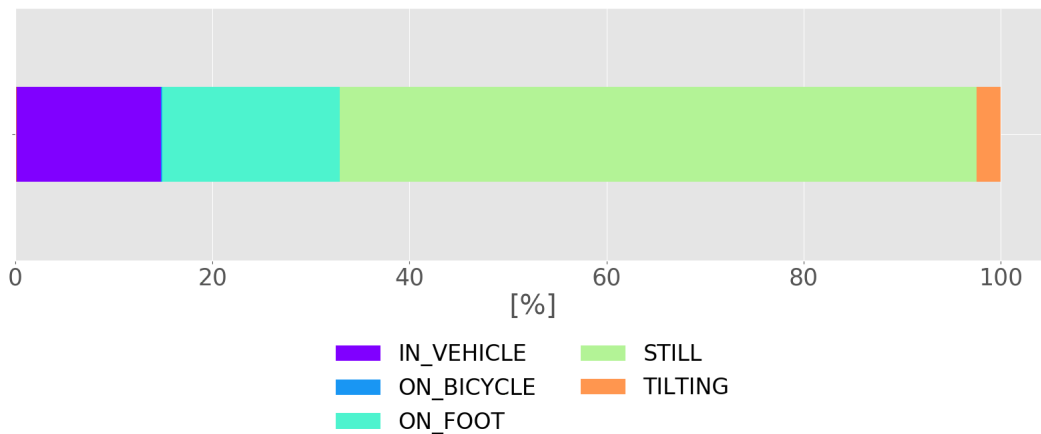


Figure 6.6 What action were you trying to accomplish in the application ?

From Figure 6.5, we observe the cumulative actions per selected 7 apps executed by the users. Chrome, Facebook, Instagram and Maps are more used to consume content than to share it. Messenger and WhatsApp, from the category Communication (Table 6.1), are essentially used for reading and writing messages. Spotify user's action is shared between, consuming music and controlling the application. We conclude from it that the apps are mostly used, as they were designed from the beginning and new functionalities (like video call in WhatsApp or Messenger) are not likely to be used.

6.3.2 What you Say vs. What You Do: Real Use of Mobile Apps

From the overall application usage data in Figure 6.7, we observe that for U1, only WhatsApp is part of its top 5 apps, its top 10 integrate Maps and Chrome. U2's top 5 include WhatsApp, Facebook and

Chrome, its top 10 contains Maps and Messenger. U3's top 5 incorporate Chrome and WhatsApp, its top 10 include Spotify. WhatsApp is the only app of our set in U4's top 5, and its top 10 include Chrome as well. U5's top 5 contains Messenger and Spotify, and its top 10 include Instagram, Facebook and WhatsApp. If we compare this data with the responses from their entry survey, where they listed their top 5 apps, only U5 use all the apps that trigger an EMA. U1, U2 and U3 listed 3 of our apps set and U4 just one.

6.3.3 EMA Context Survey

The mean time spent on apps is $17 (\pm 70)[s]$. We now focus on short app session (10 minutes or less, representing 89% of all sessions) and present in Figure 6.8 the time spent in each app by users. Chrome is the app where most time is consumed by users (22 ± 102)[m]. Text base conversations (WhatsApp and Messenger) are quite fast (4 ± 30 [m], 6 ± 45 [m] respectively), as choosing a song to play in Spotify (4 ± 11 [m]). The average time spent to reply to the EMA is $17.8 (\pm 3.9)[s]$. Depending on the app, a user will spend more or less time to reply to our EMA, as presented in Table 6.2. We observe a high correlation (>0.8) between the expectation and experience MOS ratings.

Table 6.2 Time spend to reply to EMA in each app [s]

	Mean	Std	Median[s]
WhatsApp	4.3	119.1	10.64
Instagram	13.1	8.1	9.38
Messenger	14.2	17.3	10.1
Chrome	14.7	12.3	11.07
Facebook	16	20.05	10.04
Maps	21.3	56.2	12.11
Spotify	23	70.9	10.49

6.3.4 Limitations

The small sample of participants is the main issue and does not allow for a total validation of the representative value of our results. We plan to open this study to a large number of people in coming weeks.

6.4 Discussion

The average number of surveys filled (7) vs maximum trigger possible (12) can be explained by two user habits. Messaging app, e.g. WhatsApp and Messenger, enables to reply to messages directly to the incoming message via built-in notification. The mQoL Lab is not able (yet) to detect this specific use case. The second habit is explained by the user app usage. Users are launching the apps in sessions, while our policy disables a notification if the previous one has been used less than 20 minutes before.

6.5 Conclusive Remarks and Future Work Areas

This study on user's actions showed the feasibility of mQoL Lab to gather interesting data relative to our goals. mQoL Lab methodological strength lies in the event-based triggering of EMA. It reduces the memory bias by firing a EMA just after the app use of interest. This mechanism may be exploited further in other studies to trigger EMA after the change of context-of interest - for example, disabled/enabled network interface or Bluetooth, changing from indoor to outdoor, change in lighting condition. Such instrumentation will enable to closer monitor physical changes in the individual's behaviour and environment that, supported by EMA based self-reports, may help to better understand the mental, physical or state of individual. we are planning to leverage the mQoL lab in such use cases in the near future.

Acknowledgements

SNSF (157003) (2015-2019) and *AAL CoME* (2014-7-127)

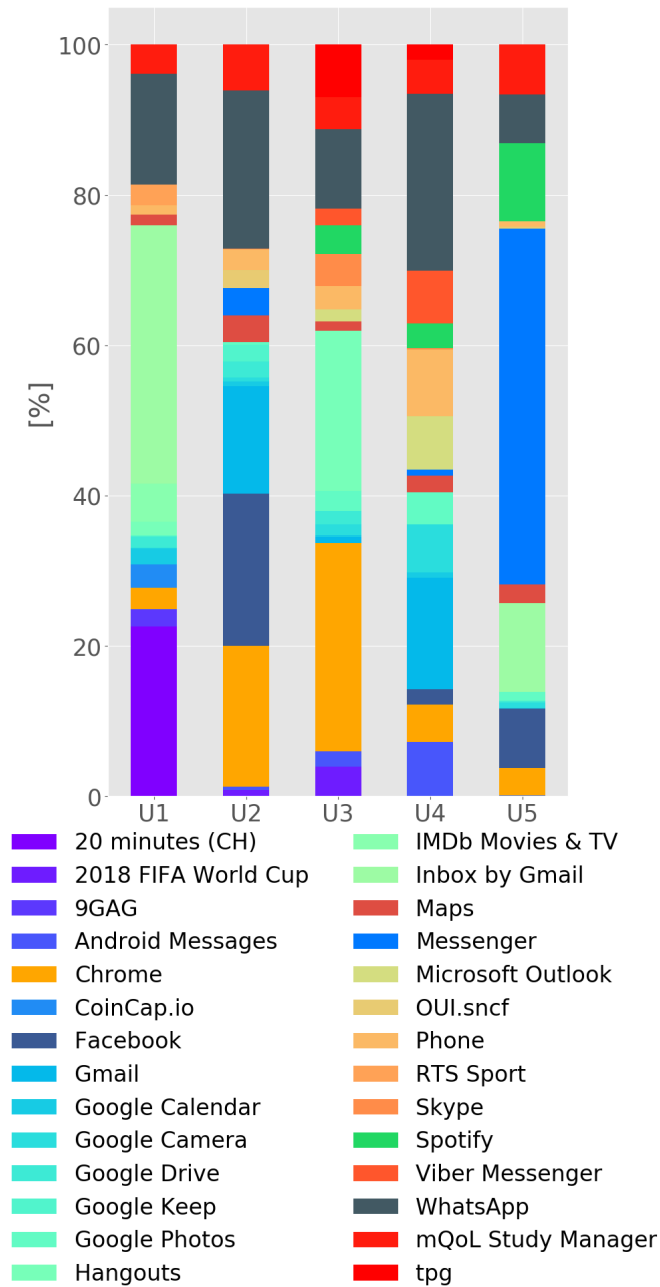


Figure 6.7 Distribution of top 10 apps per user

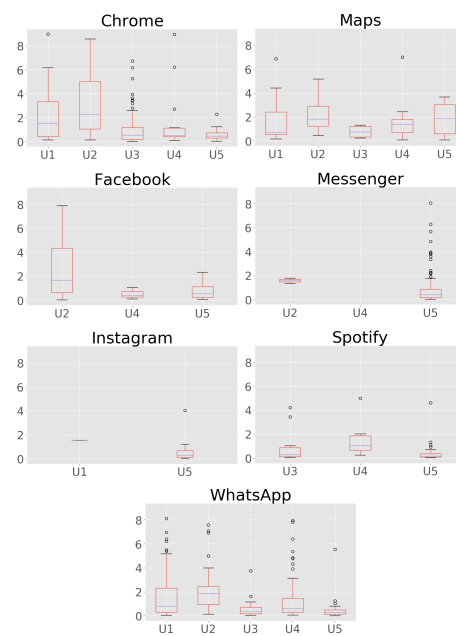


Figure 6.8 Time spend in each app per user [min]

Chapter 7

Article V: Predicting Quality of Experience of Popular Mobile Applications from a Living Lab Study

Published in Proceedings of the 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), Jun. 2019, doi: 10.1109/QoMEX.2019.8743306. vspace5mm

Chapter Contents

7.1	Introduction	102
7.2	Related Work	103
7.3	User Study	104
7.3.1	Study Protocol	104
7.3.2	Ecological Momentary Assessment (EMA)	104
7.3.3	Demographics	105
7.3.4	Data Collected	107
7.3.5	Collected Data Summary	107
7.4	QoE Prediction Model	110
7.4.1	Features	110
7.4.2	QoE/MOS Classification	110
7.4.3	Classifier	112
7.4.4	Results	112
7.5	Discussion	112
7.6	Limitations	113
7.7	Conclusions and Future Work	114

Abstract

In this paper, we present a hybrid method (qualitative and quantitative) to model and predict the Quality of Experience (QoE) of mobile applications used on Wi-Fi or cellular network. Our 33 living lab participants rated their mobile applications' QoE in various contexts for four weeks resulting in a total of 5,663 QoE ratings. At the same time, our smartphone logger (mQoL-Log) collected background information such as network information, user activity, battery statistics and more. We focused this study on frequently used and highly interactive applications including Google Chrome, Google Maps, Spotify, Instagram, Facebook, Facebook Messenger and WhatsApp. After pre-processing the dataset, we used classical machine learning techniques and algorithms (Extreme Gradient Boosting) to predict the QoE of the application usage. The results showed that our model can predict the user QoE with 94 ± 0.77 accuracy. Surprisingly, after the following top three features: session length, battery level and network QoS, the user activity (e.g., if walking) and intended action to accomplish with the app were the most predictive features. Longer application use sessions often have worse QoE than shorter sessions.

7.1 Introduction

Smartphone applications are used every day in different contexts since their introduction. The majority of them depend on an Internet connection and multiple other factors, linked to the smartphone device hardware itself (processor, memory, operating system stability, video buffer, etc.) to become an enjoyable experience to the users. Quality of Experience (QoE) is defined by [Le Callet et al. \(2012\)](#) as the degree of delight or annoyance of the user of an application or a service. QoE is profoundly shaped by the user expectations [Sackl and Schatz \(2014a\)](#), previous experiences and Quality of Service (QoS), as well as other user's context (e.g., mobility level). Collecting samples of QoE rating in-situ has always been difficult, due to many potentially confounding factors. The majority of the QoE studies are done in the lab or via crowdsourcing, where participants have to execute tasks given by the researcher and later rate their experience. In our work, unlike previous crowdsourcing-based works, we focused on collecting QoE rating and smartphone data from living lab participants unobtrusively in their daily life contexts. We then used this data to train a computational model to predict "Good" or "Bad" QoE. This paper is structured as follows: the related work is included in [Section 7.2](#). [Section 7.3](#) introduces our study, its protocol and lists the data collected. [Section 7.4](#) exposes the methodology of our work, the construction of our QoE prediction model and show its output. [Section 7.5](#) discuss the model, its results and our findings. [Section 7.6](#) presents the limitation of our study and [section 7.7](#) concludes the paper.

7.2 Related Work

The authors of [De Moor et al. \(2010\)](#) created a framework to quantify mobile QoE in a living lab setting, but they do not take into account the evolution in the smartphone landscape and the services that it generated. Past work has been done on assessing mobile broadband quality in the field by [Schatz and Egger \(2011\)](#), i.e., on laptops. The authors asked the participant to browse websites and to download files, then rate their experience. The network traffic was transferred from the broadband operator to the authors' network shaper. Their following work ([Schatz et al., 2011](#)) mapped QoE ratings and the acceptability of a web service. The factors influencing mobile application QoE are various, as showed by [Ickin et al. \(2012\)](#) in their study assessing users' perceived experience in-the-wild. The ratings were done at random times after random application usage. Quantifying QoE became more focus on specific web services with [Casas et al. \(2012\)](#); they collected QoE rating of Facebook and Youtube usage, using 3.5G mobile broadband communication on a laptop and focused on QoS metrics under their control. The authors of [Sackl et al. \(2015\)](#) focused on Web QoE ratings of a photo gallery website in a laboratory setting. In the case of video streaming on smartphone ([Wamser et al., 2015](#)) developed own Android application to collect QoE ratings of Youtube videos. The participants were invited to their lab to use specific smartphones connected to a Wi-Fi network for which the authors adjusted the network QoS, notably its available bandwidth. Other attempts have been made for quantifying QoE on Android smartphone devices. [Chen et al. \(2014\)](#) propose a tool monitoring of multiple layers, including a QoE-aware User Interface (UI) controller injected in the Android application, the overall network QoS of the device and the 4G/LTE modem state obtained using a cellular network diagnosis tool from the modem chip maker. The first study on QoE of mobile applications in real cellular networks was done by [Casas et al. \(2015c, 2016\)](#). They later combined test lab study results and QoE ratings from a study in-the-wild about various mobile services ([Casas et al., 2015a](#)). They focused mainly on QoS and the annotation of their participants to derive bandwidth thresholds for good/bad QoE on cellular networks. However, their participants were instructed to effectuate specific tasks (e.g. watch a video on Youtube and explore a map on Google Maps) and rated the QoE of that specific task in the authors' application that collected network flows information, e.g. Radio Access Technology (RAT). The flow metrics are not available on the Android OS anymore; Google removed those API as a security concern in 2016. The dataset from this study was used later in [Casas et al. \(2017a\)](#) to predict QoE comparing different machine learning classifiers. Decision tree-based was proven to get the best results. The authors followed in [Casas et al. \(2018\)](#) for predicting QoE with the benefit of ensemble models via their stacking approach. Given the state of the art, we are in an unique position to provide methods and tools that enable accurate quantification and prediction of mobile QoE in a living lab settings, in the users' daily life contexts.

7.3 User Study

Adult participants were recruited via an ad campaign on the University of Geneva mailing list. We selected the ones using Android smartphones for the longest and using a set of apps from different categories, that are highly interactive and popular (top 5 on the Google Play Store in their categories as messaging, social network, music, navigation and internet browsing). We recruited 33 participants (13 females, 1 non-disclosed) for 28 consecutive days. The study has been run in two languages at the University to allow more people to join.

7.3.1 Study Protocol

We designed the study protocol, which was approved by the university ethical committee, and followed the best practices and recommendations of the Qualinet White Paper [Hossfeld et al. \(2014\)](#). Our in-house Android application: mQoL-Log did not require **root** access on the participant smartphone, which complicated collecting low-level network data and kernel statistics. Each participant had a training session in our lab on a test device before the experiment. A web page with the content of the training was available online for later reading by participants. Once the participant has installed mQoL-Log on his/her smartphone, the user joined our study in the application (select from a list of ongoing QoL Lab studies), gave consent for the data collection, accepted and allowed the Android permissions that mQoL-Log required. Then, the user had to fill a demography survey. A feedback channel was included in mQoL-Log for participants to contact us. mQoL-Log also logged the duration that users take to provide their overall QoE feedback. We also had tools to monitor incoming data during the experiment. If a user stopped rating, the researcher could trigger a notification remotely on the user screen to motivate the user.

7.3.2 Ecological Momentary Assessment (EMA)

EMA has been used to gather QoE rating in-the-wild by [Ickin et al. \(2012\)](#) in other studies but were randomly triggered for all the applications. We implemented EMAs in our mQoL-Log via surveys after a specific application usage, hence limiting the error for the QoE rating caused by memory effect. The number of EMAs per day that all application usage could trigger was limited to 12. EMAs were only launched in waking hours (i.e., outside 9 PM to 7 AM), and the time interval between two consecutive EMAs was set up to 20 minutes. If the previous EMA was not filled when a new one is triggered, the previous one is dismissed. Users replied to the surveys by clicking on the notification at the screen's top, opening our application and providing their feedback. As shown below, some questions requested binary responses, other multiple choices and finally for the QoE rating we used the Mean Opinion Score (MOS) [ITU-T Recommendation P.800.1 \(2019\)](#), a subjective rating scale from 1 to 5 mapped to the following rating: poor, bad, poor, fair, good, excellent. The slider on the screen allowed for continuous rating as it offered higher sensitivity on the nearest target (e.g. 3.5 is between fair and good, 3.8 is close to good). [Fig. 7.1](#) represents a user selecting 5 on the scale. We simplified

our QoE rating questions to the maximum. All the questions of the EMA were the same regardless of the type of application. Smartphone applications have various actions, the third question allowed us to understand more about the user's purpose in the app. The questions and their possible responses are as follows:

1. Did your usage of app name at use start time go as expected? *Yes/No/I am not sure.*
2. How was your last usage session of app name at use start time? *Slider (MOS 1 to 5 with a colour scale from red to green).*
3. What action were you trying to accomplish? *CONSUME content, SHARE or create content, READ text message, WRITE text message, CONTROL an app (start/stop music), VIDEO call or AUDIO call.* The user can select multiple choices.
4. Did your last usage of app name at use start time meet your expectations? *Slider (MOS 1 to 5 with a colour scale from red to green).*
5. If something went wrong, please tell us more about it. *Free text entry.*

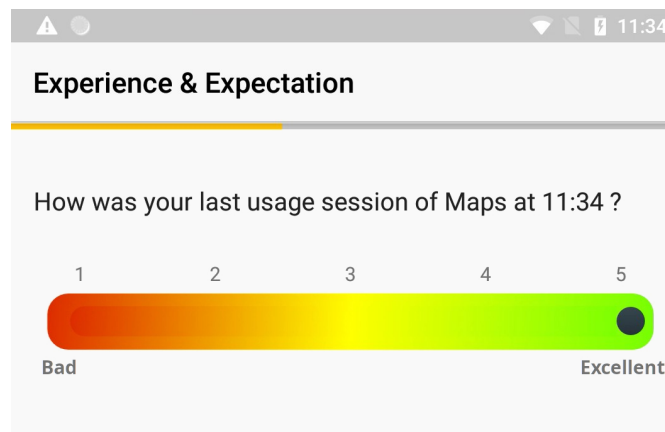


Figure 7.1 User selecting QoE=5 on the scale after using (Google) Maps

7.3.3 Demographics

The age distribution of the 33 participants is as follows. Twenty were young adults (two between 18-20 y.o. and 18 between 21-29 y.o.), followed by eight participants between 30-39 y.o., two between 40 and 49 y.o., two participants between 50-59 y.o. and one non-disclosed. The basics statistics regarding QoE ratings and participation are available in Table 7.1. A minority of participants did not use the designed applications enough to trigger all the possible EMA per participation days, and others did not respond to the triggered EMAs. We obtain an average rate of recall $75\% \pm 22$ (number of rated/number of triggered EMAs).

Table 7.1 Study Participation Metrics

user	$MOS_{mean\pm std}$	trig	ans	recall
1	5.0±0.0	131	27	20.6
2	3.88±0.52	133	121	90.9
3	3.84±0.62	313	271	86.5
4	3.88±0.77	64	52	81.2
5	4.91±0.46	309	290	93.8
6	4.96±0.21	178	172	96.6
7	4.73±0.75	92	87	94.5
8	4.0±0.0	193	136	70.4
9	4.98±0.16	213	196	92.0
10	4.91±0.47	336	324	96.4
11	3.81±0.55	230	200	86.9
12	4.98±0.19	159	123	77.3
13	4.51±0.87	347	318	91.6
14	4.86±0.39	241	215	89.2
15	4.19±0.67	380	120	31.5
16	4.9±0.44	136	22	16.1
17	4.68±0.53	251	145	57.7
18	4.91±0.52	230	207	90.0
19	4.57±0.77	298	223	74.8
20	4.24±0.65	354	131	37.0
21	4.79±0.57	150	139	92.6
22	4.96±0.25	334	266	79.6
23	4.97±0.33	289	185	64.0
24	4.95±0.25	269	142	52.7
25	4.99±0.21	504	369	73.2
26	3.59±0.62	228	221	96.9
27	4.83±0.59	126	110	87.3
28	4.9±0.45	299	264	88.2
29	4.97±0.2	318	277	87.1
30	4.84±0.5	289	164	56.7
31	4.95±0.33	232	188	81.0
32	4.93±0.31	82	71	86.5
33	4.96±0.4	184	160	86.9

7.3.4 Data Collected

The mQoL-Log background phone logger collected various time-stamped data. The Table 7.2 includes their descriptions and triggers. Before the user was asked to provide a QoE rating, once they finished their application usage session, mQoL-Log performed a network reachability test, also known as *ping*, to the application server corresponding to the app being used to provide Round-Trip Time (RTT) data as an indication of QoS level. The RTT is the time that takes a packet to go through the network to the host including the time that the host reply to the mobile client. The ping is done six times, and the first is discarded in case it was subjected to the DNS resolution. We set up a ping time out of 60 seconds, if the RTT is higher mQoL-Log stops the test. Besides the RTT, another important QoS feature is the overall network traffic of the smartphone. To gather the network connection flows information (TCP and UDP, source ip, destination ip, ports, TCP states, etc.) we collected the output of the Linux *netstat* command, which did not need root access.

7.3.5 Collected Data Summary

We collected 5,936 EMAs over the study period of 28 days, and filtering the incomplete and fallacious EMAs (e.g. incomplete answers), we obtained 5,663 exploitable ratings. On average a participant rated 180 ± 85 application QoE over the study period, overall 6 ± 3 per a day. We rounded the distribution of the application usage rating to the superior integer and present the result in Fig 7.2. The aggregate distribution of those QoE rating was as follow: 0.40 % of 1, 0.90% of 2, 5.01% of 3, 19.72 of 4 and 73.95 % of 5. They display the high imbalance of the dataset. The prevalence of “Good” QoE ($MOS \in [3.5$ to $5]$) in the data was 93.7%. The network connection type distribution of our dataset is available in 7.4. The “Handover” label covers the Internet connection transition the cellular network technologies (e.g. LTE to HSPA+; horizontal) and between Wi-Fi and cellular network (vertical). More than half, 53%, of our dataset was composed of QoE rating on a Wi-Fi connection, followed by 33% on LTE, 7% handover, 3% disconnected, 2% on HSPA+ and the last 2% on EDGE, UMTS, HSPA and HSDPA. We found that the participant activities during the application to be “still” at 52%, followed by “on foot” with 20%, “tilting” at 16%, “in vehicle” with 11%. The distribution of the intended action to accomplish was as follows: 36% of “consuming content”, 24% “reading and writing messages”, 16% “reading messages” only, 6% of non-labeled, 5% “write” only, 2% “consume and read” and “consume and write”. The remainder of the dataset contains the actions “audio”, “video”, “control application”, in total less than 1%. “Bad” QoE ($MOS \in [1$ to $3.5]$) was more present in the non-labelled actions as “consuming content” than “reading and writing messaging” and their respective single actions “writing messages” and “reading messages”. “Good” QoE was more prevalent for more actions for all applications. On average, a study participant took more time to rate “Bad” QoE (1.04 ± 5.2 min) than “Good” QoE (0.36 ± 3.7 min).

Table 7.2 mQoL-Log: Background Logger Data Collection

Name	Definition	Trigger
Network	Wi-Fi level, Wi-Fi BSSID, Wi-Fi SSID, Wi-Fi interface speed, Cell ID, Cell operator, Cell strength, Cell Radio Access Technology, Cell network code, Internet connection status, netstat (TCP network statistics), IP address, Cell bandwidth up and down stream, proxy information, domain name. Number of packet and bytes send and received on each wireless interfaces, DNS's IP address, routing table information	Changes in network connection state and during user app usage.
Ping/RTT	Active proving of the application used Internet server. A ping is executed 6 times. We derive statistics (mean, stdev and variance) from this test.	When app usage session start.
Battery	Battery state (e.g. charging, full, discharging), Battery level, Battery temperature.	Changes in battery state.
Apps	Application name on the user screen.	Changes of the application on screen.
Activity	User physical activity from the Google Play Services Activity (still, tilting: between two states, in vehicle, on bicycle, on foot, running).	Changes in user activity.
Touches	Number of user touches on the screen and duration during an usage session.	Screen Event based: new smartphone session.

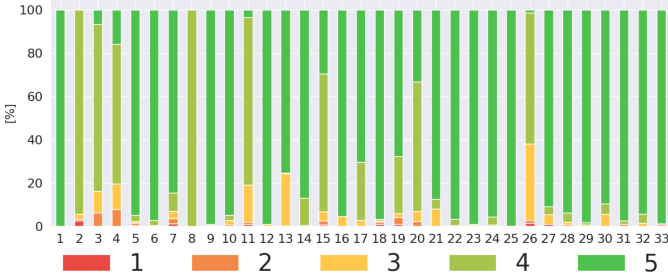


Figure 7.2 Application usage QoE/MOS rating distribution.

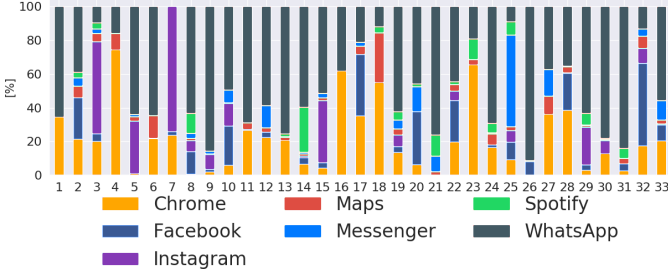


Figure 7.3 Application use distribution.

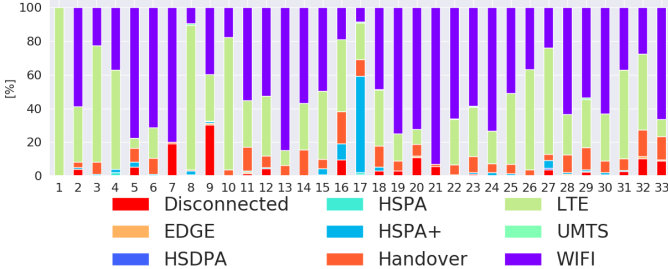


Figure 7.4 Network connectivity distribution.

7.4 QoE Prediction Model

Our goal was to predict the QoE of smartphone application usage based on the data collected from the smartphone and the QoE ratings from the participant as ground truth. We started by selecting features from our collected dataset (Table 7.2).

7.4.1 Features

From the different data collected in the background, we selected and aggregated features based around the timestamp of the application usage session of interest, hypothesizing that these features relate to user QoE. The time-based aggregation was done using a time window of two minutes centered on the event rated time. If data was unavailable for a feature, we added one more minute to the time window until data were found. We performed feature engineering to extract information from multiple features, e.g. knowing the IP address allowed us to know if the application operated over IPv6 or IPv4. We extracted the TCP states distinct count for each app usage. The list of all the network features used in the model is presented in Table 7.3. Further, we aggregated the QoE rating with the user's physical activity, the battery state, the application session duration, the application name and the action that the participant was trying to accomplish.

7.4.2 QoE/MOS Classification

Following the work of [Larson and Delespaul \(1992\)](#), we normalised the QoE rating values per user. We used one hot encoding on our categorical features to prepare them for the classifiers. The target variable was the EMA's QoE rating value. From our pre-analysis of the dataset, we observed a high imbalance on the QoE MOS rating class. No over-sampling method can be applied because of the prevalence of each class. We decided to reduce the number of classes by labelling each sample in terms of "Good" QoE $MOS \in [3.5 \text{ to } 5]$ and "Bad" QoE $MOS \in [1 \text{ to } 3.5[$, as previously employed by [Schatz and Egger \(2011\)](#). Therefore, the prevalence of "Good" QoE MOS is 93.7% and 6.3 % of "Bad" QoE. We followed a classic machine learning method by randomly selecting 70% of our data as our training dataset. The resulting 30% was split into two to obtain our validation dataset (15%) and testing dataset (15%). We conducted a randomised search cross-validation (CV=20) to optimise our model parameters. That means that the 70-15-15 split has been run 20 times by repeating the random selection of our training, validation and testing dataset, hence covering our entire dataset, this is called random permutations cross-validation (shuffle & split). After each split, we have applied SMOTE from [Chawla et al. \(2002\)](#) on the training dataset to overcome the imbalance issue. Under-sampling and over-sampling methods were tried unsuccessful. The training dataset had now 50% of "Good" QoE label (n=3718, no rating was lost), and 50% of "Bad" QoE label (n=3718); we gained 3472 artificial ratings. We scaled our training dataset, to remove the mean and scales to unit variance, as some classifiers (e.g. k-nearest neighbours) have issues with data of different unit size. This scaler was used on the validation and testing dataset. These datasets were not over-sampled.

Table 7.3 Network Features

Feature	Description (during the event to rate)
is_connected	Connection status from Android
Connection type	Network connection type described in Fig 7.4
Wi-Fi Level	Signal Strength of connected Access Point(dbm)
Wi-Fi Speed	Wi-Fi interface speed at the EMA(Mbps)
Cell strength	Signal Strength of connected cell tower(dbm)
linkDownBw	Cell downstream bandwidth (Kbps)
linkUpBw	Cell upstream bandwidth (Kbps)
win_div_net	Aggregation window for network events around the application usage time.
rxt_packets_time	Packet received per seconds during win_div_net.
txt_packets_time	Packet sent per seconds during win_div_net.
rxt_bytes_time	Bytes received per seconds during win_div_net.
txt_bytes_time	Bytes sent per seconds during win_div_net.
RTT mean	mean Round-Trip Time of the 5 pings
RTT variance	variance Round-Trip Time of the 5 pings
TCP states count during win_div_net	LISTEN, SYN-SENT, SYN-RECEIVED, ESTABLISHED, FIN-WAIT-1, FIN-WAIT-2, CLOSE-WAIT, CLOSING, TIME-WAIT, CLOSED Postel (1981)

7.4.3 Classifier

We ran a candid (non-optimized) 10x fold cross-validation on our training and validation dataset to select the algorithm with the best performance for our classification problem between: k-nearest neighbours (KNN), decision tree (DT), random forest (RF), logistic regression (LR), stochastic gradient descent (SGD), Naive Bayes (NB), gradient boosting (GB) and extreme gradient boosting (XGB). We selected the Area Under the Receiver Operating Characteristic Curve (AUC) as the metrics to find out the best classifier of our list. It expresses how accurate a model can distinguish between classes (e.g. classifying correctly the classes with minimum confusion). Table 7.4 shows the AUC of our best four classifiers on the validation classes dataset. Hence we have selected the extreme gradient boosting from [Chen and Guestrin \(2016\)](#) to most accurately predict “Good” or “Bad” QoE.

Table 7.4 AUC score on the validation dataset (all users)

Classifier	LR	KNN	SGB	XBG
$AUC_{mean \pm std}$	0.8 ± 0.04	0.8 ± 0.03	0.8 ± 0.04	0.82 ± 0.04

7.4.4 Results

For all 20 repetitions of training/validation/testing splits, we have obtained an average AUC of 0.83 ± 0.3 using the best classifier from our random search on the test data with 0.94 ± 0.01 accuracy. We derived the features importance in the XBG classifier: after the duration of application usage, battery level and QoS features, we found out that the user’s actions trying to be accomplished are relevant (e.g., send text versus view video), as well as the user physical activity (e.g., walking). The application does not matter in our case (low feature importance). To evaluate the model accuracy for the whole population, we then used the “leave one (participant) out” method and split our dataset into training, validation and testing dataset arbitrarily. One user data was used as the testing dataset, enabling us to generalise our model. We followed the same procedure as before. The average AUC of the tests dataset of our prediction model is presented in Fig 7.5. We found that the mean AUC of the test datasets to be 0.64 ± 0.20 and the accuracy 0.93 ± 0.07 .

7.5 Discussion

Some users (**1** and **8**) always provided the same ratings, even if they were rating different applications as seen by the distribution in Fig 7.3. User **25** provided only one rating of “Bad” QoE. Those users data could be discarded to obtain a more balanced dataset, but our previous attempt to predict QoE shows that more data, even imbalanced, make for a more accurate model in the end. User’s QoE is an aggregation of various variables. From our research, we found that an inexpensive predictive model can be build using QoS information, application name, user task data (user’s intent), and physical activity.

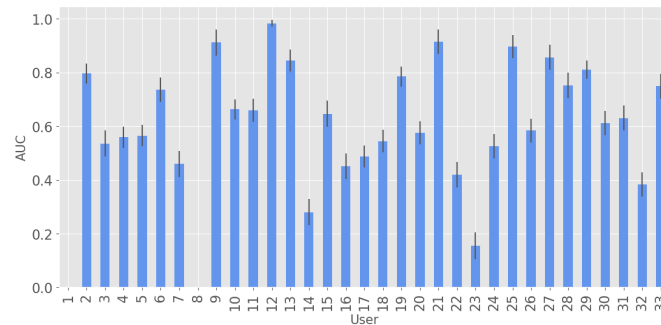


Figure 7.5 Model AUC for each user as test datasets.

The activity is playing an important role in the user experience, e.g. a participant will not rate the same if he/she is using an application walking or staying still. There are following factors influencing the living lab study, and potentially the collected data quality: the other aspect of the user's context (e.g. mental state), previous experience, surroundings, operating system updates and new available features. The app developer should optimise their application to seamlessly handle "Bad" QoE depending on what the user want to accomplish inside the application. "Bad" QoE ratings are higher when the user is 'writing' and 'tilting' between physical activity, hence with this information, the developer could provide a better way of inputting text in their messaging application (e.g. proposing predefined short answer from a substantial touch area). What is important for an application developer is that with better QoE a user is more effective; spend less time on the application accomplishing the intended tasks faster, but also use more features in an application.

7.6 Limitations

This study was only possible on Android devices, as data collection is more difficult on iOS. We can not generalise our finding for another operating system platform. The collection of the number of frames dropped by an application's UI would have been a plus to understand the hardware status. The subset of applications that we focus on did not include high bandwidth need which was studied by others, particularly regarding video QoE on smartphones. The landscape of smartphone application is evolving each day with new innovative services, modelling QoE for each new application and their underlying features' would be a hassle. Hence we tried to generalise QoE prediction based on user action within an application. The limitation is that the user's momentary emotion and stress level can influence the annotation of their application usage, as well as a participant can rate a "Good" QoE application usage negatively because of the content of the application. Even if the participants were told about this effect, the underlying influence could be a confounding factor for our model.

7.7 Conclusions and Future Work

In this paper, we presented an attempt to model and predict smartphone application QoE from a living lab study. We showed that collecting in-situ QoE rating and collecting smartphone background data enable us using common machine learning techniques, to build a predictive model for “Good” and “Bad” QoE. In the future, we want to implement our pre-trained prediction model inside an Android application. The application will predict if in the near future (e.g., 5 minutes) the QoE of current application usage session will be “Good” or “Bad”. If the prediction shifts because of the context (e.g. train inside a tunnel), the application can inform the user and prepare itself for the change. Those predictions, rated by the user, would allow us to use reinforcement learning to enhance our model comparable to recommendation systems.

Acknowledgment

SNSF MIQmodel (157003) (2015-2019) & AAL CoME (2014-7-127). The University of Geneva ethics committee approved this work under *CUREG.201803.02* on 02/07/2018.

Chapter 8

Article VI: Towards Accurate Models for Predicting Smartphone Applications' QoE with Data from a Living Lab Study

Published in *Quality and User Experience*, vol. 5, no. 1, Springer Journal, Oct. 2020, doi: 10.1007/s41233-020-00039-w.

Chapter Contents

8.1	Introduction	117
8.2	Related work	119
8.2.1	Quantifying QoE on Laptop	119
8.2.2	Smartphone Applications' QoE	119
8.2.3	User's Expectation and QoE	120
8.3	The Approach: User Study	122
8.3.1	Study Protocol	122
8.3.2	Ecological Momentary Assessment (EMA)/MOS	122
8.3.3	Smartphone-Based Data Collected	123
8.3.4	Collected Data Summary	124
8.4	Building QoE Prediction Models	128
8.4.1	Features	130
8.4.2	QoE/MOS Classification	130
8.4.3	Classifier Selection	131
8.4.4	Candid Model (XBG)	133
8.4.5	Filter Time to Reply (TR)	133
8.4.6	Unlabeled Tasks (ULT)	133
8.4.7	Filter Features Aggregation Time (FA)	133
8.4.8	Merged Filter Replies (TR) and Features Aggregation (FA) Time (TRFA)	135
8.4.9	Meta-features Selection (MT)	135
8.4.10	Expectation (EX)	136
8.4.11	On-device Prediction (OD)	136

8.5	Results	137
8.5.1	Overview of Previous Work	137
8.5.2	Scenarios' Results	137
8.6	Discussion	139
8.6.1	Ratings Quality	140
8.6.2	Features Wranglings	140
8.6.3	On-device Prediction	140
8.6.4	Recommendations for the Application Developers	141
8.6.5	Modelling Highlights	141
8.7	Study Limitations	142
8.8	Conclusions and Future Work Areas	143

Abstract

Progressively, smartphones have become the pocket Swiss army knife for everyone. They support their users needs to accomplish tasks in numerous contexts. However, the applications executing those tasks are regularly not performed as they should, and the user-perceived experience is altered. In this paper, we present our approach to model and predict the Quality of Experience (QoE) of mobile applications used over Wi-Fi or cellular network. We aimed to create predictive QoE models and to derive recommendations for mobile application developers to create QoE aware applications. Previous works on smartphone applications' QoE prediction only focus on qualitative or quantitative data. We collected both qualitative and quantitative data in-the-wild through our living lab. We ran a 4-week-long study with 38 Android phone users. We focused on frequently used and highly interactive applications. The participants rated their mobile applications' expectations and QoE and in various contexts resulting in a total of 6,086 ratings. Simultaneously, our smartphone logger (mQoL-Log) collected background information such as network information, user physical activity, battery statistics, and more. We apply various data aggregation approaches and feature selection processes to train multiple predictive QoE models. We obtain better model performance using ratings acquired within 14.85 minutes after the application usage. Additionally, we boost our models' performance with the users' expectations as a new feature. We create an on-device prediction model with on-smartphone features. We compare its performance metrics against the previous model. The on-device model performs below the full feature model. Surprisingly, among the following top three features: the intended task to accomplish with the app, the application's name (e.g., WhatsApp, Spotify), and network Quality of Service (QoS), the user physical activity is the most important feature (e.g., if walking). Finally, we share our recommendations with the application developers, and we discuss the implications of QoE and expectations in mobile application design.

8.1 Introduction

Smartphone applications are used all the time in various contexts since their introduction. The majority of them depend on an Internet connection and other factors, linked to the smartphone hardware and software (e.g., processor, memory, video buffer) to become an enjoyable experience for the users. Quality of Experience (QoE) is defined by [Le Callet et al. \(2012\)](#) as "the degree of delight or annoyance of the user of an application or service" QoE is profoundly shaped by user expectations ([Sackl and Schatz, 2014a](#)), previous experience (i.e., expectations), and Quality of Service (QoS), (e.g., network speed), as well as other user contexts (e.g., mobility level). Collecting samples of QoE ratings has always been difficult due to potentially confounding factors, including the user's immediate context. The majority of the QoE studies are done in-lab or in-situ via crowdsourcing, where participants have to execute tasks given by the researchers and later rate their experience. The previous work focused on repetitive actions orchestrated by the study authors, executed by the study participant. Once finished, the action experience was rated and collected.

In our work, we focus on collecting QoE and expectation ratings from living lab participants unobtrusively in their daily life contexts. We propose a method to build smartphone applications' QoE prediction models with data collected discreetly through a smartphone logger (mQoL-Log) and assisted by the Ecological Momentary Assessment (EMA) methodology in-the-wild. We focus on both qualitative and quantitative research for modeling QoE. This hybrid method added contextual information to design a QoE prediction model. This approach enables us to make recommendations to smartphone application developers. We train multiple prediction models to predict "High" or "Low" QoE. We perform extensive data wrangling and cleaning to build better predicting models with higher performance than our previous work (De Masi and Wac, 2019). We investigate different machine learning algorithms, as previously shown to be used in the literature (Casas et al., 2018).

Through our work, we present a path toward building an accurate QoE model from a dataset obtained in-the-wild. Contrary to an in-the-lab study, where researchers have full control over the stimulus applied to their participants. In-the-wild studies create new challenges, e.g., external factors influence the experiment. They have to be leveraged to extract knowledge from the collected data.

Current smartphone sensing technologies and the recent development in machine learning tools enable our path toward predictive QoE models executed directly on the smartphone. In the future, a smartphone should be able to anticipate the change in its QoE and clearly notify its users about the possible disappointment, (via, e.g., notification or even a screen color change).

The process of building QoE predictive models requires filtering and aggregating the raw data produced by the participants. We review the quality of our rated applications' QoE samples and the performances of our aggregation solutions. We focus on the time between the application's usage event and its rating. We explore the unlabeled tasks in the application and look into their usability to create an improved model. We investigate the impact of expectation on QoE and its correlation. We review the possibility of an on-device prediction model and its shortcoming. We investigate the most predictive features from our models to derive what factors affect smartphone applications' QoE in-the-wild. Once established, we create a model to predict QoE based on these factors. To test the importance of various features for the QoE model, the paper describes an iterative model building methodology (i.e., including data filtering and wrangling techniques, as well as model evaluation metrics).

Modeling QoE in-the-wild requires data collection from various perspectives. mQoL-Log allows for context monitoring on Android smartphones. It enables extensive data collection of critical external factors linked to participant annotation and the observed experiences. Smartphone-based context monitoring can become an issue if the privacy and experience of the participants are impacted. Hence, we design a study protocol to reduce the participant's burden. If the data collection endeavour has not been carefully planned, the study can directly impact the user's day to day smartphone-based activities and experiences.

This paper is structured as follows: the related work is in section 8.2. Section 8.3 introduces our study, its protocol, and lists the data collected. Section 8.4 exposes the methodology of our work, the construction of the QoE prediction models, and shows its output. Section 8.6 discusses the models, its

results, and our findings. Section 8.7 presents the limitation of our study and section 8.8 concludes the paper. In this paper, the term “accurate” refers to the QoE models’ performance (high accuracy: value close to 1, low: value close to 0), as used by the machine learning and QoE community.

8.2 Related work

We focused our literature review on three main categories of papers. Firstly, we looked at past QoE studies with mobile Internet devices. At that time, the most used mobile Internet devices were laptops connected to broadband Internet. Secondly, we investigated the work of quantifying QoE on smartphones. We explored the work of applications’ QoE on smartphones from framework to model. In our last category, we reviewed the works linking expectation and QoE, especially for smartphone as a platform.

8.2.1 Quantifying QoE on Laptop

In 2011 [Schatz and Egger \(2011\)](#) assessed mobile broadband quality in-situ and in-lab on laptops. The authors asked the participants to browse websites and to download files, then rate their experience. The network traffic was transferred from the broadband operator to the authors’ network shaper. Their following work the same year ([Schatz and Egger, 2011](#)) mapped QoE ratings and the acceptability of a web service. The previously cited works were all done with instructions to rate tasks experience arranged by the authors. Their study setup modified the typical experience of their participants. [Casas et al. \(2012\)](#) focused their work on the specific web services’ QoE. In their study, they collected QoE ratings from Facebook browsing and YouTube usage in-the-wild. Their 33 participants used a laptop with a 3.5G mobile broadband connection provided by the authors for 31 days. The traffic was rerouted to the authors’ network before accessing the Internet. They applied traffic shaping to influence the participants’ QoE. The participants rated the quality of the connection and the overall experience on a MOS scale, as well as the acceptability of the service. They focused their approach on QoS metrics (e.g., downlink bandwidth, traffic volume, video resolution) to compute MOS scores expressing QoE. They did not collect any context information except the physical location of the participants (home, work, university, outdoor, and other) manually reported. [Sackl et al. \(2015\)](#) focused on Web QoE ratings of a photo gallery website and on overall quality for three uses: browsing a news website, uploading a large file, and exploring different cities in Google Maps in a laboratory setting. They modulated the bandwidth and its stability to observe the participants quality perception.

8.2.2 Smartphone Applications’ QoE

[De Moor et al. \(2010\)](#) created a framework to quantify mobile QoE for all smartphone applications in a living lab setting. They evaluated the QoE of Java platform applications in the implementation of their framework. However, they did not take into account the evolution in the cellphone landscape to the smartphone and its generated services. The mobile Java platform was getting obsolete at that

time (2010). The first Apple smartphone (iPhone) was released in 2007; the first Android device (HTC Dream) was available in 2008. They advocated for long-term and user-centric perspective QoE studies in living labs, without operationalizing it. The factors influencing mobile application QoE are various, as shown by [Ickin et al. \(2012\)](#) with their study assessing users' perceived experience in-the-wild. The ratings were provided at random times after any application usage. The authors presented factors that impacted the users' QoE, such as the application interface design, performance, the battery efficiency, the in-application features, and application name. They also exposed user-centric ones such as connectivity cost, user routines, and user lifestyle. Other attempts have been made for quantifying QoE on Android OS smartphone devices. [Chen et al. \(2014\)](#) proposed a tool to monitor multiple QoE factors, including a QoE-aware User Interface (UI) controller injected in the Android OS application, the overall network QoS of the device, and the 4G/LTE modem state obtained by a cellular network diagnosis tool from the modem chipmaker. In the case of video streaming on smartphones, [Wamser et al. \(2015\)](#) developed its Android OS application to collect QoE ratings of YouTube videos. The participants were invited to their lab to use specific smartphones connected to a Wi-Fi network for which the authors adjusted the network QoS, notably its available bandwidth. The first study on QoE of mobile applications in real cellular networks was done by [Casas et al. \(2015c, 2016\)](#). They later combined in-lab study results from [Casas et al. \(2015a\)](#) and QoE ratings from a study in-the-wild about various mobile services ([Casas et al., 2015c](#)). They focused mainly on QoS and the annotation of their participants to derive bandwidth thresholds for good/bad QoE on cellular networks. However, their participants were instructed to perform specific tasks (e.g., watch a video on YouTube and explore a map on Google Maps), possibly interfering with their normal smartphone usage and creating a bias. They rated the QoE of that specific task in the authors' application. It collected network flow information, e.g., Radio Access Technology (RAT). The flow metrics are not available anymore without root access on the Android OS. Google removed those APIs as a security concern in 2016. The dataset from this study was used later by [Casas et al. \(2017a\)](#) in 2017 to predict QoE comparing different machine learning classifiers. Decision Tree-based classifiers were proven to get the best results. The authors followed with by [Casas et al. \(2018\)](#) and by [Casas \(2018\)](#) to predict QoE with the benefit of ensemble models via their stacking approach [Casas et al. \(2017b\)](#). In all their work, they did not integrate physical activity, user habits (e.g., time spent in the application), expectations or active network testing.

8.2.3 User's Expectation and QoE

Even though the expectation is pointed out in QoE models, its assessment in Internet-based services is rare. [Sackl et al. \(2012\)](#) proposed an experiment to test user expectations and QoE on wireless 3G connections versus an ADSL Internet access. The participants were directed to test various internet usage scenarios, mainly browsing websites and playing videos on a laptop. The authors modulated the QoS and provided on-screen Internet connection type label (i.e., wireless 3G or wireline ADSL) to the user. They showed that expectations, such as QoE, are relying on usage scenarios and applica-

Table 8.1 Study EMAs Questions

Questions	Answer	Type	Features
Did your usage of app name at use start time went as expected?	Yes/No/I am not sure	Single choice	Expectation
How was your last usage session of app name at use start time?	MOS 1 to 5 with a color scale from red to green	Slider	Application's QoE
What action were you trying to accomplish?	CONSUME content, SHARE or create content, READ text message, WRITE text message, CONTROL an app (start/stop music), VIDEO call or AUDIO call	Multiple choices	Task
Did your last usage of app name at use start time meet your expectations?	MOS 1 to 5 with a color scale from red to green	Slider	Expectation MOS
If something went wrong, please tell us more about it.	Text	Free text entry	Anecdotal

tions. However, the lab-based study did not take into account how many of their participants used 3G or ADSL in their day-to-day life. In their later work [Sackl and Schatz \(2014a\)](#) were able to improve two Web QoE models (Google Maps and file download) using expectation related data gathered via questionnaires before their in-lab experiment. After each test, the participants rated the experienced quality using a 5-point Absolute Category Rating (ACR) scale. The two models integrated two expectation types: desired expectation for the Google Maps application and adequate expectation for downloading online files. The desired expectation is mostly constant over time, contrary to the adequate expectation prone to change depending on the current context ([Zeithaml et al., 1993](#)). The other input in the model was the downlink bandwidth (DLBW). The authors targeted the MOS rating given by their participants; as they integrated user expectations in QoE assessment, they increased the MOS prediction accuracy in their models. [Sackl et al. \(2017\)](#) investigated user expectations and QoE in the context of networked multimedia. They showed how QoE could be integrated into QoE research. They focus on expectations before a task. Their experiments were lab-based in a controlled environment. The expectation is often influenced by the novelty of the user's context and its past experiences (i.e., fulfillment or disappointment).

The majority of the past work was in-lab, without taking into account the importance of external contextual factors influencing the QoE. The in-the-wild studies were not conducted unobtrusively. They focused on network QoS (i.e., flow size and throughput) and did not integrate user behavior in the application (e.g., time spent and task to accomplish) and expectations. Given the state of the art,

we are in a unique position to provide insight to enable accurate modeling and prediction of mobile QoE in a living lab setting in the users' daily life context.

8.3 The Approach: User Study

To find the factors affecting smartphone applications' QoE in-the-wild and to create models to predict QoE, we conducted a user study. We present our study protocol (section 8.3.1), then we describe the tools enabling us to collect the data (sections 8.3.2 - 8.3.3). Finally, we summarize the data acquired in the study (section 8.3.4).

8.3.1 Study Protocol

In our study, participants rated their smartphone application usage QoE in a minimally intrusive manner on their Android OS smartphones using our application. Adult participants were recruited via an ad campaign on the University of Geneva (UNIGE), Centre Universitaire d'Informatique mailing list. We selected the ones using Android OS smartphones for the longest and using a set of apps from different categories that are highly interactive and popular. The picked applications on the Google Play Store were respectively, in the top 5 for their categories: messaging, social network, music, navigation, and Internet browsing. The categories correspond to the listing on the Google Play Store. They are Google Chrome, Google Maps, Spotify, Instagram, Facebook, Facebook Messenger, and WhatsApp. Those applications are used on millions of devices. Our selection was based on minimizing the effect created by the introduction of new applications to our participants (i.e., limiting bias) and to maximize our recruiting pool. In our set of selected applications, we found common tasks available to the user, even in applications from different categories. For example, it is possible to share content in all the applications. However, only some allow listening to audio content: Spotify (song), WhatsApp (voice message), Messenger (voice message), and Facebook (songs in the timeline). We recruited 38 participants (15 females, two non-disclosed) along November to December 2018 (P1-P38). The study ran for 28 consecutive days in two languages common at UNIGE, to allow more people to join. The participants were invited to install our homemade application mQoL-Lab which integrated our data logger mQoL-Log.

8.3.2 Ecological Momentary Assessment (EMA)/MOS

EMAs originated in psychology as a momentary assessment of an individual's state or emotion (Shiffman et al., 2008), hence limiting the errors caused by memory effect on the ratings. EMAs were used to gather QoE ratings in-the-wild (Ickin et al., 2012). We implemented EMAs in our mQoL-Lab via surveys after a specific application usage detected by mQoL-Log.

The number of EMAs triggered per day was limited to 12. EMAs were only launched in the waking hours (i.e., outside 21:00 to 7:00), and the time interval between two consecutive EMAs was set up to 20 minutes. If the previous EMA was not filled when a new one is triggered, the previous one

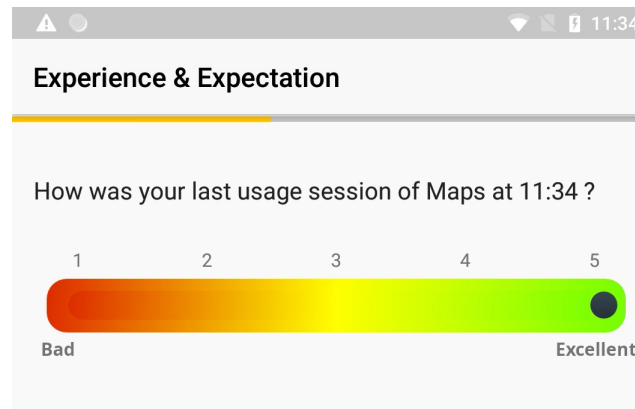


Figure 8.1 User Selecting QoE=5 on the Scale After Using (Google) Maps

was dismissed. Users replied to the survey by clicking on the notification at the screen's top. The EMA questions and their possible responses in our study are available in Table 8.1. As shown below, some questions requested binary responses, while others multiple choices. Finally, for the QoE rating, we used the Mean Opinion Score (MOS) (Union, 1993), a subjective rating scale from 1 to 5 mapped to the following rating: poor (1), bad, fair, good, excellent (5). The slider on the screen allowed for continuous rating as it offered a higher definition on the nearest target (e.g., 3.5 is between 'fair' and 'good', 3.8 is close to 'good' for acceptability of a service (Schatz et al., 2011)). On the application level, the slider scale contains a two decimal digit precision (e.g. 1.00 to 5.00). We round this information to one decimal, as the participant does not see this level of detail on his/her screen. Figure 8.1 represents a user selecting 5 on the MOS scale. The expectation questions are new; in past work on expectation and QoE (ITU-T Recommendation P.800.1, 2019; Sackl et al., 2017), study participants were asked to rank affirmations about their expectation of a specific Web service (e.g. "What do you expect from a Video on Demand Provider?"). The provided questionnaires were modified for our use case, as they advised in their work. We diverged by assessing the user's expectation fulfillment after the experience. Hence, we investigated whether their expectation impacts their experience. If so, does this factor impact QoE modeling. In past studies, the questions of the EMA were the same regardless of the type of application, while our third EMA question allowed us to understand the user's purpose in the app. We decide to reduce the number of classes by labeling each sample in terms of "High" QoE MOS above 3.5 included and "Low" QoE MOS below 3.5, as recommended by the ITU ITU-T Recommendation P.800.1 (2019) and previously employed by Schatz et al. (2011).

8.3.3 Smartphone-Based Data Collected

The mQoL-Log background phone logger (De Masi et al., 2016) collected various timestamped data. Our first trial study on smartphone applications' QoE (De Masi and Wac, 2018) indicated the importance of the user's actions to be accomplished inside an application. The Table 8.2 includes the data collectors' descriptions and triggers. QoS is a part of QoE (Le Callet et al., 2012). Hence, the

network-related data were important. Once the participants finished their application usage session, mQoL-Log performed a network reachability test, also known as *ping*, to the application server corresponding to the app. A ping provides Round-Trip Time (RTT) [ms] data as an indication of the QoS level. The RTT is the time that takes a packet to go from the client through the network to the host, including the time for the host reply to arrived at the client. The ping is done six times, and the first is discarded in case it was subjected to a DNS resolution. We set up a ping time out threshold of 60 seconds. mQoL-Log stopped the test if the threshold was met. The pings were executed at the beginning of the application usage, for which the QoE/EMA was being triggered. Besides the RTT, another important QoS feature is the overall network traffic of the smartphone. To gather the network connection flows information (TCP and UDP, source IP, destination IP, ports, TCP states), we collected the output of the Linux *netstat* command, which did not need root access (see details in section 8.4.1). We purposely recorded the ones proven by the literature (in-lab and in-the-wild on other platforms) to be an accurate indicator of smartphone application's QoE.

Table 8.2 mQoL-Log: Background Logger Data Collection

Name	Definition	Trigger
Network	Wi-Fi level, Wi-Fi BSSID, Wi-Fi SSID, Wi-Fi interface speed, Cell ID, Cell operator, Cell strength, Cell Radio Access Technology, Cell network code, Internet connection status, netstat (TCP network statistics), IP address, Cell bandwidth up and down stream, proxy information, domain name. Number of packets and bytes sent and received on the wireless interface, DNS's IP address, routing table information	Changes in network connection state and during user app usage
Ping/RTT	Active probing of the application Internet server. A ping is executed 6 times. We derive statistics (mean, stdev, and variance) from this test.	When the app usage session starts
Battery	Battery state (e.g., charging, full, discharging), battery level, battery temperature	Changes in battery state
App Name	Application name on the user screen	Changes of the application on the screen
Physical Activity	User physical activity from the Google Play Services activity (still tilting: between two states, in-vehicle, on a bicycle, on foot, running)	Changes in user activity
Touches	Number of user touches on the screen and duration during a usage session	Screen event-based: a new smartphone session

8.3.4 Collected Data Summary

The age distribution of the 38 participants is as follows. Twenty-two were young adults (three between 18-20 year-olds and 19 between 21-29 year-olds), followed by ten participants between 30-39 year-olds, two between 40 and 49 year-olds, two participants between 50-59 year-olds and two nondisclosed their age.

Table 8.3 Study Participation Raw Metrics

User ID	MOS [mean±sem]	Trig [n]	Ans [n]	P_{rate} [%]
1	5.0±0.02	131	27	20.6
2	3.88±0.52	133	121	90.9
3	3.84±0.62	313	271	86.5
4	3.88±0.77	64	52	81.2
5	4.91±0.46	309	290	93.8
6	4.96±0.21	178	172	96.6
7	4.73±0.75	92	87	94.5
8	4.0±0.0	193	136	70.4
9	4.98±0.16	213	196	92.0
10	4.91±0.47	336	324	96.4
11	3.81±0.55	230	200	86.9
12	4.98±0.19	159	123	77.3
13	4.51±0.87	347	318	91.6
14	4.86±0.39	241	215	89.2
15	4.19±0.67	380	120	31.5
16	4.9±0.44	136	22	16.1
17	4.68±0.53	251	145	57.7
18	4.91±0.52	230	207	90.0
19	4.57±0.77	298	223	74.8
20	4.24±0.65	354	131	37.0
21	4.79±0.57	150	139	92.6
22	4.96±0.25	334	266	79.6
23	4.97±0.33	289	185	64.0
24	4.95±0.25	269	142	52.7
25	4.99±0.21	504	369	73.2
26	3.59±0.62	228	221	96.9
27	4.83±0.59	126	110	87.3
28	4.9±0.45	299	264	88.2
29	4.97±0.2	318	277	87.1
30	4.84±0.5	289	164	56.7
31	4.95±0.33	232	188	81.0
32	4.93±0.31	82	71	86.5
33	4.96±0.4	184	160	86.9
34	4.0±0.0	40	37	92.5
35	4.62±0.68	78	72	92.3
36	4.79±0.58	167	88	52.6
37	4.57±0.76	184	134	72.8
38	4.05±0.23	62	21	33.8
ALL	4.61±0.43	222±104	166±89	75±23
Σ		8445	6308	74.7

As the number of EMA per day was limited to 12, only a maximum of 336 ratings could have been collected per user. A minority of participants ($N=5$) used the designed applications enough to trigger all the possible EMA per day. The participants did not respond (Ans) to all the triggered (Trig) EMAs. We obtained an average rate of assessments $P_{rate}=75\pm 23\%$ (where P_{rate} is the number of provided ratings /number of triggered EMAs per participant). Table 8.3 specify the mean \pm std MOS score of the participants and their P_{rate} . Only 24 participants obtained a P_{rate} higher than the aggregated P_{rate} . We collected 6308 EMAs. We obtain 6,086 exploitable EMAs after filtering the incomplete and erroneous EMAs (e.g., incomplete answers in one of the first two questions), we remove 3.5% of EMAs.

We round up the distribution of the application usage rating to the closest integer value and present the result in Figure 8.2. The aggregated distribution of those QoE ratings is as follows: 0.39 % of 1, 0.89% of 2, 5.03% of 3, 20.22% of 4, and 73.45 % of 5. The ratings display a high imbalance in the dataset. The prevalence of "High" QoE MOS is 93.5% and 6.5 % of "Low" QoE for D38.

The network connection distribution of our dataset is available in Figure 8.3. We defined handover as a change in the networking technology (e.g., cellular to Wi-Fi, Wi-Fi to cellular, and EDGE to LTE). The "Handover" label covers the Internet connection transition the cellular network technologies (e.g., LTE to HSPA+; horizontal) and between Wi-Fi and cellular network (vertical). Our application collected samples whenever the phone was used, including samples over cellular networks during the participants' commute and other mobility events. More than half of our dataset (61%) was composed of QoE ratings on a Wi-Fi connection, followed by 31% on LTE, 2% handover, 3% disconnected, 2% on HSPA+ and the last 2% on EDGE, UMTS, HSPA, and HSDPA.

We found that the participants' physical activities during the application to be "still" at 52%, followed by "on foot" with 20%, "tilting" at 16%, "in vehicle" with 11%. The per-user distribution of activity is in Figure 8.4.

The distribution of the intended action to accomplish was as follows: 42% of "consuming content", 24% "reading and writing messages", 18% "reading messages" only, 6% of non-labeled and 5% "write" only. The remainder of the dataset contains actions like "audio", "video", "control application"; in total less than 2%. "Low" QoE ($MOS\in[1 \text{ to } 3.5]$) was more present for the non-labeled actions like "consuming content" than for "reading and writing messaging" and own single actions "writing messages" and "reading messages". "High" QoE was more prevalent for more actions for all applications.

We observed the same rating behavior for the different age categories, except for the oldest (50 to 59 years old) - they all rated "High" QoE higher than 90% of the time. The latter rate "High" QoE only in 84% of their application usage. "time to reply" is defined as the amount of time between the end of application usage and the moment when the participant started to reply to our EMA, the overall mean is 14.85 ± 0.95 minutes. The oldest group provided the rating fastest, 94% lower than the mean "time to reply" overall categories. The 30-39 group answered 91% under the same threshold. The younger groups followed, 89% for the 21-29 and 84% for the 18-20. The 40-49 group took longer to answer. Only 77% of their ratings were given under the threshold. On average, a study participant took more time to rate "Low" QoE (1.02 ± 0.26 min, mean \pm sem) than "High" QoE (0.36 ± 0.05 min).

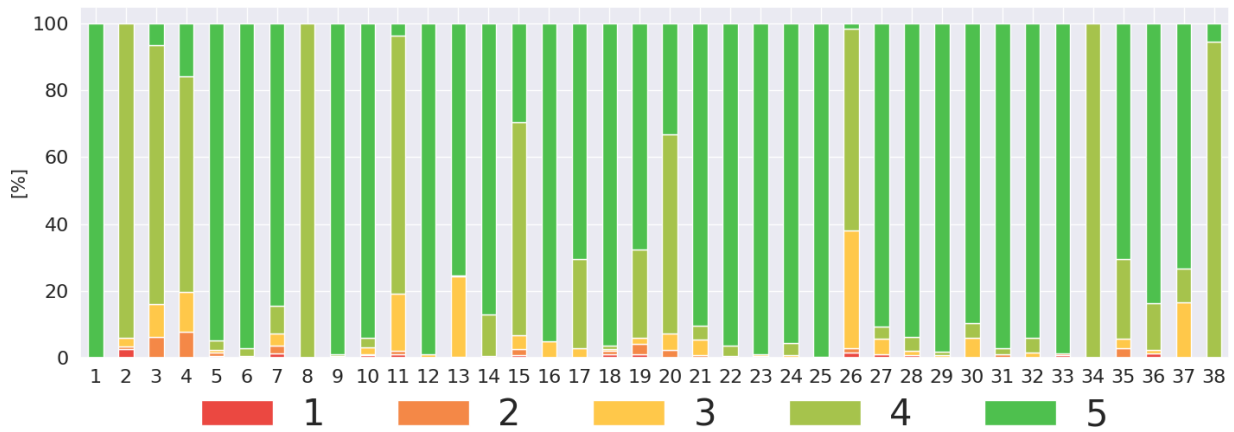


Figure 8.2 Application Usage QoE/MOS Rating Distribution Per Participant

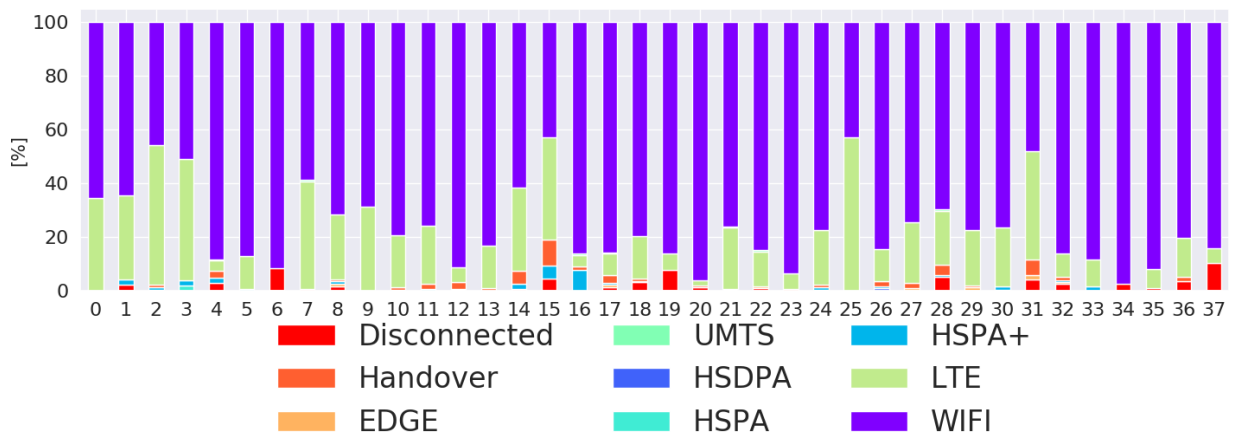


Figure 8.3 Network Connectivity Distribution Per Participant

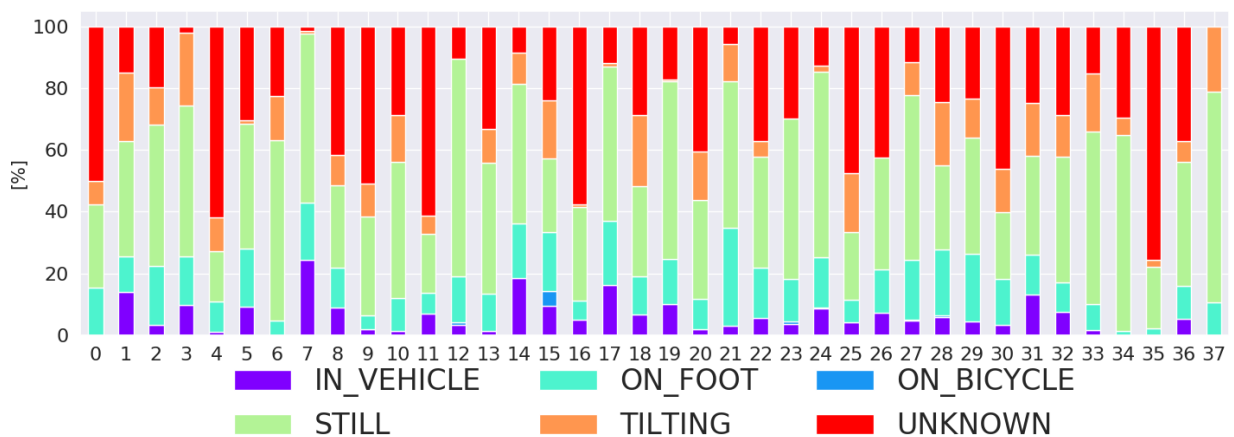


Figure 8.4 Physical Activity Distribution Per Participant

We defined the “screen session” as the amount of time between the screen turning on, and the screen turns off automatically (system timeout) or from a user’s action (manual locking). Several

applications are generally used one after another during those sessions. We computed the mean time spent inside the same screen session as 6.95 ± 0.2 minutes. On average, the users spent 2.34 ± 1.53 minutes in the selected applications of this study. On average, a participant rated 166 ± 89 application usage over the study period, overall 6 ± 3 per day. We name D38 the fully cleaned collected dataset.

The youngest group spent more time than any other groups (16.67 ± 2.46 minutes) in a session. Surprisingly, the oldest group (50-59 years old) spent 11.12 ± 1.74 minutes, coming second. They are followed by the 21-29 group with 6.76 ± 0.27 minutes. The 30-39 group spent on average 5.39 ± 0.47 minutes in “screen session”. Finally, the 40-49 group spent less time in the sessions (4.89 ± 0.57 minutes).

We investigated the users' expectation distribution from their answers to the first EMA question. Overall, only in 2% of application usage session participants were not sure about their expectations. We found that 95% of their application usage session went as expected. In “Low” QoE application's session, 76% was unexpected. In “High” QoE application's session, we found 96% of expectation matched. We found a moderate positive relationship (Hinkle et al., 2003) with a correlation between the expectation and the QoE rating of 0.595.

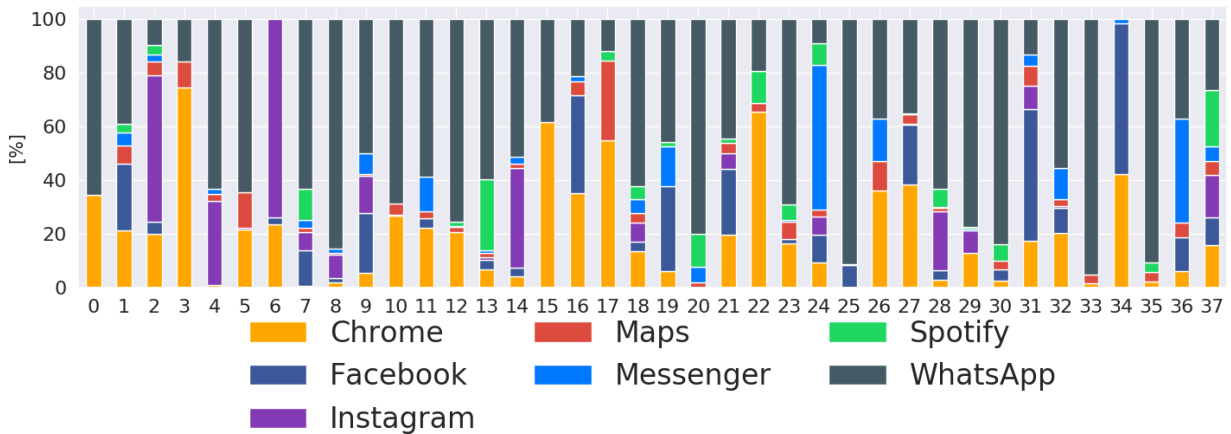


Figure 8.5 Applications Distribution Per Participant

8.4 Building QoE Prediction Models

Our goal is to predict the “High” or “Low” QoE of smartphone application usage based on the on-board data collected from the smartphones labeled by the participants' QoE ratings. The latter is used as a ground truth. We formulate it as a classification problem. We start by selecting features from our collected dataset (Table 8.2). The features are the ones influencing the user in an application usage session, and hence they are the input data of our model. We motivate our feature selection in the following subsection 8.4.1. We apply the most appropriate learning practices during our modeling. We split our data into training, validation (used for classifier hyperparameters' optimization), and testing datasets before any oversampling (no-leakage). We explore the features' importance in our

prediction models. The importance is provided by the eXtreme Gradient Boosting library (Chen and Guestrin, 2016) powering the models. For each model, we report its predictive accuracy, AUC, and recall to evaluate its performance in section 8.5. For the previously listed metrics, values closest to 1 are optimal.

Figure 8.6 summarizes our process pipeline. We went through the pipeline eight times. The first time was to select the best machine learning classifier for our QoE prediction model. The other time, it was for building models with different input features and data aggregation methods.

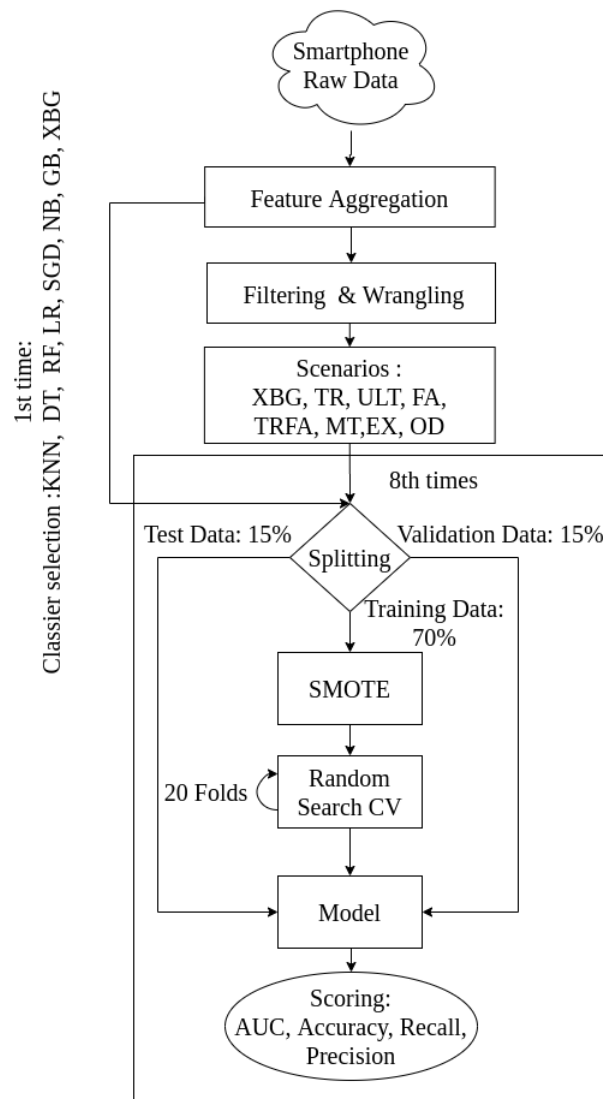


Figure 8.6 QoE Modeling Process Pipeline

8.4.1 Features

This section described part of Figure 8.6, the data aggregation, filtering, and wrangling blocks. From the different data collected in the background, we select and aggregate features centered on the beginning time of the application usage session of interest, hypothesizing that these features relate to the user QoE. The time-based aggregation was done using a time window of two minutes centered on the QoE rating. This time window is selected based on the average time spent by users in the applications selected in the study from our dataset (2.34 ± 1.53 [min]). If data was unavailable for a feature, we added one more minute (± 30 seconds before and after the current time window) until the data were found. During the reprocessing of our dataset, we filter the data collected via mQoL-Log to remove incomplete and erroneous data (e.g., an application usage session of ten hours).

We pick the networking features in our dataset. Therefore, we perform feature engineering to extract information from multiple features, e.g., knowing the IP address allows us to know if the application operates over IPv6 or IPv4. We extract the TCP states' distinct count for each app usage. The list of all network features used in the model is presented in Table 8.4. mQoL-Log recorded the network changes as they occurred, even during the application usage. Throughout the aggregation process of the network features, the potential handover information was encoded. Further, we aggregate the QoE rating with the battery state, the application session duration, the application name, and the task to fulfill the need of the participant in the application. Additional external factors influencing QoE are context-based. Hence, we select the features with high context (e.g., physical activity).

8.4.2 QoE/MOS Classification

Following the work of [Larson and Delespaul \(1992\)](#), we normalize the QoE rating values per each user. We use one-hot encoding on our categorical features (i.e., network type, application name, physical activity, and task) to prepare them for the classifiers. We follow a classic machine learning method by randomly selecting stratified 70% of our data as our training dataset. The resulting 30% is split into two to obtain our validation dataset (15%) and testing dataset (15%). We conduct a randomized search cross-validation ($cv = 20$) to optimize our model parameters. That means that the 70-15-15 split has been run 20 times by repeating the random selection of our training, validation, and testing dataset, hence covering our entire dataset. This is called "random permutation cross-validation (shuffle and split)" ([Arlot and Celisse, 2010](#)). The distribution of "High" and "Low" QoE ratings are preserved in the validation and testing dataset.

After each split, as presented in Figure 8.6 we apply SMOTE ([Chawla et al., 2002](#)) on the training dataset to overcome the imbalance issue via over-sampling. We perform downsampling in our pre-analysis. The models trained on smaller datasets are not able to generalize as the ones after SMOTE. The training dataset has 50% of "High" QoE labels ($n=3981$, no rating was lost), and 50% of "Low" QoE labels ($n=3981$); we gain 3702 "artificial" ratings. We scale our training dataset to remove the mean and scale to unit variance, as some classifiers (e.g., K-nearest neighbors) have issues with data of different unit sizes. This scaler is used on the validation and testing dataset.

Table 8.4 Network features as Collected via mQoL-Log During Application Usage

Features	Description	Type/Unit
is_connected	Connection status from Android OS	Boolean
Connection type	Network connection type (Figure 8.3)	Categorical
Wi-Fi level	Signal strength of connected Access Point	Float/dbm
Wi-Fi speed	Wi-Fi interface speed	Float/Mbps
Cell strength	Signal strength of connected cell tower	Float/dbm
Cellular down bandwidth	Cell downstream bandwidth	Float/Kbps
Cellular bandwidth	Cell upstream bandwidth	Float/Kbps
win_div_net	Aggregation window for network events around the application usage time	Int/minutes
rxt_packets_time	Packets received per second during win_div_net	Float/pps
txt_packets_time	Packets sent per second during win_div_net	Float/pps
rxt_bytes_time	Bytes received per second during win_div_net	Float/Bps
txt_bytes_time	Bytes sent per second during win_div_net	Float/Bps
RTT_{mean}	Mean Round-Trip Time of the 5 pings	Float/minutes
$RTT_{variance}$	variance Round-Trip Time of the 5 pings	Float/minutes
Netstats: TCP states count during win_div_net	LISTEN, SYN-SENT, SYN-RECEIVED, ESTABLISHED, FIN-WAIT-1, FIN-WAIT-2, CLOSE-WAIT, CLOSING, TIME-WAIT, CLOSED Postel (1981)	Int/Categorical

8.4.3 Classifier Selection

The process in Figure 8.6 is done in this section.

Table 8.5 QoE Prediction: Metrics on the Validation Dataset for Multiple Common Classifiers

Classifier	AUC	Accuracy [%]	Precision [%]	Recall [%]
k-Nearest Neighbours (KNN)	0.796±0.036	0.758±0.014	0.969±0.007	0.765±0.014
Decision Tree (DT)	0.612±0.056	0.913±0.016	0.952±0.01	0.955±0.012
Random Forest (RF)	0.758±0.029	0.908±0.014	0.955±0.01	0.946±0.01
Logistic Regression (LR)	0.775±0.024	0.823±0.008	0.967±0.005	0.839±0.01
Stochastic Gradient Descent (SGB)	0.773±0.027	0.84±0.006	0.964±0.006	0.86±0.005
Naive Bayes (NB)	0.685±0.019	0.087±0.012	0.956±0.039	0.021±0.005
Gradient Boosting (GB)	0.774±0.027	0.916±0.013	0.951±0.011	0.959±0.007
eXtreme Boosting Gradient (XBG)	0.816±0.017	0.931±0.01	0.952±0.009	0.975±0.006

In the first pass through our process in Figure 8.6, we investigated the most accurate classification algorithm for our goal. We ran a *candid* (non-optimized) 10x fold cross-validation on our training and validation dataset to select the algorithm with the best performance for our classification problem between: K-Nearest Neighbors (KNN), Decision tree (DT), Random Forest (RF), Logistic Regression (LR), Stochastic Gradient Descent (SGD), Naive Bayes (NB), Gradient Boosting (GB) and eXtreme Boosting Gradient (XBG).

We selected the Area Under the Curve (AUC) [0-1, no dimension] as the metrics to find out the more accurate classifier of our list. It expresses how accurate a model can distinguish between classes (e.g., classifying the classes correctly with minimum confusion). It measures the entire two-dimensional area underneath the receiver operating characteristic (ROC) curve. The ROC is a curve presented in a graph (x-axis: False Positive Rate, y-axis: True Positive Rate) presenting the classification model's performance. The "accuracy" is the fraction of predictions our models found out right, defined for binary classification as the sum of true positives and true negatives divided by the sum of true positives, true negatives, false positives, and false negatives. The "precision" is defined in % as the true positives divided by the sum of true positives and false positives. It is the proportion of the actual correct positive identification. The "recall" is defined in % as the true positives divided by the sum of true positives and false negatives.

Table 8.5 shows the AUC of our classifiers on the validation dataset. We find that the XGB classifier performs better in the AUC, accuracy and recall metrics. Hence, we select XGB as the base classifier to predict the most accurately "High" or "Low" QoE in this work. XBG has been the most accurate algorithm used in classification problems based on tabular datasets. Boosted tree algorithms, as XBG, have shown their performance on QoE prediction in the past work (Casas et al., 2017a).

We tested eight data filtering and wrangling scenarios (section 8.4.4 to section 8.4.11) assuming we can obtain better prediction results on our D38 dataset before applying SMOTE. The same classifier

is used in all scenarios. This step of the process is shown in Figure 8.6 in the block “Model”. We examined in section 8.5 the performance of all the models.

8.4.4 Candid Model (XBG)

We run the same machine learning method as described in Section 8.4.2, with the same features used in De Masi and Wac (2019). After the aggregation, we exploit the 6,086 ratings (D38) to train the model before SMOTE. We use this model as our referent model (*XBG*), to compare the models constructed with the filtered D38 dataset in the following scenarios. On the contrary, for the more accurate model presented in section 8.4.3, *XBG*'s hyperparameters are optimized. We present its performance in Table 8.5.

8.4.5 Filter Time to Reply (TR)

Like in section 8.4.4, we use the same features, but in this scenario, we filter our dataset D38. Specifically, we remove the ratings where the user did not answer to the EMA notification after a specific time threshold. We determine the threshold per a dataset, as the mean overall response time after the EMA was triggered. We remove 953 observations with a threshold at 14.85 minutes, leaving us with 5133 ratings with a similar distribution 93.51% of “High” QoE and 6.49% of “Low” QoE. Our hypothesis was as follows. Participants' ratings are influenced by the time difference between the app use and its rating (EMA). Hence by removing the ratings distant from the events, we anticipate achieving better performance for our model. We present the distribution of the time to reply per user in Table 8.6. The maximum times are so high as participants would reply in the morning to a notification from the past day.

8.4.6 Unlabeled Tasks (ULT)

As previously shared, 6% of the samples of D38 have unlabelled tasks (i.e., the unanswered question for “What action were you trying to accomplish?”) for a given application. We try to fill those samples with the most common user-selected task per application during the study for D38. We hypothesize that the influences of those samples could allow for a more accurate predicting model. The most common task per application are as follows: WhatsApp and Messenger are used to “READ” and “WRITE” messages. Spotify for consuming music. Chrome, Instagram, Maps, and Facebook were used to “CONSUME” different types of content. We attempt to derive a QoE model assuming that, as presented later in the results section.

8.4.7 Filter Features Aggregation Time (FA)

We propose to execute the same method used in section 8.4.5. However, instead of filtering the time delta between the application usage and its rating, we filter the dataset (D38) based on the time between the EMA and the times the aggregated features were generated (e.g., `win_div_net` feature for

Table 8.6 Times to Reply (TR) to EMAs Per User

User ID	Time to Reply mean \pm sem [min]	Minimum [min]	Maximum [min]
1	32.11 \pm 20.86	0.02	641.41
2	10.84 \pm 10.86	0.02	666.18
3	7.14 \pm 5.73	0.02	524.20
4	0.58 \pm 0.14	0.01	4.11
5	12.25 \pm 12.4	0.01	870.12
6	18.41 \pm 16.89	0.01	774.19
7	3.5 \pm 1.93	0.03	140.19
9	1.39 \pm 0.87	0.02	50.16
10	6.64 \pm 7.15	0.02	544.88
11	6.85 \pm 2.64	0.01	112.61
12	14.53 \pm 7.8	0.01	511.51
13	40.05 \pm 30.29	0.03	1282.35
14	3.66 \pm 2.7	0.01	203.57
15	21.85 \pm 12.6	0.02	774.97
16	32.86 \pm 31.29	0.01	1569.44
17	12.6 \pm 11.96	0.01	755.89
18	27.47 \pm 11.27	0.01	234.51
19	31.99 \pm 21.11	0.03	813.43
20	6.76 \pm 7.23	0.01	592.61
21	1.17 \pm 0.86	0.01	59.88
22	2.45 \pm 1.46	0.02	80.83
23	11.2 \pm 8.48	0.02	515.09
24	5.96 \pm 2.4	0.01	62.58
25	11.39 \pm 11.19	0.02	543.64
26	23.41 \pm 21.07	0.01	1006.80
27	28.51 \pm 22.69	0.03	1088.75
28	6.32 \pm 2.39	0.05	111.54
29	5.83 \pm 5.53	0.02	553.98
30	6.56 \pm 7.06	0.02	583.67
31	12.44 \pm 8.96	0.01	551.27
32	15.33 \pm 6.42	0.03	140.73
33	15.47 \pm 9.25	0.01	551.64
34	15.84 \pm 13.34	0.02	802.23
35	15.72 \pm 14.33	0.01	794.24
36	4.98 \pm 2.26	0.01	105.39
37	65.23 \pm 42.98	0.03	1732.94
38	13.13 \pm 12.54	0.01	768.73
D38	14.85 \pm 0.95	0.02 \pm 0.01	577.89 \pm 0.54

Table 8.7 Aggregation Features (FA) Threshold in Minutes for Each mQoL-Log service

mQoL-Log Collection services (Table 8.2)	FA_{38} [min]
Network	0.56 ± 0.50
Ping/RTT	3.87 ± 0.53
Battery	0.11 ± 0.03
Physical Activity	4.03 ± 0.22
Touches	0.56 ± 0.50

the network data collection service). As Android OS was terminating our data logger from time to time, and since not all users allowed the application to upload all collected data to completion, we remove the samples in which the features were collected too far off the rated event. We define the far-off threshold with a lower bound and higher bound as $\Delta t_{mean \pm std}$ for each data collection service, Δt is the time difference between the event and the time of collection for the data collection services as described in Table 8.2. Each threshold was applied to its corresponding data collection services from Table 8.2. It reduced the dataset to 4082 samples (2003 samples were removed). Table 8.7 shows the respective thresholds in minutes per data collection service. It represents the mean \pm standard error (sem). We do not test other cut-off thresholds.

8.4.8 Merged Filter Replies (TR) and Features Aggregation (FA) Time (TRFA)

We removed samples that matched two filters from two different blocks (aggregation and filtering) of our process pipeline (Figure 8.6): the feature time aggregation (FA) from section 8.7 and the time to the replied threshold (TR) as described in section 8.6. For both filters, the same previous thresholds are respectably applied. It reduced our dataset from 6,086 samples to 3,701 samples. The same features as before are used to train the model.

8.4.9 Meta-features Selection (MT)

From our previous modeling attempts (sections 8.4.4 - 8.4.8), we evaluate the most predictive features and only used this subset for training our model. We aggregate all the models generated by the 20 folds with a random grid search for each of our previous attempts: filtering based on the time difference between the event and the participants' annotation (TR), filtering the features based on time of collections (FA) and mapping the non-labeled task (section 8.4.6, ULT). We generate a total of 60 models ($3 * 20$ folds). We extract the importance of each feature for each model and compute the mean \pm std error for each. We select only the features with average importance higher than 1% (arbitrary threshold). The importance is available through the XGBoost library used to train the model. Table 8.8 contains the features and their importance. "Aggregated packet traffic stats" represent the aggregation of different features from the packet statistics (i.e., the features with the suffix "_times" in Table 8.4).

Table 8.8 Aggregated Features Importance from *TR*, *FA* and *ULT*

Feature Name Table 8.2 and 8.4	Importance Mean± sem [%]	On device Android Feasibility
Task	38.37±2.64	✓
Physical activity	9.03±0.53	✓
Application name	12.04±0.77	✓
Battery level	2.94±0.12	✓
Cell strength	4.70±0.12	✓
Network type	3.05±0.45	✓
is_connected	2.15±0.10	✓
Cellular down bandwidth	1.10±0.11	✓
Cellular Up bandwidth	1.75±0.12	✓
Aggregated packets traffic stats	1.20±0.10	
win_div_net	3.44±0.12	✓

8.4.10 Expectation (EX)

As we found a moderate correlation between the expectation and the perceived QoE (section 8.3.4), we propose to use this information as an additional feature in our new *EX* model. We use the same base features as presented in *XBG* (section 8.4.4). The expectation is based on past application sessions, e.g., prior knowledge about context and an event (i.e., “Low” QoE on WhatsApp when connected to the university Wi-Fi) (Sackl et al., 2017).

8.4.11 On-device Prediction (OD)

We selected the features that could be used to predict QoE directly on the device, transparently for the user. Some previously used features generated via feature engineering during aggregation (i.e., packet traffic stats) or duration of application usage are not information the Android OS application can obtain. Hence, we decided to base our features for this model on the ones from Table 8.8 and the features used to train our candid model.

At the time of this writing, Android OS version 10 has been released in November 2019. It includes new security measures. It is not possible to access the “netstat” command output as before, and the priority of background service execution has been modified. A new limitation was introduced with Android 10, the long-running background network services are restricted by the system. Hence, we can not use an active ping probe. We remove the features that could not be integrated into an on-board smartphone model. The on-board accessible features were as follows: battery level, user physical activity, application, task in the application, Android network manager “is_connected” attribute, network connection type, Wi-Fi level, Wi-Fi speed, cell strength, cell bandwidth up and downstream.

Table 8.9 summarized all features used in the previous scenarios and with the closest related work, i.e., building a smartphone application’s QoE prediction model (Casas et al., 2017b). As the *XBG*, *TR*,

FA, and *ULT* models share the same features but used several aggregations and filtering methods, we group them into one group G for figure clarity.

8.5 Results

In this section, we summarize our past results from [De Masi and Wac \(2019\)](#) and present the output from our new scenarios (sections [8.4.4](#) - [8.4.11](#)).

8.5.1 Overview of Previous Work

In our past work ([De Masi and Wac, 2019](#)), we had 33 participants and collected 5663 ratings. We build on QoE prediction model named XBG_{33} . Its performance metrics are as follows AUC of 0.8388 ± 0.279 and accuracy $0.939 \pm 0.007\%$. We derived the importance of the features in the most accurate XBG classifier and found that the duration of application usage, battery level, and QoS features, user's tasks to be accomplished are relevant (e.g., send text versus consuming content), as well as the user physical activity (e.g., walking) to predict QoE. The participant task, in the application itself is more important than the application used.

8.5.2 Scenarios' Results

We repeat the same process from section [8.4.2](#) for all our scenarios, only using the XBG algorithm. Hence, we train eight models with random hyperparameters search with 20 cross-validations, done 20 times to cover the full dataset.

We test the performance of our models on the test datasets. Our results are shown in [Table 8.10](#). It contains the two main metrics we selected (AUC and accuracy), as well as precision and recall. We want to validate our results statistically. We apply a pairwise t-test to the metrics, as they are normally distributed for all scenarios. The null hypothesis H_0 is as follows: there is no statistically significant difference between the scenarios' metrics. Considering that we are making multiple comparisons, we have to use the Bonferroni adjustment to select the correct cutoff to determine whether H_0 has to be strongly rejected. The base *alpha* is 0.05, and adjusted (for 36 comparisons) is $\alpha = 0.001$. If the p-values computed are inferior to α , H_0 is rejected. Hence, the difference between the scenarios' metrics is statistically significant.

The candid model XBG underperforms on both metrics against the results presented by [De Masi and Wac \(2019\)](#), which are negligible (p-values for both metrics superior to 0.1) with differences of 0.017 for AUC and 0.001 for accuracy. Filtering the time difference between the event and the participant rating time (i.e., scenario TR) allowed for more accurate models. The model TR scores higher on AUC and accuracy.

The filtering based on the feature time (FA) aggregation window data allowed for similar performance, FA 's AUC is 0.001 higher than XBG 's. The model created with both filters on the aggregated

Table 8.9 Model's Features Per Scenario

Perspectives	Features	G^1	MT	$TRFA$	EX	OD	Casas et al. (2017b)
Context	Physical activity	✓	✓	✓	✓	✓	
	Location						✓
User	Task	✓	✓	✓	✓	✓	
	Expectation				✓		
	Duration user session	✓	✓	✓			
System	Application name	✓	✓	✓	✓	✓	✓
	Battery level	✓	✓	✓	✓	✓	
	Cell strength	✓	✓	✓	✓	✓	✓
	Network type	✓	✓	✓	✓	✓	✓
	IP version	✓	✓	✓	✓		
	is_connected	✓	✓	✓	✓	✓	
	Cellular down bandwidth	✓	✓	✓	✓	✓	
	Cellular up bandwidth	✓	✓	✓	✓	✓	
	Handover	✓		✓	✓		
	Netstats	✓		✓	✓		
	Aggregated packets traffic stats	✓	✓		✓		✓
	RTT_{mean}	✓		✓	✓		
	RTT_{var}	✓		✓	✓		
	Wi-Fi level	✓		✓	✓	✓	
	Wi-Fi speed	✓		✓	✓	✓	
	win_div_net	✓	✓	✓	✓	✓	
	Cell ID						✓
	Cell Operator						✓
TCP flow ratio						✓	
Duration flow						✓	

Table 8.10 QoE Models Performance on Test Dataset for Each Scenario

Scenario	AUC	Accuracy [%]	Precision [%]	Recall
XBG_{33}	0.829 ± 0.028	0.939 ± 0.008	0.953 ± 0.007	0.984 ± 0.004
XBG	0.812 ± 0.033	0.938 ± 0.007	0.952 ± 0.007	0.984 ± 0.004
TR	0.83 ± 0.033	0.938 ± 0.007	0.953 ± 0.006	0.983 ± 0.004
ULT	0.723 ± 0.034	0.929 ± 0.009	0.94 ± 0.009	0.987 ± 0.003
FA	0.813 ± 0.038	0.924 ± 0.008	0.94 ± 0.006	0.98 ± 0.008
$TRFA$	0.822 ± 0.041	0.926 ± 0.011	0.944 ± 0.01	0.978 ± 0.007
MT	0.801 ± 0.03	0.931 ± 0.008	0.949 ± 0.008	0.979 ± 0.005
EX	0.874 ± 0.027	0.956 ± 0.007	0.967 ± 0.006	0.987 ± 0.004
OD	0.76 ± 0.037	0.925 ± 0.01	0.95 ± 0.006	0.971 ± 0.01

time and participants' reply time (scenario TRFA) $TRFA$ scored between TR and FA on AUC, and the same pattern repeated for accuracy.

The model created with the unlabeled tasks (ULT) filled by the most common task per app per user (ULT) performs the worst on the AUC metric. The meta-features selection filtering (MT) performance is only of 0.801 for AUC and 0.931 accuracy.

Table 8.8 shows the selected features and their importance. The “user accomplished task” is the most important feature. The on-device prediction model OD scored with 0.76 on AUC and 0.925 on accuracy. The “feasible” features selected for this model did not allow for higher performance.

Integration expectation (EX) in our model is beneficial to predict QoE, EX scores higher on all the metrics, with an AUC of 0.874 ± 0.027 , 0.956 ± 0.007 accuracy, 0.967 ± 0.006 for precision and a recall of 0.987 ± 0.004 . We compute the p-value for each metric compared with the candid scenario result (XBG), for all metrics except recall we found $p < \alpha$ ($p_{AUC} = 8.900e-08$, $p_{accuracy} = 8.049e-10$, $p_{precision} = 1.415e-08$ and $p_{recall} = 0.028$). Our last result showed that expectation is linked to QoE for interactive application, during a living lab study in-the-wild.

Overall, the better model to predict QoE is EX . We compare AUC, accuracy and precision metrics of EX to all the other models and found $p < \alpha$. For recall, the p-values from the comparisons with XBG_{33}, XBG, TR, ULT and FA are inferior to α .

8.6 Discussion

In this section, we discuss our findings from building QoE prediction models with several feature filtering and aggregation methods. First, we discuss the rating quality and their influence over our models (section 8.6.1), then we argue about our features choices and aggregation method (section 8.6.2). Then, we discuss the performances and the implementation of our on-device prediction model

¹ XBG, TR, FA , and ULT models

(section 8.6.3). Finally, we share our recommendations for smartphone application developers (section 8.6.4).

8.6.1 Ratings Quality

In light of our results, we saw that the annotators' rating quality varies from one to another. Users 1 and 8) always provided the same ratings, even if they were rating different applications, as seen by the distribution in Figure 8.5. User 25 provided only one rating of "Low" QoE. The previous assumption that those users' ratings could be discarded to obtain an increase in QoE model performance had been wrong, as showed by our attempt with the *TR* model. A way to solve this issue in our following studies would be to test the participants with fake-EMAs. We could ask them to rate a false application usage (e.g., with a wrong time or wrong application name) and observe if they communicate about the bogus questions.

Filtering the samples where the participants replied much later after the application usage occurred resulted in better models. The threshold for rating reliability (i.e., trust in the participant annotation) was an answer provided within 14.85 minutes. The rating's reliability is taken into account before selecting the participant's data for training a model. Convincing the participant to provide a rating just after the application usage is challenging.

8.6.2 Features Wranglings

The model *ULT* scores less than the other models, as we see the task accomplished in the application by the user is the most important feature for QoE prediction, as far as we are concerned. The importance of user tasks to accomplish is ranked first in Table 8.8. Hence, our method to retrieve the unlabeled task samples to train our model was wrong.

The *MT* model, trained with a reduced set of features, obtaining higher accuracy and AUC than *ULT*. The meta-features selection (section 8.4.9) based on the features' importance from the previous model shows that eXtreme Gradient Boosting could automatically select the essential features for building a model.

We compared our features with the ones from the related work (Casas et al., 2017b). We saw a higher focus on system-based features (i.e., QoS). Their models were solely trained on one application per model, on cellular networks, with high-precision QoS data. We used more features based on the user needs in the application and its context.

8.6.3 On-device Prediction

Machine learning prediction is often directly executed in the cloud. But an Internet connection is not always available depending on the user's context (e.g., mobile connection in a train tunnel is not available). Hence, we built an on-device prediction model to mitigate this context strain. The limited set of features available for making an on-device prediction model does not perform as well as the other models. The netstats command output was of high importance feature in our past models. The

knowledge of the current TCP session states and UDP flows made the models score better. Android has many APIs to query the network state, but none of them is fine-grained. We trained another on-device model with ratings provided in the same time-frame as the TR models' thresholds. We observed the same behavior as before. Namely, the models with filtered response have higher accuracy. In this case (e.g., scenario TR+OD), the model performs worse than the OD model with 0.782 ± 0.027 in AUC, 0.932 ± 0.008 accuracy and 0.952 ± 0.007 precision, but it obtains an increase of recall with 0.977 ± 0.006 .

On-device prediction resolve issues linked to data privacy, the input information does not leave the smartphone and the XBG model runs directly on the mobile devices. However, it has shortcomings. The device has to be powerful enough to handle a high number of predictions simultaneously when the phone is processing its normal workload (already running on-screen application and background services). It also consumes extra energy and processor time.

8.6.4 Recommendations for the Application Developers

The application developer should optimize their application to seamlessly handle "Low" QoE, depending on what the user wants to accomplish with the application. "Low" QoE ratings are higher when the user is "writing" and "tilting" between physical activity, hence with this information, the developer could provide a better way of inputting text in their messaging application (e.g., proposing predefined short answer from a half screen size touch area). What is essential for an application developer is that with better QoE, a user is more effective; spends less time on the application accomplishing the intended tasks faster, but potentially also uses more features in an application.

We propose three recommendations for the application developer. They firstly should constantly and accurately monitor the current user context. A change in physical activity, battery consumption, network type, or time spent in their application are a great indicator of QoE. Android OS API allows accessing those data in a simple way via APIs. Secondly, they should integrate a mitigation solution in the case of "Low" QoE. If a model as ours is complex to orchestrate, a cheaper solution for a heavy network application can be a simple ping to their server. If the main action in their application is impacted, they should provide real-time information to the user concerning the issue (e.g., notification to retry with a countdown). As we found out, the model built with expectation as a feature performed better (higher AUC, accuracy, recall, and precision); expectation plays a significant role in QoE. Thirdly, the application developer should use common design and usability patterns provided by the OS maker to optimize expectations and, by doing so, QoE.

8.6.5 Modelling Highlights

Quality and the quantity of data is vital in obtaining a representative model. Our study focused on modeling the Quality of Experience of smartphone users, with their provided ground truth and their smartphone's data. Overall, the data collection tools have to be tested under multiple contexts to limit the loss of data caused by network instability and participant environment. Once the data are acquired, their quality has to be controlled. It is evident from the models created with reliable data

e.g., given the higher availability of the ground truth, one can obtain higher scoring models. Hence, during modeling, the features' selection, data wrangling, and aggregation steps must be carefully executed to limit model building constrains. The human aspects, such as the user expectation, need to be conscientiously included in the experiment design and the later data analysis. For example, the memory from past application used experiences could create bias when the participant assesses its momentary experience. Thus, following those recommendations permit the creation of QoE models from in-the-wild studies data.

8.7 Study Limitations

We consider the following study limitations. First, related to the devices used. This study was only possible on Android OS devices, as data collection is more difficult on iOS. We cannot thus generalize our findings for another operating system platform. Additionally, the collection of the number of frames dropped by an application's UI would have been a plus to understand the hardware status. New security protection and updated background service execution policy are problematic for data collection without root access. The policy occasionally killed our mQoL-Log collection services to reduce the energy consumption on the smartphone. We estimate we have lost 3.5% of valuable data as a result.

The second limitation was in our choice of applications. We did not include high bandwidth need, which was studied by other (Casas et al., 2018), particularly video QoE consumption on smartphones. The landscape of smartphone applications is evolving each day with new innovative services, modeling QoE for each new application, and their underlying features' are not a scalable method. Hence, we tried to generalize QoE prediction based on user action within an application. The limitation is that the user's momentary emotion and stress level can influence the annotation of their application usage, as well as a participant, can rate a "High" QoE application usage negatively because of the content of the application. The participants were told to avoid this effect, but then it could still influence our models. Smartphone operating system (OS) makers created APIs to obtain the user's context to allow application developers to write immersive "smart" applications. We use those APIs to gather the participant's context. Hence, we trust the data validity provided by the OS. Another limitation is the EMA's questions. They could leave room for interpretation. Hence, they should be updated for our next study to remove this undesirable effect. Furthermore, the dataset collected does not contain a representative population. The uneven age group distribution of our study participants is a limitation of this work's representativeness. We could not make any conclusions based on demographic information. The presented use case, as well as the tools leveraged to collect the data, the smartphone hardware, and the set of participants are specific to our study. As the reproducibility of our results can be challenged, the presented path towards building smartphone application's QoE models' is a first step toward accurate models. Overall, this and similar in-the-wild studies are prone to such limitations, and the number of participants (e.g., implying higher cost per a participant), the study's

duration (i.e., much longer and intrusive than in-the-lab study) and the survey (EMA) respondent fatigue may have further impacted our results.

8.8 Conclusions and Future Work Areas

In this paper, we presented an attempt to model and predict smartphone application QoE from a living lab study, with 38 participants for four weeks. We showed that collecting in-situ QoE rating and collecting smartphone background data enables us to use common machine learning techniques to build an accurate predictive model for “High” and “Low” QoE. We investigated multiple data filtering scenarios that generated more accurate models in different scenarios. The data preparation (e.g., filtering and aggregation) allowed an improvement in our QoE models’ performance. The filtering of the participant QoE ratings was overall beneficial to the models. Namely, the models were performing better when trained on ratings provided closer to the application usage time. We investigated the factors influencing QoE in our dataset. Our results showed that rating the application usage session, just after the usage, permitted more reliable models. The task to accomplish with the application by the user (i.e, user’s intent), and the application itself are important factors, testifying on the difficulty of generalization for this type of all-application QoE model, contrary to the per-app QoE model. We determined that application developers should have user expectations in mind when designing an application. We found expectation based QoE models to perform better. The mobile operating systems and their applications are more than ten-years-old. Their users now have a high expectation of how the application and the system will behave. We extended our work to the challenging domain of on-device prediction models, its difficulty, and its performance. Overall, our hybrid qualitative and quantitative method performed accurately to model QoE. In the future, we plan to implement a production-ready pre-trained prediction model integrating more features inside our Android application as the user’s position (on-device only). The application will predict if, in the near future (e.g., 5 minutes), the current QoE application usage session will be “High” or “Low”. If the prediction shifts because of the context (e.g., train inside a tunnel), the application can inform the user and prepare itself for the change. Those predictions, rated by the user, would allow us to use reinforcement learning to enhance our model comparable to recommendation systems. We also plan to integrate other factors influencing living lab study and potentially the collected data quality: the other aspects of the user’s context (e.g., mental state), previous experience, surroundings, operating system updates, and newly available features.

Acknowledgment

The authors thank the study’s participants and funding agencies SNSF MIQmodel (157003), AAL GUARDIAN (6-120-CP) and H2020 WellCo (769765).

Chapter 9

Article VII: Less Annoying: Quality of Experience of Commonly Used Mobile Applications

Published in Proceedings of the 13th ACM multimedia systems conference (MMSys), June 2022. doi: 10.1145/3524273.3528183.

Chapter Contents

9.1	Introduction	146
9.2	Related Work	148
9.3	Methodology	149
9.3.1	Study Protocols (S1, S2)	149
9.3.2	Modeling QoE (S1)	151
9.4	ExpectQoE Evaluation and User Study Results (S2)	152
9.4.1	Demographic: S2	152
9.4.2	Factors Influencing QoE	153
9.4.3	ExpectQoE Model Evaluation Dataset	155
9.4.4	QoE Model Performance (T2)	156
9.4.5	ExpectQoE Ratings Analysis	156
9.4.6	ExpectQoE Notification Effectiveness	156
9.4.7	Expectation Impact	158
9.4.8	Impact of ExpectQoE on Application Usage Duration	158
9.4.9	ExpectQoE Model: QoE Levels and Features (S2)	159
9.5	Discussion and Limitations	159
9.6	Conclusions	161

Abstract

In recent years, research on the Quality of Experience (QoE) of smartphone applications has received attention from both industry and academia due to the complexity of quantifying and managing it. This paper proposes a smartphone-embedded system able to quantify and notify smartphone users of the expected QoE level (high or low) during their interaction with their devices. We conducted two in-the-wild studies for four weeks each with Android smartphones users. The first study enabled the collection of the QoE levels of popular smartphone applications' usage rated by 38 users. We aimed to derive an understanding of users' QoE level. From this dataset, we also built our own model that predicts the QoE level for application category. Existing QoE models lack contextual features, such as duration of the user interaction with an application and the user's current physical activity. Subsequently, we implemented our model in an Android application (called expectQoE) for a second study involving 30 users to maximize high QoE level, and we replicated a previous study (2012) on the factors influencing the QoE of commonly used applications. The expectQoE, through emoji-based notifications, presents the expected application category QoE level. This information enable the user's to make a conscious choice about the application to launch. We then investigated whether if expectQoE improved the user's perceived QoE level and affected their application usage. The results showed no conclusive user-reported improvement of their perceived QoE due to expectQoE. Although the participants always had high QoE application usage expectations, the variation in their expectations was minimal and not significant. However, based on a time series analysis of the quantitative data, we observed that expectQoE decreased the application usage duration. Finally, the factors influencing the QoE on smartphone applications were similar to the 2012 findings (e.i., the factors are similar but their meaning as evolve). However, we observed the emergence of digital wellbeing features as facets of the users' lifestyle choices.

9.1 Introduction

Smartphones are an integral part of modern life. They enable users to access online services to communicate and exchange information around the world. They allow them to create or consume content in different contexts; however, the experience can be impacted by the smartphone user's context. This context often changes due to circumstances such as the physical activity of the user, varying the smartphone application user experience as a result (De Masi and Wac, 2019). As such, the term Quality of Experience (QoE) was coined to mirror the known Quality of Service (QoS, (Telecommunication Union)) concept from the telecommunication and networking domains. Whereas the QoS only focuses on quantitative information. The QoE measurement is an expansion of the QoS and includes qualitative information related to the experience itself, and thus prioritizes the end-user. QoS focused on metrics obtained on user-end device and networking device (e.g., jitter, amount of packet error and dropped) which transport content. Contrary to QoE, which focuses on the experience encompassed in

the content. The QoE is defined by the Qualinet White Paper (Le Callet et al., 2012) as “an application or service user’s degree of delight or annoyance”.

Many previous works (Casas et al., 2016; Chen et al., 2014; Fiedler et al., 2010) have only focused on quantifying the smartphone applications and web browsing QoE based on QoS metrics within laboratory settings where the authors simulated external factors (e.g., a bandwidth limitation or reduced video bitrate), missing important contextual factors such as user’s habits and current activity. To bridge this gap, we first aimed to quantify the QoE of smartphone applications. As such, we employed a mixed-methods approach and collected application usage QoE ratings via an in-the-wild study (S1) with 38 participants over four weeks. During this study, we deployed an Android logger application, named mQoL-Lab (Berrocal et al., 2020b), that collected context information. The users had the opportunity to rate their application usage through Ecological Momentary Assessment (EMA; (Stone and Shiffman, 1994)). From the collected dataset, we built a QoE classification model that predicts the application category QoE level between two labels: high or low. The labels correspond to the level of acceptability from the end-user perspective (Schatz et al., 2011). The QoE classification model is based on features from three different perspectives: user (e.g., intent to accomplish), system (e.g., QoS metrics) and context (e.g., user physical activity). Related previous works only went as far as estimating the QoE level of video or social media applications on smartphones (Casas et al., 2017b), to help telecommunication providers offer better network conditions through core network parameters. This information was often processed after the study and thus retained from the end-users.

Moreover, the telecommunication providers continuous upgrade of network equipment leading to a better user experience will always be limited by the users’ current smartphone hardware, the network protocols, and the physics of wireless broadband. There are techniques to reduce user annoyance based on hardware technology upgrades or architectural system design (e.g., microservices and edge computing). However, these techniques could fail when the service provider and the network are inaccessible, or when the user’s intent is unpredictable and requires real-time access to the content (i.e., the content is impossible to cache). Hence, software approaches based on HCI technics could be a potential solution which have not been employed to grand extend yet. One of the solution in place is the indicator on smartphone which always presents the user’s network state. Accordingly, users expect internet-enabled applications to be slow when no bars are shown.

We conducted our study with the hypothesis that an intervention approach could influence smartphone users’ behaviour. The users would attempt to avoid annoying experiences and limit their application usage duration. Previous studies have shown that notifications are capable of influencing smartphone users by communicating information about the intervention topic (Mathur et al., 2016; Mehrotra et al., 2016; Pielot et al., 2017) (e.g. reducing exposure to low interest notification or pushing user to engage less in certain games). Thus, we approached maximizing users’ QoE by providing notifications that aimed at limiting their exposure to applications with a low predicted QoE level. We implemented our QoE level prediction model (expectQoE) into our mQoL-Lab and conducted a second four-week study (S2) with 30 participants to investigate the model’s influence with notifications to present to the user the predicted QoE level.

Besides QoE models, we focus on understanding the factors influencing users' experiences as they vary through context and time. The 2012 study of [Ickin et al. \(2012\)](#) on this subject identified these factors through a user study. Hence, we replicated part of their work for S2. We focused on previously defined factors influencing the QoE of smartphones and the factors current evolution.

9.2 Related Work

Assessing users' perceived experience of smartphone applications in-the-wild has been performed by [Casas et al. \(2015c, 2016\)](#). The researchers focused on the QoE of smartphone applications in cellular networks. They labeled their data in the field, but their participants were instructed to accomplish a specific task. Such study design can impact participants' annotation process. In a later work, the authors modeled smartphone application QoE ([Casas et al., 2017a](#)) with success (95% accurate), although they did not deploy or test their models outside a laboratory setting. Furthermore, they limited their focus to video and audio streaming only.

[Schwind et al. \(2020\)](#) used the MONROE ([Alay et al., 2018](#)) hardware platform to collect network and audio and video streaming metadata from online services on public transport (trains and buses) nodes in European countries. The study focused on video streaming and modeling the QoE based on video streaming bitrates. However, this work did not represent real user and smartphone interactions but only the results of multiple network tests in different contexts (i.e., mobility induced cell tower changes). Other researchers attempted to measure and predict network quality on trains ([Kaup et al., 2017](#)) by measuring QoS metrics. However, they did not factor the user's context, the application used, or previous user's experience into their models. Moreover, their model was based on a dataset collected on a specific train ride. Summarizing, the previous listed works focused on QoE estimation only.

We propose to use intervention to limit the smartphone users' exposure to annoying application experience. Interventions on smartphones are primarily present in health-based research. For instance, a smartphone-based intervention showed success in motivating physical activity in a student population ([Muntaner-Mas et al., 2021](#)), to change the participants' behavior through notifications ([Morrison et al., 2017](#)). Notifications are considered as an intervention tool due to their unexpected nature and informational content ([Mehrotra et al., 2016](#)). Additionally, smartphone-based interventions studies have employed notification to promote digital wellbeing ([Monge Roffarello and De Russis, 2019](#); [Vanden Abeele, 2021](#)) with some success. While QoE modeling has been researched for smartphone applications, maximizing the QoE through notifications has not. Our work addresses this research gap, offering a solution via a notification to reduce user annoyance preemptively.

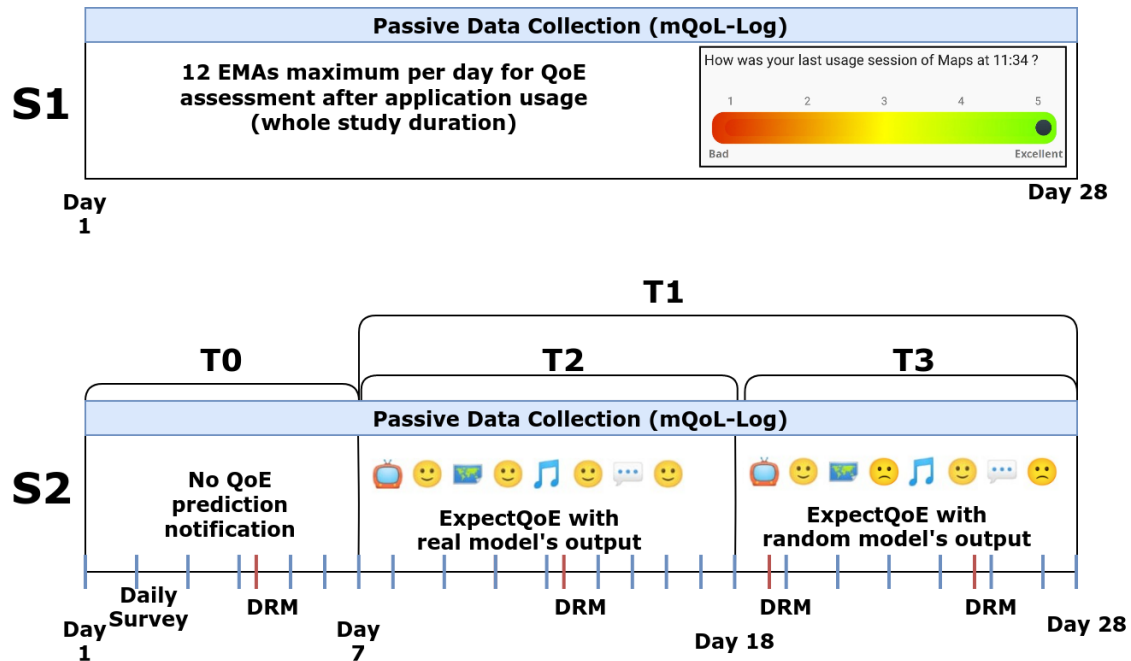


Figure 9.1 Study S1 and S2 Timeline and Research Methods

9.3 Methodology

9.3.1 Study Protocols (S1, S2)

Figure 9.1 depicts our approach and the study protocol for S1 (2018) and S2 (2021). As defined, S1 was for building the QoE model, while S2 was for model validation. The participant's information is presented respectively in Section 9.3.2.1 for S1 and Section 9.4.1 for S2. The participants were recruited on the university's campus via flyers, mailing lists, and social network posts. We focused on Android smartphone users who had lived at least five years in the Great Geneva region, at the border of France and Switzerland. Participants were required to use a minimum of four target applications daily, and a maximum of all, and have the most recent OS version on their smartphones. We focused on popular smartphone applications: Instagram, WhatsApp, Spotify, Facebook, Chrome, Facebook Messenger, and Google Maps. These applications were selected due to their high installation number from the Google Play Store and their previous selection in past work (Casas et al., 2016). Once selected, the participants were invited to our laboratory for a demonstration of the logger and of the tasks to accomplish during the studies. Each participant gave consent before enrolling in the study and downloading the mQoL-Lab application. This Android logger was employed in both studies to collect passive information on the user context. Tables 9.1 presents the data collected during S1 and the data used for estimating the QoE level during S2. Since S2 was more recent (2021 versus 2018), the availability of the data had changed. Previously available information like low-level network statistics were removed. mQoL-Lab also enables the collection of smartphone usage QoE level ratings (i.e., via EMAs) in situ (De Masi and Wac, 2018). In S1, the participants annotated the QoE level of their last ap-

plication usage through a 5-point Mean Opinion Score (MOS) (ITU-T Recommendation P.800.1, 2019). This method was previously used in QoE smartphone studies (Casas et al., 2015c; Ickin et al., 2012; Schatz and Egger, 2011). An EMA was triggered by an application events (i.e., closing or switching). The question “How was your last usage session of {‘app name’} at {‘time’}?” enabled the participants to know what application they were rating. The MOS scale contained the followings scores: poor (1), bad (2), fair (3), good (4), and excellent (5). The EMA also contained multiple choice questions about the intent the user wished to accomplish with the application: consume content, share or create content, read text message, write text message, control an app (e.g., start or stop music), video call, or audio call. We limited the number of EMAs to 12 per day during waking hours between 7:00 and 21:00, and included a 20-minute timeout between consecutive EMAs to reduce the study burden. We categorized the obtained data in the studies into three perspectives (Table 9.1) (i) User centric: qualitative data obtained from the user; (ii) System centric: quantitative data obtained from smartphone sensors linked to the smartphone hardware and software state (e.g., network QoS); (iii) Context centric: quantitative data obtained from smartphone sensors and characterized by a strong in-situ nature.

Table 9.1 Information Collected per Study (mQoL-Lab)

Perspective	Domain	Raw Features Available	S1	S2
User	Intent	Intent the user tried to accomplish by launching an application	✓	
	Application	Name of the application launched by the user	✓	✓
	Session	Duration of the application usage session (ms)	✓	
System	Network	Cell and Wi-Fi signal strength, Wi-Fi speed, Cellular up and down bandwidth, Active ping test to measure the round-trip time to the University server, handover, IP version, aggregated traffic packet statistics	✓	✓
		Netstats (i.e., TCP states per socket)	✓	
		Battery	Energy level from the battery (capacity in %)	✓
Context	Physical Activity	User’s physical activity from Google Activity Recognition (walking, running, still, on bicycle, in vehicle)	✓	✓

The protocol for S2 differed, as shown in Figure 9.1. S2 timeline is composed of three distinct periods, T0 was the baseline period, it contains the participants’ application usage habits (i.e., passive collection). Then T1 marked the beginning of the intervention period with expectQoE. T1 was composed of two periods, T2 in which the participants were notified with the real output of our QoE model. Contrary to T3, in which the participants received random QoE level. The periods in the S2 timeline enable us to capture the supposed influence of expectQoE on the participant’s application usage. T0 represents the baseline data. T1 focus on the influence of expectQoE across T2 (real model output) and T3 (random output). The predicted QoE levels in T2 were estimated using our model built from the data collected in S1. S2 study also included three types of questionnaires (Ickin et al., 2012): weekly Day Reconstruction Method (DRM) (Kahneman et al., 2004), semi-structured interviews, daily online QoE surveys at the end of the day at a random time between 19:30 and 20:00, and notifications

that presented the expectQoE level of the four application categories which the participants could rate (12 maximum per day from 8:00 to 21:00, 5 minutes minimum between two notifications). The question was asked as follows: “Did your application usage sessions meet your expectations?”. A slider was used to answer from 1 (not at all) to 5 (enormously). The weekly remote DRM interviews were conducted to discuss their previous 24 hours of smartphone usage and any other recent annoying experience on participants’ smartphones. With the daily surveys, we queried the participants about their overall QoE, expectations, stress level, and usefulness of the expectQoE system. We used an MOS scale from 1 to 5 for each indicator except for the stress level, which used a scale from 0 to 10 (Berrocal et al., 2020a). For S2 study, we developed a notification-based EMA. The expectQoE presented the application categories and predicted QoE level via emojis. Emojis surrogate plain text (Lu et al., 2016) and have been heavily used to generate notification-to-application interactions (Tauch and Kanjo, 2016). We mapped each application category to a specific emoji: communication and social 🗨️, music and audio 🎵, video 📺, and travel and local 🌍. In notification buttons were used to provide feedback with the thumbs-up 👍 and thumbs-down 👎 emoji. The QoE level was indicated via the slightly-smiling-face emoji 😊 for a predicted high-level QoE and the worried emoji 😟 for a predicted low-level QoE. The expectQoE notification content contained the dyad “category, predicted QoE” concatenated for all the categories in an emoji sequence (e.g., Figure 9.1). To preemptively limit participants’ fatigue from seeing the same content in the notification area, we randomized the order of the four dyads. The notification was triggered after an application usage started and disappeared once the user rated it. The emojis were not updated due to a limitation from the Android notification’s nature. Each update would have created a new notification which could have visually disturbed the participant. Moreover, they were only available beginning Day 7 (T2) of the study. Furthermore, after Day 18 (T3) of the study, the QoE levels presented were randomized, thus enabling us to test the impact of our model (T2) versus a balanced distribution of random QoE level as output (T3, assumed 50%/50% for low/high QoE).

9.3.2 Modeling QoE (S1)

9.3.2.1 S1 Dataset

In S1, the age distribution of the 38 participants is as follows. Thirteen were young adults (two between 18-20 y.o. and eleven between 21-29 y.o.), followed by ten participants between 30-39 y.o., two between 40 and 49 y.o., two participants between 50-59 y.o. and two non-disclosed. The gender distribution is as follows: fifteen were women, twenty-one men and one non-disclosed. Furthermore, the participants’ education level (i.e., last successful diploma obtained) was as follow, and five participants had a PhD degree, followed by fifteen participants who had a master’s degree, then four had a bachelor’s degree, twelve had a high school diploma or equivalent, and finally two participants had no diploma.

We collected 6,308 ratings (166 ± 89 per participant) of application usage QoE. Only five participants triggered the maximum possible number of EMAs. At the end of the study, the participants answered $75 \pm 2\%$ of the triggered EMAs. In general, the participants rated their QoE as good (MOS

> 4). The QoE ratings were mapped into two groups: high and low in accordance with Schatz et al.'s user accessibility threshold (Schatz et al., 2011), and due to the ratings' imbalance. More than two categories would lead to an even more imbalanced dataset than binary setting, impacting the model performances. Also, the choice of the threshold's values between the categories would have to be validated. The binary classification approach enables the construction of a robust model metrics wise (i.e., AUC). Ratings higher than or equal to 3.5 were classified as high; all other ratings were classified as low. The prevalence of high QoE levels was 93.5% versus 6.5% for low QoE levels.

9.3.2.2 ExpectQoE Features

We identified features to build a QoE-level prediction model based on the data collected during S1. Contrary to previous works, we required that the model had to make predictions directly on the devices. Accordingly, some aggregated information was inaccessible due to time constraints (e.g. time-based aggregated feature: application usage duration). We aimed at classifying the QoE levels of the following application categories: communication (e.g., WhatsApp and Facebook Messenger) and social (e.g., Instagram, Twitter, and Facebook), music and audio (e.g., Spotify), video (e.g., YouTube and Netflix), and travel and local (e.g., Google Maps). These categories represent more than 60% of the applications launched on smartphones (Böhmer et al., 2011). We selected other features that were accessible on-the-fly on Android 12. The selected features were presented in Table 9.1 column labeled S2.

9.3.2.3 ExpectQoE Building

The model we applied follows the on-device model construction presented by De Masi and Wac (2020). Our model could be subject to overfitting since the S1 dataset contains a higher amount of high-level QoE annotation than low. Hence, we undersampled the S1 dataset, resulting in maintaining 386 samples for each class. We split the dataset into a training set (70%) and a testing set (30%). We applied ten-fold cross-validation and trained the model with the XGBoost algorithm (Chen and Guestrin, 2016), which has been proven to perform efficiently with tabular data. We repeated the same process 10 times. During each fold, the undersampling selected different samples from the majority class. We obtained an average Area Under the Curve (AUC) $75 \pm 7\%$ (higher is better) on the test dataset to classify the QoE level (high/low). We exported the model with the highest AUC (82%) into our logger application for S2.

9.4 ExpectQoE Evaluation and User Study Results (S2)




9.4.1 Demographic: S2

The age distribution of the 30 participants is as follows. Four were young adults (two between 18-20 y.o. and 11 between 21-29 y.o.), followed by eleven participants between 30-39 y.o., one between 40 and 49 y.o., one participant between 50-59 y.o. and one non-disclosed. The gender distribution

is as follows. Nine were women, followed by nineteen men and two non-disclosed. Furthermore, the participants' education level is as follows: four participants had a PhD degree, followed by nine participants who had a master's degree, then seven had a bachelor's degree. Finally, seven participants had a high school diploma or equivalent, and only one participant did not have any degree. Two participants chose to not answers this question

9.4.2 Factors Influencing QoE

Table 9.2 S2: Participants Ratings to ExpectQoE Predictions

Participant	QoE Level	Response	Answered	Triggered			
ID	T0 [%] High/Low	T1 Rate [%]	T1 [n]	T1 [n]	T1 [%] S2 Total [%]	T2 [%] Model [%]	T3 [%] Random [%]
0	0.66/0.34	86	19	22	0.26/0.74	0.25/0.75	0.27/0.73
1	0.75/0.25	80	202	252	0.49/0.51	0.76/0.24	0.19/0.81
2	0.77/0.23	48	121	252	0.66/0.34	0.62/0.38	0.89/0.11
3	0.96/0.04	48	122	252	0.48/0.52	0.38/0.62	0.63/0.37
4	0.82/0.18	69	24	35	1.00/0.00	1.00/0.00	0.00/0.00
5	0.98/0.02	79	200	252	0.99/0.01	0.99/0.01	0.00/0.00
6	0.89/0.11	73	185	252	0.53/0.47	0.54/0.46	0.52/0.48
7	0.86/0.14	69	173	252	0.32/0.68	0.74/0.26	0.10/0.90
8	0.89/0.11	87	219	252	0.57/0.43	0.57/0.43	0.57/0.43
9	0.69/0.31	90	226	252	0.49/0.51	0.35/0.65	0.66/0.34
10	0.92/0.08	98	118	120	0.98/0.02	1.00/0.00	0.98/0.02
11	0.92/0.08	87	219	252	0.97/0.03	0.96/0.04	0.97/0.03
12	0.99/0.01	62	155	252	0.56/0.44	0.70/0.30	0.44/0.56
13	0.95/0.05	77	59	77	0.49/0.51	0.56/0.44	0.40/0.60
14	0.96/0.04	99	128	129	0.23/0.77	0.35/0.65	0.19/0.81
15	0.90/0.10	60	131	220	0.43/0.57	0.79/0.21	0.06/0.94
16	0.91/0.09	100	252	252	0.98/0.02	0.98/0.02	0.99/0.01
17	0.88/0.12	41	104	252	0.99/0.01	1.00/0.00	0.98/0.02
18	0.89/0.11	54	135	252	0.72/0.28	0.72/0.28	0.00/0.00
19	1.00/0.00	92	88	96	0.9/0.1	0.89/0.11	0.91/0.09
20	0.9/0.1	78	197	252	0.59/0.41	0.57/0.43	0.64/0.36
21	0.92/0.08	98	135	138	0.16/0.84	0.19/0.81	0.11/0.89
22	0.9/0.1	52	92	178	0.36/0.64	0.48/0.52	0.12/0.88
23	0.83/0.17	100	252	252	0.56/0.44	0.54/0.46	0.60/0.40
24	0.97/0.03	86	96	111	0.45/0.55	0.60/0.40	0.34/0.66
25	0.82/0.18	60	152	252	0.47/0.53	0.77/0.23	0.32/0.68
26	1.0/0.0	97	68	70	0.96/0.04	0.92/0.08	1.00/0.00
27	0.92/0.08	98	52	53	0.81/0.19	0.84/0.16	0.75/0.25
28	0.9/0.10	65	165	252	0.83/0.17	0.88/0.12	0.78/0.22
29	0.92/0.08	86.0	217.0	252.0	0.18/0.82	0.14/0.86	0.23/0.77
ALL	0.88/0.11	77±18	144±65	193±82	0.61/0.29	0.67/0.33	0.48/0.52

In 2012, [Ickin et al. \(2012\)](#) ran a study with 29 Android smartphone users for four weeks focusing on understanding smartphone QoE. The authors employed the DRM method to analyze the relations and causality between QoE annotations collected during the study, QoS, and context. Two independent coders clustered the terms with the most affinity. In the end, they distinguished seven factors influencing QoE. We used the seven factors as a template during our S2 weekly interviews. Overall, we collected 120 expressions from the 30 participants. Two researchers familiar with the QoE and

smartphones domain coded the expressions, and the measure of agreement was greater than 96%. Overall, we found similarities with the past work, yet the factor meanings have changed with time.

9.4.2.1 Application interface design

The application's interface design was commented on often. The participants enjoyed the interface of the notification bar. Contrary to [Ickin et al. \(2012\)](#), the participants complained about the content of the applications versus their mood at that time (e.g., announcement of the death of a family member via an application).

9.4.2.2 Application performance

Twenty-four participants reported problems with sharing photos and streaming videos for example: "the videos were not loading" (P22), "it's problematic, the connection is bad, on YouTube I have to wait a lot for a video to load" (P26). The participants were also conscious of the capacity of the network to which they were connected. In particular, twenty-three participants commented on roaming between countries and the time needed for their smartphone to connect to a new network. One participant experienced low QoE due to network roaming problems (P15). Two participants had to set up the cell network manually due to their proximity to a foreign cell tower (at the border). Only three participants reported playing video games on their devices. Overall, the participants were able to discern whether the performance of an application was due to the application itself or to an underlying network problem. That is different from the 2012 study, where such distinction was not made by the participants.

9.4.2.3 Battery

The batteries were able to sustain the smartphones for more than a day with high utilization from the participants. However, one participant reported carrying an extra battery when travelling in another country as their smartphone was used to guide their group and thus consumed more energy due to using GPS (P26).

9.4.2.4 Phone features

The participants reported enjoying the camera quality. Four participants used the hotspot function to share through their 4G connection with their friends or other devices (e.g., laptop or game console) when their home Internet delivered low QoE.

9.4.2.5 Applications and data connectivity cost

More than half the participants (17) mentioned having an unlimited mobile data subscription, hence their use of the hotspot feature. Overall, they were satisfied with the free applications available. However, four participants paid for application subscriptions that enhanced the application features.

Contrary to [Ickin et al. \(2012\)](#), when this services were not available on smartphone, seven participants subscribed to multiple streaming services. Overall, we found that Spotify (11 participants) and Netflix (11 participants) were among the most used services.

9.4.2.6 Routine

Twenty-five participants reported following an identical routine in the morning and in the evening. In both cases, they used a set of applications, often communication (e.g., WhatsApp, email), before starting or finishing their day, which corresponds to the findings of [Ickin et al. \(2012\)](#) regarding user routines.

9.4.2.7 Lifestyle

We observed a trend in the lifestyle factor that was not seen in 2012 ([Ickin et al., 2012](#)). Namely, participants limited their interaction with smartphones during work and at night. We identified four levels of this digital wellbeing behavior: (i) Smartphone physically inaccessible (one participant): The device is placed outside the bedroom at night, limiting accessibility; (ii) Plane mode (three participants): All access to any network is disabled; (iii) Data network off (four participants): The Wi-Fi and cell data access are turned off; the smartphone user can still receive a call, but Internet applications are unavailable; (iv) Smartphone enabled (seven participants): The “Do not disturb” mode stops all notifications from appearing on the screen, removes its signaling modalities (vibration and sound), and limits application usage based on time: “I put the phone on silence when I arrive at work” (P20). The applications run and are synchronized in the background. Additionally, contrary to past work, we found that many participants used more diverse applications that supported their lifestyle such as finance, spirituality, and health applications.

9.4.3 ExpectQoE Model Evaluation Dataset

Overall, we collected 64,179 application usage sessions from the thirty participants with 464 unique applications. The participants used mostly communication applications throughout the study which corresponds to the findings of [Böhmer et al. \(2011\)](#) at the beginning of the smartphone revolution. The ten most launched applications by the participants were WhatsApp (13.7%), Instagram (7.4%), Chrome (6.1%), Telegram (6%), Snapchat (3.8%), Gmail (3.7%), Phone dialer (2.9%), YouTube (2.9%), Message (SMS and MMS, 2.4%) and Facebook (2.3%). These applications represent 52% of the total launch application during S2. The top ten applications in which participants spend the most time correspond partially to the ten most launched applications [Böhmer et al. \(2011\)](#).

The Table 9.2 summarized the participants interactions with expectQoE in T1. Overall, they triggered on average 193 ± 82 expectQoE notifications due to their application usage. Only twenty participants used their smartphones enough to trigger the maximum amount of expectQoE notifications (252) throughout the study, impacting the amount of rating collected. However, the participants rated expectQoE on average 144 ± 65 times. The categories for which expectQoE provided the QoE level

prediction in T2 and T3 were communication and social, music and audio, travel and local and video player. Overall, these categories represent on average $65 \pm 15\%$ of the total application usage of S2 participants (T0+T1).

We examined the distribution of the expectQoE predictions for the real model output (T2) and the random model (T3). Overall, we observed that during the random model period (T3), the high and low QoE-level predictions were equally distributed among the participants (high QoE level: $49 \pm 5\%$; low QoE level: $51 \pm 5\%$). However, when the real model predicted the QoE (T2), the standard deviation was much higher (high QoE level: $48 \pm 24\%$; low QoE level: $52 \pm 24\%$), indicating a high variation in QoE level for the participants during T2. As such, eleven participants had a low QoE level distribution higher than 60%, contrary to only eight participants with a high QoE level with the same threshold.

9.4.4 QoE Model Performance (T2)

To investigate the validity of the QoE model predictions, we compare the daily reported QoE from the participants against the model's aggregated output per day in T2.

We employ the Kolmogorov-Smirnov test (Berger and Zhou, 2014), a non-parametric and distribution-free test. We found that for twenty participants the predicted QoE distribution and reported daily QoE were similar ($p < 0.04$). Indicating that over the day, the model prediction is consistent with the real the participants' feedback. However, for ten participants these distributions are not statistically significant ($p > 0.7$); the model output does not match their experience.

9.4.5 ExpectQoE Ratings Analysis

Table 9.2 presents the ratings given by the participants for each expectQoE notification they received during S2. Overall, the participants rated the expectQoE prediction an average of $77 \pm 18\%$ of the time. Their ratings were more positive (thumbs-up, 61%) than negative (thumbs-down, 29%) during the total duration of S2. Only 13 participants rated the expectQoE notification more negatively. This can be explained by the participants' expectation or experience were different from expectQoE notification (i.e., expectQoE misclassified the QoE level).

We compared the ratings for two distinct periods T2 and T3. We observed a higher mean of thumbs-up ($67 \pm 25\%$) during T2 than during T3 ($48 \pm 35\%$). Both ratings in T2 and T3 are normally distributed (T2 $p < 0.033$, and T3 $p < 0.006$). However, three participants did not provide any ratings during T3 and their data were discarded. Then, we applied a Student's t-test (Dix, 2020) to verify the statistical significant of this difference, and we found $p < 0.025$; hence, we affirmed that the model performance (T2) was better than a random process (T3) from the end-user point of view.

9.4.6 ExpectQoE Notification Effectiveness

Above we analyzed the answers from our participants regarding the notification's effectiveness. However, a quantitative analysis of application usage collected by the logger could assess it better. Therefore, we employed the Multiple Convergent Cross Mapping (MCCM; van Berkel et al. (2020)) to ana-

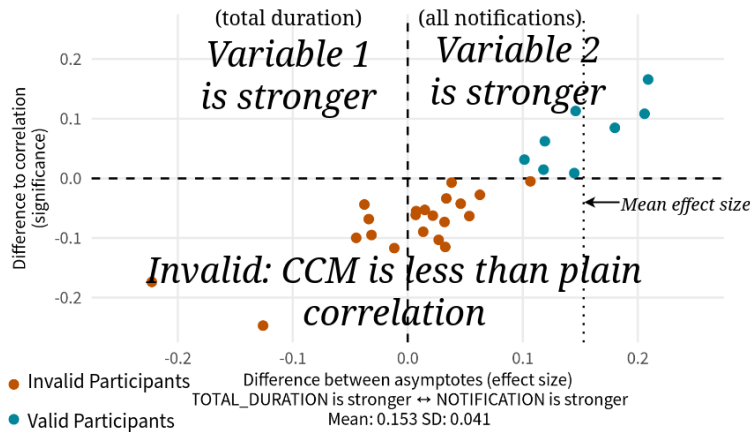


Table 9.3 S2: Causality Between Application Usage Duration and expectQoE Notifications

	expectQoE Notification: Effect Size (and STD)		
	All	High	Low
T2 Model	7/30 Valid 0.197 (0.14)	5/30 Valid 0.053 (0.15)	7/30 Valid 0.107 (0.098)
T3 Random	9/30 Valid 0.214 (0.139)	8/30 Valid 0.155 (0.116)	10/30 Valid 0.158 (0.088)

Figure 9.2 S2: Effect size and Amount of Valid Analysis for Causality Between Application Usage Duration and ExpectQoE Notifications During T1

lyze the causality between the notifications from expectQoE and smartphone use in terms of application usage duration. This method was developed to better understand the interactions between users and technology over time by differentiating causality from correlation based on quantitative data obtained in-the-wild. MCCM is built on the Convergent Cross Mapping (CCM; Sugihara et al. (2012)) method and extended from the ecology domain based on empirical dynamic modelling methods (EDM; Ye et al. (2015)). EDM allows conceptualizing multiple users’ behavior as complex nonlinear dynamical systems. CCM is used to investigate the causal relationship between two variables in a time series (i.e., which variable drives the other one) from a complex system. We looked for a positive convergence of the CCM values to determine whether the values were above direct correlation (difference in correlation greater than 0, means the correlation is significant). Then, we were able to establish which variable had a stronger effect by looking at how well one variable forecasts the other. By doing so, we established the direction of causality between the two variables. MCCM enables the aggregation of multiple CCM analyses graphically, processed with each participant time series, by comparing the difference in correlation and asymptotes (convergence point between the two variables forecast) from the CCM results indicating the effect size.

We repeated the MCCM analysis for two time periods (T2, T3) during which we tested three expectQoE notification-based variables (all QoE, high only, low only) against the participants’ application usage durations. Figure 9.2 presents the MCCM visualization. Each point represents a participant, and only the blue points (for valid analysis of a participant) are used to compute the mean effect size. The orange points indicate participants for whom the MCCM analysis failed. The thin dashed line corresponds to the mean effect size. Figure 9.2 shows that the notification received during the entire S2 duration had an impact on application usage duration for eight participants of thirty. The “invalid” participants’ MCCM analysis are explained by the auto-correlated nature of their data (i.e., the ap-

plication usage duration and the notifications are correlated, but not significant). Thus, the MCCM results are not exploitable on their dataset.

We present the overall results in Table 9.3, we include the amount of valid participants for each analysis. We observe that expectQoE had an impact on the application usage duration since the effect size is positive (Figure 9.2 X-axis). The effect size is overall higher with the random model (T3) than with the real model (T2). However, for both models, the expectQoE notifications with low QoE level have a stronger driving effect (0.107 and 0.158) than the high QoE level (0.053 and 0.155). The effect size is stronger in T2 for low QoE level comparing to high QoE level. Hence, the notification impacted more the the application usage duration in T2. Overall, we observed a decrease in application duration usage from $33.7 \pm 8[s]$ in T0 to $28 \pm 8[s]$ in T2.

9.4.7 Expectation Impact

We explored the expectQoE notification's influence on the participants' daily application usage expectations (in T0, T2 and T3). The average reported answer was 4 ± 0.49 overall, and 4 ± 0.04 in T0, 3.8 ± 0.05 in T2 and 4 ± 0.06 in T3. We focused the analysis on the participant's mean ratings during three different periods: before any use of expectQoE (T0), during expectQoE use with the real QoE model outputs (T2) and finally during expectQoE use with the random model outputs (T3). A one-way Analysis of Variance (ANOVA) of the reported satisfaction expectation ratings was carried out for each period to test if expectQoE influenced the participants' satisfaction. The distribution of the ratings within each period is normally distributed (T0: $p < 0.006$, T2: $p < 0.05$, T3: $p < 0.005$). The results show that the period (T0, T2, T3) has an impact on the participant's ratings, with $F(3,90) = 3.57$, $p < 0.03$. A Tukey post-hoc test (by setting the $\alpha = 0.05$) revealed that the participant's satisfaction increased significantly from T0 to T2. However, there is no statistically significant difference between the other two pairs of periods (T0, T3 and T2, T3). Hence, we conclude that expectQoE output during T2 has a significant impact on the participant expectations.

9.4.8 Impact of ExpectQoE on Application Usage Duration

The MCCM results have shown a partial driving force from the expectQoE system on the application usage duration of the participants after a close inspection of valid and invalid participants. We did not find other participant specific characteristics. Both subgroup of participants shared similar mean application session duration ($p < 0.03$, with one-way ANOVA) with overlapping standard deviation (valid: $97 \pm 421[s]$, invalid: $109 \pm 475[s]$). We extended our analysis based on those findings. We gathered all the application sessions from S2's participants and filtered out the applications not in the category presented by expected QoE (kept 41,585 sessions from a total of 63,529 sessions). Then, we grouped the sessions by periods: T0, T2, and T3. Interestingly, we found that 64% of the participants decreased the duration of their application session during T2 for all the application categories ($p < 0.02$, decrease of $-1.2 \pm 178[s]$) compared to T0 ($59.3 \pm 121[s]$). Additionally, only the communication and social category applications were used less in T3 than T0. However we found that the other categories

of application were used more ($p < 0.001$, increase of $+39.2 \pm 190[s]$). Finally, we observed an increase in application session duration for 65% of the participants ($p < 0.01$, increase of $+40.5 \pm 217[s]$) between T2 and T3. The increase is an unexpected outcome. However, it could be explained by the participant's fatigue in the study or the random QoE level shown during T3, negatively influencing their attitude despite expectQoE.

9.4.9 ExpectQoE Model: QoE Levels and Features (S2)

The expectQoE model achieved high accuracy for the majority of the S2 participants (i.e., Section 9.4.7). Hence, we quantified the QoE of all the application usage sessions (63529 sessions in S2) using the features previously selected (Table 9.1). Table 9.2 shows the QoE level per participants in T0. We found that overall in S2 the majority of the session was of high QoE level ($86\% \pm 7$) and a minority of low QoE ($13\% \pm 7$).

In order to better understand our results, the statistical significance was derived using a one-way ANOVA test, as both high and low median QoE level follow a normal distribution (high: $p < 0.002$, low: $p < 0.008$). Our analysis revealed that the median application duration was lower in low QoE sessions, 26 ± 9 seconds, contrary to 29 ± 10 seconds for high QoE sessions ($p < 0.001$). Hence, the participants spend more time in sessions rated as high QoE across all S2 sessions.

Furthermore, we explored the impact of the participant's physical activity and network state on the QoE level. The Radio Access Technology (RAT, e.g., Wi-Fi, LTE, UMTS, ...) does not influence the QoE level, and twenty-four participants had the same top RAT distribution for high and low QoE sessions. Then, we focused on the cell signal strength and the Wi-Fi signal (dBm). There were no significant differences between high and low QoE sessions based on cell signal strength values (dBm). Further analysis showed that the median Wi-Fi signal strength was lower in low QoE session -72 ± 3 dBm (weak, one bar on screen) than in high QoE session -65 ± 3 dBm (fair, two bars on screen) ($p < 0.01$). Finally, on average, the participants obtained a high QoE when their physical activity was "still" ($64\% \pm 19$) and lower on the other activities ($36\% \pm 16$) like walking. However, these results were not significant ($p > 0.1$).

9.5 Discussion and Limitations

In summary, our research aimed to explore the influence of a QoE level notification system (expectQoE) on smartphone application users. We verified expectQoE influence through qualitative and quantitative data. The second aim was to ascertain whether the factors influencing the QoE of smartphones have changed in the last decade. First, the analysis of the real model QoE predictions against the participants' reported QoE level yielded significant results: expectQoE influenced the participants' application usage duration. However, a relatively low performance was reported by the participants, which is different from the performance obtained during the model building phase. It could be caused by several aspects: limited training data which incorporate all the possible features' combination; bias

caused by model generalization; model structure and parameter tuning. A personalized participant model could be another approach to enhance the model performance. The model would be fine-tuned for each participant via reinforcement learning or hyperparameters optimization. The resulting models would embody one's way of perceiving the level of QoE. The user's intent may play a role in this context, where the utilitarian needs to satisfy their intent is ranked higher than their needs of hedonic satisfaction (Kahneman et al., 1999) and the knowledge that QoE is going to be low.

Second, we found a statistical difference in the system perception on the participants between the real model predictions (T2) and than the random model (T3), validating the model performance. Third, the MCCM analysis found that the expectQoE notifications drive their application usage duration. Also, the effect size is stronger when the notifications contain low QoE indications. On one hand, this could be explained by smartphone's users preemptively limiting the time they spend in an application to reduce their predicted annoyance. On the other hand, the MCCM results are difficult to generalize (from only 17 valid participants over 30, Table 9.3).

Four, we found a significant trend in the application usage duration once the intervention started (i.e., decrease application duration). However, this effect is unsustainable in time due the participant fatigue in the study or the impact of the random notification (T3), decreasing their trust in expectQoE. Additionally, the QoE level was high overall. Hence, the need for expectQoE interactions may be only suitable and useful for specific contexts (e.g., physical activity changes, roaming, and optimizing smartphone use duration to satisfy the user's intent faster and reduce their smartphone usage).

Fifth, the factors influencing the QoE of popular smartphone applications remain unchanged since documented in 2012. However, we found that smartphone users are more network conscious and care about the impact of their smartphone usage on their digital wellbeing. Also, they subscribe to multiple streaming services and often have unlimited internet access (no data cap). These changes can be linked to the smartphone entering the plateau of productivity (i.e., mainstream adoption) (Gartner, 2021), in contrast to 2012 (Ickin et al., 2012), in which smartphones were on the rise of adoption.

Finally, we expect the implications of our work can help smartphone application developers to enhance their software by going beyond simple network state indications, and include a QoE aware system, which is capable to preemptively notify to their users an approaching low QoE event. Application developers could learn from our work by making context an intrinsic information source in their application performance evaluation. Hence, enabling them to build a better experience metric than scroll jank (e.g., visual hiccup and artifact) and startup latency.

Most limitations in the work presented in this paper arise from our choice to collect data in-the-wild with limited disturbance to our participants. Moreover, the dataset gathered in S1 and used for S2 modeling efforts was limited by the application usage collected and rated, mostly communication applications. We believe that in our case, the impact is limited due to communication being the most used application category. Nonetheless, the categories may not be sufficient as such applications contain services that depend on distinct network models, e.g. a video chat and a voice call have different network needs (Tsolkas et al., 2017). Hence, the category may be good for a small-scale study focusing on popular applications used for the main service (e.g., WhatsApp for text message

and not for video conference). We believe that user's intent within the application should be the focus in future studies. As well, the root cause of low QoE events should be explored. Although we found an impact of expectQoE on our participants' application use duration, the interaction model we used (dyad of emoji with randomized placement) could have influenced the participants if they expected to observe the categories' emojis in the same place. Furthermore, the applications observed in S1 were all internet-enabled. Thus, the model implemented in S2 focuses on this type of application. However, the method to collect data, build and deploy a QoE model presented could be applied to the non-internet application. The network quality indicator (bars) are insufficient for the user to assess its expected QoE level for offline application due to the multiple factors influencing their experience. Finally the timing of S1 and S2; S1 happened in 2018 and S2 in 2021. During these three years, the Android system evolved. The system upgrades could have impacted our results. However, the habits of the S2 participants were always compared with the data gathered in T0 (baseline usage). Hence, our overall findings are valid and the impact of this time difference is limited.

9.6 Conclusions

Through a mixed method of qualitative and quantitative data collection in which the participants were active in the research by providing information directly and indirectly, we presented our research regarding the effectiveness of a QoE-based notification system to limit smartphone users' burden in case of low QoE. First, this required gathering application usage QoE levels in situ (S1). Second, it required building a QoE classifier from the data obtained during S1. This classifier was then included in our smartphone logger, providing the participants in S2 with the expected QoEs through notification. Our results showed that the participants reported higher satisfaction when expectQoE showed the real model predictions (T2) rather than random QoE level (T3). However, a global model have limited prediction capabilities. Hence, our future work includes building dynamically personalized QoE model based on the user application usage and context habits. Additionally, we investigated whether the expectQoE notifications had an impact on participant application usage by employing a MCCM analysis on a time series constructed from different periods of S2. Overall, we found that expectQoE decreased application usage duration for some participants. The influence was stronger when low QoE notification was shown. We also identified some features (e.g., Wi-Fi strength and physical activity) that impacted the overall QoE level of the participants during S2. Third, we presented changes in the factors influencing the QoE of the smartphone. We found that all factors are still applicable. However, they have evolved with new smartphone usages (e.g., streaming audio and video content). Additionally, smartphone users are now more network and wellbeing conscious than ever.

Acknowledgment

SNSF MIQmodel (157003) (2015-2019), AAL Guardian (2019-6-120-CP), Swissuniversities P-13 project and UNIGE COINF (2018-2020) . The University of Geneva ethics committee approved those studies under CUREG.201803.02 and CUREG,201909.12.

Chapter 10

Discussion and Conclusion

Chapter Contents

- 10.1 Discussion of the Results 164**
 - 10.1.1 Features Impacting Smartphone Users’ Connectivity 164
 - 10.1.2 Forecasting Application Usage 165
 - 10.1.3 Context Factors Influencing QoE 167
 - 10.1.4 Quantifying QoE On-device 170
 - 10.1.5 User’s Factors Influencing QoE 171
 - 10.1.6 Managing User Expectation 174
- 10.2 Limitations 175**
 - 10.2.1 Studies In-The-Wild 175
 - 10.2.2 Data Augmentation 176
 - 10.2.3 Study 2 Protocol Design 177
 - 10.2.4 Generalization of Impact of ExpectQoE Intervention 178
 - 10.2.5 5G Connectivity 178
 - 10.2.6 Popular Applications And Programming Language Update 179
- 10.3 Future Work Areas 179**
 - 10.3.1 Operating System Implementation 179
 - 10.3.2 QoE Continuous Indication 180
 - 10.3.3 Enhancing QoE Models 180
 - 10.3.4 Services 181
- 10.4 Conclusion 181**

In this chapter, we answer our research questions separately, compare our results to the state of the art, discuss the results and conclude on this thesis.

10.1 Discussion of the Results

In this section, we discuss our results in the context of each research question defined in Section 1.2. We summarize the relevant conclusions and experiences from the articles examined and expand the discussion. We organize the results in several subsections. To answer to RQ1s we present the features that impact the connectivity of smartphone users in Section 10.1.1. We discuss our algorithm and method to forecast application launch to answer RQ1b in Section 10.1.2. Then, we examine the context factors influencing QoE and discuss our context-aware user-centric QoE quantifying model for popular mobile applications to answer RQ2 throughout Section 10.1.3. We present challenges in quantifying the QoE of smartphones on-device to answer RQ3 in section 10.1.4. Additionally, we explore current user factors that influence QoE to answer RQ4 in Section 10.1.5. Finally, we present our QoE notification system, *expectQoE*, and its implications for managing QoE in Section 10.1.6.

10.1.1 Features Impacting Smartphone Users' Connectivity

Most smartphone applications access the mobile network to download content (e.g., video, image and audio) and to access services hosted on the cloud (e.g., machine learning inferences and API call). Therefore, connectivity to a network that provides Internet access is essential to understand the QoE of smartphone applications. However, past works focused on QoS metric, such as Round-Trip Times (i.e., time duration of a network packet to go from a client to a server and to come back). We proposed to study the features that impact the daily connectivity level of a smartphone user.

After analysis, we were able to answer the first research question:

RQ1a - What are the features (i.e., factors) that affect smartphone users' connectivity in-the-wild, over time? [...]

From all the smartphone data collected (Chapter 4, Table 4.4 and Table 4.5, 15 features) during three user studies, we found that the following four features embodies an individual's connectivity level over time:

- **Network access technology:** Wi-Fi or cellular (e.g., 3G, 3.5G, 4G or 5G)
- **Overall data consumption:** the amount of uploaded and downloaded data through both smartphone network interface
- **Signal strength:** corresponding to the power of the radio signal received by the smartphone modem (in dBm)

- **Participants' mobility patterns while connected:** derived from the network cell towers and Wi-Fi access points

Our findings reflected that mobile connectivity and smartphone network activity were consistent with the daily activities of the participants (i.e., physical activity and mobility), affecting their smartphone application usage. On average for all periods, a connection to the Internet is present $93 \pm 0.8\%$ (“mean \pm stddev”) of the time.

We also compared the features of connectivity levels from the three different cohorts over time, collected during past studies in our lab. Our results showed that the feature network access technology influenced each cohort differently. The smartphone users from the last cohort (period 2020) connected more often to 4G networks to access the Internet, compared with the older cohorts (period 2015–2017 and 2018–2019), in which the Wi-Fi connections were used most of the time. Other than the feature network access technology, we observed consistent daily and nocturnal patterns in the feature data consumption in all three cohorts, which confirms the finding in [Walelgne et al. \(2020\)](#). Among all participants in the three cohorts, low data consumption was found during the night and higher consumption during the evening and early morning. These patterns reflected how people used their devices and consequently how it impacted their connectivity over time. We also observed bursts of higher throughput, which could be originating from video consumption (leisure) or video conferencing (business or/and leisure). This finding could be valuable for network operators. For example, they can use this information to rent more bandwidth from their network provider for specific time periods, thus enabling a high-quality video conferencing experience for most users. The 5G technology offers higher bandwidth and leads to the creation of new services ([Narayanan et al., 2021](#)). However, the QoE of such service (i.e., augmented reality (AR) and virtual reality (VR) applications) can be impacted by the QoS ([Bosk et al., 2021](#)). Hence, smart resource allocation based on smartphone user habits help network operators to provide to their user a better user experience.

In summary, the four features (i.e., signal strength, network access technology, overall data consumption and mobility patterns while connected) we extracted and the observed connectivity patterns could lead to an enhanced system to preemptively improve user experience by sizing their network dynamically based on the schedule of their users (i.e., their application usage). Surprisingly the four features were affecting the users' connectivity for all the different time period. Hence, the technological advancement during those periods did not enable a lower data consumption (via better compression algorithm), better signal strength (via beamforming) or transparent handovers between network access technologies.

10.1.2 Forecasting Application Usage

As we observed from the literature (Chapter 2), applications play an important role in modeling QoE. The second part of the first research question focuses on forecasting application usage. Application forecasting can predict which application a smartphone user will launch on their device. Hence, we explored how to forecast a ranked list of future applications launched by a smartphone user, from

a limited set of features, and specifically focusing the previous chains of applications used, without context information. Context information can contain private and sensitive information, hence we wished to reduce their need and maintain high forecasting performances. Previous work focused on forecasting applications to reduce the burden of the end-user and the frequency of meaningless application usage (Roffarello and De Russis, 2021).

However, most of the literature on application forecasting used a large amount of contextual and sensitive user information (i.e., location). We have shown that the amount of information could be reduced without impacting the model and its performance. The methods are also difficult to compare due to different testing metrics and protocols (i.e., data split). They do not provide enough information to implement their methods with another data set. Finally, the previous works are all limited in the temporal forecast space (Li et al., 2022), that is, they only predict one application at a time. The ability to forecast all the applications which can be launched, in a session, provides an advantage in the QoE and wellbeing domain. It enables QoE-enable service makers and wellbeing intervention designers to develop better strategies based on these predictions, helping to eliminate potential failures before they happen. A forecast further into the future allows for a wider margin of maneuver for preemptive set-up, which can be adapted based on context change. For example, a QoE-enable service providing live information could use the forecast to optimise its decision-making algorithms (e.g., extended catching or preparing for annoying experiences). As such, the forecast can have an impact on the user experience of smartphones.

RQ1b - How to accurately forecast the usage of smartphone users' applications with application usage record?

To answer RQ1b, we first defined a method to build and test a forecasting model built from application usage records. The method, the ranking algorithm, the model implementation are presented in Chapter 5. Our proposed method enable to compare and test multiple forecasting model trained on application chains. The method was implemented on two different datasets following detail filtering operations. The operations' attribute (e.g., threshold, cutoff values and limits) were selected based on techniques employed in the application forecasting literature. Our ranking approach ($F1 : 62 \pm 6\%$, $\text{mean} \pm \text{stddev}$) performed better than the decision tree-based model ($F1 : 36\%$) for the maximum chain's length per user . Our forecasting implementation performed better than a standard decision tree algorithm that forecasts only the next application based on the previous one. Furthermore, contrary to the literature, which trains one model and then tunes it to a specific user, we trained one model per user to enable a faster generalization of the model and limit overfitting.

However, the implementation was not deployed on smartphones, which could have affected the protocol of Study 2 (S2). Also, the model suffers from a cold start problem. The model must be first trained with one week of application usage history. Hence, in a deployment scenario, the model can only provide inference once a week of data has been collected. The cold start problem could be resolve by forecasting with model created on other users during the first data collection period.

Overall, our method and algorithm can enable the development of new predictive applications that can provide information to the operating system (e.g., caching data before an application starts for a faster launch). As well, forecasting applications may facilitate the creation of digital wellbeing intervention to reduce problematic application usage (Roffarello and De Russis, 2022). We propose to integrate the application forecasting model into QoE-enabled services to enhance the end-user smartphone experience. A application could be loaded in the background and process the user inputs faster if it was forecast that the user will launch it.

10.1.3 Context Factors Influencing QoE

First, we identified the contextual factors already presented in the literature (Chapter 2). As most of the previous work focused on QoS attributes (Casas et al., 2016), we went beyond and proposed factors from the following three perspectives:

- User: intent to satisfy, expectation, application launched, its duration, and the users' interaction with the screen (number of touch and duration)
- System: networking information (QoS), active network testing (ping) and battery state (level, temperature, total capacity, and condition: charging or discharging)
- Context: physical activity of the smartphone user

Collecting samples of QoE ratings in-situ has always been difficult due to many potentially confounding factors. Most QoE studies were carried out in a laboratory or through crowd-sourcing, where participants had to perform specific tasks given by the researchers and then assess their experience (Sackl et al., 2016). Different from previous crowdsourcing-based works, we focused on collecting QoE ratings and smartphone data from living lab participants unobtrusively in their daily life contexts. This allowed us to answer the first part of the second research question:

RQ2a - What contextual factors should be considered to create an accurate and context-aware QoE user-centric quantifying model for popular mobile applications?

We illustrated the feasibility of collecting context factors to model the smartphone application QoE in Chapter 6. In summary, our qualitative results showed that contextual factors, such as network connectivity, smartphone user physical activity, and application usage, could be collected without impacting the participants usual smartphone interaction behavior.

In-the-wild, we observed how smartphone users were using their applications to accomplish different tasks and satisfy multiple intents (i.e., text communication or content sharing). Those habits were observed for five participants in our pre-S1 study (De Masi and Wac, 2018). We discovered that smartphone users consumed content (images, text, audio, and video) on Chrome, Spotify, Facebook,

Instagram, and Google Maps. Messenger and WhatsApp were used mainly to read and write messages. The analysis of the data collected in the pre-study confirmed the feasibility of collecting factors that influence the QoE of popular mobile applications. Moreover, the same results were observed later after the execution of S1 with 38 participants.

We also answered RQ2a quantitatively. We used data from Study 1 to train a computational model with the XGBoost algorithm (Chen and Guestrin, 2016) to classify the QoE application level between “Low” and “High” (i.e., a classification problem). XGBoost has been the industry and research standard algorithm for classification task on tabular data (Shwartz-Ziv and Armon, 2022). The performance and accuracy of the model were evaluated by cross-validation with a focus on the Area Under the Curve (AUC) metric. AUC provides an aggregate measure of performance across all possible classification thresholds (i.e., True Positive Rate and False Positive Rate). Our results indicate that the computational model is accurate, after a 20 folds cross-validation, the average AUC was 0.83 ± 0.3 and accuracy was 0.94 ± 0.01 after a hyperparameter random search using the best classifier available. The input features are important in identifying the contextual factors that influence QoE. Since XGBoost models are based on decision trees, they are explainable to a certain extent. As such, we investigated the most important features to classify a smartphone application experience. We analyzed the importance of the features of three QoE models (De Masi and Wac, 2020). The three models were constructed based on the same features. However, the filtering and aggregation schemes of the input data were distinct. Features with importance less than 1% were excluded (arbitrary threshold). In summary, the contextual factors that should be considered to create a user-centered accurate and context-aware QoE model for popular mobile applications are the following (ranked by importance distribution with mean \pm standard error [%]):

- **The task to accomplish/intent to satisfy in the application by the smartphone user** ($38.37 \pm 2.64\%$). We observe that the user’s intent (i.e., consume content, share or create content, read text message, write text message, control an app (start/stop music), video call or audio call) is essential in building a user-centric quantifying model. Previous QoE studies asked participants to assess their experience on application usage provoked by researchers (Casas et al., 2015a), missing the meaningfulness of the user’s intent. The origin or trigger of the intent can be twofold: 1) the users’ conscious need; they could be lost and wish to locate with a map application. Additionally, 2) the intent could have been triggered by a notification on the smartphone. The smartphone notifications are unpredictable (Mehrotra et al., 2016) (i.e., except certain cases, like alarms).
- **The application launched** ($12.04 \pm 0.77\%$). Unsurprisingly, the application launched is one of the most important features. However, the same intent could be satisfied through multiple available applications (e.g., sending a message via WhatsApp or Messenger). The selected application weighs on the experience. Motivation to launch an application is still an ongoing research topic, as presented in Chapter 5.

- **The user physical activity while interacting with an application** ($9.03 \pm 0.53\%$). Physical activity in which the smartphone user participates is important (De Masi and Wac, 2019). For example, a user seated in the tram will use their device differently than one walking or standing still (e.g., impact on self-awareness). Additionally, the type and duration of the content consume plays a role (e.g., headphone for listening to music and landscape mode to watch videos). Overall, the attention needs are different when a user interacts with a smartphone when walking (Argin et al., 2020; Liebherr et al., 2020; Takeuchi et al., 2016).
- **The network QoS (bandwidth, signal strength, network state, and packets traffic statistics)** ($2.90 \pm 1.08\%$). From the literature (Chapter 2.4.3) and RQ1a, we know that the QoS information is essential to obtain a meaningful QoE model. However, the models build from S1 have shown a low importance in network QoS. The overall high quality of the study participants' networks could be a factor influencing this finding.
- **The battery level** ($2.94 \pm 0.12\%$). The battery is an important component of the smartphone. Ickin et al. (2012) had previously presented its influence on smartphone users. However, the battery has multiple attributes. We know that the battery level (% on the screen corner) is linked to assessing its application usage experience. The battery temperature, charging state, and capacity had a negligible impact on modeling the smartphone user QoE.

The three top-ranked factors are user-centric and context-centric, contrary to the last two, which are system-centric. They affect the user's focus and its immediate context. Such context-aware model should be accurate enough to limit and reduce the user burden. As well, user burden is different for each context and profiles. Hence, the accuracy threshold for QoE user-centric quantifying model is a function based on one's burden limits, applications usage, context and profiles.

We were then able to use these contextual factors to answer the second part of the research question:

RQ2b - How to accurately predict the smartphone application user's QoE with a context-aware model?

Different models were illustrated and tested in Chapter 8. Furthermore, we provided instructions to create contextual QoE classification models. We found that our proposed process pipeline to build such models performed well through different modeling scenarios. The model build from the user expectation rating feature (i.e., S1 EMA question 3) lead to better classification performance, with an higher AUC score (0.874 ± 0.027) than the other models (e.g., candid model AUC: 0.829 ± 0.028), which was statistically significant ($p < \alpha$, $p_{AUC} = 0.001$). However, this feature was not included in RQ2a due to her active collection nature (i.e., the user has to input it).

In summary, the results showed that collecting in-situ QoE ratings and collecting smartphone background data enables us to use common machine learning techniques to build an accurate (per-

formance metric-wise) predictive model for “High” and “Low” QoE. As we wished for a more accurate model (i.e., higher performance, particularly the AUC metric), we investigated multiple input data filtering schemes that generated more accurate models in different scenarios. We observed that data preparation (e.g., filtering and aggregation) improved the performance of our QoE models (AUC: 0.822 ± 0.041 , on average $+0.03$ superior against the other models). Due to the nature of the data collection method, the input data is often unusable directly after collection because a participant may not respect the protocol which explains how to rate their experience.

The prediction is done for the current inputs available. Hence, no forecasting was done to know the QoE level in the immediate future (e.g., in the next 5 minutes). However, knowing the current QoE level in real time, even before the user feel annoyed, allow for action to take place, and could limit the user burden.

10.1.4 Quantifying QoE On-device

QoE-enabled services are often hosted on cloud infrastructure (Kiani Mehr et al., 2021). As such, these services are inaccessible where mobile networks are unavailable, making them useless. The services could also collect sensitive information which could be problematic for a future user. Therefore, QoE-enabled services that focus on managing end-user expectations should be implemented on smartphones first.

Toward our end goal to quantify and facilitate the enhancement smartphone QoE, we presents our work on the challenging domain of on-device prediction models in Chapter 8. Overall, we were able to create a model to quantify QoE on-device (best model AUC:82%). However, we observed a lower performance ($\Delta_{AUC} = -0.6\%$) for on-device models contrary to the models built with the full feature set presented in Chapter 7, due to the restrictive feature set available in the on-device context. Indeed, some user-centric and system-centric features were impossible to access due to their nature. For example the aggregation time window size to generate statistics is based on the end of the experience, which can only be determined offline (i.e., at the end of the application usage). Hence, this time window is incompatible with an on-device prediction. Other features, such as low level sockets information were removed in new OS versions due to security concerns (Android, 2022). We implemented our on-device model in mQoL-Lab, our research mobile application. In Chapter 9 we present the results of its implementation during S2. From the feedback of the participants, we found that the model correctly quantified QoE for most of them ($> 60\%$ participants, i.e., matching QoE levels). However, we hypothesized that better results could be obtained by tuning the model for each user (i.e., profiling).

RQ3 - What are the challenges to quantify QoE on-device with the users’ context factors in-the-wild?

As a concrete response to the third research question, we list and discuss the challenges that we encountered below:

- **Feature Availability** Some features, such as extended network information (e.g., TCP socket states), are not available with the new versions of the Android operating system. Others, like the aggregated packet traffic statistics, could not be computed, as the application usage needs to be finished to be included in the statistics. Therefore, these features were discarded for the on-device model. The on-board accessible features were as follows: battery level, user physical activity, application, user's intent (consume media, communicate via audio, video or text, share content), network connection type, Wi-Fi level, Wi-Fi speed, cell strength, cell bandwidth up and downstream. Overall, the performance of the model was negatively affected by this action. The on-device model performs worse (AUC: 0.76 ± 0.037) than the base model (AUC: 0.829 ± 0.028 , Chapter 8).
- **Prediction Timeframe** All smartphone application usage embodies the users' interaction with their device. Due to the nature of the experience, the QoE level may change during the application usage. Therefore, the temporality of the prediction could affect the model output. As such, we prioritized an approach to quantify the instantaneous QoE level, i.e., when the application is launched or placed in the foreground.
- **User Intent** The user intent could not have been asked before each application usage to the participant. This request would have generated high user annoyance. However, our analysis showed that the intent is essential to quantify QoE. Thus, we proposed a scheme that selected the probable intent based on the intent's frequency for QoE prediction obtained in S1.

In summary, we found that the main challenge is the availability of features. Smartphone features are now becoming more difficult to obtain due to privacy and security concerns of makers. Some important features, such as session duration and aggregated packet traffic statistics, were impossible to generate. QoE on-device models should be designed and operationalized to access a higher level of users' information, and application-generated data should be protected with special permissions (e.g., same as the location permission). Another means to employ problematic features in to train the model directly on the device. Hence, the data stay on the device. Federated learning could be use to solve this problem in the future. Furthermore, the models created from the limited set of features performed worst than their counterparts. The duration of an application session and aggregated statistics could be interpolated to balance this issue. Additionally, the temporality of a smartphone experience and the user's context are fast-changing (i.e., long, short or micro-session and situational context as described by [Ferreira et al. \(2014\)](#)). Hence, such a QoE classification model should continuously update its inputs to improve its overall performance.

10.1.5 User's Factors Influencing QoE

Chapter 9 examines the fourth research question empirically through the use of weekly Day Reconstruction Method session during S2 (2021). The last assessment of the factors influencing user QoE on smartphones was done more than ten years ago ([Ickin et al., 2012](#)). Since then, the hardware of

smartphones and mobile networks has been upgraded. In addition, applications became more ubiquitous in our daily lives and its users became more knowledgeable about their devices (Shaheen et al., 2017), which could have impact the factors influencing QoE. During this thesis, we collected the S2's participants inputs about these factors in 2021.

RQ4 - What is the evolution of the most influential factors for the user's QoE on smartphones since their last assessment?

We learned that the factors influencing the QoE of smartphone applications were identical to those identified in previous work (Ickin et al., 2012). However, this factors evolved with new smartphone usages and the evolution in telecommunication network technology (e.g., streaming audio and video content). Additionally, recent smartphone users are more network and wellbeing-conscious than before. Moreover, they subscribe to multiple streaming services and often have unlimited internet access (no data cap). These changes can be linked to the smartphone entering the plateau of productivity (i.e., mainstream adoption, as presented by Gartner (2021), in contrast to 2012 (Ickin et al., 2012), in which smartphones were on the rise of adoption. We discuss each factor in the following below:

- **Application interface design** Overall, the design of the application interface has been frequently discussed. The participants enjoyed the interface of the notification bar. However, they complained about the content of the applications versus, for example, their mood at that time (e.g., announcement of the death of a family member via an application). The non-homogeneity of the Android application interface is a complex issue (Kang et al., 2016). Google is trying to solve it with their Material Design guideline Google (2014) and Android Jetpack (Google, 2018), to build better applications faster. These guidelines were found to be followed for a majority of applications available on the Google Play Store (Yang et al., 2021).
- **Application performance** Participants reported problems with sharing photos and streaming videos. The participants also understood the need for the network, they were connected to. Additionally participants commented on roaming between countries and the time it takes for their smartphones to connect to a new network. However, even by knowing their current context, the will to satisfy their intent was strong. Participants reported they had set up the cell network manually due to their proximity to a foreign cell tower (at the border). The participants were able to discern whether the performance of an application was due to the application itself or to an underlying network problem. That is different from the 2012 study, where the participants did not distinguish between the two. In summary, the applications' performances were due to external factors and not the applications themselves. As such, the applications used by the participants were stable enough to trigger annoyance due to the network, and not other sources. However, this could be explained by the hardware performance of the smartphone, which could manage a resource-hungry and not optimized application (Liu et al., 2014). On low-

level smartphone, the applications' processes may use all the available resources and cause a bottleneck impacting the users through visual hiccup (e.g., scroll jank).

- **Battery** The battery capacity was enough to sustain the smartphones for more than a day. Previous work had shown the importance of battery management for streaming (Ickin, 2015; Ickin et al., 2013), and pointed toward optimizations in this context. Android has battery energy saving policies (Android, 2022) that reduces the overall consumption of unused applications. Hence, the impact of battery life on QoE was diminished. Additionally, the development of new technologies in battery management and optimization of smartphone energy (for example, screen technology, as presented by Dash and Hu (2021)) made battery factors less important than before.
- **Phone features** Overall, the smartphones now have the same basic hardware features (e.g., GPS, camera, Wi-Fi, Bluetooth, 4G or 5G), except for the storage extension (e.g., micro-sd) and the headphone jack (Buckle, 2019). The discussed phone features are now based on the available application and services (software). Smartphone makers have been pushing for more features, like integration of machine learning system in their OS (Google, 2022) for photography (e.g., segmentation, recognition and filtering), and toward natural language processing tasks (e.g., spell checker, translation, speech recognition, text-to-speech) on the device throughout special computer chip (e.g., Tensor processor).
- **Applications and data connectivity cost** We found that a small majority (N = 17/30) of the participants did not buy any application. However, 22 participants were paying for a multimedia subscription service (e.g., Netflix and Spotify). Furthermore, they also had an unlimited data cap. The data cap changed from 2012, this type of contract was expensive, and the connection was slow (pre-4G era). Hence, streaming services on smartphones were used more on Wi-Fi than on cellular connection (Detti et al., 2012).
- **Routine** A high number (N=25/30) of participants reported following an identical routine in the morning and the evening. In both cases, they used a set of applications, often communication (e.g., WhatsApp, email), before starting or finishing their day, which corresponds to the findings of Ickin et al. (2012) about user routines. This habit makes smartphone use more widespread (Oulasvirta et al., 2012). This routine has also been observed in the field of addictology related to smartphones (Haug et al., 2015; Kwon et al., 2013b; Samaha and Hawi, 2016).
- **Lifestyle** As shown in Chapter 9, we observed an evolution of the lifestyle factor. As previously, we found applications supporting the day-to-day activities and life of the participant. However, the applications were more diverse (e.g., financial, spiritual and health applications). From the reviewed literature (Gonçalves et al., 2020; Lunde et al., 2018; Vaz et al., 2021), we know that such applications can have a negative or positive impact on the user's lifestyle. Nonetheless, the participants only reported applications with a positive effect. Thus, it could indicate that

they were not aware of the negative aspects or neglecting to report negative effect. As such, the dependence on an application may burden the user's life and raise their degree of annoyance with a smartphone application (Lowe-Calverley and Pontes, 2020).

In summary, in this section, we listed and discussed seven factors influencing the QoE of smartphone applications and their evolution or stalling since 2012. The factors were previously identified by Ickin et al. (2012). Surprisingly the factors are still valid after ten years. As well, the participants who reflected on their application usage limited their exposure to their smartphone, during night or at home after work. Smartphone user are more wellbeing conscious than ever.

10.1.6 Managing User Expectation

In Chapter 9 we present S2 and the *expectQoE* notification system to answer the fifth research question in detail. Managing the end-user smartphone application QoE's has been done in the past (see Chapter 2.5). However, the methods developed focus on the adaptation of content, i.e., limiting its size, and the adjustment of the network protocols mechanisms (e.g., use of multihoming with QUIC toward).

RQ5 - What interaction method can contribute to manage smartphone users' expected QoE in-the-wild?

Hence, we proposed employing an interaction method which places the user in the center, allowing them to decide which action they wish to perform to manage their expectation. The users expect their application to always have high QoE by default. Hence, they feel more annoyed when low nonfrequent QoE events present themselves. Based on the information provided by the *expectQoE* notification system, the end-user determines whether they wish to launch an application that could create an annoying experience. *ExpectQoE*, through emoji-based notifications, presents the expected QoE level of multiple application categories. We then investigated whether the intervention with *expectQoE* facilitate the user's QoE level and affected their application usage. The experimental results showed a limited user-reported improvement in their perceived QoE due to *expectQoE* (N=20/30). Although participants always had high QoE application usage expectations, the variation in their expectations caused by *expectQoE* was minimal ($p < 0.04$), and was not statistically significant ($p > 0.7$). However, based on a time series analysis of the quantitative data, we observed that *expectQoE* slightly affected the application usage duration for a majority of users (N=18/30) when a low QoE was predicted, 64% of the participants decreased the duration ($p < 0.02$, T2: $-1.2 \pm 178[s]$) compared to the baseline (T0: $59.3 \pm 121[s]$). In general, we found that the expectations of smartphone users could be influenced by employing *expectQoE*. The notification informing the smartphone user about the level of QoE affected the duration of application usage.

In summary, we found that a notification system could represent the expected QoE levels for categories of smartphone applications. In conjunction with this thesis, we implemented this system in our study application and tested its effectiveness in-the-wild. We obtained a positive result to

conclude partially that, indeed, this interaction method does help manage and reduce the smartphone users' expectations in case of low expected QoE.

10.2 Limitations

After exposing our answers to the research questions, we present the limitations we identified in our work below. We also propose some actions and ideas to resolve them.

10.2.1 Studies In-The-Wild

Studies in-the-wild offer researchers a better overview of daily activities of their participants by collecting their behavior, interaction, and actions (Rogers and Marshall, 2017). However, multiple challenges arise from this approach. Firstly, the studies' duration in-the-wild is longer than in lab studies, although it depends on the research question. Hence, finding participants for such can be challenging. Another limitation is the number of participants. The ITU (ITU-T, 2021) recommends at least 35 participants for the in-the-wild study (e.i., S1 had 37 and S2 had 30 participants). However, this number was selected to assess audiovisual content and not smartphone applications. Furthermore, the described ITU protocol does not consider the context, the quality of the assessment, and the intent of the participant in their guidelines. Another challenge in recruiting for a smartphone-based QoE study in-the-wild is the hardware and software differences of smartphones. The observed population may have different devices that can affect their daily habits, assessment of QoE, and usage of applications (Böhmer et al., 2011). Hence, in some context, the population are given the same standard device. From the software perspective, we instructed our participants not to upgrade their Android system during the two studies. The only updates possible were for the smartphone application and the security update. Other updates can create a bias if participants update their applications and then assess their QoE at a different time during the same period (Taylor and Martinovic, 2017).

Moreover, the total control of the operating system platform is problematic due to the energy policy of Android (Dash and Hu, 2021) and the deployment of applications through the Google Play Store (Google, 2019). Our application to survey the participants and collect data passively, mQoL-Lab, had to be set up so the Android system would not deactivate it, even if the application is never shown on the foreground (i.e., on the screen, contrary to the background, not visible on-screen).

The results obtained from the S1 and S2 data analysis considered the participants' answers as the ground truth when they assessed their application QoE level and the *expectQoE* impact. However, to study user-centric QoE, no other data could be used to assess the truthfulness of their answers.

Finally, due to the recruitment process and the incentive to participate in the two studies, both populations of S1 and S2 are close in age and gender proportion. However, the populations are not representative of the smartphone user population. The majority of the population of our studies were students and young professionals between 20 to 35 years old. The population was not normally

distributed according to age groups, gender, and socioeconomic background. Hence, our findings are limited and can not be generalized.

10.2.2 Data Augmentation

10.2.2.1 Resampling

The data augmentation and resampling techniques were implemented on the data connectivity dataset presented in Chapter 4. The resampling process involved assigning minute-based intervals to each collected sample. During minutes without any available data, the last recorded data prior to that time was duplicated until a new sample becomes available. This approach enables the comparison of connectivity factors at different points in time. Data collection on the Android platform operates on an event-driven basis, utilizing a subscribe/publish model. Consequently, only when a change occurs, new data is published for collection by the Android logger. The augmentation process fills in the gaps between the previous and newly recorded events. It is important to note that this method assumes that the absence of data indicates a lack of change in both smartphone usage and connectivity within the user context. The absence of data can be attributed to two main reasons: either the logger was terminated by the operating system or the logger's battery safety mode (activated automatically when the battery level reaches 30%) was triggered. This augmentation method was not validated. Hence our findings could be limited, in particular for study participants which often let their smartphone reach the battery threshold.

10.2.2.2 Imbalanced Ratings

The ratings collected during $S1$ and $S2_{T0}$ had a high imbalance, respectively the distribution was 93.5% of high QoE and 6.5 % of low for $S1$, and 88% of high QoE and 11% of low QoE. Which follows the distribution previously reported by [Ickin et al. \(2012\)](#). Hence, we could assume that smartphone application users have a delightful experience most of the time. This high imbalance is problematic for building a classification model. The training algorithm behind the model needs a balanced class to limit overfitting (that is, always predicting the majority class). A full research domain named imbalance learning ([He and Garcia, 2009](#)) tackles this issue. Two main techniques were developed from this domain: over-sampling (e.g., generating or repeating samples until the classes are balanced) and under-sampling (e.g., removing or filtering samples until the classes are balanced)

We used the Synthetic Minority Over-sampling Technique (SMOTE) ([Chawla et al., 2002](#)) for over-sampling the low QoE class. SMOTE has been a standard since it became available, as shown by [Fernandez et al. \(2018\)](#). The deep learning SMOTE algorithm ([Dablain et al., 2022](#)) has been developed to improve its performance. However, [Goorbergh et al. \(2022\)](#) found that SMOTE could impact their model on the risk of ovarian cancer due to amplifying the minority class unnecessarily. Although, in their model, the consequences could cause harm (e.g., over-treating a healthy person), the direct impact is limited in our context. The model's performance could be decreased, but it would not impact the features' importance. Also, in most applications of imbalanced learning, the objective is not to

obtain a better representation of the domain (i.e., risk of ovarian cancer), but to bias models towards greater sensitivity on skewed classes (e.g., low QoE). Finally, the model developed and deployed during S2 was built via under-sampling. The performance was worse than using SMOTE (over-sampling) but the model only learned from real events and not from synthetic data. Although, the imbalance could have been higher. Participants may have reported their annoying experiences more often, because they remember them better than the good ones (Kensinger and Corkin, 2003). This means that during S1, a participant could have discarded a notification asking for a rating because its experience was high. Thus, this behaviour could have accentuate the imbalance between high and low QoE. Nevertheless, the same participant would have automatically reported their low experience level. The non-rated experiences could skew S1's results.

10.2.3 Study 2 Protocol Design

The S2 protocol (Chapter 1, Section 1.4.2) includes the first week of data collection without any intervention. Then, for eleven days, our QoE prediction model predicted the expected QoE output, and the participants were notified. Once this period is finished, the participants are notified with a random QoE level for ten days. The output is included in the Android notification and remains until the participants rate it. We wanted to observe *expectQoE* impact per user over time with this protocol. Additionally, the periods were in the same order for all participants (i.e., they all had T2 and then T3, none got T3 and then T2, no cross-over). However, this choice in the protocol could impact the aggregated results. Another way to observe the impact would have been to create a control group with random QoE level and another group with the QoE model outputs. Nevertheless, our conclusions from S2 are still valid per participant due to how your analysis was done. We compared the data of each participant to themselves. For example, to find out how the impact of *expectQoE* on the application usage duration between T0 and T2, we assessed one participant who collected data during the two distinct periods. Then the results were aggregated. This study design is named within-subjects design (Keren, 1993; List et al., 2011). Each participant experiences all conditions (i.e., with and without the real model output). We tested the same participants repeatedly to assess differences between conditions. Longitudinal studies often use within-subjects designs to assess changes within the same individuals over time (Caruana et al., 2015).

We also asked the participants about their daily QoE level (ACR MOS scale, 1 to 5). This aggregated assessment could impact the performance analysis of the model in-the-wild. The assessment was based on their full day. As such, it could have missed low or high-level experience, making this part of the analysis less robust. The variance between high and low QoE events may have influenced their overall ratings. Additionally, the daily QoE level was asked each day. The query's schedule was fixed in the evening hours. The participants' assessment may have been incorrect and affected by their most recent experience. In the future, for this type of study, the QoE should be continuously assessed during the day at random intervals. Alternatively, the same protocol established in S1 could be used partially.

The participants would receive the *expectQoE* notification, and in the same application session, they would receive an EMA asking to rate their QoE level for a specific application.

10.2.4 Generalization of Impact of ExpectQoE Intervention

Chapter 9 presents the *expectQoE* system to prepare the user of the smartphone application with low QoE, based on our QoE model built in Chapter 8. The S2 results have shown that such a system impacts smartphone user habits. We found that the *expectQoE* notifications drive, for some participants via the MCCM analysis (van Berkel et al., 2020), their application usage duration down. When we investigated the overall application use duration for all participants, we found that overall the participants decreased the duration of their application session after the start of the notification (T2) for all the applications compared to the baseline.

Furthermore, the effect size is stronger when notifications contain low QoE indications. On one hand, this could be explained by smartphone users preemptively limiting the time they spend in an application to reduce their predicted (by *expectQoE*) annoyance. On the other hand, the analyses were only valid for 17 participants in total. Thus, the results from the MCCM analysis are difficult to generalize.

10.2.5 5G Connectivity

Chapter 2 provides an overview of the connectivity status observed among smartphone users before 5G. However, it is essential to note that although 5G technology has been increasingly deployed and adopted, it may not be the ultimate solution for enhancing smartphone QoE. Despite the anticipated benefits stemming from the new transmission methods, advanced signal processing techniques, enhanced cell roaming capabilities, and increased bandwidth that 5G offers, it is crucial to consider several factors that limit its effectiveness in improving QoE.

Firstly, network coverage remains a significant challenge for 5G adoption. While 5G networks are gradually expanding, they are often available only in select areas (i.e., need more towers than 4G for the same area), limiting the overall accessibility and consistency of high-quality connections across various regions and locations. Moreover, the high-frequency bands used in 5G technology have limited penetration capabilities, resulting in potential connection interruptions when users are in indoor environments or surrounded by obstacles like buildings or trees.

Secondly, the compatibility between devices and 5G networks poses another hurdle. Although new smartphones are being manufactured with 5G capabilities, many older devices lack the necessary hardware components to support this technology. Consequently, a significant portion of smartphone users may not immediately benefit from 5G networks, thus limiting the overall impact on QoE improvement.

Finally, the potential congestion issues in 5G networks need to be carefully addressed. As the number of users adopting 5G technology increases, the network's capacity may face strain, leading to lower data speeds and potential degradation of QoE. Additionally, the quality of service experi-

enced by users may differ depending on network traffic and the service provider's resources, further influencing the overall QoE.

Therefore, while 5G technology holds promise in improving smartphone QoE, it is important to consider the limitations regarding network coverage, device compatibility, and potential congestion issues before concluding that it represents an all-encompassing solution.

10.2.6 Popular Applications And Programming Language Update

As presented in Chapter 7, the choice of the popular application was made in 2018 through the top application installed from the Google Play Store (Statista, 2021a). The applications were also selected due to their interactive features and the functionalities in the application (Gutierrez et al., 2011). A selection had to be made due to the nature of QoE to allow comparison of QoE assessments of the same application by different participants. However, the popularity of an application changes with time (Zhao et al., 2019a). Hence, other popular applications emerged in the top application since our work started (e.g., TikTok, Zoom, and Reddit). Additionally, the development platforms and tools used by the application companies have also evolved (Google, 2021). In 2017, Google announced that Kotlin would be the second programming language supported by Android (Titus, 2017). Since then, more developers have been writing Android applications in this language due to its advantages over Java. As well, Google released the Flutter framework in 2017. Flutter enables natively compiled, multi-platform applications from a single codebase (e.g., the same code can be used on Android, iOS, Windows and Linux). The impact of the programming language on the application end-user QoE has not been studied.

Finally, the timing of S1 and S2 is problematic; S1 happened in 2018 and S2 in 2021. During these three years, the Android system evolved. The system upgrades could have impacted our results. However, the habits of the S2 participants were always compared with the data collected in T0 (baseline usage). Hence, our general findings are valid, and the impact of this time difference is limited.

10.3 Future Work Areas


In this section, we discuss and identify a few potential research directions based on our findings. We focus on strengthening the performance of QoE machine learning models, as well as the management and optimization of QoE systems for smartphone applications. Below, we list and comment on those research directions, respectively.

10.3.1 Operating System Implementation

Operating system makers may wish to integrate such QoE alert systems directly into the core of their OS. Their users would always be informed about the possible low QoE due to external factors. The development of smartphone OS is a complex and difficult task (Li et al., 2022). The integration of new data collection systems and models could become a burden in the long term. Such models should

be continually updated based on smartphone user behavior and habits (e.g., by tuning the model, [Brownlee \(2019\)](#)). Furthermore, the OS maker should make new APIs available to monitor low-level factors that influence QoE. On installation, an application could request permission to access pervasive network information directly at the kernel level. For example, an Extended Berkeley Packet Filter (eBPF, [Scholz et al. \(2018\)](#)) program could be included in the applications which need this low-level information. eBPF allows user-written extensions in the kernel processing path. The kernel executes custom eBPF programs supplied by the application, effectively moving the kernel functionality into the userspace.

10.3.2 QoE Continuous Indication

The two indicators always present on the screen for communication with smartphones are Wi-Fi and cell signal indicators. However, following a possible inclusion of a user-centric QoE model in the operating system. We have demonstrated the effectiveness of such indicators through S2 (Chapter 9). The indicator would use the model output integrated in the operating system. Current indicators show the evolution of signal strength by increasing and reducing the size of their attributes (i.e., bars  or arcs). The UI of such indicator should follow the best practices for its design ([Yablonski, 2020](#)). For example, the von Restorff effect ([Wallace, 1965](#)) should be observed: *When multiple similar objects are present, the one that differs from the rest is most likely to be remembered.* Therefore, the QoE indicator should be different enough to stand out on top of the classical Wi-Fi and cell signal strength indicators.

10.3.3 Enhancing QoE Models

As shown in Chapter 9, all participants did not highly rate the predictions of the QoE model. This issue could originate from the model training data or the model choice and its features. The computational model was trained with data collected from 39 participants who have references for an annoying or delightful full experience. As such, smartphone application ratings encompass a bias. This bias could be assessed by profiling the users. For example, via the entry survey administered at the beginning of S1 or by classifying users based on their level of knowledge of the smartphone (i.e., expert annotator with a higher weight). The selected model type (XGBoost, [Chen and Guestrin \(2016\)](#)) has been proven to offer the best performance and training time (i.e., low) for classification and regression tasks on tabular data. However, a new combination of deep learning models and XGBoost performs better than XGBoost alone ([Shwartz-Ziv and Armon, 2022](#)).

Hence, we propose four ways to enhance the future generation of the model for predicting the QoE level of smartphone applications:

- **Expert Annotator** An annotator would follow strictly defined guidelines for rating sessions of smartphone usage. Each session would be recorded for playback for other expert annotators to validate the final rating. However, this method creates a higher financial cost (i.e., paying

the expert annotator) and privacy cost (i.e., capturing a video of the user's smartphone screen) of running QoE studies. Finally, the ratings obtained would be of better quality.

- **Personalized Model** The model could be fine-tuned per user, based on their previous experience and context. As such, a profile would be constructed based on their application usage (Xia et al., 2020b; Zhao et al., 2019a). This profile would be updated after each new input from the user and its direct feedback about the model. A reinforcement learning (Sutton and Barto, 2018) algorithm could be employed to facilitate the creation of such user-based models. The reward function would be the positive or negative feedback of the smartphone user.
- **Federated model** Federate learning involves statistical training models on remote devices, such as smartphones, while keeping data localized. The model weights built from the local data are then sent to the cloud. This process enables the creation of custom models tailored to specific data. Moreover, such models have been developed for QoE modeling (Ickin et al., 2021). However, the proposed method fails to address the user context.
- **Model Algorithm** As shown in Chapter 6 the choice of algorithm for classification matters. As such, another algorithm or model architecture could offer better classification performances.

10.3.4 Services

Our findings shows the feasibility of QoE-enable services and systems which can influence smartphone application experience. Contrary to *expectQoE* which is inferring own device, QoE-service can also be hosted on the cloud, depending on the use case. In particular, the QoE models built with our method are explicable. We can identify the feature with the highest importance, and so those that impact a smartphone user. As such, QoE-enabled services that provide recommendations, based on context, are feasible. For example, Alice has been with a mobile operator for many years and uses Zoom every day on her commute via train. However, her QoE is poor on different parts of her ride. As well, she has to have the video option activated during her call (i.e., work policy). A QoE-enable recommendation service would be able to inform her to change mobile operator or to take the next train, as it was able to find the factor, which contribute to her annoyance. As well, another QoE-enable services possible is tailored models based on crowdsourced data trained with federated learning. This service would push QoE models on demand, allowing for on-device inference. Each client could rate the current models until the system creates the best model for their profile (e.g., working commuter).

10.4 Conclusion

This thesis studied a user-centric approach to quantify and facilitate smartphone application QoE. It presents *expectQoE*, a QoE-enable system, including a QoE model trained on a plurality of external factors and end-user intents. The motivation came from the limitation observed in smartphone application QoE. That is the end-user is neglected, and their possible experience level is computed from

network test metrics only, limiting such models to a narrow area of smartphone applications (e.g., high bandwidth need as video). In this thesis, we tackled QoE quantification and facilitation problem centered around human factors.

To solve the QoE quantification and facilitation problem, we designed, developed, and evaluated the *expectQoE* system to implement an interactive method that aims to reduce smartphone user annoyance. We conducted two in-the-wild studies to explore the QoE smartphone applications within real-life scenarios, using a mixed methods approach to collect qualitative and quantitative data to assess our methods and systems. The facilitation is done via leveraging the expectation of the application users. The contextual factors were identified to build an accurate and context-aware QoE user-centric quantification model. We tested different methods to identify the influential factors and model structures due to the complexity of assessing QoE in-the-wild. The model was trained to predict the expected QoE. With the trained model implemented in *expectQoE*, we demonstrated the feasibility and influence of such system to decrease smartphone user annoyance in a low QoE context.

The scientific contribution of this thesis includes a prediction QoE model, executed by a platform portable system. Our method and system facilitates intervention toward limiting exposure to poor interaction with a smartphone application. Besides, we identified the factors influencing user QoE on smartphones, leveraging the day reconstruction method to collect qualitative data.

With the findings of this thesis, we can conclude that implementation of QoE-enable systems and services, such as *expectQoE*, in combination with user-centric expectation management, has the potential to affect smartphone users positively, due to the lingering effect of small annoyances, which has an higher bothering impact from short and nonfrequent low QoE events than the expected “normal” high QoE application usage.

References

- T. Abar, A. Ben Letaifa, and S. El Asmi. Heterogeneous Multiuser QoE Enhancement Over DASH in SDN Networks. *Wireless Personal Communications*, 114(4):2975–3001, Oct. 2020. ISSN 1572-834X. doi:10.1007/s11277-020-07513-w. URL <https://doi.org/10.1007/s11277-020-07513-w>.
- N. Abramson. THE ALOHA SYSTEM: another alternative for computer communications. *Fall Joint Computer Conference*, 37:281–285, Jan. 1977. doi:10.1145/1478462.1478502.
- V. Aggarwal, E. Halepovic, J. Pang, S. Venkataraman, and H. Yan. Prometheus: Toward Quality-of-Experience Estimation for Mobile Apps from Passive Network Measurements. *Proc. of the Workshop on Mobile Computing Systems and Applications - ACM HotMobile*, pages 1–6, 2014. doi:10.1145/2565585.2565600. URL <http://dl.acm.org/citation.cfm?id=2565585.2565600>. ISBN: 9781450327428.
- A. K. Akodu, S. R. Akinbo, and Q. O. Young. Correlation among smartphone addiction, craniovertebral angle, scapular dyskinesis, and selected anthropometric variables in physiotherapy undergraduates. *Journal of Taibah University Medical Sciences*, 13(6):528–534, Dec. 2018. ISSN 1658-3612. doi:10.1016/j.jtumed.2018.09.001.
- O. Alay, V. Mancuso, A. Brunstrom, S. Alfredsson, M. Mellia, G. Bernini, and H. Lonsethagen. End to End 5G Measurements with MONROE: Challenges and Opportunities. *IEEE 4th International Forum on Research and Technologies for Society and Industry, RTSI 2018 - Proceedings*, 2018. doi:10.1109/RTSI.2018.8548510. ISBN: 9781538662823.
- L. Amour, M. I. Boulabiar, S. Souihi, and A. Mellouk. An improved QoE estimation method based on QoS and affective computing. In *2018 International Symposium on Programming and Systems (ISPS)*, pages 1–6, Apr. 2018. doi:10.1109/ISPS.2018.8379009.
- Android. Android 10 Release Notes, 2022. URL <https://source.android.com/setup/start/android-10-release>.
- G. Argin, B. Pak, and H. Turkoglu. Between Post-Flâneur and Smartphone Zombie: Smartphone Users' Altering Visual Attention and Walking Behavior in Public Space. *ISPRS International Journal of Geo-Information*, 9(12):700, Dec. 2020. ISSN 2220-9964. doi:10.3390/ijgi9120700. URL <https://www.mdpi.com/2220-9964/9/12/700>. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.
- S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(none):40–79, Jan. 2010. ISSN 1935-7516. doi:10.1214/09-SS054. URL <https://projecteuclid.org/journals/statistics-surveys/volume-4/issue-none/A-survey-of-cross-validation-procedures-for-model-selection/10.1214/09-SS054.full>. Publisher: Amer. Statist. Assoc., the Bernoulli Soc., the Inst. Math. Statist., and the Statist. Soc. Canada.

- R. A. Armstrong. When to use the Bonferroni correction. *Ophthalmic & Physiological Optics: The Journal of the British College of Ophthalmic Opticians (Optometrists)*, 34(5):502–508, Sept. 2014. ISSN 1475-1313. doi:[10.1111/opo.12131](https://doi.org/10.1111/opo.12131).
- P. Arnau-Gonzalez, T. Althobaiti, S. Katsigiannis, and N. Ramzan. Perceptual video quality evaluation by means of physiological signals. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, May 2017. doi:[10.1109/QoMEX.2017.7965651](https://doi.org/10.1109/QoMEX.2017.7965651). ISSN: 2472-7814.
- J. Aru, R. Rutiku, M. Wibrál, W. Singer, and L. Melloni. Early effects of previous experience on conscious perception. *Neuroscience of Consciousness*, 2016(1), Jan. 2016. ISSN 2057-2107. doi:[10.1093/nc/niw004](https://doi.org/10.1093/nc/niw004). URL <https://doi.org/10.1093/nc/niw004>.
- G. Association. The State of Mobile Internet Connectivity 2019, 2019. URL <https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2019/07/GSMA-State-of-Mobile-Internet-Connectivity-Report-2019.pdf>.
- D. Astels. *Test Driven development: A Practical Guide*. Prentice Hall Professional Technical Reference, 2003. ISBN 978-0-13-101649-1.
- R. Baeza-Yates, D. Jiang, F. Silvestri, and B. Harrison. Predicting The Next App That You Are Going To Use. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15*, pages 285–294, New York, NY, USA, Feb. 2015. Association for Computing Machinery. ISBN 978-1-4503-3317-7. doi:[10.1145/2684822.2685302](https://doi.org/10.1145/2684822.2685302). URL <https://doi.org/10.1145/2684822.2685302>.
- N. Ballou, V. R. Warriar, and S. Deterding. Are You Open? A Content Analysis of Transparency and Openness Guidelines in HCI Journals. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*, pages 1–10, New York, NY, USA, May 2021. Association for Computing Machinery. ISBN 978-1-4503-8096-6. doi:[10.1145/3411764.3445584](https://doi.org/10.1145/3411764.3445584). URL <https://doi.org/10.1145/3411764.3445584>.
- C. G. Bampis and A. C. Bovik. Feature-based prediction of streaming video QoE: Distortions, stalling and memory. *Signal Processing: Image Communication*, 68:218–228, Oct. 2018. ISSN 0923-5965. doi:[10.1016/j.image.2018.05.017](https://doi.org/10.1016/j.image.2018.05.017). URL <https://www.sciencedirect.com/science/article/pii/S0923596518303679>.
- S. Baraković and L. Skorin-Kapov. Survey of research on Quality of Experience modelling for web browsing. *Quality and User Experience*, 2(1):6, July 2017. ISSN 2366-0147. doi:[10.1007/s41233-017-0009-2](https://doi.org/10.1007/s41233-017-0009-2). URL <https://doi.org/10.1007/s41233-017-0009-2>.
- J. E. Bardram. The CARP Mobile Sensing Framework – A Cross-platform, Reactive, Programming Framework and Runtime Environment for Digital Phenotyping. *arXiv:2006.11904 [cs]*, June 2020. URL <http://arxiv.org/abs/2006.11904>. arXiv: 2006.11904.
- A. Barea, X. Ferre, and L. Villarroel. Android vs. iOS Interaction Design Study for a Student Multiplatform App. In C. Stephanidis, editor, *HCI International 2013 - Posters' Extended Abstracts*, Communications in Computer and Information Science, pages 8–12, Berlin, Heidelberg, 2013. Springer. ISBN 978-3-642-39476-8. doi:[10.1007/978-3-642-39476-8_2](https://doi.org/10.1007/978-3-642-39476-8_2).
- L. W. Barsalou. *Perceptions of Perceptual Symbols*, 1999.
- E. Bañuelos-Lozoya, G. González-Serna, N. González-Franco, O. Fragosó-Díaz, and N. Castro-Sánchez. A Systematic Review for Cognitive State-Based QoE/UX Evaluation. *Sensors*, 21(10):3439, Jan. 2021. doi:[10.3390/s21103439](https://doi.org/10.3390/s21103439). URL <https://www.mdpi.com/1424-8220/21/10/3439>. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

- V. W. Berger and Y. Zhou. Kolmogorov–Smirnov Test: Overview. In *Wiley StatsRef: Statistics Reference Online*. American Cancer Society, 2014. ISBN 978-1-118-44511-2. doi:10.1002/9781118445112.stat06558. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat06558>. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat06558>.
- C. Bergmeir and J. M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, May 2012. ISSN 0020-0255. doi:10.1016/j.ins.2011.12.028. URL <https://www.sciencedirect.com/science/article/pii/S0020025511006773>.
- A. Berrocal, W. Concepcion, S. D. Dominicis, and K. Wac. Complementing Human Behavior Assessment by Leveraging Personal Ubiquitous Devices and Social Links: An Evaluation of the Peer-Ceived Momentary Assessment Method. *JMIR mHealth and uHealth*, 8(8):e15947, 2020a. doi:10.2196/15947. URL <https://mhealth.jmir.org/2020/8/e15947/>. Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.
- A. Berrocal, V. Manea, A. De Masi, and K. Wac. mQoL Lab: Step-by-Step Creation of a Flexible Platform to Conduct Studies Using Interactive, Mobile, Wearable and Ubiquitous Devices. *Procedia Computer Science*, 175:221–229, Jan. 2020b. ISSN 1877-0509. doi:10.1016/j.procs.2020.07.033. URL <http://www.sciencedirect.com/science/article/pii/S1877050920317130>.
- M. Bian and L. Leung. Linking Loneliness, Shyness, Smartphone Addiction Symptoms, and Patterns of Smartphone Use to Social Capital. *Social Science Computer Review*, 33(1):61–79, Feb. 2015. ISSN 0894-4393. doi:10.1177/0894439314528779. URL <https://doi.org/10.1177/0894439314528779>. Publisher: SAGE Publications Inc.
- K. Bilal and A. Erbad. Impact of Multiple Video Representations in Live Streaming: A Cost, Bandwidth, and QoE Analysis. In *2017 IEEE International Conference on Cloud Engineering (IC2E)*, pages 88–94, Apr. 2017. doi:10.1109/IC2E.2017.20.
- M. Bode and D. B. Kristensen. The digital doppelgänger within: A study on self-tracking and the quantified self movement. *Assembling Consumption: Researching actors, networks and markets*, pages 119–135, 2016. URL <https://portal.findresearcher.sdu.dk/en/publications/the-digital-doppelg%C3%A4nger-within-a-study-on-self-tracking-and-the->. Publisher: Routledge.
- M. Bosk, M. Gajić, S. Schwarzmann, S. Lange, R. Trivisonno, C. Marquezan, and T. Zinner. Using 5G QoS Mechanisms to Achieve QoE-Aware Resource Allocation. In *2021 17th International Conference on Network and Service Management (CNSM)*, pages 283–291, Oct. 2021. doi:10.23919/CNSM52442.2021.9615557. ISSN: 2165-963X.
- E. Boz, B. Finley, A. Oulasvirta, K. Killki, and J. Manner. Mobile QoE prediction in the field. *Pervasive and Mobile Computing*, 59:101039, Oct. 2019. ISSN 1574-1192. doi:10.1016/j.pmcj.2019.101039. URL <http://www.sciencedirect.com/science/article/pii/S1574119218307673>.
- M. L. Bracken and B. M. Waite. Self-Efficacy and Nutrition-Related Goal Achievement of MyFitnessPal Users. *Health Education & Behavior*, 47(5):677–681, Oct. 2020. ISSN 1090-1981. doi:10.1177/1090198120936261. URL <https://doi.org/10.1177/1090198120936261>. Publisher: SAGE Publications Inc.
- T. G. Browne. Biofeedback and Neurofeedback. In H. S. Friedman, editor, *Encyclopedia of Mental Health (Second Edition)*, pages 170–177. Academic Press, Oxford, Jan. 2016. ISBN 978-0-12-397753-3. doi:10.1016/B978-0-12-397045-9.00121-X. URL <https://www.sciencedirect.com/science/article/pii/B978012397045900121X>.

- J. Brownlee. Tune Hyperparameters for Classification Machine Learning Algorithms, Dec. 2019. URL <https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/>.
- C. Buckle. Which Smartphone Features Really Matter to Consumers?, Jan. 2019. URL <https://blog.gwi.com/chart-of-the-week/smartphone-features-consumers/>.
- M. Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer. Falling asleep with Angry Birds, Facebook and Kindle: a large scale study on mobile application usage. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 47–56, New York, NY, USA, Aug. 2011. Association for Computing Machinery. ISBN 978-1-4503-0541-9. doi:10.1145/2037373.2037383. URL <https://doi.org/10.1145/2037373.2037383>.
- J. A. Cannon-Bowers and E. Salas. Reflections on shared cognition. *Journal of Organizational Behavior*, 22(2):195–202, 2001. ISSN 1099-1379. doi:10.1002/job.82. Place: US Publisher: John Wiley & Sons.
- H. Cao and M. Lin. Mining smartphone data for app usage prediction and recommendations: A survey. *Pervasive and Mobile Computing*, 37:1–22, June 2017. ISSN 1574-1192. doi:10.1016/j.pmcj.2017.01.007. URL <http://www.sciencedirect.com/science/article/pii/S1574119217300421>.
- A. Capponi, C. Fiandrino, B. Kantarci, L. Foschini, D. Kliazovich, and P. Bouvry. A Survey on Mobile Crowdsensing Systems: Challenges, Solutions, and Opportunities. *IEEE Communications Surveys Tutorials*, 21(3):2419–2465, 2019. ISSN 1553-877X. doi:10.1109/COMST.2019.2914030. Conference Name: IEEE Communications Surveys Tutorials.
- X. Carbonell, A. Chamarro, U. Oberst, B. Rodrigo, and M. Prades. Problematic Use of the Internet and Smartphones in University Students: 2006–2017. *International Journal of Environmental Research and Public Health*, 15(3), Mar. 2018. ISSN 1661-7827. doi:10.3390/ijerph15030475. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5877020/>.
- J. M. Carroll and J. R. Olson. Chapter 2 - Mental Models in Human-Computer Interaction¹¹This chapter appeared in its entirety and is reprinted from *Mental Models in Human Computer Interaction: Research Issues about What the User of Software Knows*, J.M. Carroll and J.R. Olson, Editors,-The report of the workshop on software human factors: Users mental models, Nancy Anderson, chair, sponsored by the Committee on Human Factors, Commission on Behavioral and Social Sciences and Education, National Research Council, published by the National Academy Press, 1987. In M. Helander, editor, *Handbook of Human-Computer Interaction*, pages 45–65. North-Holland, Amsterdam, Jan. 1988. ISBN 978-0-444-70536-5. doi:10.1016/B978-0-444-70536-5.50007-5. URL <https://www.sciencedirect.com/science/article/pii/B9780444705365500075>.
- E. J. Caruana, M. Roman, J. Hernández-Sánchez, and P. Solli. Longitudinal studies. *Journal of Thoracic Disease*, 7(11):E537–E540, Nov. 2015. ISSN 2072-1439. doi:10.3978/j.issn.2072-1439.2015.10.63. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4669300/>.
- P. Casas. On the Analysis of Network Measurements Through Machine Learning: The Power of the Crowd. In *TMA 2018 - Proceedings of the 2nd Network Traffic Measurement and Analysis Conference*, 2018. ISBN 978-3-903176-09-6. doi:10.23919/TMA.2018.8506486.
- P. Casas, A. Sackl, S. Egger, and R. Schatz. YouTube & Facebook Quality of Experience in mobile broadband networks. *2012 IEEE Globecom Workshops, GC Wkshps 2012*, pages 1269–1274, 2012. doi:10.1109/GLOCOMW.2012.6477764. ISBN: 9781467349413.
- P. Casas, B. Gardlo, and M. Seufert. Taming QoE in cellular networks: From subjective lab studies to measurements in the field. *Network and Service*, 2015a. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7367364.

- P. Casas, R. Schatz, F. Wamser, M. Seufert, and R. Irmer. Exploring QoE in Cellular Networks : How Much Bandwidth do you Need for Popular Smartphone Apps ? In *AllThingsCellular*, pages 13–18, 2015b. ISBN 978-1-4503-3538-6. doi:[10.1145/2785971.2785978](https://doi.org/10.1145/2785971.2785978). Issue: 1.
- P. Casas, M. Varela, P. Fiadino, M. Schiavone, H. Rivas, and R. Schatz. On the analysis of QoE in cellular networks: From subjective tests to large-scale traffic measurements. *IWCMC 2015 - 11th International Wireless Communications and Mobile Computing Conference*, pages 37–42, 2015c. ISSN 0022-3417 (Print). doi:[10.1109/IWCMC.2015.7289054](https://doi.org/10.1109/IWCMC.2015.7289054). ISBN: 9781479953448.
- P. Casas, M. Seufert, F. Wamser, B. Gardlo, A. Sackl, R. Schatz, and others. Next to You: Monitoring Quality of Experience in Cellular Networks from the End-devices. *IEEE Transactions on Network and Service Management*, 4537(c):1–1, 2016. ISSN 1932-4537. doi:[10.1109/TNSM.2016.2537645](https://doi.org/10.1109/TNSM.2016.2537645). URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7423787>.
- P. Casas, A. D’Alconzo, F. Wamser, M. Seufert, B. Gardlo, A. Schwind, P. Tran-Gia, and R. Schatz. Predicting QoE in cellular networks using machine learning and in-smartphone measurements. *2017 9th International Conference on Quality of Multimedia Experience, QoMEX 2017*, 02152:3–8, 2017a. doi:[10.1109/QoMEX.2017.7965687](https://doi.org/10.1109/QoMEX.2017.7965687). ISBN: 9781538640241.
- P. Casas, J. Vanerio, and K. Fukuda. GML learning, a generic machine learning model for network measurements analysis. In *2017 13th International Conference on Network and Service Management (CNSM)*, pages 1–9, Nov. 2017b. doi:[10.23919/CNSM.2017.8255998](https://doi.org/10.23919/CNSM.2017.8255998). ISSN: 2165-963X.
- P. Casas, M. Seufert, N. Wehner, A. Schwind, and F. Wamser. Enhancing machine learning based QoE prediction by ensemble models. *Proceedings - International Conference on Distributed Computing Systems*, 2018-July:1642–1647, 2018. doi:[10.1109/ICDCS.2018.00186](https://doi.org/10.1109/ICDCS.2018.00186). ISBN: 9781538668719.
- P. Casas, S. Wassermann, N. Wehner, M. Seufert, and T. Hossfeld. Not all Web Pages are Born the Same Content Tailored Learning for Web QoE Inference. In *2022 IEEE International Symposium on Measurements & Networking (M&N)*, pages 1–6, July 2022. doi:[10.1109/MN55117.2022.9887781](https://doi.org/10.1109/MN55117.2022.9887781). ISSN: 2639-5061.
- O. Celenk, T. Bauschert, and M. Eckert. Machine Learning based KPI Monitoring of Video Streaming Traffic for QoE Estimation. *ACM SIGMETRICS Performance Evaluation Review*, 48(4):33–36, 2021. ISSN 0163-5999. doi:[10.1145/3466826.3466839](https://doi.org/10.1145/3466826.3466839). URL <https://doi.org/10.1145/3466826.3466839>.
- G. Cencetti, G. Santin, A. Longa, E. Pigani, A. Barrat, C. Cattuto, S. Lehmann, M. Salathe, and B. Lepri. Digital Proximity Tracing in the COVID-19 Pandemic on Empirical Contact Networks. *medRxiv*, page 2020.05.29.20115915, July 2020. doi:[10.1101/2020.05.29.20115915](https://doi.org/10.1101/2020.05.29.20115915). URL <https://www.medrxiv.org/content/10.1101/2020.05.29.20115915v2>. Publisher: Cold Spring Harbor Laboratory Press.
- A. Chamberlain, A. Crabtree, T. Rodden, M. Jones, and Y. Rogers. Research in the wild: understanding ‘in the wild’ approaches to design and development. In *Proceedings of the Designing Interactive Systems Conference, DIS ’12*, pages 795–796, New York, NY, USA, June 2012. Association for Computing Machinery. ISBN 978-1-4503-1210-3. doi:[10.1145/2317956.2318078](https://doi.org/10.1145/2317956.2318078). URL <https://doi.org/10.1145/2317956.2318078>.
- M. Chan. Mobile-mediated multimodal communications, relationship quality and subjective well-being: An analysis of smartphone use from a life course perspective. *Computers in Human Behavior*, 87:254–262, Oct. 2018. ISSN 0747-5632. doi:[10.1016/j.chb.2018.05.027](https://doi.org/10.1016/j.chb.2018.05.027). URL <http://www.sciencedirect.com/science/article/pii/S074756321830253X>.
- N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, June 2002. ISSN 1076-9757.

- J. Chen, J. Lieffers, A. Bauman, R. Hanning, and M. Allman-Farinelli. The use of smartphone health apps and other mobile health (mHealth) technologies in dietetic practice: a three country study. *Journal of Human Nutrition and Dietetics*, 30, Jan. 2017. doi:[10.1111/jhn.12446](https://doi.org/10.1111/jhn.12446).
- M. Chen, Y. Hao, S. Mao, D. Wu, and Y. Zhou. User Intent-Oriented Video QoE with Emotion Detection Networking. In *2016 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, Dec. 2016. doi:[10.1109/GLOCOM.2016.7842364](https://doi.org/10.1109/GLOCOM.2016.7842364).
- Q. A. Chen, H. Luo, S. Rosen, Z. M. Mao, K. Iyer, J. Hui, K. Sontineni, and K. Lau. QoE Doctor : Diagnosing Mobile App QoE with Automated UI Control and Cross-layer Analysis. In *Proceedings of the 2014 Conference on Internet Measurement Conference - IMC '14*, pages 151–164, New York, New York, USA, Nov. 2014. ACM Press. ISBN 978-1-4503-3213-2. doi:[10.1145/2663716.2663726](https://doi.org/10.1145/2663716.2663726). URL <http://dl.acm.org/citation.cfm?id=2663716.2663726>.
- T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, pages 785–794, New York, NY, USA, Aug. 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi:[10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785). URL <https://doi.org/10.1145/2939672.2939785>.
- M. Ciman and K. Wac. Individuals' Stress Assessment Using Human-Smartphone Interaction Analysis. *IEEE Transactions on Affective Computing*, 9(1):51–65, Jan. 2018. ISSN 1949-3045. doi:[10.1109/TAFFC.2016.2592504](https://doi.org/10.1109/TAFFC.2016.2592504). Conference Name: IEEE Transactions on Affective Computing.
- M. Ciman and K. Wac. Smartphones as Sleep Duration Sensors: Validation of the iSenseSleep Algorithm. *JMIR mHealth and uHealth*, 7(5):e11930, May 2019. doi:[10.2196/11930](https://doi.org/10.2196/11930). URL <https://mhealth.jmir.org/2019/5/e11930>. Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.
- M. Ciman, K. Wac, and O. Gaggi. iSenseStress: assessing stress through human-smartphone interaction analysis. In *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth '15*, pages 84–91, Brussels, BEL, May 2015. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). ISBN 978-1-63190-045-7.
- J. W. Creswell and A. Tashakkori. Editorial: Differing Perspectives on Mixed Methods Research. *Journal of Mixed Methods Research*, 1(4):303–308, Oct. 2007. ISSN 1558-6898. doi:[10.1177/1558689807306132](https://doi.org/10.1177/1558689807306132). URL <https://doi.org/10.1177/1558689807306132>. Publisher: SAGE Publications.
- D. Dablain, B. Krawczyk, and N. V. Chawla. DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2022. ISSN 2162-2388. doi:[10.1109/TNNLS.2021.3136503](https://doi.org/10.1109/TNNLS.2021.3136503). Conference Name: IEEE Transactions on Neural Networks and Learning Systems.
- S. J. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. page 2, San Jose, CA, Aug. 1992. doi:[10.1117/12.135952](https://doi.org/10.1117/12.135952). URL <http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.135952>.
- M. Dasari, S. Vargas, A. Bhattacharya, A. Balasubramanian, S. R. Das, and M. Ferdman. Impact of Device Performance on Mobile Internet QoE. In *Proceedings of the Internet Measurement Conference 2018, IMC '18*, pages 1–7, New York, NY, USA, Oct. 2018. Association for Computing Machinery. ISBN 978-1-4503-5619-0. doi:[10.1145/3278532.3278533](https://doi.org/10.1145/3278532.3278533). URL <https://doi.org/10.1145/3278532.3278533>.

- P. Dash and Y. C. Hu. How much battery does dark mode save? an accurate OLED display power profiler for modern smartphones. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '21, pages 323–335, New York, NY, USA, June 2021. Association for Computing Machinery. ISBN 978-1-4503-8443-8. doi:10.1145/3458864.3467682. URL <https://doi.org/10.1145/3458864.3467682>.
- B. Davison and H. Hirsh. Predicting sequences of user actions. *AAAI Technical Report*, pages 5–12, 1998. URL http://scholar.google.com/scholar?hl=en&q=personal+information+trail+machine+learning&btnG=Search&as_sdt=0,39&as_ylo=&as_vis=0#9.
- J. De Letter, A. Zheleva, M. Maes, A. All, L. De Marez, and W. Durnez. What did you expect? *Quality and User Experience*, 6(1):5, Apr. 2021. ISSN 2366-0147. doi:10.1007/s41233-021-00045-6. URL <https://doi.org/10.1007/s41233-021-00045-6>.
- A. De Masi and K. Wac. You're Using This App for What? A mQoL Living Lab Study. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp '18, pages 612–617, New York, NY, USA, Oct. 2018. Association for Computing Machinery. ISBN 978-1-4503-5966-5. doi:10.1145/3267305.3267544. URL <https://doi.org/10.1145/3267305.3267544>.
- A. De Masi and K. Wac. Predicting Quality of Experience of Popular Mobile Applications from a Living Lab Study. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, June 2019. doi:10.1109/QoMEX.2019.8743306. ISSN: 2472-7814.
- A. De Masi and K. Wac. Towards accurate models for predicting smartphone applications' QoE with data from a living lab study. *Quality and User Experience*, 5(1):10, Oct. 2020. ISSN 2366-0147. doi:10.1007/s41233-020-00039-w. URL <https://doi.org/10.1007/s41233-020-00039-w>.
- A. De Masi and K. Wac. The Importance of Smartphone Connectivity in Quality of Life. In K. Wac and S. Wulfovich, editors, *Quantifying Quality of Life: Incorporating Daily Life into Medicine*, Health Informatics, pages 523–551. Springer International Publishing, Cham, 2022. ISBN 978-3-030-94212-0. doi:10.1007/978-3-030-94212-0_23. URL https://doi.org/10.1007/978-3-030-94212-0_23.
- A. De Masi, M. Ciman, M. Gustarini, and K. Wac. mQoL smart lab: quality of life living lab for interdisciplinary experiments. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, pages 635–640, New York, NY, USA, Sept. 2016. Association for Computing Machinery. ISBN 978-1-4503-4462-3. doi:10.1145/2968219.2971593. URL <https://doi.org/10.1145/2968219.2971593>.
- K. De Moor, I. Ketyko, W. Joseph, T. Deryckere, L. De Marez, L. Martens, and G. Verleye. Proposed framework for evaluating quality of experience in a mobile, testbed-oriented living lab setting. *Mobile Networks and Applications*, 15(3):378–391, 2010. ISSN 1383469X. doi:10.1007/s11036-010-0223-0. ISBN: 1383-469X.
- T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens. Quantifying the influence of rebuffering interruptions on the user's quality of experience during mobile video watching. *IEEE Transactions on Broadcasting*, 59(1):47–61, 2013. ISSN 00189316. doi:10.1109/TBC.2012.2220231.
- B. De Ridder, B. Van Rompaey, J. K. Kampen, S. Haine, and T. Dilles. Smartphone Apps Using Photoplethysmography for Heart Rate Monitoring: Meta-Analysis. *JMIR cardio*, 2(1):e4, Feb. 2018. ISSN 2561-1011. doi:10.2196/cardio.8802.
- D. de Ridder, F. Kroese, and L. van Gestel. Nudgeability: Mapping Conditions of Susceptibility to Nudge Influence. *Perspectives on Psychological Science*, 17(2):346–359, Mar. 2022. ISSN 1745-6916. doi:10.1177/1745691621995183. URL <https://doi.org/10.1177/1745691621995183>. Publisher: SAGE Publications Inc.

- Deboitte. Global Mobile Consumer Survey 2018 - Results for Switzerland, 2018. URL <https://www2.deloitte.com/ch/en/pages/technology-media-and-telecommunications/articles/global-mobile-consumer-survey.html>.
- Z. Deljac and M. Randić. A method to minimize the maintenance costs of a broadband access network based on optimal steady-state probability mixture distribution of proactive and reactive maintenance. *Computers and Electrical Engineering*, 102:108280, Sept. 2022. ISSN 0045-7906. doi:10.1016/j.compeleceng.2022.108280. URL <https://www.sciencedirect.com/science/article/pii/S0045790622005080>.
- K. Dery, D. Kolb, and J. MacCormick. Working with connective flow: how smartphone use is evolving in practice. *European Journal of Information Systems*, 23(5):558–570, Sept. 2014. ISSN 0960-085X. doi:10.1057/ejis.2014.13. URL <https://doi.org/10.1057/ejis.2014.13>. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1057/ejis.2014.13>.
- A. Detti, M. Pomposini, N. Blefari-Melazzi, S. Salsano, and A. Bragagnini. Offloading cellular networks with Information-Centric Networking: The case of video streaming. In *2012 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 1–3, June 2012. doi:10.1109/WoWMoM.2012.6263734.
- A. K. Dey. Understanding and Using Context. *Personal and Ubiquitous Computing*, 5(1):4–7, Feb. 2001. ISSN 1617-4909. doi:10.1007/s007790170019. URL <https://doi.org/10.1007/s007790170019>.
- A. K. Dey, K. Wac, D. Ferreira, K. Tassini, J.-H. Hong, and J. Ramos. Getting closer: an empirical investigation of the proximity of user to their smart phones. In *Proceedings of the 13th international conference on Ubiquitous computing*, UbiComp '11, pages 163–172, New York, NY, USA, Sept. 2011. Association for Computing Machinery. ISBN 978-1-4503-0630-0. doi:10.1145/2030112.2030135. URL <https://doi.org/10.1145/2030112.2030135>. tex.ids: Dey2011b.
- A. Dix. Statistics for HCI: Making Sense of Quantitative Data. *Synthesis Lectures on Human-Centered Informatics*, 13(2):1–181, Apr. 2020. ISSN 1946-7680, 1946-7699. doi:10.2200/S00974ED1V01Y201912HCI044. URL <https://www.morganclaypool.com/doi/10.2200/S00974ED1V01Y201912HCI044>.
- T. M. T. Do and D. Gatica-Perez. Where and what: Using smartphones to predict next locations and applications in daily life. *Pervasive and Mobile Computing*, 12:79–91, June 2014. ISSN 1574-1192. doi:10.1016/j.pmcj.2013.03.006. URL <http://www.sciencedirect.com/science/article/pii/S1574119213000576>.
- Z. Du, D. Liu, and L. Yin. User in the loop: QoE-oriented optimization in communication and networks. In *2017 6th International Conference on Computer Science and Network Technology (ICCSNT)*, pages 420–424, Oct. 2017. doi:10.1109/ICCSNT.2017.8343731.
- J. Dunaway and S. Soroka. Smartphone-size screens constrain cognitive access to video news stories. *Information, Communication & Society*, 24(1):69–84, Jan. 2021. ISSN 1369-118X. doi:10.1080/1369118X.2019.1631367. URL <https://doi.org/10.1080/1369118X.2019.1631367>. Publisher: Routledge _eprint: <https://doi.org/10.1080/1369118X.2019.1631367>.
- U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J.-N. Antons, K. Y. Chan, N. Ramzan, and K. Brunnström. Psychophysiology-Based QoE Assessment: A Survey. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):6–21, Feb. 2017. ISSN 1941-0484. doi:10.1109/JSTSP.2016.2609843. Conference Name: IEEE Journal of Selected Topics in Signal Processing.

- P. R. Ericsson. Benefits of mobile communication in rural and developing areas, Sept. 2010. URL <https://www.ericsson.com/en/press-releases/2010/9/benefits-of-mobile-communication-in-rural-and-developing-areas>. Last Modified: 2017-06-22T14:00:11+00:00.
- B. Esteves, K. Fraser, S. Kulkarni, O. Conlan, and V. Rodríguez-Doncel. Extracting and Understanding Call-to-actions of Push-Notifications. In P. Rosso, V. Basile, R. Martínez, E. Métais, and F. Meziane, editors, *Natural Language Processing and Information Systems*, Lecture Notes in Computer Science, pages 147–159, Cham, 2022. Springer International Publishing. ISBN 978-3-031-08473-7. doi:10.1007/978-3-031-08473-7_14.
- Fairphone. Fairphone, 2021. URL <https://www.fairphone.com/fr/>.
- A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61:863–905, Apr. 2018. ISSN 1076-9757. doi:10.1613/jair.1.11192. URL <https://www.jair.org/index.php/jair/article/view/11192>.
- D. Ferreira, J. Goncalves, V. Kostakos, L. Barkhuus, and A. K. Dey. Contextual experience sampling of mobile application micro-usage. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, MobileHCI '14, pages 91–100, New York, NY, USA, Sept. 2014. Association for Computing Machinery. ISBN 978-1-4503-3004-6. doi:10.1145/2628363.2628367. URL <https://doi.org/10.1145/2628363.2628367>.
- D. Ferreira, V. Kostakos, and A. K. Dey. AWARE: Mobile Context Instrumentation Framework. *Frontiers in ICT*, 2, 2015. ISSN 2297-198X. doi:10.3389/fict.2015.00006. URL <https://www.frontiersin.org/articles/10.3389/fict.2015.00006/full>. Publisher: Frontiers.
- M. Fiedler, T. Hossfeld, and P. Tran-Gia. A generic quantitative relationship between Quality of Experience and Quality of Service. *Blekinge Tekniska hogskola*, 24(March):36–41, 2010. ISSN 0890-8044. doi:10.1109/MNET.2010.5430142. ISBN: 0890-8044 VO - 24.
- M. Furini, S. Mirri, M. Montangero, and C. Prandi. Privacy perception and user behavior in the mobile ecosystem. In *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good*, GoodTechs '19, pages 177–182, New York, NY, USA, Sept. 2019. Association for Computing Machinery. ISBN 978-1-4503-6261-0. doi:10.1145/3342428.3342690. URL <https://doi.org/10.1145/3342428.3342690>.
- M. Gao, B. Ai, and Y. Niu. QoE-Aware Coordinated Caching for Adaptive Video Streaming in High-speed Railways. In *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*, pages 1–6, Nov. 2020. doi:10.1109/VTC2020-Fall49728.2020.9348815. ISSN: 2577-2465.
- A. C. Garcia and S. Casas. Quality of Experience in Mobile Applications: A Systematic Mapping of Metrics and Tools. *International Journal of Interactive Mobile Technologies (ijIM)*, 14(08):126–139, May 2020. ISSN 1865-7923. URL <https://www.online-journals.org/index.php/i-jim/article/view/12819>. Number: 08.
- Gartner. Gartner Hype Cycle Research Methodology, 2021. URL <https://www.gartner.com/en/research/methodologies/gartner-hype-cycle>.
- S. Gonçalves, P. Dias, and A.-P. Correia. Nomophobia and lifestyle: Smartphone use and its relationship to psychopathologies. *Computers in Human Behavior Reports*, 2:100025, Aug. 2020. ISSN 2451-9588. doi:10.1016/j.chbr.2020.100025. URL <https://www.sciencedirect.com/science/article/pii/S2451958820300257>.

- Google. Material Design, 2014. URL <https://material.io/design>.
- Google. Jetpack Compose, 2018. URL <https://developer.android.com/jetpack/compose>.
- Google. Developer Policy Centre, 2019. URL <https://play.google.com/about/developer-content-policy/>.
- Google. Select a category and tags for your app or game - play console help, 2020. URL <https://support.google.com/googleplay/android-developer/answer/113475?hl=en>.
- Google. Modern Android Development, 2021. URL <https://developer.android.com/modern-android-development>.
- Google. Machine Learning - Build smarter apps with machine learning, 2022. URL <https://developer.android.com/ml>.
- R. v. d. Goorbergh, M. van Smeden, D. Timmerman, and B. Van Calster. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *arXiv:2202.09101 [stat]*, Feb. 2022. URL <http://arxiv.org/abs/2202.09101>. arXiv: 2202.09101.
- M. Gustarini and K. Wac. Ubiquitous inference of mobility state of human custodian in people-centric context sensing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 725–728, New York, NY, USA, Sept. 2012. Association for Computing Machinery. ISBN 978-1-4503-1224-0. doi:[10.1145/2370216.2370375](https://doi.org/10.1145/2370216.2370375). URL <https://doi.org/10.1145/2370216.2370375>.
- M. Gustarini, S. Ickin, and K. Wac. Evaluation of challenges in human subject studies ”in-the-wild” using subjects’ personal smartphones. *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication - UbiComp '13 Adjunct*, pages 1447–1456, 2013. doi:[10.1145/2494091.2496041](https://doi.org/10.1145/2494091.2496041). URL <http://dl.acm.org/citation.cfm?doid=2494091.2496041>. ISBN: 9781450322157.
- M. Gustarini, M. P. Scipioni, M. Fanourakis, and K. Wac. Differences in smartphone usage: Validating, evaluating, and predicting mobile user intimacy. *Pervasive and Mobile Computing*, 33:50–72, Dec. 2016a. ISSN 1574-1192. doi:[10.1016/j.pmcj.2016.06.003](https://doi.org/10.1016/j.pmcj.2016.06.003). URL <http://www.sciencedirect.com/science/article/pii/S1574119216300682>.
- M. Gustarini, K. Wac, and A. K. Dey. Anonymous smartphone data collection: factors influencing the users’ acceptance in mobile crowd sensing. *Personal and Ubiquitous Computing*, 20(1): 65–82, Feb. 2016b. ISSN 1617-4917. doi:[10.1007/s00779-015-0898-0](https://doi.org/10.1007/s00779-015-0898-0). URL <https://doi.org/10.1007/s00779-015-0898-0>.
- A. Gutierrez, R. G. Dreslinski, T. F. Wenisch, T. Mudge, A. Saidi, C. Emmons, and N. Paver. Full-system analysis and characterization of interactive smartphone applications. In *2011 IEEE International Symposium on Workload Characterization (IISWC)*, pages 81–90, Nov. 2011. doi:[10.1109/IISWC.2011.6114205](https://doi.org/10.1109/IISWC.2011.6114205).
- S. Haug, R. P. Castro, M. Kwon, A. Filler, T. Kowatsch, and M. P. Schaub. Smartphone use and smartphone addiction among young people in Switzerland. *Journal of Behavioral Addictions*, 4(4):299–307, Dec. 2015. ISSN 2063-5303. doi:[10.1556/2006.4.2015.037](https://doi.org/10.1556/2006.4.2015.037).
- J. Hausmann and K. Wac. Activity Level Estimator on a Commercial Mobile Phone: A Feasibility Study. In *Proc. International Workshop on Frontiers in Activity Recognition using Pervasive Sensing*, 2011.

- H. He and E. A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, Sept. 2009. ISSN 1558-2191. doi:[10.1109/TKDE.2008.239](https://doi.org/10.1109/TKDE.2008.239). Conference Name: IEEE Transactions on Knowledge and Data Engineering.
- A. Hegde, M. Vijayalakshmi, and G. Jayalaxmi. QoE Aware Video Adaptation For Video Streaming in 5G Networks. In *2022 IEEE International Conference on Data Science and Information System (ICDSIS)*, pages 1–8, July 2022. doi:[10.1109/ICDSIS55133.2022.9915912](https://doi.org/10.1109/ICDSIS55133.2022.9915912).
- J. M. Hektner, J. A. Schmidt, and M. Csikszentmihalyi. *Experience sampling method: Measuring the quality of everyday life*. Experience sampling method: Measuring the quality of everyday life. Sage Publications, Inc, Thousand Oaks, CA, US, 2007. ISBN 1-4129-2557-6 (Paperback); 1-4129-4923-8 (Hardcover); 9781412925570 (Paperback); 9781412925570 (Hardcover). Pages: xiii, 352.
- D. E. Hinkle, W. Wiersma, and S. G. Jurs. Applied statistics for the behavioral sciences, 2003. URL <http://catalog.hathitrust.org/api/volumes/oclc/50716608.html>. Place: Boston, Mass.; [London].
- K. Hornbæk. Some Whys and Hows of Experiments in Human–Computer Interaction. *Foundations and Trends® in Human–Computer Interaction*, 5(4):299–373, June 2013. ISSN 1551-3955, 1551-3963. doi:[10.1561/1100000043](https://doi.org/10.1561/1100000043). URL <https://www.nowpublishers.com/article/Details/HCI-043>. Publisher: Now Publishers, Inc.
- J. Hosek, P. Vajsar, L. Nagy, M. Ries, O. Galinina, S. Andreev, Y. Koucheryavy, Z. Sulc, P. Hais, and R. Penizek. Predicting user QoE satisfaction in current mobile networks. *2014 IEEE International Conference on Communications, ICC 2014*, pages 1088–1093, 2014. doi:[10.1109/ICC.2014.6883466](https://doi.org/10.1109/ICC.2014.6883466). ISBN: 9781479920037.
- T. Hosfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, and M. Fiedler. The memory effect and its implications on Web QoE modeling. *2011 23rd International Teletraffic Congress (ITC)*, pages 103–110, 2011. ISBN: 978-1-4577-1187-9.
- T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia. Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing. *IEEE Transactions on Multimedia*, 16(2), 2014. ISSN 15209210. doi:[10.1109/TMM.2013.2291663](https://doi.org/10.1109/TMM.2013.2291663). ISBN: 1520-9210 VO - 16.
- T. Hossfeld, P. E. Heegaard, M. Varela, and L. Skorin-Kapov. Confidence Interval Estimators for MOS Values. 2018. URL <http://arxiv.org/abs/1806.01126>. arXiv: 1806.01126.
- T. Hossfeld, P. E. Heegaard, M. Varela, L. Skorin-Kapov, and M. Fiedler. From QoS Distributions to QoE Distributions: a System’s Perspective. pages 1–7, 2020. URL <http://arxiv.org/abs/2003.12742>. arXiv: 2003.12742.
- T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel. Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force ”Crowdsourcing”. *COST Action IC1003 European Network on Quality of Experience in Multimedia Systems and Services*, 2014. URL <https://hal.archives-ouvertes.fr/hal-01078761>. tex.ids: hossfeldBestPracticesRecommendations2014.
- T. Hoßfeld, M. Fiedler, and J. Gustafsson. Betas: Deriving quantiles from MOS-QoS relations of IQX models for QoE management. In *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, pages 1011–1016, May 2017. doi:[10.23919/INM.2017.7987430](https://doi.org/10.23919/INM.2017.7987430).
- T. Hoßfeld, P. E. Heegaard, L. Skorin-Kapov, and M. Varela. Fundamental Relationships for Deriving QoE in Systems. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, June 2019. doi:[10.1109/QoMEX.2019.8743339](https://doi.org/10.1109/QoMEX.2019.8743339). ISSN: 2472-7814.

- J. Huang, F. Qian, Y. Guo, Y. Zhou, Q. Xu, Z. M. Mao, S. Sen, and O. Spatscheck. An in-depth study of LTE: effect of network protocol and application behavior on performance. *ACM SIGCOMM Computer Communication Review*, 43(4):363–374, Sept. 2013. ISSN 0146-4833. doi:10.1145/2534169.2486006. URL <http://dl.acm.org/citation.cfm?id=2534169.2486006>. Publisher: ACM ISBN: 978-1-4503-2056-6.
- K. Huang, C. Zhang, X. Ma, and G. Chen. Predicting mobile application usage using contextual information. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp '12*, pages 1059–1065, New York, NY, USA, Sept. 2012. Association for Computing Machinery. ISBN 978-1-4503-1224-0. doi:10.1145/2370216.2370442. URL <https://doi.org/10.1145/2370216.2370442>.
- A. Huet, A. Saverimoutou, Z. B. Houidi, H. Shi, S. Cai, J. Xu, B. Mathieu, and D. Rossi. Deployable models for approximating web QoE metrics from encrypted traffic. *IEEE Transactions on Network and Service Management*, pages 1–1, 2021. ISSN 1932-4537. doi:10.1109/TNSM.2021.3073672. Conference Name: IEEE Transactions on Network and Service Management.
- S. Ickin. *Quality of Experience on Smartphones Selim Ickin*. PhD thesis, 2015. ISBN: 9789172953031.
- S. Ickin, K. Wac, M. Fiedler, L. Janowski, H. Jin-Hyuk, and a. K. Dey. Factors influencing quality of experience of commonly used mobile applications. *Communications Magazine, IEEE*, 50(April): 48–56, 2012. ISSN 0163-6804. doi:10.1109/MCOM.2012.6178833. ISBN: 0163-6804.
- S. Ickin, K. Wac, and M. Fiedler. QoE-based energy reduction by controlling the 3g cellular data traffic on the smartphone. In *2013 22nd ITC Specialist Seminar on Energy Efficient and Green Networking (SSEEGN)*, pages 13–18, Nov. 2013. doi:10.1109/SSEEGN.2013.6705396.
- S. Ickin, M. Fiedler, K. Wac, P. Arlos, C. Temiz, and K. Mkocha. VLQoE : Video QoE instrumentation on the smartphone. *Multimedia tools and applications*, 74(2):381–411, 2015. doi:10.1007/s11042-014-1919-0. Publisher: Blekinge Institute of Technology, Department of Communication Systems Publisher: Springer.
- S. Ickin, M. Fiedler, and K. Vandikas. QoE Modeling on Split Features with Distributed Deep Learning. *Network*, 1(2):165–190, Sept. 2021. doi:10.3390/network1020011. URL <https://www.mdpi.com/2673-8732/1/2/11>. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.
- M. T. R. Insights. Connectivity and QoL, Oct. 2017. URL <https://www.technologyreview.com/2017/10/24/148193/connectivity-and-qol/>.
- ITU-T. P.913 : Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment, June 2021. URL <https://www.itu.int/rec/T-REC-P.913>.
- ITU-T Recommendation P.800.1. Mean Opinion Score (MOS) Terminology. pages 1–8, 2019.
- J. Iyengar and M. Thomson. QUIC: A UDP-Based multiplexed and secure transport, May 2021. URL <https://rfc-editor.org/rfc/rfc9000.txt>. Number: 9000 Series: Request for comments tex.howpublished: RFC 9000 tex.pagetotal: 151.
- U. Jekosch. Assigning Meaning to Sounds — Semiotics in the Context of Product-Sound Design. In J. Blauert, editor, *Communication Acoustics*, pages 193–221. Springer, Berlin, Heidelberg, 2005. ISBN 978-3-540-27437-7. doi:10.1007/3-540-27437-5_8. URL https://doi.org/10.1007/3-540-27437-5_8.
- W. Jiang, Z. Sarsenbayeva, N. van Berkel, C. Wang, D. Yu, J. Wei, J. Goncalves, and V. Kostakos. User Trust in Assisted Decision-Making Using Miniaturized Near-Infrared Spectroscopy. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*, pages 1–16, New York, NY, USA, May 2021. Association for Computing Machinery. ISBN 978-1-4503-8096-6. doi:10.1145/3411764.3445710. URL <https://doi.org/10.1145/3411764.3445710>.

- S. L. Jones, D. Ferreira, S. Hosio, J. Goncalves, and V. Kostakos. Revisitation analysis of smart-phone app use. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '15*, pages 1197–1208, New York, NY, USA, Sept. 2015. Association for Computing Machinery. ISBN 978-1-4503-3574-4. doi:10.1145/2750858.2807542. URL <https://doi.org/10.1145/2750858.2807542>.
- P. Juluri, V. Tamarapalli, and D. Medhi. Measurement of Quality of Experience of Video-on-Demand Services: A Survey. *IEEE Communications Surveys Tutorials*, 18(c):401–418, Jan. 2015. ISSN 1553-877X. doi:10.1109/COMST.2015.2401424. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7035000>.
- S. Jumisko-Pyykkö and T. Vainio. Framing the Context of Use for Mobile HCI. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 2(4):1–28, 2010. ISSN 1942-390X. URL https://econpapers.repec.org/article/iggjmhci0/v_3a2_3ay_3a2010_3ai_3a4_3ap_3a1-28.htm. Publisher: IGI Global.
- D. Kahneman and J. Riis. Living, and thinking about it: Two perspectives on life. *The science of well-being*, 1:285–304, 2005.
- D. Kahneman, E. Diener, and N. Schwarz, editors. *Well-Being: Foundations of Hedonic Psychology*. Russell Sage Foundation, 1999. ISBN 978-0-87154-424-7. URL <https://www.jstor.org/stable/10.7758/9781610443258>.
- D. Kahneman, A. B. Krueger, D. A. Schkade, N. Schwarz, and A. A. Stone. A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science*, 306(5702):1776–1780, Dec. 2004. ISSN 0036-8075, 1095-9203. doi:10.1126/science.1103572. URL <https://science.sciencemag.org/content/306/5702/1776>. Publisher: American Association for the Advancement of Science Section: Report.
- Y. Kang, Y. Zhou, M. Gao, Y. Sun, and M. R. Lyu. Experience Report: Detecting Poor-Responsive UI in Android Applications. In *2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE)*, pages 490–501, Oct. 2016. doi:10.1109/ISSRE.2016.16. ISSN: 2332-6549.
- K. Katevas, H. Haddadi, and L. Tokarchuk. SensingKit: Evaluating the Sensor Power Consumption in iOS Devices. In *2016 12th International Conference on Intelligent Environments (IE)*, pages 222–225, Sept. 2016. doi:10.1109/IE.2016.50. ISSN: 2472-7571.
- K. Katevas, I. Arapakis, and M. Pielot. Typical phone use habits: intense use does not predict negative well-being. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–13, Barcelona Spain, Sept. 2018. ACM. ISBN 978-1-4503-5898-9. doi:10.1145/3229434.3229441. URL <https://dl.acm.org/doi/10.1145/3229434.3229441>.
- R. Katz and F. Callorda. The economic contribution of broadband, digitization and ICT regulation, 2018. URL https://www.itu.int/en/ITU-D/Regulatory-Market/Documents/FINAL_1d_18-00513_Broadband-and-Digital-Transformation-E.pdf.
- E. Kaufmann, O. Cappé, and A. Garivier. On the Complexity of A/B Testing. In *Proceedings of The 27th Conference on Learning Theory*, pages 461–481. PMLR, May 2014. URL <https://proceedings.mlr.press/v35/kaufmann14.html>. ISSN: 1938-7228.
- F. Kaup, F. Fischer, and D. Hausheer. Measuring and predicting cellular network quality on trains. In *2017 International Conference on Networked Systems (NetSys)*, pages 1–8, Mar. 2017. doi:10.1109/NetSys.2017.7903960.
- K. Kavli. Kavli HUMAN Project, 2016. URL <http://kavlihumanproject.org>.

- E. A. Kensinger and S. Corkin. Effect of Negative Emotional Content on Working Memory and Long-Term Memory. *Emotion*, 3(4):378–393, 2003. ISSN 1931-1516. doi:[10.1037/1528-3542.3.4.378](https://doi.org/10.1037/1528-3542.3.4.378). Place: US Publisher: American Psychological Association.
- G. Keren. Between- or within-subjects design: A methodological dilemma. In *A handbook for data analysis in the behavioral sciences: Methodological issues*, pages 257–272. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1993. ISBN 978-0-8058-1036-3 978-0-8058-1037-0.
- I. Ketykó, K. De Moor, T. De Pessemier, A. J. Verdejo, K. Vanhecke, W. Joseph, L. Martens, and L. De Marez. QoE measurement of mobile YouTube video streaming. *Proceedings of the 3rd workshop on Mobile video delivery - MoViD '10*, page 27, 2010. doi:[10.1145/1878022.1878030](https://doi.org/10.1145/1878022.1878030). URL <http://portal.acm.org/citation.cfm?doid=1878022.1878030>. ISBN: 9781450301657.
- N. A. Khan, M. A. Habib, and S. Jamal. Effects of smartphone application usage on mobility choices. *Transportation Research Part A: Policy and Practice*, 132:932–947, Feb. 2020. ISSN 0965-8564. doi:[10.1016/j.tra.2019.12.024](https://doi.org/10.1016/j.tra.2019.12.024). URL <https://www.sciencedirect.com/science/article/pii/S0965856418302131>.
- S. Kiani Mehr, P. Jogalekar, and D. Medhi. Moving QoE for monitoring DASH video streaming: models and a study of multiple mobile clients. *Journal of Internet Services and Applications*, 12(1):1, Apr. 2021. ISSN 1869-0238. doi:[10.1186/s13174-021-00133-y](https://doi.org/10.1186/s13174-021-00133-y). URL <https://doi.org/10.1186/s13174-021-00133-y>.
- K. J. Kim and S. S. Sundar. Does Screen Size Matter for Smartphones? Utilitarian and Hedonic Effects of Screen Size on Smartphone Adoption. *Cyberpsychology, Behavior, and Social Networking*, 17(7):466–473, July 2014. ISSN 2152-2715. doi:[10.1089/cyber.2013.0492](https://doi.org/10.1089/cyber.2013.0492). URL <https://www.liebertpub.com/doi/full/10.1089/cyber.2013.0492>. Publisher: Mary Ann Liebert, Inc., publishers.
- M.-Y. Kim. The Effects of Smartphone Use on Life Satisfaction, Depression, Social Activity and Social Support of Older Adults. *Journal of the Korea Academia-Industrial cooperation Society*, 19(11):264–277, 2018. ISSN 1975-4701. doi:[10.5762/KAIS.2018.19.11.264](https://doi.org/10.5762/KAIS.2018.19.11.264). URL <https://www.koreascience.or.kr/article/JAKO201809454741298.page>. Publisher: The Korea Academia-Industrial cooperation Society.
- J. Kjeldskov and M. B. Skov. Was it worth the hassle? ten years of mobile HCI research discussions on lab and field evaluations. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, MobileHCI '14, pages 43–52, New York, NY, USA, Sept. 2014. Association for Computing Machinery. ISBN 978-1-4503-3004-6. doi:[10.1145/2628363.2628398](https://doi.org/10.1145/2628363.2628398). URL <https://doi.org/10.1145/2628363.2628398>.
- M. Koo and H. Skinner. Challenges of internet recruitment: a case study with disappointing results. *Journal of Medical Internet Research*, 7(1):e6, Mar. 2005. ISSN 1438-8871. doi:[10.2196/jmir.7.1.e6](https://doi.org/10.2196/jmir.7.1.e6).
- V. Kostakos, D. Ferreira, J. Goncalves, and S. Hosio. Modelling Smartphone Usage: A Markov State Transition Model. In *UbiComp*, pages 486–497, 2016. ISBN 978-1-4503-4461-6. doi:[10.1145/2971648.2971669](https://doi.org/10.1145/2971648.2971669). tex.ids= Kostakos2016.
- D. Kumar, S. Jeuris, J. E. Bardram, and N. Dragoni. Mobile and Wearable Sensing Frameworks for mHealth Studies and Applications: A Systematic Review. *ACM Transactions on Computing for Healthcare*, 2(1):8:1–8:28, Dec. 2021. ISSN 2691-1957. doi:[10.1145/3422158](https://doi.org/10.1145/3422158). URL <https://doi.org/10.1145/3422158>.
- M. Kwon, D.-J. Kim, H. Cho, and S. Yang. The Smartphone Addiction Scale: Development and Validation of a Short Version for Adolescents. *PLOS ONE*, 8(12):e83558, Dec. 2013a. ISSN 1932-6203. doi:[10.1371/journal.pone.0083558](https://doi.org/10.1371/journal.pone.0083558). URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0083558>. Publisher: Public Library of Science.

- M. Kwon, J.-Y. Lee, W.-Y. Won, J.-W. Park, J.-A. Min, C. Hahn, X. Gu, J.-H. Choi, and D.-J. Kim. Development and Validation of a Smartphone Addiction Scale (SAS). *PLOS ONE*, 8(2):e56936, Feb. 2013b. ISSN 1932-6203. doi:[10.1371/journal.pone.0056936](https://doi.org/10.1371/journal.pone.0056936). URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0056936>. Publisher: Public Library of Science.
- S. Lall, M. Agarwal, and R. Sivakumar. A YouTube Dataset with User-level Usage Data: Baseline Characteristics and Key Insights. In *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pages 1–7, June 2020. doi:[10.1109/ICC40277.2020.9148782](https://doi.org/10.1109/ICC40277.2020.9148782). ISSN: 1938-1883.
- R. Larson and P. A. E. G. Delespaul. Analyzing Experience Sampling data: A guidebook for the perplexed. In *The experience of psychopathology: Investigating mental disorders in their natural settings*, pages 58–78. Cambridge University Press, New York, NY, US, 1992. ISBN 978-0-521-40339-9. doi:[10.1017/CBO9780511663246.007](https://doi.org/10.1017/CBO9780511663246.007).
- N. Lathia, K. Rachuri, C. Mascolo, and G. Roussos. Open source smartphone libraries for computational social science. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*, UbiComp '13 Adjunct, pages 911–920, New York, NY, USA, Sept. 2013. Association for Computing Machinery. ISBN 978-1-4503-2215-7. doi:[10.1145/2494091.2497345](https://doi.org/10.1145/2494091.2497345). URL <https://doi.org/10.1145/2494091.2497345>.
- J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. M. T. Do, O. Dousse, J. Eberle, and M. Miettinen. From big smartphone data to worldwide research: The Mobile Data Challenge. *Pervasive and Mobile Computing*, 9(6):752–771, Dec. 2013. ISSN 1574-1192. doi:[10.1016/j.pmcj.2013.07.014](https://doi.org/10.1016/j.pmcj.2013.07.014). URL <http://www.sciencedirect.com/science/article/pii/S1574119213000965>.
- J. Lazar, J. H. Feng, and H. Hochheiser. Chapter 15 - Working with human subjects. In J. Lazar, J. H. Feng, and H. Hochheiser, editors, *Research Methods in Human Computer Interaction (Second Edition)*, pages 455–491. Morgan Kaufmann, Boston, Jan. 2017. ISBN 978-0-12-805390-4. doi:[10.1016/B978-0-12-805390-4.00015-7](https://doi.org/10.1016/B978-0-12-805390-4.00015-7). URL <https://www.sciencedirect.com/science/article/pii/B9780128053904000157>.
- P. Le Callet, S. Möller, and P. Andrew. Qualinet White Paper on Definitions of Quality of Experience. *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*, (March), 2012.
- S. Lee and H. Cha. User interface-level QoE analysis for Android application tuning. *Pervasive and Mobile Computing*, 40(C):382–396, Sept. 2017. ISSN 1574-1192. doi:[10.1016/j.pmcj.2017.04.004](https://doi.org/10.1016/j.pmcj.2017.04.004). URL <https://doi.org/10.1016/j.pmcj.2017.04.004>.
- J. Li, L. Krasula, P. L. Callet, Z. Li, and Y. Baveye. Quantifying the Influence of Devices on Quality of Experience for Video Streaming. In *2018 Picture Coding Symposium (PCS)*, pages 308–312, June 2018. doi:[10.1109/PCS.2018.8456304](https://doi.org/10.1109/PCS.2018.8456304). tex.ids= Li2018a ISSN: 2472-7822.
- T. Li, T. Xia, H. Wang, Z. Tu, S. Tarkoma, Z. Han, and P. Hui. Smartphone App Usage Analysis: Datasets, Methods, and Applications. *IEEE Communications Surveys Tutorials*, pages 1–1, 2022. ISSN 1553-877X. doi:[10.1109/COMST.2022.3163176](https://doi.org/10.1109/COMST.2022.3163176). Conference Name: IEEE Communications Surveys Tutorials.
- X. Li, X. Wang, and G. Xiao. A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications. *Briefings in Bioinformatics*, 20(1):178–189, Jan. 2019. ISSN 1477-4054. doi:[10.1093/bib/bbx101](https://doi.org/10.1093/bib/bbx101). URL <https://doi.org/10.1093/bib/bbx101>.
- Z.-X. Liao, Y.-C. Pan, W.-C. Peng, and P.-R. Lei. On mining mobile apps usage behavior for predicting apps usage in smartphones. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, CIKM '13, pages 609–618, New York, NY, USA, Oct. 2013. Association

- for Computing Machinery. ISBN 978-1-4503-2263-8. doi:10.1145/2505515.2505529. URL <https://doi.org/10.1145/2505515.2505529>.
- N. Liberati. Teledildonics and New Ways of “Being in Touch”: A Phenomenological Analysis of the Use of Haptic Devices for Intimate Relations. *Science and Engineering Ethics*, 23(3):801–823, June 2017. ISSN 1471-5546. doi:10.1007/s11948-016-9827-5. URL <https://doi.org/10.1007/s11948-016-9827-5>.
- M. Liebherr, P. Schubert, S. Antons, C. Montag, and M. Brand. Smartphones and attention, curse or blessing? - A review on the effects of smartphone usage on attention, inhibition, and working memory. *Computers in Human Behavior Reports*, 1:100005, Jan. 2020. ISSN 2451-9588. doi:10.1016/j.chbr.2020.100005. URL <https://www.sciencedirect.com/science/article/pii/S2451958820300051>.
- S. Linxen, C. Sturm, F. Brühlmann, V. Cassau, K. Opwis, and K. Reinecke. How WEIRD is CHI? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, number 143, pages 1–14. Association for Computing Machinery, New York, NY, USA, May 2021. ISBN 978-1-4503-8096-6. URL <https://doi.org/10.1145/3411764.3445488>.
- J. A. List, S. Sadoff, and M. Wagner. So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*, 14(4):439, Mar. 2011. ISSN 1573-6938. doi:10.1007/s10683-011-9275-7. URL <https://doi.org/10.1007/s10683-011-9275-7>.
- W. Liu, Y. Cao, and R. W. Proctor. How do app icon color and border shape influence visual search efficiency and user experience? Evidence from an eye-tracking study. *International Journal of Industrial Ergonomics*, 84:103160, July 2021. ISSN 0169-8141. doi:10.1016/j.ergon.2021.103160. URL <https://www.sciencedirect.com/science/article/pii/S0169814121000780>.
- Y. Liu, C. Xu, and S.-C. Cheung. Characterizing and detecting performance bugs for smartphone applications. In *Proceedings of the 36th International Conference on Software Engineering*, ICSE 2014, pages 1013–1024, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 978-1-4503-2756-5. doi:10.1145/2568225.2568229. URL <https://doi.org/10.1145/2568225.2568229>.
- E. Lowe-Calverley and H. M. Pontes. Challenging the Concept of Smartphone Addiction: An Empirical Pilot Study of Smartphone Usage Patterns and Psychological Well-Being. *Cyberpsychology, Behavior, and Social Networking*, 23(8):550–556, Aug. 2020. ISSN 2152-2715. doi:10.1089/cyber.2019.0719. URL <https://www.liebertpub.com/doi/full/10.1089/cyber.2019.0719>. Publisher: Mary Ann Liebert, Inc., publishers.
- E. H. C. Lu, Y. W. Lin, and J. B. Ciou. Mining mobile application sequential patterns for usage prediction. *Proceedings - 2014 IEEE International Conference on Granular Computing, GrC 2014*, pages 185–190, 2014. doi:10.1109/GRC.2014.6982832. Publisher: IEEE ISBN: 9781479954643.
- X. Lu, W. Ai, X. Liu, Q. Li, N. Wang, G. Huang, and Q. Mei. Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '16*, pages 770–780, New York, NY, USA, Sept. 2016. Association for Computing Machinery. ISBN 978-1-4503-4461-6. doi:10.1145/2971648.2971724. URL <https://doi.org/10.1145/2971648.2971724>.
- C. Lucchese, C. I. Muntean, F. M. Nardini, R. Perego, and S. Trani. RankEval: Evaluation and investigation of ranking models. *SoftwareX*, 12:100614, July 2020. ISSN 2352-7110. doi:10.1016/j.softx.2020.100614. URL <https://www.sciencedirect.com/science/article/pii/S2352711020303277>.
- K. Lukoff, C. Yu, J. Kientz, and A. Hiniker. What Makes Smartphone Use Meaningful or Meaningless? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):22:1–22:26, Mar. 2018. doi:10.1145/3191754. URL <https://doi.org/10.1145/3191754>.

- P. Lunde, B. B. Nilsson, A. Bergland, K. J. Kværner, and A. Bye. The Effectiveness of Smartphone Apps for Lifestyle Improvement in Noncommunicable Diseases: Systematic Review and Meta-Analyses. *Journal of Medical Internet Research*, 20(5):e9751, May 2018. doi:[10.2196/jmir.9751](https://doi.org/10.2196/jmir.9751). URL <https://www.jmir.org/2018/5/e162>. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.
- C. Luo, J. Goncalves, E. Velloso, and V. Kostakos. A Survey of Context Simulation for Testing Mobile Context-Aware Applications. *ACM Computing Surveys*, 53(1):21:1–21:39, Feb. 2020. ISSN 0360-0300. doi:[10.1145/3372788](https://doi.org/10.1145/3372788). URL <https://doi.org/10.1145/3372788>.
- B. M. Marlin and R. S. Zemel. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems, RecSys '09*, pages 5–12, New York, NY, USA, Oct. 2009. Association for Computing Machinery. ISBN 978-1-60558-435-5. doi:[10.1145/1639714.1639717](https://doi.org/10.1145/1639714.1639717). URL <https://doi.org/10.1145/1639714.1639717>.
- D. Marsden and D. Littler. Repertory grid technique – An interpretive research framework. *European Journal of Marketing*, 34(7):816–834, Jan. 2000. ISSN 0309-0566. doi:[10.1108/03090560010331261](https://doi.org/10.1108/03090560010331261). URL <https://doi.org/10.1108/03090560010331261>. Publisher: MCB UP Ltd.
- H. P. Martinez, G. N. Yannakakis, and J. Hallam. Don't classify ratings of affect; Rank Them! *IEEE Transactions on Affective Computing*, 5(3):314–326, 2014. ISSN 19493045. doi:[10.1109/TAFFC.2014.2352268](https://doi.org/10.1109/TAFFC.2014.2352268). Publisher: IEEE.
- A. Mathur, N. D. Lane, and F. Kawsar. Engagement-Aware Computing: Modelling User Engagement from Mobile Contexts. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '16*, pages 622–633, 2016. doi:[10.1145/2971648.2971760](https://doi.org/10.1145/2971648.2971760). URL <http://dl.acm.org/citation.cfm?doid=2971648.2971760>. ISBN: 9781450344616.
- M. H. Mazhar and Z. Shafiq. Real-time Video Quality of Experience Monitoring for HTTPS and QUIC. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 1331–1339, Apr. 2018. doi:[10.1109/INFOCOM.2018.8486321](https://doi.org/10.1109/INFOCOM.2018.8486321).
- D. McAran and N. Shaw. Insights from the Apple Human Interface Guidelines on Intuitive Interaction. In F. F.-H. Nah and K. Siau, editors, *HCI in Business, Government and Organizations*, Lecture Notes in Computer Science, pages 128–140, Cham, 2020. Springer International Publishing. ISBN 978-3-030-50341-3. doi:[10.1007/978-3-030-50341-3_11](https://doi.org/10.1007/978-3-030-50341-3_11).
- M. MediaLab. MIT Media Lab: Human Dynamics Group, 2016. URL <http://hd.media.mit.edu>.
- A. Mehrotra, M. Musolesi, R. Hendley, and V. Pejovic. Designing content-driven intelligent notification mechanisms for mobile applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '15*, pages 813–824, New York, NY, USA, Sept. 2015. Association for Computing Machinery. ISBN 978-1-4503-3574-4. doi:[10.1145/2750858.2807544](https://doi.org/10.1145/2750858.2807544). URL <https://doi.org/10.1145/2750858.2807544>.
- A. Mehrotra, V. Pejovic, J. Vermeulen, R. Hendley, and M. Musolesi. My Phone and Me: Understanding People's Receptivity to Mobile Notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1021–1032. Association for Computing Machinery, New York, NY, USA, May 2016. ISBN 978-1-4503-3362-7. URL <https://doi.org/10.1145/2858036.2858566>.
- V. Menkovski, G. Exarchakos, and A. Liotta. The value of relative quality in video delivery. *Journal of Mobile Multimedia*, 7(3):151–162, Sept. 2011. ISSN 1550-4646.

- K. Mitra, A. Zaslavsky, and C. Ahlund. Context-Aware QoE Modelling, Measurement, and Prediction in Mobile Computing Systems. *IEEE Transactions on Mobile Computing*, 14(5):920–936, 2015. ISSN 15361233. doi:[10.1109/TMC.2013.155](https://doi.org/10.1109/TMC.2013.155).
- A.-N. Moldovan and C. H. Muntean. QoE-aware video resolution thresholds computation for adaptive multimedia. In *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–6, June 2017. doi:[10.1109/BMSB.2017.7986152](https://doi.org/10.1109/BMSB.2017.7986152). ISSN: 2155-5052.
- A. Monge Roffarello and L. De Russis. The Race Towards Digital Wellbeing: Issues and Opportunities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-5970-2. URL <https://doi.org/10.1145/3290605.3300616>.
- G. S. Moreira, H. Jo, and J. Jeong. NAP: Natural App Processing for Predictive User Contexts in Mobile Smartphones. *Applied Sciences*, 10(19):6657, Jan. 2020. doi:[10.3390/app10196657](https://doi.org/10.3390/app10196657). URL <https://www.mdpi.com/2076-3417/10/19/6657>. Number: 19 Publisher: Multidisciplinary Digital Publishing Institute.
- L. G. Morrison, C. Hargood, V. Pejovic, A. W. A. Geraghty, S. Lloyd, N. Goodman, D. T. Michaelides, A. Weston, M. Musolesi, M. J. Weal, and L. Yardley. The Effect of Timing and Frequency of Push Notifications on Usage of a Smartphone-Based Stress Management Intervention: An Exploratory Trial. *PLOS ONE*, 12(1):e0169162, Jan. 2017. ISSN 1932-6203. doi:[10.1371/journal.pone.0169162](https://doi.org/10.1371/journal.pone.0169162). URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0169162>. Publisher: Public Library of Science.
- M. Mozetić, M. Stepanović, M. Milotić, and D. Živkov. Trajectory-Aware Buffering Strategy for HTTP adaptive streaming QoE enhancement for on-the-go users. In *2021 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–4, Jan. 2021. doi:[10.1109/ICCE50685.2021.9427606](https://doi.org/10.1109/ICCE50685.2021.9427606). ISSN: 2158-4001.
- A. Muntaner-Mas, V. A. Sanchez-Azanza, F. B. Ortega, J. Vidal-Conti, P. A. Borràs, J. Cantallops, and P. Palou. The effects of a physical activity intervention based on a fatness and fitness smartphone app for University students. *Health Informatics Journal*, 27(1):1460458220987275, Jan. 2021. ISSN 1460-4582. doi:[10.1177/1460458220987275](https://doi.org/10.1177/1460458220987275). URL <https://doi.org/10.1177/1460458220987275>. Publisher: SAGE Publications Ltd.
- P. G. Muscari. The Subjective Character of Experience. *The Journal of Mind and Behavior*, 6(4):577–597, 1985. ISSN 0271-0137. URL <https://www.jstor.org/stable/43853190>. Publisher: Institute of Mind and Behavior, Inc.
- A. Najjar, Y. Mualla, K. D. Singh, G. Picard, D. Calvaresi, A. Malhi, S. Galland, and K. Främling. One-to-Many Negotiation QoE Management Mechanism for End-User Satisfaction. *IEEE Access*, 9:59231–59243, 2021. ISSN 2169-3536. doi:[10.1109/ACCESS.2021.3071646](https://doi.org/10.1109/ACCESS.2021.3071646). Conference Name: IEEE Access.
- H. Nam, K. H. Kim, and H. Schulzrinne. QoE matters more than QoS: Why people stop watching cat videos. In *Proceedings - IEEE INFOCOM*, volume 2016-July, 2016. ISBN 978-1-4673-9953-1. doi:[10.1109/INFOCOM.2016.7524426](https://doi.org/10.1109/INFOCOM.2016.7524426). ISSN: 0743166X.
- A. Nandugudi, A. Maiti, T. Ki, F. Bulut, M. Demirbas, T. Kosar, C. Qiao, S. Y. Ko, and G. Challen. PhoneLab: A Large Programmable Smartphone Testbed. In *Proceedings of First International Workshop on Sensing and Big Data Mining, SENSEMINE'13*, pages 1–6, New York, NY, USA, Nov. 2013. Association for Computing Machinery. ISBN 978-1-4503-2430-4. doi:[10.1145/2536714.2536718](https://doi.org/10.1145/2536714.2536718). URL <https://doi.org/10.1145/2536714.2536718>.

- A. Narayanan, X. Zhang, R. Zhu, A. Hassan, S. Jin, X. Zhu, X. Zhang, D. Rybkin, Z. Yang, Z. M. Mao, F. Qian, and Z.-L. Zhang. A variegated look at 5G in the wild: performance, power, and QoE implications. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference, SIGCOMM '21*, pages 610–625, New York, NY, USA, Aug. 2021. Association for Computing Machinery. ISBN 978-1-4503-8383-7. doi:[10.1145/3452296.3472923](https://doi.org/10.1145/3452296.3472923). URL <https://doi.org/10.1145/3452296.3472923>.
- N. Natarajan, D. Shin, and I. S. Dhillon. Which app will you use next? collaborative filtering with interactional context. In *Proceedings of the 7th ACM conference on Recommender systems, RecSys '13*, pages 201–208, New York, NY, USA, Oct. 2013. Association for Computing Machinery. ISBN 978-1-4503-2409-0. doi:[10.1145/2507157.2507186](https://doi.org/10.1145/2507157.2507186). URL <https://doi.org/10.1145/2507157.2507186>.
- O. C. Novac, M. Novac, C. Gordan, T. Berczes, and G. Bujdosó. Comparative study of Google Android, Apple iOS and Microsoft Windows Phone mobile operating systems. In *2017 14th International Conference on Engineering of Modern Electric Systems (EMES)*, pages 154–159, June 2017. doi:[10.1109/EMES.2017.7980403](https://doi.org/10.1109/EMES.2017.7980403).
- A. Offenwanger, A. J. Milligan, M. Chang, J. Bullard, and D. Yoon. Diagnosing Bias in the Gender Representation of HCI Research Participants: How it Happens and Where We Are. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, number 399, pages 1–18. Association for Computing Machinery, New York, NY, USA, May 2021. ISBN 978-1-4503-8096-6. URL <https://doi.org/10.1145/3411764.3445383>.
- C. Olaverri-Monreal and J. Gonçalves. Capturing mental models to meet users expectations. In *2014 9th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–5, June 2014. doi:[10.1109/CISTI.2014.6877006](https://doi.org/10.1109/CISTI.2014.6877006). ISSN: 2166-0727.
- A. J. Oliner, A. P. Iyer, I. Stoica, E. Lagerspetz, and S. Tarkoma. Carat: collaborative energy diagnosis for mobile devices. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems, SenSys '13*, pages 1–14, New York, NY, USA, Nov. 2013. Association for Computing Machinery. ISBN 978-1-4503-2027-6. doi:[10.1145/2517351.2517354](https://doi.org/10.1145/2517351.2517354). URL <https://doi.org/10.1145/2517351.2517354>.
- K. Opoku Asare, A. Visuri, and D. S. T. Ferreira. Towards early detection of depression through smartphone sensing. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, UbiComp/ISWC '19 Adjunct*, pages 1158–1161, New York, NY, USA, Sept. 2019. Association for Computing Machinery. ISBN 978-1-4503-6869-8. doi:[10.1145/3341162.3347075](https://doi.org/10.1145/3341162.3347075). URL <https://doi.org/10.1145/3341162.3347075>.
- A. Oulasvirta, T. Rattenbury, L. Ma, and E. Raita. Habits make smartphone use more pervasive. *Personal and Ubiquitous Computing*, 16(1):105–114, Jan. 2012. ISSN 1617-4917. doi:[10.1007/s00779-011-0412-2](https://doi.org/10.1007/s00779-011-0412-2). URL <https://doi.org/10.1007/s00779-011-0412-2>.
- L. Pandey and A. S. Arif. Enabling Text Translation Using the Suggestion Bar of a Virtual Keyboard. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 4352–4357, Oct. 2020. doi:[10.1109/SMC42975.2020.9282879](https://doi.org/10.1109/SMC42975.2020.9282879). ISSN: 2577-1655.
- S. Petrangeli, T. Wauters, and F. D. Turck. QoE-Centric Network-Assisted Delivery of Adaptive Video Streaming Services. In *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, pages 683–688, Apr. 2019. ISSN: 1573-0077.
- G. Pibiri, C. Mc Goldrick, and M. Huggard. Expected Quality of Service (eQoS) A network metric for capturing end-user experience. In *IFIP Wireless Days*, pages 1–6. IEEE, Nov. 2012. ISBN 978-1-4673-4402-9. doi:[10.1109/WD.2012.6402873](https://doi.org/10.1109/WD.2012.6402873). URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6402873>.

- R. W. Picard. Affective computing: challenges. *International Journal of Human-Computer Studies*, 59(1):55–64, 2003. ISSN 1071-5819. doi:[10.1016/S1071-5819\(03\)00052-1](https://doi.org/10.1016/S1071-5819(03)00052-1). URL <https://www.sciencedirect.com/science/article/pii/S1071581903000521>.
- M. Pielot, B. Cardoso, K. Katevas, J. Serrà, A. Matic, and N. Oliver. Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):91:1–91:25, Sept. 2017. doi:[10.1145/3130956](https://doi.org/10.1145/3130956). URL <https://doi.org/10.1145/3130956>.
- J. Postel. Transmission control protocol. Technical Report 793, RFC Editor, Sept. 1981. URL <https://www.rfc-editor.org/rfc/rfc793.txt>. Series: Request for comments tex.howpublished: Internet Requests for Comments tex.howpublished: RFC 793 tex.pagetotal: 91.
- Z. Qin, Y. Wang, Y. Xia, H. Cheng, Y. Zhou, Z. Sheng, and V. C. M. Leung. Demographic information prediction based on smartphone application usage. In *2014 International Conference on Smart Computing*, pages 183–190, Nov. 2014. doi:[10.1109/SMARTCOMP.2014.7043857](https://doi.org/10.1109/SMARTCOMP.2014.7043857).
- K. U. Rehman Laghari and K. Connelly. Toward total quality of experience: A QoE model in a communication ecosystem. *IEEE Communications Magazine*, 50(4):58–65, Apr. 2012. ISSN 1558-1896. doi:[10.1109/MCOM.2012.6178834](https://doi.org/10.1109/MCOM.2012.6178834). Conference Name: IEEE Communications Magazine.
- A. M. Roffarello and L. De Russis. Towards detecting and mitigating smartphone habits. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '19 Adjunct, pages 149–152, New York, NY, USA, Sept. 2019. Association for Computing Machinery. ISBN 978-1-4503-6869-8. doi:[10.1145/3341162.3343770](https://doi.org/10.1145/3341162.3343770). URL <https://doi.org/10.1145/3341162.3343770>.
- A. M. Roffarello and L. De Russis. Understanding, Discovering, and Mitigating Habitual Smartphone Use in Young Adults. *ACM Transactions on Interactive Intelligent Systems*, 11(2):13:1–13:34, 2021. ISSN 2160-6455. doi:[10.1145/3447991](https://doi.org/10.1145/3447991). URL <https://doi.org/10.1145/3447991>.
- A. M. Roffarello and L. De Russis. Understanding and Streamlining App Switching Experiences in Mobile Interaction. *International Journal of Human-Computer Studies*, 158:102735, Feb. 2022. ISSN 1071-5819. doi:[10.1016/j.ijhcs.2021.102735](https://doi.org/10.1016/j.ijhcs.2021.102735). URL <https://www.sciencedirect.com/science/article/pii/S1071581921001531>.
- Y. Rogers and P. Marshall. Research in the Wild. *Synthesis Lectures on Human-Centered Informatics*, 10(3):i–97, Apr. 2017. ISSN 1946-7680. doi:[10.2200/S00764ED1V01Y201703HCI037](https://doi.org/10.2200/S00764ED1V01Y201703HCI037). URL <https://www.morganclaypool.com/doi/abs/10.2200/S00764ED1V01Y201703HCI037>. Publisher: Morgan & Claypool Publishers.
- H. E. Ross. On the possible relations between discriminability and apparent magnitude., Dec. 1997. URL <http://cogprints.org/697/>. Pages: 187-203 Volume: 50.
- W. B. Rouse and N. M. Morris. On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin*, 100(3):349–363, 1986. ISSN 1939-1455. doi:[10.1037/0033-2909.100.3.349](https://doi.org/10.1037/0033-2909.100.3.349). Place: US Publisher: American Psychological Association.
- J. Ruiz, Y. Li, and E. Lank. User-defined motion gestures for mobile interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 197–206, New York, NY, USA, May 2011. Association for Computing Machinery. ISBN 978-1-4503-0228-9. doi:[10.1145/1978942.1978971](https://doi.org/10.1145/1978942.1978971). URL <https://doi.org/10.1145/1978942.1978971>.

- A. Sackl and R. Schatz. Got what you want? Modeling expectations to enhance web QoE prediction. *2014 6th International Workshop on Quality of Multimedia Experience, QoMEX 2014*, (1):57–58, 2014a. doi:[10.1109/QoMEX.2014.6982291](https://doi.org/10.1109/QoMEX.2014.6982291). ISBN: 9781479965366.
- A. Sackl and R. Schatz. Evaluating the influence of expectations, price and content selection on video quality perception. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 93–98, Sept. 2014b. doi:[10.1109/QoMEX.2014.6982302](https://doi.org/10.1109/QoMEX.2014.6982302).
- A. Sackl, P. Zwickl, S. Egger, and P. Reichl. The role of cognitive dissonance for QoE evaluation of multimedia services. In *2012 IEEE Globecom Workshops*, pages 1352–1356, Dec. 2012. doi:[10.1109/GLOCOMW.2012.6477779](https://doi.org/10.1109/GLOCOMW.2012.6477779). ISSN: 2166-0077.
- A. Sackl, P. Casas, R. Schatz, L. Janowski, and R. Irmer. Quantifying the Impact of network bandwidth fluctuations and outages on Web QoE. In *QoMEX*, volume 20, pages 1–6, 2015. ISBN 978-1-4799-8958-4. doi:[10.1109/QoMEX.2015.7148078](https://doi.org/10.1109/QoMEX.2015.7148078).
- A. Sackl, B. Gardlo, and R. Schatz. Size does Matter. Comparing the Results of a Lab and a Crowdsourcing File Download QoE Study. pages 127–131, Aug. 2016. doi:[10.21437/pqs.2016-27](https://doi.org/10.21437/pqs.2016-27). URL http://www.isca-speech.org/archive/PQS_2016/abstracts/1570275415.html.
- A. Sackl, R. Schatz, and A. Raake. More than I ever wanted or just good enough? User expectations and subjective quality perception in the context of networked multimedia services. *Quality and User Experience*, 2(1):3, 2017. ISSN 2366-0139. doi:[10.1007/s41233-016-0004-z](https://doi.org/10.1007/s41233-016-0004-z). URL <http://link.springer.com/10.1007/s41233-016-0004-z>. tex.ids: Sackl2017a ISBN: 2366-0139, 2366-0147 publisher: Springer International Publishing.
- M. Samaha and N. S. Hawi. Relationships among smartphone addiction, stress, academic performance, and satisfaction with life. *Computers in Human Behavior*, 57:321–325, Apr. 2016. ISSN 0747-5632. doi:[10.1016/j.chb.2015.12.045](https://doi.org/10.1016/j.chb.2015.12.045). URL <http://www.sciencedirect.com/science/article/pii/S0747563215303162>.
- R. Schatz and S. Egger. Vienna surfing : assessing mobile broadband quality in the field. In *Proceedings of the first ACM SIGCOMM workshop on Measurements up the stack - W-MUST '11*, page 19, New York, New York, USA, Aug. 2011. ACM Press. ISBN 978-1-4503-0800-7. doi:[10.1145/2018602.2018608](https://doi.org/10.1145/2018602.2018608). URL <http://dl.acm.org/citation.cfm?id=2018602.2018608>.
- R. Schatz, S. Egger, and A. Platzer. Poor, good enough or even better? Bridging the gap between acceptability and QoE of mobile broadband data services. *IEEE International Conference on Communications*, (May 2014):6, 2011. ISSN 05361486. doi:[10.1109/icc.2011.5963220](https://doi.org/10.1109/icc.2011.5963220). ISBN: 9781612842332.
- D. Scholz, D. Raumer, P. Emmerich, A. Kurtz, K. Lesiak, and G. Carle. Performance Implications of Packet Filtering with Linux eBPF. In *2018 30th International Teletraffic Congress (ITC 30)*, volume 01, pages 209–217, Sept. 2018. doi:[10.1109/ITC30.2018.00039](https://doi.org/10.1109/ITC30.2018.00039).
- A. Schwind, C. Midoglu, O. Alay, C. Griwodz, and F. Wamser. Dissecting the performance of YouTube video streaming in mobile networks. *International Journal of Network Management*, 30(3):e2058, 2020. ISSN 1099-1190. doi:<https://doi.org/10.1002/nem.2058>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/nem.2058>. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/nem.2058>.
- S. C. Seow. *Designing and Engineering Time: The Psychology of Time Perception in Software*. Addison-Wesley Professional, Apr. 2008. ISBN 978-0-13-270251-5. Google-Books-ID: jyhezugDiNQC.
- A. Seufert, S. Schröder, and M. Seufert. Delivering User Experience over Networks: Towards a Quality of Experience Centered Design Cycle for Improved Design of Networked Applications. *SN Computer Science*, 2(6):463, Sept. 2021. ISSN 2661-8907. doi:[10.1007/s42979-021-00851-x](https://doi.org/10.1007/s42979-021-00851-x). URL <https://doi.org/10.1007/s42979-021-00851-x>.

- M. Seufert, S. Wassermann, and P. Casas. Considering User Behavior in the Quality of Experience Cycle: Towards Proactive QoE-Aware Traffic Management. *IEEE Communications Letters*, 23(7): 1145–1148, July 2019. ISSN 1558-2558. doi:[10.1109/LCOMM.2019.2914038](https://doi.org/10.1109/LCOMM.2019.2914038). Conference Name: IEEE Communications Letters.
- S. Shaheen, A. Cohen, and E. Martin. Smartphone App Evolution and Early Understanding from a Multimodal App User Survey. In G. Meyer and S. Shaheen, editors, *Disrupting Mobility: Impacts of Sharing Economy and Innovative Transportation on Cities*, Lecture Notes in Mobility, pages 149–164. Springer International Publishing, Cham, 2017. ISBN 978-3-319-51602-8. doi:[10.1007/978-3-319-51602-8_10](https://doi.org/10.1007/978-3-319-51602-8_10). URL https://doi.org/10.1007/978-3-319-51602-8_10.
- Z. Shen, K. Yang, W. Du, X. Zhao, and J. Zou. DeepAPP: a deep reinforcement learning framework for mobile application usage prediction. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems, SenSys '19*, pages 153–165, New York, NY, USA, Nov. 2019. Association for Computing Machinery. ISBN 978-1-4503-6950-3. doi:[10.1145/3356250.3360038](https://doi.org/10.1145/3356250.3360038). URL <https://doi.org/10.1145/3356250.3360038>.
- S. Shiffman, A. A. Stone, and M. R. Hufford. Ecological Momentary Assessment. *Annual Review of Clinical Psychology*, 4(1):1–32, 2008. ISSN 1548-5943. doi:[10.1146/annurev.clinpsy.3.022806.091415](https://doi.org/10.1146/annurev.clinpsy.3.022806.091415).
- C. Shin, J.-H. Hong, and A. K. Dey. Understanding and prediction of mobile application usage for smart phones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp '12*, pages 173–182, New York, NY, USA, Sept. 2012. Association for Computing Machinery. ISBN 978-1-4503-1224-0. doi:[10.1145/2370216.2370243](https://doi.org/10.1145/2370216.2370243). URL <https://doi.org/10.1145/2370216.2370243>.
- D. Shin. Cross-analysis of usability and aesthetic in smart devices: what influences users' preferences? *Cross Cultural Management: An International Journal*, 19(4):563–587, Oct. 2012. ISSN 1352-7606. doi:[10.1108/13527601211270020](https://doi.org/10.1108/13527601211270020). URL <http://www.emeraldinsight.com/doi/10.1108/13527601211270020>. Publisher: Emerald Group Publishing Limited.
- R. Shwartz-Ziv and A. Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81: 84–90, May 2022. ISSN 1566-2535. doi:[10.1016/j.inffus.2021.11.011](https://doi.org/10.1016/j.inffus.2021.11.011). URL <https://www.sciencedirect.com/science/article/pii/S1566253521002360>.
- L. Silver. Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally, Feb. 2019. URL <https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/>.
- E. Song, T. Pan, C. Jia, T. Huang, and Y. Liu. WebQMon.ai: Threshold-Oblivious On-Line Web QoE Assessment Using LSTM Neural Networks. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1033–1034, Apr. 2019. doi:[10.1109/INFOCOMW.2019.8845175](https://doi.org/10.1109/INFOCOMW.2019.8845175).
- C. Stanik, M. Haering, C. Jesdabodi, and W. Maalej. Which App Features Are Being Used? Learning App Feature Usages from Interaction Data. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*, pages 66–77, Aug. 2020. doi:[10.1109/RE48521.2020.00019](https://doi.org/10.1109/RE48521.2020.00019). ISSN: 2332-6441.
- Statista. Google Play Store: number of apps 2021, 2021a. URL <https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/>.
- Statista. Global YouTube viewing time share by device 2021, 2021b. URL <https://www.statista.com/statistics/1173543/youtube-viewing-time-share-device/>.
- Statista. Smartphone market share 2022, 2022. URL <https://www.statista.com/statistics/271496/global-market-share-held-by-smartphone-vendors-since-4th-quarter-2009/>.

- A. A. Stone and S. Shiffman. Ecological momentary assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine*, 16(3):199–202, 1994. ISSN 1532-4796(Electronic),0883-6612(Print). doi:[10.1093/abm/16.3.199](https://doi.org/10.1093/abm/16.3.199). Publisher: Lawrence Erlbaum Place: US.
- W. Strijbosch, O. Mitas, M. van Gisbergen, M. Doicaru, J. Gelissen, and M. Bastiaansen. From Experience to Memory: On the Robustness of the Peak-and-End-Rule for Complex, Heterogeneous Experiences. *Frontiers in Psychology*, 10:1705, 2019. ISSN 1664-1078. doi:[10.3389/fpsyg.2019.01705](https://doi.org/10.3389/fpsyg.2019.01705). URL <https://www.frontiersin.org/article/10.3389/fpsyg.2019.01705>.
- G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch. Detecting Causality in Complex Ecosystems. *Science*, 338(6106):496–500, Oct. 2012. doi:[10.1126/science.1227079](https://doi.org/10.1126/science.1227079). URL <https://www.science.org/lookup/doi/10.1126/science.1227079>. Publisher: American Association for the Advancement of Science.
- I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA, Dec. 2014. MIT Press.
- R. S. Sutton and A. G. Barto. *Reinforcement Learning, second edition: An Introduction*. MIT Press, Nov. 2018. ISBN 978-0-262-35270-3. Google-Books-ID: uWV0DwAAQBAJ.
- G. Szabó, S. Rácz, D. Bezzera, I. Nogueira, and D. Sadok. Media QoE enhancement With QUIC. In *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 219–220, Apr. 2016. doi:[10.1109/INFCOMW.2016.7562075](https://doi.org/10.1109/INFCOMW.2016.7562075).
- M. Tahaei and K. Vaniea. Recruiting Participants With Programming Skills: A Comparison of Four Crowdsourcing Platforms and a CS Student Mailing List. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–15, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9157-3. doi:[10.1145/3491102.3501957](https://doi.org/10.1145/3491102.3501957). URL <https://doi.org/10.1145/3491102.3501957>.
- N. Takeuchi, T. Mori, Y. Suzukamo, N. Tanaka, and S.-I. Izumi. Parallel processing of cognitive and physical demands in left and right prefrontal cortices during smartphone use while walking. *BMC Neuroscience*, 17(1):9, 2016. ISSN 1471-2202. doi:[10.1186/s12868-016-0244-0](https://doi.org/10.1186/s12868-016-0244-0). URL <https://doi.org/10.1186/s12868-016-0244-0>.
- C. Tauch and E. Kanjo. The roles of emojis in mobile phone notifications. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, pages 1560–1565, New York, NY, USA, Sept. 2016. Association for Computing Machinery. ISBN 978-1-4503-4462-3. doi:[10.1145/2968219.2968549](https://doi.org/10.1145/2968219.2968549). URL <https://doi.org/10.1145/2968219.2968549>.
- V. F. Taylor and I. Martinovic. To Update or Not to Update: Insights From a Two-Year Study of Android App Evolution. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, pages 45–57, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 978-1-4503-4944-4. doi:[10.1145/3052973.3052990](https://doi.org/10.1145/3052973.3052990). URL <https://doi.org/10.1145/3052973.3052990>.
- I. Telecommunication Union. E.800 : Definitions of terms related to quality of service. URL <https://www.itu.int/rec/T-REC-E.800-200809-I>.
- X. Teng, H. Pham, and D. R. Jeske. Reliability Modeling of Hardware and Software Interactions, and Its Applications. *IEEE Transactions on Reliability*, 55(4):571–577, Dec. 2006. ISSN 1558-1721. doi:[10.1109/TR.2006.884589](https://doi.org/10.1109/TR.2006.884589). Conference Name: IEEE Transactions on Reliability.

- L. F. Tiotsop, E. Masala, A. Aldahdooh, G. V. Wallendael, and M. Barkowsky. Computing Quality-of-Experience Ranges for Video Quality Estimation. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3, June 2019. doi:[10.1109/QoMEX.2019.8743303](https://doi.org/10.1109/QoMEX.2019.8743303). ISSN: 2472-7814.
- J. Titus. Google I/O 2017: Empowering developers to build the best experiences across platforms, May 2017. URL <https://android-developers.googleblog.com/2017/05/google-io-2017-empowering-developers-to.html>.
- C. Tossell, P. Kortum, A. Rahmati, C. Shepard, and L. Zhong. Characterizing web use on smartphones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2769–2778, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 978-1-4503-1015-4. doi:[10.1145/2207676.2208676](https://doi.org/10.1145/2207676.2208676). URL <https://doi.org/10.1145/2207676.2208676>.
- C. Tossell, P. Kortum, C. Shepard, A. Rahmati, and L. Zhong. Exploring Smartphone Addiction: Insights from Long-Term Telemetric Behavioral Measures. *International Journal of Interactive Mobile Technologies (ijIM)*, 9(2):37–43, Mar. 2015. ISSN 1865-7923. URL <https://online-journals.org/index.php/i-jim/article/view/4300>. Number: 2.
- J. A. Tran, K. S. Yang, K. Davis, and A. Hiniker. Modeling the Engagement-Disengagement Cycle of Compulsive Phone Use. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–14, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-5970-2. doi:[10.1145/3290605.3300542](https://doi.org/10.1145/3290605.3300542). URL <https://doi.org/10.1145/3290605.3300542>.
- D. Tsolkas, E. Liotou, N. Passas, and L. Merakos. A survey on parametric QoE estimation for popular services. *Journal of Network and Computer Applications*, 77(October 2016):1–17, 2017. ISSN 10958592. doi:[10.1016/j.jnca.2016.10.016](https://doi.org/10.1016/j.jnca.2016.10.016). URL <http://dx.doi.org/10.1016/j.jnca.2016.10.016>. Publisher: Elsevier.
- I. T. I. Union. *General Aspects of Quality of Service and Network Performance*, 1993.
- N. van Berkel, S. Dennis, M. Zyphur, J. Li, A. Heathcote, and V. Kostakos. Modeling interaction as a complex system. *Human-Computer Interaction*, 36(2021):1–27, Jan. 2020. ISSN 0737-0024, 1532-7051. doi:[10.1080/07370024.2020.1715221](https://doi.org/10.1080/07370024.2020.1715221). URL <https://www.tandfonline.com/doi/full/10.1080/07370024.2020.1715221>.
- M. H. van Velthoven, J. Powell, and G. Powell. Problematic smartphone use: Digital approaches to an emerging public health problem. *DIGITAL HEALTH*, 4:2055207618759167, Jan. 2018. ISSN 2055-2076. doi:[10.1177/2055207618759167](https://doi.org/10.1177/2055207618759167). URL <https://doi.org/10.1177/2055207618759167>. Publisher: SAGE Publications Ltd.
- M. M. P. Vanden Abeele. Digital Wellbeing as a Dynamic Construct. *Communication Theory*, 31(4):932–955, Nov. 2021. ISSN 1050-3293. doi:[10.1093/ct/qtaa024](https://doi.org/10.1093/ct/qtaa024). URL <https://doi.org/10.1093/ct/qtaa024>.
- C. L. Vaz, N. Carnes, B. Pousti, H. Zhao, and K. J. Williams. A randomized controlled trial of an innovative, user-friendly, interactive smartphone app-based lifestyle intervention for weight loss. *Obesity Science & Practice*, 7(5):555–568, 2021. ISSN 2055-2238. doi:[10.1002/osp4.503](https://doi.org/10.1002/osp4.503). URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/osp4.503>. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/osp4.503>.
- C. L. Ventola. Mobile Devices and Apps for Health Care Professionals: Uses and Benefits. *Pharmacy and Therapeutics*, 39(5):356–364, May 2014. ISSN 1052-1372. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4029126/>.

- A. J. Verdejo, K. De Moor, I. Ketyko, K. T. Nielsen, J. Vanattenhoven, T. De Pessemer, W. Joseph, L. Martens, and L. de Marez. QoE estimation of a location-based mobile game using on-body sensors and QoS-related data. In *2010 IFIP Wireless Days*, pages 1–5, Oct. 2010. doi:[10.1109/WD.2010.5657756](https://doi.org/10.1109/WD.2010.5657756). ISSN: 2156-972X.
- A. Visuri, N. van Berkel, T. Okoshi, J. Goncalves, and V. Kostakos. Understanding smartphone notifications' user interactions and content importance. *International Journal of Human-Computer Studies*, 128:72–85, Aug. 2019. ISSN 1071-5819. doi:[10.1016/j.ijhcs.2019.03.001](https://doi.org/10.1016/j.ijhcs.2019.03.001). URL <https://www.sciencedirect.com/science/article/pii/S1071581919300205>.
- K. Wac. From Quantified Self to Quality of Life. In H. Rivas and K. Wac, editors, *Digital Health: Scaling Healthcare to the World*, Health Informatics, pages 83–108. Springer International Publishing, Cham, 2018. ISBN 978-3-319-61446-5. doi:[10.1007/978-3-319-61446-5_7](https://doi.org/10.1007/978-3-319-61446-5_7). URL https://doi.org/10.1007/978-3-319-61446-5_7.
- K. Wac. Quality of Life Technologies. In M. Gellman, editor, *Encyclopedia of Behavioral Medicine*, pages 1–2. Springer, New York, NY, 2019. ISBN 978-1-4614-6439-6. doi:[10.1007/978-1-4614-6439-6_102013-1](https://doi.org/10.1007/978-1-4614-6439-6_102013-1). URL https://doi.org/10.1007/978-1-4614-6439-6_102013-1.
- K. Wac, A. van Halteren, and D. Konstantas. QoS-Predictions Service: Infrastructural Support for Proactive QoS- and Context-Aware Mobile Services (Position Paper). In R. Meersman, Z. Tari, and P. Herrero, editors, *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, Lecture Notes in Computer Science, pages 1924–1933, Berlin, Heidelberg, 2006. Springer. ISBN 978-3-540-48276-5. doi:[10.1007/11915072_100](https://doi.org/10.1007/11915072_100).
- K. Wac, M. Gustarini, J. Marchanoff, M. Fanourakis, C. Tsiourti, M. Ciman, J. Hausmann, and G. Pinar. Mqol: experiences of the 'mobile communications and computing for quality of life' living lab. *2015 17th International Conference on E-health Networking, Application Services (HealthCom)*, (i):177–181, 2015a. doi:[10.1109/HealthCom.2015.7454494](https://doi.org/10.1109/HealthCom.2015.7454494). ISBN: 978-1-4673-8325-7.
- K. Wac, G. Pinar, M. Gustarini, and J. Marchanoff. More mobile & not so well-connected yet: Users' mobility inference model and 6 month field study. In *2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pages 91–99. IEEE, Oct. 2015b. ISBN 978-1-4673-9283-9. doi:[10.1109/ICUMT.2015.7382411](https://doi.org/10.1109/ICUMT.2015.7382411). URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7382411>.
- K. Wac, G. Pinar, M. Gustarini, and J. Marchanoff. Smartphone users mobile networks quality provision and VoLTE intend: Six-months field study. In *2015 IEEE 16th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 1–9. IEEE, June 2015c. ISBN 978-1-4799-8461-9. doi:[10.1109/WoWMoM.2015.7158169](https://doi.org/10.1109/WoWMoM.2015.7158169). URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7158169>.
- E. A. Walelgne, A. S. Asrese, J. Manner, V. Bajpai, and J. Ott. Understanding Data Usage Patterns of Geographically Diverse Mobile Users. *IEEE Transactions on Network and Service Management*, pages 1–1, 2020. ISSN 1932-4537. doi:[10.1109/TNSM.2020.3037503](https://doi.org/10.1109/TNSM.2020.3037503). Conference Name: IEEE Transactions on Network and Service Management.
- W. P. Wallace. Review of the historical, empirical, and theoretical status of the von Restorff phenomenon. *Psychological Bulletin*, 63(6):410–424, 1965. ISSN 1939-1455. doi:[10.1037/h0022001](https://doi.org/10.1037/h0022001). Place: US Publisher: American Psychological Association.
- F. Wamser, M. Seufert, P. Casas, R. Irmer, P. Tran-Gia, and R. Schatz. YoMoApp: A tool for analyzing QoE of YouTube HTTP adaptive streaming in mobile networks. *2015 European Conference on Networks and Communications, EuCNC 2015*, pages 239–243, 2015. doi:[10.1109/EuCNC.2015.7194076](https://doi.org/10.1109/EuCNC.2015.7194076). ISBN: 9781467373593.

- S. Wassermann, N. Wehner, and P. Casas. Machine Learning Models for YouTube QoE and User Engagement Prediction in Smartphones. *ACM SIGMETRICS Performance Evaluation Review*, 46(3): 155–158, Jan. 2019. ISSN 0163-5999. doi:10.1145/3308897.3308962. URL <https://doi.org/10.1145/3308897.3308962>.
- A. Wattanapisit, C. H. Teo, S. Wattanapisit, E. Teoh, W. J. Woo, and C. J. Ng. Can mobile health apps replace GPs? A scoping review of comparisons between mobile apps and GP tasks. *BMC medical informatics and decision making*, 20(1):5, Jan. 2020a. ISSN 1472-6947. doi:10.1186/s12911-019-1016-4.
- A. Wattanapisit, T. Tuangratananon, and S. Wattanapisit. Usability and utility of eHealth for physical activity counselling in primary health care: a scoping review. *BMC family practice*, 21(1):229, 2020b. ISSN 1471-2296. doi:10.1186/s12875-020-01304-9.
- A. White, D. S. Thomas, N. Ezeanochie, and S. Bull. Health Worker mHealth Utilization: A Systematic Review. *Computers, informatics, nursing : CIN*, 34(5):206–213, May 2016. ISSN 1538-2931. doi:10.1097/CIN.000000000000231. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4860109/>.
- WHOQOL. WHOQOL - Measuring Quality of Life| The World Health Organization, 2012. URL <https://www.who.int/tools/whoqol>.
- R. Wu, J. Chen, C. Lu Wang, and L. Zhou. The influence of emoji meaning multiplicity on perceived online review helpfulness: The mediating role of processing fluency. *Journal of Business Research*, 141:299–307, Mar. 2022. ISSN 0148-2963. doi:10.1016/j.jbusres.2021.12.037. URL <https://www.sciencedirect.com/science/article/pii/S0148296321009516>.
- S. Wulfovich, M. Fiordelli, H. Rivas, W. Concepcion, and K. Wac. “I Must Try Harder”: Design Implications for Mobile Apps and Wearables Contributing to Self-Efficacy of Patients With Chronic Conditions. *Frontiers in Psychology*, 10, 2019. ISSN 1664-1078. doi:10.3389/fpsyg.2019.02388. URL <https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02388/full>. Publisher: Frontiers.
- J. Xia, G. Cheng, D. Guo, and X. Zhou. A QoE-Aware Service-Enhancement Strategy for Edge Artificial Intelligence Applications. *IEEE Internet of Things Journal*, 7(10):9494–9506, Oct. 2020a. ISSN 2327-4662. doi:10.1109/JIOT.2020.2996422. Conference Name: IEEE Internet of Things Journal.
- T. Xia, Y. Li, J. Feng, D. Jin, Q. Zhang, H. Luo, and Q. Liao. DeepApp: Predicting Personalized Smartphone App Usage via Context-Aware Multi-Task Learning. *ACM Transactions on Intelligent Systems and Technology*, 11(6):64:1–64:12, Oct. 2020b. ISSN 2157-6904. doi:10.1145/3408325. URL <https://doi.org/10.1145/3408325>.
- C. Xiang, D. Liu, S. Li, X. Zhu, Y. Li, J. Ren, and L. Liang. HiNextApp: A Context-Aware and Adaptive Framework for App Prediction in Mobile Systems. In *2017 IEEE Trustcom/BigDataSE/ICSS*, pages 776–783, Aug. 2017. doi:10.1109/Trustcom/BigDataSE/ICSS.2017.312. ISSN: 2324-9013.
- Y. Xu, M. Lin, H. Lu, G. Cardone, N. Lane, Z. Chen, A. Campbell, and T. Choudhury. Preference, context and communities: a multi-faceted approach to predicting smartphone app usage patterns. In *Proceedings of the 2013 International Symposium on Wearable Computers, ISWC '13*, pages 69–76, New York, NY, USA, Sept. 2013. Association for Computing Machinery. ISBN 978-1-4503-2127-3. doi:10.1145/2493988.2494333. URL <https://doi.org/10.1145/2493988.2494333>.
- J. Yablonski. *Laws of UX: Using Psychology to Design Better Products & Services*. May 2020. ISBN 978-1-4920-5531-0. URL https://shop.aer.io/oreilly/p/Laws_of_UX/9781492055266-9149.

- B. Yang, Z. Xing, X. Xia, C. Chen, D. Ye, and S. Li. Don't Do That! Hunting Down Visual Design Smells in Complex UIs Against Design Guidelines. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 761–772, May 2021. doi:[10.1109/ICSE43902.2021.00075](https://doi.org/10.1109/ICSE43902.2021.00075). ISSN: 1558-1225.
- H. Ye, R. J. Beamish, S. M. Glaser, S. C. H. Grant, C.-h. Hsieh, L. J. Richards, J. T. Schnute, and G. Sugihara. Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proceedings of the National Academy of Sciences*, 112(13):E1569–E1576, Mar. 2015. ISSN 0027-8424, 1091-6490. doi:[10.1073/pnas.1417063112](https://doi.org/10.1073/pnas.1417063112). URL <https://www.pnas.org/content/112/13/E1569>. Publisher: National Academy of Sciences Section: PNAS Plus.
- D. Yu, Y. Li, F. Xu, P. Zhang, and V. Kostakos. Smartphone App Usage Prediction Using Points of Interest. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4): 174:1–174:21, Jan. 2018. doi:[10.1145/3161413](https://doi.org/10.1145/3161413). URL <https://doi.org/10.1145/3161413>.
- Y. Yu, T. Xia, H. Wang, J. Feng, and Y. Li. Semantic-aware Spatio-temporal App Usage Representation via Graph Convolutional Network. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):101:1–101:24, Sept. 2020. doi:[10.1145/3411817](https://doi.org/10.1145/3411817). URL <https://doi.org/10.1145/3411817>.
- V. A. Zeithaml, L. L. Berry, and A. Parasuraman. The Nature and Determinants of Customer Expectations of Service. *Journal of the Academy of Marketing Science*, 21(1):1–12, Jan. 1993. ISSN 0092-0703. doi:[10.1177/0092070393211001](https://doi.org/10.1177/0092070393211001). URL <http://link.springer.com/10.1177/0092070393211001>.
- S. Zhao, J. Ramos, J. Tao, Z. Jiang, S. Li, Z. Wu, G. Pan, and A. K. Dey. Discovering Different Kinds of Smartphone Users Through Their Application Usage Behaviors. pages 498–509, 2016. ISBN: 9781450344616.
- S. Zhao, S. Li, J. Ramos, Z. Luo, Z. Jiang, A. K. Dey, and G. Pan. User profiling from their use of smartphone applications: A survey. *Pervasive and Mobile Computing*, 59:101052, Oct. 2019a. ISSN 1574-1192. doi:[10.1016/j.pmcj.2019.101052](https://doi.org/10.1016/j.pmcj.2019.101052). URL <https://www.sciencedirect.com/science/article/pii/S1574119219300124>.
- S. Zhao, Z. Luo, Z. Jiang, H. Wang, F. Xu, S. Li, J. Yin, and G. Pan. AppUsage2Vec: Modeling smartphone app usage for prediction. *Proceedings - International Conference on Data Engineering*, 2019-April(December):1322–1333, 2019b. ISSN 10844627. doi:[10.1109/ICDE.2019.00120](https://doi.org/10.1109/ICDE.2019.00120). ISBN: 9781538674741.
- L. Zimmermann. “Your Screen-Time App Is Keeping Track”: Consumers Are Happy to Monitor but Unlikely to Reduce Smartphone Usage. *Journal of the Association for Consumer Research*, 6(3):377–382, July 2021. ISSN 2378-1815. doi:[10.1086/714365](https://doi.org/10.1086/714365). URL <https://www.journals.uchicago.edu/doi/abs/10.1086/714365>. Publisher: The University of Chicago Press.
- X. Zou, W. Zhang, S. Li, and G. Pan. Prophet: what app you wish to use next. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication, UbiComp '13 Adjunct*, pages 167–170, New York, NY, USA, Sept. 2013. Association for Computing Machinery. ISBN 978-1-4503-2215-7. doi:[10.1145/2494091.2494146](https://doi.org/10.1145/2494091.2494146). URL <https://doi.org/10.1145/2494091.2494146>.

Appendices

Appendix A

Informed Consent Form: Study 1

<p style="text-align: center;">UNIVERSITY of GENEVA Center for Informatics Quality of Life Technologies Lab Switzerland Research Consent Form</p>	<p style="text-align: center;">Investigators</p> <p style="text-align: center;">Alexandre de Masi (PhD student) Marios Fanourakis (PhD student) Allan Berrocal (PhD student) Oscar Dabrowski (PhD student) Prof. Katarzyna Wac</p>
<p style="text-align: center;">Protocol Director : Assoc. Prof. Katarzyna Wac www.qol.unige.ch</p>	
<p>Protocol Title : <i>Studying the Subjective and Objective Experience of Mobile Applications Used in Different Contexts of Daily Life</i></p>	

DESCRIPTION: You are invited to participate in a research study on how you use and experience your smartphone. We will be assessing the impact of your smartphone usage on your expectations and experiences. By studying this, we hope to create guidelines to help designers of smartphone applications to provide a stress-free experience.

We will collect data from the Android OS smartphone you will carry for 28 days (4 weeks). The study will involve that you answer a short survey (appearing automatically on the screen of your mobile phone) throughout the day after a mobile application is used (email, web, gaming, etc.). The experiment constitutes of no more than 10-12 surveys being requested to be answered after the use of an application. The survey will ask you to rate your mobile application usage experience from bad (1) to excellent (5), if the usage went as expected, what action you were trying to accomplish with the application and if you can/want to tell us more about your smartphone experience at this particular point of time (via free text entry).

We will also collect automatically and unobtrusively on your phone the following information, specifically upon the changing context (e.g., smartphone indicates a movement, a WiFi network change, new application opened on a phone): current time, operator network's cell, cell network's signal strength, phone battery level, charging patterns, current running applications on the screen, operator network status and its performance, WiFi network status, headphone status, screen brightness level and orientation. The phone will also log the amount of data being sent and received on the network interface. We do not collect your detailed GPS-based location data, but we request the position in term of cell tower, enabling us to know the antenna signal power received by the smartphone. The content of visited websites or messages, and so on, will never be recorded, just the fact that you are using a particular application, e.g., a browser or a messaging application. The content of phone calls and messages sent/received will not be collected, neither will the phone numbers. No audio or video recordings will be made by us, only the logging of data on your mobile phone.

We require you to own an Android OS mobile phone (Android version 5.0) and use it like normal. We require that you answer the surveys appearing on your mobile phone to the best of your availability. If you choose to stop participating, you are free to do so at any time.

PRIVACY, THE USE AND STORAGE OF DATA: The information collected on you as a participant and the acquired data are confidential. The data will be used for research

<p style="text-align: center;">UNIVERSITY of GENEVA Center for Informatics Quality of Life Technologies Lab Switzerland Research Consent Form</p>	<p style="text-align: center;">Investigators</p> <p style="text-align: center;">Alexandre de Masi (PhD student) Marios Fanourakis (PhD student) Allan Berrocal (PhD student) Oscar Dabrowski (PhD student) Prof. Katarzyna Wac</p>
<p style="text-align: center;">Protocol Director : Assoc. Prof. Katarzyna Wac www.qol.unige.ch</p>	
<p>Protocol Title : <i>Studying the Subjective and Objective Experience of Mobile Applications Used in Different Contexts of Daily Life</i></p>	

purposes only. The analysis results might be the subject of scientific publications, always respecting the strict anonymity of participants.

Each participant is assigned a code number. No information identifying the person is attached to the data. Only project managers keep a list of the correspondence between the code and your identity (including your contact) along the project duration. The experimenter and project managers are strictly bound by professional secrecy with regard to data and connections between data and subjects. It will be stop being processed on 31 December 2021 and will be deleted in 31 December 2026 at the latest.

All the data will be archived on our secure servers, with access protection and separate backup, under the responsibility of the Protocol Director. In addition, at your written request, your data can be erased at any time without any negative consequences to you.

We also inform you that the anonymous raw recordings, excluding any data that could lead to recognizing you, might be made accessible on an Open Science platform. This would allow sharing data with other researchers as a collaborative research tool.

TIME INVOLVEMENT: Your participation will take approximately 4 weeks (i.e., 28 days). Your participation in this experiment will take at most 30 minutes per week via short maximum 2 minutes surveys that appear on your screen and you will fill during the day.

RISKS AND BENEFITS: The risks and discomfort associated with your participation in this study are no greater than those ordinarily encountered in daily life or while carrying a cell phone. The benefits, which may reasonably be expected to result from this study, are understanding of your smartphone usage experience and expectations and understanding of your own habits. We cannot and do not guarantee or promise that you will receive any benefits from this study.

PAYMENTS: You will not receive a payment for your participation. Legitimate participants will participate in the lottery of three small prizes in form of Internet coupons. A lottery will take place each 3 months and all winners will be informed via the application itself.

UNIVERSITY of GENEVA Center for Informatics Quality of Life Technologies Lab Switzerland Research Consent Form	Investigators Alexandre de Masi (PhD student) Marios Fanourakis (PhD student) Allan Berrocal (PhD student) Oscar Dabrowski (PhD student) Prof. Katarzyna Wac
Protocol Director : Assoc. Prof. Katarzyna Wac www.qol.unige.ch	
Protocol Title : <i>Studying the Subjective and Objective Experience of Mobile Applications Used in Different Contexts of Daily Life</i>	

PARTICIPANT'S RIGHTS: If you have read this form and have decided to participate in this project, please understand that your participation is voluntary and you have the right to withdraw consent or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled. Your data will be deleted if you withdraw your participation. You have the right to refuse to answer particular questions. Your individual privacy will be maintained in all published and written data resulting from the study. The results of this research study may be presented at scientific or professional meetings or published in scientific journals. You must be at least 18 years old to participate.

ACCESS TO RESULTS: In case of interest from you as for the results of the research, you can contact Protocol Director, Katarzyna Wac <Katarzyna.Wac@unige.ch>, +41 22 379 02 42 until 31 December 2021. Only overall results of the study will be able to be transmitted.

CONTACT INFORMATION: Questions: If you have any additional questions, concerns or complaints about this research, its procedures, risks and benefits, or if you wish to opt-out from the study, please contact the Protocol Director, Katarzyna Wac <Katarzyna.Wac@unige.ch>, +41 22 379 02 42.

INDEPENDENT CONTACT: The research has been approved by CUI/ISS of University of Geneva. If you are not satisfied with how this study is being conducted, or if you have any concerns, complaints, or general questions about the research or your rights as a participant, please contact Center for Informatics Director Giovanna Di Marzo <Giovanna.DiMarzo@unige.ch>, to speak to someone independent from the research team.

ENGAGEMENT OF THE PARTICIPANT

I have read and understood all the information above YES NO

I accept that my data will be used for scientific purposes and that the results of the research aggregating my data and the data of other participants will be published in scientific journals or books. The data will remain anonymous and no identity information will be given.

YES NO

<p>UNIVERSITY of GENEVA Center for Informatics Quality of Life Technologies Lab Switzerland Research Consent Form</p>	<p style="text-align: center;">Investigators</p> <p style="text-align: center;">Alexandre de Masi (PhD student) Marios Fanourakis (PhD student) Allan Berrocal (PhD student) Oscar Dabrowski (PhD student) Prof. Katarzyna Wac</p>
<p>Protocol Director : Assoc. Prof. Katarzyna Wac www.qol.unige.ch</p>	
<p>Protocol Title : <i>Studying the Subjective and Objective Experience of Mobile Applications Used in Different Contexts of Daily Life</i></p>	

I accept that my data and that the results of the research aggregating my data will be used for educational purposes (courses and seminars for the training of students or professionals subject to professional secrecy. YES NO

I authorize the sharing of my data with researchers from Quality of Life Technologies Lab, University of Geneva, Switzerland. YES NO

I have voluntarily chosen to participate in this research. I have been informed that I may withdraw at any time without providing any justification and may, if necessary, request the destruction of my personal data. This consent does not relieve the organizers of the research of their responsibilities. I retain all my rights guaranteed by law. YES NO

A signed and dated copy of this consent form is for you to keep.

PARTICIPANT SIGNATURE _____ DATE _____

ENGAGEMENT OF THE RESEARCHER

The information on this consent form and the answers I gave to the participant accurately describes the project. I undertake to conduct this study. I undertake that the research participant will receive a copy of this consent form.

UNIGE REPRESENTATIVE SIGNATURE _____

Appendix B

Informed Consent Form: Study 2

<p>UNIVERSITY of GENEVA Center for Informatics Quality of Life Technologies Lab Switzerland Research Consent Form</p>	<p>Investigators</p> <p>Alexandre De Masi (PhD student) Igor Matias (PhD student) Prof. Katarzyna Wac</p>
<p>Protocol Director : Prof. Katarzyna Wac www.qol.unige.ch</p>	
<p>Protocol Title : <i>Studying the Subjective and Objective Experience of Mobile Applications Used During The Day in Switzerland</i></p>	

DESCRIPTION: You are invited to participate in a research study on how you use and experience your smartphone during your day. We will be assessing the impact of your smartphone usage on your expectations and experiences. By studying this, we hope to create free guidelines to help designers of smartphone applications to provide a stress-free experience.

We will collect data from the Android OS smartphone you will carry for 28 days (4 weeks). The study will involve that you answer a short notification (appearing automatically on the screen of your mobile phone) throughout your day before and/or after a mobile application is used (email, web, gaming, etc.). The experiment consists of no more than 12 surveys being requested to be answered during the day. The notification contains possible quality of experience for multiple application categories. You will be asked to provide positive or negative feedback on the notification. This notification will start during the second week of the study.

Every evening for all 28 days, you will have to answer another type of survey. You will be asked about the actions you were trying to accomplish with the applications and if you can/want to tell us more about your overall experience of your smartphone (via free text entry) and your stress level.

We will also collect automatically and unobtrusively on your phone the following information, specifically upon the changing context (e.g., smartphone indicates a movement, a WiFi network change, new application opened on a phone): current time, location, cell network's signal strength, phone battery level, charging patterns, currently running applications on the screen, operator network status and its performance, WiFi network status, headphone status, application notification, screen status and orientation. The phone will also log the amount of data being sent and received on the network interface.

We do not collect your detailed GPS-based location data. The content of visited websites or messages, and so on, will never be recorded, just the fact that you are using a particular application, e.g., a browser or a messaging application. The content of phone calls and messages sent/received will not be collected, neither will the phone numbers. No audio or video recordings will be made by us, only the logging of data on your mobile phone.

We require you to own an Android OS mobile phone (Android version 9.0) and use it like normally you do. We require that you answer the surveys appearing on your mobile phone to

<p>UNIVERSITY of GENEVA Center for Informatics Quality of Life Technologies Lab Switzerland Research Consent Form</p>	<p style="text-align: center;">Investigators</p> <p style="text-align: center;">Alexandre De Masi (PhD student) Igor Matias (PhD student) Prof. Katarzyna Wac</p>
<p style="text-align: center;">Protocol Director : Prof. Katarzyna Wac www.qol.unige.ch</p>	
<p>Protocol Title : <i>Studying the Subjective and Objective Experience of Mobile Applications Used During The Day in Switzerland</i></p>	

the best of your availability. If you choose to stop participating, you are free to do so at any time.

PRIVACY, THE USE, AND STORAGE OF DATA: The information collected from you as a participant and the acquired data are confidential. The knowledge output generated from the data will be used for research purposes only. The analysis results might be the subject of scientific publications, always respecting the strict anonymity of participants.

Each participant is assigned a random identification number (RID). No information identifying the person is attached to the data. The experimenter and project managers are strictly bound by professional secrecy with regard to data and connections between data and subjects.

All the data will be archived on our secure servers, with access protection and separate backup, under the responsibility of the Protocol Director. In addition, at your written request (via anonymous email or letter containing your RID), your data can be erased at any time without any negative consequences to you.

We also inform you that the anonymous raw recordings, excluding any data that could lead to recognizing you, might be made accessible on an Open Science platform. This would allow sharing data with other researchers as a collaborative research tool.

TIME INVOLVEMENT: Your participation will take approximately 4 weeks (i.e., 28 days). Your participation in this experiment will take at most 5 minutes per week via short maximum 20 seconds surveys that appear on your screen and you will fill during the day.

RISKS AND BENEFITS: The risks and discomfort associated with your participation in this study are no greater than those ordinarily encountered in daily life or while carrying a cell phone. The benefits, which may reasonably be expected to result from this study, are understanding of your smartphone usage experience and expectations and understanding of your own habits. We cannot and do not guarantee or promise that you will receive any benefits from this study.

PAYMENTS: You will not receive payment for your participation.

PARTICIPANT'S RIGHTS: If you have read this form and have decided to participate in this

<p>UNIVERSITY of GENEVA Center for Informatics Quality of Life Technologies Lab Switzerland Research Consent Form</p>	<p>Investigators</p> <p>Alexandre De Masi (PhD student) Igor Matias (PhD student) Prof. Katarzyna Wac</p>
<p>Protocol Director : Prof. Katarzyna Wac www.qol.unige.ch</p>	
<p>Protocol Title : <i>Studying the Subjective and Objective Experience of Mobile Applications Used During The Day in Switzerland</i></p>	

project, please understand that your participation is voluntary and you have the right to withdraw consent or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled. If you decide to terminate or interrupt your participation in the study, you will be asked whether you authorize the researchers to keep and use the data already collected, or whether you prefer that your data be destroyed. In case you withdraw your participation and request your data to be deleted, you have one month after the last day of the study to send your request.

You have the right to refuse to answer particular questions. Your individual privacy will be maintained in all published and written data resulting from the study. The results of this research study may be presented at scientific or professional meetings or published in scientific journals. You must be at least 18 years old to participate (on the day of enrollment).

ACCESS TO RESULTS: In case of interest from you as for the results of the research, you can contact Protocol Director, Katarzyna Wac <Katarzyna.Wac@unige.ch>, +41 22 379 02 46 until 31 December 2021. Only the overall results of the study will be able to be transmitted.

CONTACT INFORMATION: Questions: If you have any additional questions, concerns, or complaints about this research, its procedures, risks, and benefits, or if you wish to opt out of the study, please contact the Protocol Director, Katarzyna Wac <Katarzyna.Wac@unige.ch>, +41 22 379 02 46.

INDEPENDENT CONTACT: The research has been approved by CUI/ISS of the University of Geneva. If you are not satisfied with how this study is being conducted, or if you have any concerns, complaints, or general questions about the research or your rights as a participant, please contact Center for Informatics Director Prof. Giovanna Di Marzo <Giovanna.DiMarzo@unige.ch>, to speak to someone independent from the research team.

ENGAGEMENT OF THE PARTICIPANT

I have read and understood all the information above

YES NO

I accept that my data will be used for scientific purposes and that the results of the research aggregating my data and the data of other participants will be published in scientific journals

<p>UNIVERSITY of GENEVA Center for Informatics Quality of Life Technologies Lab Switzerland Research Consent Form</p>	<p>Investigators</p> <p>Alexandre De Masi (PhD student) Igor Matias (PhD student) Prof. Katarzyna Wac</p>
<p>Protocol Director : Prof. Katarzyna Wac www.qol.unige.ch</p>	
<p>Protocol Title : <i>Studying the Subjective and Objective Experience of Mobile Applications Used During The Day in Switzerland</i></p>	

or books. I have understood that the data will remain anonymous and no identifiable information will be given. YES NO

I accept that my data and that the results of the research aggregating my data will be used for educational purposes (courses and seminars for the training of students or professionals subject to professional secrecy. YES NO

I have voluntarily chosen to participate in this research. I have been informed that I may withdraw at any time without providing any justification and may, if necessary, request the destruction of my collected data. This consent does not relieve the organizers of the research of their responsibilities. I retain all my rights guaranteed by law. YES NO

Signature:

Date:

