Date of acceptance Grade
September 2018

Instructor
Petteri Nurmi, Huber Flores

Analysis of the Impact of Performance on Apps Retention

Wladimir Agustín, Zúñiga Corrales

Helsinki September 27, 2018

UNIVERSITY OF HELSINKI
Department of Computer Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Computer Science

Tekijä — Författare — Author
Wladimir Agustín, Zúñiga Corrales

Työn nimi — Arbetets titel — Title

Analysis of the Impactof Performance on Apps Retention

Oppiaine — Läroämne — Subject
Computer Science

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu thesis	September 27, 2018	69 pages + 0 appendices

Tiivistelmä — Referat — Abstract

The non-stopping expansion of mobile technologies has produced the swift increase of smartphones with higher computational power and sophisticated sensing and communication capabilities have provided the foundations to develop apps on the move with PC-like functionality. Indeed, nowadays apps are almost everywhere and their number has increased exponentially with Apple AppStore, Google Play and other mobile app marketplaces offering millions of apps to users. In this scenario, it is common to find several apps providing similar functionalities to users. However, only a fraction of these applications have a long-term survival rate in app stores.

Retention is a metric widely used to quantify the lifespan of mobile apps. A higher app retention corresponds to higher adoption and level of engagement. While existing scientific studies have analysed mobile users' behaviour and support the existence of factors that influence apps retention, the quantification about how do these factors affect long-term usage is still missing. In this thesis, we contribute to these studies quantifying and modelling one of the critical factors that affects app retention: performance. We deepen the analysis of performance based on two key-related variables: network connectivity and battery consumption. The analysis is performed by combining two large-scale crowdsensed datasets. The first includes measurements about network quality and the second about app usage and energy consumption.

Our results show the benefits of data fusion to introduce richer contexts impossible of being discovered when analysing data sources individually. We also demonstrate that, indeed, high variations of these variables together and individually affect the likelihood of long-term app usage. But also, that retention is regulated by what users consider reasonable standards of performance, meaning that improvement of latency and energy consumption does not guarantee higher retention. To provide further insights, we develop a model to predict retention using performance-related variables. Its accuracy in the results allows generalising the effect of performance in long-term usage across categories, locations and moderating variables.

ACM Computing Classification System (CCS):
A.1 [Introductory and Survey],
G.3: [Probability and statistics],
I.6.4: [Model Validation and Analysis]

Avainsanat — Nyckelord — Keywords
Combined Data Analysis, Crowdsensing, Mobile Apps Retention and Performance.

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

1 Introduction

The continous proliferation of mobile technologies allows us to have mobile apps for almost everything. Nowadays, the number of apps has increased exponentially with Apple AppStore, Google Play and other mobile app marketplaces offering millions of apps to users[1]. In this scenario, it is common to find numerous apps having similar functionalities. However, only a fraction of them are used for long periods of time[2]. Specifically, studies on mobile app usage suggest that over a quarter of installed apps are only used once[3], and even apps used for more than a day are unlikely to stay relevant longer than a fortnight [SLP+18].

Retention is widely used to measure the success of mobile apps. It represents the fraction of users continuing to use an app after certain period of time since first use [Pan17]. While low retention of apps is well known[4], surprisingly little is known about factors that cause people to stop using applications. Indeed, existing work has mostly focus on exploring the usability factors that cause poor user experience and negative perception in app usage. Other work has explored partially the impact of performance in app retention but without quantifying how different variations in performance impact the behavior of users. For example, performance related characteristics and technical problems have been shown to be leading factors that affect emotional response of users and increase app's abandon rate[5]. Some studies complement this view by demonstrating that these factors are a significant source of frustration and a common complaint in app reviews [IWF+12, KSNH15]. Despite these efforts, there is still little understanding about how performance related factors influence the retention of apps.

Improving our understanding of how users are influence by variations in performance would be of significant academic and commercial interest [CLH+14a, FLL+13a].

[1] https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/
https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/

[2] https://andrewchen.co/new-data-shows-why-losing-80-of-your-mobile-users-is-normal-and-that-the-best-apps-do-much-better/

[3] http://info.localytics.com/blog/app-user-retention-improves-in-the-us

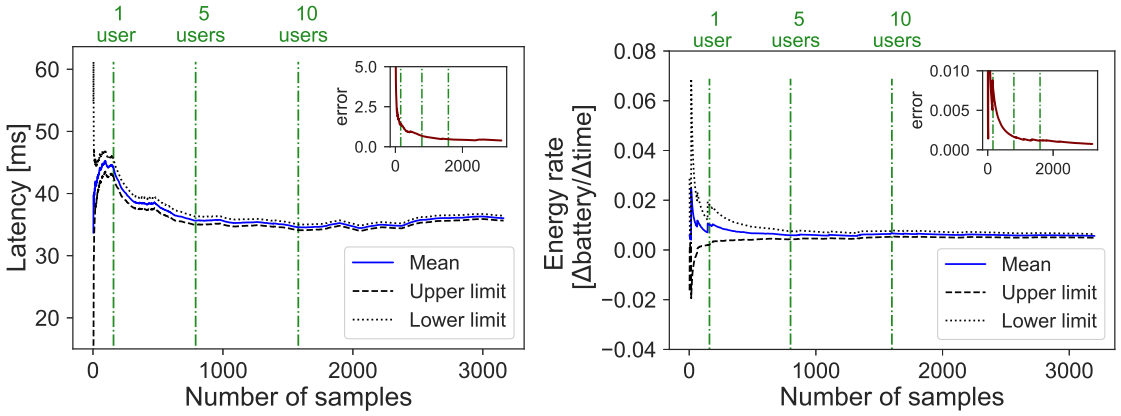
[4] http://andrewchen.co/new-data-shows-why-losing-80-of-your-mobile-users-is-normal-and-that-the-best-apps-do-much-better/
http://info.localytics.com/blog/the-8-mobile-app-metrics-that-matter

[5] https://blog.appdynamics.com/featured/deletetheapp-digital-leaders-raise-consumer-expectations-new-heights/

Moreover, it would deepen our knowledge about mobile interactions and how they are influenced by context [BJSH15, BHS+11, dRBHV12] and provide app developers with useful insights about the most relevant sensitive issues on performance that can be key to the long term success of short term failure of apps [ALvK+14, RPA+12]. More importantly, a better understanding about performance issues will improve the development of mobile apps in early stages, which in turn will minimise the cost of adapting and tuning the functionality of applications.

In the case of combined impact of performance factors, we have found no work that has attempt to quantify this combined effect and its influence of app usage. Indeed, existing partial work focus solely on a single factor. For instance, studies based on network level and active monitoring on the user's device [AHP+14a, BSA+13, GK15] ignore other factors that could also be perceived by the users at the same time. In the first case, network measurements are difficult to be associated with specific apps and do not consider factors like energy at all. Besides, they are influenced by mobility, network operator, communication technology and available network infrastructure [FHN+17]. In the second case, active client monitoring acquires a higher collection of performance factors and identifies apps associated with them. However, it depends on the user's context, i.e. if energy consumption is being affected by ambient temperature, mobility and network connectivity [AHP+14a, BSA+13, GK15].

While individual devices can collect relevant information, they are likely to cover only a small set of usage contexts. A large number of samples is required to increase the number of relevant contexts – something that would need long-time experiments or require heavy battery sampling when performed on individual devices [FHN+17]. We illustrate this in Figures 1a and 1b showing the level of accuracy of the characterisation based on the number of samples for two performance factors: network latency (Figure 1a) and energy consumption (Figure 1b). In both cases, we observe that hundreds, or even up to a thousand, samples are required for accurately characterising a performance variable — something that is unfeasible if we collect measurements from individual devices only. Crowdsensing is a potential way to overcome this problem. It is a data collection method that takes advantage of the popular existing apps' distribution channels, such as mobile app marketplaces, to collect information from mobile users devices such a way to increment the number of samples and enrich the diversity of contexts by increasing the number of the users that share their data. One of the issues of crowdsensing is the energy-side effects of collecting data from mobile device's sensors continuously. Piggybacking solve



(a) Latency

(b) Energy-intensiveness

Figure 1: Accuracy of mean estimate latency and energy-intensiveness as a number of samples and users.

this issue by implementing opportunistic collection, which means that sensing takes place only when the device is in use [LCZ+13]. Unlike in intrusive client-side monitoring, by piggybacking the overhead of data collection is minimised and separate instrumentation of the devices is avoided.

In this thesis, we address the gap of existing studies quantifying and modelling the effects of performance in apps retention, i.e., whether users are willing to continue using an app or not after experiencing variations in performance. We analyse two performance-related variables: latency and energy. Both variables are considered important factors to examine user's behaviour related to apps functioning [FLL+13a, IWF+12]. For example, energy consumption is recognised as one of the causes of frustration and stress, forcing users to take countermeasures to extend the battery life of their devices [BCCR14, PLNT16, RQZ07]. The high popularity of energy saving applications [OIS+13], and incorporation of energy draining alerts and control mechanisms in the new versions of mobile operative systems further show the importance of studying energy consumption. Latency, on the other hand, is a well-known factor affecting lag on user interactions and increasing frustration [CHL06, DSA+11].

We perform the analysis using two large-scale datasets of crowdsensed measurements collected by piggybacking. The first dataset contains information about network performance at different locations measured in terms of latency [SMS13] and the second about app usage and energy consumption [OIS+13]. Besides, in this thesis, we study the combined effect of the performance factors. To capture this effect, we

fuse the individual datasets using statistical survey analysis methods, specifically hot deck multiple imputation [Rei93, AL10]. We analyse samples from two locations, Finland and EST-USA, with different network and device characteristics. We limit our analysis to these locations as they have the highest combined number of samples across both datasets.

Our results about performance factors influencing app retention demonstrate that high performance decreases the likelihood of retaining an application. Reversely, we also show that improvement of latency or battery consumption does not guarantee the higher retention as long as performance remains reasonable. However, reasonable performance depends on what users are accustomed to experience, having the influence of factors like app categories, network features and location. For instance, different network infrastructures result in different expectations. We discover Finland to have a lower expected latency than EST-USA due to faster network connectivity. As part of the thesis we also develop a model that predicts retention based on the expected performance. We demonstrate the benefits of our model showing that it not only predicts retention accurately but also generalises well across location, app category and other factors moderating the effect of performance.

1.1 Goals

Based on the state of the art we outline the goals of this thesis. We summarise our goals as follows:

- Perform a large-scale analysis of mobile app performance using combined information from two big data independent crowdsensed datasets collected by piggybacking.
- Analyse how do performance variations affect users' behaviour measuring mobile apps' retention for different levels of latency and energy performance.
- Analyse the relationship between latency and energy performance and their combined effect on long-term app usage.
- Quantify and model the impact of performance on app retention both for individual and combined latency/energy effect.

1.2 Contributions

Our contributions encompass the goals of this thesis. The following sums up the contributions:

- We perform a large-scale analysis of mobile app performance not seen before in any previous studies. We carry out combined data analysis of two crowdsensed datasets showing the benefits of using crowdsensing and combined data to provide richer contexts of analysis.
- We quantify the effect of latency and energy consumption on apps retention. In general, high latency and high energy decrease app retention. However, we also show that higher latency or battery performance is not always a synonym of higher retention given that performance metrics are regulated by different factors.
- We demonstrate that latency and energy have a significant effect on mobile app retention and latency is usually the first critical factor perceived during the use of applications.
- We use our findings to build a model that uses performance metrics to predict the overall and the individual categories retention. In the first case, the model provides a good fit having an error of 1.4 percentage points (measured using Mean Absolute Error MAE). In the second case, accuracy depends on the amount of data and usage characteristics, with an error of 8.9 percentage points for the category with the lowest number of samples.

1.3 Outline

This thesis is structured as follows:

- Chapter 2 reviews the state-of-the-art about mobile apps usage, crowdsensing and combined data analysis.
- Chapter 3 describes the datasets, preprocessing steps and metrics used in this thesis.
- Chapter 4 analyses in detail the impact of latency and energy on mobile apps retention.
- Chapter 5 extends the results by developing a model for predicting retention based on performance metrics.
- Chapter 6 presents the summary and conclusions of this thesis.

2 State-of-the-Art

In this chapter, we review the recent research in areas that we consider relevant for our study. We start describing the studies related to mobile app quality to highlight the importance of identifying user's behaviour associated with the quality of experience. Then, we review work about mobile app performance as one of the critical factors contributing to long/low-term usage. Next, we go deeper into performance related variables that we use in this thesis. We introduce, first, the studies related to the effect of energy on retention and abandonment, and second, we do the same at the network infrastructure level. Next, we show a literature survey to validate the use of crowdsensed datasets in our analysis and the benefits of mobile crowdsensing to capture a richer amount of contexts. To finalise this review, we explore latest research that provides some insights about data fusion and combined data analytics.

2.1 Mobile App Quality

Main goals of the recent studies on mobile app quality are oriented to analyse the users' perception to mobile apps. Usability studies [RCT+07], contextual inquiries, e.g. using experience sampling [FGK+14], sensor data logging [ORMR12], interviews [IWF+12], and text mining on user reviews [FLL+13b] are the most used techniques.

According to the in-situ case study by Rogers et al. [RCT+07], their goal was to understand and improve the usability and situated user experience of a mobile learning device. They found out that many of the interface changes that were subsequently made to the application led to enhanced usability and encouraged to a different kind of user experience. Surprisingly, their study revealed how the environment, e.g. the time of year, can have an impact on the user experience.

Ferreira et al. [FGK+14] studied micro-usage of the apps by carrying out a study using the experience sampling method (ESM) tool to collect in situ real-time qualitative data of an application. By micro-usage is meant brief bursts of interaction with applications that tend to last less than 15 seconds. The goal of the study was to find out how users manage their time on the smartphone interacting with the device. Their findings suggest that about approximately 40% of application launches last less than 15 seconds and happen most frequently when the user is at home and alone. The most frequently used and micro-used applications were for

people's social connections. By studing and revealing reasons why some apps are used so briefly, Ferreira et al. contributed to a broader understanding of mobile phone usage practices.

User's perception is seen mainly as the essential part of existing studies, but the factors that affect it are normally ignored. According to Chen et al. [CLH+14b] app ratings play the most crucial role whether the app is downloaded or not. Complementary, Ickin et al. [IWF+12] show that the factors that influence quality perceptions are bugs, performance issues, and poor match with user needs. In general, app's perception is affected by these factors, but studies lack explaining the causes of changes in behaviour. This thesis deepens the analysis of these causes. We focus on the effect of performance-related variables and why should it be taken into account.

2.2 Mobile App Performance

In the previous subsection performance was shown as one of the key values affecting users' perception. The recent studies about the mobile app performance model are based on the steps employed by the user's task [RPA+12] and the response time perceived by the user. Network communication and processing costs are seen the most relevant aspects causing bottlenecks in app performance.

The research of Ravindranath et al. [RPA+12] introduces performance bottlenecks and failures with critical paths for user transactions and exception paths when apps fail during a transaction. The study diagnoses them with a system called AppInsight. It provides information about the critical path through the code for every user transaction to be used to improve the user experience. The critical path identifies the portions of the code that directly impact user-perceived latency. However, the critical path may not always accurately characterise user experience. Data from AppInsight shows that mobile apps have a tremendous amount of concurrency, with many asynchronous calls and several parallel threads in a typical user transaction.

Then again some of the studies have tried to relate network performance metrics with user satisfaction [GK15, AHP+14a]. They have focused on using machine learning to predict user response times and capturing features, such as bitrate, jitter and delay [BH10, BSA+13, NFS00]. Aggarwal et al. [AHP+14a] took the first step to address the challenges to understand how network-level parameters will influence a specific apps Quality of Experience (QoE) by presenting a novel approach to estimate app QoE using passive network measurements. According to their study

automatically learned QoE models are promising, g.e. semi-supervised learning approaches may be able to reduce data collection bias and improve accuracy on data from a real user.

There have been attempts to alleviate bottlenecks through different kinds of approaches for dynamic resource augmentation that rely only on the device's resources [SN01] or remote infrastructure [FSK+17]. The main focus of these studies has been modelling and improving performance, but the problem is that they do not measure the level at which it starts to influence user perceptions. This thesis explores the relationship between mobile app performance and app retention but also contributes to measuring the level at which performance effect becomes critical for users.

2.3 Effect of Energy

According to human interface studies, 80% of mobile users are willing to take action to improve their battery life [RQZ07]. Recent studies have shown that some of the causes of unnecessarily high energy consumption of applications can be for example the environment, settings of the smartphone [PLNT15a] or programming problems [PJHM12]. These so-called energy bugs, including system configurations, user behaviour, power-hungry apps etc., produce increased battery drain that causes frustration among users and can render devices unusable [OIS+13]. Energy-hungry applications that reduce battery life are more likely to get uninstalled by users [ALvK+14], or even replaced by a different app.

There are various causes for poor battery life and rapid energy drain, e.g. extensive use of resources by running applications or the device operating system itself, causing a need of recharge a mobile device even several times per day. Either way, rapid energy drain contributes negatively to the user experience [ALvK+14]. In order to prevent deleting an application, there are attempts trying to improve energy consumption by controlling processes on the device, or helping users to identify energy-hungry applications [OIS+13] and raising the level of energy awareness among users [ALvK+14].

The research of Rahmati et al. [RQZ07] is the first to study HBI, human-battery interaction, i.e. how human users deal with limited battery lifetime. Their work in understanding HBI provides a new approach for improving the usability of mobile phones by examining various aspects of HBI, including charging behaviour, bat-

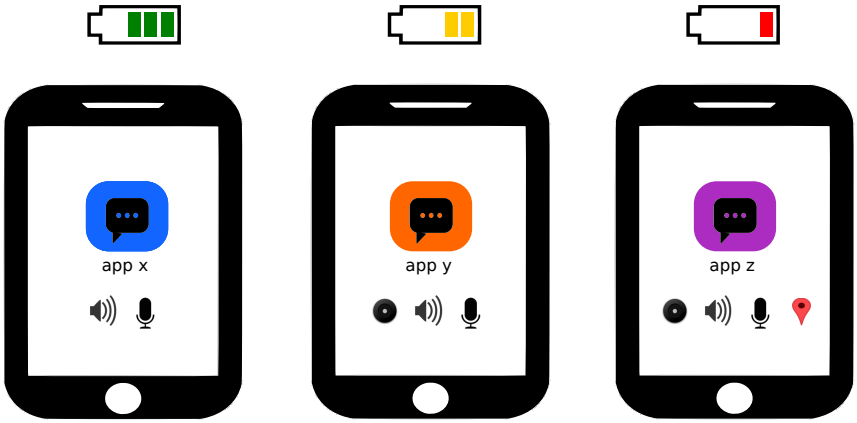


Figure 2: *Energy effect:* Battery consumption correlates with the amount of device's instrumentation used by the apps.

tery indicators, user interfaces for power-saving settings, user knowledge and user reaction. Complementary user studies show that mobile phone users take different strategies to deal with limited battery lifetime, and often they have inadequate knowledge of the power characteristics of a mobile phone. The users fall into two categories: those who regularly charge their phone regardless of the charge level and those who charge it based on charge level feedback from the phones battery interface. Their study shows that battery interfaces impose cognitively and technologically challenging loads to the user, leading to under-utilised power-saving settings, under-utilised battery energy and dissatisfied users. To improve HBI they suggest to providing intuitive and accurate battery interfaces and proper information through the user interface and user manual.

Indeed, energy load correlates with the usage of the device instrumentation required for running an application (figure 2). High energy consumption, however, could be produced for reasons not exclusively related to usage. Some apps continue draining considerable amounts of battery even when they are running in the background. In that sense, energy has been shown to be an active source of frustration and a cognitive burden as users actively seek to prolong their battery lifetime [BCCR14, PLNT16, RQZ07].

In general, while these studies describe the impact of energy performance in users behaviour and introduce strategies related to mitigate energy consumption, they do not quantify this effect. This thesis contributes to these studies considering energy as one of the performance-related variables which effect will be quantified and modelled.

2.4 Effect of Network Quality

Several studies show that network performance is possible of being measured using factors like throughput [NCKB+14] and signal strength [HQG+12a]. However, latency is considered as the most useful metric used for describing the network performance due it is associated with the usual complaints of the users. Latency influences the response time of mobile applications continuously affecting the observed performance. Latency represents the interval of time starting when a user makes a request in an application and ending when the user receives the response from the app [Bra91].

The influence of latency on user experience in desktop contexts and within specific application categories, e.g. online gaming [CHL06, TWLC11], education and video streaming [War09, MCC11] is the main focus of the previous studies of network quality. They have shown that latency affects the duration of session times, and the users tend to look actively for new ways to improve latency levels.

Among the users, latency has made real-time interaction and collaboration more difficult, e.g., in Second Life, which is a popular multi-user virtual world platform used in education. Second Life's visual experience is rendered in real time that places excessive stress on the graphic capabilities and bandwidth at the user end. The critical components of the end-user experience, specifically the frame rate, are often compromised which can lead to a situation called 'lag'. Lags disturb and frustrate users by slowing the experience [War09].

Game players struggle often with lag, but their perceptions and reactions are less studied. An Internet survey carried out by Po-Han Tsen et al. [TWLC11] aims to understand lag from the point of view of the players. Their findings show that players suffer from lag during gameplay and find it disturbing. Most of the players lack the required technical background to detect causes of lag and are looking for solutions, for example, a diagnostic tool, to identify and mitigate the problem.

While these studies show the effect of latency in desktop contexts, this cannot be necessarily applied to mobile context due to users are more dependent on the network quality. The abrupt changes in network quality depend on the technology (WiFi, 3G, LTE) and traffic conditions. The effect of changes in network quality also influences battery life, raising the impact of network quality on users.

Similarly to energy, in general, the studies do not quantify the effect of energy app retention. This thesis includes energy, the second performance related variable, to

be quantified and modelled.

2.5 Mobile Crowdsensing and User Behaviour

Crowdsensing is a method for acquiring useful information using a large number of users who share the data of their mobile devices [GYL11]. According to Ganti et al. [GYL11], it refers to the monitoring of large-scale phenomena where sensing is autonomous, and user involvement is minimal, although the applications collect sensitive sensor data of individuals(figure 3). In mobile crowdsensing, the population of mobile devices, the type of sensor data each can produce, and the quality regarding accuracy, latency and confidence can change all the time due to device mobility, variations in their energy levels and communication channels, and device owner's preferences. The challenge in identifying patterns from a large amount of data is usually application specific, and it involves specific data mining algorithms. The question of preserving the security and privacy of an individual, but at the same time enable applications data collection, is essential.

Characterisation of mobile malware incidence rates [TLN+14], identification of users' personality traits related with well-being perception [SRRM+17], and users' attitude towards the influence of context over different mobile resources and systems [FHN+17, PLNT15b] are the main trends in the recent research related with crowdsensing.

Peltonen et al. [PLNT15b] offer a novel approach for constructing energy models from crowdsourced measurements. It provides new insights into battery usage and demonstrates the complexity of the relationships between different factors and battery discharge. The large-scale dataset of crowdsourced battery discharge measurements was collected using a collaborative energy diagnostic system Carat. The results show the validity of using crowdsourced measurements for constructing battery models through a combination of large-scale analysis of a dataset containing battery discharge, system state and hardware power measurements. The models constructed by their approach capture the combined effects of multiple factors simultaneously and provide a characterisation of the energy state of a mobile device. This study offers a cost-effective alternative for modelling battery consumption that captures complex interdependencies affecting battery consumption in everyday use and can be used to understand long-term effects of sensor and battery management strategies on battery life.

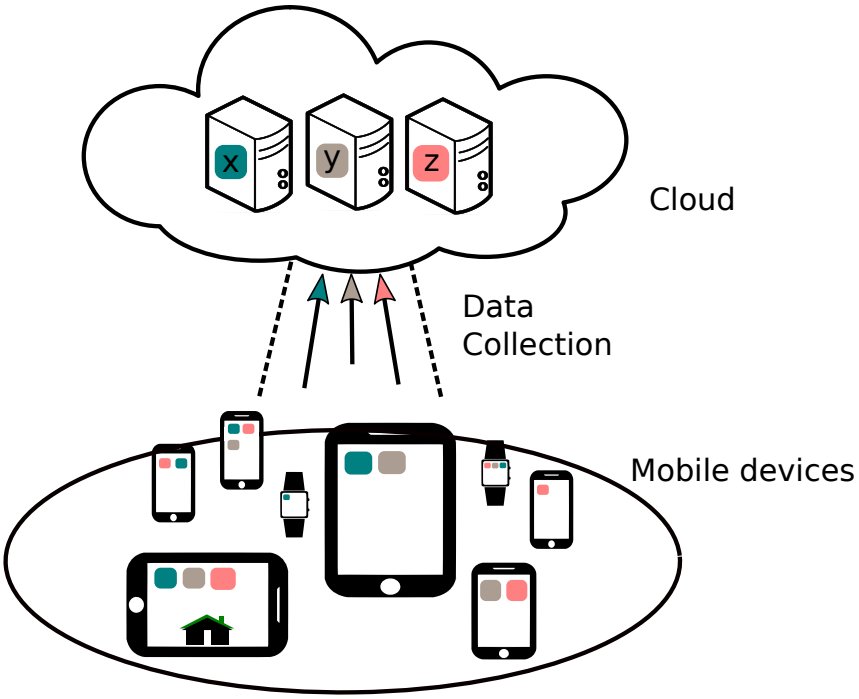


Figure 3: *Crowdsensing:* Applications collect information shared by mobile users regarding to their device instrumentation for different purposes.

Another application that uses the characterisation of network connectivity parameters based on crowdsensed measurements from users at different locations is Netradar. According to Sonntag et al., [SMS13] almost all of the existing network measurement tools and web-based services focus on network bandwidth, sometimes latency, and only report the results to the end user. Depart from other applications, Netradar measures and shares the quality of mobile internet connection. In addition to network bandwidth and latency, it measures network connectivity parameters such as location, download and upload speeds, and signal strength. By seeking, combining and analysing all possible data about the connectivity, it is designed to help end users to gather and share data about mobile networks and devices, and researchers to understand better the mobile network connectivity. The former studies do no explain the reasons behind its variation, although different factors affect and limit the mobile downstream bandwidth. Sonntag et al. have identified five major factors: radio technology, coverage, congestion caused by other users, the mobile phone itself and handovers.

In this work, we raise the stakes by showing that the combination of passive measurements from two distinct crowdsensed datasets can be used to expand the number of contexts, understand the behaviour of mobile users and improve the modelling of

the influence of different factors on app retention.

2.6 Combined Data Analysis

Different performance factors are typically closely linked with each other. For example, latency results in energy drain, and networking technology has a major impact on energy use [HQG+12b, PLNT15b, SMS13] As different performance factors are typically correlated and have non-trivial interdependencies that need to be taking into account while analysing them [PLNT15b], the analysis of events considering independent factors could hide the diversity of execution contexts or smooth interdependencies.

Hot deck multiple imputation is a widely used method for aligning two datasets that overlap only partially [AL10, Rei93, Mye11]. The idea in hot deck imputation is to fill in missing values (in the combined set) with items that are similar (in the individual datasets, figure 4).

According to Andrigde and Little [AL10], hot deck imputation is a conventional technique for handling item non-response, when a sampled unit does not respond to the entire survey or a particular question. After the incomplete dataset is filled with the missing values, it can be analysed with traditional analysis methods. Respondent, i.e. the donor, is similar to the non-respondent, i.e. the recipient, concerning characteristics observed by both cases. Depending on the used technique, the donor is selected randomly from a set of potential donors (i.e. random hot deck methods) or a single donor is identified and values are imputed from that case (i.e. deterministic hot deck methods). Such as in all imputation methods, the result of hot deck imputation is a rectangular data set. Compared to other imputation methods, it is less sensitive to model misspecification than other method based on a parametric model. It makes implicit assumptions through the choice of metric to match donors to recipients and the variables included in this metric, and does not rely on model fitting for the variable to be imputed. This method has various strengths, such as it imputes real values, avoids strong parametric assumptions, can incorporate covariate information, and provide good interferences for linear and non-linear statistics. One of its weaknesses is that it requires good matches of donors to recipients that reflect available covariate information.

In this work, we use combined data analysis to study the influence of performance related variables in long-term app usage. We combined two individual datasets, the

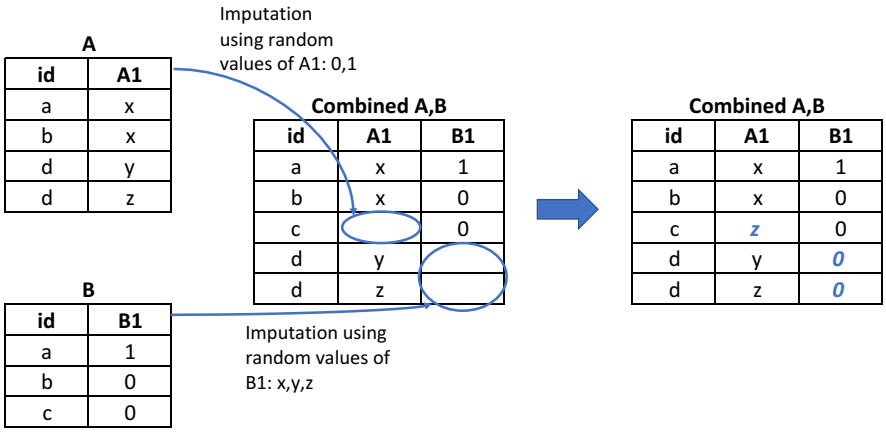


Figure 4: Hot deck with random imputation.

first including measurements of latency and the second of energy consumption.

2.7 Summary

We conducted a literature review about the state-of-the-art about mobile applications usage and mobile crowdsensing. First, we showed that actual studies, oriented to describe users' behaviour during apps usage, rely on that their behaviour is affected by the quality of experience. Next, we observed some studies focused on understanding and modelling performance metrics of the users' satisfaction. Then, we reviewed literature that analyses latency and energy as relevant factors of performance. Finally, we introduced the actual research on mobile crowdsensing and combined data analysis by showing their importance for improving factor's characterisation through collecting a higher amount of contexts of users' behaviour. In general, while state-of-the-art shows a big-picture about users' response to performance changes, it lacks quantifying and modelling the effects of latency and energy performance on apps retention, as well as in studying their combined effect. Table 1 shows in detail how does this thesis complement the state-of-the-art.

In the next chapter we describe the datasets and how we prepare them for the analysis.

Topic	Study	Study contribution	Limitations	Thesis Contribution
Mobile App Quality	[RCT+07]	Impact of environmental factors on user experience.	Studies based on users' perception.	Analysis of the causes for changes in users' behaviour based on performance factor.
	[FGK+14]	Understanding mobile phone usage practices.	Factors affecting usage are ignored.	
	[IWF+12]	Present bugs, performance and user needs as key factors that influence quality perceptions.	Lack on explain the causes for changes in behaviour.	
	[CLH+14b]	Show the significance of app ratings for download decision.		
Mobile App Performance	[RPA+12]	Present how the effect of network communication and processing cost cause app performance issues.	Lack on measure where do performance start to affect perception.	Quantification and modelling of the effect of performance in apps usage, but also identification of the location where performance starts to affect perception.
	[GK15] [AHP+14a]	Show the relation between network performance and user satisfaction.		
	[AHP+14a]	Estimate QoE based on passive network measurements.		
	[SN01] [FSK+17]	Model and improve performance based on devices' features or remote infrastructure.		
Effect of Energy	[PLNT15a] [PLNT16] [PJHM12]	Present the causes of unnecessarily high energy consumption.	Lack on measure the effect and where do performance issues start to affect apps' usage.	Analysis, quantification and modelling of the effect of energy and latency in apps usage, both variables individually and together.
	[PLNT16] [RQZ07] [ALvK+14] [OIS+13]	Present users' behaviour to energy-hungry apps, and strategies to prevent swift energy drain.		
Effect of Network Quality	[NCKB+14]	Measure network performance based on throughput	Usual users complaints point to latency as the most appropriate factor for studying app performance. Studies are focused on desktop contexts. Lack on measure the effect and where do performance issues start to affect usage.	
	[HQG+12a]	and signal strength.		
	[CHL06] [TWLC11] [War09] [MCC11]	Show the influence of latency in users' behaviour.		
Mobile Crowdsensing and User Behaviour	[GYL11]	Explain crowdsensing approach and applications.	Studies are focused on specific topics. Analysed data comes only from one app.	Study of users' behaviour based on combined data analysis using large-scale crowdsensed datasets.
	[TLN+14]	Characterization of mobile malware.		
	[SRRM+17]	Indentification of users personality traits.		
	[PLNT15b]	Construction of energy models.		
	[SMS13]	Characterisation of network connectivity parameters.		
	[FHN+17] [PLNT15b]	Contexts' influence depending on mobile resources and systems.		
Combined Data Analysis	[HQG+12b] [PLNT15b] [SMS13]	Introduce the importance of doing combined data analysis.	Studies have not been tested using crowdsensed data.	
	[Rei93] [AL10] [Mye11]	Introduce hot deck imputation method and its applications.		

Table 1: Summary of current state of the art solutions that address the influence of performance issues in app retention.

3 Datasets, Preparation and Metrics

In this chapter, we describe the two crowdsensed datasets used in this thesis and their metrics. Besides, we introduce the method to calculate retention rate. In the rest of this chapter, we detail how we combine the data sources and validate the combined dataset.

3.1 Datasets Overview

In this thesis, we analyse the impact of performance on application retention using two performance variables: latency and energy consumption. However, understanding performance based on these two factors is challenging due to performance is sensitive to a user's context, we require a significant amount of data to characterise execution contexts representative of typical everyday situations. To guarantee the capture of higher number of contexts, we consider to use two large-scale datasets collected through crowdsensing: NetRadar [SMS13] and Carat [OIS+13][6]. We use NetRadar to obtain measurements of the network performance in distinct areas. Carat, on the other hand, gives us information about apps' performance regarding their usage and energy consumption.

The analysis of retention considering latency and energy consumption by separate could hide the diversity of execution contexts or smooth interdependencies. In that sense, it is necessary to consider analysis to be performed using combined information. Combination enriches the quality of the contexts improving the understanding about network quality and battery consumption influence in long-term app usage. We combine NetRadar and Carat datasets employing coarse-grained location, timezone and cellular network information. Considering only the combined measurements also ensures the usage contexts captured are similar across our analysis. After combining the datasets, we focus our study on country granularity level considering only those countries with the highest amount of data for our analysis. In the intersection of the two datasets, 91% of the data is from Finland and the USA, and 93% of the USA data is from Eastern USA. This mainly due to demographics of the user populations of the mobile apps which were used to collect measurements. As a result, we focus our analysis in Finland and USA (EST - Eastern Standard Time). Here hereafter we refer sampled data from the USA as EST-USA. The datasets we used for this study are described in Table 2. In the following, we go into detail

[6]urlhttp://carat.cs.helsinki.fi and `http://www.netradar.com/`.

Component	**NetRadar** [SMS13]	**Carat** [OIS+13]	**Combined**
Combined Fields	Timestamp, Time zone, MNC [Mobile Network Code], MCC [Mobile Country Code]	Timestamp, Time zone, MNC, MCC	All of the left
Fields	Avg. 5s latency [milliseconds], GPS location	Running applications list, Battery level[%]	All of the left
Samples	875,907	19,608,938	Latency: 1,000,058 Energy: 2,819,748
Users	-	25,402	1,241
Apps	-	48,770	243
Start Time	01/01/2016	01/07/2016	01/07/2016
End Time	31/12/2016	31/12/2016	31/12/2016

Table 2: Summary statistics of application usage and network connectivity datasets.

about datasets, preprocessing and preparation, previously analysing the effect of performance in apps retention.

3.2 Performance – Related Variables

3.2.1 Network Latency

We include as the first related-variable of this thesis the *latency*. We analyse its role in the performance, and how variations in this factor impact app's attrition.

Data Description For the analysis, we use latency samples from NetRadar[SMS13]. This app periodically collects data of mobile network status from end-user devices and provide detailed information about network conditions. We restrict our analysis on cellular network connectivity as Wi-Fi has higher bandwidth than cellular technologies and as its performance has less variation overall [DHD10, DNSB14]. Another reason to limit on cellular networks is that it guarantees our analysis to capture a broad range of usage contexts and higher spectrum of mobility patterns. The dataset comprises 875,907 samples collected from January to December 2016, and the information includes the following fields:

- Latency in milliseconds [ms],
- Timestamp in Unix epoch time format,
- GPS coordinates as longitude, latitude pairs,

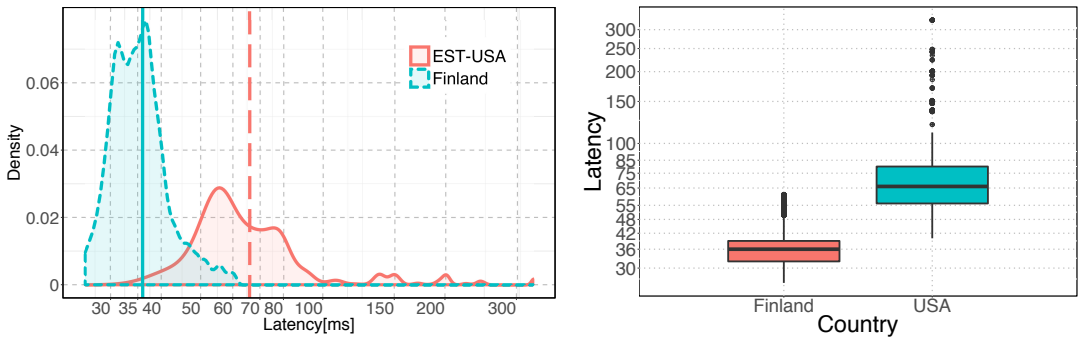


Figure 5: *Comparison of latency distributions between Finland and EST-USA:* There is a clear difference between the expected latency experienced by users in EST-USA and Finland. However, latency corresponds to 2G, 3G and LTE networks, being the last the common zone of the highest density of both countries.

- Mobile network identifiers described by mobile country code (MCC), mobile network code (MNC), cell identification number (Cellid) and local area code (LAC).

Data Cleaning and Preprocessing We perform the dataset cleaning individually for Finland and EST-USA samples in two steps. In the first cleaning step, we validate NetRadar cells information using OpenCellid [7] database. We combine NetRadar and OpenCellid dataset using as matching field the Cell Global Identity (CGI), which is given by the combination of the MCC, MNC, LAC and cell id. We keep only samples of Netradar with valid CGI id. The use of Cell Global Identity effectively allows identifying a particular cell tower[Val98]. As second cleaning-step, we filter NetRadar samples from not invalid and atypical values. We consider only the samples with values of latency greater than zero and within the percentiles 2.5 to 97.5. This method guarantees to separate outliers (less common/atypical values) from the dataset.

The estimated latency $Lat_{u,a}$ for a user u using an application a is obtained by calculating the median of latencies lat_t of u's samples at each time t, t = 1,2, ...n, when application a was used ($a \in s_t$):

$$Lat_{u,a} = median(lat_{t=1}, lat_{t=2}, ..., lat_{t=n}), \quad (1)$$

[7]The OpenCellid project is the largest collaborative open-data repository of cell towers location. The project aims to offer GSM localisation data collected from diverse sources including mobile apps and network providers. More information can be found in `https://opencellid.org`.

where time $t = n$ corresponds to the last sample when the app was active. We use the median since it is more robust against extreme values than average.

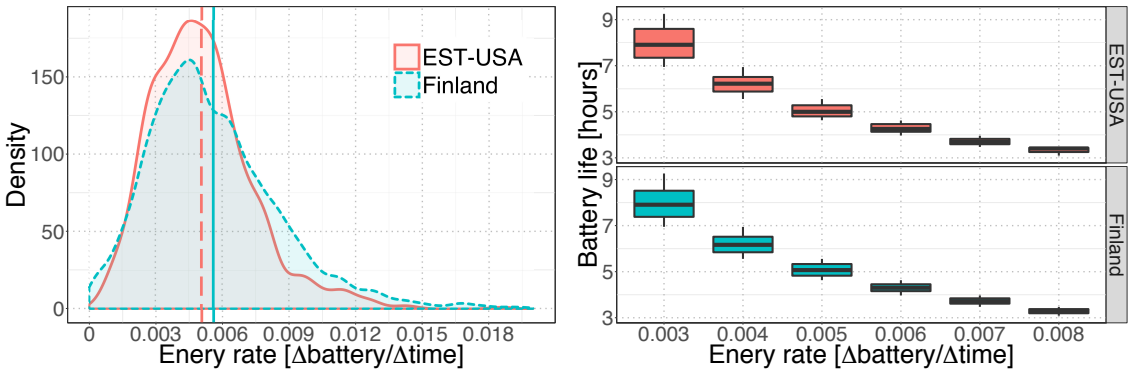
Figure 5 compares latency distributions of the dataset for Finland and EST-USA. We observe the two locations having distinct latency distributions with Finland showing lower expected latency than EST-USA (median 36ms vs 66ms). The overall variation within Finland is several orders of magnitude smaller with the majority of values being within 45ms. For EST-USA, latency is mostly in the range between 40 to 100ms, with smaller peaks at 140ms and even at 200ms. While the values of the distributions are different, their shape is similar with both being long-tailed and skewed towards lower values. We get the same overview looking at the kurtosis and skewness coefficients for both countries having 12.4 and 2.2 for Finland, 2.7 and 0.6 for the US, respectively.

To put the values into context, most latencies for Finland are below 70ms which is within LTE network range [FHN+18]. For EST-USA, the majority of values is within LTE range, but we can also observe values over 100ms which are likely to correspond to 3G connectivity - or even 2G at the end of the tail. This would suggest differences in network infrastructure, or mobile subscriptions, within the two locations. These differences in latency distributions and characteristics of the underlying network infrastructure motivate us to consider the two locations separately in our analysis.

3.2.2 Energy Consumption

We consider *energy* consumption as the second performance factor to be analysed in this thesis. Similarly, as in for latency, we analyse its effects on users behaviour related to long-term app usage.

Data Description As a source of energy consumption measurements, we consider Carat [OIS+13], a popular mobile energy-awareness application. Besides energy consumption, Carat collects application usage data and hence it is used also as a source of retention data. Carat collects samples whenever battery level changes. The dataset includes 19,608,938 samples collected from 1,241 users between July and December 2016. In Carat, each sample contains measurements of the current battery level, timestamp of the sample, list of running applications, and additional attributes such as temperature, device uptime, battery state, and mobile network information. Note that Carat does not collect GPS measurements, but only contains country information as given by the mobile network such as mobile country code



(a) (b)

Figure 6: *Energy rate distribution and battery life influence:*(a) Energy rate is similarly distributed for both locations, the consumption rate is normally less than 0.008 and the mean is located around 0.005. (b) The energy consumption rate shown as active battery life in both countries. In general, these apps would drain the battery in less than 9 hours with constant use.

(MCC) and mobile network code (MNC).

Data Cleaning and Preprocessing Energy consumption is not trivial of being determined. Some studies have developed different mechanisms for estimating this parameter, i.e. by measuring the use of network, CPU and display energy during the use of the applications [MKC12]. We use the same approach as for measuring the energy consumed by an app. We use this method due to it has been previously validated for working with Carat dataset.

In this context, as a unit of analysis, we consider *energy rates* which correspond to the relative change in battery in a given time interval. Formally, let Δb denote change in battery between successive samples, and let Δt denote the difference in timestamps. In that sense, we define energy rate as the mean change in battery over the interval of time, i.e.

$$e = \frac{\Delta b}{\Delta t}. \tag{2}$$

The energy rates calculated using Carat dataset provide information about the battery level over time including battery charging and discharging periods. Then, we validate e only considering samples in the following conditions:

- Battery level has increased, the rate is positive (negative rate indicates charging).
- Battery state is not charging (AC or USB).
- Device uptime has increased from the last sample (the device has not been turned off in between).

We also limit our analysis to samples collected from Android devices as information about running applications cannot be accessed on other platforms and because the sampling granularity on Android devices is better than on iOS devices. Similarly, we consider the interval between two samples from the same device in the Carat dataset for energy consumption, as in the original paper [OIS+13].

To study energy consumption of an application a, we take all the rates e_a containing a in the list of currently running apps and calculate the mean rate $\overline{e_a}$ for the application. Since energy consumption fluctuates due to environmental variables (e.g., Wi-Fi and strength of cellular signal) and specific system settings (e.g., screen brightness or use of location tracking), and other currently running applications, we represent energy consumption using the 95% standard error of the mean (SEM) confidence interval [OIS+13] given by

$$e_{a,range} = \overline{e_a} \pm h \cdot \frac{\sigma_a}{\sqrt{n_a}}, \tag{3}$$

where $h = 1.96$ is the confidence interval coefficient, σ_a is the standard deviation, and n_a is the number of samples containing a.

Figure 6a compares the distribution of energy consumption between Finland and EST-USA. Compared to latency, the distributions are closer to each other (mean rate 0.0056 for Finland cf. 0.0051 for EST-USA). Also, the variance of the distributions is similar, with 44% and 45% of applications exceeding the average in Finland and EST-USA, respectively. Energy consumption peaks for both countries are near e = 0.005 %/s. The highest concentration is located before the mean, after which values decreases, indicating that most energy consumption is just below the mean, and apps over the mean in energy consumption are increasingly rare as we progress right on the figure.

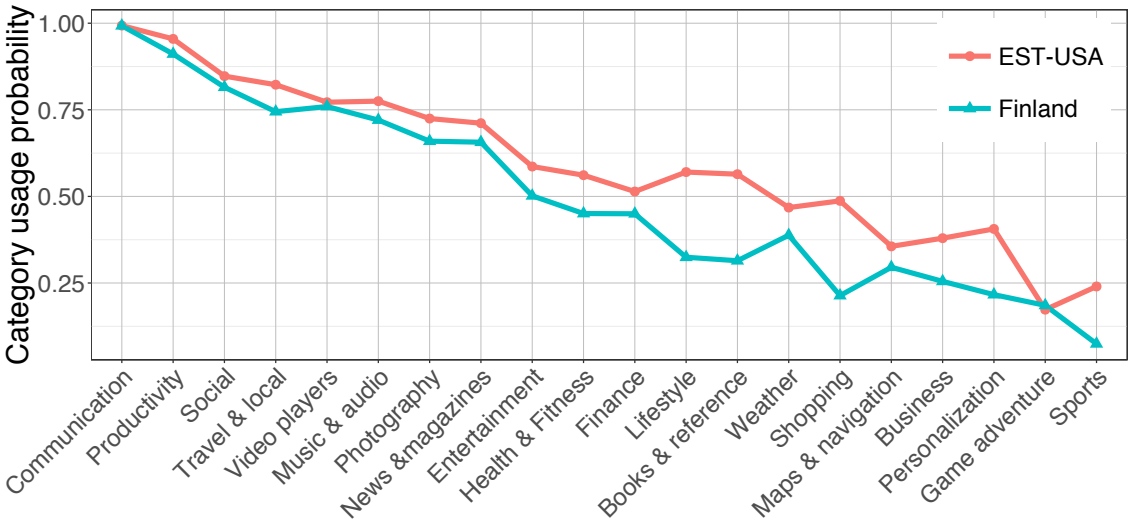


Figure 7: Application category use profiles are different in USA and Finland.

To put these values into context, figure 6b compares distributions of expected battery life in terms of expected battery life, which corresponds to the time an application can run continuously on the device before draining the battery. The means of the two locations would correspond to around 5 hours of battery life, while most apps are in the range of 5-9 hours. The histograms show the range of active battery life corresponding to a given e value. As expected, energy rate and battery life keep a negative correlation. In the worst case (around 0.008) the expected battery time is around 3 hours, while for applications with the most common r (around 0.005) this value goes closer to 6 hours. Note that Carat flags applications with very heavy energy drain as hogs and recommends the user to remove them [OIS+13]. For this reason applications with very low expected battery life are rare in the dataset.

In the analysis of the diverse contexts related to this variable, we have to identify an essential factor, while the two locations have similar energy consumption distributions, they differ in terms of application usage patterns. This is illustrated in Figure 7, which shows the usage frequency of each category for both locations. Overall, EST-USA has higher overall application usage, and the two locations differ in terms of the relative importance of application categories. In our analysis, we separately consider the moderating effect of application category on the relationship between performance and retention.

3.3 Retention Rate

The main focus of our work is on analysing and quantifying how performance-related factors affect long-term user behaviour. As a measure of user behaviour, we consider n day retention rate, which is the fraction of users continuing to use an app n days since first use. Retention is widely used to measure the success of apps as higher retention corresponds to higher adoption and level of engagement [SLP+18] and vice versa. As a source of retention information, we use the list of running applications collected by Carat (see the previous section).

3.3.1 Data Cleaning and Preprocessing

The list of applications captured by Carat contains all the apps running on the device, including those related to the operating system and pre-installed by the manufacturer. To ensure our analysis focuses on actual usage, we filter the applications using two criteria. First, we extract the category of each app from Google Play Store 2016 (the dataset contains 54,758 apps and 50 categories) and remove all apps for which no category information is found (either because they are not available on Google Play, or have no category set) these mainly correspond to operating system related applications. We also use this list for classifying the apps to categories. Second, we remove applications matching specific filtering rules and that have not been seen on the foreground on the device. These mainly correspond to pre-loaded applications such as Samsung Security Policy (com.policydm), Google Exchange Services (com.google.android.exchange) and Messages Android (com.android.mms).

3.3.2 Estimation of Retention Rate

To estimate retention, let d_a^u denote the number of days between the first and last use of an application a by user u in the Carat data. Since an app could have been used before or after the retention interval, we apply a rigorous criteria [SLP+18] to ensure retention is not influenced by the data collection period considering only the cases where we have measurements for 7 days before first use and 7 days after last use. In that sense, we ensure that the time the app was seen is indeed the entirety of the user's usage of the app, and not, for example, either the final, or first part of such data for that user. For instance, a retention interval that estimates app usage between July 15th - December 16th, takes into consideration an extra gap from July 8th - December 24th.

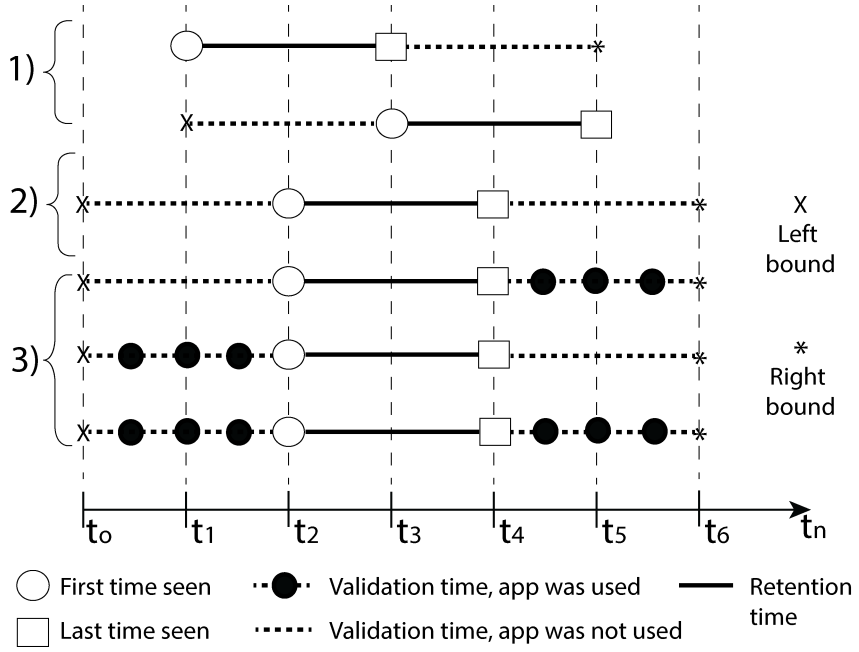


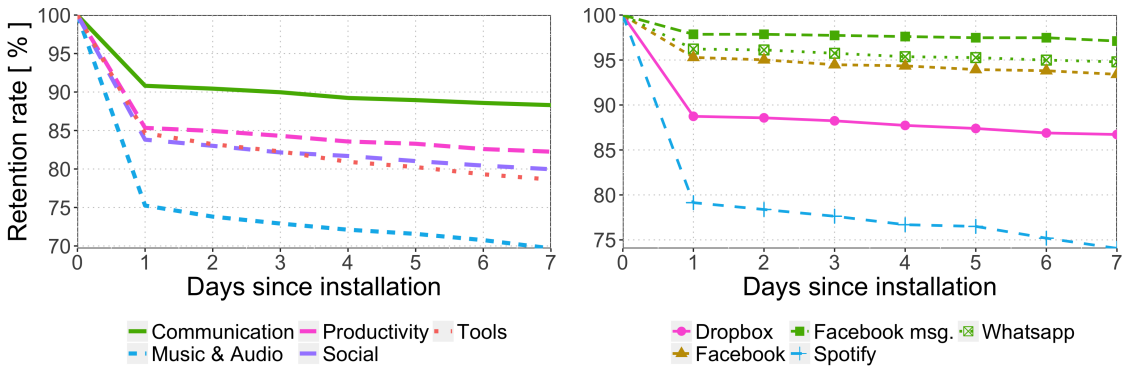
Figure 8: *Method to validate samples:* only type 2) samples are considered as valid for analysis.

Figure 8 depicts the validation method, we consider valid only the apps samples *type* 2), given that they include full left and right validation periods without usage traces. Similarly, to ensure the estimated retention patterns are sufficiently robust, we only consider users that have at least 14 days of data (Carat has been installed at least for 14 days), and apps that have at least 10 users that have used them for more than a day, in order to avoid abnormal retention fluctuations caused by a single user.

The n day retention rate of a, denoted r_n^a, is then given by the fraction of users whose retention time d_a^u is higher than n, i.e.,

$$r_n^a = \frac{\#U_{a,n}}{\#U_a} \tag{4}$$

where U_a is the set of users that use a, and $U_{a,n} \subseteq U_a$ is the subset of users for whom $d_a^u \geq n$. Note that retention rate is cumulative so a user with $d_a = 3$ days also contributes to day 1 and 2 retention rate. In that sense, for an app a, a retention rate equals to 0.82 on day 3, r_3^a, represents that on day three only the 82% of users continue using the application, while the 18% of them have stopped using it.



(a) (b)

Figure 9: *Retention rate of top 5 (a) categories and (b) applications*: The highest drop in retention occurs during the first day.

To further illustrate retention rate, we calculate the retention rate of the top 5 categories and mobile applications during the first seven days since installation. Figure 9 shows the results. We observe the correlation between retention rate and days since users installed apps. Long-term usage decreases linearly over, but slope changes independently for each category and app. However, in all the cases, the highest decrease in retention occurs on day one indicating that the most significant abandonment occurs on the first day of usage, up to 25% for categories and 20% for applications. Afterwards, retention decreases slowly, 4% to 6% per day for categories, and 2% to 6% for applications.

Figure 9 also illustrates that on day 7, applications like Facebook maintains a retention rate over 95%, while Spotify shows a retention of 75%. A similar pattern occurs for their categories by day 7 (see Figure9a), retention for Communication category is around 87%, while retention for Music and Audio drops below 70%. This suggests the influence that popular applications have on their categories [SLP+18].

To further motivate the need to consider the two locations, Finland and EST-USA, individually in the analysis, Table 3 and 4 compares the average retention over the first 7 days across both locations for the 5 most popular categories and applications, respectively. From table 3 we observe the two locations to have distinct retention patterns with Finland having higher mean retention than EST-USA, but also much higher variation. To highlight differences in usage across the locations, as part of the tables, we have also included the number of users, number of samples and summary statistics of the performance variables for the apps. While the number of users tends to be lower in EST-USA than in Finland, we observe that the number of samples to

		Retention(r)			Energy			Latency	
Category	Location	Mean	St.dev.	Users	Mean	St.dev.	Samples	Median	Samples
Communication (C)	Finland	94.1	2.3	993	0.0049	0.0014	3,184,554	34	1,197,260
	EST-USA	78.1	7.5	83	0.0052	0.0010	474,340	79	4,044
Productivity (P)	Finland	90.1	3.8	719	0.0049	0.0014	1,726,808	37	622,071
	EST-USA	78.1	7.5	63	0.0048	0.0018	328,787	79	1,654
Tools (T)	Finland	87.5	4.8	697	0.0059	9.0E-04	1,614,639	36	606,531
	EST-USA	76.8	8.1	73	0.0050	0.0015	376,269	89	2,585
Social (S)	Finland	88.7	4.3	684	0.0056	0.0016	1,367,938	34	532,110
	EST-USA	76.4	8	68	0.0051	0.0011	294,461	68	2,779
Music (M)	Finland	79.7	7.2	552	0.0051	0.0020	292,685	35	118,019
	EST-USA	64.4	12.3	42	0.0037	0.0011	118,454	101	885

Table 3: Variation of retention rate, energy and latency, and influence of app performance in retention for top 5 categories..

		Retention(r)			Energy			Latency	
App	Location	Mean	St.dev.	Users	Mean	St.dev.	Samples	Median	Samples
Whatsapp (C)	Finland	98.3	0.8	815	0.0047	0.0014	980,097	35	405,203
	EST-USA	66.5	13.8	25	0.0050	3.0E-04	45,281	79	588
Facebook (C)	Finland	98	0.9	617	0.0039	0.0018	1,109,947	32	389,834
messenger	EST-USA	95	2.1	55	0.0052	0.0016	239,841	108	2,178
Facebook (S)	Finland	96.9	1.5	565	0.0036	0.0021	955,718	37	345,073
app.	EST-USA	86.7	5.4	59	0.0060	1.0E-04	225,832	77	1897
Dropbox (P)	Finland	94	2.6	428	0.0056	0.0019	606,958	40	204,430
	EST-USA	78.1	8.8	24	0.0038	1.0E-04	102,060	79	316
Twitter (N)	Finland	92.7	3.2	323	0.0042	0.0017	192,232	37	71,067
	EST-USA	67.8	13.8	26	0.0045	0.0012	38,192	70	314

Table 4: Variation of the retention rate, energy and latency, and influence of app performance in retention for top 5 applications.

characterise each performance variable is enough for both locations.

Retention also describes cross-country popularity of categories and applications exposing the diverse contexts of usage. For instance, in table 4 we observe Facebook Messenger to be well-positioned in Finland and EST-USA; however, Whatsapp presents higher acceptance in Finland than EST-USA. In terms of retention behaviour, Dropbox has lower latency, but higher energy drain in Finland than EST-USA, which suggests that this app is used under differing situations. The differences in retention and usage patterns further serve to illustrate why do measurements have to be analysed separately.

3.4 Combining Datasets

Carat and NetRadar datasets capture individual performance factors and thus cannot be used to quantify their combined effect. To analyse and quantify the combined effect of performance factors, we therefore need to combine measurements in the two

datasets. We perform the combination using hot deck multiple imputation. The study of the combined dataset allows discovering the effects of energy and latency *together* on retention, after which it is possible to understand the predominance between them and quantify the impact on long-term usage of apps. Combination process, however, is not trivial due it could produce a side-effect of losing original dataset properties. In effect, each feature in the combined dataset needs to keep significant representativeness respecting its individual data source; otherwise, fusion will not reflect valid contexts of use. In the following, we describe in detail the combination process and verification tests we performed to asses dataset validity for this study.

3.4.1 Data Fusion

We used the common features in NetRadar and Carat (see Sections 3.2.1 and 3.2.2) for the datasets combination. The sampling periods of the datasets differ, and hence we first needed to align them temporally. We performed the alignment by creating hourly bins and mapping each sample in NetRadar and Carat to the closest bin. We added time zone in NetRadar using geocoding of the GPS coordinates of the mobile cells. As a result, we fused the datasets using a combination of timestamp, and coarse-grained location information given by Mobile Country Code (MCC), Mobile Network Code (MNC) and reverse geocoding of the GPS (time zone from the cellular coverage) as (longitude, latitude) pairs. This method effectively corresponds to hot deck statistical matching.

After temporal alignment and (MCC, MNC, Time zone) tuples matching, we calculated hourly latency values for a given location as medians of all matching measurements across the datasets. Median is more appropriate to determine the hourly latency rate as it calculates the most typical communication speed experienced by users without being affected by network outliers or the location of the base stations. The measurements in the combined dataset are summarised in Table 2. In total, the combined dataset comprises 243 applications and 1,241 users from July to December 2016. This is translated in terms of samples to $1,000,058$ measurements for analysing latency, and $2,819,748$ measurements for analysing energy. The reason for differing sample counts for energy and latency is that we perform the matching separately for each application and category considered in our analysis. As the energy dataset is originally larger, this results in a higher total sample count of energy.

		Energy		Latency	
Dataset	Location	Mean	stddev	Mean	stddev
Individual	Finland	38.27	12.31	0.0053	0.012
	EST-USA	87.45	61.43	0.0059	0.016
Combined	Finland	36.7	6.6	0.0056 0	.003
	EST-USA	84.4	53.4	0.0051	0.0023

Table 5: Mean and standard deviation (stddev) of latency a energy for individual and combined datasets.

3.4.2 Validity

We demonstrate the validity of the combined dataset by comparing statistical characteristics extracted from the combined data against those extracted from the individual datasets. First, we compare mean latency and energy of the individual datasets to those of the combined set. Table 5 shows mean and standard deviation values for both latency and energyare closely aligned, suggesting that the statistical characteristics of the individual datasets are preserved in the fusion. Second, we compare the sample distributions between the combined and individual datasets using Kolmogorov-Smirnov distribution tests. The Kolmogorov-Smirnov test, KS, is a nonparametric statistical method for comparing two samples in terms of their probability distribution. A p-value < 0.05 indicates that we have to reject the null hypothesis, Ho, which is that both samples come from the same distribution. The results of KS test show that no statistically significant differences were found (latency: Finland $KS = 0.104$; EST-US $KS = 0.096$, $p > 0.05$; energy: Finland $KS = 0.04$; EST-US $KS = 0.05$, $p > 0.05$).

3.4.3 Representativeness

Eastern USA samples represent the 93% of the overall USA samples. Besides validating the quality of fusion, it is necessary to compare whether the energy distribution of Eastern USA is representative of the USA as a whole. We perform this by comparing the energy distributions of all samples from the USA and those matched to Eastern USA based on timezone information. Figure 10 compares the distributions of USA and EST-USA. Both follow similar distribution, which is confirmed with the results of Kolmogorov-Smirnov distribution test: no significant differences were found between samples($KS = 0.06$, $p > 0.05$).

Certainly, the total number of users and applications in the combined dataset is

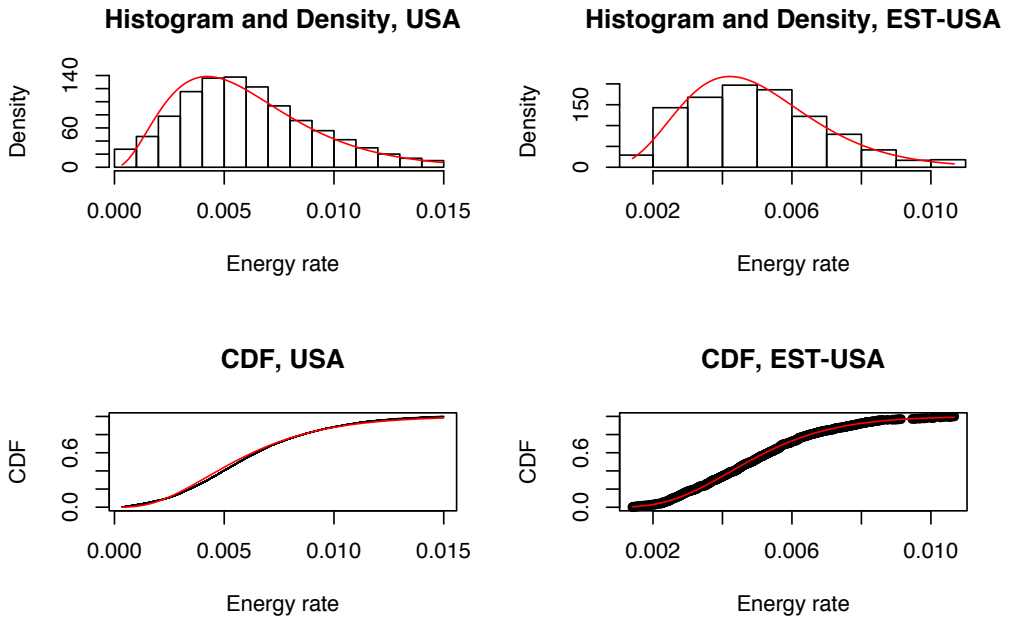


Figure 10: Energy distribution of the USA compared to EST-USA.

smaller when compared to each dataset individually. Data reduction is caused for pruning invalid samples, i.e. containing different cellular information. However, we guarantee combined dataset to be representative on its own of the usage pattern behaviours and locations that are captured in each dataset even at local level. For instance, figure 11 illustrates the distribution of the total number of days a user keeps an app using latency from different cities of Finland. We consider Helsinki, Turku and Tampere, as they have the highest number of samples. We observe that, indeed, samples are similarly distributed, and the combined dataset preserves individual dataset's characteristics.

3.5 Summary

Unlike common methods that study performance related variables independently, in this thesis, we examine combined effect based on the fusion of two individual datasets. While data fusion allows capturing more rich contexts than using datasets individually, the combination process can result in decreasing the number of samples. Despite that, increasing contexts richness produces a more realistic representation of app usage, facilitating the quantification of performance effect on long-term apps usage and model development.

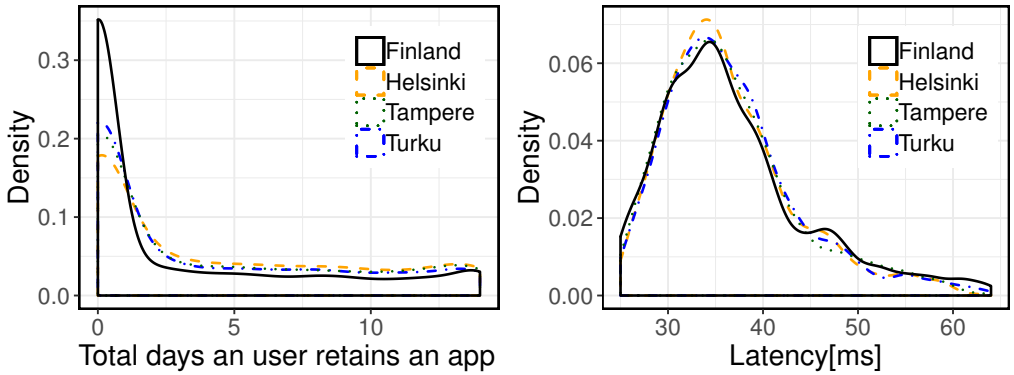


Figure 11: Distribution of the total number of days a user keeps an app using latency from different locations in Finland.

In this chapter, we described the datasets and metrics used in our study. First, we introduced a general description of crowdsensed datasets. Second, we put into context latency and energy explaining the importance of these factors, introducing the methods we used to measure these variables and describing the large-scale datasets characteristics. Third, we introduced the retention rate and the method to calculate this parameter. Finally, we described the data combination process and the methods to validate this dataset respecting to the individual datasets.

In the next chapter, we go into detail about the analysis of performance-related variables on apps retention.

4 The Impact of Performance on Retention

In this chapter, we analyse and quantify the impact of performance-related factors on long-term user behaviour. In the analysis, we focus on latency and energy as two of the main performance variables. We use Netradar and Carat datasets as described in the previous section (see sections 3.2.1, 3.2.2 and 3.4). We consider the impact of latency and energy both individually and together showing that both have a significant influence on retention. To understand how these factors behave in different contexts, we proceed to quantify the point where performance effect becomes significant. Specifically, we identify the point where retention starts to decrease among latency and energy with performance below the threshold. We refer to this as the "critical point". We demonstrate that the critical point is different for energy and latency, and is moderated by user expectations, app functionality, and location. Also, improvement of performance beyond this point could produce a negligible effect on retention. We end this chapter by analysing the combined effect of latency and energy, showing their complex relationship where neither variable alone is capable of explaining retention.

As discussed in the previous section, we focus our analysis in two locations, Finland and Eastern USA, using measurements from the combined latency-energy dataset. We start performing the analysis of the five more popular categories and applications of the datasets. Later we extend the study analysing the effects of performance on highly-rated but less popular apps.

4.1 Application Performance Influences Retention

We begin our analysis by demonstrating and quantifying the overall influence of latency and energy as *individual* performance-related variables on retention, analysing their combined effect in Sec. 4.5. Both latency and energy have been shown to affect user experience [IWF+12], and hence to have an indirect effect on long-term user behaviour. However, whether they have a direct effect on retention has not been previously established nor quantified. We start the analysis considering the five most popular application categories (Communication, Productivity, Tools, Music & Audio, and Social) and applications (Dropbox, Facebook Messenger, Whatsapp, Facebook, Twitter).

		Significance: Energy			Significance: Latency		
Category	Location	Day 1	Day 7	Day 15	Day 1	Day 7	Day 15
Communication (C)	Finland	0.028	0.005	0.006	0.003	0.685	0.921
	EST-USA	0.019	0.026	0.244	0.313	0.001	0.053
Productivity (P)	Finland	0.654	0.600	0.378	0.033	0.001	0.002
	EST-USA	0.263	0.636	0.756	3.0E-04	1.0E-04	0.584
Tools (T)	Finland	3.0E-04	0.001	1.0E-04	2.0E-04	1.0E-05	0.007
	EST-USA	0.059	0.005	5.0E-04	0.499	0.029	0.016
Social (S)	Finland	0.223	0.284	0.027	0.010	0.246	0.0669
	EST-USA	2.0E-04	4.0E-05	0.003	0.0612	0.022	0.099
Music (M)	Finland	0.004	0.050	0.304	0.803	0.288	5.0E-04
	EST-USA	0.027	0.389	0.908	0.244	0.007	0.013

Table 6: Results of statistical analysis of the importance of performance on retention for top 5 categories. Darker colors reflect statistical significance ($p \leq 0.05$).

		Significance: Energy			Significance: Latency		
App	Location	Day 1	Day 7	Day 15	Day 1	Day 7	Day 15
Whatsapp (C)	Finland	0.028	0.040	0.011	0.007	0.066	0.109
	EST-USA	0.315	0.194	0.724	0.460	0.293	0.125
Facebook (C)	Finland	0.027	0.107	0.007	0.203	0.638	0.381
messenger	EST-USA	0.050	0.017	0.011	0.186	0.096	0.158
Facebook (S)	Finland	0.239	0.431	0.022	0.009	0.001	0.002
app.	EST-USA	0.013	0.004	0.009	0.035	0.050	0.010
Dropbox (P)	Finland	0.665	0.478	0.792	0.039	0.004	1.0E-04
	EST-USA	0.377	0.216	0.134	0.105	0.074	0.313
Twitter (N)	Finland	0.089	0.040	0.147	0.231	0.065	0.232
	EST-USA	0.471	0.033	0.077	0.030	0.041	0.198

Table 7: Results of statistical analysis of the importance of performance on retention for top 5 applications. Darker colors reflect statistical significance ($p \leq 0.05$).

4.1.1 Statistical Analysis

For each day of the retention period ($1-15$ days), we compare the mean performance of those who stop using the application and those who retain it. We assess overall effect using the Kruskal-Wallis test, also known as non-parametric ANOVA. This statistical test is used to evaluate significant differences between two or more groups. The evaluation is performed based on the comparison of mean ranks of the groups. Kruskal-Wallis test assumes as the null hypothesis, H_0, which means that groups come from the same populations (groups have the same means). On the other hand, alternative hypothesis, H_1, states that at least one of the means of the groups differs. The level of significance of the difference between groups is measured using p-value coefficient. The lowest p-value, the highest significance. We use a $p - value \leq 0.05$ to accept H_1.

Table 6 and 7 show, respectively, at category and app level the results of the statis-

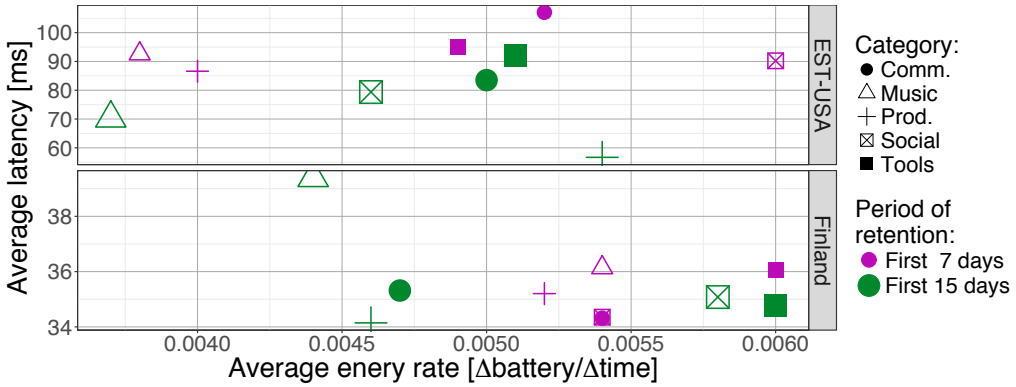
tical test and summary statistics about retention and performance values. From the results, we observe that performance indeed affects retention, but this effect is moderated by application category and popularity of the app. The effects also reflect different interaction patterns across app categories. For example, messaging apps (Facebook messenger and Whatsapp) that require users to wait for response are not influenced by latency, but energy drain has a significant effect on their retention. On the other hand, productivity apps, which often are used for shorter periods of time, have a significant effect on latency but not on energy.

While comparing effects of different days, we observe (from Table 6 and 7) users having different levels of tolerance for poor performance depending on app category. As an example, Music applications show no effect on latency at day 1, and even at day 7 they only show an effect in Eastern USA where latency overall is higher than in Finland. However, at day 15 latency has a significant effect for both locations. Similarly, we observe the effects of energy to be higher for later days in the case of Facebook and Twitter, suggesting users are willing to tolerate more performance issues with them, potentially involving other factors that are more important during the first few days, such as user experience.

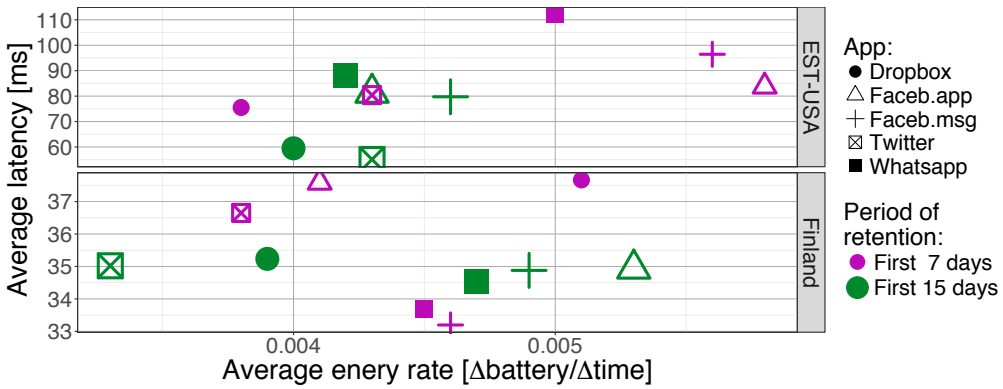
Latency and energy consumption effect change according to the location. In general, almost 60% of the cases are significantly affected by performance at category-level, while 50% at app-level. Latency slightly highlights as the factor with the highest influence. The 56% of the significant cases corresponds to this variable. A similar situation occurs at category-level with 56%. However, at application-level proportions are equal both for latency and energy. At app-level, during the first-day latency represents the majority of the significant cases. Curiously, at day 15 the number of significant cases is higher for energy consumption, pointing again to latency as the factor with the highest influence on performance. This behaviour suggests that latency is usually perceived first by the users. Finally, we observe for specific categories and applications influence is not significant, showing that in some cases users continue using the applications despite variations in performance.

4.1.2 Identifying Factors Relationship

To further provide evidence about the relation of performance factors and retention rate we compare the mean performance of latency and energy during the first 7 and 15 days. The analysis is applied at category and app level. For each day we calculate the expected performance of latency and energy by averaging over all



(a)



(b)

Figure 12: Influence of latency and energy for the first 7 and 15 days of retention at category (a) and application (b) level.

users. Expected performance effectively represents the mean performance of users continuing to use the application.

Figure 12 shows the mean performance of latency and energy for the first 7 and 15 days of retention. Interestingly, for most apps and categories users that abandon apps early tend to experience lower performance than users with long-term usage. For instance, increased latency, rather than those who continue using the application. From the figure, we observe that long-term used applications are located in the middle of the spectrum of energy and latency. We observe a similar pattern at category-level; however, we notice in some cases high dominance of a single performance factor, meaning that most likely users will continue using apps of this category despite changes in performance of one factor. For instance, in EST-USA all the categories show an average latency lower for users with 15 days of retention, despite the variation in energy rate. Three apps present a similar pattern in Finland.

This behaviour proposes again that latency dominates energy in terms of being the first effect perceived by users. Additionally, figure 12b reveals effect predominance in each location at app-level. While for 7-day-retention latency and energy are highly dispersed, for 15-day-retention both factors tend to be together. Hence, for EST-USA, performance locates in less than 90 ms for latency, and between 0.004 and 0.0045 units/sec for energy. On the other hand, for Finland, although energy values are sparse for 15-day-retention, latency values are below 35.5 ms. Certainly, the two performance variables are regulated by context.

4.2 Level of Critical Point in Performance

The previous section demonstrated that both latency and energy have an overall effect on retention. We next analyse the relationship between performance-related variables and retention in more detail. We show that we can identify clear points where decrease in performance results in lower retention. We refer to these points as *critical points* of performance. By identifying the critical points, we can understand the behaviour of users related to variations in latency and energy consumption during the use of applications. The analysis both individually and together can establish more accurately how, through these two variables, performance affects retention. Conversely, we demonstrate that improving performance beyond this point has no influence on retention.

To carry out the analysis, we split users into two performance groups using a threshold v on the two performance factors and compare the retention in the two groups using a test of proportions (i.e., a two-tailed z-test). The test of proportions is a statistical method to obtain the degree of significance of comparing two populations. The difference between groups is significant when $p - value < 0.05$, rejecting the null hypothesis H_0 about evidence is not sufficient to say that proportions of the groups are different.

We iterate over different values of v considering values between the 10^{th} and 90^{th} percentile identifying the range of values where retention is significantly different. We omit the lowest and highest 10 percentiles as these resulted in the smaller group having insufficient data to evaluate statistical significance. We show percentiles instead of exact performance values due to values change across categories and applications depending on which samples include the category or application. In the following, we refer to the two user groups as *high* and *low* depending on which side of v the average performance of users in the corresponding group is.

4.2.1 Differences on Diverse Performance Levels

The behaviour of retention related to performance variations responds to the diversity of users, locations and type of applications. This diversity moderates the effect of latency and energy, making that it appears at different levels, here the importance of identifying the location where *high* and *low* groups start influencing retention.

The results of our analysis are shown as series of heatmaps in figure 14. In the heatmaps, the intensity of colours reflects statistical significance. Darker colours reflect higher significance than lighter ones, as given by the test of proportions. The analysis of the variation of colours reflects the behaviour of groups *high* and *low*, we postpone the exact quantification of performance into the next section. We also notice that for some categories and applications the result is not entirely conclusive, indicating that performance was not a deciding factor, i.e. latency for Tools in EST-USA. These differences are likely produced by short and infrequent interaction patterns with the applications in corresponding categories, resulting in other factors, like usability, being the dominant decision behind retention.

We observe that location plays an important role on how a category characterises performance as users application usage differs between locations. The analysis of the overall sample indicates for Finland that retention for users from group *low* is higher around the 50th percentile, being latency the first factor becoming significant. A similar pattern is observed in EST-USA in terms of latency as the factor which is perceived first but after the 60th percentile. More in detail, latency results at category-level are more consistent, especially for Finland, starting latency effect around 50 and 70 percentiles. EST-USA seems to react to latency changes from early percentiles, but this reaction takes different values within the percentile scale. On the other hand, energy results at category-level are similar to latency-level, especially for Finland.

4.2.2 Location of Different Performance Levels

After understanding the behaviour of groups *high* and *low*, we determine the percentile of latency and energy after which test of proportions indicates a significant difference ($p - value < 0.05$) in retention rate between low and high groups. More in detail, table 8 shows the results. We see clear differences in the points where performance start to influence retention. Mirroring the results of the previous section, we see that categories and applications moderate results.

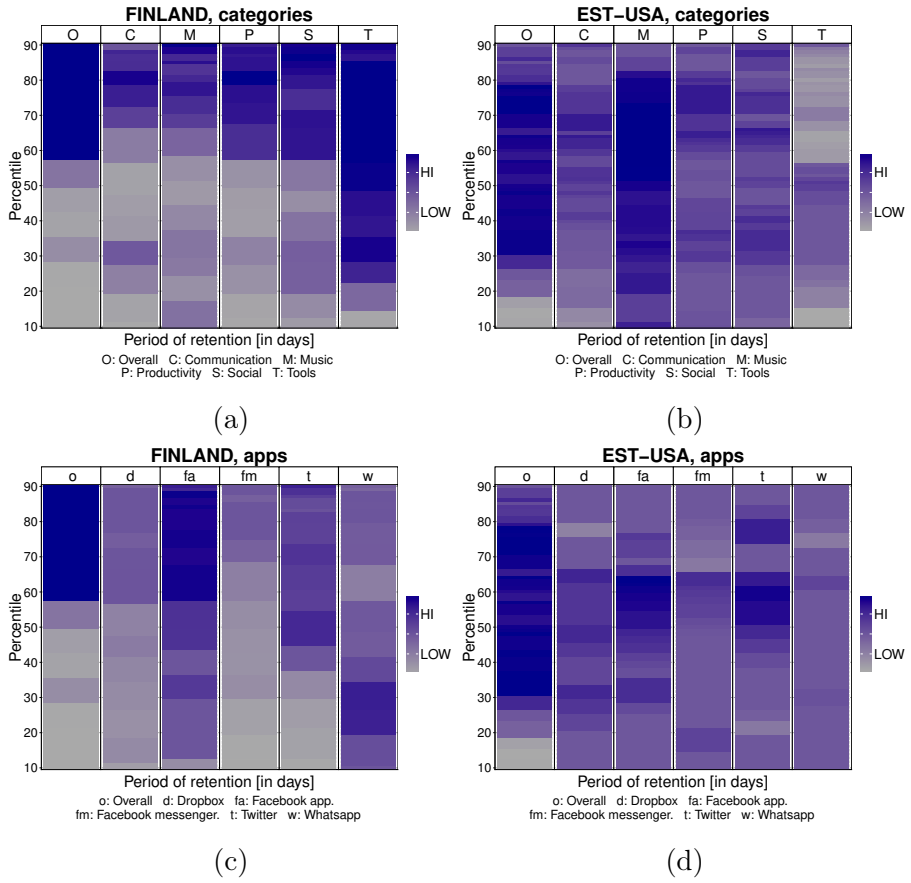


Figure 13: Retention rate difference of *high* and *low* groups proportions for latency. Categories: O: Overall, C: Communications, P: Productivity, T: Tools, S: Social, M: Music; Apps: o: overall, w: Whatsapp, fm: Facebook Msg., fa: Facebook App., d: Dropbox, t: Twitter

Besides categories and applications, we also observe location to moderate the level where performance starts to influence retention heavily. In Finland, significant differences start to occur only at higher percentiles, whereas in Eastern USA significance starts to appear earlier. As an example, latency higher than 60^{th} percentile has a significant effect of retention across all application categories in Finland, whereas in the USA the effect is significant already from 30^{th} percentile onwards. Similarly, energy starts to have an effect at a much earlier percentile in Eastern USA than in Finland. For latency, this difference can be partially explained by differences in network infrastructure, with users in Finland having lower latency and less variability than users in Eastern USA. However, for energy, this is not the case with the distributions being similar across the two locations (see figure 6). Consequently, this suggests that users at different locations either assign different importance to

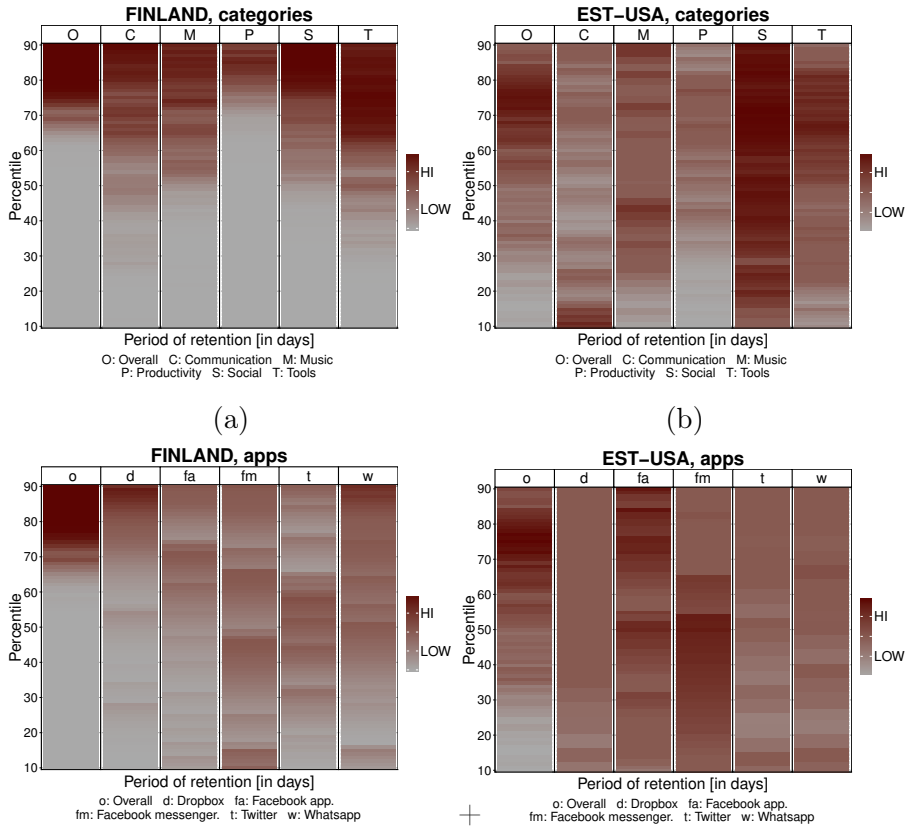


Figure 14: Retention rate difference of *high* and *low* groups proportions for energy. Categories: O: Overall, C: Communications, P: Productivity, T: Tools, S: Social, M: Music; Apps: o: overall, w: Whatsapp, fm: Facebook Msg., fa: Facebook App., d: Dropbox, t: Twitter

energy or have different levels of tolerance.

4.3 Difference in the Effect of Performance

In the previous section, we understood the relationship between retention and related variables: latency and energy. We identified the critical point where decreasing performance affects retention. Besides, we demonstrated that performance does not always affect retention, and location, app and category usually moderate its effects. We extend the results of the previous section quantifying the level at which critical point starts to be perceived by the users. We use this information to establish the area of good enough performance and analyse the influence of the combined effect.

Area, Factor	**O**	**C**	**P**	**T**	**S**	**M**	**w**	**fm**	**fa**	**d**	**t**
Finland, L	63	67	58	26	65	67	10	84	30	57	45
Finland, E	69	63	78	50	67	63	50	43	69	78	51
EST-USA, L	24	32	10	18	25	10	10	10	10	10	31
EST-USA, E	52	10	64	25	11	29	10	10	10	10	10

Table 8: Retention rate difference of *high* and *low* groups proportions for app categories and apps: L: Latency, E: Energy; O: Overall, C: Communications, P: Productivity, T: Tools, S: Social, M: Music, w: Whatsapp, fm: Facebook Msg., fa: Facebook, d: Dropbox, t: Twitter.

	Critical Point, EST-US			Critical Point, Fin		
Category	Energ.	Lat.	ΔP	Energ.	Lat.	ΔP
Communication	10	27	-17	64	67	-3
Productivity		10	10	79	58	21
Tools	24	28	-4	50	23	27
Social	11	16	-5	67	58	93
Music	29	10	19	64	67	-3

Table 9: Critical Point (CP) and ΔP for top 5 categories.

	Critical Point, EST-US			Critical Point, Fin		
App	Energ.	Lat.	ΔP	Energ.	Lat.	ΔP
Whatsapp		10	10	74	11	63
Facebook Msg.	10	10	0	84		84
Facebook App.	19	10	9	87	30	57
Dropbox	46	10	36	80	57	23
Twitter	70	31	39		45	45

Table 10: Critical Point (CP) and ΔP for top 5 applications.

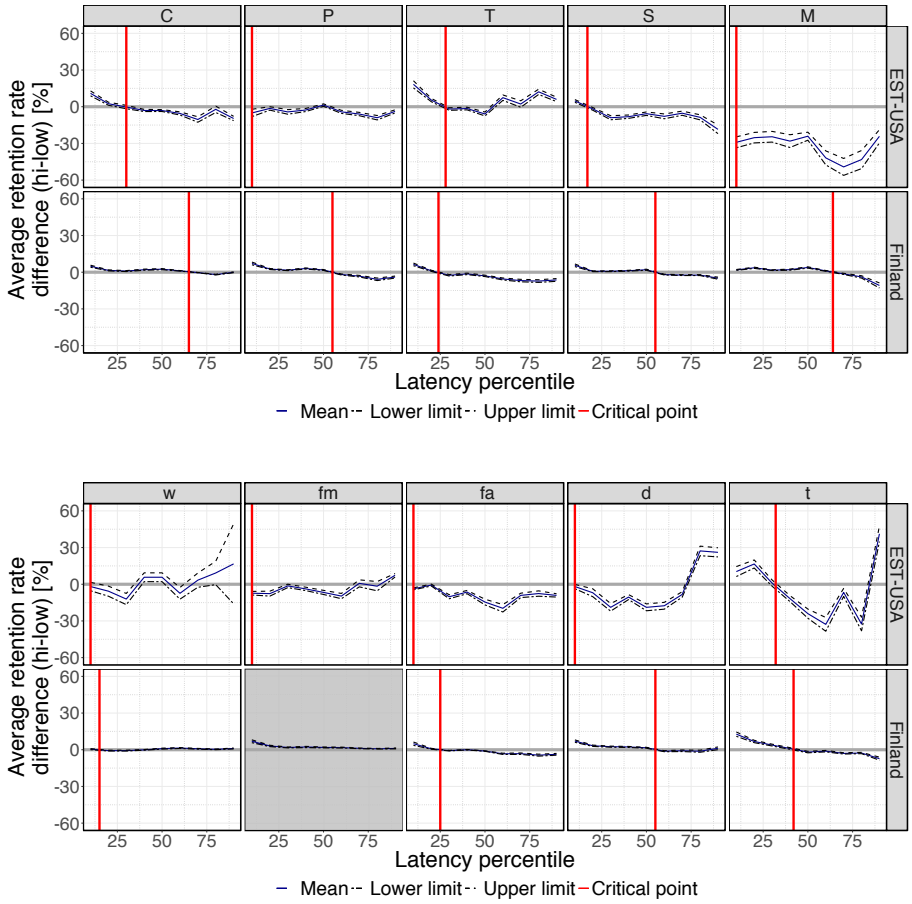


Figure 15: Average retention difference for *high* and *low* groups for Latency Categories: O: Overall, C: Communications, P: Productivity, T: Tools, S: Social, M: Music; Apps: o: overall, w: Whatsapp, fm: Facebook Msg., fa: Facebook App., d: Dropbox, t: Twitter

4.3.1 Quantification of Critical Point

To perform this analysis, we first calculate the difference in retention percentage, $\Delta r_{high,low}$, between high and low groups during the first t days since and app a was used for first time,

$$\Delta r_{high,low} = \overline{r_{high}} - \overline{r_{low}}. \tag{5}$$

As $\Delta r_{h,l}$ approaches for the first time to zero we get close to the critical point; then when the difference turns to negative, it indicates that the *low* group experience a better level of retention. Since $\Delta r_{h,l}$ can variate because of the number of samples and size of the groups we use the same approach of section 3.2.2 to calculate the

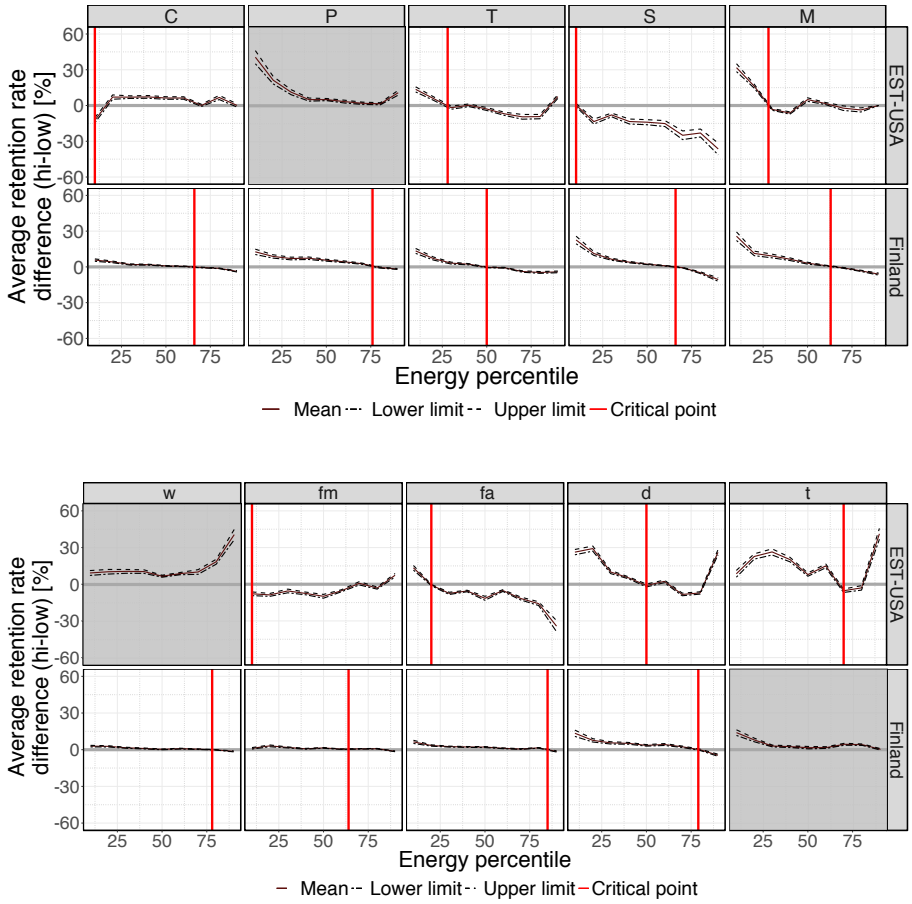


Figure 16: Average retention difference for *high* and *low* groups.for Energy. Categories: O: Overall, C: Communications, P: Productivity, T: Tools, S: Social, M: Music; Apps: o: overall, w: Whatsapp, fm: Facebook Msg., fa: Facebook App., d: Dropbox, t: Twitter

mean confidence interval.

We illustrate the average retention difference using a figure that depicts the percentile where changes become significant for the first time and can start to be quantified (Figure 16). The critical point (depicted as a vertical line). The points in the negative $y - axis$ correspond to cases where retention is higher for users that experienced a better performance. A critical point closer to the $y - axis$ describes that the influence of the performance is perceived faster. On the other hand, a negative value of $\Delta r_{h,l}$ shows the influence of performance to decrease retention. As more negative $\Delta r_{h,l}$, higher the effect on retention. The cases where the performance factor does not seem to influence the retention (average retention rate difference is

always higher than zero) are coloured with a grey background.

In the figure 16, we show the results for the top 5 categories and applications both for latency and energy in Finland and EST-USA. We can observe the strength of significance to change considerably across locations, categories, and applications. We also see that, depending on the category, percentile at which performance differences become significant varies between energy and latency, with one factor usually having a significant effect on retention much earlier than the other. Latency is typically the factor that shows significant effect first. In most of the cases, variations in performance affect retention, only in four cases (one in categories and three in apps) performance does not affect retention, meaning that other factors can be affecting app retention more that battery and energy. Most of these cases correspond to apps of communication category, such as Whatsapp and Facebook messenger, which are commonly used within social circles and whose usage is moderated by level of social activity [SN13]. In the case of categories, Productivity generally includes applications used for working purposes. In both scenarios, even if the performance of apps would be not-optimal, it is unlikely to stop using them. In the first case because replacing them would require the user's entire social circle to migrate to a new service. In the second case, because the use of these apps depends on organisations' policies.

Finally, regarding location, the figure shows a sharper behaviour for Finland, which would be related to the fact that this location has a higher number of samples compared to EST-USA. As seen in the previous section, we confirmed that user' context regulates the critical point. While we quantify the point where performance starts to influence retention, we demonstrate that this point changes according to the category, app and location. However, in general, a similar response of retention behaviour can be observed in performance variations both for latency and energy.

4.3.2 Comparing Critical Point Area

After quantifying critical points, we deepen our study analysing the differences in significance. In the previous section, we identified the factor which is typically perceived first by the users, latency. However, the analysis of how the critical points of latency and energy are related allows determining the relationship between these factors. We compare the difference in the critical points CP between energy and latency, represented as the difference in percentiles ΔP. We start our analysis for the overall samples of each location, which means that we consider the collection of

all applications across all categories for both, Finland and Eastern USA. Next, we apply the same analysis at the category and application level.

The analysis over the overall samples allows having a better understanding of the relationship between latency and energy. Figure 17 illustrates the results of our analysis. Interestingly, when the effect of latency is perceived first, the area of ΔP covers a broader percentile range than in the opposite case. Indeed, when demonstrating that latency is the first factor to affect the retention. On the other hand, energy becomes significant only at many later percentages, normally after the 50th percentile.

The overall analysis for the Finland sample indicates that for latency, retention for users of the group *low* is higher around the 50th percentile. However, for EST-USA, due to the high variability of measurements, the point converges around the 20th percentile. To validate our results further, we remove 10% of outliers data (less frequent) from the left and right tails of the EST-USA distribution and perform the same experiment again. When using the EST-USA dataset with of 10% removed data, we obtain the critical point closer to the 50th percentile.

The relation of the factors at category-level and application-level is explored in Tables 9 and 10. We analyse this relation by calculating a (Kendall) correlation between CPs for both factors. We observe a positive correlation between CPs for both categories (0.41, $p - value = 0.04$) and apps (0.62, p-value = 0.05), the lowest p-value, the highest significance of the correlation. From the table, we see a greater difference between energy and latency when latency is perceived first.

We extended our analysis to compare the size of the latency/energy critical point windows for categories and applications. The average size of the window for categories is 9.11 percentiles while for apps is 37.83, which means that the size of the windows at application-level is three times bigger than category-level. Intuitively, the energy consumption of a particular application may take a long time for the user to discover, while network conditions can change rapidly within seconds and minutes. Therefore poor latency can be discovered much quicker than high energy consumption. Higher latency may also affect the energy consumption of the device, which can result in retention decreasing faster. Because latency is a shorter-term phenomenon than battery life, the decreased retention is easily attributed to latency variations instead of both energy consumption and latency.

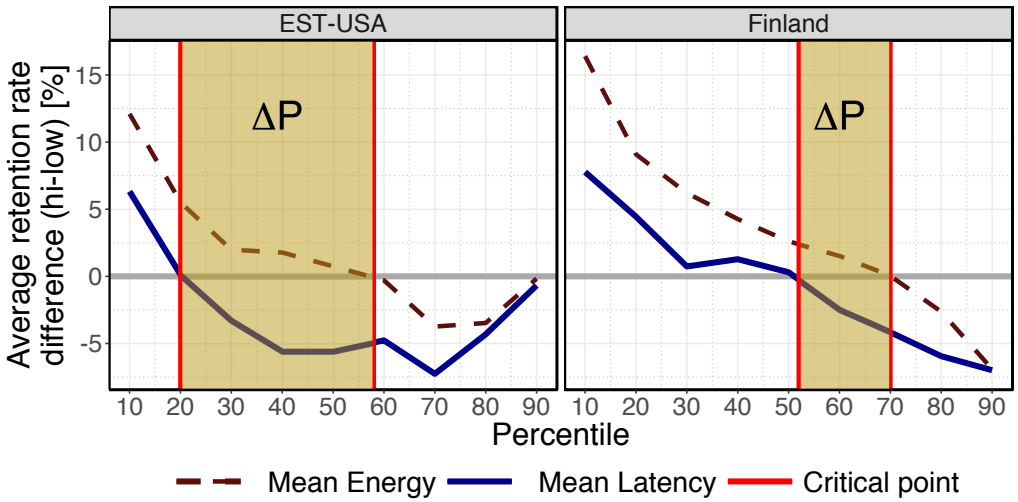


Figure 17: Overall average retention difference for *high* and *low* groups combining latency and energy "critical point" thresholds.

4.4 Effects of Performance on Highly-Rated but Less Popular Apps

Additionally to energy and latency, other factors affect retention of mobile applications, such as app utility, app functionality, and user interface design. To further validate our study, we analyze the influence of performance in 10 applications that are not within the ten most popular, which means to have lower popularity but high user satisfaction. We use the star ranking of Google Play Store for measuring users satisfaction. Furthermore, we verify that low punctuation corresponds to factors that are not related to app performance. As these applications have predominantly received high ratings, users are likely satisfied with functionalities and user interface design of apps. For our selection of apps, we also ensured that retention is not related to popularity, and any negative ratings would not be caused by differences in functionality between commercial and free versions of the app, e.g., due to a high amount of advertisement or limited functionality.

We applied the same method of Section 4.2 to calculate the critical point, CP and average retention rate difference, $DeltaP$. Table 18 shows the results. From the table, we see for all apps, excepting Zedge, that the window between the critical latency point and critical energy point is tinier for apps with a higher retention rate. We also observe that latency has a higher impact than energy for applications with high dependency on displaying on-line content, such as Viaplay. On the other hand, retention of apps used for personalisation, such as Zedge, is more affected by

App	Google ⋆	r	CP,E	CP,L	ΔP
ag	4.5	90.2	10	22	-12
aa	4.5	95.2	33	14	19
f	4.4	84.1	37	90	-53
ac	4.5	90	21	12	9
ae	4.6	88.2	71	31	40
e	4.4	79.2	49	10	39
v	4.0	84.3	58	17	41
h	4.4	71.2	50	28	22
s	4.5	70.6	56	21	35
z	4.6	92	10	73	-63

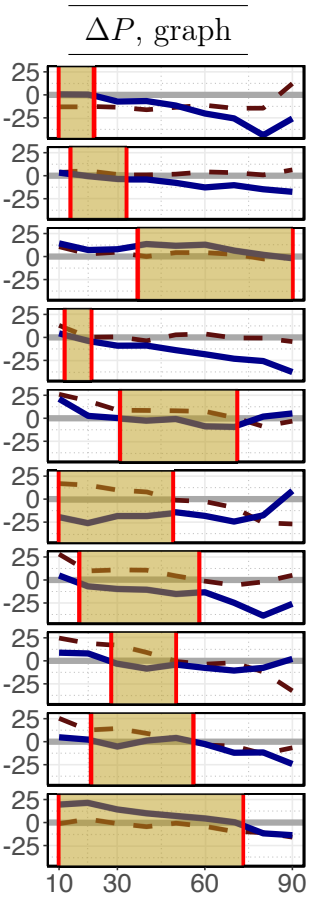


Figure 18: Effect of performance in case study applications. r: retention, CP: Critical Point ,E: Energy, L: Latency; App: ag: AVG antivirus, aa: Avast antivirus, f: Firefox, ac: Avast cleanup, ae: Aliexpress, e: Ebay, v: Viaplay, h: Here WeGo, s: Sports tracker, z: Zedge.

energy variations.

In most of the cases, critical point behaves similarly for apps offering similar functionality. For example, eBay and Aliexpress have a lower critical point in latency than energy. Both applications are focused on online shopping meaning that they are used only intermittently. The importance of latency is understandable for these apps. However, for utility apps, such as AVG and Avast, no clear patterns can be identified. Indeed, Avast is more sensitive to latency whereas AVG is more sensitive to energy. In summary, our results show that even for applications with high user ratings critical points can be identified and quantified. Our analysis suggests that, indeed, performance has a significant influence on retention, but also that importance of latency and energy is dependent on the app.

4.5 Latency and Energy both Affect Retention

In this section we measure the joined impact of performance-related variables. We combine energy consumption and latency by performing a cost-benefit analysis that looks at the combined effect on retention when the importance of individual factors is varied. To perform the analysis, we define a linear cost function that determines the overall effect of the two performance factors as a weighted combination of their individual effect.

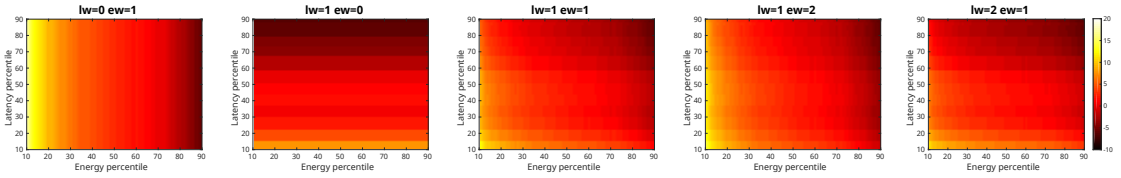
We consider different relative weightings to see how the importance of individual factors affects retention. Formally, let r_l and r_e denote the differences in retention between the *high* and *low* groups (See Section 4.2), and lw and ew the weights of latency and energy, respectively. Given energy e and l, we calculate retention for a given performance level, denoted $R(e, l)$, using

$$R(e, l) = \frac{r_e \cdot ew + r_l \cdot lw}{ew + lw}. \tag{6}$$

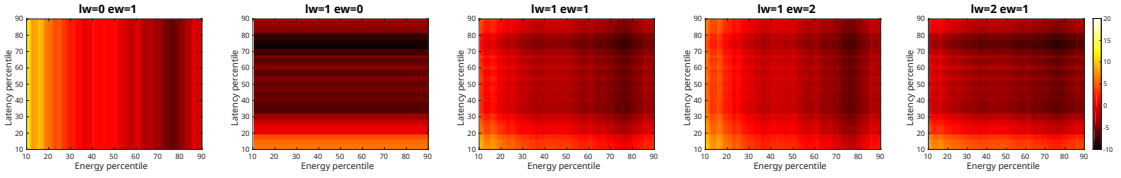
Figure 19 shows the results of our analysis as a series of heatmaps. Each heatmap shows the combined effect of latency (y-axis) and energy (x-axis) on retention for different percentiles (10-90) and different weights lw and ew. In the figure, lighter colours reflect retention improvement, and darker ones aggravated retention. The scale is in percentage units of retention.

The first two heatmaps on both lines show the effects of energy and latency individually by setting the weights of latency and energy, $lw = 0$ and $ew = 0$, respectively. When only one performance factor is considered, the effect of performance on retention is approximately linear as can be seen from the two first heatmaps. When either performance variable has higher importance, the effect quickly becomes non-linear with neither variable dominating the other. When the importance of latency is twice as high as that of energy, the effect on retention is slightly higher than in the opposite case. However, even in this case, there is a lot of variation and a highly complex relationship between the two performance variables.

We also observe the effect of location moderating performance effect, being this smoother for Finland. EST-USA shows a higher weakness to latency and energy variations. For both locations, when analysing factors independently (first two graphs in the right of each country), we confirm that latency effect (second graph) is perceived



(a) Finland



(b) EST-USA

Figure 19: Retention behaviour for performance-related factor given different weights of latency lw and energy ew. The high variation when combining latency and energy effects reflects a high-complex relationship.

first, but also that its effect on having lower retention is stronger than energy (first graph). On the other hand, when studying the combined effect, both countries show worse retention (darker colour) as we increase the percentile towards the right/up.

4.6 Summary

In this chapter, we studied in detail the effect of performance on mobile apps retention using the combined dataset from section 3.4. The analysis was applied both at category and application level for Finland and EST-USA. First, we demonstrated that performance affects long-term apps usage having a point where performance becomes critical and starts affecting retention. Second, we quantified this critical point CP for latency, energy and both combined in each location, app and category. Third, we studied the behaviour of the critical point CP and the relationship between the critical point of latency and energy. We extended this analysis to ten less popular but high-rated applications. Finally, we analysed the relationship between latency and energy performance. In general, we showed that poor performance increases apps abandonment. However, we also showed that having better performance does not imply increasing retention rate due to performance is regulated by location, apps category, usage patterns and user's perception. We demonstrated latency as the factor that is usually perceived first (see section 4.3). Combined analysis showed the high-complex relation between latency and energy(see section 4.5).

In the next chapter, we complement the analysis by developing a model for predicting retention based on performance-related variables both individually and combined.

5 Modeling the Effect of Performance on Retention

In this chapter, we use our results for modelling the retention based on the performance-related variables both individually and together. We apply the model for both locations and at category-level.

At this point, we have quantified the effect of performance on retention. As we mentioned in the introduction, relying on crowdsensing to characterise app performance allows quantifying the point at which users start to be concern about performance, but also to develop a more realistic model of the users' behaviour. In this chapter, we turn our attention to modelling and predicting the degree in which the performance factors affect retention. To accomplish this, we build a generic model that takes as input current performance values and predicts the likely retention difference between those with better or worse performance. We demonstrate that our model has a good performance for predicting the retention of the overall data.

5.1 Model Specification

In the general model, we consider that app retention is influenced by M factors $F_i, |1 <= i <= M$. Each factor F_i has a performance threshold ϵ_i. The changes in ϵ_i affect the overall retention. In that scenario, ϵ_i represents the starting point to quantify how a decrease in the performance of a factor impacts app retention, the critical point CP (See the section: 4.3). By analysing the changes in performance relative to ϵ_i, it is possible to estimate the amount of influence that a performance level has on retention. For estimating the retention, we use a step function as depicted in equation 7, where x is the performance value for a factor, e.g., 30 ms for latency; and $g_i(x)$ is an exponential probability function that approximates the retention rate of the factor given expected performance. We use an exponential function given that it fits better with the data distribution (See: figure 20, an example of testing diverse probability functions in energy data).

$$R_i(x) = \begin{cases} 0, & x \leq \epsilon_i \\ g_i(x), & x > \epsilon_i \end{cases} \tag{7}$$

We then quantify the overall impact on retention by aggregating the influence of each factor. The overall impact of app performance on retention, R, is determined

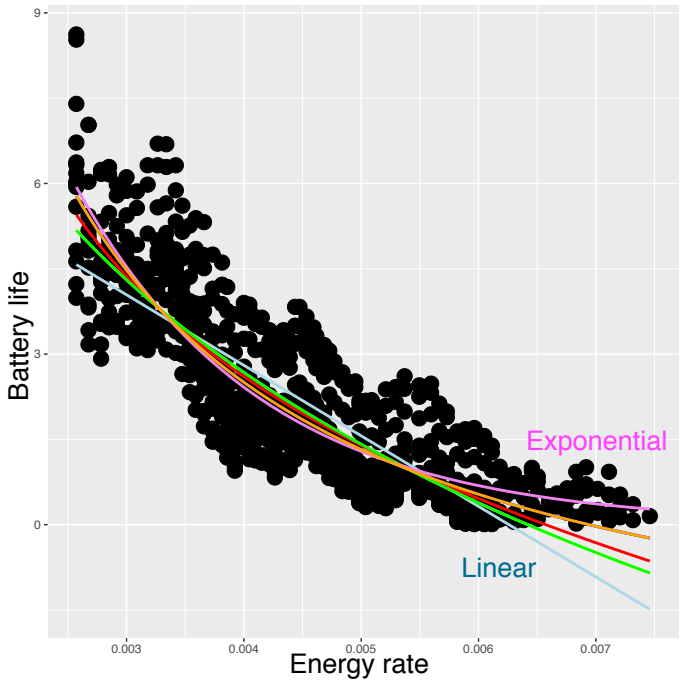


Figure 20: Function validation, an example test for energy. The exponential function curve is the one that better fits the data distribution.

by the factor $F_i \in M$ whose influence on retention is highest, i.e. $R = max(R_i)$. The expected retention rate is then calculated from the uninfluenced retention rate curve $r(x)$ by division:

$$r'(x) = \frac{(x)}{R}. \tag{8}$$

5.2 Experimental Setup

To validate the performance of our model we testing its predictions for data that was not seen before. We apply 80/20 data split approach, which means that data is divided in two subsets. Hence, 80% of the data is used for training and the remaining 20% for testing. We first apply data split for each country, we use this as our baseline. We further validate model performance between Finland and EST-USA subsets. We train our model with data from Finland and predict EST-USA

Retention estimation baseline	O	C	M	P	S	T
Latency						
Finland (80%) → (20%)	0.91	0.54	1.67	0.98	0.83	1.72
EST-USA (80%)→(20%)	1.90	3.33	10.03	2.86	3.15	2.13
Battery						
Finland (80%) → (20%)	0.63	0.30	0.75	0.73	1.12	1.08
EST-USA (80%)→(20%)	1.51	2.49	3.87	1.69	3.70	2.77

Table 11: MAE for categories (model data → predicted), O: Overall, C: Communications, M: Music, P: Productivity, S: Social, T: Tools

retention based on expected performance, and vice versa (Cross-country). We also analyse the effect of mixing data from Finland and EST-USA (Mixed) into a single subset to predict retention using the same 80/20 approach. We then compare the performance of our model when predicting retention based on combined factors.

5.3 Individual Factor Prediction

Table 11 shows the results of baseline for individual factor prediction. The error is measured in percentage points. We observe our model predicting retention values with low error rate especially for Finland as it is the country with the highest number of samples. We then explore Cross-country validation. Table 12 shows the results for both latency and energy. We notice that error rate increases slightly compared to country baseline. The model has an average overall prediction MAE of 2.25, which depicts an overhead of 46% when compared with the baseline. We observe, however, a small error window in retention-based expected latency for (EST-USA → Finland). In this case, error is reduced by 5%. The slightly overhead is due to critical points are regulated by contexts (location, application, categories), as we see in section 4. Consequently, performance starts to be perceived by users in each country differently. For instance, when training our model with latency data from Finland to predict EST-US; while latency in Finland goes around $30 - 45$ ms, the ground truth of EST-USA consists of values around $51 - 147$ ms (see section 3.2.1). As a result, the model reduces its accuracy prediction rate.

On the other hand, when analysing application category separately, we observe a higher error rate particularly when trying to generalise data from EST-USA to Finland. As we demonstrated before, prediction errors are influenced by the number

Retention estimation	O	C	M	P	S	T
(Cross-country) Latency						
Finland → EST-USA	3.58	5.00	31.99	3.76	7.13	2.29
EST-USA → Finland	1.66	6.49	37.06	3.39	8.15	2.01
(Cross-country) Battery						
Finland → EST-USA	2.29	5.84	4.98	2.21	15.49	4.61
EST-USA → Finland	1.53	1.30	5.67	2.18	21.29	4.18
(Mixed Finland + EST-USA)						
Latency (80%)→(20%)	0.82	0.50	1.10	0.85	0.61	1.49
(Mixed Finland + EST-USA)						
Battery (80%)→(20%)	0.28	0.27	0.83	0.56	1.11	1.04
(Mixed Finland + EST-USA)						
Combined Latency+						
Battery (80%)→(20%)	0.29	0.45	0.97	0.35	0.47	0.51

Table 12: MAE for categories (model data → predicted), O: Overall, C: Communications, M: Music, P: Productivity, S: Social, T: Tools

of samples and applications of the training subset (see Tables 3 and 4). Besides, the applications in each category can differ between locations. For instance, Music category data of Finland come from the *Spotify* app, while for EST-USA, most of the data come from the *Pandora Music* app (not available for installation in Finland). This explains why Music category has the highest rate of error. On the other hand, categories with similar usage patterns results are well aligned across the two countries, like in Productivity category.

Finally, we analyse the performance of our model when we consider data of Finland and EST-USA as a whole for predicting retention. We illustrate the results on Table 12. We observe that performance of our model significantly improves when data from both countries is used for training purposes. Indeed, we observe very accurate predictions with marginal errors up to 1.49 for all categories both for latency and energy.

5.4 Combined Factor Prediction

As we performed in section 4.5, we analyse combined effect on retention of latency and energy. Since accuracy of prediction is clearly improved when we use data from

both countries (mixed sample), we apply 80/20 approach using the mixed sample for both energy and latency. As a result, combined effect shows that retention is constrained by the factor that influences more performance. In other words, the factor that is perceived first by users. We illustrate the results of combined prediction in table 12. From the table, we observe an improvement in the overall retention prediction compared to individual factors results (Baseline, Cross-country and Mixed). The maximum marginal error for mixed sample is 0.51 for all the categories, which depicts around a 50% reduction in error when compared with our mixed model that performs the best in individual factor analysis.

5.5 Summary

In this chapter, we developed a predictive model that estimates retention through latency and energy performance, both individually and combined. The model was applied for Finland, Eastern-USA and the overall (Finland, EST-USA) sample. The results showed high accuracy for the overall and combined-latency/energy samples. It also demonstrated the dependency between prediction accuracy and amount of data and category apps.

In the next chapter, we present the discussion of this thesis.

6 Discussion

In this thesis we analysed apps retention based on performance variables. We selected latency and energy due to both are important factors that contribute to apps attrition. While our results show the response of retention to performance changes, as well as quantify and model this effect, some topics derived from our analysis need to be discussed in parallel. For instance, latency and energy are not the unique factors affecting apps retention, combined data analysis enriches contexts variety but highly-reduces the number of samples, and energy efficiency models can impact retention by affecting perception that users have about the applications. This chapter deepens in the discussion of our study.

6.1 Other Factors that Influence App Retention

In this thesis we have discussed two performance-related factors, energy and latency, and we have studied their relationship with app retention. As we have pointed out, besides from these two factors, more elements could impact on retention and user perceptions. These factors, for example include device type, device branch, communication infrastructure, service back-end provisioning, app design, app reliability, in-app ads, location, and so on, to list a few of them. In this study we focused on energy and latency as they are seen the most important to affect app performance and user experience. We are dealing with a very complex problem with a vast number of factors, and further studies focusing on the effect of other factors are required to obtain a full understanding of how users make decisions on app usage. With the results of this thesis, we are one step closer to fully understand how users make decisions on app usage.

6.2 On Data Validity

We used the Carat application as source of energy measurements. The data collected by this application is biased towards active use because it records samples whenever the battery level changes, and may not be able to record data when the phone is in sleep mode, depending on the operating system version. The resulting battery life values represent the remaining time for actively using the device with a given application running 100% of the time. In the dataset, the most common e (around 0.005) represents an active battery life of 5 hours. To mitigate these

biases, we ensured selecting the location and apps with the most samples to foster better characterization of performance factors. Similar considerations apply to the NetRadar dataset used as source of latency data, which is predominantly collecting data whenever users explicitly request network performance assessment or periodically at user configurable intervals (between 1 and 120 minutes).

6.3 Data Quantity

Results of our model validation suggest that the number of training samples is critical for ensuring high-quality predictions regarding the extent to which performance affects retention. In the case of Finland, data from several hundreds of users was obtained while for the US only a few tens of users were retained after data fusion. Our data was collected from two mobile applications that have been in long-term usage worldwide, suggesting that crowdsensing is indeed essential for capturing a sufficient quantity of measurements and different contexts. However, our results also highlight the difficulties when *multiple* crowdsensing datasets need to be combined in that their intersection might be small, limiting the power of analyses carried on it.

6.4 Data Collection Mechanisms

The analysis used in this thesis relies on device monitoring. Compared to other data collection mechanisms, it is less intrusive and costly way to collect massive amounts of data in the wild. In general, data collection mechanisms are of critical importance when gathering information. The level of intrusiveness of data collection mechanisms influences the cost, quality and amount of information collected. Existing methods rely on application instrumentation, device and network monitoring to collect performance data. In this study, our methodology takes a step further by using passive data measurements obtained from different device monitoring to model complex contexts. Indeed, complex contexts that require capturing multiple parameters at the same time can be computationally exhaustive for a device, e.g., high-energy drain or induce performance degradation. Our methodology can be applied to analyse complex contexts where traditional crowdsensing falls short, and it is unfeasible to apply.

6.5 Fusion of Large-Scale Passive Data

In this work, we combine passive measurements from NetRadar and Carat datasets (data fusion) to study the effect that different performance factors have on retention. While we ensure that statistically the dataset combination is representative by analyzing and estimating similarity metrics of each dataset individually, we experienced a high reduction of available samples for analysis of the fused dataset, which was mainly due to limited coverage of USA in the NetRadar dataset. However, data fusion is necessary to ensure the variety and quality of contexts that we study. For example, the location, time, operators and communication technologies, etc., must be matched between the records of the two datasets. In other words, there is a trade-off between data size and data quality.

Besides, the individual nature of each dataset (NetRadar - network connectivity, Carat - App usage) also acted as a filter in the combination process, as further manipulation was required to match attributes in both datasets, e.g., reverse geocoding in the GPS of NetRadar to match the time zones of Carat data records. We were able to model the combined relation that energy and latency have on retention by merging the two datasets. Our methodology also provided insights about the relationship between performance factors that were initially hidden, but revealed when the different sources were combined, data leaks, similar to recent observations[8]. Specifically, we observed that when latency starts affecting retention, significant variations in energy efficiency are possible before retention is affected further. However, when battery life issues cause lower retention, latency can vary less before retention degrades further.

6.6 Users can Affect Results

Another way to obtain the combined dataset would be to collect all of the included fields directly using a single app that combines the functionality of Carat and Netradar. However, such an app would connect latency and energy consumption, and might bias users towards exploring that relationship as well, affecting the results. In addition, users typically troubleshoot one issue at a time on their device, so dedicated apps for network performance and energy awareness may gather more users separately than when combined.

[8]https://www.technologyreview.com/the-download/610086/fitness-app-data-is-revealing-military-bases-to-enemy-fighters/

6.7 Energy Efficiency Models can Influence Retention

As we mentioned earlier, some of the applications can be energy-hungry, and using applications can affect to a battery life of smartphones making extending it a primary research topic in academy and industry. Mobile devices are equipped with awareness mechanisms that monitor energy consumption based on applications usage and resources utilization. Based on this information, smartphones can decide whether to stop, outsource, or moderate the execution of tasks to save energy. While these mechanisms indeed induce gains in energy as the computation of tasks are reduced, they can foster collateral damage in the perception that users have towards apps. This observation suggests that app performance is diminished and augmented dynamically based on application usage. For instance, iOS devices implement a low power mode mechanism that reduces the computation of applications in the background to save energy[9]. This can potentially affect retention as the responsiveness of applications is degraded. By using our model in conjunction with energy efficiency models, it is possible to equip smartphones with a smarter mechanism that can save energy without degrading performance to an extent in which it is not tolerable to users anymore.

6.8 Influence of Performance Depends on Usage Patterns

As we found out during the research, application usage patterns can be very different. For example, Dropbox is a productivity application, which mostly runs on the background synchronizing photos. When the users interacts with it, they do so to find or share a file, using it infrequently and for a short period, which results in the smaller influence of performance degradation. On the other hand, Facebook and Twitter apps provide a continuous feed of updates enabling users to spend hours reading, watching, and interacting with content. Not only does this presenting a larger window of opportunity for performance issues to manifest, but this highlights how different usage patterns are likely to influence the importance of different performance factors.

Next we introduce the finally chapter with the summary of the findings and the thesis conclusion.

[9]http://www.tapsmart.com/tips-and-tricks/guide-understanding-low-power-mode-on-iphone-ios-10/

7 Summary and Conclusion

In this chapter, we present the summary of our results and the conclusion. One of the principal metrics used to describe app's success is retention. It shows the percentage of users continuing to use an app n days since the first use. While state-of-the-art has demonstrated that performance is a key factor that influences retention, this effect still has not been quantified. This problem is addressed by this thesis, contributing to quantify the impact of performance in retention, but also to model this effect. The impact is studied using two key factors that affect performance: latency and energy. Both factors are analysed individually and together through the combination of two large-scale crowdsensed datasets, the first including measurements about network performance and the second about app usage and battery consumption. We perform the analysis for Finland and Eastern USA which are the locations with the most representative number of samples. The same consideration is taken to perform the analysis at category and application level selecting the top five of each level. We quantify and model the effect of performance in long terms application usage. Our results demonstrate that low performance increases low retention likelihood, but also that improvement in performance not necessarily ensures high retention because it depends on what users consider a reasonable level of performance. In the following, we summarise our results and present their impact and open issues.

7.1 Datasets Combination

Unlike the actual studies about performance effect that analyse latency and energy independently, this thesis analyses performance effects using a combined latency/energy dataset. Data fusion allows acquiring richer contexts than individual datasets. Besides, it facilitates to obtain a better understanding about users' behaviour, deeper analysis to quantify performance effect (see section 4.5) and factors interaction (see section 4.3) and more accurate model (see section 5.4). However, data fusion is far to be trivial; low datasets alignment produces the decreasing of samples respecting to original data sources (see table 2), which is one of the collateral effects of data fusion. Indeed, combination methods still need to overcome challenges related to mitigation of data/context loss and guarantee representativeness of original data.

7.2 Performance on Long-term Mobile Apps Usage

Besides to confirm the results of previous research demonstrating that performance, measured in terms of latency and energy, significantly affects mobile apps retention, we quantify this effect for latency and energy, both individually and combined, by finding the critical point CP beyond performance degradation starts to affect retention. We demonstrate that long-term usage decreases when performance is poor, nevertheless we show that performance effect depends on location, app category and user's expectations. Hence, these factors regulate the location of the critical point CP. The quantification of performance effect also contributes to understanding users' response to factors variation. For example, we demonstrate latency to be the factor which is normally perceived faster by the users(see section 4.3). Similarly, when analysing the response at category and application level, we observe some categories to be more sensitive to variations in latency than energy (see section 4.2). While we analyse the combined effect, we show its high significance, but also its high complexity (see section 4.5) . Quantification of performance effect and CP' location can help to expand our learning about the combined effect and performance variables. Indeed, quantification of other factors affecting app retention is pending.The quantification approach used in this thesis can be useful to analyse the other factors and consequently improve our understanding of users' behaviour.

7.3 Modeling

We use the quantification of the effect of performance in retention to develop a model that predicts retention based on the expected performance at category-level. We use the samples from Finland, EST-USA and both combined predicting the retention values for locations individually and cross-country. The performance of our model is validated for latency, energy and both factors combined. Individual factor prediction presents marginal errors up to 1.49 for all categories when considering location-combined data (see section 5.3) On the other hand, combined factor prediction shows a maximum marginal error for the combined sample of 0.51 for all categories (see section 5.4) While our model has, in general, a good performance, prediction's accuracy depends on apps category overlapping, which limits the power of our model. Hence, it would be necessary first to establish the similarity between categories of each location to know the expected prediction error. Availability of apps is different for each location. Future studies should consider this issue to

improve cross-country model generalisation.

To summarise, this thesis contributes to studies focused on understanding mobile apps usage. We quantify and model the influence of performance in apps retention. Long-term usage is demonstrated to be affected by performance, but the latter is moderated by factors like app usage, location, app category and user's preferences. The performance was studied by combining two large-scale crowdsensed datasets that contain measurements of latency and energy, respectively. We demonstrate that normally latency effects are perceived first by users than energy. Finally, our model estimates retention based on the expected performance showing good performance when applied at the cross-country level.

References

AHP+14a Aggarwal, V., Halepovic, E., Pang, J., Venkataraman, S. and Yan, H., Prometheus: Toward quality-of-experience estimation for mobile apps from passive network measurements. *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications*. ACM, 2014, page 18.

AHP+14b Aggarwal, V., Halepovic, E., Pang, J., Venkataraman, S. and Yan, H., Prometheus: Toward quality-of-experience estimation for mobile apps from passive network measurements. *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications*, HotMobile '14, New York, NY, USA, 2014, ACM, pages 18:1–18:6, URL `http://doi.acm.org/10.1145/2565585.2565600`.

AL10 Andridge, R. R. and Little, R. J., A review of hot deck imputation for survey non-response. *International statistical review*, 78,1(2010), pages 40–64.

ALvK+14 Athukorala, K., Lagerspetz, E., von Kügelgen, M., Jylhä, A., Oliner, A. J., Jacucci, G. and Tarkoma, S., How Carat Affects User Behavior: Implications for Mobile Battery Awareness Applications. *CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2014, ACM.

BBV09a Balasubramanian, N., Balasubramanian, A. and Venkataramani, A., Energy consumption in mobile phones: A measurement study and implications for network applications. *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, IMC '09, New York, NY, USA, 2009, ACM, pages 280–293, URL `http://doi.acm.org/10.1145/1644893.1644927`.

BBV09b Balasubramanian, N., Balasubramanian, A. and Venkataramani, A., Energy consumption in mobile phones: a measurement study and implications for network applications. *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*. ACM, 2009, pages 280–293.

BCCR14 Banerjee, A., Chong, L. K., Chattopadhyay, S. and Roychoudhury, A., Detecting energy bugs and hotspots in mobile apps. *Proceedings of the 22Nd ACM SIGSOFT International Symposium on Foundations of*

Software Engineering, FSE 2014, New York, NY, USA, 2014, ACM, pages 588–598.

BH10 Brooks, P. and Hestnes, B., User measures of quality of experience: why being objective and quantitative is important. *IEEE Network Magazine*, 24,2(2010).

BHS+11 Böhmer, M., Hecht, B., Schöning, J., Krüger, A. and Bauer, G., Falling asleep with angry birds, facebook and kindle: A large scale study on mobile application usage. *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, 2011.

BJSH15 Baeza-Yates, R. A., Jiang, D., Silvestri, F. and Harrison, B., Predicting the next app that you are going to use. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015*, 2015, pages 285–294, URL `http://doi.acm.org/10.1145/2684822.2685302`.

Bra91 Bradner, S., Benchmarking terminology for network interconnection devices. Technical Report, 1991.

BSA+13 Balachandran, A., Sekar, V., Akella, A., Seshan, S., Stoica, I. and Zhang, H., Developing a predictive model of quality of experience for internet video. 43,4(2013), pages 339–350.

CHL06 Chen, K.-T., Huang, P. and Lei, C.-L., How sensitive are online gamers to network quality? *Communications of the ACM*, 49, pages 34–38.

CLH+14a Chen, N., Lin, J., Hoi, S. C. H., Xiao, X. and Zhang, B., Ar-miner: mining informative reviews for developers from mobile app marketplace. *36th International Conference on Software Engineering, ICSE '14, Hyderabad, India - May 31 - June 07, 2014*, 2014, pages 767–778, URL `http://doi.acm.org/10.1145/2568225.2568263`.

CLH+14b Chen, N., Lin, J., Hoi, S. C., Xiao, X. and Zhang, B., Ar-miner: mining informative reviews for developers from mobile app marketplace. *Proceedings of the 36th International Conference on Software Engineering*. ACM, 2014, pages 767–778.

CTX09 Chen, K.-T., Tu, C.-C. and Xiao, W.-C., Oneclick: A framework for measuring network quality of experience. *INFOCOM 2009, IEEE*. IEEE, 2009, pages 702–710.

DHD10 Deshpande, P., Hou, X. and Das, S. R., Performance comparison of 3g and metro-scale wifi for vehicular network access. *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 2010, pages 301–307.

DNSB14 Deng, S., Netravali, R., Sivaraman, A. and Balakrishnan, H., Wifi, lte, or both?: Measuring multi-homed wireless internet performance. *Proceedings of the 2014 Conference on Internet Measurement Conference*. ACM, 2014, pages 181–194.

dRBHV12 de Reuver, M., Bouwman, H., Heerschap, N. and Verkasalo, H., Smartphone measurement: do people use mobile applications as they say they do? *International Conference on Mobile Business, ICMB 2012, Delft, The Netherlands, June 21-22, 2012*, 2012, page 2, URL `http://aisel.aisnet.org/icmb2012/2`.

DSA+11 Dobrian, F., Sekar, V., Awan, A., Stoica, I., Joseph, D., Ganjam, A., Zhan, J. and Zhang, H., Understanding the impact of video quality on user engagement. *SIGCOMM Comput. Commun. Rev.*, 41,4(2011), pages 362–373. URL `http://doi.acm.org/10.1145/2043164.2018478`.

FGK+14 Ferreira, D., Goncalves, J., Kostakos, V., Barkhuus, L. and Dey, A. K., Contextual experience sampling of mobile application micro-usage. *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services*, MobileHCI '14, New York, NY, USA, 2014, ACM, pages 91–100, URL `http://doi.acm.org/10.1145/2628363.2628367`.

FHN+17 Flores, H., Hui, P., Nurmi, P., Lagerspetz, E., Tarkoma, S., Manner, J., Kostakos, V., Li, Y. and Su, X., Evidence-aware mobile computational offloading. *IEEE Transactions on Mobile Computing*, PP,99(2017), pages 1–1.

FHN+18 Flores, H., Hui, P., Nurmi, P., Lagerspetz, E., Tarkoma, S., Manner, J., Kostakos, V., Li, Y. and Su, X., Evidence-aware mobile computational

offloading. *IEEE Transactions on Mobile Computing*, 17,8(2018), pages 1834–1850.

FHTG10 Fiedler, M., Hossfeld, T. and Tran-Gia, P., A generic quantitative relationship between quality of experience and quality of service. *IEEE Network*, 24,2(2010), pages 36–41.

FLL+13a Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J. and Sadeh, N., Why people hate your app: Making sense of user feedback in a mobile app store. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, New York, NY, USA, 2013, ACM, pages 1276–1284, URL `http://doi.acm.org/10.1145/2487575.2488202`.

FLL+13b Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J. and Sadeh, N., Why people hate your app: Making sense of user feedback in a mobile app store. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pages 1276–1284.

FSK+17 Flores, H., Su, X., Kostakos, V., Riekki, J., Lagerspetz, E., Tarkoma, S., Hui, P., Li, Y. and Manner, J., Modeling mobile code acceleration in the cloud. *Proceedings of the Annual IEEE International Conference on Distributed Computing Systems (ICDCS 2017)*, (Atlanta, GA, USA), June 5-8, 2017.

GK15 Gabale, V. and Krishnaswamy, D., MobInsight: On Improving The Performance of Mobile Apps in Cellular Networks. *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, Republic and Canton of Geneva, Switzerland, 2015, International World Wide Web Conferences Steering Committee, pages 355–365, URL `https://doi.org/10.1145/2736277.2741138`.

GYL11 Ganti, R. K., Ye, F. and Lei, H., Mobile crowdsensing: current state and future challenges. *IEEE Communications Magazine*, 49,11(2011).

HQG+12a Huang, J., Qian, F., Gerber, A., Mao, Z. M., Sen, S. and Spatscheck, O., A close examination of performance and power characteristics of 4g lte networks. *Proceedings of the 10th international conference on Mobile systems, applications, and services*. ACM, 2012, pages 225–238.

HQG+12b Huang, J., Qian, F., Gerber, A., Mao, Z. M., Sen, S. and Spatscheck, O., A close examination of performance and power characteristics of 4g lte networks. *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, MobiSys '12, New York, NY, USA, 2012, ACM, pages 225–238, URL `http://doi.acm.org/10.1145/2307636.2307658`.

HXT+10 Huang, J., Xu, Q., Tiwana, B., Mao, Z. M., Zhang, M. and Bahl, P., Anatomizing application performance differences on smartphones. *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, MobiSys '10, New York, NY, USA, 2010, ACM, pages 165–178, URL `http://doi.acm.org/10.1145/1814433.1814452`.

IWF+12 Ickin, S., Wac, K., Fiedler, M., Janowski, L., Hong, J.-H. and Dey, A. K., Factors influencing quality of experience of commonly used mobile applications. *IEEE Communications Magazine*, 50,4(2012).

KSNH15 Khalid, H., Shihab, E., Nagappan, M. and Hassan, A. E., What do mobile app users complain about? *IEEE Software*, 32,3(2015), pages 70–77.

LCZ+13 Lane, N. D., Chon, Y., Zhou, L., Zhang, Y., Li, F., Kim, D., Ding, G., Zhao, F. and Cha, H., Piggyback crowdsensing (pcs): energy efficient crowdsourcing of mobile sensor data by exploiting smartphone app opportunities. *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2013, page 7.

MCC11 Mok, R. K., Chan, E. W. and Chang, R. K., Measuring the quality of experience of http video streaming. *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on*. IEEE, 2011, pages 485–492.

MKC12 Mittal, R., Kansal, A. and Chandra, R., Empowering developers to estimate app energy consumption. *Proceedings of the 18th annual international conference on Mobile computing and networking*. ACM, 2012, pages 317–328.

Mye11 Myers, T. A., Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication Methods and Measures*, 5,4(2011), pages 297–310.

NCKB+14 Nikravesh, A., Choffnes, D. R., Katz-Bassett, E., Mao, Z. M. and Welsh, M., Mobile network performance from user devices: A longitudinal, multidimensional analysis. *International Conference on Passive and Active Network Measurement*. Springer, 2014, pages 12–22.

NFS00 Narayanan, D., Flinn, J. and Satyanarayanan, M., Using history to improve mobile application adaptation. *IEEE Workshop on Mobile Computing Systems and Applications*, Monterey, CA, USA, December 7–8, 2000.

OIS+13 Oliner, A. J., Iyer, A. P., Stoica, I., Lagerspetz, E. and Tarkoma, S., Carat: Collaborative energy diagnosis for mobile devices. *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2013, page 10.

ope Opencellid project, `https://opencellid.org/`. Accessed: 2018-04-6.

ORMR12 Oulasvirta, A., Rattenbury, T., Ma, L. and Raita, E., Habits make smartphone use more pervasive. *Personal and Ubiquitous Computing*, 16,1(2012), pages 105–114. URL `https://doi.org/10.1007/s00779-011-0412-2`.

Pan17 Panosian, H., Design your app. In *Learn iOS Application Distribution*, Springer, 2017, pages 95–123.

PHZ+11 Pathak, A., Hu, Y. C., Zhang, M., Bahl, P. and Wang, Y.-M., Fine-grained power modeling for smartphones using system call tracing. *Proceedings of the sixth conference on Computer systems*, New York, NY, USA, 2011, ACM, pages 153–168, URL `http://doi.acm.org/10.1145/1966445.1966460`.

PJHM12 Pathak, A., Jindal, A., Hu, Y. C. and Midkiff, S., What is keeping my phone awake? Characterizing and detecting no-sleep energy bugs in smartphone apps. *Mobisys*, 2012.

PLNT15a Peltonen, E., Lagerspetz, E., Nurmi, P. and Tarkoma, S., Energy modeling of system settings: A crowdsourced approach.

2015 IEEE International Conference on Pervasive Computing and Communications (PerCom), PerCom'15. IEEE, 2015, pages 37 – 45, URL http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7146507.

PLNT15b Peltonen, E., Lagerspetz, E., Nurmi, P. and Tarkoma, S., Energy modeling of system settings: A crowdsourced approach. *Pervasive Computing and Communications (PerCom), 2015 IEEE International Conference on*. IEEE, 2015, pages 37–45.

PLNT16 Peltonen, E., Lagerspetz, E., Nurmi, P. and Tarkoma, S., Constella: Crowdsourced system setting recommendations for mobile devices. *Pervasive and Mobile Computing*, 26,Supplement C(2016), pages 71 – 90. URL http://www.sciencedirect.com/science/article/pii/S1574119215001959. Thirteenth International Conference on Pervasive Computing and Communications (PerCom 2015).

RCT+07 Rogers, Y., Connelly, K., Tedesco, L., Hazlewood, W., Kurtz, A., Hall, R. E., Hursey, J. and Toscos, T., Why it's worth the hassle: The value of in-situ studies when designing ubicomp. *UbiComp 2007: Ubiquitous Computing*, Krumm, J., Abowd, G. D., Seneviratne, A. and Strang, T., editors, Berlin, Heidelberg, 2007, Springer Berlin Heidelberg, pages 336–353.

Rei93 Reilly, M., Data analysis using hot deck multiple imputation. *The Statistician*, pages 307–313.

RPA+12 Ravindranath, L., Padhye, J., Agarwal, S., Mahajan, R., Obermiller, I. and Shayandeh, S., Appinsight: Mobile app performance monitoring in the wild. *OSDI*, volume 12, 2012, pages 107–120.

RQZ07 Rahmati, A., Qian, A. and Zhong, L., Understanding human-battery interaction on mobile phones. *Proceedings of the 9th International conference on Human computer interaction with mobile devices and services*, New York, NY, USA, 2007, ACM, pages 265–272, URL http://doi.acm.org/10.1145/1377999.1378017.

SAA+15 Spetebroot, T., Afra, S., Aguilera, N., Saucez, D. and Barakat, C., From network-level measurements to expected quality of experience:

The skype use case. *2015 IEEE International Workshop on Measurements Networking (M N)*, Oct 2015, pages 1–6.

SLP+18 Sigg, S., Lagerspetz, E., Peltonen, E., Nurmi, P. and Tarkoma, S., Exploiting usage to predict instantaneous app popularity: Trend filters and retention rates. *ACM Transactions on the Web (TWEB)*, 2018.

SMS13 Sonntag, S., Manner, J. and Schulte, L., Netradar-measuring the wireless world. *Modeling & Optimization in Mobile, Ad Hoc & Wireless Networks (WiOpt), 2013 11th International Symposium on*. IEEE, 2013, pages 29–34.

SN01 Satyanarayanan, M. and Narayanan, D., Multi-fidelity algorithms for interactive mobile applications. *Wireless Networks*, 7,6(2001), pages 601–607.

SN13 Salehan, M. and Negahban, A., Social networking on smartphones: When mobile phones become addictive. *Computers in Human Behavior*, 29,6(2013), pages 2632–2639.

SRRM+17 Servia-Rodríguez, S., Rachuri, K. K., Mascolo, C., Rentfrow, P. J., Lathia, N. and Sandstrom, G. M., Mobile sensing at the service of mental well-being: a large-scale longitudinal study. *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pages 103–112.

TLN+14 Truong, H. T. T., Lagerspetz, E., Nurmi, P., Oliner, A. J., Tarkoma, S., Asokan, N. and Bhattacharya, S., The company you keep: Mobile malware infection rates and inexpensive risk indicators. *23nd International World Wide Web Conference, WWW '14, Seoul, Korea, April 7-11, 2014*, 2014, pages –.

TWLC11 Tseng, P. H., Wang, N. C., Lin, R. M. and Chen, K. T., On the battle between lag and online gamers. *Proceedings of the 2011 IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR)*, 2011.

Val98 Valentine, E. L., Location based screening in a mobile telecommunications system, September 29 1998. US Patent 5,815,808.

War09 Warburton, S., Second life in higher education: Assessing the potential for and the barriers to deploying virtual worlds in learning and teaching. *British journal of educational technology*, 40,3(2009), pages 414–426.