

Alma Mater Studiorum Università di Bologna  
Archivio istituzionale della ricerca

Designing Robust Models for Behaviour Prediction Using Sparse Data from Mobile Sensing: A Case Study of Office Workers' Availability for Well-being Interventions

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

Kucukozer-Cavdar, S., Taskaya-Temizel, T., Mehrotra, A., Musolesi, M., Tino, P. (2021). Designing Robust Models for Behaviour Prediction Using Sparse Data from Mobile Sensing: A Case Study of Office Workers' Availability for Well-being Interventions. ACM TRANSACTIONS ON COMPUTING FOR HEALTHCARE, 2(4), 1-33 [10.1145/3458753].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/870407> since: 2022-02-26

*Published:*

DOI: <http://doi.org/10.1145/3458753>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

**Kucukozer-Cavdar, Seyma and Taskaya-Temizel, Tugba and Mehrotra, Abhinav and Musolesi, Mirco and Tino, Peter, Designing Robust Models for Behaviour Prediction Using Sparse Data from Mobile Sensing: A Case Study of Office Workers' Availability for Well-Being Interventions (2021), in ACM Transactions on Computing for Healthcare, vol. 2, n. 4, pp. 1-33**

The final published version is available online at: <https://doi.org/10.1145/3458753>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

*This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)*

***When citing, please refer to the published version.***

# Designing Robust Models for Behaviour Prediction using Sparse Data from Mobile Sensing: A Case Study of Office Workers' Availability for Wellbeing Interventions

SEYMA KUCUKOZER-CAVDAR and TUGBA TASKAYA-TEMIZEL, Middle East Technical University, Turkey

ABHINAV MEHROTRA, Samsung AI Centre, United Kingdom

MIRCO MUSOLESI, University College London, United Kingdom, The Alan Turing Institute, United Kingdom, and University of Bologna, Italy

PETER TINO, The University of Birmingham, United Kingdom

Understanding in which circumstances office workers take rest breaks is important for delivering effective mobile notifications and make inferences about their daily lifestyle, e.g., whether they are active and/or have a sedentary life. Previous studies designed for office workers show the effectiveness of rest breaks for preventing work-related conditions. In this paper, we propose a hybrid personalised model involving a kernel density estimation model and a generalised linear mixed model to model office workers' available moments for rest breaks during working hours. We adopt the experience-based sampling method through which we collected office workers' responses regarding their availability through a mobile application with contextual information extracted by means of the mobile phone sensors. The experiment lasted 10 workdays and involved 19 office workers with a total of 528 responses. Our results show that time, location, ringer mode, and activity are effective features for predicting office workers' availability. Our method can address sparse sample issues for building individual predictive behavioural models based on limited and unbalanced data. In particular, the proposed method can be considered as a potential solution to the "cold-start problem", i.e., the negative impact of the lack of individual data when a new application is installed.

CCS Concepts: • **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; • **Computing methodologies** → **Kernel methods**; *Ensemble methods*; **Classification and regression trees**.

Additional Key Words and Phrases: personalised modelling, cold-start problem, rest break prediction, hybrid model

## 1 INTRODUCTION

Mobile devices have been widely used for health intervention delivery in recent years. There have been several projects focusing on when and how to deliver effective intervention messages (e.g., [25, 49]). One of the main areas where health-related interventions have been adopted is wellbeing at workplaces, especially in offices in which employees tend to have a sedentary behaviour during working hours [32, 37, 52]. Sedentary workstyle results in work-related musculoskeletal disorders (WRMSDs) such as carpal tunnel syndrome [56] or repetitive strain injury (RSI) [35]. It has a negative impact on employees' health and also causes economic costs to organisations. For that purpose, many organisations and agencies provide educational material and training to their employees about WRMSDs and how to prevent them, and also support interventions for behaviour change [42]. The most common solution for preventing WRMSDs is to have simple micro breaks from repetitive work [4, 26, 32, 42]. However, notification timing is important since taking breaks from repetitive work may decrease the productivity of employees if the intervention is not delivered at the right moment.

---

Authors' addresses: Seyma Kucukozer-Cavdar, seymakucukozer@gmail.com; Tugba Taskaya-Temizel, ttemizel@metu.edu.tr, Middle East Technical University, Dumlupinar Bulvari, Ankara, Turkey; Abhinav Mehrotra, a.mehrotra1@samsung.com, Samsung AI Centre, Cambridge, United Kingdom; Mirco Musolesi, m.musolesi@ucl.ac.uk, University College London, Gower Street, London, United Kingdom, The Alan Turing Institute, London, United Kingdom, University of Bologna, Bologna, Italy; Peter Tino, P.Tino@cs.bham.ac.uk, The University of Birmingham, Birmingham, United Kingdom.

When the health interventions regarding rest breaks are delivered through mobile phones, circumstances such as the time of the day and location in which office workers tend to take rest breaks become important contextual information that can be extracted by means of mobile phone sensors. This information is then used to infer the most appropriate time to send a notification. Several studies have shown that notifications sent based on time, users' activity, location, mobile application usage, cognitive context or ringer mode are more likely to be accepted and acted upon by users [9, 15, 22, 30, 31, 33, 39, 40, 43, 44, 48].

Self-reported data collection for user modelling studies (e.g., for modelling interruptibility or break availability) is a time-consuming but an essential task. In the recent years, it has been shown that models built on individual data (i.e., individual models) are superior to the models that are built on all users' data (i.e., general or population models) in predicting interruptibility since user characteristics vary across populations [18, 61]. However, individual models have an obvious drawback: a certain amount of data has to be collected in order to start making predictions. A limited amount of data per user can usually be obtained due to dropouts and missed/ignored messages/notifications. In addition, the number of data points per user varies significantly: this creates a bias in modelling, i.e., some users might be over-represented (or under-represented) in the training dataset. The superiority of individual models over generic models has been well-demonstrated when there is sufficient data [61]. In other words, the model suffers from the so-called *cold-start problem*. However, the effectiveness of individual-level models with unbalanced limited data has not been investigated in the past. This is a topic of crucial importance, since addressing this problem is fundamental for the implementation of these mechanisms in practical applications.

The goal of this paper is to address the issues listed above. In other words, the aim of this work can be summarised in the following research question: *"Is it possible to develop a model for inferring availability of office workers and their willingness of having rest breaks using information from mobile phone sensors by considering the variety in the number and characteristics of the responses and addressing the cold start problem at the same time?"* In particular, in order to answer this question, we propose a *hybrid model*, which is able to take into account both the specific individual and the entire population of users simultaneously. It is composed of two stages: first, we computed the kernel density estimates of each user's self-reported break availability response with its corresponding timestamp. In mobile interventions, user responses (including non-responses) with their corresponding times can be quite helpful to infer users' availability. For instance, at certain times of the day, users may be more willing to reply to notifications regardless of their current location.

We first conducted a survey on 19 office workers and requested them to install our mobile sensing application, which delivers break reminder notifications. These include questions for identifying the context of the office workers. The experiment was conducted for ten working days. The mobile application used in the study collected mobile sensor data in the background during the experiment. We obtained 528 valid responses that were used to construct our hybrid model using generalised linear mixed modelling (GLMM). Our solution is based on a novel generic *population-* and *individual-level* model for unbalanced and limited amount of data.

The main contributions of this paper can be given as follows:

- (1) In this study, we propose a hybrid model, which is based on a novel approach. It relies on both individual-level and population-level data to offer a possible solution to the cold-start problem. We modelled availability of office workers with the features selected (kernel density estimates, location parameters, ringer mode parameters, physical activity, and mobile application usage) with GLMM using a Markov chain Monte Carlo method. The main advantage of GLMM resides in its ability to take into consideration both within and between subjects' factors successfully. The use of kernel density estimation facilitated the incorporation of time information into a linear model. Besides, using GLMM at the first stages of an interruption might be a solution to the cold-start problem of individual-level models when there is unbalanced limited data to predict a user's availability.

- (2) We propose new features to represent the interaction between users and indoor locations for modelling user's availability for taking breaks. In the literature of context-aware modelling, GPS data has been utilised extensively for identifying location [8, 43]. However, there is a limited number of studies focusing on the characterisation of indoor locations for context awareness. In this study, we define several parameters based on the indoor locations of users such as the duration that they spend in each location, visit frequency to locations, and location similarity with their workplaces. We investigate the effects of those parameters on the availability for taking rest breaks.

The findings of our study can be summarised as follows: (a) time of the day, activity, and time duration of a certain action have significant effects on break availability of office workers for a rest break, (b) GLMM has a higher predictive power than individual or generalised random forest models when there is a lack of sufficient and repeated-measures data, (c) consequently, GLMM can be considered as a potential method for dealing with the “cold-start period”, which affects individual-level models.

We believe that researchers and practitioners in the area of mobile systems for healthcare will benefit from the findings of our study: in particular, developers who work mainly in the mobile health domain might consider contextual variables such as time spent in action and activity as input for the design of intelligent mechanisms for delivering notifications and messages to users, for example in the design of effective positive behaviour interventions for healthcare, especially in office contexts.

## 2 RELATED WORK

We now review the state of the art in two areas that are relevant to this study, namely interruptibility and studies about break reminders and physical activity promotion in workplaces.

### 2.1 Interruptibility

The most prominent factors that impact or are associated with interruptibility are time [30, 33, 39, 43, 50, 53], physical activity [22, 24, 30, 33, 34, 36, 53], location [8, 9, 13, 30, 39, 43, 46, 48], application usage [33, 34, 36, 40, 44], and ringer mode [8, 31, 33, 40, 50].

In terms of time granularity, researchers have considered hour of the day or other subdivision of the day, such as morning or evening. There are contradictory results about time being significant for interruptibility: the majority of the studies showed time as a significant variable [30, 33, 39], whereas other variables rather than time have been found significant on interruptibility in a few studies [43, 50, 53]. Those studies suggest that the effect of time varies.

One of the significant features for detecting available moments is physical activity. It has been found that users are more interruptible during activity transitions [22, 24]. In recent works [30, 33], activity type (e.g., walking, standing, running) obtained from mobile devices has been used for an interruptibility management system. A number of recent studies (see, for example, [34, 36, 53]) have also used information about the type of activity to predict opportune moments for notification delivery.

Location is generally collected by GPS [30, 43], user feedback [9, 13, 39, 48] or both [8]. The main limitation of GPS is that it is unreliable for indoor localisation [55]. Poppinga et al. [43] had to mark 90% of the locations they collected as “no location” because they were indoors. Ter Hofte [48] and Choi et al. [9] collected the locations of users with experience-based sampling (ESM) solely, whereas Pejovic and Musolesi [39] also used location changes with Wi-Fi and Bluetooth in addition. Likewise, Chang and Tang [8] classified locations obtained from GPS as “work” or “home” locales and discussed the receptivity of users based on different locales. Finally, Exler et al. [13] presented location groups to users and measured their receptivity in different locations. These studies show that users' receptivity to messages or interruptions varies based on the location.

Ringer mode is another modality discussed in recent works. Specifically, Chang and Tang [8] analysed the ringer mode settings in different contexts. Their findings suggest that individuals use different ringer modes in different locales. In [31, 40], researchers found that users respond to messages faster when their phones are in vibrate mode compared to ring or silent mode. Similarly, Turner, Alan, and Whitaker [50] showed that people are more receptive when their phones are in sound mode or vibrate mode compared to silent mode.

Finally, Fischer et al. [14] found that users tend to respond more quickly to notifications after an episode of mobile use. Oh, Jalali and Rain [33] considered the application usage duration as an input for a solution for smart notifications. In addition, Park et al. [36] and Okoshi et al. [34] studied how device usage can be informative for the estimation of break-points.

## 2.2 Studies Related to Break Reminders in Office Environments

Most office workers face repetitive movements and/or static postures in their work lives. Many studies [4, 16, 32] suggest simple breaks for preventing medical conditions related to computer use (e.g., carpal tunnel syndrome, tension neck syndrome). Healy et al. [21] showed the association between taking frequent breaks from sitting and a healthier metabolism. Numerous studies have proved the effectiveness of reminders sent through applications for desktop computers or mobile devices [5, 7, 10, 47, 52]. In particular, they found that timing is a critical factor.

Van Dantzig et al. [52] showed the importance of taking breaks and how persuasive strategies might be effectively adopted to increase individuals' willingness to take them. Berque et al. [5] designed a software system that persuades users to avoid immoderate typing speeds, to use typing shortcuts, and to take breaks from typing in order to prevent RSI. The system reminded participants when they exceeded a typing speed and provide them with appropriate feedback. The results show that the feedback provided by the system has a positive effect on typing behaviour, and shortcuts for words are used more effectively. As a different perspective, Taylor et al. [47] suggested two main determinants leading to physical activity behaviour in office environments: (1) attitudes, behavioural and social determinants, and (2) environmental and policy determinants. Knowledge, behavioural management skills, self-efficacy, enjoyment, perceived benefits, perceived barriers, and social support from family, coworkers and friends constitute the first category, whereas workplace norms and "culture", management support and available physical space constitute the latter. Hence, they emphasised the workplace routines have an important role for workers to take breaks or do exercises. Because of that, the social determinants in offices are explored in this study. Cooley and Pedersen [10] conducted a study with 46 participants for increasing non-purposeful movement breaks at work in order to reduce prolonged sitting times. They designed a persuasive software that reminds employees to take a break from their sitting times. They concluded that reminders should be unobtrusive for the system to be successful in the long term.

Epstein, Avrahami, and Biehl [12] investigated not only the cost of interruptions on work tasks, but also on disrupting a break itself. The authors also discussed the importance of understanding in which circumstances office workers take rest breaks. A very recent study [54] describes how to improve the design of systems for activity tracking in workplaces including those for measuring social dynamics and extracting daily routines of the employees.

Other studies have investigated interruptibility in an office context. Züger and Fritz [60] used psycho-physiological sensors in order to identify software developers' interruptibility with an experiment conducted in a lab and a field study. They were able to successfully predict interruptibility. Moreover, they found positive correlation between interruptibility and mental load. Another study [59] focused on reducing in-person interruptions in the workplace combining a physical interruptibility indicator (a LED light) with an automatic interruptibility measure using computer activities. Even though they focused on interruptibility in an office context, their methods differ from our study. We focus on detecting available moments for breaks of different duration using mobile sensing.

All the studies showed the importance of personal and social factors in a workplace for effective interventions. As discussed in Section 2.2, the solutions presented by these studies show that break reminder applications are effective especially for work-related conditions such as RSI. However, these systems should strike a balance between healthy behaviour in the office and productivity, by minimising the messages sent to the users in inappropriate moments. In particular, the studies in Section 2.1 offer solutions for detecting interruptibility using context information using data from computers, smartphones or wearable sensors.

The goal of our study is to develop a system exploring the trade-off between effectiveness of the interventions and the number of notifications/messages sent to the users.

### 3 RESEARCH METHOD AND EXPERIMENTAL DESIGN

Mobile-based ESM was adopted throughout the study. ESM is used for capturing and recording human behaviour as it happens in their natural settings [11]. Hence, the data obtained by ESM has higher validity and less bias compared to other methods [38]. For these reasons, ESM is indeed a powerful methodology for capturing users' natural feelings and thoughts.

We designed a user experiment for investigating the factors related to rest break availability of office workers. The main steps of the experiment design are as follows: (1) Delivery of the pre-experiment questionnaire, which consisted of questions about demographic information, routine break times, and daily routines in mobile phone usage; (2) Installation of the mobile sensing application; (3) ESM sampling during 10 work days; (4) Model fitting.

#### 3.1 Pre-Experiment Questionnaire

We delivered a pre-experiment questionnaire to collect demographic information (age, gender, occupation, job role, and educational background) to understand the work routines of the participants. The work start/end hours and the participants' favourite times for breaks, type of the break (e.g., social, tea/coffee, lunch), location of the breaks (in the same office, outside the office on the same/different floor, or outside the building) and availability for a physical exercise in those breaks were also collected in order to understand work routines of the participants. Finally, the participants were asked to state in which ringer mode they use their mobile phones at each possible situation (e.g., in a meeting, in a rest break) during work hours.

#### 3.2 Design of the Mobile Sensing Application

We developed a mobile sensing application to remind rest breaks to employees during office hours, which is also used for collecting a variety of contextual information. The application works on mobile devices equipped with the Android operating system version 4.2 or higher. It has a basic user interface with three main menus. In the first menu, 5-minute and 10-minute-long two videos showing basic office exercises are presented to users, one of which can be preferred according to the duration of their break. In the second one, the reminder messages are shown. In the third one, an explanation about the study is displayed and a communication box is provided to the participants in order for them to access researchers for their questions. Sample screenshots are depicted in Figure 1.

In the background, the mobile application collects Wi-Fi access point information, GPS location, accelerometer, activity type returned by the Google Activity Recognition API, ringer mode, Google calendar data, screen activity, and application usage information. After the Android API level 21, the UsageStatsManager API requires system-level permission and is not granted to third-party apps. Users are required to make manual security changes from settings and some might prefer to opt out. Consequently, the screen on/off times and user's screen presence are also captured at all times. Through this information we were able to detect when a user started using their mobile phone, which applications were used, and when their interaction with their phone ended. In particular,



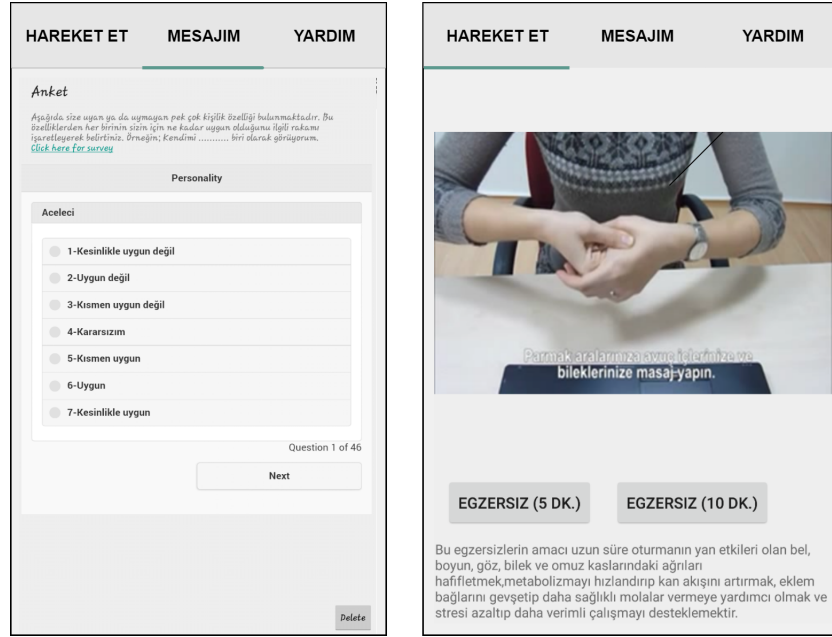


Fig. 1. Screenshots of the mobile sensing application. *Left*: A screenshot of the form used for questions. *Right*: A screenshot showing a video used to promote office exercises.

for those who made the necessary security changes, the mobile application continuously stores which and when application packages have been used.

The ringer mode is recorded every time it is changed. Activity data is only captured when significant motion (i.e., a sensor is triggered when a change is identified in the user's location) [1] is detected. The sampling starts whenever the participant starts moving and the activity type (e.g., walking, driving, and running) is stored in one-minute intervals until the user stops. In each state transition from *still* to *moving* and *moving* to *still*, Wi-Fi access point data including BSSID, SSID, frequency and level information is recorded. For the phones without a significant motion sensor activity, accelerometer data is recorded every three seconds for one minute in five-minute intervals. A server application was developed to automatically download the sensor data from each user at the end of the day.

### 3.3 Notification Delivery and ESM Questions

We delivered the ESM questions to the participants through the mobile sensing application each time we sent a rest break reminder; hence, we used ESM questions as the reminders themselves. The ESM questions used in the study are:

- **Question 1:** How long is your current break duration or for how long are you able to take a break? *Options:* 1 = cannot take a break now, 2 = less than 5 minutes, 3 = between 5-15 minutes, 4 = more than 15 minutes
- **Question 2:** What are you doing now? *Options:* In a meeting, working on computer, in a tea/coffee break, in a lunch/snack break, on road, in a social break (chatting, talking on the phone etc.), in a bathroom break, in a smoking break, following the media (news, magazines etc.), in a health break, in a praying break, and other.



The reminder delivery algorithm is designed to dispatch notifications only during working hours. First, the algorithm discards the time slots when the user has a calendar event. Then, it selects two notification times from the preferred time slots of the users, if stated in the pre-experiment questionnaire. Then, the remaining four notification times are randomly picked. If a participant has no preference, all six notification times are randomly chosen. Each time a reminder is sent, a motivational message for taking a break or doing a simple exercise is also included. More specifically, the notification times must satisfy the following constraints for each user: (1) The notification times must be delivered between work start time and end time. (2) The notification times must not overlap with calendar events. (3) If stated, two of the notification times must be selected from the preferred time slots of the user. For example, if a break is scheduled between 12:00 and 13:00, a random time is picked from this interval. (4) There must be at least one hour between two notification times. The pseudocode of the notification delivery algorithm is given in Appendix A.

When a reminder is sent with an ESM questionnaire, the message is seen in the notification bar as a regular notification. The intensity of the notifications changes based on the ringer mode of devices, i.e., it is based on system default settings. For example, if the device is in vibrate mode, the system notifies the user with a vibration. The participants were not asked to disable the “mute” mode. There were no re-prompts if users missed a prompt. Three possible scenarios may happen after the arrival of the message: (1) user sees the message and responds to it; (2) user sees the message but does not respond; or (3) user does not see the message. We refer to the time a user takes to respond to a notification as “response notification time”. If the notification is not responded in 15 minutes after its delivery, it is automatically deleted as in a previous study [31].

### 3.4 Procedure

**3.4.1 Ethics Committee Approval.** The ethics committee approval for data collection through mobile phones and questionnaires were obtained from the university’s research centre for applied ethics.

**3.4.2 Participant Recruitment.** Participants were recruited with convenience sampling method. The experiment was advertised among the students of a graduate institute in a university and promoted in social media. The participants were directed to the website of the experiment, in which we presented the pre-experiment questionnaire and the mobile application download links to them. The participants were granted a coffee coupon if they filled the pre-questionnaire and replied to at least 25% of the ESM messages, and that coupon was the only compensation for taking part in the experiment.

**3.4.3 Pilot Phase.** The main aim of the pilot phase is to evaluate all the experiment steps to oversee any potential technical problems such as scaling that might be encountered during the main experiment, which was planned with a larger participant pool. The experiment was conducted with five participants having different mobile phones in terms of brands and Android versions, and working in different workplaces. Three of the participants were male and two of them were female. Four participants were affiliated with different universities. One of them was an engineer working in a corporate company. A summary of the descriptive statistics about the participants is given in Table 1.

The procedure in this phase was conducted exactly in the same way as the main experiment. Nevertheless, two additional datasets were collected, which are specific to this phase. This is due to the fact that this additional data collection was time-consuming and cannot be easily extended to a large participant pool. The first dataset was constructed based on the answers of the five participants who were requested to fill a form indicating their location information (e.g., home, work, outdoor) at the end of each day. To facilitate the process, we automatically identified all the locations where each user spent more than five minutes each day and presented them with the associated timestamp. Finally, the participants labelled them. The second dataset comprised both the raw activity data and the labelled activity data by Google Activity API. These additional datasets were used for the following

Table 1. Descriptive information of the participants

		Pilot Phase	Main Experiment
Number of participants		5	14
Gender	Female	2	6
	Male	3	8
Age	Min	27	24
	Max	34	42
	Median	31	30.5
	Mean	30.4	31.73
	Std. dev.	2.33	5.25
Work duration per day	Min	7	6
	Max	8.5	10
	Median	8	9
	Mean	7.9	8.92
	Std. dev.	0.49	1.16
Work experience (years)	Min	1.42	0.08
	Max	8.25	21.50
	Median	7	3
	Mean	5.45	4.18
	Std. dev.	2.94	4.63
Occupation	Engineer	1	8
	Academics	4	3
	Bank personnel	0	1
	IT specialist	0	1
	Technician	0	1
Work sector	Private	3	7
	Government	2	5
	Freelance	0	2

tasks in the study: (1) to label activity types later to be used to categorise them into still and moving actions (see Section 4.2.2), (2) to cluster user locations later to be used for finding location similarity and frequency (see Section 4.2.3 for details), which requires a parameter setup. The optimal parameter (a.k.a. threshold) was used when we investigated the statistical relationships between the features and the break availability using repeated measures correlation (see Table 5) as the results need to be reported based on factual data. The details of the collected data from the pilot participants can be seen in Table 2.

**3.4.4 Experiment.** After conducting the pilot phase, the main experiment was advertised to a larger pool of participants. The system setup and the procedure were exactly the same as in the pilot phase. In total, 46 individuals responded to the pre-experiment survey in full. Thirty-eight of them successfully installed the mobile application. Six participants dropped out of the experiment. Fourteen of the 32 participants had a response rate of 25% or higher, hence, the data obtained from them were included in the analyses. The descriptive information of the participants is given in Table 1. None of the participants were affiliated with the research group.

Table 2. Descriptive statistics of the collected data during the pilot phase and the main experiment

		Pilot Phase	Main Experiment
Number of ESM responses		131	397
Location labels	Number of data points	415	N/A
	Min data points per individual	51	N/A
	Max data points per individual	95	N/A
	Median data points per individual	59	N/A
	Mean data points per individual	66.43	N/A
	Std. dev. data points per individual	17.51	N/A
Activity data	Smartphones with Google Activity API support	5	9
	Smartphones without Google Activity API support	0	5
Ringer mode	ESM messages sent at sound mode	106	230
	ESM messages sent at vibrate mode	19	161
	ESM messages sent at silent mode	6	6
	Number of participants whose base state is sound	5	7
	Number of participants whose base state is vibrate	0	6
	Number of participants whose base state is silent	0	1

#### 4 DATA COLLECTED AND FEATURE EXTRACTION

We consider five main features for modelling and characterising user behaviour as suggested in previous works, namely answer time [30, 33, 39, 43, 50, 53], activity [22, 30, 33, 34, 36], location [8, 13, 30, 39, 43, 48], application usage [33, 34, 36, 40, 44], and ringer mode [8, 31, 33, 40, 50]. In the following sections, we discuss each factor in detail.

##### 4.1 Collected Data

In total, 921 ESM messages were sent to all users throughout the experiment (pilot phase is included). A total of 292 messages were sent in the preferred time slots of each user as stated in the pre-experiment questionnaire. The participants responded to 528 ESM messages, 131 of them were collected from the pilot participants as shown in Table 2. In the following sections, we report the whole data (pilot and main part). The details of the data collected in the pilot phase and the second part of the experiment can be seen in Table 2.

Ninety-seven of 292 messages sent in the preferred time slots were labelled as “can take a break” by the participants whereas 37 of them were labelled as “cannot take a break”, and 158 of the messages were not answered. Similarly, 208 of 629 messages sent in a random time were labelled as “can take a break”, 37 of them were labelled as “cannot take a break”, and 301 of the messages were not responded.

The response rates of each user are given in Figure 2 (top). The response rates are grouped into two categories corresponding to the cases in which the message is sent in preferred time slots and randomly, respectively. In the same figure (middle), the positive (i.e., “can take a break”) and negative (i.e., “cannot take a break”) response rates for preferred and random messages are displayed. The number of positive responses is higher than that of negative ones overall. This may be due to the fact that the participants responded to the messages when they were usually available and ignored the notification messages at other times.

The bottom figure shows the average duration of preferred times per day in minutes for each user. In the pre-experiment questionnaire, some users reported a wider range of time intervals whereas some did not report any, such as Users 15 and 16. The figures indicate that there is no significant difference between the notifications sent at preferred and at random times. There might be two possible explanations for this result. Although some

types of break might take place at around the same time over different days, such as lunch or praying breaks, others may shift or change due to the nature of the tasks that users are carrying out on a specific day. Another reason could be that some users did not report their daily schedule correctly when filling out the questionnaire.

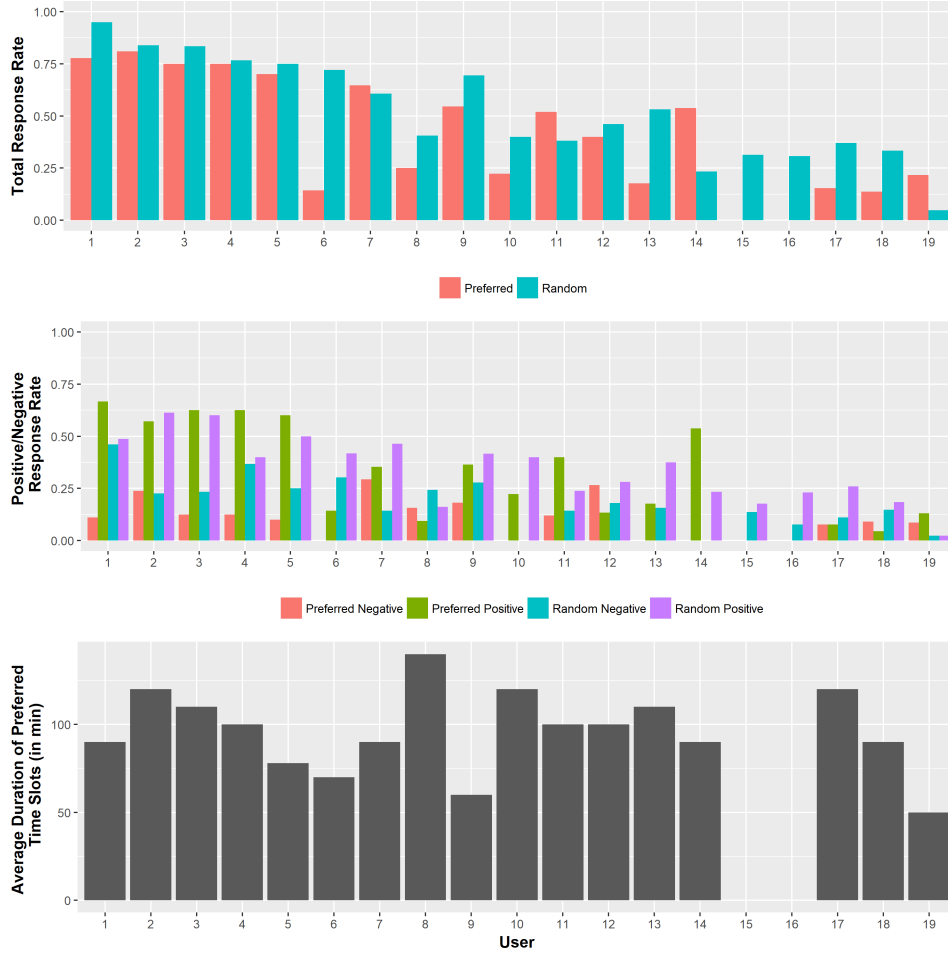


Fig. 2. *Top*: Total response rates of users to the messages sent in preferred and random time slots. *Middle*: Positive and negative response rates of users to the messages sent in preferred and random time slots. *Bottom*: Average duration of preferred time slots of users during work hours per day (in minutes).

## 4.2 Training and Parameter Setting Phase

In this section, activity classification and modelling user locations based on the user labelling in the pilot phase is explained. Table 3 shows the abbreviations that are used throughout this paper.

**4.2.1 Modelling User Activity.** User activity is defined as whether a user is *moving* or *still* when notifications are sent. To identify user activity, *still periods*, i.e., the periods during which the mobile phone is detected as being

Table 3. Abbreviations

Abbreviation	Description
$N$	Notifications set
$SP$	Still periods set
$L$	Locations set
$t_{start_j}$	Start time of $j^{th}$ still time period ( $\forall j \in SP$ )
$t_{end_j}$	End time of $j^{th}$ still time period ( $\forall j \in SP$ )
$t_{submission_n}$	Submission time of response to notification $n$ ( $\forall n \in N$ )
$t_{workStart}$	Work start time
$t_{workEnd}$	Work end time
$activity_n$	Activity when notification $n$ is sent ( $\forall n \in N$ )
$TSA_n$	Time spent in action when the notification $n$ is responded ( $\forall n \in N$ )
$LS_n$	Similarity between the location where the notification $n$ is responded and the base location ( $\forall n \in N$ )
$LF$	Visit frequency of the location where the notification $n$ is responded ( $\forall n \in N$ )
$AP$	Wi-Fi access points set
$AU_{tn}$	Application usage in $t$ minutes prior to the notification $n$ ( $\forall n \in N, \forall t \in \{5, 10, 15, 30, 45, 60\}$ )
$T_b$	Probability density of break availability category $b$ ( $\forall b \in \{1, 2, 3, 4\}$ )
$RM$	Ringer mode
$RMC$	Ringer mode change
$CI$	Credible interval
$f_i(a)$	RSSI of access point $a$ overheard at location $i$ ( $\forall a \in AP, \forall i \in L$ )

still for more than five minutes were extracted. Every *still period*  $j$  has a start  $t_{start_j}$  and end time  $t_{end_j}$ . The start time of a *still period* is the moment when the activity is read as *still* with an accuracy of 100% by the Google Activity library. The end time of a *still period* is the time when the activity type returns a value other than *still*. A *still period set* ( $SP$ ) includes all *still period* data.

Equation 1 shows how the activity was labelled:  $activity_n$  denotes the activity of the user when notification  $n$  is received. If the user's submission time of response  $t_{submission_n}$  to notification  $n$  is between any start time and end time of still periods, then the activity is labelled as “*still*”. Otherwise, it is labelled as “*moving*”.

$$activity_n = \begin{cases} still, & \text{if } t_{start_j} \leq t_{submission_n} \leq t_{end_j}, j \in SP, n \in N. \\ moving, & \text{otherwise} \end{cases} \quad (1)$$

**4.2.2 Predicting Activity from Accelerometer Data.** The smartphones of five participants in the main experiment did not show significant motion sensor activity, hence, their activity types could not be directly recorded in the database. We therefore fitted a model to classify their activity types based on their raw accelerometer data. Ustev et al. [51] used three features for classifying activity data: standard deviation of magnitude of accelerometer data, variance of the magnitude of accelerometer data, and mean  $z$  value of accelerometer data using kNN. For our case, the mean, standard deviation, and variance were computed upon the samples of the raw accelerometer data which is collected every three seconds for one minute in five-minute intervals, i.e., the window size of the subsamples is one-minute for every five minutes. We also used kNN with a value of  $k = 51$  for which we obtained the best results (the value of  $k$  was set using cross-validation). We also used a support vector machine (SVM) classifier with radial basis kernel for predicting activity (the kernel width of the SVM model was obtained through cross-validation). The classifiers were trained and validated on (disjoint) data obtained in the pilot study. The out-of-sample accuracy of the kNN model is 81.40%, whereas the accuracy of the SVM model is slightly

better: 82.43%. Hence, the classification of activity was carried out using the SVM model for the five participants in the main experiment. After the classification of the activity, the ongoing activity when notifications were sent was classified using Equation 1.

**4.2.3 Modelling User Location.** GPS sensor data is unreliable for indoor localisation [55]. Since the participants of the study spend most of their time indoor (in their offices), we first need to differentiate indoor locations where a user has been. This has been done using an existing RSSI fingerprinting based on Wi-Fi access points (APs) in the literature.

Briefly, this method computes the similarity between two locations based on the RSSI of Wi-Fi APs recorded at two locations. The similarity computation is given in Equation 2. Wi-Fi APs recorded at locations  $l_1$  and  $l_2$  are denoted as  $AP_1$  and  $AP_2$ ,  $AP = AP_1 \cup AP_2$ .  $f_i(a)$  denotes the RSSI of AP,  $a \in AP$ , recorded at location  $l_i$ .

$$S = \frac{1}{|AP|} \sum_{a \in AP} \frac{\min(f_1(a), f_2(a))}{\max(f_1(a), f_2(a))} \quad (2)$$

For each user in the experiment, an  $M \times M$  similarity matrix was computed where  $M$  is the number of *still periods* of that user. At first, each *still period* was considered as a single location. However, the user may be in the same location at different times of the day. In order to identify the locations of a user in still periods, hierarchical clustering was used since it enables to adopt a threshold value for identifying clusters. Different threshold values were attempted for identifying clusters. The accuracy of the method was computed by comparing labels extracted by means of the clustering algorithm with those assigned by the participants in the pilot study. The total number of the labelled locations was 415. The threshold values of .05, .10, .15, .20, .30, .40 and .50 were used for clustering. The best accuracy (71.33%) was obtained when the threshold value was set to .15. Hence, if the similarity of two still periods is higher than .15, then they are considered as the same location. The accuracy dropped significantly when the threshold was greater than .20.

It is worth mentioning that the main reason for not achieving a higher accuracy value may be related to errors in the labelling process during the pilot study. In fact, the pilot study participants filled the location form provided to them, however, they may have forgotten the locations where they had been for a very short period of time, so that they may have specified wrong location in the form for that time period. Since the data obtained from the participants gave the most accurate result when the threshold was set to .15, the two still time periods, which had a similarity higher than .15, were considered as the same location. However, given the issues related to clustering, we consider all the values of the threshold of the clustering algorithm, as discussed in the next section.

### 4.3 Feature Space

We defined the feature space before modelling the data. Since we were not able to detect the exact locations of users, we introduced new features for representing locations in terms of duration, similarity, and frequency using the location clusters. The following features are calculated for all users who participated in the experiment.

**4.3.1 Time Spent in Action.** After defining user locations inferentially by clustering based on the similarities, the locations where users answered ESM messages were extracted. Then, we defined the model features based on the locations and the activity of the users. The first one is the time spent in action (TSA). It quantifies the duration (in minutes) of the action when the ESM message is answered. The action is considered as “*still*” or “*moving*” based on the still time periods calculated. The duration calculation is given in Equation 3.

$$TSA_n = \begin{cases} t_{submission_n} - t_{start_j}, & \text{if } activity_n = \text{“still”} \\ t_{submission_n} - t_{end_{j-1}}, & \text{if } activity_n = \text{“moving”} \end{cases}, j \in SP, n \in N \quad (3)$$

**4.3.2 Location Similarity.** Since the experiment was conducted during the working hours of the participants, it is assumed that the location each participant spent most of her time is her workplace and named as her “*base location*”. It is labelled as the place where the sum of the *TSAs* for all the notifications is the highest for still periods. Location similarity is defined as the similarity between the location where the ESM message arrived and the user’s base location. If the user is in the base location (i.e., in her workplace) when the message has arrived, then the location similarity is set to one. On the other hand, if the user is in a totally different place (e.g., in a cafeteria for lunch), then the location similarity is set to zero. The location similarity is calculated using Equation 2.

**4.3.3 Location Frequency.** Location frequency is defined as the total number of visits to the location  $i$  throughout the experiment over the total number of visits to all locations throughout the experiment. For example, while the frequency of visits to the base location is expected to be the highest, the frequency of the locations where participants have their lunch is expected to be significantly lower, such as 10 visits (over the total number of visits) if the individual had lunch every day at the same place during the experiment.

**4.3.4 Ringer Mode Status.** In the pre-experiment questionnaire, we collected the ringer mode in which participants were keeping their mobile phones during their office hours and breaks. Figure 3 shows the responses of the participants. It can be seen that users keep their phones in different modes in different situations. In praying, health breaks, and in meetings, the participants set the ringer mode of their mobile phones either to silent or vibrate but not to sound mode.

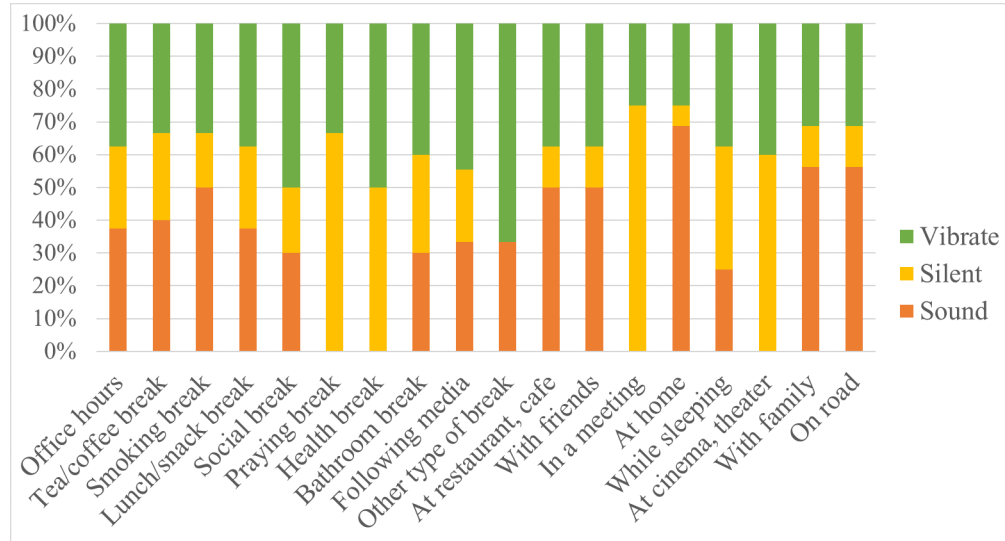


Fig. 3. Ringer modes kept in office breaks and different locations

Out of the 528 ESM messages, 337 messages arrived when the mobile phone was set to sound mode, 180 arrived in vibrate mode and 11 in silent mode. Since the number of data points for the silent mode is significantly lower than the others, silent and vibrate mode data are combined in a *silent-or-vibrate* mode category to be used in the modelling phase.

We calculated the total duration of the adoption of each ringer mode over the experiment for each user. The results showed that 14 users kept their mobile devices in the sound mode more than 50% of the total duration and



the remaining five of them kept them in vibrate mode. As a result, we used two variables linked to the ringer mode data. In the first case, we just used “*ringer mode*” whether it is in sound mode or silent-or-vibrate mode. In the second case, we defined a variable named “*ringer mode change*” showing whether the ringer mode has changed compared to the *base state*, which is determined as the state in which a user mostly keeps her mobile phone during work hours. For this variable, we consider the silent mode as it is (i.e., we did not merge with vibrate mode for ringer mode change detection). In our dataset, the number of silent and vibrate modes were lower compared to sound mode. For this reason, we merged them for developing our model. This was the only data pre-processing we performed. We hypothesise that the ringer state change occurs when a significant state change in the user’s daily routine is about to happen.

**4.3.5 Application Usage.** We define a phone usage session as the time spent between screen-on and off [3, 28]. We extracted the information concerning the phone usage of each user in terms of start time, end time, duration of the session (end time - start time) and inter-event times. We also merged the sessions where inter-event time is less than five seconds as in previous studies [3, 28]. We investigated usage sessions with a duration of 5, 10, 15, 30, 45, and 60 minutes before message delivery. We collected application types and names in phone usage sessions; however, we only included phone usage sessions as application usage variables in this study.

**4.3.6 Break Types and Break Availability.** We selected break availability of users as the target variable for prediction. Recall that we also asked participants to report the ongoing task when notifications arrived. The percentages of the number of responses with respect to break availability and break type categories are given in Table 4. The participants stated that they cannot take a break, or they can take up to 5-minutes break when they are working (51.75% and 30.26% respectively). Similarly, when they are in a meeting, they stated that they are not available (89.19%). Note that when participants are in a lunch/snack break, they mostly marked their availability selecting the “more than 15 minutes” option (71.43%). We can also see that the duration of social, bathroom or tea/coffee breaks is approximately 5-15 minutes. We performed Fisher’s Exact Test on break type and break availability variables since sample sizes are small and the data is unequally distributed among the cells of the contingency table, and the result shows that break availability and break types are significantly related ( $p < .001$ ). Since break availability gives a strong indication of the break type, we focused on predicting break availability.

## 5 PROPOSED APPROACH

Our proposed model is composed of two stages: in the first stage, we focus on modelling time, then, in the second stage, we model break availability with time and other variables described in the previous section. The flowchart of the model is given in Figure 4.

Before continuing with modelling phase, we first split our dataset into training and test sets using repeated random sub-sampling validation (i.e., repeated hold-out). The training and test sets are constructed with the proportions of 30-70, 40-60, 50-50, 60-40, and 70-30 with stratified sampling, which enables to balance class proportions in each set. Twenty random sets per configuration were used.

### 5.1 Modelling Response Submission Time

First, user response submission time is converted into a numeric variable corresponding to a decimal representation of the number of hours in a day from 0 to 23.99. In other words, a decimal representation is used for representing fractions of hours. For example, if a user responded to a notification at 12:30:46, the response submission time in hours is equal to 12.51. Then, we considered the self-reported duration of the break availability obtained from ESM because there are certain time intervals during which users prefer to take macro or micro breaks. We used kernel density estimation (KDE) to estimate the probability density of the break duration versus response

Table 4. Percentages of the number of responses with respect to break availability over their reported task/break type during the notifications considering the entire dataset.

Current Task/Break vs. Break Availability	Cannot take a break	Less than 5-min	5-15 min	More than 15-min
Working	51.75%	30.26%	17.54%	.44%
Tea/coffee break	2.38%	38.10%	45.24%	14.29%
Other	32.43%	13.51%	10.81%	43.24%
Praying break	10.00%	10.00%	40.00%	40.00%
Following the media	.00%	26.32%	57.89%	15.79%
Bathroom break	.00%	44.44%	50.00%	5.56%
Health break	25.00%	.00%	25.00%	50.00%
Smoking break	28.57%	42.86%	28.57%	.00%
Social break	5.88%	20.59%	50.00%	23.53%
In a meeting	89.19%	5.41%	5.41%	.00%
Lunch/snack	3.90%	11.69%	12.99%	71.43%
On road	33.33%	6.67%	33.33%	26.67%

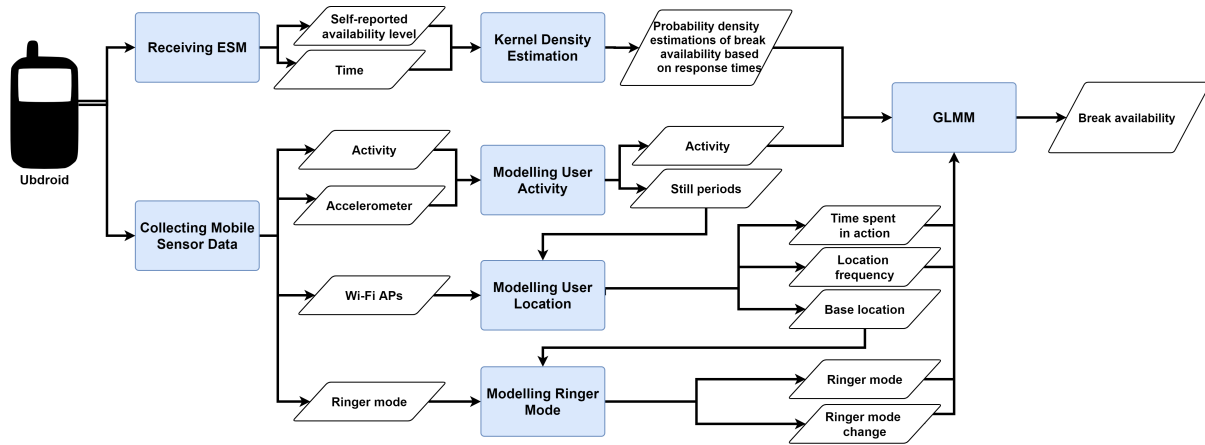


Fig. 4. Flowchart of the hybrid model proposed

submission time. KDE uses density estimation to find the optimal cut-points, automatically adapts sub-intervals to the data and finds the number of bins by the cross-validated log-likelihood [45]. Hence, we did not include time as a discretized variable as in previous studies.

As an example, in Figure 5 it is possible to observe a user's KDE plot. The figure shows that the user has mostly macro breaks (more than 15 minutes) at 12:00. Similarly, the tendency to have micro breaks (5 minutes or less) is higher between 14:00-16:00. The user does not have breaks or did not respond to any message before 10:00 or after 17:00 so that no information is present for those periods.

We used the *kde* function of the *ks* package in R Software. We used Gaussian kernel, where the bandwidth is selected with the plug-in bandwidth selector (a reliable method for bandwidth selection [23]). Since break availability times and durations vary with each individual, we fitted a 2-D KDE on each user's training dataset

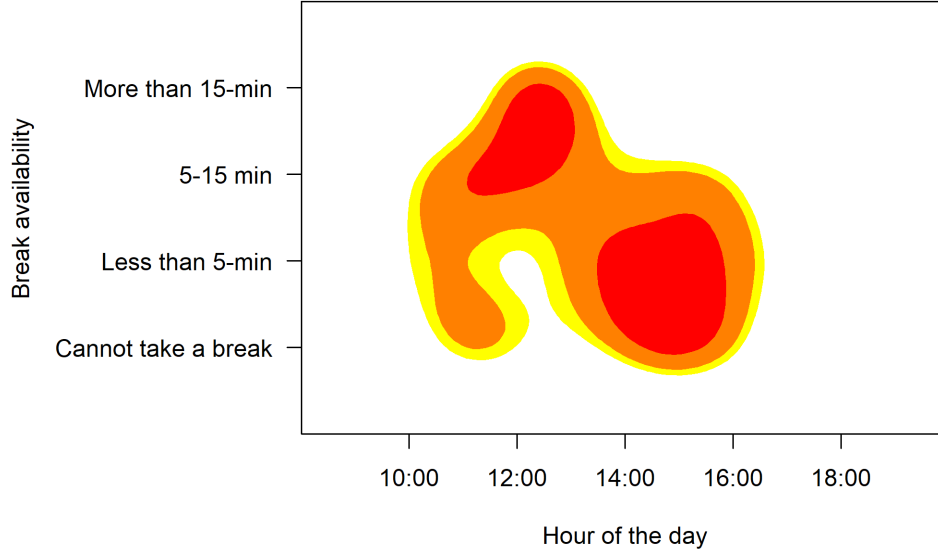


Fig. 5. Density plot of a user's break availability levels based on hours of the day.

comprising self-reported break availability responses with their corresponding timestamps. Then, we defined four new variables associated to break availability predictions called  $T1$ ,  $T2$ ,  $T3$  and  $T4$  corresponding to “Cannot take a break”, “Less than 5 minutes”, “Between 5-15 minutes”, and “More than 15 minutes”. Finally, for each ESM notification time in the training dataset, the notification-time-conditional probability density function predictions of  $T1$ ,  $T2$ ,  $T3$  and  $T4$  were obtained for the test set, and included in the mixed model analysis later. In other words, we build the 2-D density Parzen window estimator using the training set. Then we apply it to make predictions on the test set, conditional on the particular times found in the test set.

## 5.2 Modelling and Predicting Break Availability

After modelling the response submission time in the first phase, we then applied a method for predicting break availability with the time variables generated from KDE, and other variables stated in Section 4.3. We selected Generalised Linear Mixed Models (GLMM) as our main model. We discuss the reasons beyond this choice below.

**5.2.1 Why GLMM?** Generalised linear models (GLM) extends linear models by handling response variables with non-normal distribution [20, 29]. Generalised linear mixed models (GLMM) incorporate random effects (individuals, populations, species or vials with a large number of levels [20]) in GLMs. In its simplest form, a GLMM can be written as in Equation 4 where  $y$  is a  $N \times 1$  column vector, the target variable ( $N$  is the total number of data points in the dataset);  $x$  is the matrix of fixed predictors with the dimension of  $N \times p$  ( $p$  is the number of predictors);  $\beta$  is a  $p \times 1$  column vector of the fixed-effects coefficients;  $z$  is the design matrix with the dimension of  $N \times q$  corresponding to the  $q$  random predictors (accounting for the random complement to the fixed  $x$ );  $b$  is a  $q \times 1$  vector of the random effects (the random complement to the fixed  $\beta$ ); and  $\beta_0$  is a  $N \times 1$  residuals vector. For our study,  $y$  is the break availability level as the target variable,  $x$  represents the predictors, namely the activity, location similarity, time spent in action etc., and  $z$  indicates the participants as the random predictors.

$$y = \beta_0 + x\beta + zb \quad (4)$$

GLMMs can deal with numerous response distributions and repeated-measures observations. Specifically, the use of GLMMs is appropriate when there are many levels (e.g., individuals, species), few data on each level, or when the number of samples for each level is not the same. Since GLMM incorporates random effects (i.e., individuals), it is possible to obtain an individual prediction from GLMM. It simply fits separate linear regression lines for each random effect included. A Bayesian framework using Markov Chain Monte Carlo (MCMC) methods is a convenient way to fit a GLMM specifically for modelling non-Gaussian data because integrating over the random effects is compelling [20].

Since our dataset consisted of multiple responses of participants, the data points could not be considered as independent. Besides, the number of responses is not the same for each participant. Moreover, the response variable is ordinal, which means that its distribution is not Gaussian. For these reasons, the assumptions of approaches such as ANOVA have been violated. In this case, it is suggested to use approaches such as generalised linear mixed models (i.e., hierarchical or multilevel modelling) [2].

There is a variety of methods for Bayesian model selection including deviance information criterion (DIC), widely applicable information criterion (WAIC), leave-one-out cross-validation or  $K$ -fold cross-validation [41]. Our performance evaluation was performed with the R package named *MCMCglmm* [20]. We used cross-validation for model selection as the package does not return the pointwise log-likelihood directly for the WAIC calculation. Hence, we used 9-fold cross-validation. We considered all users' data in each fold (i.e., we stratified users' responses to each fold). In the case of 10-fold, we did not have a user's data in the test dataset. The convergence of the Markov chains is checked using the Gelman-Rubin diagnostic criterion for which the value of 1.002 (or below) indicates the convergence [17]. Finally, autocorrelation between consecutive iterations in the chain should be less than .10, which indicates the chain has mixed well [20]. After several trials where we looked for MCMC convergence and consistency among runs using a random search, a burn-in value of 8000, a thinning interval of 50, and the number of MCMC iterations of 50000 were selected. We used weakly informative priors for an ordinal response variable. We set the variance component to 1 as in [20]. Finally, Gelman-Rubin diagnostics were close to 1, and autocorrelation plots were stationary, which means that the autocorrelation between consecutive samples in the chain is low enough for the convergence. One of the autocorrelation plots is given in Figure 6 for the variable *location frequency*.

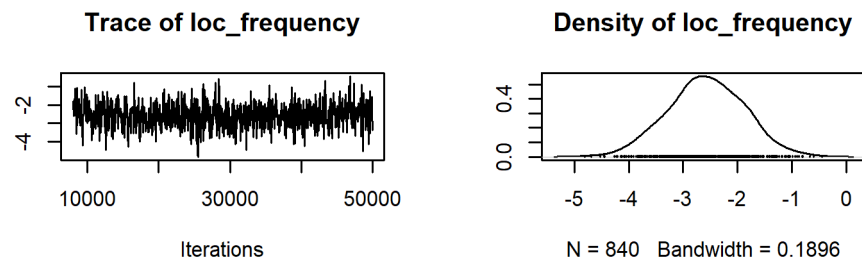


Fig. 6. Posterior distribution of a model parameter. *Left*: Time series of a parameter in the model as MCMC iterates (note that the range of x-values is from 8000 to 50000). *Right*: probability density estimates of the parameter. The peak of the distribution (the posterior mode) is the most likely value.

**5.2.2 Comparison of GLMM with Other Methods and Individual Models.** Because our dataset consists of personal data, the level of the variables used in the analyses may be quite different from user to user. For example, one may take regular breaks, hence the number of his/her activity switches between *still* and *moving* categories might be higher than another person who mostly spends his/her time at his/her desk. In such situations, previous studies [53, 61] showed deficiencies on transferring general models to individual-level ones. For this reason, individual models (i.e., trained only on one user’s data) usually lead to more accurate results. On the other hand, individual models may suffer from the lack of sufficient individual data for training in the beginning, which is named as the “cold-start problem” [61].

As a comparator, we used the random forest classifier. For training random forest models, we used both general (with all participants) and individual data. We generated user-specific models with random forest classifier on each participant’s data separately. For all random forest classifiers (general and individual) in the study, the number of trees was set to 500 since it gave the most accurate results among several values (50, 100, 150, 250, 500, 750, 1000) when cross-validated on a dataset reserved at the beginning of the experiment and not used later to form test sets. We also calculated the baseline performance with the majority classifier that always predicts the class with the highest number of data points.

## 6 DATA ANALYSIS AND RESULTS

In this section, we report the results of our hybrid model for inferring break availability of office workers using mobile sensors by considering the cold start problem, the variety in the number and characteristics of the responses, and repeated-measures in the data. We also investigate how the model is comparable to individual and population-level models.

### 6.1 Model and Feature Set Selection using GLMM

We first analyse the repeated-measures correlations (*rmcorr*) [2] between the predictor variables and break availability using the data obtained from all the users both in the pilot study and the main experiment. *rmcorr* shows the linear association between the variables. GLMM is inherently a linear model, so we consider the correlation results for feature set selection instead of variable selection with Gini or other metrics. Table 5 shows the correlation results, which provides insights about which variables are related to break availability.

As it is possible to observe in Table 5, *location similarity (LS)*, *time spent in action (TSA)*, and *location frequency (LF)* are the significantly related variables to the break availability at the .001 level. We also included *5-min application usage (AU<sub>5</sub>)* in the models even it is not significantly related to break availability (although having the highest  $r_{rm}$  among all application usage variables) in order to ensure that we have at least one representative variable for application use. Hence, we include a combination of those variables in the GLMM in addition to the response submission time variables. However, in order to prevent multicollinearity issues, variable pairs that have a correlation coefficient with higher than .70 are not included. Location frequency (*LF*) and location similarity (*LS*) are highly correlated ( $r_{rm} = .870, p < .001$ ). We select *LF* is selected since it has a higher relation to break availability ( $r_{rm} = -.152, p < .001$ ) than *LS* ( $r_{rm} = -.140, p < .001$ ).

MCMC GLMM fits are built iteratively with the KDE predictions of each break availability level (duration) at a given time  $T_1$ ,  $T_2$ ,  $T_3$  and  $T_4$  together with *TSA*, *LF*, *activity (A)*, *ringer mode (RM)*, *ringer mode change (RMC)*, and *5-min application usage (AU<sub>5</sub>)* parameters, which are given to the models as fixed components ( $x$  in Equation 4). The random component of the models ( $bz$  part in Equation 4) refers to the users (19 participants). The response variable ( $y$  in Equation 4) is the break availability with four levels.

Table 6 summarises the models fitted for predicting break availability with several combinations of the covariates. The simplest model (Model 1) consists of  $T_1$ ,  $T_2$ ,  $T_3$ ,  $T_4$ , *TSA* and *LF* as the fixed components. Then, in

Table 5. Repeated measures correlation (*rmcorr*) coefficients for the predictor and the target variables

		<i>LS</i>	<i>LF</i>	<i>TSA</i>	<i>AU<sub>5</sub></i>	<i>AU<sub>10</sub></i>	<i>AU<sub>15</sub></i>	<i>AU<sub>30</sub></i>	<i>AU<sub>45</sub></i>	<i>AU<sub>60</sub></i>
<i>LS</i>	<i>r<sub>rm</sub></i>	1	.870	.119	-.120	-.129	-.088	.093	.067	.052
	<i>p</i>		<.001	.007	.006	.004	.05	.04	.13	.24
<i>LF</i>	<i>r<sub>rm</sub></i>		1	.114	-.105	-.104	-.079	.065	.048	.028
	<i>p</i>			.01	.02	.02	.08	.14	.28	.53
<i>TSA</i>	<i>r<sub>rm</sub></i>			1	-.059	-.065	-.079	-.002	.013	.010
	<i>p</i>				.18	.15	.07	.97	.77	.82
<i>AU<sub>5</sub></i>	<i>r<sub>rm</sub></i>				1	.849	.739	.086	.042	.043
	<i>p</i>					<.001	<.001	.05	.35	.34
<i>AU<sub>10</sub></i>	<i>r<sub>rm</sub></i>					1	.893	.107	.070	.065
	<i>p</i>						<.001	.02	.11	.14
<i>AU<sub>15</sub></i>	<i>r<sub>rm</sub></i>						1	.134	.092	.083
	<i>p</i>							.002	.04	.06
<i>AU<sub>30</sub></i>	<i>r<sub>rm</sub></i>							1	.930	.854
	<i>p</i>								<.001	<.001
<i>AU<sub>45</sub></i>	<i>r<sub>rm</sub></i>								1	.957
	<i>p</i>									<.001
<i>Break</i>	<i>r<sub>rm</sub></i>	-.140	-.152	-.226	.071	.066	.019	-.022	-.065	-.054
<i>Availability</i>	<i>p</i>	.001	<.001	<.001	.109	.139	.667	.627	.144	.226

Model 2, *AU<sub>5</sub>* is added as another fixed component. In Model 3, *A* is added instead of *AU<sub>5</sub>*. In Models 4 and 5, *RM* and *RMC* are added as another fixed component separately and all were included in Model 6.

Table 6. Models fit upon different covariates for predicting break availability

Model No	Covariates
1	<i>T1 + T2 + T3 + T4 + TSA + LF</i>
2	<i>T1 + T2 + T3 + T4 + TSA + LF + AU<sub>5</sub></i>
3	<i>T1 + T2 + T3 + T4 + TSA + LF + A</i>
4	<i>T1 + T2 + T3 + T4 + TSA + LF + A + RM</i>
5	<i>T1 + T2 + T3 + T4 + TSA + LF + A + RMC</i>
6	<i>T1 + T2 + T3 + T4 + TSA + LF + A + RM + RMC</i>

We used accuracy and macro F1 scores as performance metrics. In macro F1, all classes are considered as equally important, which is the case in our domain. But, there might other domains where one class might be considered more important, then, weighted F1 score might be a better choice. For instance, for prediction of break availability of doctors, finding non-interruptible moments might be more important than correctly estimating the duration of the availability.

All the models are evaluated on five datasets, four of which are subsamples of the original dataset (Table 7). The main reason for running the models on different subsamples is to avoid spurious results affected by the variations in the number of responses. Hence, in each subsample, we eliminated the data points of the participants with the highest and lowest response rates incrementally. To be more specific, the first sample is the full original dataset

consisting of all the responses of users ( $N=528$  with 19 users). In the second dataset, two users are removed from the first dataset. These users are those with the highest and lowest number of ESM responses, and so on. As a result, the four sampled datasets include  $N=528$  with 19 users (the original dataset),  $N=469$  with 17 users (Dataset 2),  $N=410$  with 15 users (Dataset 3) and  $N=350$  with 13 users (Dataset 4) respectively. Finally, the fifth dataset consists of the users who have Google Activity API in their mobile phones, which means that five users whom we predicted their activities from the first dataset ( $N=369$  with 14 users).

Table 7. Accuracy values and macro F1 scores of the GLMM used to predict the break availability for different models on different datasets.

Model No	Accuracy				
	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
1	44.21% $\pm$ 7.79%	45.31% $\pm$ 7.90%	46.20% $\pm$ 7.89%	46.55% $\pm$ 6.84%	47.36% $\pm$ 6.05%
2	44.13% $\pm$ 7.71%	45.00% $\pm$ 8.13%	45.98% $\pm$ 8.13%	46.52% $\pm$ 7.05%	47.82% $\pm$ 6.59%
3	44.48% $\pm$ 9.50%	45.85% $\pm$ 9.50%	46.15% $\pm$ 9.94%	46.13% $\pm$ 9.19%	47.19% $\pm$ 6.26%
4	44.92% $\pm$ 8.25%	45.96% $\pm$ 8.63%	47.50% $\pm$ 9.42%	47.10% $\pm$ 8.93%	47.63% $\pm$ 6.15%
5	45.01% $\pm$ 8.72%	46.36% $\pm$ 9.34%	47.48% $\pm$ 8.63%	<b>47.88% <math>\pm</math> 8.47%</b>	<b>48.16% <math>\pm</math> 5.85%</b>
6	<b>45.23% <math>\pm</math> 8.62%</b>	<b>46.57% <math>\pm</math> 9.43%</b>	<b>47.68% <math>\pm</math> 8.69%</b>	47.58% $\pm$ 8.16%	47.90% $\pm$ 5.74%
Model No	Macro F1 Scores				
	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
1	52.58% $\pm$ 3.49%	55.10% $\pm$ 4.92%	54.07% $\pm$ 5.56%	51.76% $\pm$ 6.60%	47.59% $\pm$ 7.00%
2	51.29% $\pm$ 4.13%	54.21% $\pm$ 5.20%	53.21% $\pm$ 4.84%	51.45% $\pm$ 5.64%	46.98% $\pm$ 7.51%
3	53.72% $\pm$ 4.72%	57.64% $\pm$ 4.62%	54.26% $\pm$ 5.35%	51.86% $\pm$ 8.17%	47.59% $\pm$ 5.51%
4	53.52% $\pm$ 4.90%	56.10% $\pm$ 5.22%	54.46% $\pm$ 5.23%	52.58% $\pm$ 5.73%	46.55% $\pm$ 6.19%
5	51.62% $\pm$ 8.89%	58.63% $\pm$ 6.65%	<b>57.88% <math>\pm</math> 6.19%</b>	<b>56.02% <math>\pm</math> 6.43%</b>	58.29% $\pm$ 5.91%
6	<b>54.69% <math>\pm</math> 4.82%</b>	<b>59.07% <math>\pm</math> 5.10%</b>	57.75% $\pm$ 5.00%	55.93% $\pm$ 6.32%	<b>58.33% <math>\pm</math> 6.30%</b>

We fitted the same models using different threshold values used for clustering users' locations in order to investigate the effects of threshold selection. Note that although in Section 4.2.3, we found that the optimal threshold value for location labelling was .15 based on the pilot phase data, we did not use it here since we included all the participants' data. Instead, we used all possible threshold values during modelling. The ideal case would be to record all the locations precisely and look for the relationships. However, in the wild, it is not the case. Because of that, in our experiment, the modelling phase did not use any predetermined thresholds and the experiments were repeated with a range of possible values for these features. As a result, the values of  $LF$  and  $TSA$  variables in five datasets change according to seven threshold values (.05, .10, .15, .20, .30, .40, and .50), which resulted in 35 different runs for each model. Since we have six models and used 9-fold cross-validation, this resulted in a total of  $35 \times 6 \times 9 = 1890$  runs. Table 7 summarises these runs with the mean and standard deviation of the accuracy values and macro F1 scores for each model and dataset.

In order to test whether there is a statistical difference between the models, we applied Friedman Test on the accuracy and macro F1 values of the models. The results show that six models are significantly different ( $\chi^2(5) = 19.857, p < .05$  for accuracy values,  $\chi^2(5) = 21.114, p < .001$  for F1 scores). Binary comparisons of models using Wilcoxon signed rank tests with the Bonferroni p-value adjustment revealed that Model 1 and 2 are not significantly different based on their accuracy values ( $V = 14036, p = .610$ ); however, they are different based on F1 scores ( $V = 16118, p < .05$ ). For this reason, we selected Model 1, since it has higher F1 scores than Model 2 has. Similarly, Model 1 and Model 3 are not significantly different based on their accuracy values ( $V = 15430, p = .660$ ).



but they are different based on F1 scores, which means that adding  $A$  to the model variables leads to better results. Model 3 and 4 are statistically different based on their accuracy values ( $V = 9770.5, p < .001$ ), but their F1 scores are only different at .10 level of significance ( $V = 14142, p < .10$ ). This means that  $RM$  is also effective on predicting break availability. Models 4 and 5 are not significantly different based on the accuracy values ( $V = 20962, p = .34$ ). However, Model 5's F1 score is significantly higher than Model 4's ( $V = 2486, p < .001$ ). Finally, Model 5 and 6 are not statistically different from each other ( $V = 17000, p = .772$  for accuracy values,  $V = 17432, p = .797$  for F1 scores); yet, we selected Model 6 for the remainder of our analysis, since it has the highest accuracy values and F1 scores among all six models over all datasets.

## 6.2 Prediction Results

Table 8 shows the posterior distributions of each parameter in Model 6 with their posterior means, 95% credible intervals (2.5 and 97.5 percentiles of the posterior distribution), and the significance values ( $p$ ).  $T1$  and  $T4$  have a significant effect on break availability prediction. It means that the KDE of break availability for “Cannot take a break” and “More than 15 minutes” are more effective in predicting the output variable. The inverse relationship between  $T1$  and the output shows that an increase in the likelihood for the estimation of break availability  $BA$  level 1 (i.e., “Cannot take a break”) is a sign of a decrease in the break availability. The positive relationship between  $T4$  and break availability similarly shows an increase in the likelihood for the estimation of  $BA$  level 4 (i.e., “More than 15 minutes”), which suggests an increase in the break availability. The  $T2$  and  $T3$  variables, which correspond to  $BA$  level 2 and 3 respectively, are not significant for predicting break availability levels.

Table 8. Posterior means, 95% credible intervals (CI) and  $p$  values of parameters for Model 6

Parameters	Posterior mean	95% CI	$p$
(Intercept)	2.01	(1.17, 2.92)	<.001
$T1$	-1.48	(-1.93, -.97)	<.001
$T2$	-.19	(-.61, .26)	.41
$T3$	-.23	(-.72, .30)	.34
$T4$	1.14	(.64, 1.58)	<.001
$A[MOVING]$	.22	(.09, .37)	.002
$RM[SILENT]$	-.15	(-.32, .00)	.07
$RMC[CHANGE]$	-.19	(-.40, .05)	.10
$LF$	-.54	(-1.11, .04)	.07
$TSA$	-.11	(-.22, -.01)	.03

The model outputs show that there is a negative relation between break availability and  $LF$  (posterior mean=-.54; 95% CI [-1.11, .04]), and significant negative relation between break availability and  $TSA$  (posterior mean=-.11; 95% CI [-.22, -.01]). The results imply that as the location frequency increases, the duration of the breaks decreases and vice versa. Similarly, as the time spent in action increases, the duration of the breaks decreases or vice versa. Based on the values of the magnitude, we can say that  $LF$  has a higher effect than  $TSA$ .

In addition to the location parameters, we consider the impact of the activity on the break availability. The model considers “still” category as the basis and calculates the posterior mean of the “moving” category as .22 with 95% CI= [.09, .37]. It means that users tend to take longer breaks when they are moving. Similarly, there is an inverse relationship with ringer mode change and break availability. The relationship with ringer mode is also found as positively correlated. Users have a longer break when their phones are in sound mode. The results obtained from the pre-experiment questionnaire support these findings since users mostly change their ringer

modes when they attend a meeting or when they do not want to be notified in other words in the situations when they cannot take a break. Figure 7 shows that the ringer state change (specifically vibrate→silent, sound→silent, sound→vibrate) occurs mostly when users cannot take a break or can take a break with a duration of less than 5 minutes.

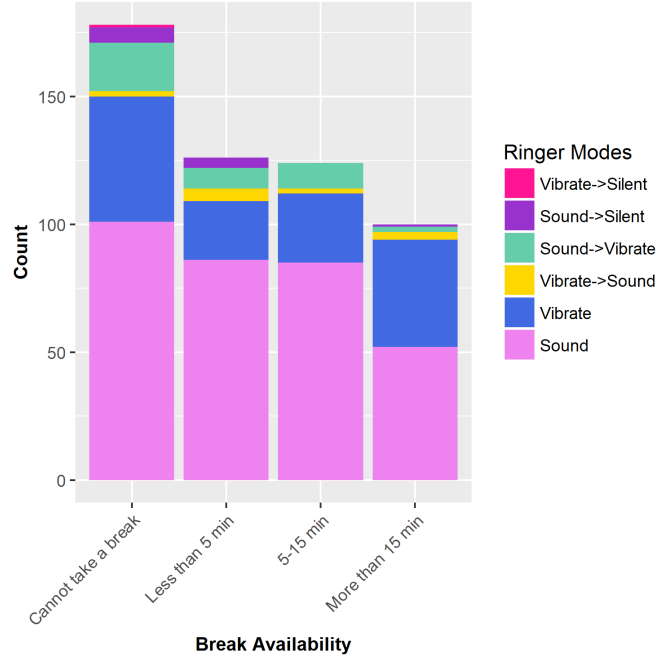


Fig. 7. Number of counts of each break availability category grouped by ringer mode change. The ringer mode on the left hand side of the arrow shows the base ringer mode (ringer mode which user keeps his/her mobile phone in general), whereas the ringer mode on the right shows the current ringer mode when ESM questionnaire is answered. The ringer modes without arrows show the unchanged ringer modes.

### 6.3 Comparison Results

The variables used in the evaluation of the random forest models (population and individual level) were the same as those reported in the analysis of the GLMM results discussed above:  $T1$ ,  $T2$ ,  $T3$ ,  $T4$ ,  $A$ ,  $RM$ ,  $RMC$ ,  $LF$  and  $TSA$ . We report the accuracy and the macro F1 scores of the models as the performance metrics by averaging 20 runs. We also performed 1-person-out cross-validation (CV) on the dataset in order to re-validate the performance of the classifiers. The biases caused by individuals that were uniformly distributed in other splits were avoided by performing 1-person-out CV. In Table 9, the mean and standard deviations of the accuracy values and the F1 scores obtained from six different classifiers for predicting break availability are presented.

As can be seen from the table, GLMM predicts break availability levels better than all other classifiers for all training percentages. As expected, the accuracy obtained using 1-person-out CV is lower since the individuals' own data are highly informative and we do not use them in 1-person-out CV. In order to compare the accuracy values of the classifiers, we conducted statistical tests on the accuracy values and the F1 scores obtained from all runs. The Shapiro-Wilk Test shows that the accuracy values and the F1 scores of the models do not distribute

Table 9. Comparison of model accuracy values and macro F1 scores for predicting break availability

Training Percentage	Accuracy					
	GLMM	GLM	General RF (Selected Variables)	General RF (All Variables)	Individual RF	Baseline
30%	53.47% ± 12.70%	45.01% ± 11.20%	45.73% ± 14.47%	48.08% ± 14.19%	33.42% ± 14.10%	33.87% ± 12.06%
40%	52.86% ± 13.36%	44.44% ± 12.04%	46.69% ± 14.50%	48.79% ± 14.06%	33.16% ± 14.61%	34.28% ± 11.89%
50%	52.36% ± 14.13%	43.58% ± 12.62%	44.95% ± 15.76%	48.04% ± 15.32%	34.38% ± 15.03%	33.15% ± 13.11%
60%	53.07% ± 16.19%	45.31% ± 14.93%	46.99% ± 17.99%	49.41% ± 17.29%	33.96% ± 18.00%	33.86% ± 15.09%
70%	53.51% ± 19.88%	45.64% ± 18.80%	46.79% ± 19.74%	49.77% ± 19.86%	34.42% ± 19.07%	31.51% ± 16.46%
LOOCV	49.36% ± 16.56%	41.86% ± 13.76%	46.63% ± 17.27%	51.25% ± 17.22%	NA	32.73% ± 13.80%
F1 Scores						
30%	55.57% ± 13.62%	55.15% ± 13.81%	51.83% ± 13.92%	53.10% ± 13.55%	51.64% ± 13.92%	50.03% ± 13.00%
40%	57.77% ± 13.33%	56.98% ± 14.56%	53.79% ± 13.65%	54.87% ± 13.58%	51.71% ± 14.85%	49.99% ± 13.49%
50%	57.38% ± 13.47%	57.14% ± 13.75%	53.62% ± 13.84%	55.09% ± 14.50%	52.19% ± 15.38%	48.73% ± 14.45%
60%	61.23% ± 15.12%	61.19% ± 15.28%	57.39% ± 16.01%	58.76% ± 15.44%	54.93% ± 15.89%	50.15% ± 15.73%
70%	63.35% ± 16.92%	63.25% ± 17.29%	60.76% ± 17.15%	62.51% ± 16.63%	57.14% ± 16.87%	48.87% ± 16.02%
LOOCV	55.51% ± 17.82%	51.55% ± 13.80%	54.53% ± 17.77%	53.58% ± 16.59%	NA	53.10% ± 8.47%

normally ( $p < .001$ ), hence, the Friedman Test was performed. The results of Friedman Test show that the average accuracy values obtained from four classifiers are significantly different for the prediction of break availability ( $\chi^2(3) = 1908.315, p < .001$ ). Then, Wilcoxon signed rank tests were conducted for binary comparisons of the models as post-hoc tests using the Bonferroni p-value adjustment. The results of Wilcoxon signed rank tests show that the accuracy obtained from GLMM is significantly higher than the accuracy of general random forest model with all variables ( $V = 16208, V = 17809, V = 16292, p < .05$  for 40%, 50% and 60% training percentages respectively), individual random forest model's accuracy ( $V = 38379, V = 36349, V = 31784, V = 31109, V = 27688$  for 30%, 40%, 50%, 60% and 70% training respectively and  $p < .001$  for all training percentages), and the baseline model accuracy ( $V = 37962, V = 30413, V = 27844, V = 26580, V = 23828$  for 30%, 40%, 50%, 60% and 70% training respectively and  $p < .001$  for all training percentages). Although there is not a significant difference between accuracy values of GLMM and general RF models, F1 scores of GLMM is significantly higher than general RF with selected variables according to the Friedman Test results ( $V = 39826, V = 39297, V = 37336, V = 36336, V = 25469$  for 30%, 40%, 50%, 60% and 70% training respectively and  $p < .001$  for all training percentages). GLMM also outperforms general RF with all variables in F1 scores ( $V = 38046, V = 39345, V = 35059, V = 35923, V = 25518$  for 30%, 40%, 50%, 60% and 70% training respectively and  $p < .001$  for all training percentages). The overall accuracy and F1 of GLMM are higher than GLM's although the difference between GLM and GLMM performances was not found significant.

#### 6.4 Performance of Individual Models

We also report the accuracy values of each user's model with their means and standard deviations in Table 10. Seventy percent of training data was used for the runs reported in the table, and in total, 20 runs were made. The bold values in the table show the highest accuracy among the four classifiers. The results are given based on the participants' number of responses ( $N$ ) in descending order. The participants whose number of responses is less than 20 ( $N=7$  users) are not included in the table because such a limited number of data points might not be sufficient for individual models to learn the target category.

The accuracy values obtained from the individual random forest classifiers were not as high as those obtained from GLMM and the general random forest classifier most of the time. The individual random forest classifiers

Table 10. The average and standard deviation of accuracy values for predicting break availability levels with GLMM, GLM, general RF, individual RF, and baseline classifier

User No	Number of Responses	GLMM	GLM	General RF	Individual RF	Baseline
U15	50	<b>51.25% ± 11.74%</b>	50.67% ± 11.11%	31.25% ± 9.78%	13.75% ± 10.28%	26.25% ± 8.54%
U17	49	<b>53.21% ± 9.33%</b>	27.67% ± 11.5%	40.36% ± 12.70%	33.57% ± 9.14%	37.14% ± 9.25%
U16	47	<b>50.33% ± 13.80%</b>	37.50% ± 15.00%	43.00% ± 13.17%	45.33% ± 10.13%	25.33% ± 11.49%
U03	45	61.70% ± 11.42%	49.29% ± 10.07%	<b>61.79% ± 7.06%</b>	61.07% ± 9.12%	43.57% ± 10.59%
U06	37	<b>50.00% ± 14.89%</b>	48.18% ± 13.55%	46.82% ± 12.26%	43.64% ± 13.06%	27.73% ± 13.02%
U12	36	50.91% ± 11.94%	<b>51.36% ± 12.26%</b>	39.09% ± 11.83%	34.09% ± 9.27%	49.09% ± 14.86%
U07	32	53.50% ± 15.65%	50.50% ± 14.68%	<b>56.50% ± 15.31%</b>	52.00% ± 15.08%	26.50% ± 8.13%
U01	31	40.56% ± 13.62%	31.67% ± 14.54%	<b>46.67% ± 13.77%</b>	38.33% ± 13.23%	19.44% ± 7.10%
U04	31	<b>47.78% ± 12.54%</b>	43.89% ± 16.31%	37.78% ± 13.20%	23.89% ± 9.72%	31.11% ± 10.57%
U13	27	<b>51.88% ± 14.21%</b>	49.38% ± 14.32%	40.63% ± 17.62%	31.25% ± 13.75%	45.00% ± 14.28%
U08	25	<b>50.71% ± 15.00%</b>	49.29% ± 15.70%	47.86% ± 18.11%	30.71% ± 16.24%	40.71% ± 21.88%
U05	24	<b>55.00% ± 18.11%</b>	45.71% ± 18.88%	32.14% ± 15.28%	35.71% ± 15.72%	17.14% ± 8.79%

provide the prediction accuracy results that are even worse than the baseline classifier for six users. Furthermore, GLMM predicted nine users' break availability most accurately among five classifiers. In this comparison, GLMM appears to be the best performing method to predict the break availability of the participants of our study.

### 6.5 Cold-Start User Evaluation

In order to compare the performance of the models in addressing the cold-start problem, we employed an iterative modelling for each of them. As reported in Table 9, we employed LOOCV removing each participant's data from the dataset, trained the models with the other remaining participants' data, and validated the models on that removed participant's data. In cold-start user evaluation, we added each participant's daily data to training iteratively and tested the model on the consequent days of the same user's data. For example; as seen in the upper left of Figure 8, we added the first two days' data of a participant to the training set (remember that there are already remaining participants' all days-data is in the training set), and validated the model performance on the same participant's remaining 8-days data. Note that we repeated this process for the same twelve users reported in Table 10 because of the limited number of data. Iteratively, we employed this method for each day, and reported the performance of each model for Day 2, Day 4, Day 6, and Day 8 using macro F1 scores (hence tested on the remaining 8, 6, 4 and 2 days of the individuals respectively). In Figure 8, each line represents a user, orange lines show the participants whose GLMM performance is higher than RF, and the purple lines indicate the opposite cases. The number of users that are associated to a better GLMM performance is higher on Day 2. This indicates that GLMM is able to predict most of our participants' responses even on Day 2. As days pass by, the RF model improves its performance, since the amount of available data for each user also increases. The same analysis was also conducted without using location frequency parameters to observe their effects on prediction results during cold-start evaluation. This way also enabled us to prevent any biases that might be caused by the threshold selection. The resulting figure is given in Appendix B. It can be seen that there is no significant difference between Figure 8 and Appendix B indicating that random forest model does not outperform GLMM even in the absence of location frequency variable. Location frequency was not found as significant in Table 8,

thus excluding it might not have affected cold-start evaluation of the classifiers. Given these results, GLMM can be considered as a potential solution to the cold-start problem.

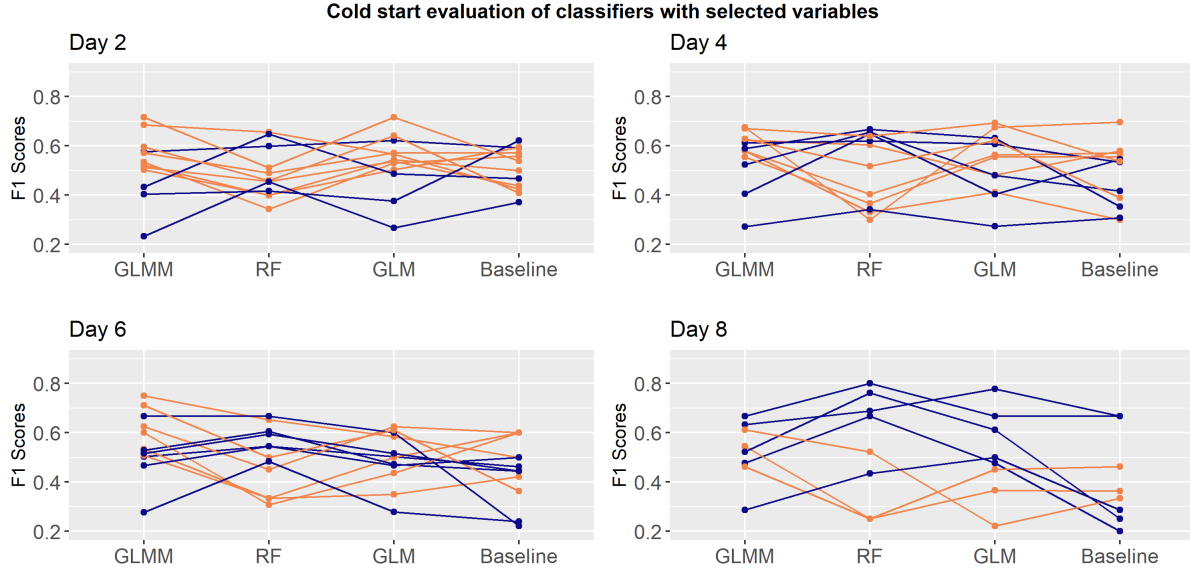


Fig. 8. F1 scores of GLMM, RF, GLM and baseline classifiers for predicting 8, 6, 4 and 2 days of the experiment for each individual using their own data on the first 2, 4, 6 and 8 days of their data respectively with LOOCV. *Upper-left*: shows the F1 scores of the models trained on users' 2-days data and tested on 8-days data. *Upper-right*: F1 scores of models trained on users' 4-days data and tested on 6-days data. *Bottom-left*: F1 scores of models trained on users' 6-days data and tested on 4-days data. *Bottom-right*: F1 scores of models trained on users' 8-days data and tested on 2-days data. The orange lines indicate the users whose GLMM performance is higher than RF and purple lines refer to those whose RF model performance is higher.

## 7 DISCUSSION

In this study, we built a hybrid model for inferring availability of office workers for having rest breaks using mobile phone sensing. The focus of this work is on the cold-start problem, the impact of the variance of the number and characteristics of user responses, and the repeated-measures design of the data. We compared our hybrid solution with individual and general models. In this section, we discuss our findings with respect to each feature used in our model in detail.

### 7.1 Time

Previous studies are mainly based on two types of time representation, namely hour of the day [16, 18, 25] or part of the day (e.g., morning, evening) [30, 33, 50]. However, if there is no linear relationship between target variable and time, time information may not be found significant in linear models, although it may actually be the case. For this reason, we converted time as a likelihood of each break availability level with a 2-D KDE. This conversion facilitated the modelling of time using generalised linear mixed models. Based on our results, time appears to be an effective factor in predicting break availability. Specifically, in our study, office workers tend to have longer breaks especially at midday (between 12-2 PM). They also tend to have shorter breaks at about 2-4

PM. The results presented in this study might be specific to the population under consideration, but we believe that the methodology we propose has general applicability.

## 7.2 Location

Previous studies showed that location is an important context information for predicting interruptibility [8, 9, 13, 30, 39, 43, 46, 48]. Since it was not possible to detect the exact indoor location of the users, we defined metrics in order to better understand user's presence in specific places. We investigated the effects of location in different measures: location frequency (i.e., how frequently a user visits a place) and location similarity (how similar a place is to the user's workplace). Based on our results, location frequency is associated to break availability of office workers.

Our results show that users tend to have longer breaks at locations where they visit less frequently. These locations may be the places where they have lunch breaks. Our preliminary results show that participants have longer breaks when they are in lunch/snack breaks. The model output supports this finding. Similarly, when they have just arrived at a location, they tend to have longer breaks. To the best of our knowledge, this is the first work that uses location parameters in terms of duration, frequency and similarity metrics for predicting the availability of office workers. Previous studies suggest that location-based notifications should be used for increasing user's receptivity [13]. Our work contributes to this area of work, providing researchers and practitioners with more insights about the factors linked to location and users' receptivity.

## 7.3 Physical Activity

The activity of users (i.e., whether the user is still or moving) affects the break availability as found in the previous studies [22, 30, 33, 34, 36]. Our results are in line with these findings. We have found that users are more likely to have a break when they are already moving. When they are not moving, the tendency of not taking a break is higher. Similarly, our results show that the time spent in action has a significant effect on the availability for a rest break. When the users are about to start the activity, the possibility of taking longer rest breaks is higher compared to the periods when they are about to finish their activity. The longer the time they spend in an activity, the less likely they will take a rest break.

## 7.4 Ringer Mode

In the pre-experiment questionnaire, the participants stated that they keep their mobile phones in different ringer modes in different contexts as in [8]. More specifically, none of the participants keeps their phones in sound mode while in a meeting or in a prayer break. A change of the ringer mode of a phone is an indicator of a change in context. The person may be attending a meeting or going for a break during which she/he does not want to be interrupted. Considering such situations, we recorded the duration of each ringer mode status from the sensor data we collected. We extracted the ringer mode used by the participants most of the time and logged the changes with respect to that ringer mode.

Different from previous work, we kept a variable for ringer mode change in addition to original ringer mode values, and investigated its effect. We hypothesised that a change in the ringer mode affects users' context and implicitly their availability. Our results show that ringer mode and ringer mode change are supplementary parameters for identifying break availability complementary to location, time, and activity. According to our results, users do not tend to take a break when the ringer mode is in silent-or-vibrate mode and when there is a change in ringer mode. Since most of the users keep their mobile phones in sound mode, the change may mean that they do not want to be interrupted. Hence, ringer mode change to vibrate or silent might be interpreted as a potential indicator of unavailability. The results presented in Fig. 6 support this finding. Note that the number of

changes from sound to silent and the number of changes from sound to vibrate are higher in the “cannot take a break” and “less than 5-min” categories.

## 7.5 Application Usage

Previous work showed that an increase in application usage may be an indicator for an available moment to take a break or responsiveness [14, 28]. According to our results, the models based on application usage variables are characterised by worse performance with respect to the others. In other words, application usage does not seem linked to break availability in this study. This may be due to cultural or environmental factors. Since we collected our dataset in office environments during working hours, employees may not be able to use their mobile phones even when they are available for a break. In addition, other factors that have not been included in the study, such as application type, might explain this result. A notification might have arrived while users are taking notes on their mobile phones during a meeting. In such situations, application type or category may become a more important feature than application usage. One of the previous studies [44] had found similar results showing the importance of application type on interruptibility.

## 7.6 Individual-level Models, Data Sparsity and the “Cold-start Problem”

We compared GLMM with random forest classifier, which is commonly used as a machine learning algorithm at the basis of intelligent notification mechanisms (e.g., [53, 61]) and also with a baseline classifier. Previous studies had a higher number of data points than ours (e.g., minimum 127 for an individual in [61]) and this might have led to a better performance for individual models. The positive performance trend of the random forest models also suggest that they could result in a better performance compared to GLMM if the study has lasted longer. Our results show that GLMM may be preferable compared to individual random forest models when sufficient data is not available. GLMM is an appropriate choice in the presence of data sparsity. It also incorporates both individual and general-level mean. Hence, when there is a new user with an insufficient number of data points, the general-level mean might be used for that user at first until the amount of data is sufficient to fit an individual-level mean. GLMM is an effective method, which does not violate any statistical assumption of having repeated-measures data. It has been widely used in numerous recent studies especially in botany or biology domain [19, 27]. Its main limitation is its linearity. In the case of abundant data, other methods such as deep learning algorithms or ensemble methods could be applied on the data. Collecting human data in experiments is compelling specifically for researchers because of the challenges in recruiting participants or participants opting out from the experiment etc. Our solution might be one of the solutions for those challenging situations. We believe that this brings a solution to the cold-start problem stated in [61].

GLMM can be used in the cold-start period, then, when sufficient data is accumulated, individual models can supersede them. As Wozniak, Grana and Corchado [58] stated in their paper: “Referring to classification problems, Wolpert’s theorem [57] has a specific lecture: there is not a single classifier modeling approach which is optimal for all pattern recognition tasks, since each has its own domain of competence”. Each problem has different characteristics and it is almost impossible to find a general solution, however, we offer a novel solution: implementing GLMM at the first stages of modelling, then based on the performance, switching to another classifier such as random forest or deep learning methods. For instance, recommendation systems also make use of switching algorithms such as in the study of [6]. Hence, such an approach can be used in our domain as well. As a future step in our research agenda, we plan to explore this approach using a larger dataset, possibly with a different population.

## 7.7 Break Types vs. Break Availability

Our data showed that certain types of breaks have different characteristics in terms of duration. Lunch/snack breaks last more than 15 minutes, whereas social or coffee/tea breaks last approximately 10-15 minutes. Users



marked themselves “not available” when they are in the middle of working or meeting. Therefore, it is possible to infer the type of activity and how this is related to the predictability of users’ availability. However, the use of break types for prediction was not successful with GLMM due to the limited sample size and similar characteristics of certain break types. It is worth noting that our results are also in line with a recent study of receptivity for health interventions [9].

## 8 IMPLICATIONS

In this article, we have presented a hybrid model for the prediction of break availability of office workers, based on both individual and population-level data. Personalised models have been increasingly used in ubiquitous computing. Their drawback is that they require a large amount of individual data for training. This might be difficult, especially when an application is initially installed or at the beginning of a study. Our results show that using a hybrid model that addresses both within- and between-subjects’ factors, is an effective method for modelling unbalanced and sparse data. Besides, we have shown that we obtained significant results thanks to the repeated-measures nature of the dataset without any violation of statistical assumptions such as the independence of data points, having a Gaussian distribution.

This study also provides researchers and practitioners with general insights for the design of mobile intervention systems, especially for those targeting workplace settings. In particular, this work has extensively analysed and discussed the factors that should be considered in the design of such applications. We have shown that some contextual information is particularly informative, namely features related to user locations or activities, such as the time spent in action and the frequency of visits to a given location.

## 9 LIMITATIONS AND FUTURE WORK

One limitation of the current work is the lack of accurate location tracking. Since we were not able to detect exact locations, we clustered the locations of Wi-Fi access points. Even though we have tested several threshold values for location clustering, the results presented in this study might depend on the characteristics of the signals of Wi-Fi access points used by the participants of our study. In the wild, the first option is that several methods with different threshold values might be used to generate a collective output for prediction. In the second option, the method can be used with the default threshold. In the third option, users may be asked to label their important locations and they could be employed in the method.

In this study, we only considered application usage in terms of duration. Other features, such as application category and inter-event time between usage sessions might be included in future work. Moreover, in this study we only considered data from mobile sensors for predicting the availability of rest breaks. Other information coming from wearable sensors, including physiological sensors, might be used in order to improve the accuracy of the prediction. The response time (the time between the notification sent to users and the response submitted) was not included in this study. This aspect is part of our future research agenda.

Another main limitation of the study is related to the composition of the population of the study. Since we recruited participants based on convenience sampling, we were able to recruit only participants who commonly work in the same city. More in general, the data was collected among participants of a single country, hence, there may be cultural characteristics that impact the results of our study. Having said that, we stress the fact that the contribution of this work is methodological and can apply to populations with different backgrounds.

Besides, the size of the dataset is limited, even though it is worth noting that GLMM can handle a limited number of data points. Indeed, we were able to collect a small number of data points for some users. The main reason for not being able to collect a high number of data points could be associated with the burden of replying to six ESM messages in a day. For this reason, the participants with a relatively low threshold of 25% of response rate were included in the study. Similarly, since the pilot phase of the experiment required significant manual effort,

we could not recruit more than five participants. They labelled their locations for each day of the experiment. This might have been particularly burdensome for the participants of the pilot phase. Nevertheless, we made an effort to choose pilot participants who work in different locations (e.g., in a private company, in academia) in order to develop a more robust clustering for locations. Compliance could be improved with a better design, which reduces the burden for participants. The performance results for both the GLMM and individual classifiers could also be indeed improved with a higher number of data points.

Because of its sparsity, our dataset did not contain a sufficient number of data points for different activity types (e.g., driving, walking); hence, we had to classify activities into two categories: “still” and “moving”. The effects of different activity types on break availability could be investigated separately with a larger amount of data. In fact, this is directly related to the pilot participants’ data as we discussed above. If we had a higher number of participants in the pilot phase, we would have been able to train our activity classifier with other activity types as well. In addition, the sampling rate for the older phones which do not support the “significant motion sensor” could have been increased. However, the choice of selection of the sampling rate was done taking into consideration the capacity of the phones. Since we were not able to collect a higher amount of data, this might be considered as a limitation of our analysis.

Finally, we point out that in our approach, we went through a series of pre-processing steps, such as location clustering and the application of activity recognition algorithms for identifying users’ physical activity. This was necessary due to data sparsity. In general, from a practical point of view, to apply the proposed method in the same way we did, it is necessary to derive user locations and calculate the similarity or frequencies of the locations. On the other hand, the results in Table 8 show that time variables are statistically significant, whereas the higher  $p$  value of location frequency (LF) suggests that it may not be as critical as the other variables. The results shown in the figures in Appendix B demonstrate that the models built without location variables perform well compared to the models built with them. Due to potential security and privacy concerns, users might not be willing to share their location information with third party applications. Mobile phone usage over time, however, is less affected by privacy issues. KDE models with time information performed well even with a lower amount of data in the training set. As a consequence, although it is necessary to collect location information, the model performance is not highly dependent on the location variables as seen in the results.

## 10 CONCLUSION

In this study we have presented a method for addressing the issues related to the availability of a limited amount of personal data for building individual predictive behavioural model through mobile sensing. In particular, the approach can be considered a solution to the “cold-start problem”, i.e., the negative impact of the lack of data about an individual (for example, when a new application is installed or at the beginning of a study) on the performance of machine learning classifiers.

We have discussed a specific case study, i.e., the prediction of the availability of office workers to take rest breaks using data collected through mobile sensing. In particular, we have developed a novel hybrid model, which relies on data from individuals and the general population. We have compared the results of our model and found that GLMM is able to provide more accurate results than an individual random forest model, typically adopted in previous studies.

We have also discussed which context information is informative for the prediction of rest break availability. Specifically, location, time, activity and ringer mode status are associated with users’ availability. This result is very important for designing effective positive behaviour intervention for the wellbeing of employees. Finally, we believe that the contribution of this study is methodological in nature: even if the results we presented might be specific to the population taken into consideration, the methodology is of general applicability and can be adopted for applications in different situations, involving users from different backgrounds.

## ACKNOWLEDGMENTS

This work is supported by The Scientific and Technological Research Council of Turkey under Tubitak BIDEB-2219 grant no 1059B191500728.

## REFERENCES

- [1] Android Developers Web Page 2020. *Motion sensors*. Retrieved 2020-03-09 from [https://developer.android.com/guide/topics/sensors/sensors\\_motion#sensors-motion-significant](https://developer.android.com/guide/topics/sensors/sensors_motion#sensors-motion-significant)
- [2] Jonathan Z. Bakdash and Laura R. Marusich. 2017. Repeated measures correlation. *Frontiers in Psychology* 8, MAR (2017), 1–13. <https://doi.org/10.3389/fpsyg.2017.00456>
- [3] Nikola Banovic, Christina Brant, Jennifer Mankoff, and Anind Dey. 2014. ProactiveTasks: The short of mobile device use sessions. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices and Services*. ACM Press, New York, USA, 243–252. <https://doi.org/10.1145/2628363.2628380>
- [4] Ronald De Vera Barredo and Kelly Mahon. 2007. The effects of exercise and rest breaks on musculoskeletal discomfort during computer tasks: an evidence-based perspective. *Journal of Physical Therapy Science* 19, 2 (2007), 151–163. <https://doi.org/10.1589/jpts.19.151>
- [5] Dave Berque, Jimmy Burgess, Alexander Billingsley, ShanKara Johnson, Terri L. Bonebright, and Brad Wethington. 2011. Design and evaluation of persuasive technology to encourage healthier typing behaviors. In *Proceedings of the 6th International Conference on Persuasive Technology Persuasive Technology and Design: Enhancing Sustainability and Health*. ACM Press, New York, USA, 1–10. <https://doi.org/10.1145/2467803.2467812>
- [6] Daniel Billsus and Michael J Pazzani. 2000. User modeling for adaptive news access. *User modeling and user-adapted interaction* 10, 2-3 (2000), 147–180.
- [7] Scott A. Cambo, Daniel Avrahami, and Matthew L. Lee. 2017. BreakSense: Combining physiological and location sensing to promote mobility during work-breaks. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3595–3607. <https://doi.org/10.1145/3025453.3026021>
- [8] Yung-Ju Chang and John C. Tang. 2015. Investigating mobile users' ringer mode usage and attentiveness and responsiveness to communication. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 6–15. <https://doi.org/10.1145/2785830.2785852>
- [9] Woohyeok Choi, Sangkeun Park, Duyeon Kim, Youn-kyung Lim, and Uichin Lee. 2019. Multi-stage receptivity model for mobile just-in-time health intervention. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, Vol. 3. ACM, 1–26. <https://doi.org/10.1145/3328910>
- [10] Dean Cooley and Scott Pedersen. 2013. A pilot study of increasing nonpurposeful movement breaks at work as a means of reducing prolonged sitting. *Journal of Environmental and Public Health* 2013 (2013). <https://doi.org/10.1155/2013/128376>
- [11] Mihaly Csikszentmihalyi and Reed Lardon. 1983. The experience sampling method. *New Directions for Methodology of Social and Behavioral Science* 15 (1983), 41–56.
- [12] Daniel A Epstein, Daniel Avrahami, and Jacob T Biehl. 2016. Taking 5: Work-breaks, productivity, and opportunities for personal informatics for knowledge workers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. 673–684. <https://doi.org/10.1145/2858036.2858066>
- [13] Anja Exler, Marcel Braith, Andrea Schankin, and Michael Beigl. 2016. Preliminary investigations about interruptibility of smartphone users at specific place types. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct (UbiComp '16)*. 1590–1595. <https://doi.org/10.1145/2968219.2968554>
- [14] Joel E. Fischer, Chris Greenhalgh, and Steve Benford. 2011. Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11)*. 181–190. <https://doi.org/10.1145/2037373.2037402>
- [15] Joel E. Fischer, Nick Yee, Victoria Bellotti, Nathan Good, Steve Benford, and Chris Greenhalgh. 2010. Effects of content and time of delivery on receptivity to mobile interruptions. In *Proceedings of the 12th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '10)*. 103–112. <https://doi.org/10.1145/1851600.1851620>
- [16] Traci Galinsky, Naomi Swanson, Steven Sauter, Robin Dunkin, Joseph Hurrell, and Lawrence Schleifer. 2007. Supplementary breaks and stretching exercises for data entry operators: a follow-up field study. *American Journal of Industrial Medicine* 50, 7 (jul 2007), 519–27. <https://doi.org/10.1002/ajim.20472>
- [17] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis* (2nd ed. ed.). Chapman & Hall, Boca Raton, FL, USA.
- [18] Miriam Gil, Pau Giner, and Vicente Pelechano. 2012. Personalization for unobtrusive service interaction. *Personal and Ubiquitous Computing* 16, 5 (jun 2012), 543–561. <https://doi.org/10.1007/s00779-011-0414-0>

- [19] Alexandre Girard, Nathalie Br  heret, Ga  lle Bal, Jean-Gabriel Mavoungou, Jean-F  lix Tchibinda, Fils Makaya, and Marc Girondot. 2021. Unusual sexual dimorphism and small adult size for olive ridley sea turtles are linked to volumetric geometric constraints. *Marine Biology* 168, 1 (2021), 1–11.
- [20] J. D. Hadfield and S. Nakagawa. 2010. General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology* 23, 3 (2010), 494–508. <https://doi.org/10.1111/j.1420-9101.2009.01915.x>
- [21] Genevieve N Healy, David W Dunstan, Jo Salmon, Ester Cerin, Jonathan E Shaw, Paul Z Zimmet, and Neville Owen. 2008. Breaks in sedentary time: beneficial associations with metabolic risk. *Diabetes Care* 31, 4 (apr 2008), 661–6. <https://doi.org/10.2337/dc07-2046>
- [22] Joyce Ho and Stephen S. Intille. 2005. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. 909–918. <https://doi.org/10.1145/1054972.1055100>
- [23] M. C. Jones, J. S. Marron, and S. J. Sheather. 1996. A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.* 91, 433 (mar 1996), 401. <https://doi.org/10.2307/2291420>
- [24] Kyohei Komuro, Yuichiro Fujimoto, and Kinya Fujita. 2017. Relationship between worker interruptibility and work transitions detected by smartphone. In *Human-Computer Interaction. Interaction Contexts. HCI 2017. Lecture Notes in Computer Science*. Springer, Cham, 687–699. [https://doi.org/10.1007/978-3-319-58077-7\\_53](https://doi.org/10.1007/978-3-319-58077-7_53)
- [25] N. Lathia, V. Pejovic, K. Rachuri, C. Mascolo, M. Musolesi, and P. J. Rentfrow. 2014. Smartphones for large-scale behaviour change interventions. *IEEE Pervasive Computing* 12, 3 (2014), 66–73. <https://doi.org/10.1109/MPRV.2013.56>
- [26] Christine Leah. 2011. *Exercises to reduce musculoskeletal discomfort for people doing a range of static and repetitive work*. Technical Report. Health and Safety Laboratory. <https://www.hse.gov.uk/research/rpdf/rr743.pdf>
- [27] Po-An Lin, Sulav Paudel, Amin Afzal, Nancy L Shedd, and Gary W Felton. 2021. Changes in tolerance and resistance of a plant to insect herbivores under variable water availability. *Environmental and Experimental Botany* 183 (2021), 104334.
- [28] Akhil Mathur, Nicholas D. Lane, and Fahim Kawsar. 2016. Engagement-aware computing: modelling user engagement from mobile contexts. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 622–633. <https://doi.org/10.1145/2971648.2971760>
- [29] Charles E Mcculloch. 2003. Generalized linear mixed models. In *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics. [http://www.jstor.org/stable/pdf/4153190.pdf?refreqid=excelsior\[%\]3A90e9e2c63f204843d73cc61ef7589b51](http://www.jstor.org/stable/pdf/4153190.pdf?refreqid=excelsior[%]3A90e9e2c63f204843d73cc61ef7589b51)
- [30] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016. PrefMiner: Mining user’s preferences for intelligent mobile notification management. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, 1223–1234. <https://doi.org/10.1145/2971648.2971747>
- [31] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My phone and me: understanding people’s receptivity to mobile notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1021–1032. <https://doi.org/10.1145/2858036.2858566>
- [32] Dan Morris, A.J. Bernheim Brush, and Brian R. Meyers. 2008. SuperBreak: using interactivity to enhance ergonomic typing breaks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM Press, 1817–1826. <https://doi.org/10.1145/1357054.1357337>
- [33] Hyungik Oh, Laleh Jalali, and Ramesh Jain. 2015. An intelligent notification system using context from real-time personal activity monitoring. In *IEEE International Conference on Multimedia and Expo*. 1–6. <https://doi.org/10.1109/ICME.2015.7177508>
- [34] Tadashi Okoshi, Hiroki Nozaki, Jin Nakazawa, Hideyuki Tokuda, Julian Ramos, and Anind K. Dey. 2016. Towards attention-aware adaptive notification on smart phones. *Pervasive and Mobile Computing* 26 (2016), 17–34. <https://doi.org/10.1016/j.pmcj.2015.10.004>
- [35] K. T. Palmer. 2001. Use of keyboards and symptoms in the neck and arm: evidence from a national survey. *Occupational Medicine* 51, 6 (2001), 392–395. <https://doi.org/10.1093/occmed/51.6.392>
- [36] C. Park, J. Lim, J. Kim, S.-J. Lee, and D. Lee. 2017. "Don’t bother me. I’m socializing!": a breakpoint-based smartphone notification system. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. 541–554. <https://doi.org/10.1145/2998181.2998189>
- [37] Sharon Parry and Leon Straker. 2013. The contribution of office work to sedentary behaviour associated risk. *BMC Public Health* 13, 1 (2013), 296. <https://doi.org/10.1186/1471-2458-13-296>
- [38] Veljko Pejovic, Neal Lathia, Cecilia Mascolo, and Mirco Musolesi. 2016. Mobile-based experience sampling for behaviour research. In *Emotions and Personality in Personalized Services*. Springer, Cham, 141–161. <http://arxiv.org/abs/1508.03725>
- [39] Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: designing intelligent prompting mechanisms for pervasive applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 897–908. <https://doi.org/10.1145/2632048.2632062>
- [40] Martin Pielot, Karen Church, and Rodrigo de Oliveira. 2014. An in-situ study of mobile phone notifications. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices and Services*. 233–242. <https://doi.org/10.1145/2628363.2628364>

- [41] Juho Piironen and Aki Vehtari. 2017. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing* 27, 3 (2017), 711–735. <https://doi.org/10.1007/s11222-016-9649-y>
- [42] Zinta Podniece, S Heuvel, and B Blatter. 2008. *Work-related musculoskeletal disorders: prevention report*. Technical Report. European Agency for Safety and Health at Work. [https://osha.europa.eu/en/publications/reports/en\\_TE8107132ENC.pdf](https://osha.europa.eu/en/publications/reports/en_TE8107132ENC.pdf)
- [43] Benjamin Poppinga, Wilko Heuten, and Susanne Boll. 2014. Sensor-based identification of opportune moments for triggering notifications. *IEEE Pervasive Computing* 13, 1 (2014), 22–29. <https://doi.org/10.1109/MPRV.2014.15>
- [44] Alireza Sahami Shirazi, Niels Henze, Tilman Dingler, Martin Pielot, Dominik Weber, and Albrecht Schmidt. 2014. Large-scale assessment of mobile notifications. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. 3055–3064. <https://doi.org/10.1145/2556288.2557189>
- [45] Yu Sang, Heng Qi, Keqiu Li, Yingwei Jin, Deqin Yan, and Shusheng Gao. 2014. An effective discretization method for disposing high-dimensional data. *Information Sciences* 270 (2014), 73–91. <https://doi.org/10.1016/j.ins.2014.02.113>
- [46] Hermann Stern, Viktoria Pammer, and SN Lindstaedt. 2011. A preliminary study on interruptibility detection based on location and calendar information. In *Proceedings of the 3rd Workshop on Context-Systems Design, Evaluation and Optimisation*.
- [47] Wendell C. Taylor, Ross Shogog, Vincent Chen, David M. Rempel, Marybeth Pappas Baun, Cresendo L. Bush, Tomas Green, and Nicole Hare-Everline. 2010. The Booster Break program: Description and feasibility test of a worksite physical activity daily practice. *Work* 37, 4 (2010), 433–443. <https://doi.org/10.3233/WOR-2010-1097>
- [48] G. H. Ter Hofte. 2007. Xensible interruptions from your mobile phone. In *Proceedings of the 9th International Conference on Human Computer Interaction with Mobile Devices and Services*. 178–181. <https://doi.org/10.1145/1377999.1378003>
- [49] Piiastiina Tikka and Harri Oinas-Kukkonen. 2016. RightOnTime: The role of timing and unobtrusiveness in behavior change support systems. In *International Conference on Persuasive Technology (Persuasive '16)*. Springer International Publishing, 327–338. [https://doi.org/10.1007/978-3-319-31510-2\\_28](https://doi.org/10.1007/978-3-319-31510-2_28)
- [50] Liam D. Turner, Stuart M. Allen, and Roger M. Whitaker. 2017. Reachable but not receptive: Enhancing smartphone interruptibility prediction by modelling the extent of user engagement with notifications. *Pervasive and Mobile Computing* 40 (2017), 480–494. <https://doi.org/10.1016/j.pmcj.2017.01.011>
- [51] Yunus Emre Ustev, Ozlem Durmaz Incel, and Cem Ersoy. 2013. User, device and orientation independent human activity recognition on mobile phones. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct*. ACM Press, New York, USA, 1427–1436. <https://doi.org/10.1145/2494091.2496039>
- [52] Saskia Van Dantzig, Gijs Geleijnse, and Aart Tjimen Van Halteren. 2013. Toward a persuasive mobile application to reduce sedentary behavior. *Personal and Ubiquitous Computing* 17, 6 (2013), 1237–1246. <https://doi.org/10.1007/s00779-012-0588-0>
- [53] Aku Visuri, Niels Van Berkel, Chu Luo, Jorge Goncalves, Denzil Ferreira, and Vassilis Kostakos. 2017. Predicting interruptibility for manual data collection: a cluster-based user model. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 12. <https://doi.org/10.1145/3098279.3098532>
- [54] Dhaval Vyas, Thilina Halloluwa, Nikolaj Heinzler, and Jinglan Zhang. 2019. More than step count: designing a workplace-based activity tracking system. *Personal and Ubiquitous Computing* (sep 2019), 1–15. <https://doi.org/10.1007/s00779-019-01305-1>
- [55] He Wang, Ahmed Elgohary, and Romit Roy Choudhury. 2012. No need to war-drive: Unsupervised indoor localization. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services (MobiSys '12)*. 197–210. <https://doi.org/10.1145/2307636.2307655>
- [56] R. Williams and M. Westmorland. 1994. Occupational cumulative trauma disorders of the upper extremity. *American Journal of Occupational Therapy* 48, 5 (1994), 411–420. <https://doi.org/10.5014/ajot.48.5.411>
- [57] David H Wolpert. 2002. The supervised learning no-free-lunch theorems. In *Soft computing and industry*. Springer, 25–42.
- [58] Michał Woźniak, Manuel Graña, and Emilio Corchado. 2014. A survey of multiple classifier systems as hybrid systems. *Information Fusion* 16 (2014), 3–17.
- [59] Manuela Züger, Christopher Corley, André N Meyer, Boyang Li, Thomas Fritz, David Shepherd, Vinay Augustine, Patrick Francis, Nicholas Kraft, and Will Snipes. 2017. Reducing interruptions at work: A large-scale field study of flowlight. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 61–72. <https://doi.org/10.1145/3025453.3025662>
- [60] Manuela Züger and Thomas Fritz. 2015. Interruptibility of software developers and its prediction using psycho-physiological sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2981–2990. <https://doi.org/10.1145/2702123.2702593>
- [61] Manuela Züger, Sebastian C. Müller, André N. Meyer, and Thomas Fritz. 2018. Sensing interruptibility in the office: a field study on the use of biometric and computer interaction sensors. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 591. <https://doi.org/10.1145/3173574.3174165>

## A NOTIFICATION DELIVERY ALGORITHM

---

**Algorithm 1:** Notification delivery algorithm.
 

---

```

delete one hour from the beginning and the end of work day of users;
set  $d$  (day number) = 1;
set  $n$  (notification number) = 1;
while not end of user set do
  while  $d \leq 10$  do
    import calendar events of user for day;
    calculate empty slots by extracting calendar events from user's work day;
    check empty slots for day ( $e_d$ );
    if  $e_d = \emptyset$  then
      skip day;
      increase  $d$ ;
    while  $n \leq 6$  do
      check preferred time slots of user collected in the questionnaire ( $p$ );
      if  $p = \emptyset$  then
        continue;
      else
        if  $n > 2$  then
          continue;
        else
          pick a random hour ( $r$ ) from  $p$ ;
          delete  $r$  from  $e_d$ ;
          if  $r \in e$  then
            continue;
          else
            go back checking  $p$  step;
      select  $r$  from  $e$ ;
      set notification time as selected random hour;
      delete one hour before and after selected notification time from  $e_d$ ;
      increase  $n$ ;
    increase  $d$ ;
  continue with next user;

```

---



B COLD-START EVALUATION WITHOUT LOCATION VARIABLES

