Trajectories of Depression: Unobtrusive Monitoring of Depressive States by means of Smartphone Mobility Traces Analysis

Luca Canzian
University of Birmingham, UK
l.canzian@cs.bham.ac.uk

Mirco Musolesi
University College London, UK
University of Birmingham, UK
m.musolesi@ucl.ac.uk

ABSTRACT

One of the most interesting applications of mobile sensing is monitoring of individual behavior, especially in the area of mental health care. Most existing systems require an interaction with the device, for example they may require the user to input his/her mood state at regular intervals. In this paper we seek to answer whether mobile phones can be used to unobtrusively monitor individuals affected by depressive mood disorders by analyzing only their mobility patterns from GPS traces. In order to get ground-truth measurements, we have developed a smartphone application that periodically collects the locations of the users and the answers to daily questionnaires that quantify their depressive mood. We demonstrate that there exists a significant correlation between mobility trace characteristics and the depressive moods. Finally, we present the design of models that are able to successfully predict changes in the depressive mood of individuals by analyzing their movements.

Author Keywords

Mobile Sensing; Depression; Spatial Statistics; GPS Traces

ACM Classification Keywords

H.1.2. Models and Principles: User/Machine Systems; J.4 Computer Applications: Social and Behavioral Sciences

INTRODUCTION

According to a recent report by the World Health Organization [9], in high-income countries up to 90% of people who die by suicide are affected by mental disorders, and depression is the most common mental disorder associated with suicidal behavior. More generally, depressive disorders do not only affect the personal life of individuals and their families and social circles, but they also have a strong negative economic impact [28]. In fact, according to a study by the European Depression Association [9], 1 in 10 employees in the United Kingdom had taken time off at some point in their working lives because of depression problems. Currently, psychologists rely mainly on self-assessment questionnaires and phone/in-site interviews to diagnose depression and monitor its evolution. This methodology is time-consuming, expensive, and prone to errors, since it often relies on the patient's recollections and self-representation. As a consequence, changes in the depression state may be detected with delay, which makes intervention and treatment more difficult.

Several recent projects have investigated the potential use of mobile technologies for monitoring stress, depression and other mental disorders (see, for example, [25, 6, 31, 24, 36, 1, 5, 39], providing new ways for supporting both patients and healthcare officers [8, 20]. Indeed, mobile phones are ubiquitous and highly personal devices, equipped with sensing capabilities, which are carried by their owners during their daily routine [19]. However, existing works mostly rely on periodic user interaction and self-reporting. Our goal is to build systems that *minimize* and, if possible, *remove* the need for user interaction.

We focus on a specific type of data that can be reliably collected by almost any smartphone in a robust way, namely *location information*, and we investigate how it is possible to correlate characteristics of human mobility and depressive state. Indeed, interview-based studies have shown that depression leads to a reduction of mobility and activity levels (see, for example, [34]). Previous work has shown the potential of using different smartphone sensor modalities to assess mental well-being. However, the focus was on the activity level detected with the accelerometer sensor [31], voice analysis using the microphone [24], colocation using Bluetooth and WiFi registration patterns [25], and call logs [5]. In this paper instead we focus on the characterization (also from a statistical point of view) and exploitation of *mobility data collected by means of the GPS receivers embedded in today's mobile phones*. More specifically, this work for the first time addresses the following key questions: *is there any correlation between mobility patterns extracted from GPS traces and depressive mood?* Is it possible to devise unobtrusive smartphone applications that collect and exploit *only* mobility data in order to automatically infer a potential depressed mood of the user over time?

In order to answer these questions, we need to *quantitatively* characterize the movements of the user over a certain time interval and correlate them to a *numeric* indicator of the depressed mood of a user. For this reason, we first extract *mobility traces* for a user and we define and compute *mobility metrics* that summarize key features of the user movement pat-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
UbiComp '15, September 07-11, 2015, Osaka, Japan
© 2015 ACM. ISBN 978-1-4503-3574-4/15/09 $15.00
DOI: http://dx.doi.org/10.1145/2750858.2805845

terns over time. We then use the questions from the widely-used "PHQ-8" depression test [16, 18, 15] in order to quantify depressive states.

In order to obtain ground truth measurements about the correlation of mobility patterns and depressive states, we have developed an Android application for smartphones – *MoodTraces* [27] – that periodically collects the locations of the users. MoodTraces collects also the answers to 8 daily questions from the "PHQ-8" depression test that the users are asked to take, concerning the occurrence of specific depressive symptoms in the current day. The answers are used to compute a daily integer score for each user ranging from 0 to 24, which we call *PHQ score*. For each user, we then analyze how the mobility metrics and the PHQ score vary in time, proving that there exists a significant correlation between them. Driven by these results, we then investigate whether it is possible to predict changes in the PHQ score from variations in the mobility metrics. To achieve this goal, we train and test personalized classification models for each user. An extensive evaluation shows that, for most of the users, these personalized models are able to accurately detect changes in the PHQ score exploiting only mobility metrics. It is worth noting that, after a training phase, these models are able to monitor the depressive state of individuals without requiring a direct interaction with the device. This is particularly important for patients with serious depressive conditions who might not be willing (or, unfortunately, sometimes able) to actively report their condition using the mobile device. Moreover, a generic, albeit less accurate, training model built on data collected in pilot studies like this might be used in order to remove the need of a training phase in the deployment of the application.

To summarize, the contribution of this paper is threefold. *First*, we design an energy-efficient Android application to collect mobility data and assess the presence of a depressed mood, and we deploy it and collect data from 28 users. *Second*, we define a set of mobility metrics that can be extracted from the mobility traces of the users and, using the ground truth data collected by means of the Android application, we identify a significant correlation between the changes of such metrics and the variations in the PHQ score. Such a correlation ranges from 0.336 to 0.432 when the mobility metrics are computed over a period of 14 days. *Third*, we train and evaluate personalized and general machine learning models to predict PHQ score changes from mobility metrics variations, obtaining very good prediction accuracies. For example, when the mobility metrics are computed over a period of 14 days, the general model achieves sensitivity and specificity values of 0.74 and 0.78 (respectively), whereas the average sensitivity and specificity values of the personalized models are 0.71 and 0.87 (respectively).

RELATED WORK

Given the increasing availability of mobility traces extracted by means of GPS phones or WiFi registration records, we have witnessed a growing interest in the investigation of the properties of human movement, with the goal of identifying patterns or developing prediction models [37, 2]. Mobility and other contextual information are increasingly collected by means of mobile phone sensing applications [7], i.e., by means of sensors (such as GPS, accelerometers, etc.), which are embedded in today's smartphones.

In particular, in the pervasive and ubiquitous computing community, several projects have investigated the use of smartphone data for the automatic detection and prediction of psychological states and mental health conditions [22, 32, 23, 6].

Stress monitoring using smartphones has been studied in [24, 1, 36, 5]. More specifically, user locations and social interactions (through Bluetooth proximity, phone calls and SMS logs) are exploited by the authors of [1] to detect differences between stressful and non stressful periods, showing that behavior changes can be automatically detected by means of mobile phones.

The detection and monitoring of bipolar disorder by means of mobile technologies are discussed in [10, 11, 29, 12]. The authors in [10] reports the results of a 6 months trial with 6 patients suffering from bipolar disorder, in which they record subjective and objective data (including self-assessments, activity, sleep, and phone usage) and inform both the patient and clinicians on the importance of the different data items according to the patient's mood. In [11, 29, 12] the authors describe multiple real-life studies of the use of smartphone based sensors for state monitoring of bipolar disorder. [11] shows initial evidence that relatively simple features derived from location, motion and phone call patterns are a good indication of state transitions; [29] analyzes how the episodes of the diseases correlate to the sampled data and suggest that personalized models are better suited to detect early signs of the onset of a bipolar episode; [12] studies the detection of state changes, achieving a precision and recall of 96% and 94%, respectively, and state recognition accuracy of 80%.

To the best of our knowledge this work for the first time demonstrates that it is possible to devise metrics that can be used to capture the correlation between mobility patterns extracted from GPS traces and depressive mood. Moreover, our paper differentiates from the above cited literature in a number of different ways. *First*, the works in [24, 23, 6, 1, 36, 5] focus on mood or stress detection and not on the analysis and prediction of variations of the depressive states of individuals. *Second*, most of the cited works (i.e., [23, 6, 1, 36, 5, 10]) require user interaction. Conversely, our work aims to develop applications for automatic and unobtrusive depression diagnosis and monitoring that do not require any user interaction, and to achieve this goal we focus only on aggregate metrics that are extracted solely from mobility data. *Third*, some of the cited works (e.g., [1, 10, 11, 29]) aim at identifying correlations between the collected data and the target parameter, whereas in our work we also propose *predictive* mechanisms in order to forecast depressive mood changes from mobility data. *Fourth*, the projects presented in [10, 11, 29, 12] analyze a population of individuals suffering bipolar disorder and their aim is to detect changes between severe depression states and severe mania states. Instead, in our work we collect data from a general population, inside which only few individuals suffer from a severe form of depression. As a

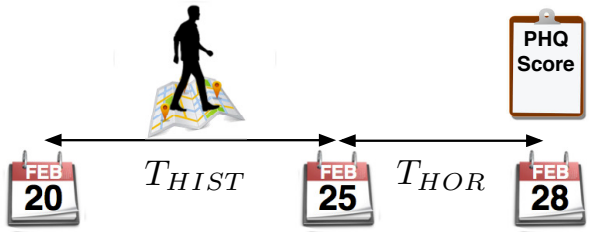


Figure 1: T_{HIST} is the duration of the interval over which we compute the metrics, whereas T_{HOR} represents how much in advance we compute the metrics. In other words, with respect to the example in the figure, we try to answer the following question: is there a relationship between the user mobility behavior from 20 February until 25 February and the PHQ score result obtained on day 28 February?

consequence, our technique can also be applied to monitor the PHQ score of a user that does not suffer from any form of depression (i.e., having a low PHQ score), and for an early depression diagnosis in case the condition of the user worsens.

OUR APPROACH

The key question of this paper, which is illustrated in Fig. 1, is whether the mobility behavior of an individual can give information about his/her current depressive state, which is quantified by a PHQ score. In order to answer this question, we first need to introduce the key definitions and notations that we use in this work. In the following we provide a formal definition of *mobility traces* and we define a set of *mobility metrics*, i.e., a set of statistical summaries characterizing the movement of individuals.

Formal Definition of Mobility Traces

We consider the *mobility trace* of a user as a sequence of stops and moves. This is a widely used definition of mobility trace (see for example [38]). A stop represents a geographic location in which the user spends a certain interval of time. Formally, we define a *stop place* (shortly, *place*) as a tuple:

$$Pl = \langle ID, t^a, t^d, C \rangle \,, \tag{1}$$

where ID is an identifier, t^a is the time of arrival, t^d is the time of departure, and C is a latitude-longitude pair. For a specific user we define the mobility trace $MT(t_1, t_2)$ for the time interval $[t_1, t_2]$ as the sequence of places visited by that user during that time interval:

$$MT(t_1, t_2) = (Pl_1, Pl_2, \ldots Pl_{N(t_1,t_2)}) \,, \tag{2}$$

where $N(t_1, t_2)$ is the total number of visited places in the interval $[t_1, t_2]$. The time references satisfy the following inequalities: $t_1^a \geq t_1$, $t_i^d < t_{i+1}^a$, $\forall\, i = 1, \ldots, N(t_1, t_2) - 1$, and $t_{N(t_1,t_2)}^d \leq t_2$.[1] The time gaps between places, i.e., the intervals $\left[t_i^d, t_{i+1}^a\right]$, correspond to periods of time in which the user is moving from one place to another. Whenever we do not specify the time interval of a mobility trace, we implicitly consider a time interval equal to the period of study.

[1]We use the subscript i to denote a parameter of the i-th place. For example, t_i^a represents the time of arrival of the place Pl_i.

Mobility Metrics

In this work, the mobility traces are used to compute a set of *mobility metrics*, which represent an aggregate information about the user movements. The considered metrics have been designed to capture behavioural patterns associated with depression such as reduced mobility [34] and, more generally, limited willingness of performing different activities, which usually involve movements to various locations [18]. It is quite interesting to note that the proposed metrics are also able to capture behavioral characteristics that might not be visible through traditional questionnaire-based methods. We consider the following mobility metrics, which are associated to a specific user and time interval $[t_1, t_2]$.

1) The total distance covered $D_T(t_1, t_2)$. We formally define this distance as follows:

$$D_T(t_1, t_2) = \sum_{i=1}^{N(t_1,t_2)-1} d(C_i, C_{i+1}) \,, \tag{3}$$

where $d(C_i, C_{i+1})$ is the geodesic distance between the two latitude-longitude pairs C_i and C_{i+1}.

2) The maximum distance between two locations $D_M(t_1, t_2)$. It represents the maximum span of the area covered in the interval $[t_1, t_2]$. More formally:

$$D_M(t_1, t_2) = \max_{i,j \in \{1,\ldots,N(t_1,t_2)\}} d(C_i, C_j) \,. \tag{4}$$

3) The radius of gyration $G(t_1, t_2)$. This metric is used to quantify the coverage area and is defined as the deviation from the centroid of the places visited in the interval $[t_1, t_2]$ [13]. We weight the contribution of each place by the time spent in that place. Let $T_i = t_i^d - t_i^a$ the time spent in the i-th place, $\tilde{T} = \sum_{i=1}^{N(t_1,t_2)} T_i$ the total time spent in different places, and C_{cen} the coordinates of the centroid of the places visited in the interval $[t_1, t_2]$, then:

$$G(t_1, t_2) = \sqrt{\frac{1}{\tilde{T}} \sum_{i=1}^{N(t_1,t_2)} T_i \cdot d(C_i, C_{cen})^2} \,. \tag{5}$$

4) The standard deviation of the displacements σ_{dis}. With "displacement" we refer to the distance between one place and the subsequent one. Let $D_{dis} = \frac{1}{N(t_1,t_2)-1} \sum_{i=1}^{N(t_1,t_2)-1} d(C_i, C_{i+1})$ the average displacement, then:

$$\sigma_{dis} = \sqrt{\frac{1}{N(t_1, t_2) - 1} \sum_{i=1}^{N(t_1,t_2)-1} \left(d(C_i, C_{i+1}) - D_{dis}\right)^2} \,. \tag{6}$$

5) The maximum distance from home $D_H(t_1, t_2)$. Among all the places visited by the user during the period of study, we assign the label "home" to the cluster in which the user was found most often at 02:00, 06:00 and 20:30 during weekdays. Let C_H the coordinates of such a cluster, then:

$$D_H(t_1, t_2) = \max_{i \in \{1,\ldots,N(t_1,t_2)\}} \{d(C_i, C_H)\} \,. \tag{7}$$

6) The number of different places visited $N_{dif}(t_1, t_2)$. Let $\mathbb{1}_{ij}$ the indicator function, which is equal to 1 if $ID_i = ID_j$, otherwise it is equal to 0. The we can formally define $N_{dif}(t_1, t_2)$ as:

$$N_{dif}(t_1, t_2) = \sum_{i=1}^{N(t_1,t_2)} \max\left\{1 - \sum_{j \neq i} \mathbb{1}_{ij}\ , 0\right\} . \quad (8)$$

7) The number of different significant places visited $N_{sig}(t_1, t_2)$. We assign the label "significant place" to each of the 10 most visited places among all the places visited by the user during the period of study. $N_{sig}(t_1, t_2)$ quantifies how many of these significant places are visited in the interval $[t_1, t_2]$. Let $ID_{s_1}, \ldots, ID_{s_{10}}$ the IDs of the significant places, then:

$$N_{sig}(t_1, t_2) = \sum_{j=1}^{10} \min\left\{\sum_{i=1}^{N(t_1,t_2)} \mathbb{1}_{is_j}\ , 1\right\} . \quad (9)$$

8) The routine index $R(t_1, t_2)$. We define this new metric with the goal of quantifying how different the places visited by the user during the time interval $[t_1, t_2]$ are with respect to the places visited by the user during the same time interval in other days. To formally define this metric we need to introduce some further notation. Given the mobility trace $MT(t_1, t_2)$, we define the augmented mobility trace $\overline{MT}(t_1, t_2)$ as the mobility trace in which the gaps between places are filled with "mobility places", and we denote by ID_m the IDs of these places.

We define the difference function $f^t(\overline{MT}_1, \overline{MT}_2)$ between two (augmented) mobility traces $\overline{MT}_1 = \overline{MT}(t_1, t_2)$ and $\overline{MT}_2 = \overline{MT}(t_3, t_4)$ at time instant t as:

$$f^t(\overline{MT}_1, \overline{MT}_2) = \begin{cases} 0 & \text{if } t \notin [t_1, t_2] \text{ or } t \notin [t_3, t_4] \\ & \text{or } ID^{1,t} = ID^{2,t} \\ 1 & \text{otherwise} \end{cases} \quad (10)$$

where we use the notation $ID^{i,t}$ to denote the ID of the place belonging to the i-th trace and visited by the user at time t. Basically, for two overlapping mobility traces, $f^t(\overline{MT}_1, \overline{MT}_2)$ checks whether the place at time instant t is different for the two traces.

Now we can define the average difference $f(\overline{MT}_1, \overline{MT}_2)$ between $\overline{MT}_1$ and $\overline{MT}_2$ as the fraction of the overlapping time in which the places in $\overline{MT}_1$ are different from the places in $\overline{MT}_2$:

$$f(\overline{MT}_1, \overline{MT}_2) = \frac{1}{\bar{t}_2 - \bar{t}_1} \int_{\bar{t}_1}^{\bar{t}_2} f^t(\overline{MT}_1, \overline{MT}_2) \mathrm{d}t\ , \quad (11)$$

where $\bar{t}_1 \triangleq \max\{t_1, t_2\}$ and $\bar{t}_2 \triangleq \min\{t_3, t_4\}$, i.e., $[\bar{t}_1, \bar{t}_2]$ represents the overlapping period.

We define the translated augmented mobility trace $\overline{MT}^t(t_1, t_2)$ as the mobility trace that is obtained from $\overline{MT}^t(t_1, t_2)$ by translating all the times in advance by an amount equal to t.

Let t_{init} and t_{fin} be the time of arrival and the time of departure of the first and last places visited by the user in the period of study, respectively. Let T be the length of one day, n_1 the number of (whole) days elapsed from t_{init} to t_1, and n_2 the number of (whole) days elapsed from t_2 to t_{fin}. Finally, we define the routine index as:

$$R(t_1, t_2) = \frac{1}{n_2 + n_1} \left[\sum_{i=1}^{n_1} f\Big(\overline{MT}(t_{init}, t_{fin}), \overline{MT}^{iT}(t_1, t_2)\Big) + \sum_{i=1}^{n_2} f\left(\overline{MT}(t_{init}, t_{fin}), \overline{MT}^{-iT}(t_1, t_2)\right) \right] . \quad (12)$$

In words, the routine index $R(t_1, t_2)$ represents the average difference between the mobility behavior of the user in $[t_1, t_2]$ and the mobility behavior of the same user in other days in the same (daily) time interval. Notice that $0 \leq R(t_1, t_2) \leq 1$.

Finally, we would like to remark the fact that these metrics are able to capture also the absence of movement, which might be associated to specific mental health states, for example, a person that stays at home for subsequent days or that do not go far from his/her home because of his/her depressive mood.

MOODTRACES APPLICATION

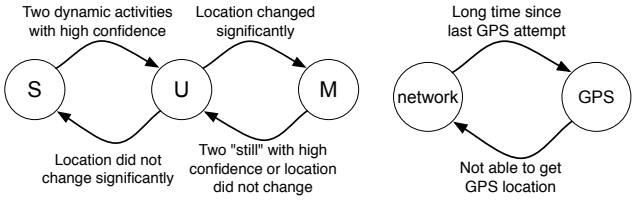
Overview

MoodTraces is an Android application for mobile phones that automatically samples phone's sensors to retrieve the current location, which is represented by a time reference, a longitude value, and a latitude value. Additional information about the phone usages and user activities extracted using the Android API are also collected, but they are not analyzed directly in this work, given its specific focus on mobility pattern analysis. Activity information is used to optimize the sampling process as discussed below. In addition to sensor based data, MoodTraces collects the answers that the users provide to daily questionnaires. It is worth noting that this application collects information about the user mood only to get the ground-truth data. All the collected data is sent via a secure transmission protocol to a secure server located at our institution. We exploit the asynchronous delay-tolerant data transfer strategy of the ES Data Manager Library [32, 21] to transmit the collected data in an energy-efficient way and to avoid any cost on participating users.

Location Collection Process

One of the key challenges for mobile sensing applications is their energy use and impact on battery life. Among all the data collected by MoodTraces, GPS-based location data are by far the most expensive in terms of energy consumption [3, 26]. In order to limit the energy impact of the location data collection, we exploit the collected activity data[2] to trigger location sampling only when the user moves from one place

[2]We collect activity data with a sampling rate of 1 minute. However, to conserve battery, activity reporting stops when the device is "still" for an extended period of time, and it resumes once the device moves again. This only happens on devices that are equipped with a motion trigger sensor, a sensor that automatically wakes the device to notify when significant motion is detected.



(a) State-based sampling rate. (b) State-based provider.

Figure 2: State machines to select the location sampling rate and the location provider.

to another. Our approach is represented by the state machine in Fig. 2a.

The three states that we consider and the corresponding location sampling rates adopted are described in the following.

Static (S): in this state we never sample location data. This state corresponds to all the situations in which we are confident that the user is remaining in the same place (e.g., during working hours a user may remain for an extended period inside the building in which he/she works; it is not necessary to sample location data continuously in this case).

Moving (M): in this state we sample location data with a sampling rate of $T_s = 5$ minutes. This state corresponds to the situations in which we are confident that the user is moving from one place to another (e.g., when the user moves from his/her working place to his/her home).

Undecided (U): as soon as we enter this state we get the user location, and then we get another user location after $T_s = 5$ minutes. Based on the distance between the two locations, we decide whether to transit to the **S** or to the **M** state. This state corresponds to all the situations in which we are not sure whether the user is moving from one place to another (e.g., the phone may detect that the user is walking, but we are not sure whether the user is walking from his/her working place to his/her home, or whether the user is walking from one office to another office at his/her workplace). When MoodTraces is installed, and whenever the smartphone is rebooted, we initialize the state to **U**.

We transit from **S** to **U** whenever two consecutive activity samples indicate that the user is performing a "dynamic activity" (i.e., in vehicle, on bicycle, on foot, running, walking) with a confidence level larger than $P_{th} = 50\%$. We transit from **M** to **U** whenever either 1) we get two consecutive location samples whose distance is smaller than $d_{th} = 250$ m; or 2) two consecutive activity samples indicate that the user is "still" with a confidence level larger than $P_{th} = 50\%$. Finally, from **U** we transit to **M** if the two location samples that we collect inside the **U** state are more than $d_{th} = 250$ m apart, otherwise we transit to **S**.

In addition to the sampling rate, the choice of the "location provider" has also an important impact on the energy consumption. Indeed, in Android it is possible to sample location data by using either the "GPS provider", which is accurate but energy-demanding and limited in indoor environment, or the "network provider", which provides less accuracy but is much less energy-demanding than the GPS provider. In order to trade-off accuracy with energy consumption, MoodTraces implements the simple state machine represented in Fig. 2b. Each state represents the provider that will be used in case MoodTrace requests a location sample. By default, in order to have accurate location we use the GPS provider. If we do not receive any location update in 1 minute after a request (e.g., the GPS receiver is not able to track the GPS signal because the user is moving using a subway), then we switch to the network provider. We switch back to the GPS provider after 30 minutes we entered the network provider state. MoodTraces subscribes also to a "passive location provider", meaning that it passively receives location updates when other applications or services request them, independently of the states of the two state machines represented in Fig. 2.

Questionnaire Collection Process

The PHQ-9 is a widely used and extensively studied 9-item questionnaire for assessing and monitoring depression severity [16, 18, 15]. A systematic review of the PHQ-9 is reported in [18]. It shows that, in 19 studies, the PHQ-9 achieves sensitivity values ranging from 0.77 to 0.88, and specificity values ranging from 0.88 to 0.94. As a comparison, [4] identifies 12 studies that address the performance of the Hospital Anxiety and Depression Scale (HADS) and reports a mean sensitivity of 0.83 and a mean specificity of 0.79, and similar sensitivity and specificity values are achieved by the General Health Questionnaire (GHQ).

The PHQ-8 is an 8-item questionnaire that omits the ninth item of the PHQ-9 questionnaire. [18] reports correlation values of 0.998 and 0.997 in two different studies between PHQ-9 scores and PHQ-8 scores, and [16] suggests to use the PHQ-8 instead of the PHQ-9 in clinical samples in which the risk of suicidality is felt to be extremely low or if data is being gathered in a self-administered fashion. Since both of these conditions are met in our study, and because of its brevity and its structure, a PHQ-8 based daily questionnaire is used in MoodTraces to assess the presence of a depressed mood and how this varies in time.

The PHQ-8 is composed of 8 questions: for each of them the participant is asked to report the frequency of the occurrence of a specific depressive symptom during the last 2 weeks. Based on the provided answers, each question is associated with a score between 0 and 3. The final *PHQ score* is computed by adding the contributions of all questions; hence, it ranges from 0 to 24. Cut points of 5, 10, 15 and 20 are considered to diagnose mild, moderate, moderately severe and severe levels of depressive symptoms, respectively [17, 16, 18, 15]. As stated in [18], these categories were chosen both for pragmatic reasons, in that the cut points of 5, 10, 15, and 20 are simple for clinicians to remember and apply, and for empirical reasons, in that using different cut points did not noticeably change the associations between increasing PHQ-9 severity and measures of construct validity. These thresholds has been assessed with respect to the diagnoses provided by independent structured mental health professional in a sample of 580 patients. For more details we refer the interested reader

to [18]. Given the fact that the accuracy of the PHQ-8 test is comparable to the accuracy of other tests, the main reason that led us to the adoption of the PHQ-8 test is its brevity and its structure that allow for the design of a simple daily "yes-no" questionnaire in which the user is asked whether each of the depressive symptoms occurred during that day. In this way, the PHQ score for a certain day can be computed by looking at the answers provided during the previous two weeks and computing the frequencies for each depressive symptom.

During each day the questionnaire is available from 16:00 until 2:00 of the subsequent day.[3,4] A notification at 16:00 informs the user that the questionnaire is available, following by other notifications at 20:00 and 23:30 in case the user has not completed the questionnaire by that time. The questionnaire takes less than 1 minute to complete and can be completed in multiple sessions. Notice that we collect also the answers to the questionnaires that are only partially completed.

According to the best practice in survey design [30, 35, 33] we collect the completion time for each provided answer and the questionnaire includes a trap question (i.e., a question having a known answer), asking whether the user is at home or at work. The validity of the answer is checked throughout the collected location data. The answer completion times and the trap questions allowed us to verify the quality of the collected data and to discard non reliable questionnaires.

Recruitment Process

MoodTraces has been available for the general public for free in the GooglePlay Store [27] since September 3, 2014, and, at the time of writing of this paper, it is still available. The study presented in this paper refers to the data collected from September 3, 2014, to June 14, 2015. In this period we had a total of 184 installs and 46 users had MoodTraces running in their phones at June 14, 2015.

At the beginning MoodTraces has been used only by few researchers with the goal of fixing bugs, tuning parameters, and adding features to improve the user experience. We have advertised our study starting from the end of November, exploiting different resources: academic mailing lists, Twitter, Facebook, Reddit, and charities. To promote the application and to give incentives to the users to reply to the questionnaires, we committed to select (through a lottery) one winner of a Nexus 5 mobile phone and five winners that have received a 10 pounds Amazon voucher each among all the participants that have completed the daily questionnaire at least 50 times in a two-month span. Finally, we remark that we received the full approval of the Ethics Review Board of our institution before starting the recruitment process, and that all the documentation, including the consent form and the information sheet for the participants, are available on request.

DATA PROCESSING

In this section we describe how we processed the data to calculate the PHQ score and the mobility metrics for each user on a daily basis. It is worth noting that this procedure was discussed and approved with colleagues that are leading experts in suicidal prevention studies and in clinical practice.

PHQ Score Computation

For each user, we exploit the answers to the daily questionnaires that the user provides in order to compute his/her PHQ score on a daily basis.

First, in order to improve the reliability of the collected answers, we void 1) the questionnaires for which the trap question has been replied erroneously, and 2) the questionnaires that are replied too quickly, as in [35] and [33] we identify these questionnaires as the 10% questionnaires with lowest Speeder Index. To compute the Speeder Index we first calculate the median completion time for each answer across all the questionnaires of all the users. Then, to each answer we assign a value of 1 if the completion is at least equal to the median, otherwise we assign a value equal to the ratio between the answer completion time and the median completion time. Finally, for each questionnaire the Speeder Index is computed as the average value of all the provided answers.

Second, to compute the PHQ score on a given day x we retrieve all the questionnaires that have been submitted from day $x-13$ until day x and we count how many times each depressive symptom occurred in this time interval. Since some answers are missing for some days (e.g., because the users did not submit the questionnaire or submitted an incomplete questionnaire), we deal with missing answers by using a linear interpolation to compute the occurrence frequency of the corresponding symptom. For example, if in a 2-week period the user replies 12 times to a specific answer indicating that the corresponding symptom occurred 6 times, then we assign an occurrence frequency for that symptom equal to 7 days over 14. The linear interpolation is adopted only if the user replies to at least 80% of the answers, otherwise we do not compute a PHQ score for the corresponding day. Our approach follows the current practices in the area, as it is possible to observe in previous other studies based on PHQ-9, in which missing values were replaced with the mean value of the remaining items if the number of missing items was below 20% [15, 18].

Third, based on the occurrence frequency we assign the appropriate score to each depressive symptom (see [18]), and the PHQ score is computed as the sum of the scores of all depressive symptoms.

Finally, to remove cyclicity effects, from the computed PHQ score we subtract the average PHQ score obtained by that user in that day of the week. In order to simplify the presentation, in the remainder of the paper we use the terms PHQ score to indicate the deviations from the average values (with the exception of Fig. 3 that shows the histograms of the real PHQ scores).

Mobility Traces and Mobility Metrics Computation

The location collection process adopted by MoodTraces, in addition of being energy-efficient, is suitable to detect stop

[3] We chose to make the survey available from late afternoon because it asks whether depressive symptoms have occurred in the current day, hence it cannot be taken at the beginning of the day.

[4] All the times are in the user local time.

places. Indeed, when the user remains in a place for an extended period of time MoodTraces enters the state **S** as discussed above. If the user performs some minor movements (e.g., walking from one office to another office at his/her workplace), MoodTraces enters the **U** state and then goes back to **S**, because it realizes that the user did not move more than $d_{th} = 250$ m away from the previous place. On the other hand, if the user performs major movements then MoodTraces enters the **M** state after transiting on the **U** state. Hence, by looking at the evolution of the state machine we can identify the intervals in which the user stops in a place from the periods in which the user moves.

Since we are not interested in identifying two locations that are in close proximity as different places (e.g., two different offices in the same building), we register the departure from a place if and only if both the transition $\mathbf{S} \rightarrow \mathbf{U}$ *and* the transition $\mathbf{U} \rightarrow \mathbf{M}$ occur; in this case the departure time T^d is set equal to the time in which the transition $\mathbf{S} \rightarrow \mathbf{U}$ occurs. Analogously, we register the arrival in a place if both the transition $\mathbf{M} \rightarrow \mathbf{U}$ *and* the transition $\mathbf{U} \rightarrow \mathbf{S}$ occur, and the arrival time T^a is set equal to the time in which the transition $\mathbf{M} \rightarrow \mathbf{U}$ occurs. As for the coordinates of the place, we set them equal to the centroid of all locations collected in the time interval $[T^a, T^d]$ having an accuracy of at least $d_{acc} = 200$ m. Since the threshold $d_{th} = 250$ m is used to determine the transition from the state **U**, the geographic area corresponding to a place can be approximately quantified by a circle centered in the coordinates and having a radius of $d_{th} = 250$ m.

Similarly to the case of the questionnaire collection process, also the location collection process is prone to missing values. Let $[\,\underline{t}, \bar{t}\,]$ a time interval in which location data are not available due to external factors (e.g., the phone is switched off or the location services are disabled). We deal with this situation by assigning the time interval $[\,\underline{t}, \bar{t}\,]$ to a stop place if the last location collected before the time $\underline{t}$ is in close proximity to the first location collected after the time $\bar{t}$, i.e., if their minimum distance is below $d_{th} = 250$ m (this is the case, for example, of a user that switches off his/her phone before sleeping, and switches it on the next morning). Otherwise that interval will be associated to user movements (this is the case, for example, of a phone with a fully depleted battery during a trip). However, to avoid to use low quality data, we do not compute the mobility trace (and, as a consequence, the mobility metrics) associated to a time interval $[t_1, t_2]$ such that the sum of the periods in which location data are missing is larger than half the target interval, i.e., $\frac{t_2 - t_1}{2}$.

Finally, notice that many places in which a user stops correspond to geographic locations that the user visits more than once during the period of study (e.g., home, work, bus stop, etc.). It is convenient to use the same ID and coordinates for the places corresponding to the same geographic location. To achieve this goal we adopt the following simple clustering algorithm:

1. Find Pl_i and Pl_j such that $ID_i \neq ID_j$ and $d(C_i, C_j) = \min_{n \neq m}\{d(C_n, C_m)\}$, where $d(C_n, C_m)$ denotes the distance between the coordinates of the places Pl_n and Pl_m;

2. If $d(C_i, C_j) < d_{acc}$ then assign $ID_j \leftarrow ID_i$, $C_i \leftarrow y$, and $C_j \leftarrow y$, where y is the centroid of all locations collected in the time intervals $[T_i^a, T_i^d]$ and $[T_j^a, T_j^d]$ having an accuracy of at least $d_{acc} = 200$ m;

3. Iterate **1.** and **2.** until convergence.

At each step the above algorithm finds, among all the stop places, the two places that are closest to each other. If they are in close proximity (closer then $d_{acc} = 200$ m) they are given the same ID and their centroid is recomputed, otherwise the algorithm terminates. Convergence is guaranteed because the algorithm iterates a maximum of $N(t_1, t_2) - 1$ times, after which all places are clustered under the same ID. Once we get the mobility trace of a user, the mobility metrics for a certain time interval $[t_1, t_2]$ can easily be computed adopting Eqs. (3) to (9) and (12). Finally, as we did for the PHQ scores, in order to remove cyclicity effects, from each computed mobility metrics we subtract the average mobility metric obtained by that user in that interval in that day of the week.

Combining Mobility Metrics and PHQ Scores

For each user, we compute the time series of the PHQ scores and of the mobility metrics,

$$\begin{aligned} \mathbf{PHQ} &= \left(PHQ^1, PHQ^2, \ldots, PHQ^N\right) \\ \mathbf{D_T}(T_{HIST}, T_{HOR}) &= \left(D_T^1(T_{HIST}, T_{HOR}), \ldots\right) \\ \mathbf{D_M}(T_{HIST}, T_{HOR}) &= \left(D_M^1(T_{HIST}, T_{HOR}), \ldots\right) \\ &\vdots \end{aligned} \tag{13}$$

where N is the total number of days the user utilized MoodTraces in the period of study, PHQ^i is the PHQ score associated to the i-th day since the user downloaded MoodTraces, and D_T^i, D_M^i, etc., are the mobility metric associated to the PHQ score PHQ^i. We associate each PHQ score to the mobility metrics through the use of the *time history* T_{HIST} and the *time horizon* T_{HOR} parameters, which define the interval over which the mobility metrics are computed, as illustrated in Fig. 1. The parameter T_{HIST} represents the length of the interval over which the mobility metric is computed, whereas T_{HOR} represents how much in advance with respect to the PHQ score the mobility metric is computed. Hence, for each mobility metric we must compute one sequence of values for each considered combination of the parameters T_{HIST} and T_{HOR}.

We define the vector $\left[PHQ^i, D_T^i(T_{HIST}, T_{HOR}), D_M^i(T_{HIST}, T_{HOR}), \ldots\right]$ as the i-th instance. For each user, we remove the instances for which either the PHQ score or the mobility metrics cannot be computed. Finally, we exclude from the study all the users having less than $N_{inst} = 20$ instances. As a consequence of this filtering procedure, the evaluation presented in this paper is generated from a dataset of 28 users, which is briefly described in the following.

The final dataset, generated from the filtering procedure described above, includes 28 users: 15 male and 13 female.

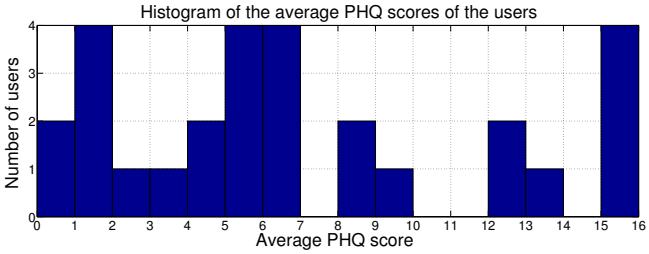


Figure 3: Histograms of the average PHQ score of the users.

Many of these are linked to the academic environment (7 students, 3 PhD students, and 7 researchers/lecturers), but there are also individuals working in the private sector, artists, and retired people. The average age is 31 years old. On average each user has been monitored for 71 days, and we were able to compute 63 PHQ scores for each user. Figure 3 shows the histogram of the average PHQ scores of the users. Most of the users do not suffer, on average, of depressive problems (average PHQ score below 5), or they suffer of mild depressive symptoms (average PHQ score from 5 to 10). However, there are also some users suffering of moderate (average PHQ score from 10 to 15) and moderately severe (average PHQ score from 15 to 20) depressive symptoms. These users experience also peaks on the daily PHQ score overpassing the severe depression cut point value 20.

EVALUATION

In this Section we analyze the data collected by means of MoodTraces. We first analyze how the mobility metrics and the PHQ score of each user jointly vary in time, proving that there exists a significant correlation among them. Then we train and test personalized regression and classification models for each user to investigate whether it is possible to predict changes in the PHQ score from variations in the mobility metrics.

Correlation Analysis

We exploit the sequences defined by Eq. (13) to analyze the correlation[5] between each mobility metric and the PHQ score, for each user and for different values of the time history parameter T_{HIST}, which is measured in days. We also compute the p-value associated to each correlation value, representing the probability of getting a correlation as large as the observed value by random chance (i.e., when the true correlation is zero). For this analysis we set the time horizon $T_{HOR} = 0$ days, i.e., the last day of the time interval over which we compute a specific mobility metric is equal to the day in which we compute the associated PHQ score.

Table 1 shows the average values (among the users) of the absolute correlations and p-values, for each mobility metric and for $T_{HIST} = 1$ day (i.e., the mobility metrics are computed over the same day of the corresponding PHQ score) and $T_{HIST} = 14$ days (i.e., the mobility metrics are computed considering a time span of two weeks before the day of the corresponding PHQ score). We consider the absolute value of the correlation because it represents an ordinal measure of how strong the relationship between the mobility metric and the PHQ score is, hence it is reasonable to compute its average (this does not hold for the "signed correlation", because strong negative and positive dependencies would compensate each other by computing the average).

[5]In this work we consider the Pearson correlation [40], which is usually adopted to quantify linear dependences between variables.

Mobility metric	Average abs. correlation		Average p-value	
	$T_{HIST} = 1$	$T_{HIST} = 14$	$T_{HIST} = 1$	$T_{HIST} = 14$
D_T	0.159	0.402	0.401	0.095
D_M	0.152	**0.432**	0.425	**0.069**
G	0.160	0.343	0.422	0.197
σ_{dis}	0.147	0.417	0.431	0.088
D_H	0.199	0.358	0.297	0.168
N_{dif}	0.191	0.360	0.335	0.157
N_{sig}	0.201	0.336	0.385	0.181
R	**0.227**	0.368	**0.262**	0.138

Table 1: The averages of the absolute values of the correlations and of the p-values for different mobility metrics, for $T_{HIST} = 1$ day and $T_{HIST} = 14$ days.

For $T_{HIST} = 1$ the average correlations[6] range from 0.147 (associated to the metric σ_{dis}) to 0.227 (associated to the metric R). Though these values are significantly different from 0, since the number of instances for each user is not extremely large the corresponding average p-values are quite large. Indeed, usually a correlation value is considered significant only if the corresponding p-value is smaller than the significance level $\alpha = 0.05$, but in our case the minimum p-value for $T_{HIST} = 1$ is 0.262. For $T_{HIST} = 1$ "low level" metrics such as the total distance covered D_T or the maximum distance between two locations D_M have a smaller average correlation than the metrics that capture semantic information about the visited places, such as the number of different significant places visited N_{sig} or the routine index R. Interestingly, this situation is reversed for $T_{HIST} = 14$. This suggests that low level distance-based metrics such as D_T and D_M require the observation of the user mobility behavior for longer time intervals than metrics that incorporate some semantic about the visited places. However, for sufficiently long time intervals, they can provide stronger clues about the depressive state of the user. With the increase of T_{HIST} from 1 to 14 days the average correlation for each mobility metrics increases (ranging now from 0.336 for N_{sig} to 0.432 for D_M) and, as a consequence, the corresponding p-values decrease.

This aggregate analysis summarized in Table 1 provides interesting insights about the correlation between PHQ scores and mobility metrics, but it does not provide a complete understanding of the strength of the correlation at individual level. For this reason, we now investigate the correlations and p-values in a non-aggregate form, by plotting the histograms of their values for $T_{HIST} = 1$ and $T_{HIST} = 14$. Due to space constraints, it is not possible to show the histograms for all mobility metrics. Hence, we choose to con-

[6]To simplify the presentation, in the remainder of this section we use "average correlation" instead of "average of the absolute values of the correlations".

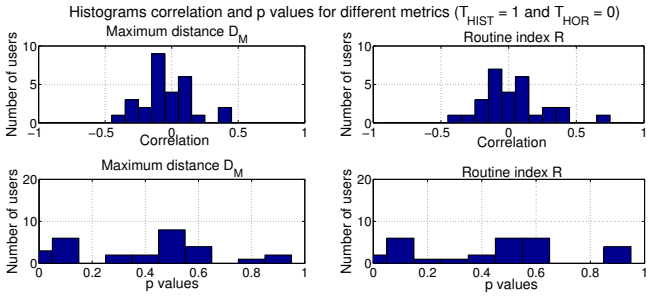


Figure 4: Histograms of the correlation and of the p-values for $T_{HIST} = 1$ day.

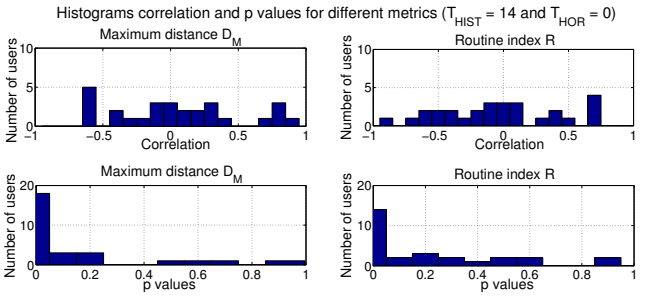


Figure 5: Histograms of the correlation and of the p-values for $T_{HIST} = 14$ day.

sider only the maximum distance between two locations D_M, whose average correlation is the second worst of all metrics for $T_{HIST} = 1$ and the best for $T_{HIST} = 14$, and the routine index R that, on the other extreme, has the best average correlation for $T_{HIST} = 1$. The histograms of the other mobility metrics are very similar to the histograms of the two selected metrics.

Fig. 4 shows the histograms of the correlation and of the p-values for $T_{HIST} = 1$ day. The histograms of the correlation associated to D_M (top-left subfigure) and to R (top-right subfigure) are distributed closely around 0, and the associated p-values (bottom-left and bottom-right subfigures) are quite large: there are only 3 users for D_M and 2 users for R having a p-value lower than $\alpha = 0.05$, corresponding to the first bin of the p-value histogram. Fig. 5 shows that for $T_{HIST} = 14$ days the correlation values for both mobility metrics are distributed more uniformly. It is important to remark that different users react differently to changes in their moods, for example an increase in the PHQ score is associated to smaller travelled distances for some users (those for which the correlation between PHQ score and D_M is negative) and larger travelled distances for others (those for which the correlation is positive). This suggests that personalized models, instead of general ones, should be used to monitor the depressive state of an individual using his/her mobility traces. Quite interestingly, Fig. 5 shows that there are 18 users for D_M and 14 users for R for which the p-value is lower than $\alpha = 0.05$, meaning that the corresponding correlation can be considered significant.

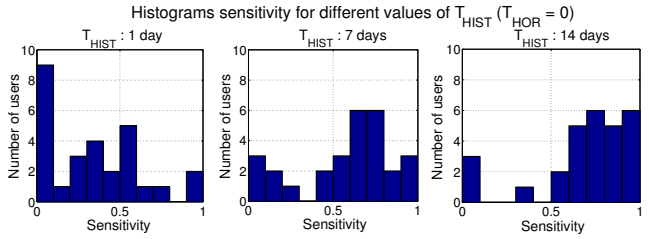


Figure 6: Histogram of the sensitivity, for $T_{HIST} = 1$, 7, and 14 days.

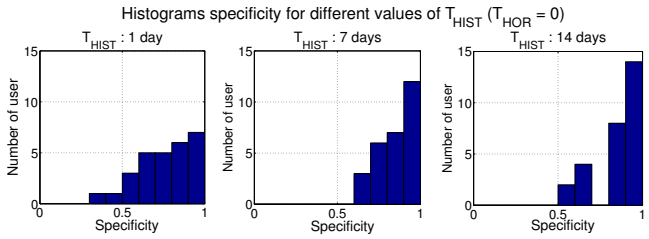


Figure 7: Histogram of the specificity, for $T_{HIST} = 1$, 7, and 14 days.

Prediction Analysis

The fundamental question that has driven this study is whether it is possible to monitor and diagnose a depressed mood by looking at the mobility trace collected from a smartphone of an individual. To answer to this question we set up the following analysis with the collected data. For each user and for each instance we compute a "label" in the following way: the label is equal to 1 if the PHQ score is larger that the average PHQ score of that user plus one standard deviation, otherwise the label is equal to 0. A label equal to 1 corresponds to the situation in which the user has a a PHQ score that is significantly larger than its usual value. Our goal is to develop models that are able to detect this situation. For each user, we train and test a personalized Support Vector Machine (SVM) classifier with a Gaussian radial basis function kernel [14]. In order to fully exploit the limited number of instances available for each user, we adopt a leave-one-out cross validation approach for training and testing. Moreover, for each training test we adopt again leave-one-out cross validation in order to optimize the value of the SVM penalty parameter C. We vary such a parameter using the exponentially growing sequence $C = 2^{-5}, 2^{-3}, \ldots, 2^{5}$.

To quantify the performance of the classification models we consider two metrics: 1) the *sensitivity* (or true positive rate), i.e., the fraction of 1 labels that are correctly classified; and 2) the *specificity* (or true negative rate), i.e., the fraction of 0 labels that are correctly classified. In our first analysis we show the histograms of the sensitivity and specificity, for different values of T_{HIST} and for $T_{HOR} = 0$. The results are shown in Figs. 6 and 7. As T_{HIST} increases, the distributions of the values of both the sensitivity and the specificity move closer to 1. On one extreme, with $T_{HIST} = 1$ day many of the trained models achieve low sensitivity and specificity performances.

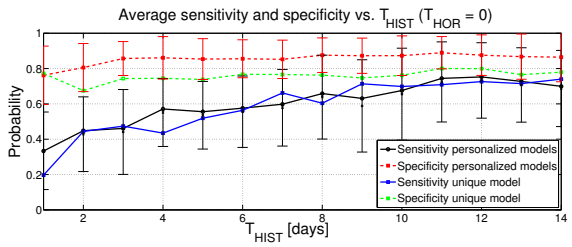


Figure 8: Average sensitivity and specificity values vs. T_{HIST}, for $T_{HOR} = 0$ days.

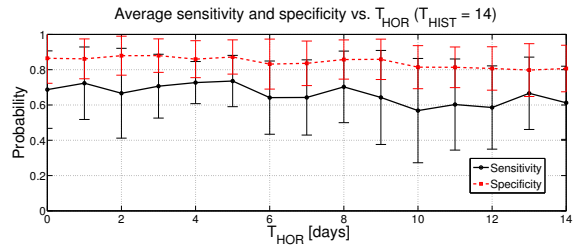


Figure 9: Average sensitivity and specificity values vs. T_{HOR}, for $T_{HIST} = 14$ days.

On the other extreme, with $T_{HIST} = 14$ days most of the trained models achieve very large sensitivity and specificity values. This means that, for most of the users, these personalized models are able to detect periods in which the users experience an unusual depressed mood (this is linked to the sensitivity), and at the same time they generate very few false alarms (this is linked to the specificity). Notice that for all the users the specificity values are larger than the sensitivity values: this is not surprising because we are trying to detect unusual PHQ scores for each users, hence the datasets are unbalanced (they contain more 0 labels than 1 labels) and, as a consequence, in order to minimize the mis-classification probability, the trained SVM models are biased toward the predictions of the 0 labels.

Next we investigate how the average (among the users) sensitivity and specificity values vary with the time interval T_{HIST}, for $T_{HOR} = 0$. The results are showed in Fig. 8. Average sensitivity and specificity values are associated with a confidence bar, which covers an interval of two standard deviation around the average value. Fig. 8 shows also the sensitivity and specificity value of a generic SVM model, which is trained and tested with the same modalities of the personalized models, but it exploits all the data collected from all the 28 users. Both the average sensitivity and specificity of the personalized models and the sensitivity and specificity of the unique model rise with the increase of T_{HIST}, and reach large values for $T_{HIST} = 14$ days. We notice that personalized models achieve better performance that the unique general model, confirming the insights derived from the correlation analysis. However, the good performance of the unique general model demonstrates the feasibility of this alternative approach, which has the advantage that it does not require the collection of labeled data from each user for training purposes, and this might increase the actual usability and acceptance of the proposed prediction tools. This represents an interesting trade-off to explore. For example, a model trained on all the data can be adopted when personalized data are not available, e.g., when a user installs an application relying on these mechanisms for the first time.

In our final analysis we fix the value of the time interval T_{HIST} to 14 days and we vary the parameter T_{HOR}; the average sensitivity and specificity values obtained (along with the corresponding confidence bars) are represented in Fig. 9. As expected, the average sensitivity and specificity values decrease as T_{HOR} increases. This is not surprising since T_{HOR} represents the prediction horizon (see Fig. 1), i.e., how much in advance we try to predict the change in the PHQ score of the user. However, it is surprising that the decay is very slow. Indeed, the average sensitivity and specificity are quite large even if we try to predict changes in the PHQ score 14 days in advance, which is for example the time span over which depressive symptoms are evaluated in the PHQ-8 test (and in many other standard test to diagnose depression). This means that the considered mobility metrics might identify early signs that can be exploited for an early detection of depressed moods.

CONCLUSIONS

In this work we have demonstrated that it is possible to observe a significant correlation between mobility patterns and depressive mood using data collected by means of smartphones. We have also shown that it is possible to develop inference algorithms as a basis for unobtrusive monitoring and prediction of depressive mood disorders.

We believe that this work represents an important starting point in this area and can be used as a basis for more application-oriented projects in the area of digital mobile interventions. For example, the techniques for automatic detection of depressive state presented in this work can be used for building systems for automatic interventions, both through technology (e.g., phone calls from healthcare officers) or traditional physical interactions. Moreover, the focus of this paper is on a specific modality, i.e., GPS location, but the results of this work can be indeed exploited to build a more refined system based on the analysis of data extracted by means of other sensors, such as accelerometers, and other sources of information, such as call and SMS logs. Finally, we plan to use the current application (or an extended version) in future studies that will focus on specific populations, such as individuals that have been clinically diagnosed as depressed.

ACKNOWLEDGEMENTS

The authors would like to thank all the participants of this study. Prof. Rory O'Connor and Dr Paul Patterson provided an invaluable contribution to the study, in particular for the design of the experiments. This work was supported through the EPSRC grant "Trajectories of Depression: Investigating the Correlation between Human Mobility Patterns and Mental Health Problems by means of Smartphones" (EP/L006340/1) and partially by the "LASAGNE" Project, Contract No. 318132 (STREP), funded by the European Commission.

REFERENCES

1. Bauer, G., and Lukowicz, P. Can smartphones detect stress-related changes in the behaviour of individuals? In *Proceedings of PERCOM '12 Workshops* (March 2012), 423–426.
2. Baumann, P., Kleiminge, W., and Santini, S. How Long Are You Staying? Predicting Residence Time from Human Mobility Traces. In *Proceedings of MobiCom '13* (2013), 231–234.
3. Ben Abdesslem, F., Phillips, A., and Henderson, T. Less is More: Energy-efficient Mobile Sensing with Senseless. In *Proceedings of MobiHeld '09* (2009), 61–62.
4. Bjelland, I., Dahl, A. A., Haug, T. T., and Neckelmann, D. The validity of the Hospital Anxiety and Depression Scale. *Journal of Psychosomatic Research 52*, 2 (2001), 69–77.
5. Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., and Pentland, A. S. Pervasive stress recognition for sustainable living. In *Proceedings of PERCOM '14 Workshops* (March 2014), 345–350.
6. Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., and Mohr, D. C. Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research 13*, 3 (August 2011).
7. Campbell, A. T., Eisenman, S., Lane, N., Miluzzo, E., Peterson, R., Lu, H., Zheng, X., Musolesi, M., Fodor, K., and Ahn, G.-S. The rise of people-centric sensing. *IEEE Internet Computing 12*, 4 (July 2008), 12–21.
8. Donker, T., Petrie, K., Proudfoot, J., Clarke, J., Birch, M.-R., and Christensen, H. Smartphones for smarter delivery of mental health programs: a systematic review. *Journal of Medical Internet Research 15*, 11 (2013).
9. European Depression Association. IDEA: Impact of Depression at work in Europe Audit, September 2012.
10. Frost, M., Doryab, A., Faurholt-Jepsen, M., Kessing, L. V., and Bardram, J. E. Supporting Disease Insight Through Data Analysis: Refinements of the Monarca Self-assessment System. In *Proceedings of UbiComp '13* (2013), 133–142.
11. Grüenerbl, A., Oleksy, P., Bahle, G., Haring, C., Weppner, J., and Lukowicz, P. Towards smart phone based monitoring of bipolar disorder. In *Proceedings of mHealthSys '12* (2012).
12. Grüenerbl, A., Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., Mayora, O., Haring, C., and Lukowicz, P. Using Smart Phone Mobility Traces for the Diagnosis of Depressive and Manic Episodes in Bipolar Patients. In *Proceedings of AH '14* (2014).
13. Hoteit, S., Secci, S., Sobolevsky, S., Pujolle, G., and Ratti, C. Estimating Real Human Trajectories through Mobile Phone Data. In *Proceedings of MDM '13*, vol. 2 (June 2013), 148–153.
14. Hsu, C.-W., Chang, C.-C., and Lin, C.-J. A Practical Guide to Support Vector Classification. Tech. rep., Department of Computer Science, National Taiwan University, 2003.
15. Kocalevent, R.-D., and Hinz, A. Standardization of the depression screener patient health questionnaire (PHQ-9) in the general population. *General Hospital Psychiatry 35*, 5 (2013), 551–555.
16. Kroenke, K., and Spitzer, R. L. The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals 32*, 9 (September 2002), 509–515.
17. Kroenke, K., Spitzer, R. L., and Williams, J. B. W. The PHQ-9: Validity of a brief depression severity measure. *J Gen Intern Med. 16*, 9 (2001), 606–613.
18. Kroenke, K., Spitzer, R. L., Williams, J. B. W., and Löwe, B. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. *General Hospital Psychiatry 32*, 4 (2010), 345–359.
19. Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. T. A survey of mobile phone sensing. *IEEE Communications Magazine 48*, 9 (2010), 140–150.
20. Lathia, N., Pejovic, V., Rachuri, K. K., Mascolo, C., Musolesi, M., and Rentfrow, P. J. Smartphones for large-scale behavior change interventions. *IEEE Pervasive Computing*, 3 (2013), 66–73.
21. Lathia, N., Rachuri, K. K., Mascolo, C., and Roussos, G. Open Source Smartphone Libraries for Computational Social Science. In *Proceedings of UbiComp '13 Adjunct* (2013), 911–920.
22. Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Alstyne, M. V. Life in the network: the coming age of computational social science. *Science 323*, 5915 (February 2009), 721–723.
23. LiKamWa, R., Liu, Y., Lane, N. D., and Zhong, L. Moodscope: building a mood sensor from smartphone usage patterns. In *Proceedings of MobiSys '13* (2013), 389–402.
24. Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., Chittaranjan, G. T., Campbell, A. T., Gatica-Perez, D., and Choudhury, T. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of UbiComp '12*, ACM (2012), 351–360.
25. Madan, A., Cebrian, M., Lazer, D., and Pentland, A. Social sensing for epidemiological behavior change. In *Proceedings of UbiComp '10*, ACM (2010), 291–300.
26. Mehrotra, A., Pejovic, V., and Musolesi, M. SenSocial: A Middleware for Integrating Online Social Networks and Mobile Sensing Data Streams. In *Proceedings of Middleware '14* (2014), 205–216.

27. MoodTraces application. `https://play.google.com/store/apps/details?id=com.nsds.moodtraces`.

28. Olesen, J., Gustavsson, A., Svensson, M., Wittchen, H. U., and Jonsson, B. The Economic Cost of Brain Disorders in Europe. *European Journal of Neurology 19*, 1 (January 2012), 155–162.

29. Osmani, V., Maxhuni, A., Grünerbl, A., Lukowicz, P., Haring, C., and Mayora, O. Monitoring Activity of Patients with Bipolar Disorder Using Smart Phones. In *Proceedings of MoMM '13* (2013).

30. Peifer, J., and Garrett, K. Best Practices for Working with Opt-In Online Panels. Tech. rep., The Ohio State University School of Communication, April 2014. `http://www.comm.ohio-state.edu/Opt-in_panel_best_practices.pdf`.

31. Rabbi, M., Ali, S., Choudhury, T., and Berke, E. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of UbiComp '11*, ACM (2011), 385–394.

32. Rachuri, K. K., Musolesi, M., Mascolo, C., Rentfrow, P. J., Longworth, C., and Aucinas, A. Emotionsense: A mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of UbiComp '10* (September 2010), 281–290.

33. Rao, K., Wells, T., and Luo, T. Speeders in a multi-mode (mobile and online) survey. *MRA's Alert! Magazine Fourth Quarter 2014* (2014).

34. Roshanei-Moghaddam, B., Katon, W., and Russo, J. The longitudinal effects of depression on physical activity. *General Hospital Psychiatry 31* (2009), 306–315.

35. Roßmann, J. Data quality in web surveys of the German longitudinal election study 2009. In *3rd ECPR Graduate Conference* (2010).

36. Sano, A., and Picard, R. W. Stress recognition using wearable sensors and mobile phones. In *Proceedings of ACII '13* (2013), 671–676.

37. Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. Limits of Predictability in Human Mobility. *Science 327* (2010), 1018–1021.

38. Spaccapietra, S., Parent, C., Damiani, M. L., Macedo, J. A. de, Porto, F., and Vangenot, C. A Conceptual View on Trajectories. *IEEE Transactions on Knowledge and Data Engineering 65*, 1 (April 2008), 126–146.

39. Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., and Campbell, A. T. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of UbiComp '14*, ACM (2014), 3–14.

40. Wasserman, L. *All of Statistics*. Springer Science & Business Media, 2011.