

SmokingOpp: Detecting the Smoking ‘Opportunity’ Context Using Mobile Sensors

SOUJANYA CHATTERJEE, University of Memphis, USA
 ALEXANDER MORENO, Georgia Institute of Technology, USA
 STEVEN LLOYD LIZOTTE, University of Utah, USA
 SAYMA AKTHER, University of Memphis, USA
 EMRE ERTIN, The Ohio State University, USA
 CHRISTOPHER P FAGUNDES, Rice University, USA
 CHO LAM, University of Utah, USA
 JAMES M. REHG, Georgia Institute of Technology, USA
 NENG WAN, University of Utah, USA
 DAVID W. WETTER, University of Utah, USA
 SANTOSH KUMAR, University of Memphis, USA

Context plays a key role in impulsive adverse behaviors such as fights, suicide attempts, binge-drinking, and smoking lapse. Several contexts dissuade such behaviors, but some may trigger adverse impulsive behaviors. We define these latter contexts as ‘*opportunity*’ contexts, as their passive detection from sensors can be used to deliver context-sensitive interventions.

In this paper, we define the general concept of ‘opportunity’ contexts and apply it to the case of smoking cessation. We operationalize the smoking ‘opportunity’ context, using self-reported smoking allowance and cigarette availability. We show its clinical utility by establishing its association with smoking occurrences using Granger causality. Next, we mine several informative features from GPS traces, including the novel location context of smoking spots, to develop the *SmokingOpp* model for automatically detecting the smoking ‘opportunity’ context. Finally, we train and evaluate the SmokingOpp model using 15 million GPS points and 3,432 self-reports from 90 newly abstinent smokers in a smoking cessation study.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**;

Additional Key Words and Phrases: Mobile Health, Context, Smoking Cessation, Intervention, GPS traces

ACM Reference Format:

Soujanya Chatterjee, Alexander Moreno, Steven Lloyd Lizotte, Sayma Akther, Emre Ertin, Christopher P Fagundes, Cho Lam, James M. Rehg, Neng Wan, David W. Wetter, and Santosh Kumar. 2020. SmokingOpp: Detecting the Smoking ‘Opportunity’

Authors’ addresses: Soujanya Chatterjee, schtrj1@memphis.edu, University of Memphis, Memphis, TN, 38152, USA; Alexander Moreno, alexander.f.moreno@gatech.edu, Georgia Institute of Technology, Atlanta, GA, 30332, USA; Steven Lloyd Lizotte, steven.lizotte@geog.utah.edu, University of Utah, Salt Lake City, UT, 84112, USA; Sayma Akther, sakther@memphis.edu, University of Memphis, Memphis, TN, 38152, USA; Emre Ertin, ertin.1@osu.edu, The Ohio State University, Columbus, OH, 43210, USA; Christopher P Fagundes, christopher.fagundes@rice.edu, Rice University, Houston, TX, 77005, USA; Cho Lam, cho.lam@hci.utah.edu, University of Utah, Salt Lake City, UT, 84112, USA; James M. Rehg, rehg@gatech.edu, Georgia Institute of Technology, Atlanta, GA, 30332, USA; Neng Wan, neng.wan@utah.edu, University of Utah, Salt Lake City, UT, 84112, USA; David W. Wetter, david.wetter@hci.utah.edu, University of Utah, Salt Lake City, UT, 84112, USA; Santosh Kumar, skumar4@memphis.edu, University of Memphis, Memphis, TN, 38152, USA.

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.

© 2020 Association for Computing Machinery.

2474-9567/2020/3-ART4 \$15.00

<https://doi.org/10.1145/3380987>

1 INTRODUCTION

‘Impulsivity’ is defined in [15] as ‘*a predisposition toward rapid, unplanned reactions to internal or external stimuli with diminished regard to the negative consequences of these reactions to the impulsive individual or to others.*’ Examples of impulsive behaviors include fights (e.g., verbal arguments and road rage), impulsive buying [36, 53], suicide attempts [43], overeating [71], binge-drinking [68], gambling [60], and smoking lapse [8]. Decades of prior works show that one of the major precipitants of such behaviors is the existence of ‘contexts’ or situational factors that may increase the risk of impulsive behaviors and in some cases, may lead to the final occurrence of the impulsive behavior. We refer to such contexts as ‘opportunity’ contexts, because automated detection of these contexts from mobile sensors gives us an opportunity to deliver novel sensor-triggered interventions.

‘Opportunity’ Context: An ‘opportunity’ context for an impulsive adverse behavior is a *Spatio-Behavioro-Temporal* context that is ‘ripe’ for the occurrence of such behaviors. We call the ‘opportunity’ contexts *Spatio-Behavioro-Temporal* because, in addition to being influenced by the (spatial, social, environmental, and economic) characteristics of the current place, the current context becomes an ‘opportunity’ context if specific behaviors have been performed at the current and prior places.

We use the phenomenon of a forest fire to characterize the ‘opportunity’ contexts because it shares several characteristics with impulsive adverse behavior. For instance, forest fire escalates quickly if the conditions are conducive (e.g., dry leaves, low moisture, and wind.). Using the inhibitor-inducer terminology from the drug development literature, an ‘opportunity’ context is created when there is an *absence* of *inhibitors* and *presence* of *inducers*. In the case of a forest fire, lack of moisture or precipitation (inhibitors) and presence of dry leaves and wind (inducers) create suitable conditions for the spread of a forest fire that may be sparked by an unattended camp fire [67]. We provide several examples of *absence of inhibitors* and *presence of inducers* in Table 1 to demonstrate the generalizability of this characterization for different impulsive adverse behaviors.

The detection of ‘opportunity’ contexts using sensors can be used to deliver sensor-triggered mobile interventions that are optimized for the current context. Doing so, however, first requires the automated detection of such ‘opportunity’ contexts. We present a computational framework that can be used for detecting ‘opportunity’ contexts passively using mobile sensors in the field environment.

Table 1. Examples of *Lack of Inhibitors* and *Presence of Inducers* for different impulsive behaviors

Impulsive Behavior	Absence of Inhibitors	Presence of Inducers
Fights (e.g., road rage)	Absence of Law enforcement or calming influencers	Access to Weapons, physical strength, supporting personnel
Impulsive Buying (online or offline)	Having credit/money on hand	Promotional incentives on products
Smoking Lapse	Smoking Allowance	Availability of cigarettes
Suicide Attempts	Being alone or away from family/friends	Availability of means to commit suicide
Overeating, binge-drinking	Not having to pay per use	Competitors or companions doing the same thing
Gambling	Having credit/funds	Winning or losing streaks

1.1 Technical Challenges

Automated detection of ‘opportunity’ contexts from mobile sensors involves the following three challenges,

Feasibility: Many of the contextual characteristics (absence of inhibitors and the presence of inducers) that define the opportunity usually cannot be directly inferred from the currently available mobile sensors. Hence, they require non-trivial computational modeling.

Validation: Evaluating the accuracy of ‘opportunity’ context detection is challenging because it can usually be ascertained only after the occurrence of impulsive adverse behavior or event, which in many cases may be rare (and sometimes tragic) events.

Temporal Precision: There may be multiple inhibitors and inducers involved in an ‘opportunity’ context. Further, each may be better characterized by its density (e.g., the density of dry leaves) or intensity (e.g., moisture level, wind speed) rather than a binary level of presence or absence. Therefore, it is non-trivial to decide when to declare the emergence of an ‘opportunity’ context.

1.2 Application to Smoking Lapse

To demonstrate the utility of the concept of ‘opportunity’ contexts, we use the case of smoking ‘opportunity’ context, which is hypothesized to increase the probability of smoking lapse.

There are several reasons for this choice. First, tobacco smoking continues to be a leading cause of preventable death in the world, causing more than 7 million deaths per year, and is projected to kill 8 million per year by 2030 [1]. On a positive note, the risk of dying from smoking-related diseases is reduced by 90% on quitting smoking before the age of 40, and approximately 7 out of 10 adult smokers in the U.S. attempt to quit smoking completely. However, only 6.2% of quit attempts are successful. The majority of participants relapse in the first few days after quitting. Hence, developing technologies for smoking cessation is of high **societal importance**.

Second, there have been several recent advances in the detection of smoking from mobile sensors, making it feasible to do real-time-real-life smoking detection [55]. This has resulted in substantial real-life mobile sensor data available to pursue this research on the detection of ‘opportunity’ contexts. We use data collected from one such recent study. This helps us partially address the **feasibility challenge**.

Third, many smoking cessation studies include self-reports multiple times daily. Some of the questions included in these self-reports enquire directly about the smoking restrictions in the current place and the availability of cigarettes. This provides us with labels of *inhibitors* and *inducers*, together with associated sensor data. We use these labels and the detection of smoking to develop a framework on how to address the **validation challenge**.

Fourth, the labels of smoking allowance at the current place and the availability of cigarettes can both be binary decisions. This makes the **temporal precision challenge** tractable. Despite there being one inhibitor and inducer in this case, the temporal precision challenge is still non-trivial because the assessment of both the inhibitor and inducer is only via self-report and does not contain any information regarding the initiation of the state transition. In particular, the self-report only indicates whether the current place allows smoking but not when the participant entered the current place. The place here can refer to stepping outdoors from a building (where smoking may be allowed). Similarly, the self-report may indicate the availability of cigarette at the moment, but it contains no information on when the individual entered a place where cigarette was available.

In summary, the problem of detecting the smoking ‘opportunity’ context demonstrates the challenges of reliably detecting a *spatio-behavioro-temporal* context using mobile sensors.

1.3 Contributions

Our work makes several new contributions in developing the SmokingOpp model. We introduce the concept of smoking ‘opportunity’ context, present its characterization (in Section 5) and measurement (in Section 5.2.1), and show that it predicts the smoking frequency (in Section 5.3). We introduce a new spatial context called

smoking spot (in Section 6.3) and perform context mining to discover informative features from GPS data for detecting the smoking ‘opportunity’ context (in Section 6). We propose segmentation of the sensor time-series into windows that are based on the change in both location and activity state (instead of location transitions alone) to obtain higher temporal precision and greater coverage in detecting smoking ‘opportunity’ context transitions (in Section 7.5).

1.4 Utilities of Detecting the Smoking ‘Opportunity’ Context

This work has several utilities for the computing and health research communities. First, passive detection of the smoking ‘opportunity’ context reduces the burden on participants and reduces the recall-bias associated with retrospective self-reports. Second, our framework for continuous sensing and detection of ‘opportunity’ contexts offers the opportunity to deliver context-sensitive sensor-triggered intervention for smoking cessation. Third, researchers can analyze multi-modal contextual data at higher temporal resolution in vulnerable contexts along with psychological contexts (e.g., stress [32], craving [16]), and assess lapse risk.

2 RELATED WORKS

Our work on the smoking ‘opportunity’ context is related to and builds upon several prior works. First, is the analysis of context associated with impulsive adverse behaviors, which has mostly been from self-reports. Second, our model for detecting the smoking ‘opportunity’ context is built upon a more in-depth analysis of mobility patterns obtained from GPS traces. We summarize prior works on context detection from GPS traces. Finally, our model can be leveraged for designing and delivering sensor-based interventions to support smoking cessation. We describe recent works on geospatial analysis of smoking exposure and context-based smoking intervention.

Research on ‘Contexts’ Associated with Impulsive Adverse Behaviors from Self-Report: The effect of ‘context’ or situational factors on impulsive adverse behavior has been extensively studied via self-reports. It has been found that smoking allowance and availability of cigarettes are significantly associated with the temptation of smoking during smoking cessation and in some cases, may lead to the occurrence of smoking lapse [64]. Works on impulsive eating/binge drinking, show that food cues (context associated with seeing, smelling food), time-of-day, leads to impulsive eating [35, 58]. Contexts such as companion-competition, peer-family influence, social gatherings, are highly conducive to binge-drinking [34]. This work [40] presents a detailed review of the contexts that may lead to impulsive buying behaviors, including having money on someone, promotional incentives, attractive advertisements, and others. These works motivate the formulation of ‘opportunity’ contexts. Explicitly, we define the smoking ‘opportunity’ context using smoking allowance as the *absence of inhibitor* and cigarette availability as *the presence of inducer*.

Context Detection from GPS Traces: The concept of context [20] and context-aware computing [7] has inspired extensive research. Recent works have focused on estimating psychological contexts such as stress [13, 14, 39] by analyzing human mobility patterns from GPS traces. Other works focus on detecting geospatial contexts to infer locations where a user may explore opportunities for performing activities such as ‘eating,’ ‘shopping,’ ‘entertainment,’ ‘sports and exercise,’ ‘fun and amusement,’ using location and activity histories of other users [74]. Detection of contexts with a potential for cue exposure utilizes GPS traces of users to quantify exposure to food outlets [17, 62]. In addition to GPS data, several other sources of data have been used to detect contexts. For example, [54] develops machine learning techniques to mine noisy data from social media and learns patterns to inform a descriptive and predictive model to infer health status. Although we leverage methods from these prior works on cleaning and processing GPS traces to detect dwell places, the development of a geodatabase of smoking spots is a novel contribution of our work.

Geo-spatial Exposure and Context-Sensing Intervention During Smoking Cessation: Prior work [37] has demonstrated the utility of creating a geodatabase of Point-of-Sale tobacco outlets by showing their association

with craving and smoking lapse during smoking cessation. Recent works [48, 51] have investigated the effects of tobacco outlet density and proximity to the user’s home on smoking lapse. They quantify exposure to smoking by the walking distance to the tobacco outlet within a pre-defined range of 250m and 500m from the user’s home. Finally, [41] identifies high-risk locations for smoking using rule-based geo-fencing and provides episodic context-triggered smoking interventions that are well accepted among users. Our work complements these works by creating a new geodatabase of smoking spots, where smoking is highly likely to occur, including micro-locations at personal places such as designated smoking areas outside office buildings or stepping outdoors at residences, that are not easily geofenced.

3 DATASET

We use the data collected in a smoking cessation study to develop the *SmokingOpp* model. The Institutional Review Board approved the study (IRB), and all the participants provided written consent. Participant demographics, inclusion criteria, study setup, and data collected appear below.

3.1 Wearable Sensors and Smartphone

Participants wore a chest-band of sensors (AutoSense [22]) consisting of ECG, respiration, and accelerometers and a wristband consisting of a 3-axis accelerometer and 3-axis gyroscope on both wrists. The participants wore the sensor-suite only during their waking hours (up to 16 hours per day, from wake till bed-time).

Participants carried a study-provided smartphone with the open-source mCerebrum software [31] installed. The study smartphone was used to communicate with the on-body sensor suite and collect self-reports via Ecological Momentary Assessments (EMA). The smartphone’s GPS sensor was utilized to collect GPS traces of participants continuously at a rate of 1 GPS point every 1 second. The GPS data was extracted from the phone at the end of the study. All the data, including wearable sensors, EMA data, GPS sensor traces, were stored in a secure server with the open-source Cerebral-Cortex [28] software installed.

3.2 Participants

Participants were 126 smokers, 55 female, and 71 male, 18+ years of age, with a mean age of 49.134 ± 13.137 years. All participants were African-American, smoked at least 3 cigarettes per day, and were motivated to quit smoking within the next 30 days of the start of the study. The participants were residents of a large city in the USA and had a valid home address and telephone numbers. All of them agreed to wear the sensor suite.

Participants were excluded if they had a contraindication for the nicotine patch (e.g., participants at risk of heart attack, angina, and other related health problems), active substance abuse or dependence issues, physically unable to wear equipment, pregnant or lactating, or currently using tobacco cessation medications. Moreover, participants who were unable to complete the entire study were excluded from the analysis.

3.3 Study Protocol

Interested participants were invited to an in-person information session where they were provided with detailed information about the study. Once enrolled at the baseline visit, participants picked a smoking quit date. They visited the lab during which they were trained in the proper use of the sensor devices and how to respond to questionnaires in the form of Ecological Momentary Assessments (EMA) via mobile phones. They wore the sensors for 4 days during the *pre-quit* phase. They received nicotine patch therapy, self-help materials, and brief quitting advice. On their set quit date, participants returned to the lab. Then they wore the sensors for 10 more days during the *post-quit* (or *smoking cessation*) phase. At the end of 10 days (14 days from the study start), participants returned to the lab and underwent biochemical verification of their smoking status. All the participants were compensated for their time and effort.

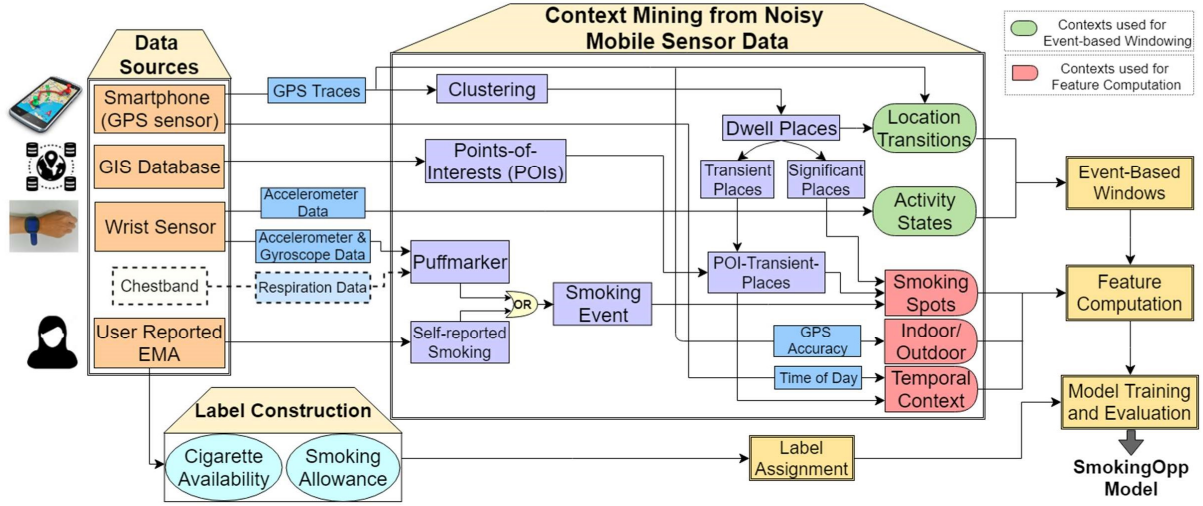


Fig. 1. Data processing pipeline for training and evaluating the SmokingOpp model. The dotted boundaries around the chestband and respiration data denote that using respiration can improve the accuracy of smoking detection, but is optional.

For uniform coverage of a day by EMA's, the day was divided into 4 blocks. The first three blocks consisted of 4 hours each, with remaining time assigned to the last block). In each block, up to 3 EMAs were triggered with a minimum separation of 30 minutes between successive prompts. Irrespective of the source (random or triggered by the detection of stress or smoking), each EMA included the two items we use for constructing the smoking 'opportunity' context labels. Thus, we are able to use each EMA for ground truth labeling.

3.4 Data Description

A total of 16,562 hours of wearable sensor data (about 900 million data points) and over 20 million GPS points across 1,519 person-days were collected from 126 participants. Participants also provided over 4,000 EMAs, with a compliance rate of 67.241% (4,136 out of 6,151 EMA's prompted, were completed by the participants). Each EMA had 53 items that took participants an average of 4.23 minutes to complete.

3.5 Data Screening for Modeling

As we use cross-subject validation, we ensure uniformity and sufficiency of self-reported data (that is used for labeling). Therefore, we select only those participants who provided a minimum of 2 self-reports on each day of the pre-quit and post-quit period. As a result, 26 participants were excluded from the analysis. We use GPS sensor data for model development. Hence, participants who had no location data for more than 3 consecutive days (since this limits us from understanding the prior locations visited for an extended amount of time) were excluded from the study. As a result, 10 participants were excluded. Thus, we are left with 90 participants. That amounts to 12,696 hours of sensor data (about 700 million data points), 15+ million GPS points and 3,432 (out of 4,964, 69.137% completion rate) completed EMAs across 1,080 participant days.

4 THE SMOKINGOPP FRAMEWORK

Figure 1 presents an overview of the entire SmokingOpp framework. It starts with the data sources (represented by the boxes on the far left), followed by the various stages of processing they go through (represented by the

boxes in the middle), and how they contribute to the final model (represented by the boxes on the far right). In Sections 5, 6, and 7, we describe further details of each component as follows. First, we define the smoking ‘opportunity’ context in terms of smoking allowance (absence of inhibitor) and cigarette availability (presence of inducer), (in Section 5.2). Second, we develop a binary measure of the smoking ‘opportunity’ context from self-report, to use as ground truth labels (in Section 5.2.1). Third, we test the clinical utility of this new measure. We use Granger Causality to test how well the smoking ‘opportunity’ context predicts smoking frequency (in Section 5.3). Fourth, we mine the noisy mobile sensor data to derive meaningful contexts and analyze their significance in detecting the smoking ‘opportunity’ context. We mine the contexts from GPS traces and GIS databases (in Section 6). Fifth, we determine the appropriate size of data segmentation (i.e., window length) that can allow efficient detection of every transition in the smoking ‘opportunity’ context with high temporal precision (in Section 7.1). Sixth, we assign ground truth labels (in Section 7.2) and compute informative features from the candidate windows (in Section 7.3) using the insights from the context mining analysis. Finally, we train (in Section 7.4) and test (in Section 7.5) the SmokingOpp model on real-life data.

5 DEFINING THE SMOKING ‘OPPORTUNITY’ CONTEXT

In the following, we first provide a brief description of smoking cessation and the role of context in triggering a smoking lapse. We then present our characterization of the smoking ‘opportunity’ context based on the user reported (via EMA) status of availability of cigarettes and smoking allowance and propose a binary measure of the smoking ‘opportunity’ context (for labels). Finally, we demonstrate, via Granger causality analysis, that the current state of cigarette availability and smoking allowance Granger-cause (are significantly predictive of) the future number of cigarettes smoked.

5.1 Smoking Cessation and the Role of Context in Triggering Smoking Lapse

The process in which smokers attempt to quit smoking and remain abstinent is termed as smoking cessation. During abstinence, nicotine deprivation causes withdrawal symptoms (such as craving [63]). In this vulnerable period, newly abstinent smokers may lapse if they experience acute stress, elation, or restlessness, or are exposed to certain ‘contexts’ (or environmental cues) such as alcohol, seeing others smoke, or cigarette butts [63, 64]. First lapses usually lead to a full relapse (when the newly abstinent smoker reverts to more regular smoking) [64]. Prior works focus on detecting internal states such as stress [32, 56] and craving [16] as well as points of interest such as tobacco point of sale and bars [37], that are highly conducive to smoking [19, 64]. We complement these prior works by detecting ‘personalized spatial contexts’ including smoking spots and micro-locations at significant dwell places that are highly conducive to smoking for each newly abstinent smoker. Individuals who are better able to cope with such at-risk situations are successful in maintaining abstinence [27].

5.2 Inhibitor-Inducer Characterization of the Smoking ‘Opportunity’ Context and Its Measurement

The **smoking ‘opportunity’ context** during smoking cessation, is a Spatio-Behavioro-Temporal context that is ‘ripe’ for the occurrence of smoking lapse. For example, a smoking ‘opportunity’ may be created due to having purchased cigarettes in the recent past (e.g., on the way home) and stepping outdoors in the current place.

We characterize the smoking ‘opportunity’ context using *smoking allowance* at the current location as the *absence of inhibitor* and *cigarette availability* as the *presence of inducer*.

Absence of Inhibitor — Smoking Allowance: Smoking allowance (or restrictions) depends on the characteristics of the current place and being indoors vs. outdoors. For example, smoking allowance near a tobacco outlet or bar is more prevalent as compared to near hospitals, schools, or places of worship, which have very specific and limited spaces of smoking allowance. But, in several cases detecting the allowance state of a place

may be non-trivial due to GPS inaccuracies and ambiguity in the allowance state in the current place. For example, the smoking allowance at personal places (e.g., own home, friend's home) may depend on the owners.

Presence of Inducer — Cigarette Availability: Cigarette Availability refers to the state when an individual has direct access to (e.g., on their person) or known pathways to access (e.g., borrowing from someone). In the absence of direct observation of purchase, detection of the change in cigarette availability is non-trivial.

Association with Impulsive Adverse Behavior — Smoking Lapse: Prior works show that during abstinence, high availability of cigarettes, or heightened awareness of cigarette availability increases the risk of smoking [47]. Smoking allowance is also associated with an increased likelihood of lapse [64].

5.2.1 Constructing a Binary Label of the Smoking ‘Opportunity’ Context. We utilize momentary self-reports, to label smoking allowance and cigarette availability at a given time in the natural environment of participants. For assessment of smoking allowance, the momentary self-report question, i.e., Ecological Momentary Assessment (EMA), ‘Right now, Is smoking allowed where you are?’ has 3 response options of ‘Smoking is forbidden,’ ‘Smoking is discouraged,’ ‘Smoking is allowed.’ For assessment of cigarette availability, the question ‘Right now, Cigarettes are available to me?’ has 5 response options of ‘Not at All,’ ‘With Extreme Difficulty,’ ‘With Difficulty,’ ‘Fairly Easily,’ ‘Easily.’ Given these categorical responses to the two items, we need a partitioning of the 15 ($=5 \times 3$) categorical response combinations to obtain a binary label of the smoking ‘opportunity’ context — ‘high’ or ‘low.’ Using decades of experience in smoking cessation research, we narrow our design choices to the following two,

- **Easily-Allowed:** We consider the smoking ‘opportunity’ context to be ‘**high**,’ if cigarettes are ‘Easily’ available and ‘Smoking is Allowed’ at the current place. For the remaining 14 response combinations, the smoking ‘opportunity’ context is labeled as ‘**low**.’
- **Easily-Fairly-Allowed:** We consider the smoking ‘opportunity’ context to be ‘**high**,’ if cigarettes are ‘Easily’ or ‘Fairly Easily’ available and ‘Smoking is Allowed’ at the current place. For the remaining 14 response combinations, the smoking ‘opportunity’ context is labeled as ‘**low**.’

We use data analysis to explore this design space and settle on a specific mapping (in Section 7.5).

5.3 Granger Causality Analysis Between the Smoking ‘Opportunity’ Context and Cigarette Smoking

Prior works show that smoking lapse is associated with the easy availability of cigarettes and exposure to pro-smoking environments (high smoking allowance) [64]. The goal of this work is to develop a model for continuous assessment of cigarette availability and smoking allowance. To do so, we first show evidence (from our data set) that smoking allowance and cigarette availability predict future smoking. As a variable to describe smoking usage, we use the number of cigarettes smoked. Several other variables are possible, such as time to first lapse or whether there is a lapse, but the number of cigarettes smoked is the easiest to model.

More precisely, assume we have three time series $\{x_t\}_{t \in \mathbb{N}}$, $\{z_t\}_{t \in \mathbb{N}}$, $\{y_t\}_{t \in \mathbb{N}}$, where x_t is *cigarette availability*, z_t is *smoking allowance*, and y_t is the *number of cigarettes smoked*. If we control for y_t up to time t , are x_t and z_t predictive of future values of y_t ? In other words, *are current cigarette availability and smoking allowance predictive of future cigarettes smoked?* If the answer is *yes*, and more recent values of cigarette availability and smoking allowance are more predictive than those further into the past, then this suggests that it would be advantageous to determine higher frequency changes in cigarette availability and smoking allowance (which we use to characterize the smoking ‘opportunity’ context, as described in Section 5.2) via mobile sensor data.

The standard setting for testing this is Granger causality [25]. In this setting, we specify a model with k lags

$$y_t = \alpha + a_1 y_{t-1} + \dots + a_k y_{t-k} + b_1 x_{t-1} + \dots + b_k x_{t-k} + c_1 z_{t-1} + \dots + c_k z_{t-k} + \epsilon_t \quad (1)$$

We would like to do the following hypothesis test. The null is

$$H_0 : b_1 = \dots = b_k = c_1 = \dots = c_k = 0 \quad (2)$$

that is, that lags of *cigarette availability* and *smoking allowance* have no effect on *number of cigarettes smoked*. The alternative hypothesis is

$$H_1 : \exists i, 1 \leq i \leq k \text{ such that } b_i \neq 0 \text{ or } c_i \neq 0 \quad (3)$$

that is, some lag of smoking availability or accessibility has an effect on number of cigarettes smoked.

However, the standard Granger causality model assumes normality of errors; our y_t is non-negative count data, while our x_t and z_t are ordinal levels. Under these conditions, the Gaussian error assumption will generally not hold. It won’t be a good approximation unless the number of values y_t can take is reasonably large, and the skewness of $y_t|x_t$ is low. Thus, using linear regression is often not appropriate. As an alternative, we can fit a generalized linear model (GLM), which relaxes the assumption of Gaussian errors. A GLM [42] has three parts; first, an exponential family distribution describing the response distribution as a function of parameters and features, second, a linear predictor, and third, a link function g .

Given the setting above, the mean of the distribution is related to our features as follows,

$$\mathbb{E}(y_t) = \mu \quad (4)$$

$$= g^{-1}(\alpha + a_1 y_{t-1} + \dots + a_k y_{t-k} + b_1 x_{t-1} + \dots + b_k x_{t-k} + c_1 z_{t-1} + \dots + c_k z_{t-k}) \quad (5)$$

and we want to test the null vs alternative hypothesis as above. The standard model for regression with count data is Poisson regression [18]. This assumes that our response, conditioned on our features, is Poisson distributed. Also, our link function is the log, with inverse link $g^{-1} = \exp$. Using the Poisson distribution may give rise to several issues. First is zero inflation; often count data exhibits more zeros than would be observed under a Poisson distribution. Second is over-dispersion; in the Poisson distribution, the variance is equal to the mean, when, in practice, the observed variance is often higher. One way to handle over-dispersion is to use negative binomial regression, which generalizes Poisson regression with an additional parameter that describes the variance.

Hypothesis Tests: We would thus like to test the following hypotheses.

- Test for zero inflation
 - H_0 : True model is Poisson
 - H_1 : True model is zero-inflated Poisson. This is a mixture model where one mixture component is a Poisson distribution, and the other mixture component takes the value 0 with probability 1.
- Over-dispersion: Poisson vs Negative Binomial
 - H_0 : True model is Poisson
 - H_1 : True model is negative binomial, which allows for more flexible variance
- Granger Causality
 - H_0 : Cigarette availability and smoking allowance have no effect on future number of cigarettes smoked, controlling for lags of cigarettes smoked
 - H_1 : They have an effect

We can test each of these in sequence. Because we have three hypothesis tests, we apply the Bonferroni correction. Our goal is to have the family-wise error rate $P(\text{reject any true } H_0) \leq \alpha$, where α is chosen and usually set to 0.05. In order to ensure this, the Bonferroni correction requires that the p-value for test i , denoted by p_i , to be $p_i < \frac{\alpha}{n}$, where n is the number of tests. For each Hypothesis test here, since we have three tests, instead of seeking a p-value of < 0.05 , we seek a p-value of $< 0.05/3 = 0.017$.

Exploratory Data Analysis: We focus on the post-quit EMA data for analysis; within and across participants, this gives us 2,388 observations. For our exploratory analysis, we focus on 400 observations. We will then perform the hypothesis test on the remaining observations.

We start with a partial autocorrelation (PACF) plot to determine how many lags to use in our model. The PACF gives the auto-correlation between the value at time t and each lag $t - k$, with linear dependence for lags in between removed. Figure 2 shows the results. Note that this is a linear relationship between lag and future, and

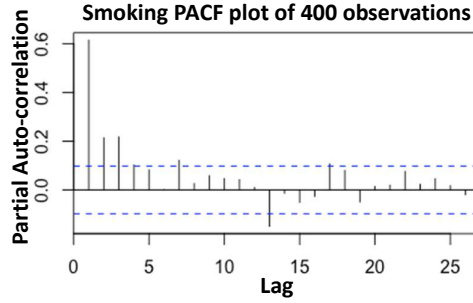


Fig. 2. Partial Auto-correlation Function plot for the number of cigarettes smoked, suggestive of *three* lags for modeling.

we are technically using a GLM, which posits a non-linear relationship; however, it can still be useful for model selection. Therefore, we use three lags.

Primary Data Analysis: First, we test for zero inflation via the *countreg* package in R, which uses the hypothesis test developed in [69]. They derive a score test, which tests whether the mixture weight for the zero-only distribution is 0. We obtain a p-value of 0.254 for this test and fail to reject the null hypothesis of no zero inflation. For the remaining tests, we do not use a zero-inflated model. For the second test, we fit both a Poisson regression and a negative binomial model. The latter has one additional parameter, which models the variance. We can thus do a likelihood ratio test, where the resulting test statistic should have a χ^2_1 distribution. We obtain a p-value of 0.001, which is below our Bonferroni corrected target p-value of 0.017. For the third test, for Granger causality, we test two negative binomial models, one with the availability and allowance lags, and one without. As these are nested models, we can also apply a χ^2_6 test, since the alternative model has six more parameters than the null model. This test gives us a p-value of $5.202e - 05$, well below the target p-value of 0.017.

Thus, we reject the null hypothesis that the lags of cigarette availability and smoking allowance have no effect when including lags of the number of cigarettes smoked. This justifies using cigarette availability and smoking allowance to quantify the smoking ‘opportunity’ context.

6 CONTEXT MINING FROM NOISY MOBILE SENSOR DATA

Prior works on detecting contextual cues for smoking lapse has focused on building a database of well-defined geolocations such as cigarette point-of-sale. Detection of exposure to such locations can then be done via geofencing. Our goal is to detect visitation to any location that may be conducive to a specific smoker. This may include microlocations at personal places such as designated smoking areas outside office buildings or stepping outdoors at residences, that are not easily geofenced.

Therefore, to detect dynamically changing exposures to smoking ‘opportunity’ contexts, we focus on identifying sensor-derived contexts that can detect dynamic changes in the presence of inducers and the absence of inhibitors for smoking. We propose a new context called smoking spots (i.e., where participants are observed to have smoked before quitting) to supplement other informative contexts from the smoking cessation literature (e.g., bars). Moreover, we note that changes in the smoking ‘opportunity’ can not only occur when transitioning from one spatial context into another but also in the same context (e.g., office building) if the participant steps outdoors/indoors. Therefore, in addition to detecting the spatial context, we also detect the activity episodes.

6.1 Deriving *Activity Episodes* from the Wrist Worn Inertial Sensors

We infer whether a subject is in motion or not from the 3-axis wrist-worn accelerometers [21, 45]. Phone accelerometer was not used because on some occasions (e.g., stepping outside of the house or a building, where

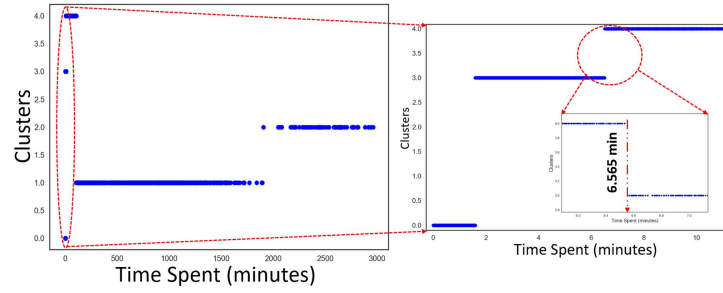


Fig. 3. GMM based clustering produces a lower bound of 6.565 minutes (of dwell time) for determining a *dwell place*.

smoking is usually permitted), the phone may not be on the person. Hence, a vital context may be missed. The accelerometry data collected during vehicular movements is not assessed. We do not consider the ‘in-vehicle’ context to be an ‘opportunity’ context, which can be addressed in future works.

6.2 Deriving *Dwell Places* and *Location Transitions* from GPS Traces

We employ a spatio-temporal clustering based on time and distance to detect *dwell places* from the noisy GPS traces of a participant. We categorize the dwell places into *significant* and *transient places* based on the frequency of visitation. We use the *dwell places* and the GPS traces of each participant to construct their location transitions.

A GPS trace is a sequence of time-stamped GPS points, $tr = [p_1, p_2, \dots, p_m]$. A GPS point $p_i = (x_i, y_i, a_i, t_i)$, $\forall 0 \leq i < m$, consists of timestamp t_i with $(t_i < t_{i+1})$, latitude and longitude (x_i, y_i) , and GPS signal accuracy at that point a_i . GPS sensor noise, poor positioning signal, and other factors lead to inherent inaccuracies in GPS traces. We de-noise the GPS traces via median filtering [75] as the gap between consecutive GPS points is much less than fifty meters even at a speed of 100 kilometers per hour due to the sampling rate of 1 Hz in our GPS traces. We perform median filtering, by substituting a GPS sample point p_i , with the median of temporal predecessor points from a window length of 2 minutes (i.e., 120 predecessor points).

6.2.1 Dwell Places. We use the de-noised GPS traces to derive *dwell places* (pl). More formally, a *dwell place* (pl) is a geographical region where a user has been for at least T_d time within a distance of D_d around it. For example, a user must stay at the gas station for some minimum time to be able to buy cigarettes there.

A pl is characterized by a set of consecutive GPS points $\{p_m, p_{m+1}, \dots, p_k\}$, with three conditions: i.) $d(p_m, p_i) \leq D_d$, ii.) $d(p_k, p_{k+1}) > D_d$, $\forall i$, where $m < i < k$, and iii.) $Interval(p_m, p_k) \geq T_d$. For computing the distance between two points p_m and p_i ($i > 1$), $d(p_m, p_i)$, we use the haversine distance [66]¹ and time interval between two points is computed as $Interval(p_m, p_k) = |t_m - t_k|$. Finally, we obtain a $pl = (x, y, t_{arr}, t_{dep})$, where, (x, y) is the centroid of pl . User’s arrival and departure times into and out of pl are represented by $t_{arr} = t_m$ and $t_{dep} = t_k$ respectively.

Next, we use 200 meters for D_d based on prior works [75, 76]. But, for T_d , there is no set threshold, as it depends on the use case scenario. Hence, we use data analysis to find an appropriate value of T_d that can distinguish transit versus dwell at a pl .

Finding the minimum dwell time (T_d): We are interested in finding the minimum dwell time T_d , which is a lower bound on the dwell time at a location to declare it as a pl ; we discard stops with a time of stay less than T_d . First, we compute the time spent (in minutes) within a distance threshold of 200m. Next, we observe that the distribution of time spent clusters around several mean values. Hence, we use a Gaussian Mixture Model (a soft

¹Haversine distance computes the greater circle distance (shortest distance on the surface of a sphere) between two points on the surface of the earth.

clustering method) to cluster the times spent within a distance threshold of D_d . We experiment using different numbers of clusters/components for the Gaussian mixture model. Finally, we find that 5 components/clusters (see Figure 3) minimizes the Bayesian Information Criterion (BIC) score. From Figure 3, we observe that Clusters 0 and 3 represent those with very low dwell times (ranging from 0.02 - 6.5 minutes). These may be from vehicular movement and walking/running. Cluster 4 represents those places where participants dwell for a considerable amount of time (ranging from 6.565 - 105.2 minutes). Finally, Clusters 1 and 2 represent those places where participants usually dwell for a longer duration. These may be places where participants spend most of their time, such as 'home' and 'work.' Based on this analysis, we use 6.565 minutes, which is the lower bound for cluster 4, as the minimum dwell time (T_d).

6.2.2 Categorization of Dwell Places into Significant and Transient Places. We categorize the *dwell places* into *significant places* and *transient places* based on the *frequency of visitation*.

Significant Places: *Dwell places*, where a participant frequently visits (e.g., 'home,' 'work') are categorized as *significant places*. We apply a density-based clustering [24, 52] on the *dwell places* to obtain the *significant places*. A participant i may have L significant places, which we denote as $SP^i = [SP_1^i, SP_2^i, \dots, SP_L^i, \dots, SP_L^i]$.

Transient Places: Transient places are the *dwell places* which are not marked as a *significant place*. For example, a 'gas station' or a 'convenience store' may be categorized as *transient places*.

6.2.3 Location Transitions. We use the *dwell places* to transform the GPS trace of a given participant into a set of location transitions, where a transition occurs between the departure from one place and the arrival at the next.

We use the location and activity transitions as events to obtain the windows for detecting the smoking 'opportunity' context (in Section 7.1). Additionally, we use the *significant and transient places* along with other information to derive the smoking spots, which are usually associated with the smoking 'opportunity.'

6.3 The Smoking Spots

We identify a new context called *Smoking Spots*. These are places where participants are observed to have smoked before quitting. In smoking cessation studies, data is usually collected from pre-quit days to generate a baseline for analysis. In this work, we leverage the pre-quit data to locate the smoking spots. Similar to the geodatabase of tobacco point-of-sale [37], the smoking spots can be of independent broader utility. This new geodatabase can be used for designing and delivering new interventions. In addition, we use these smoking spots and other informative features to detect smoking 'opportunity' contexts that may be used in new smoking cessation interventions.

We locate two types of smoking spots. We call the first type, the *personal smoking spots*, which are person-specific (frequently visited and unique to each individual, e.g., 'home') and are constructed around the *significant places* where participants usually smoke during their regular smoking period. Personal smoking spots provide evidence that smoking is usually allowed here. Cigarette availability may also be higher at these spots if another smoker shares the *significant place* or if half-burnt cigarettes may still be lying around. The second type is the *general smoking spots*, which are the *transient places* where smoking events are usually observed (potentially, from multiple participants). These are crowd-sourced, i.e., are not person-specific and may be candidates for constructing a new geodatabase of smoking spots. Examples of general smoking spots include outside of an office building or a movie theater. These spots are important as they provide evidence that smoking occurs there. Further, as multiple participants are observed to have smoked there, these spots may also provide an opportunity to borrow a cigarette from another smoker, thus increasing cigarette availability.

In what follows, we first obtain evidence for smoking allowance (from EMAs) and smoking occurrence (via user-reported or automated detection of smoking events, see Section 6.3.1) at different *dwell places*. This helps us categorize the *significant and transient places* into personal and general smoking spots, respectively. We associate

the *transient places* with specific semantic types (in Section 6.3.4) to obtain the *POI-transient-places*. We also use the semantic type to select an appropriate level of proximity to determine exposure to the general smoking spots. We assess how well our characterization of general smoking spots conforms to the widely held beliefs about smoking allowance and smoking occurrence at different types of locations (e.g., places of worship vs. bars).

6.3.1 Smoking Event. We are interested in detecting microlocations where smoking occurs. For this purpose, we first use participants’ self-reported smoking events provided in EMA questionnaires. The EMA item was worded ‘*Since the last assessment, have you smoked any cigarettes?*’, with an option to respond ‘Yes,’ if they have smoked after responding to the preceding EMA, ‘No’ otherwise. The EMA responses are timestamped. Participants may change their location between their last smoking event(s) and when they report the smoking event. To avoid this ambiguity, we utilize an EMA reported at a given *dwell place (pl)* only if the previous EMA was reported at the same location. Although this rule resolves ambiguity, it results in missing smoking events at transient places.

To increase the chances of capturing these missed smoking events (especially at transient places), we utilize puffMarker [55], which is a sensor-based model for detecting smoking episodes. PuffMarker uses the breathing patterns captured from a RIP (Respiratory Inductive Plethysmography) sensor and hand gestures captured using 6-axis inertial sensors (3-axis accelerometers and 3-axis gyroscopes) worn on both wrists. It uses inertial sensor data to identify hand-to-mouth gestures and applies a machine learning model on the corresponding respiration data to detect deep inhalation and exhalation pattern expected during smoking.

Although smoking can be detected using hand-to-mouth gesture alone [44, 61, 65]², using both breathing patterns and hand-to-mouth gesture improves the recall from 75% to 87.5% and reduces the false positive rate from 0.58 per day to 0.17 per day (see Table 2 in [55]). Also, we note that if the wrist sensor is worn on only the dominant wrist, we may miss smoking events from 7 (out of 90) participants who switch hands during smoking.

Using puffMarker, we are able to detect 663 additional smoking events not captured by self-reports. We note that some of these events can be false positives. However, given a low false-positive rate of puffMarker (1 false detection for every 6 days of sensor wearing [55]), we assume all the detected smoking events to be true events.

We also note that as is the case with any sensor-detected event, puffMarker can detect smoking events only when sensors are properly worn and sensors are active. Hence, some smoking events may remain unreported and undetected. As the use of smoking detection is only during constructing a geodatabase of smoking spots, but not needed when the SmokingOpp model is applied to detect smoking ‘opportunity’ contexts, the impact of any missed smoking events on the SmokingOpp model is tolerable.

6.3.2 Characterization of Smoking Spots. Each dwell place where at least one smoking event is detected is a candidate to be designated as a personal or general smoking spot, depending on whether the dwell place is significant or transient. As the detection of one smoking event can result from misreporting or false detection from sensors, we seek additional corroboration to designate a dwell place as a smoking spot. The first criterion is evidence that smoking is allowed at the candidate dwell place. The final criterion is the existence of a sufficient frequency of smoking events at the candidate dwell place for it to be designated as a smoking spot. In the following, we describe our approach to identify personal and general smoking spots.

6.3.3 Identifying the Personal Smoking Spots. We create a sequence of all visits to *significant places*, SP^i (see Section 6.2.2) during the pre-quit phase by participant i to identify personal smoking spots from among them. We represent this temporal sequence as $sp^i = [sp_1^i, sp_2^i, \dots, sp_j^i, \dots, sp_n^i]$, with $sp_j^i = SP_l^i \in SP^i$ for $1 \leq l \leq L$. We note

²We note that in addition to detecting the hand-to-mouth gestures from wrist-worn accelerometers, methods have been developed for detecting hand to mouth smoking gestures using other devices. One approach [72] applies a computer vision technique to detect smoking by recognizing the gesture of the person’s arms from video data. Another work [57] detects smoking gestures by using an RF-based proximity sensor placed on the person’s chest to detect the hand when it is in the vicinity of the mouth. The Personal Automatic Cigarette Tracker - PACT [59] embeds a sensor in the cigarette lighter, which records a smoking event whenever the lighter is lit.

that as sp^i is a temporal sequence, it is likely that several visits can be to the same place, i.e., $sp_j^i = sp_k^i$ for $j \neq k$. Each *significant place* visit, sp_j^i , has a corresponding time of arrival, $sp_j^i.t_{arr}$, and a time of departure, $sp_j^i.t_{dep}$.

We need evidence of smoking allowance and sufficient frequency of smoking occurrence at SP_l^i for it to be regarded as a personal smoking spot. To assess smoking allowance at SP_l^i , we use the self-reported smoking allowance in EMAs (see Section 5.2.1) at this place. A ‘*Smoking is allowed*’ response to the EMA item ‘*Right now, Is smoking allowed where you are?*’ is considered to be positive evidence. For each participant i , we assign an EMA e to a sp_j^i if $sp_j^i.t_{arr} \leq t_e \leq sp_j^i.t_{dep}$, where t_e is the initiation time of response to that EMA. We then compute the *probability of smoking allowance* at SP_l^i as the fraction of EMA responses at $\forall sp_j^i = SP_l^i$ that are positive.

Next, for evidence of smoking occurrence, we compute the *probability of smoking occurrence* at SP_l^i as $\sum_{j=1}^{n_l} \mathcal{E}_j^{SP_l^i} / n_l$, where $\mathcal{E}_j^{SP_l^i} = 1$, if at least 1 smoking episode is detected during j^{th} visit to SP_l^i , and n_l is the total number of visits to SP_l^i . We compute the *probability of smoking occurrence* as the fraction of visits to SP_l^i when smoking is observed at least once.

We use this definition instead of counting each smoking occurrence for two reasons. First, both the EMAs as well as the detection of smoking by sensors may miss several smoking episodes. Hence, using the numbers may not be fully representative of the actual smoking prevalence at different *significant places*. Second, this definition provides some in-variance to the time spent at different *significant places*.

To determine a threshold for deciding personal smoking spots, we analyze the probability of smoking allowance and the probability of smoking occurrence at all candidate SP_l^i in our data set. We want to cluster the candidate SP_l^i , which are closest to each other based on the probability of smoking allowance and occurrence. Using Agglomerative clustering [10], we observe that any SP_l^i with both probabilities of 0.4 or higher are separated. Hence, we use the threshold of 0.4 for the probability of smoking allowance and occurrence to designate a personal smoking spot (see Figure 4).

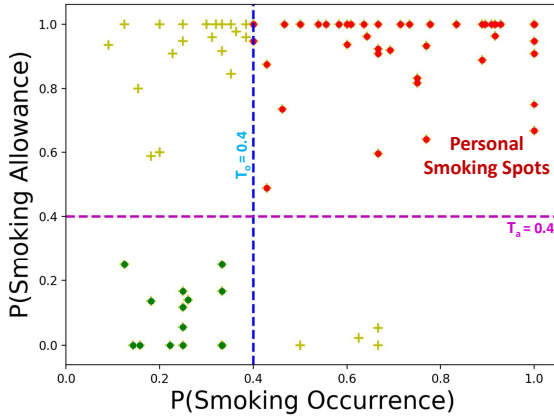


Fig. 4. *Significant places*, where the *probability of smoking allowance and occurrence* are both 0.4 or higher are designated as *Personal Smoking Spots* (marked in red).

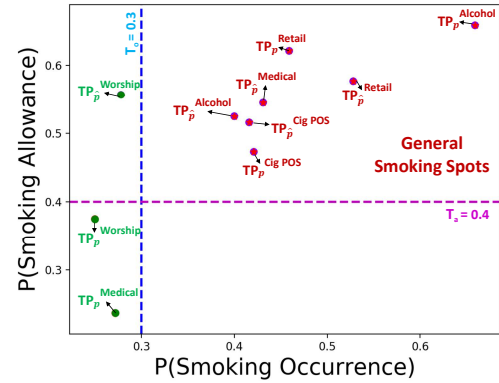


Fig. 5. *POI-transient-places* of type i (either TP_p^i or TP_p^i), where the *probability of smoking allowance and occurrence* are 0.4 and 0.3 or higher, respectively, are designated as *General Smoking Spots* (marked in red).

6.3.4 Identifying the General Smoking Spots. We create a list of all transient places (TP) visited by any participant during the pre-quit phase to identify general smoking spots from among them. In contrast with a small number of significant places (for each participant), the number of transient places visited by all the participants is quite

large. To narrow down our search for general smoking spots, we extract the semantic meaning of *transient places* to assess the smoking allowance at a candidate location to determine if it is potentially conducive to smoking.

To obtain this information, we create a customized Point-of-Interest (POI) database, which contains a list of POIs in and around the city of residence, including properties of the place, GPS location, and type. Following are a list of POIs and the corresponding databases from which they have been acquired.

POIs usually conducive to smoking: We acquire POIs such as a smoke shop, tobacco retail outlet, convenience stores, bars, and others, where smoking is usually allowed and/or there is an opportunity to purchase cigarettes. We acquire 10, 875 licensed alcohol places, including bar and alcohol stores through the state’s alcoholic beverage commission (e.g., TABC [2]), 8,768 licensed cigarette point-of-sale (POS) (e.g., tobacco shop (5, 501) and gasoline station (3, 267)), retail and convenience stores (16, 518) from the state’s comptroller (e.g., [3]).

POIs usually NOT conducive to smoking: We acquire POIs such as hospitals, churches, schools, and other similar places, where smoking is usually prohibited (barring a few designated spots), and the opportunity to purchase cigarettes is almost none. We obtain 1, 396 licensed medical services, including *hospitals, clinics, and biomedical-research labs* from the state’s comptroller. We obtain 1, 543 registered *public and private schools, colleges, universities, libraries, 60 community centers*, and 2, 510 registered places of worship, including *church, temple, mosque, others* from state’s open data portals.

Obtaining the POI database: Using ESRI ArcMap 10.x, we create three individual address locators from the TIGER [6] and STAR Map [5] data sets. Then, we create a composite locator in ArcMap to incorporate all three individual locators. First, we examine and clean all data, next we geocode them using the composite locator based on address matching. The first round of geocoding reaches an average matching of 93%, exceeding a commonly accepted geocoding threshold score of 85% (for the first run) [50]. Then, we visually inspect the tied addresses and re-match by using the TIGER file, STAR map, and Google Maps. We use the re-matching process to enhance the accuracy of the original geocode results. Finally, we merge all extracted GIS data to obtain a master POI database, which contains all the POIs with their semantic types and corresponding GPS coordinates.

Obtaining the POI-transient-places: Finally, we use this newly constructed POI database to associate semantic meaning to *all the transient places* from our dataset. As described above, we consider 6 POI semantic types — ‘alcohol,’ ‘cigarette point-of-sale (POS),’ ‘retail,’ ‘medical,’ ‘school,’ and places of ‘worship.’ In order to associate a *transient place* with a POI semantic type, we use the method described in [38]. A *transient place* associated with a POI is termed as *POI-transient-place*, indexed by $i \in \{\text{‘alcohol,’ ‘cigarette point-of-sale (POS),’ ‘retail,’ ‘medical,’ ‘worship’}\}$. Note, we do not find any EMA reported at the ‘school’ POI. Hence, the analysis of the ‘school’ type has not been possible in this work. We exclude all the other *transient places* that are without any POI association.

Next, we create a subset of *POI-transient-places* based on the proximity from their associated POI. The *POI-transient-places*, which are in a closer proximity to their associated POI are termed as *proximal-POI-transient-places*, (TP_p). In particular, $\forall TP_p^i \in TP_p$ reside within a distance of 30m from the centroid of their associated POI. The remaining POI-transient-places are referred to as non-proximal-POI-transient-places ($TP_{\hat{p}}$). In particular, $\forall TP_{\hat{p}}^i \in TP_{\hat{p}}$ reside within the buffer of 30m and 100m from the centroid of their associated POI. Depending on the proximity distance of a POI-transient-place to the POI centroid, a POI-transient-place type i may be assigned to TP_p (as TP_p^i) or $TP_{\hat{p}}$ (as $TP_{\hat{p}}^i$).

Designating POI-transient-places as General Smoking Spots: From all visits to transient places, we extract all distinct visits to POI-transient-places by any participant and represent them as $d = [d_1, d_2, \dots, d_N]$, with $d_j = TP_p^i$ or $d_j = TP_{\hat{p}}^i$ for some POI-transient-place i . We note that it is likely that different visits by same or different participants may be to the same POI-transient-place type i , i.e., $d_j = d_k = TP_p^i$ (or $TP_{\hat{p}}^i$) for $j \neq k$. Each visit d_j , has a corresponding time of arrival, $d_j.t_{arr}$, and a time of departure, $d_j.t_{dep}$.

Similar to personal smoking spot identification, we need evidence of smoking allowance and sufficient frequency of smoking occurrence at a POI-transient-place type i for it to be regarded as a general smoking spot. To assess

smoking allowance at a POI-transient-place type i across all participants (since these are crowd-sourced and not person-specific), we use the self-reported smoking allowance in EMAs. A ‘*Smoking is allowed*’ response to the EMA item ‘*Right now, Is smoking allowed where you are?*’ is considered to be positive evidence. For all participants, we assign an EMA e to a d_j if $d_j.t_{arr} \leq t_e \leq d_j.t_{dep}$, where t_e is the initiation time of response to that EMA. Similar to personal smoking spots, the *probability of smoking allowance* at a general smoking spot candidate TP_p^i (or $TP_{\hat{p}}^i$) as the fraction of EMA responses at $\forall d_j = TP_p^i$ (or $TP_{\hat{p}}^i$) that are positive.

Next, for the evidence of smoking occurrence, we compute the *probability of smoking occurrence*, at TP_p^i (or $TP_{\hat{p}}^i$) as $\sum_{j=1}^n \mathcal{E}_j / n$, where $\mathcal{E}_j = 1$, if at least 1 smoking episode is detected during j^{th} visit to a TP_p^i (or $TP_{\hat{p}}^i$), and n ($\leq N$) is the total number of visits to TP_p^i (or $TP_{\hat{p}}^i$). We compute the *probability of smoking occurrence* as the fraction of visits to TP_p^i (or $TP_{\hat{p}}^i$) when smoking is observed at least once.

To determine a threshold for general smoking spots, we analyze the probability of smoking allowance and the occurrence of all the candidate proximal and non-proximal POI-transient-places in our data set. Using Agglomerative clustering, we observe that any POI-transient-place with the probability of smoking occurrence of 0.3 or higher and the probability of smoking allowance of 0.4 or higher are separated. Hence, we use these two thresholds to designate a general smoking spot (see Figure 5).

Our analysis confirms that smoking allowance and occurrence are high around bars and alcohol stores and low at closer proximity to places of worship and hospitals. Interestingly, we observe that smoking occurrence is low at non-proximal areas of places of worship. However, we find a higher probability of smoking occurrence in non-proximal areas of hospitals.

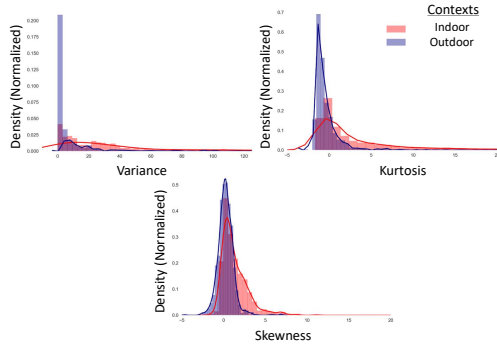


Fig. 6. Distributions of variance, skewness, and kurtosis of GPS accuracy can distinguish ‘outdoors’ vs. ‘indoors.’

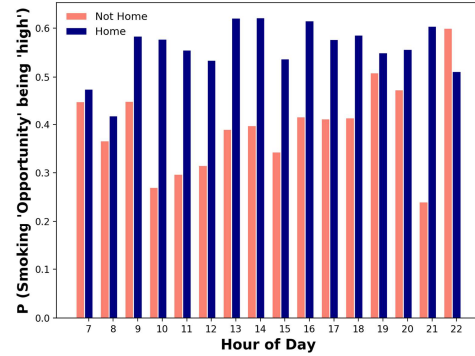


Fig. 7. The Smoking ‘opportunity’ context varies significantly across time when people are not at ‘home.’

6.4 Indoor/Outdoor Context

Smoking spots usually provide a coarse boundary of geographical region where smoking may occur. Detection of indoor vs. outdoor can provide a more fine-grained indication of momentary smoking ‘opportunity’ as tobacco control policies in several countries prohibit smoking at indoor environments [4]. The indoor environment is usually associated with low/no smoking allowance. We build upon prior works [26, 49] to find indicators or correlates of indoor/outdoor environment from GPS data. The intuition behind the estimation of indoor/outdoor from GPS data is that GPS signal accuracy deteriorates in indoor environments due to signal obstruction and

limited satellite visibility [70]. To assess the suitability of GPS accuracy as an indicator of indoor vs. outdoor, we label some time windows as ‘indoor’ and ‘outdoor’ and then test how well are they separable using GPS accuracy.

To estimate indoor/outdoor from GPS signal accuracy, first, we label time windows when participants are walking outside versus when indoors. For this purpose, we compute the speed (in meter/sec) of participants when transitioning from one dwell place to another. If the speed is greater than the maximum comfortable walking speed (1.46 meter/sec [11]) and less than minimum transportation in a vehicle (3.33 meter/sec [29]) while transitioning from one dwell place to another for a certain amount of time, we label those windows as an *outdoor* environment. Note, we do not use transport ‘inside’ a vehicle as ‘outdoor’ to remove any ambiguity. If the speed is lesser than maximum comfortable walking speed and participants are dwelling within 10 meters of home location, we label those windows as *indoor* environment.

To find informative features from GPS accuracy, we compute the variance, skewness, and kurtosis of GPS accuracy in the windows marked as *outdoor* ($n=1,488$) and *indoor* ($n=575$). Next, we perform one-tailed Mann-Whitney U test on the following samples, first, the variance, skewness, and kurtosis of GPS accuracy when *indoors*, second, the variance, skewness, and kurtosis of GPS accuracy when *outdoors*. We observe that variance ($n = 575$; $median = 28.777$; $mean = 63.093 \pm 84.312$; $p \leq 0.0001$), skewness ($n = 575$; $median = 0.874$; $mean = 1.383 \pm 1.847$; $p \leq 0.0001$), and kurtosis ($n = 575$; $median = 0.943$; $mean = 5.558 \pm 12.903$; $p \leq 0.0001$) of GPS accuracy when *indoors* is significantly greater than the variance ($n = 1,488$; $median = 1.376$; $mean = 10.441 \pm 30.604$; $p \leq 0.0001$), skewness ($n = 1,488$; $median = 0.219$; $mean = 0.233 \pm 0.945$; $p \leq 0.0001$), kurtosis ($n = 1,488$; $median = -0.896$; $mean = -0.22 \pm 3.476$; $p \leq 0.0001$) when *outdoors* (see Figure 6).

6.5 Temporal Context

In addition to exploring the role of spatial context in detecting smoking ‘opportunity,’ we also investigate the role of temporal context. We analyze the role of time of day and temporal patterns of visitation to tobacco outlets.

6.5.1 Contextualized Time of Day. We explore the impact of the hour of the day on the smoking ‘opportunity’ variations. We use the binary labels of the smoking ‘opportunity’ from Section 5.2.1 to compute the probability of the smoking ‘opportunity’ being ‘high’ during each waking hour of the day (T_h , from 7 AM to 10 PM, referred to as T_7 and T_{22} , respectively), as the fraction of EMA responses at T_h that are labeled as ‘high’ smoking ‘opportunity’.

Role of time of day on the likelihood of smoking or craving [9, 16] has been known. However, we notice that the likelihood of the smoking ‘opportunity’ being high differs when participants are at ‘home’ (within 100 meters of ‘home’ GPS coordinate), and ‘not at home’ (see Figure 7).

Key Observations: We observe a decreasing trend in the smoking ‘opportunity’ outside the home, as time progresses from T_8 to T_{11} (which is most likely because participants are usually busy doing their work during these hours). During T_{12} , T_{13} , T_{14} , the smoking ‘opportunity’ increases gradually (most likely participants may break for lunch during these hours). There is a decrease during T_{15} , T_{16} , which is followed by a rising trend during the evening hours. Second, barring the morning hours of T_8 and T_9 , we do not observe any interesting temporal trend of the smoking ‘opportunity’ when participants are at home.

6.5.2 Temporal Patterns of Visitation to Cigarette Point-of-Sale (POS). As described in Section 5.2, the high availability of cigarettes is usually associated with a ‘high’ smoking ‘opportunity.’ We are interested in finding whether the *time since* the previous visit to a POS (where there is an opportunity to purchase cigarettes) is associated with cigarette availability. We consider a visit to cigarette POS has occurred if an individual dwells for at least 6.565 minutes within the distance of 30m from the centroid of a cigarette POS [37].

As we are interested in determining the role of such visits on cigarette availability, we use the two design choices from Section 5.2.1, but only the responses to the EMA questionnaire — ‘*Right now, Cigarettes are available to me?*’ in both categories of *Easily-Allowed* and *Easily-Fairly-Allowed* to assign EMA responses to the states of ‘high’ or ‘low’ availability.

Table 2. For *Easily-Fairly-Allowed*, during the post-quit phase, the median *time since* (in *hr*) the last POS visit associated with lower cigarette availability is significantly greater ($p_value = 0.019$) than that associated with high cigarette availability.

Design Choice	Study Session (# of Low_CA, # of High_CA)	median (TSPV_Low_CA)	median (TSPV_High_CA)	mean $\pm$ sd (TSPV_Low_CA)	mean $\pm$ sd (TSPV_High_CA)	p_value
<i>Easily-Allowed</i>	Pre-quit (229, 328)	6.296	8.772	18.844 $\pm$ 23.211	20.456 $\pm$ 22.608	0.09
<i>Easily-Allowed</i>	Post-quit (576, 475)	54.501	44.456	76.526 $\pm$ 78.411	71.859 $\pm$ 84.908	0.084
<i>Easily-Fairly-Allowed</i>	Pre-quit (152, 405)	7.56	7.88	19.436 $\pm$ 22.183	19.924 $\pm$ 23.314	0.372
<i>Easily-Fairly-Allowed</i>	Post-quit (444, 607)	57.288	44.478	78.845 $\pm$ 78.623	67.726 $\pm$ 72.964	0.019

Subsequently, we analyze the effect of the *time since* the previous visit to cigarette POS on the state of cigarette availability. First, for each participant, we compute the *time since* the previous visit to cigarette POS using the following algorithm. We start from the beginning of the study for each participant and set the *time_since* counter to 0. Once we find a visit made to a cigarette POS, we keep incrementing the *time_since* counter by dwell time at intermediate dwell places and transitions until the next visit to POS is recorded. At the beginning of the next visit, we reset the *time_since* to 0 and keep on incrementing it until the beginning of the next visit. The process continues until the end of the study. The *time_since* is computed in the unit of *hour (hr)*.

We hypothesize that greater *time since* the previous visit to cigarette POS indicates a lower cigarette availability and conversely lesser *time since* the previous visit to cigarette POS indicates a higher cigarette availability. The intuition is that the availability of cigarettes (by purchase) reduces with lesser frequency of visitation to the cigarette POS. To evaluate the above hypothesis, we perform the following statistical tests and report several interesting insights. For each design choice approach (*Easily-Allowed* and *Easily-Fairly-Allowed*) and the corresponding phase of the study (pre and post-quit), we perform a one-tailed Mann-Whitney U test on the two independent samples. The first samples are *time since* the previous POS visit associated with a lower cigarette availability (*TSPV_Low_CA*). The second samples are *time since* the previous POS visit associated with a higher cigarette availability (*TSPV_High_CA*). We present the hypothesis test results in Table 2.

Key Observations: First, during the post-quit, for *Easily-Fairly-Allowed*, the median *TSPV_Low_CA* (57.288 *hr*) is significantly higher ($p_value = 0.019$) than the median *TSPV_High_CA* (44.478 *hr*). Second, the above finding implies that delaying the visit to a cigarette POS lowers the state of cigarette availability, which in turn indicates ‘low’ smoking ‘opportunity.’ Third, the median *time since* previous POS visit is higher during the post-quit as compared to the pre-quit phase. This implies that there is a noticeable behavioral change in terms of the frequency of visitation to cigarette POS.

7 THE SMOKINGOPP MODEL

For developing the SmokingOpp model, we first develop an approach to segment the continuous sensor time series into candidate windows. After locating the windows, we assign ground truth labels to it and compute features from sensor data to train the SmokingOpp classifier. Finally, we evaluate the impact of various design choices on the accuracy of detecting the smoking ‘opportunity’ context.

7.1 Event-Based Windowing for Efficient Detection of the Smoking ‘Opportunity’ Context

Our goal is to identify an appropriate segmentation of the sensor time series (into *windows*) such that every transition in the smoking ‘opportunity’ context concurs with a transition in the window. This will provide high computational efficiency and sufficient data within a window to compute robust features.

A participant’s smoking ‘opportunity’ context may change with changes in his/her location. For example, in Figure 8, we observe that the smoking ‘opportunity’ context changes from ‘low’ at the *Dwell Place A* to ‘high’ at the *Dwell Place B*. Sometimes, the smoking ‘opportunity’ context may also change in the same place (see Figure 9).

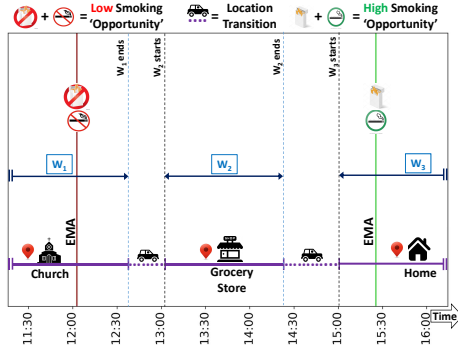


Fig. 8. Defining Smoking ‘opportunity’ context windows based on change in location.

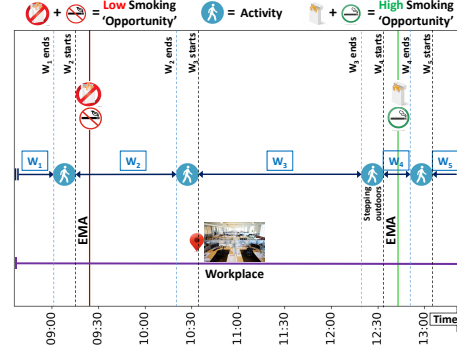


Fig. 9. Defining Smoking ‘opportunity’ context window based on both change in location and activity state.

For example, if a participant moves (detected as an activity episode) from indoor (‘low’ smoking ‘opportunity’) to an outdoor designated smoking area (‘high’ smoking ‘opportunity’) at his/her workplace.

Using only location to define windows may miss some smoking ‘opportunity’ context transitions, while defining windows by using both location and activity transitions may significantly increase the number of windows. Therefore, we consider both approaches to segment the sensor time series into windows. We refer to the first approach as *Win-Loc*. We observe from a sample in Figure 8 that there are 3 such windows, namely, W_1, W_2, W_3 . We refer to the second approach as *Win-Loc-Act*. We observe from a sample in Figure 9 that there are 5 such windows created, namely, W_1, W_2, W_3, W_4, W_5 .

7.2 Label Assignment

After determining the candidate windows, our goal is to assign labels (obtained from participants’ self-reports) to each window. First, we determine the binary labels for the smoking ‘opportunity’ context. We construct two classes of labels based on each design choice (i.e., *Easily-Allowed*, and *Easily-Fairly-Allowed*) as described in Section 5.2.1. If we find a single self-report in a window, we assign the reported label of the smoking ‘opportunity’ context to that window. If multiple homogeneous self-reports occur in a window and all of them have the same label, we assign the reported homogeneous label to that window. We exclude all other windows because of ambiguity in labels. We note that these excluded windows only constitute 7.991% of total windows.

7.3 Feature Computation

To detect the smoking ‘opportunity’ context, we compute several features from each window of sensor data. We use context mining presented in Section 6 to compute 11 window-level features.

Personal Smoking Spots (2 features): To determine the likelihood of the place in the current window to be a personal smoking spot (see Section 6.3.3), we use the binary indicator *is current place a significant place?*, and the *probability of smoking occurrence at the current place* as features.

General Smoking Spots (2 features): Similar to personal smoking spots, to determine the likelihood of the the place in the current window to be a general smoking spot (see Section 6.3.4), we use the binary indicator *is current place a TP_p^i or TP_p^i ?*, and the *probability of smoking occurrence at the current place* as features.

Indoor/Outdoor Context (3 features): As the *variance*, *skewness*, and *kurtosis* of GPS accuracy in the current window are indicators of indoor vs. outdoor (see Section 6.4), we use these three statistics as features.

Contextualized Time of Day (3 features): We use *distance to 'home' (in meters)* from the current place, and we compute two features using the hour of day at the mid point of the current window (T_h): T_h and T_h^2 , to assess linear and non-linear impact of time of day (see Section 6.5.1).

Temporal Patterns of Visitation to Cigarette POS (1 feature): We use the *time since* the last visit to a cigarette POS to indicate the behavioro-temporal evolution of cigarette availability (see Section 6.5.2).

7.4 Model Training and Evaluation

Our goal is to develop a model that can passively detect changes in smoking ‘opportunity’ contexts from continuous mobile sensor data in-the-wild. For model training and evaluation, we use sensor data and ground truth self-reports (as a label) of the smoking ‘opportunity’ context from the post-quit phase.

We consider the widely-used supervised classification models for detecting the smoking ‘opportunity’ context – Random forest based classifier [12] with 100 (RF_100) and 1000 (RF_1000) trees, SVM based classifier with RBF kernel (SVM-RBF) [73], Logistic regression (LR) classifier [30], and Adaboost classifier [23]. We use grid search for hyper-parameter optimization in each model.

For evaluation of model accuracy, we use leave-one-participant-out-cross-validation (LOPOCV), where we exclude a single participant’s data for testing and use the remaining for training purposes, and repeat the experiment for all the participants. We report the median values of the model scores, including F1, Recall, and Precision in box-plots (see Figures 10 and 11). We also evaluate the impact of windowing choices on temporal precision of the model, i.e., number of smoking ‘opportunity’ transitions detected, and the size of windows (as smaller windows may detect smoking ‘opportunity’ transitions sooner).

7.5 Experimental Results

Our goal is to evaluate different combinations of the design choices for windowing (*Win-Loc* and *Win-Loc-Act*) and labeling (*Easily-Allowed* and *Easily-Fairly-Allowed*) to observe their impact on model performance.

Win-Loc Windowing and Easily-Allowed labels: Out of 666 labeled windows, 290 (43.54%) are ‘high,’ and 376 (56.46%) are ‘low.’ Recall, precision, and F1 values appear in Figure 10 (on the left). All the models produce similar performance scores, with logistic regression having the best median F1 score of 51.67%.

Win-Loc-Act Windowing and Easily-Allowed labels: Out of 1,363 labelled windows, 658 (48.28%) are ‘high’ and 705 (51.72%) are ‘low.’ Recall, precision, and F1 values appear in Figure 10 (on the right). All the models outperform the *Win-Loc* windowing and *Easily-Allowed* approach. Adaboost produces the best median precision of 70.73%, and logistic regression produces the best median F1 of 63.07%.

Win-Loc Windowing and Easily-Fairly-Allowed labels: Out of 666 labeled windows, 373 (56.0%) are ‘high,’ and 293 (43.99%) are ‘low.’ Recall, precision, and F1 values appear in Figure 11 (on the left). The performance of all the models is better than the *Win-Loc* or *Win-Loc-Act* windowing and *Easily-Allowed* based approach. Adaboost produces the best median precision of 66.67%, and logistic regression produces the best median F1 of 67.98%.

Win-Loc-Act Windowing and Easily-Fairly-Allowed labels: Out of 1,363 labelled windows, 808 (59.28%) are ‘high’ and 555 (40.72%) are ‘low.’ Recall, precision, and F1 values appear in Figure 11 (on the right). Overall, the *Win-Loc-Act* windowing and *Easily-Fairly-Allowed* approach outperforms all the other modeling approaches. Logistic regression performs the best overall with a median F1 of 74.3%.

Key Findings: We report several key findings. First, from Table 3, we observe that the *Easily-Fairly-Allowed* labeling approach produces better results than *Easily-Allowed* approach. We observe an improvement of 16.31% and 11.23% in F1 scores for *Win-Loc* and *Win-Loc-Act* windowing, respectively, via logistic regression-based modeling. This may be because in several scenarios of *fairly easily* availability of cigarettes, even if the participants are not carrying cigarettes with them, they can borrow a cigarette (fairly or relatively easily), resulting in a ‘high’ smoking ‘opportunity’ context. Hence, labeling some of these contexts as ‘low’ smoking ‘opportunity’ context

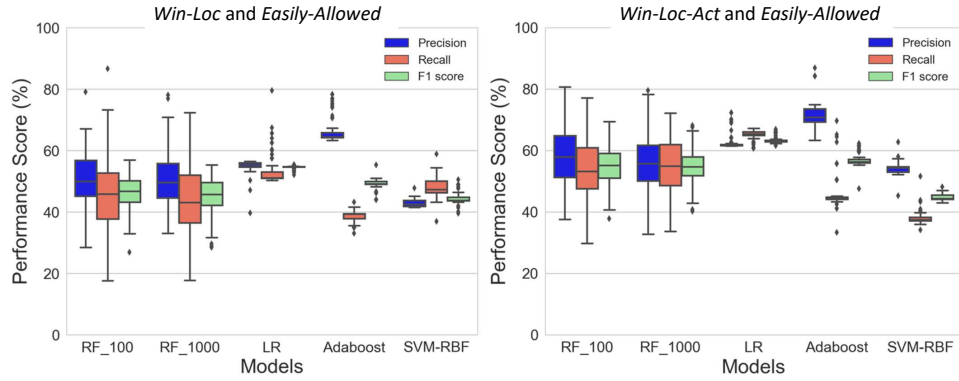
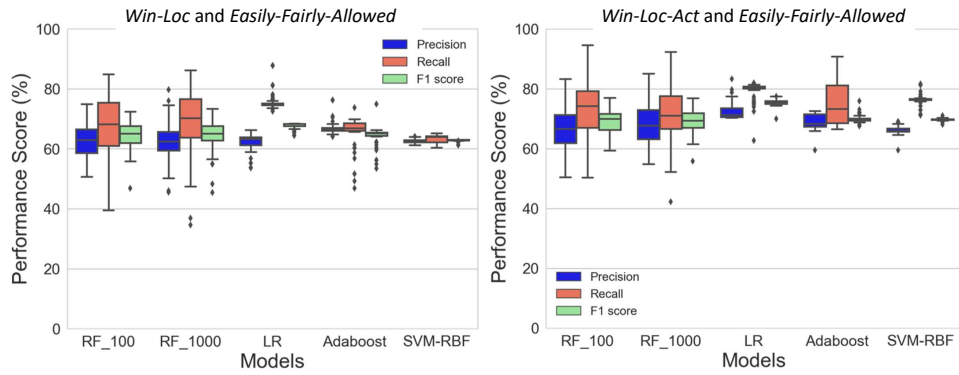
Fig. 10. Model performance for both choices of windowing when using the *Easily-Allowed* label.Fig. 11. Model performance for both choices of windowing when using the *Easily-Fairly-Allowed* label.

Table 3. Shows the performance of the Logistic Regression model for various choices of windowing and labeling approaches.

Windowing	Labeling	# of Windows		Logistic Regression		
		Labeled as ‘High’	Labeled as ‘Low’	F1	Precision	Recall
Win-Loc	Easily-Allowed	290	376	51.67	55.34	51.79
Win-Loc-Act	Easily-Allowed	658	705	63.07	62.26	65.19
Win-Loc	Easily-Fairly-Allowed	373	293	67.98	63.95	74.31
Win-Loc-Act	Easily-Fairly-Allowed	808	555	74.3	70.27	79.82

in the *Easily-Allowed* labeling approach may result in false labels. We also observe that despite having a larger number of windows in the *Win-Loc-Act* approach, we obtain a higher F1 as compared to the *Win-Loc* approach. In particular, the *Win-Loc-Act* approach achieves a 6.32% increase in the F1 as compared to the *Win-Loc* approach (with the *Easily-Fairly-Allowed* approach for computing the smoking ‘opportunity’ context).

Second, we observe that we can capture 37.864% more smoking ‘opportunity’ context transitions by using activity episodes in addition to place transitions in defining our window of assessment. In particular, we capture 206 transitions using the *Win-Loc* approach (106 transitions from ‘low’ to ‘high’ and 100 transitions from ‘high’

to ‘low’). Using *Win-Loc-Act*, on the other hand, we capture a total of 284 transitions (145 transitions from ‘low’ to ‘high’ and 139 transitions from ‘high’ to ‘low’).

Third, we observe that the average window size (in minutes) for *Win-Loc-Act* approach is 48.479 ± 55.302 (median=28.506) versus that of *Win-Loc* approach (64.545 ± 65.037 minutes (median=41.583)). Therefore, the *Win-Loc-Act* approach provides better temporal precision.

8 LIMITATIONS AND FUTURE WORKS

In this work, we introduce and define the ‘Opportunity’ context for impulsive adverse behaviors. Next, to demonstrate its utility, we develop a framework for detecting the smoking ‘opportunity’ context. We design methods to characterize the smoking ‘opportunity’ context using smoking allowance (as the absence of inhibitor) and cigarette availability (as the presence of inducer). Then, using discriminative features from sensor data, we develop a model to detect binary measures of ‘high’ or ‘low’ smoking ‘opportunity’ context.

Since this work is a first step towards detecting the smoking ‘opportunity’ context, it has several limitations that present opportunities for future research in both the UbiComp and health research communities. First, we define the concept of an ‘opportunity’ context for impulsive behaviors and provide several examples. But, we only present its application to the case of smoking cessation. Our framework may motivate ‘opportunity’ context model development for other impulsive behaviors such as impulsive eating and binge drinking.

Second, our SmokingOpp model achieves only a moderate F1 score of 74.3%. This can partly be attributed to a lack of a direct measure of the transition of the smoking ‘opportunity’ context from ‘low’ to ‘high’ and ‘high’ to ‘low’. We use place transitions and then activity episodes to capture potential indoor to outdoor transitions at the same place. These are proxy measures, neither of which directly identify all candidate transitions of the smoking ‘opportunity’ context. More direct measures, e.g., from wearable eyeglasses that can potentially detect cigarette purchases or cigarette borrowing, can potentially improve the accuracy of detecting the smoking ‘opportunity’ context. They raise exciting research opportunities in first-person computer vision modeling.

Third, this is an offline model, computed purely from observational data. To become widely useful in the society, the clinical utility of the SmokingOpp model in the management of at-risk situations of impulsive behaviors should be established by developing and evaluating sensor-triggered just-in-time mobile intervention via randomized clinical trials that can be triggered based on the detection of the smoking ‘opportunity’ context. For the intervention to be most effective, it should be triggered as soon as there is a change in the smoking ‘opportunity’ context from ‘low’ to ‘high’. Doing so requires the detection of an ‘opportunity’ context as soon as there is a transition of context. This requires the development of an online version of the SmokingOpp model.

Fourth, the SmokingOpp model can be improved over time in a real-life deployment by using active learning that can personalize the model to each individual’s data over time.

Fifth, in this work, we detect the smoking ‘opportunity’ context for cigarette smoking. Recently, there has been a rapid growth in the usage of e-cigarettes [33], and some researchers have reported the harmful effects of e-cigarette [46]. We note that the contexts or situations in which e-cigarettes are usually used may be different from that of smoking cigarettes. This may require an adaption to the SmokingOpp model to detect the e-cigarette smoking ‘opportunity’ context.

9 CONCLUSIONS

Our work introduced a novel concept of ‘opportunity’ context for impulsive adverse behaviors. We characterize these spatio-behavioro-temporal contexts which are conducive to impulsive adverse behaviors (e.g., smoking lapse, overeating/binge drinking, etc.) using inhibitors, absence of which, and inducers, presence of which, create an at-risk situation for impulsive adverse behaviors. It adds a new dimension to the fundamental notion of context and provides a new direction for context-aware applications. Impulsive adverse behaviors can have a significant

negative health impact on individuals and their friends and families. Reliable detection of ‘opportunity’ contexts passively using sensors can create novel opportunities to intervene before such adverse events occur.

As this work has shown, the behavioral component in the ‘opportunity’ context makes its reliable detection challenging. But, our framework for successfully detecting the smoking ‘opportunity’ context changes the question from ‘whether’ to ‘how well’ can these contexts be detected. Consequently, this work opens up exciting research opportunities with a potential for high societal impact.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for significantly improving the organization and presentation of this manuscript. We sincerely thank the smoking cessation research study coordinators Rebecca Stoffel, Michelle Chen, Kristi Parker, Jeffrey Ramirez, and Andy Leung. Research reported here was supported by the National Institutes of Health (NIH) under awards R01CA224537, R01MD010362, R01CA190329, R24EB025845, U01CA229437, and U54EB020404 (by NIBIB) through funds provided by the trans-NIH Big Data-to-Knowledge (BD2K) initiative, and by the National Science Foundation (NSF) under awards IIS-1722646, ACI-1640813, and CNS-1823221.

REFERENCES

- [1] Accessed April, 2019. *CDC: Smoking is the leading cause of preventable death*. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/index.htm
- [2] Accessed February, 2018. *Texas Alcoholic Beverage Commission. “Licensing.”* <https://www.tabc.state.tx.us/>
- [3] Accessed February, 2018. *Texas Comptroller. Active Cigarette/Tobacco Retailers, Open Data Portal*. <https://data.texas.gov/Government-and-Taxes/Active-Cigarette-Tobacco-Retailers/u5nd-4vpg/data>
- [4] Accessed May, 2019. *CDC: Smoking Banned indoors*. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/secondhand_smoke/protection/improve_health/index.htm
- [5] Accessed September, 2018. *Southeast Texas Addressing and Referencing Map*. www.h-gac.com/rds/gis_data/starmap
- [6] Accessed September, 2018. *TIGER/Line Shapefiles and TIGER/Line Files Technical Documentation*. <https://www.census.gov/programs-surveys/geography/technical-documentation/complete-technical-documentation/tiger-geo-line.html>
- [7] Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a better understanding of context and context-awareness. In *International symposium on handheld and ubiquitous computing*. Springer, 304–307.
- [8] Rebecca L Ashare and Larry W Hawk. 2012. Effects of smoking abstinence on impulsive behavior among smokers high and low in ADHD-like symptoms. *Psychopharmacology* 219, 2 (2012), 537–547.
- [9] Timothy B Baker, Megan E Piper, Danielle E McCarthy, Daniel M Bolt, Stevens S Smith, Su-Young Kim, Suzanne Colby, David Conti, Gary A Giovino, Dorothy Hatsukami, et al. 2007. Time to first cigarette in the morning as an index of ability to quit smoking: implications for nicotine dependence. *Nicotine & Tobacco Research* 9, Suppl_4 (2007), S555–S570.
- [10] Doug Beeferman and Adam Berger. 2000. Agglomerative clustering of a search engine query log. In *KDD*, Vol. 2000. 407–416.
- [11] Richard W Bohannon. 1997. Comfortable and maximum walking speed of adults aged 20–79 years: reference values and determinants. *Age and ageing* 26, 1 (1997), 15–19.
- [12] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
- [13] Michelle Nicole Burns, Mark Begale, Jennifer Duffecy, Darren Gergle, Chris J Karr, Emily Giangrande, and David C Mohr. 2011. Harnessing context sensing to develop a mobile intervention for depression. *Journal of medical Internet research* 13, 3 (2011), e55.
- [14] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 1293–1304.
- [15] Samuel R Chamberlain and Barbara J Sahakian. 2007. The neuropsychiatry of impulsivity. *Current opinion in psychiatry* 20, 3 (2007), 255–261.
- [16] Soujanya Chatterjee, Karen Hovsepian, Hillol Sarker, Nazir Saleheen, Mustafa al’Absi, Gowtham Atluri, Emre Ertin, Cho Lam, Andrine Lemieux, Motohiro Nakajima, et al. 2016. mCrave: Continuous estimation of craving during smoking cessation. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 863–874.
- [17] W Jay Christian. 2012. Using geospatial technologies to explore activity-based retail food environments. *Spatial and spatio-temporal epidemiology* 3, 4 (2012), 287–295.
- [18] Stefany Coxé, Stephen G West, and Leona S Aiken. 2009. The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of personality assessment* 91, 2 (2009), 121–136.

- [19] Jonathan F Deiches, Timothy B Baker, Stephanie Lanza, and Megan E Piper. 2013. Early lapses in a cessation attempt: lapse contexts, cessation success, and predictors of early lapse. *nicotine & tobacco research* 15, 11 (2013), 1883–1891.
- [20] Anind K Dey. 2001. Understanding and using context. *Personal and ubiquitous computing* 5, 1 (2001), 4–7.
- [21] Katherine Ellis, Jacqueline Kerr, Suneeta Godbole, Gert Lanckriet, David Wing, and Simon Marshall. 2014. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. *Physiological measurement* 35, 11 (2014), 2191.
- [22] Emre Ertin, Nathan Stohs, Santosh Kumar, Andrew Raij, Mustafa al’Absi, and Siddharth Shah. 2011. AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. ACM, 274–287.
- [23] Yoav Freund, Robert E Schapire, et al. 1996. Experiments with a new boosting algorithm. In *icml*, Vol. 96. Citeseer, 148–156.
- [24] Zhongliang Fu, Zongshun Tian, Yanqing Xu, and Changjian Qiao. 2016. A two-step clustering approach to extract locations from individual GPS trajectory data. *ISPRS International Journal of Geo-Information* 5, 10 (2016), 166.
- [25] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* (1969), 424–438.
- [26] Paul D Groves, HFS Martin, Kimon Voutsis, DJ Walter, and Lei Wang. 2013. Context detection, categorization and connectivity for advanced adaptive integrated navigation. The Institute of Navigation.
- [27] Chad J Gwaltney, Saul Shiffman, Mark H Balabanis, and Jean A Paty. 2005. Dynamic self-efficacy and outcome expectancies: prediction of smoking lapse and relapse. *Journal of abnormal psychology* 114, 4 (2005), 661.
- [28] Timothy Hnat, Syed Monowar Hossain, Nasir Ali, Simona Carini, Tyson Condie, Ida Sim, Mani B Srivastava, and Santosh Kumar. 2017. mCerebrum and Cerebral Cortex: A Real-time Collection, Analytic, and Intervention Platform for High-frequency Mobile Sensor Data.. In *AMIA*.
- [29] Tim Horberry, Janet Anderson, Michael A Regan, Thomas J Triggs, and John Brown. 2006. Driver distraction: The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. *Accident Analysis & Prevention* 38, 1 (2006), 185–191.
- [30] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. Vol. 398. John Wiley & Sons.
- [31] Syed Monowar Hossain, Timothy Hnat, Nazir Saleheen, Nusrat Jahan Nasrin, Joseph Noor, Bo-Jhang Ho, Tyson Condie, Mani Srivastava, and Santosh Kumar. 2017. mCerebrum: A Mobile Sensing Software Platform for Development and Validation of Digital Biomarkers and Interventions. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. ACM, 7.
- [32] Karen Hovsepian, Mustafa al’Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 493–504.
- [33] Jidong Huang, Zongshuan Duan, Julian Kwok, Steven Binns, Lisa E Vera, Yoonsang Kim, Glen Szczypka, and Sherry L Emery. 2019. Vaping versus JUULing: how the extraordinary growth and marketing of JUUL transformed the US retail e-cigarette market. *Tobacco control* 28, 2 (2019), 146–151.
- [34] Michael A Ichiyama and Marc I Kruse. 1998. The social contexts of binge drinking among private university freshmen. *Journal of Alcohol and Drug Education* 44, 1 (1998), 18.
- [35] Anita Jansen. 1998. A learning model of binge eating: cue reactivity and cue exposure. *Behaviour research and therapy* 36, 3 (1998), 257–272.
- [36] Alireza Karbasivar and Hasti Yarahmadi. 2011. Evaluating effective factors on consumer impulse buying behavior. *Asian Journal of Business Management Studies* 2, 4 (2011), 174–181.
- [37] Thomas R Kirchner, Jennifer Cantrell, Andrew Anesetti-Rothermel, Ollie Ganz, Donna M Vallone, and David B Abrams. 2013. Geospatial exposure to point-of-sale tobacco: real-time craving and smoking-cessation outcomes. *American journal of preventive medicine* 45, 4 (2013), 379–385.
- [38] Mingqi Lv, Ling Chen, and Gencai Chen. 2012. Discovering personally semantic places from GPS trajectories. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 1552–1556.
- [39] Abhinav Mehrotra and Mirco Musolesi. 2018. Using autoencoders to automatically extract mobility features for predicting depressive states. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 127.
- [40] Ganesan Muruganantham and Ravi Shankar Bhakat. 2013. A review of impulse buying behavior. *International Journal of Marketing Studies* 5, 3 (2013), 149.
- [41] Felix Naughton, Sarah Hopewell, Neal Lathia, Rik Schalbroeck, Chloë Brown, Cecilia Mascolo, Andy McEwen, and Stephen Sutton. 2016. A context-sensing mobile phone app (Q sense) for smoking cessation: a mixed-methods study. *JMIR mHealth and uHealth* 4, 3 (2016), e106.
- [42] John Ashworth Nelder and Robert WM Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* 135, 3 (1972), 370–384.
- [43] Maria A Oquendo, Hanga Galfalvy, Stefani Russo, Steven P Ellis, Michael F Grunebaum, Ainsley Burke, and J John Mann. 2004. Prospective study of clinical predictors of suicidal acts after a major depressive episode in patients with major depressive disorder or

- bipolar disorder. *American Journal of Psychiatry* 161, 8 (2004), 1433–1441.
- [44] Abhinav Parate, Meng-Chieh Chiu, Chaniel Chadowitz, Deepak Ganesan, and Evangelos Kalogerakis. 2014. Risq: Recognizing smoking gestures with inertial sensors on a wristband. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. ACM, 149–161.
- [45] Toby G Pavey, Nicholas D Gilson, Sjaan R Gomersall, Bronwyn Clark, and Stewart G Trost. 2017. Field evaluation of a random forest activity classifier for wrist-worn accelerometer data. *Journal of science and medicine in sport* 20, 1 (2017), 75–80.
- [46] Jennifer L Pearson, Amanda Richardson, Raymond S Niaura, Donna M Vallone, and David B Abrams. 2012. e-Cigarette awareness, use, and harm perceptions in US adults. *American journal of public health* 102, 9 (2012), 1758–1766.
- [47] Jane Powell, Lynne Dawkins, Robert West, John Powell, and Alan Pickering. 2010. Relapse to smoking during unaided cessation: clinical, cognitive and motivational predictors. *Psychopharmacology* 212, 4 (2010), 537–549.
- [48] Anna Pulakka, Jaana I Halonen, Ichiro Kawachi, Jaana Pentti, Sari Stenholm, Markus Jokela, Ilkka Kaate, Markku Koskenvuo, Jussi Vahtera, and Mika Kivimäki. 2016. Association between distance from home to tobacco outlet and smoking cessation and relapse. *JAMA internal medicine* 176, 10 (2016), 1512–1519.
- [49] Valentin Radu, Panagiota Katsikouli, Rik Sarkar, and Mahesh K Marina. 2014. A semi-supervised learning approach for robust indoor-outdoor detection with smartphones. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*. ACM, 280–294.
- [50] Jerry H Ratcliffe. 2004. Geocoding crime and a first estimate of a minimum acceptable hit rate. *International Journal of Geographical Information Science* 18, 1 (2004), 61–72.
- [51] Lorraine R Reitzel, Ellen K Cromley, Yisheng Li, Yumei Cao, Richard Dela Mater, Carlos A Mazas, Ludmila Cofta-Woerpel, Paul M Cinciripini, and David W Wetter. 2011. The effect of tobacco outlet density and proximity on smoking cessation. *American Journal of Public Health* 101, 2 (2011), 315–320.
- [52] Alex Rodriguez and Alessandro Laio. 2014. Clustering by fast search and find of density peaks. *Science* 344, 6191 (2014), 1492–1496.
- [53] Dennis W Rook and Robert J Fisher. 1995. Normative influences on impulsive buying behavior. *Journal of consumer research* 22, 3 (1995), 305–313.
- [54] Adam Sadilek and Henry Kautz. 2013. Modeling the impact of lifestyle on health at scale. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 637–646.
- [55] Nazir Saleheen, Amin Ahsan Ali, Syed Monowar Hossain, Hillol Sarker, Soujanya Chatterjee, Benjamin Marlin, Emre Ertin, Mustafa Al’Absi, and Santosh Kumar. 2015. puffMarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 999–1010.
- [56] Hillol Sarker, Matthew Tyburski, Md Mahbubur Rahman, Karen Hovsepian, Moushumi Sharmin, David H Epstein, Kenzie L Preston, C Debra Furr-Holden, Adam Milam, Inbal Nahum-Shani, et al. 2016. Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 4489–4501.
- [57] Edward Sazonov, Kristopher Metcalfe, Paulo Lopez-Meyer, and Stephen Tiffany. [n. d.]. RF hand gesture sensor for monitoring of cigarette smoking. In *2011 Fifth International Conference on Sensing Technology*. IEEE, 426–430.
- [58] K Schag, J Schönleber, M Teufel, S Zipfel, and KE Giel. 2013. Food-related impulsivity in obesity and Binge Eating Disorder—a systematic review. *Obesity Reviews* 14, 6 (2013), 477–495.
- [59] Philipp M Scholl, Nagihan Kücükylidiz, and Kristof Van Laerhoven. 2013. When do you light a fire? Capturing tobacco use with situated, wearable sensors. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. 1295–1304.
- [60] Roberto Secades-Villa, Victor Martínez-Loredo, Aris Grande-Gosende, and José Ramón Fernández-Hermida. 2016. The relationship between impulsivity and problem gambling in adolescence. *Frontiers in Psychology* 7 (2016), 1931.
- [61] Volkan Senyurek, Masudul Imtiaz, Prajakta Belsare, Stephen Tiffany, and Edward Sazonov. 2019. Cigarette Smoking Detection with An Inertial Sensor and A Smart Lighter. *Sensors* 19, 3 (2019), 570.
- [62] Cindy Shearer, Daniel Rainham, Chris Blanchard, Trevor Dummer, Renee Lyons, and Sara Kirk. 2015. Measuring food availability and accessibility among adolescents: Moving beyond the neighbourhood boundary. *Social Science & Medicine* 133 (2015), 322–330.
- [63] Saul Shiffman. 2005. Dynamic influences on smoking relapse process. *Journal of personality* 73, 6 (2005), 1715–1748.
- [64] Saul Shiffman, Jean A Paty, Maryann Gnys, Jon A Kassel, and Mary Hickcox. 1996. First lapses to smoking: within-subjects analysis of real-time reports. *Journal of consulting and clinical psychology* 64, 2 (1996), 366.
- [65] Muhammad Shoaib, Ozlem Durmaz Incel, Hans Scholten, and Paul Havinga. 2018. Smokesense: Online activity recognition framework on smartwatches. In *International conference on mobile computing, applications, and services*. Springer, 106–124.
- [66] Roger W Sinnott. 1984. Virtues of the Haversine. *Sky Telesc.* 68 (1984), 159.
- [67] Scott L Stephens. 2005. Forest fire causes and extent on United States Forest Service lands. *International Journal of Wildland Fire* 14, 3 (2005), 213–222.
- [68] Julia M Townshend, Nicolas Kambouropoulos, Alison Griffin, Frances J Hunt, and Raffaella M Milani. 2014. Binge drinking, reflection impulsivity, and unplanned sexual behavior: impaired decision-making in young social drinkers. *Alcoholism: Clinical and Experimental Research* 38, 4 (2014), 1143–1150.
- [69] Jan Van den Broek. 1995. A score test for zero inflation in a Poisson distribution. *Biometrics* (1995), 738–743.

- [70] Frank van Diggelen. 2002. Indoor GPS theory & implementation. In *2002 IEEE Position Location and Navigation Symposium (IEEE Cat. No. 02CH37284)*. IEEE, 240–247.
- [71] Michael W Wiederman and Tamara Pryor. 1996. Substance use and impulsive behaviors among adolescents with eating disorders. *Addictive behaviors* 21, 2 (1996), 269–272.
- [72] Pin Wu, Jun-Wei Hsieh, Jiun-Cheng Cheng, Shyi-Chyi Cheng, and Shau-Yin Tseng. 2010. Human smoking event detection using visual interaction clues. In *2010 20th International Conference on Pattern Recognition*. IEEE, 4344–4347.
- [73] Zhi-Qiang Zeng, Hong-Bin Yu, Hua-Rong Xu, Yan-Qi Xie, and Ji Gao. 2008. Fast training support vector machines using parallel sequential minimal optimization. In *2008 3rd international conference on intelligent system and knowledge engineering*, Vol. 1. IEEE, 997–1001.
- [74] Vincent W Zheng, Yu Zheng, Xing Xie, and Qiang Yang. 2010. Collaborative location and activity recommendations with GPS history data. In *Proceedings of the 19th international conference on World wide web*. ACM, 1029–1038.
- [75] Yu Zheng. 2015. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)* 6, 3 (2015), 29.
- [76] Yu Zheng, Like Liu, Longhao Wang, and Xing Xie. 2008. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 247–256.