# Scaling Crowdsourcing with Mobile Workforce : A Case Study with Belgian Postal Service

Peer-reviewed author version

Acer Günay, Utku; van den Broeck, Marc; Forlivesi, Claudio; HELLER, Florian & Kawsar, Fahim (2019) Scaling Crowdsourcing with Mobile Workforce : A Case Study with Belgian Postal Service. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3(2) (Art N° 35).

DOI: 10.1145/3328906
Handle: http://hdl.handle.net/1942/28245

# Scaling Crowdsourcing with Mobile Workforce : A Case Study with Belgian Postal Service

UTKU GÜNAY ACER and MARC VAN DEN BROECK, Nokia Bell Labs, Antwerp, Belgium
CLAUDIO FORLIVESI, ING, Belgium
FLORIAN HELLER, Hasselt University - tUL - Flanders Make, Belgium
FAHIM KAWSAR, Nokia Bell Labs, Cambridge, UK and TU Delft, Netherlands

Traditional urban-scale crowdsourcing approaches suffer from three caveats - lack of complete spatiotemporal coverage, lack of accurate information and lack of sustained engagement of crowd workers. In this paper, we argue that these caveats can be addressed by embedding crowdsourcing tasks into the daily routine of mobile workforces that roam around an urban area. As a use case, we take the bpost who deliver the letters and parcels to the citizens across entire Belgium. We present a study that explores the behavioural attributes of these mobile postal workers both quantitatively (6.3K) and qualitatively (6) to assess the opportunity of leveraging them for crowdsourcing tasks. We report their mobility pattern, workflow, and behavioural traits which collectively inform the design of a purpose-built crowdsourcing solution. In particular, our solution operates on two key techniques - route augmentation, and on-wearable interruptibility management. Together, these mechanisms enhance the spatial coverage, response accuracy and increase workers' engagement with crowdsourcing tasks. We describe these principal components in a wearable smartwatch application supported by a data management infrastructure. Finally, we report a first-of-its-kind real-world trial with ten postal workers for two weeks to assess the quality of road signs at the city centre of Antwerp. Our findings suggest that our solution was effective in achieving 89% spatial coverage and increasing response rate (83.6%) and accuracy (100%) of the crowdsourcing tasks. Although limited in scale, these and the rest of our findings highlight the way of building an efficient and purposeful crowdsourcing solution of the future.

CCS Concepts: • **Information systems** → **Crowdsourcing**; • **Human-centered computing** → *Ubiquitous computing*; *Mobile computing*; Mobile devices.

Additional Key Words and Phrases: mobile crowdsourcing, wearable computing, interruptibility, behaviour modelling



## 1 INTRODUCTION

With the emergence of the Internet of Things (IoT), local authorities increasingly use connected sensors to understand their cities in order to plan for the future societal, economic, and environmental well-being of their citizens. In the past few years, a number of cities have deployed diverse Internet of Things (IoT) systems in urban spaces that offer a quantitative view of the urban landscape (e.g., Noise, Air Pollution, Mobility, etc.). However, such systems fail to capture the qualitative aspects of the urban landscape including the relationship between the citizens and their cities.

In order to better understand the qualitative aspects of their cities, municipalities encourage their citizens to use crowdsourcing tools to provide information about the areas they live in and support a more sustainable habitat. Through crowdsourcing, user-generated content can be collected and analyzed to identify and address problems urban areas face. The rise of location-based services and the widespread use of mobile devices makes such citizen participation more feasible and accessible [17]. In addition, companies such as Gigwalk[1], FieldAgent[2] and TaskRabbit[3] pay users to perform tasks such as price checks, audits, etc. In order to motivate people to

[1] http://www.gigwalk.com
[2] https://www.fieldagent.net
[3] https://www.taskrabbit.com

Authors' addresses: Utku Günay Acer, utku_gunay.acer@nokia-bell-labs.com; Marc van den Broeck, marc.van_den_broeck@nokia-bell-labs.com, Nokia Bell Labs, Antwerp, Belgium ; Claudio Forlivesi, ING, Belgium; Florian Heller, florian.heller@uhasselt.be, Hasselt University - tUL - Flanders Make, Belgium; Fahim Kawsar, fahim.kawsar@nokia-bell-labs.com, Nokia Bell Labs, Cambridge, UK and TU Delft, Netherlands.

participate in information collection, such mechanisms may provide monetary incentives to people for each task they complete.

Traditional approaches to urban mobile crowdsourcing suffer from three major problems. Because participation is voluntary, the citizens mostly contribute information at times/ locations that are convenient to them. This leads to the *lack of spatiotemporal coverage* where most of the data is collected in highly populated areas like city centers at busy hours. The volume of data associated with suburban areas and off-hours is, on the other hand, much smaller. Another issue of mobile crowdsourcing is the *lack of accuracy* in the collected information as there is no way of inspecting the users' input. As a result, these systems cannot guarantee the quality of the data. Finally, such systems have to deal with the *lack of sustained engagement* where workers exhibit early enthusiasm but lose interest and drop out over time. This leads to only a small group of workers completing a large portion of the tasks[33], and potentially, to a bias in the collected data.

We argue that mobile workforces that roam around the city (and the country in the larger scale) can overcome all these issues if crowdsourcing tasks are embedded into the daily routines of the workers. As a case study, we take workers of Belgian Postal Services, *bpost*, that collectively traverse the entire country to deliver letters and parcels. bpost employs the largest payroll in Belgium including postal workers that deliver letters, packets, and parcels to the 4.7 million households across the country. These workers go on 9200 mail rounds, delivering up to 9 million items every day[4].

Since a workforce such as that of bpost roams around the entire country every day, utilizing them can address the coverage problem. As the workers serve in the same rounds every day, they become very aware of what happens in the area they serve and they are able to grasp the perception of citizens. This enables them to accurately interpret qualitative phenomena and report it. By embedding these crowd-sourcing tasks to their daily rounds instead of collecting data from volunteers also ensures that data collection is sustained over time.

We present a study that quantitatively and qualitatively explores the behavioural attributes of bpost workforce. It includes an analysis of a dataset that reports information from more than 6300 postal workers delivering parcels to the households. In addition, we present a contextual study of six workers to assess opportunities in their behaviour for crowdsourcing tasks.

Our analysis of their mobility pattern, workflow, and behavioural attributes suggest that postal workers can engage with mobile devices to carry out crowdsourcing tasks without affecting their primary tasks of distributing items to citizens. However, in order to ensure complete coverage, it is essential to augment their routes with crowdsourcing tasks. In addition, workers often experience situational disabilities that may prevent them from using their devices. As a result, crowdsourcing tasks need to be presented to them at opportune moments when they are indeed available to provide answers to queries.

Based on our observations, we develop a purpose-built crowdsourcing solution that operates on two key techniques - route augmentation, and on-wearable interruptibility management. The former utilizes an efficient algorithm to embed a crowdsourcing task into workers' daily route optimally. The latter models spatiotemporal physical and conversational activities of a postal worker using multi-modal sensing on a smartwatch to identify opportune moments for engaging the worker with crowdsourcing tasks. Collectively, these techniques enhance the spatial coverage and increase the accurate response rate of crowdsourcing tasks presented to the postal workers at the right location at the right time. We then take this solution in a real-world trial with ten postal workers for two weeks in the city of Antwerp. Formal evaluation of our deployment shows that our tailor-made solution increases the spatial coverage of the tasks to 89% up from 48%, increases response rate up to 83% as opposed to 57% and most importantly, providing 100% accurate responses. We also observed the engagement to increase over time. We further qualified these quantitative findings with semi-structured interviews and reflected on workers experience and impression of such a solution. Although due to the small scale and limited participants,

[4]http://corporate.bpost.be/about-us/bpost-at-a-glance?sc_lang=en

our results can not be considered conclusive, they carry profound design suggestions towards building a powerful crowd-sourcing solution of the future.

The rest of the paper is organized as follows: We present the related work in Section 2. Section 3 describes our study methodology and participants. Section 4 presents a qualitative and quantitative analysis of the behaviour of postal workers . We lay out our design considerations for our crowdsourcing solution based on the behavioural analysis and describe various components and the end-to-end system in Section 5. We then present the performance of different components and report the first-of-its-kind real-world study with postal workers for crowdsourcing in Section 6 and 7. We offer a reflection on the primary lessons that emerged from this study acknowledging its limitations in Section 8 before concluding the paper in Section 9.

## 2 RELATED WORK AND BACKGROUND

Monitoring various parameters, such as traffic, air quality, or noise, is an important source of information for local authorities to apply effective counter-measures if needed. While, to some degree, this can be done through a network of sensors distributed throughout the city, such an approach entails high costs particularly if one requires high spatial coverage, i.e. a dense network. With crowdsourcing, it is possible to leverage the ubiquity of the smart devices and use sensors on these devices to collect data. This intrinsically implies human participation as these the holders of the smart devices need to volunteer in data collection. Yet, the biggest advantage of a human in the loop, is that such spatial crowdsourcing approaches can also measure qualitative data.

One aspect of crowdsourcing, as formulated by Howe [16], is to break down complex tasks into small bits that are easy to solve by humans [13], even with only minimal contextual knowledge. Platforms like the Amazon Mechanical Turk (AMT)[5] enable programmatic access to a large crowd of human workers to solve tasks computers are not yet able to complete effectively. With mobile phones or smartphones being universally available, the idea of mobile crowdsourcing or spatial crowdsourcing is to take this approach into the real world and integrating a geospatial component to the task information, e.g., have the task be executed at a specific location [1, 5, 23]. Mobile crowdsourcing applications cover a broad range of use cases, from simple tasks like taking a picture of a specific location [1], monitoring traffic[6], helping with setting up furniture[7], to maintaining network infrastructure in rural areas [19].

The value of such systems is that it can sense both objective and subjective data. For example, the presence of a pothole [10] can be estimated by looking at accelerometer values, but the perceived quality of the road [40] can differ from such sensor-based assessment. However, certain aspects need detailed attention for the service to return useful results, which we address in the following

### 2.1 Scaling & Coverage

The spatial coverage of mobile crowdsourcing platforms ranges from a local university campus as for many research projects (e.g., [4, 5, 21]), distinct cities, to country-wide or even worldwide deployment for some commercial platforms. While the service might be available in large areas, the actual spatial coverage by the crowd workers is a challenging aspect in mobile crowdsourcing. Members of platforms that only cover a large building [5] or a University Campus [21] do not need to take large detours for successful task completion. On a larger scale, responses to mobile crowdsourcing tasks are biased towards popular places at rush-hours due to various factors like the personal interest of the participants [37], or their socio-economical status [41]. For example, the level of detail in Open Streetmap (OSM) data is usually much higher in urban areas than in rural areas [37]. However, many applications require an unbiased coverage, e.g., to get an overview after a

[5] https://www.mturk.com/
[6] https://www.waze.com
[7] https://www.taskrabbit.com

disaster [27]. An analysis by Musthag et al. [33] on an active mobile crowdsourcing platform showed that a very small number of users (10%) of so-called super agents account for over 80% of the completed tasks. The platform under investigation used a pull-based task distribution model, which, although preferred by a majority of users [1], means that users have to organize their tasks and routes themselves. One approach to achieving better spatial coverage and throughout the workers' base is to analyze the users' paths and push task requests that require only a minimal detour to the predicted path [6–8]. This increases the likelihood for successful task completion and increases fairness among the crowd workers.

In the context of smart cities, a series of projects have been deployed to investigate the usefulness of crowdsourcing or crowdsensing. FixMyStreet[8], the PotholePatrol [10], Nericell [32], and CommuniSense [40] all focus on assessing the state of roads in order to repair them. Nericell [32] monitors traffic conditions such as congestions, or stop-and-go traffic, while SignalGuru [26] determines traffic signal patterns to support a continuous driving experience. CrowdOut [2] is a tool to report traffic related infringements and problems to local authorities, such as illegally parked cars, broken signs, and signals, or road quality in general. Ear-Phone [38] and NoiseTube [28] use mobile phones to measure noise levels and create a noise map of urban areas. Safecast[9] collects radiation and air-quality data from all over the world to form a transparent open-data platform. Ushahidi[10] is a platform introduced to map reports of violence in Kenya and has evolved into a large platform for citizen participation.

## 2.2 Motivation & Incentives

For all these services, it is important to keep the crowd workers engaged with the platform, i.e., having workers complete tasks and have a crowd worker complete new tasks over time. A major factor for successful task completion is a good task assignment algorithm that takes into account the worker's context [4]. According to Alt et al. [1], users prefer pulling tasks from a database and prefer tasks that are located close to their home. However, pull-based mechanisms defer the combination of tasks and planning a route with minimal detour entirely to the user [33]. Push-based mechanisms can be fed with various information about the workers and plan and combine tasks to a meaningful bundle [21]. To increase the likelihood of tasks being completed while participants are on the move, Chen et al. [6, 7] implemented a system that assigns series of tasks to participants based on a prediction of their path. As a technical solution to increase the response rate, Vaish et al. [44] propose to reduce the complexity of tasks to a minimum and integrate the completion dialogue into the unlock-procedure of a smartphone. McSense [4] also integrates device context such as the phone's battery level into the task distribution, as workers with only minimal battery power left will probably not accept long tasks. Ren et al. [39] suggest to include social context into the distribution as well. Their experiment shows that the quality of the reports increases if the assigned tasks are from an area the worker has expressed interest for. Kandappu et al. [22] achieved 95% completion rate for tasks by creating a social network that allows workers to refer tasks they are not able to complete themselves to other workers they know well. The social component of personally knowing the referrer of a task, also increased the average detour a worker was willing to take in order to complete the task.

Wang et al. [45] assign a reputation to each crowd worker that influences both the chance to get a task assigned and the reward for successfully completing that task. This generates an intrinsic motivation to provide high-quality data as the other workers are invited to play a game to verify the accuracy of the data with the possibility to earn rewards themselves. Palacin-Silva et al. [34] compared a gamified version of an ice-reporting app for water bodies in Finland to a traditional implementation. The gamified version included an interactive map, task assignments as stories, challenges, points, leaderboards, and a feedback module. The overall involvement

[8] https://fixmystreet.org
[9] https://blog.safecast.org
[10] https://www.ushahidi.com

was significantly higher with the gamified app than with the traditional app with workers being more active, and fewer people dropping out of the worker pool.

### 2.3 Managing Interruptions

Interruptions such as notifications on mobile phones result in context switches from a primary task to another that is requested by the notification and back. These context switches entail a considerable mental demand that, depending on when the interruption happens, can lead to forgetting what the primary task actually was [9].

With the increasing popularity of mobile devices, determining the user's context to find opportune moments to present notifications became a multimodal task. As early PDAs did not have the rich set of sensors current smartphones contain, Ho et al. [14] attached two inertial measurement units (IMU) to the upper thigh and ankle of the user. These sensor readings were used to determine changes in physical activity correlating with a user-initiated task switch, and subsequently being a moment where an additional task does not interrupt an existing primary one. Similarly, Fischer et al. [12] inspected mobile-phone activity with the goal to determine opportune moments to trigger specific user responses through notifications. They found that the moment right after a mobile phone interaction, i.e., the end of a call, is the most suitable for that purpose. The user has to switch context to a different task anyway, meaning that the interruption by the notification is minimized.

To increase the chances of a notification to be perceived, recent projects aim at inferring the user's context from various smartphone sensors. PrefMiner [31] is an Android notification management library. Based on the notification title and the user's location, the library can predict whether a notification is likely to be dismissed. The user can then further adapt these automatically inferred rules to her needs.

The system by Pielot et al. [36] determines a user's state of boredom. Based on factors such as recency of last communication, time of day, and demographics, the system determines opportune moments to trigger a notification, which results in an increased reaction rate. Mathur et al. [30] used EEG readings to determine mental workload and to correlate it to a set of dimensions from smartphone sensors, demographics, usage context, and time of day. Based on these results, they implemented a classifier that qualifies the smartphone context with user engagement with a high accuracy. InterruptMe [35] is a library that learns and models interruptibility on the Android platform. For its development, the authors logged usage context based on smartphone data and asked for additional social context using a brief questionnaire that was triggered by the notification. Participants had to describe if they were in company or not, in what kind of activity they were engaged and how important that was, and how they felt. Overall, the authors conclude that not only context has to be taken into account, but also the recent interruption load in order to achieve better response rates.

We use the notification paradigm to efficiently embed crowdsourcing tasks into mobile workers' daily routines. In other words, tasks are presented to workers in the form of notifications without prompting any effort from the workers to manually search for tasks. We build upon prior work on interruptibility management for mobile devices to design a rule-based wearable-only app to present notifications at the opportune moments. The rules that drive the app operation are driven from a user study to understand the workers' behaviour while they deliver parcels.

### 2.4 Data Quality

One problem of mobile crowdsourcing is the bias in spatial coverage towards popular places [21] and the fact that potentially only a few workers contribute a large number of results [32]. TA$Ker is a real-world testbed for mobile crowdsourcing deployed on a University campus in Singapore with the goal to evaluate methods of compensation for these bias effects [21]. It demonstrates that users prefer multiple tasks being bundled and that increasing the reward for tasks assigned to less popular locations increases fairness across the worker revenue and decreases the spatial coverage bias. Ul Hassan et al. [43] proposed an adaptive task distribution algorithm

that increases coverage by selecting workers based on their location and willingness, which it learns over time. They show that this algorithm achieves a higher spatial coverage with fewer workers than the simple baseline algorithm.

Regarding the issue of trust in the accuracy of the crowdsourced data, several approaches have been presented. Kandappu et al. show that requiring a very fast response leads to more incorrect data [21]. One way to ensure the accuracy of the responses is to assign tasks to multiple workers and accept the most common response [24]. Assigning each worker a reputation, such as by Kazemi et al. [25] and Wang et al. [45] introduces an intrinsic motivation to provide good results. The system in [25] only accepts answers that have a task-related confidence level higher than a certain threshold, while the system in [45] adds a verification game component. Hoh et al. [15] achieved a better data quality by adjusting the rewards for correct data, in this case, information on free parking spots, depending on the availability of parking spots and workers. In contrast, Huang et al. [18] automatically determine the reputation of a device by comparing samples to the community average, thus solely on a data level without a human in the loop.

Janjua et al. [20] propose an approach of ensuring data integrity on a much lower level, by running critical software elements in a trusted execution environment that ensures that, e.g., the picture has been taken at the time and location its metadata suggests.

Using a mobile workforce as crowd workers establishes a baseline trust into the results. Bakewell et al. [3] analyzed how a mobile workforce, in this case, field engineers, responds to being equipped with a set of Apps that track their progress on assigned tasks over the course of the day. While it empowers the workers' autonomy, it also increases the anxiety that the collected data will be used against them if their performance is insufficient for whatever reason.

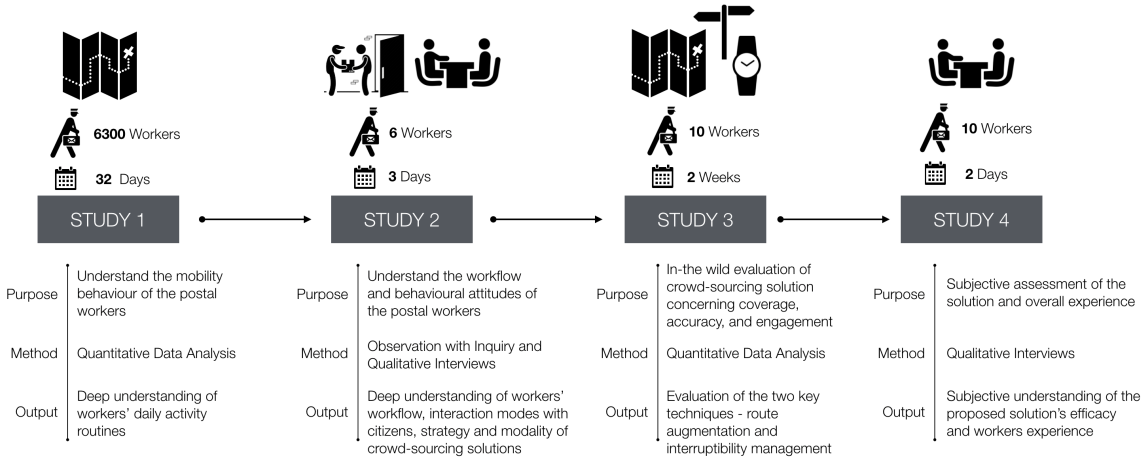



Fig. 1. Overall mixed-method methodology followed in this research including quantitative data analysis, contextual study, in-the-wild deployment and qualitative interviews

## 3 STUDY METHODOLOGY

The objective of this work is to engage a mobile workforce in crowd-sourcing tasks with an aim to address the caveats of conventional approaches, namely, lack of spatiotemporal coverage, lack of accurate information, and lack of sustainable engagement. We have approached this research in four stages applying mixed-method

study methodology and in-the-wild system deployment. The overall method is reflected in Figure 1. We began with a quantitative data analysis of 6.3K postal workers' activity traces to understand their mobility, and then we performed a contextual study with six workers to understand their workflow and behavioural attributes. Grounded on the findings of these analyses we develop a system with purpose-built components and evaluated them in a two weeks real-world deployment with ten postal workers and finally, we conclude this research with qualitative interviews with those ten workers to understand their experience and impression of such system and engagement. In the rest of this paper, we describe each of these studies in depth.

*Participants:* This research was conducted in the context of a national-scale project involving multiple stakeholders interested in the outcome of the work. One of the stakeholders was Belgian Postal Services, *bpost*, employing the largest payroll in Belgium including 15000 postal workers that deliver letters, packets, and parcels to the 4.7 million households across the country. bpost was interested in the project as it wants to enter a new market as a provider of citizen and city data for a variety of companies. As part of its innovation strategy, bpost has identified the potential of using its workforce in the field to collect a vast amount of data as well as to execute small activities beyond the traditional business of postal service as a new source of revenue stream. As such, they actively supported this work as they saw it as an excellent pilot opportunity.

The region of interest drove the selection of actual postal workers. As we discuss later, we wanted to cover both urban and rural areas and selected three potential regions - Antwerp, Brussels, and Sint-Katelijne-Waver. We conducted the contextual study in all three regions and the in-the-wild deployment study at Antwerp. bpost management decided the specific parts of these regions and allocated the workers for those regions as participants of this study. These workers, however, were not enforced instead the entire team of workers for those regions were informed, and the final participants joined the study willingly knowing this is a pilot study that might help their organisation and without any tangible incentives. bpost management did not change their daily routines or job description for the period of the study neither did they introduce any reward for their participation. As such, we should interpret the results reported hereafter as a reflection of postal workers who opt-in for this research as a gesture of goodwill towards their management and organisation.

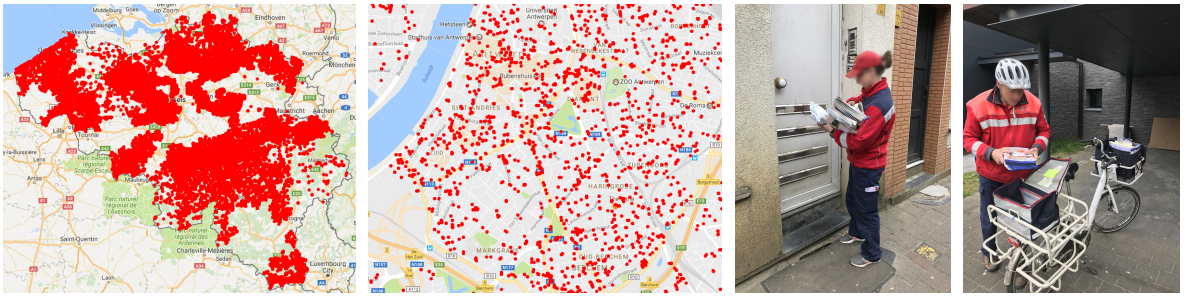
## 4 BEHAVIOURAL STUDY

We begin by reporting a mixed-method study to assess the feasibility of using a mobile postal workforce for crowdsourcing. First, we quantitatively analyze the mobility of postal workers to identify the situational opportunities for crowd-sourced tasks. Then we report a contextual study on the workflow and behaviour of the mobile postal workers. We seek to understand whether the behaviour of the postal workers during their rounds allows them willingly i) to engage with a mobile device and ii) with a citizen to perform a crowd-sourcing task, e.g., responding to a subjective question regarding their current location. The collective findings of this mixed-method study inform the design of our system for purposefully engage the postal workers in crowd-sourcing task.

### 4.1 Understanding Mobility

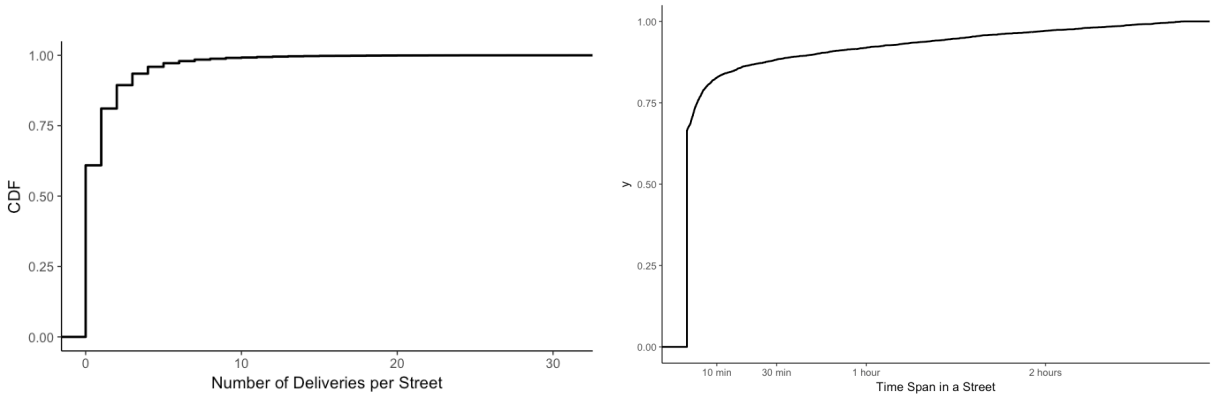
In this section, we present an analysis of a dataset provided by bpost. This dataset consists of a list of *events* from mobile terminals given to a subset of workers who deliver large packets and parcels to citizens. Such large items are delivered by workers who operate a motor vehicle such as vans, hence, we refer to these workers as *drivers*. The events in the dataset include scanning of barcodes on delivered items as well as other supplementary information such as charging of the device, etc. Events may also be geotagged with GPS coordinates.

The traces provide information about more than 6300 workers with around 350K events over the duration of 32 days. Approximately half of these events correspond to deliveries. However, not every event is tagged with GPS coordinates. On average, roughly 142K events per day include GPS coordinates and about 140K of them correspond to deliveries.



(a) Delivery locations across Belgium (b) Delivery locations in Antwerp (c) A worker on foot (d) A worker on a bike

Fig. 2. Representation of dataset used in the quantitative study and pictures from qualitative behaviour analysis




(a) (b)

Fig. 3. Distribution of number of deliveries and time difference between first and last delivery in each street. Most streets see only a small number of deliveries with short duration

As shown in Figure 2a, the deliveries with GPS coordinates are not uniformly distributed across the country. bpost geographically divides Belgium into 242 distribution centres, but not all of these centres provide valid GPS coordinates. However, those mail centres who provide valid GPS coordinates, communicate them for a large part of their deliveries. In Antwerp city centre, for example, 93% of the deliveries of our dataset have associated GPS coordinates, as shown in Figure 2b. For convenience, our further analysis focuses on the deliveries in this mail centre to evaluate the spatial coverage of drivers.

We use a local Nominatim[11] server to reverse geocode GPS coordinates to street level location information. Our analysis reveals that there is at least one delivery per day in 55% of the streets in the centre of Antwerp. The streets without a delivery are typically very short streets and/or with a very small number of residents, nevertheless, such streets are situated close to another street with a delivery.

Figure 3a shows the cumulative distribution function (CDF) of the number of deliveries in a street over the course of a day. As such, there are at most 3 deliveries in 93% of the streets in a day. Similarly, Figure 3b shows the distribution of the time difference between the first and last delivery on a street in a day. In 88% of the streets,

[11] http://wiki.openstreetmap.org/wiki/Nominatim

this time span is less than just 1 hour. For a few streets, there are parcels delivered in the afternoon rounds which leads to a larger time span.

These figures indicate that while the mobile workforce of bpost provides widespread spatial coverage, the temporal coverage is minimal since time spent in each street is usually only a fraction of the day.

Table 1. Key statistics on the behaviour of the postal workers

| | Mean | Median | Standard Deviation |
|---|---|---|---|
| Round Duration | 3.92 hrs | 3.99 hrs | 1.77 hrs |
| Number of Deliveries per Round | 35.73 | 32 | 21.59 |
| Time Between Two Deliveries | 6.78 min | 3.31 min | 12.44 min |
| Total Distance | 24.94 km | 19.71 km | 18.24 km |
| Distance Between Two Deliveries | 683.43 m | 479.1 m | 747.64 m |

The summary of key statistics about the mobility of the drivers is given Table 1. These values indicate that the drivers do not have to rush between deliveries to finish their jobs. Considering the typical distance between two delivery points and the time drivers spend between two deliveries (e.g., walking back to the truck after delivering a package), we concur that the workers will have time to engage with their mobile devices if prompted for crowd-sourcing tasks.

## 4.2 Understanding Workflow

In order to assess the workflow of the postal workers, we carried out an observational study with 6 postal workers in two urban settings, Brussels and Antwerp, and a rural area, Sint-Katelijne-Waver of Belgium. In each location, we accompanied two workers with a different mode of transportation. We conducted a user shadowing study with *fly-on-the-wall* and *observation with inquiry* techniques [29]. In the first part of the tour, we observed the postal workers without any comment or intervention. In the second part, we asked them questions on certain aspects related to their behaviour. We used a head-mounted camera to capture a workers round on video for later analysis. In addition, we used a GPS tracking application to track the route of the worker. Figures 2c and 2d show a couple of photographs taken during this study.

The mail rounds are broadly categorized into two groups based on the type of deliveries they make. In one group, workers deliver letters and small packets. In urban areas, workers in this group walk along their rounds with their caddies. In rural areas, they use small vehicles such as bikes and scooters to travel extended distances but they still walk among a group of houses. We refer to workers in this group as *pedestrians*. In the second group, workers use motor vehicles to carry and deliver large parcels. In addition, they deliver letters for premium customers, i.e., people/companies that pay to receive their mail before 9:00 am and high volume users, i.e., entities that receive a large number of mail items. These workers are referred to as *drivers*[12]. Drivers also distribute the letter bags of pedestrians to designated pick-up locations such as apartment buildings and depots, such that the pedestrians do not have to carry all letters at all times. Pedestrians pick up these bags left by the drivers along their rounds and deliver the items to their final destinations. In contrast to urban areas where a worker makes only one round per day, in rural areas, a single worker can complete multiple rounds using different modes of transportation, e.g., first on a car, then on a bike. In other words, a worker can be both a driver and a pedestrian over the course of the day. The summary of the mode of transportation of worker rounds is given in Table 2.

Pedestrians perform their round in the morning, typically completing the delivery of letters by midday. Most of the parcels are also delivered in the morning by drivers. In addition, there are also afternoon driver rounds,

[12]The previously studied dataset provides information about drivers.

Table 2. Mode of transportation for postal workers

| | Urban | Rural | Pedestrian (letters) | Driver (parcels) |
|---|---|---|---|---|
| Walk Only | ✓ | | ✓ | |
| Car Only | ✓ | | | ✓ |
| Bike/Scooter + Walk | | ✓ | ✓ | |
| Car + Walk (Two rounds) | | ✓ | ✓ | ✓ |

delivering a small number of parcels. These drivers also collect outgoing letters in mailboxes placed around the city.

The rounds of pedestrians are much shorter in comparison to those of drivers. They deliver letters to the same people almost every day. Hence, a pedestrian establishes contacts with individuals on a regular basis and is able to build trust relationships with residents along that round. This gives them opportunities to collect voluntary information, either through observation or through dialogue with the residents. A driver, on the other hand, can only establish such contacts with the representatives of premium and high volume customers that they see every day. They also frequently interact with proxies who can provide collective information such as concierges of large buildings since even though the addressee changes each time, a building with a large number of households is regularly visited by the worker. In addition, both groups of workers cross the same neighbourhoods every day. Hence, they quickly notice changes in the urban setting.

While a pedestrian does not have any time critical deliveries, drivers have a deadline for delivering mail for premium customers before 9:00 am. They also have a soft deadline for distributing pedestrian bags to the pick-up locations before 11:00 am. Still, we did not observe either of both groups of workers rush their deliveries. They frequently take breaks in the cafes and pubs along their round and have conversations with the people. Drivers in cities are an exception to this as it may be difficult for them to find parking spots for their vehicles but they still have time for conversations during deliveries.

Regarding potential interaction with mobile devices to collect information, the workers have situational disabilities that prevent them from using their mobile devices. Instead of single letters, pedestrians pick up a batch of letters from their caddies or bikes and distribute them to the mailboxes. Doing this, they leave their caddies or bikes behind and hold the letters in their hands. Figure 2c and Figure 2d represent such situational disabilities. Moreover, workers cannot use their devices if/when they operate the bikes and scooters. Similarly, drivers cannot engage with mobile devices while they are behind the steering wheel or carry parcels from the vehicle to the destination address. Still, there are opportune moments to engage with the devices. For example, it is possible to check mobile devices after pedestrians complete a batch and while they return to their caddies. In the drivers' case, most parcels require a signature and they have to wait for people to pick up the delivery after they ring the doorbell. They use their devices while waiting for the addressee to show up. All the workers noted that while it is possible they may not accept calls immediately, they get back to the caller as soon as they can.

The summary of our observations is presented in Table 3.

### 4.3 Understanding Attitude

After we finished the qualitative analysis, we conducted a small semi-structured interview with the six postal workers we followed on their daily round, a group including four males (aged between 40 and 62) and two females (aged between 40 and 45). Two of these workers were merely on foot, two were drivers only, one of them performed a walking tour after finishing a driving round and one used a scooter on the round. The walk only and car only workers did rounds in urban settings whereas the other two had rounds in a rural area. The goal

Table 3. Summary of Postal Worker behaviour

| | Pedestrian | Driver |
|---|---|---|
| Delivery with Deadlines | | ✓ |
| Voluntary Opportunity | ✓ | |
| Proxy Interaction | ✓ | ✓ |
| Situational Disadvantages | ✓ | ✓ |
| Opportunity to Use Mobile Device | ✓ | ✓ |

of the interview was to find out i) if postal workers are indeed willing to take on the extra task of collecting information in their area and ii) if so, what kind of tools and devices they would prefer to complete this task.

Regarding the first question, all six persons responded that they would agree to answer contextual questions when they are out on the street delivering letters and packages. In addition, they said they could ask questions to residents they serve. More specifically, we asked whether they would feel comfortable asking people about their perception of the subjective properties of their neighbourhood, e.g., the safety of the area they live in, etc. All subjects said they would not have any objections to such queries. However, two workers added that they do not see the value of asking this kind of questions, but they would do it anyway if requested by their management. None of the workers thought their primary task would prevent them from collecting such information.

We also presented them with possible use cases where they would receive a question at a certain location during their tour. These use cases include both quantitative and qualitative questions, and their answers require both objective observations (presence of a solar panel, condition of traffic signs, etc.), and subjective perception (safely or cleanliness of a street, etc.). One of the questions we suggested was the following: *Does the house at this address have solar panels on the roof?*.

After explaining that this kind of question could cause harm to residents because failure to report a solar panel installation could result in paying a high penalty for tax avoidance, five of them responded that they would not endanger their social trust relationship with the people they serve. One remarked that he would not want people to think he was a traitor.

Then, we inquired about the type of device they would like to use to collect this information, and we presented them with four possible choices: (a) smartphone, (b) smartwatch, (c) tablet, (d) in-house bpost scanning device. The most popular choice was the bpost terminal. The workers agreed that they were not tech savvy. The drivers, in particular, noted that they would feel more comfortable with a device they already know and operate. Their second choice was to use a smartwatch, followed by the smartphone. None of the respondents opted for the use of a tablet.

## 4.4 Summary and Design Implications

Our mixed-method study covered a wide range of issues in understanding postal workers mobility, workflow, their relationship with citizens, attitude to mobile crowd-sourcing tasks, etc. In this section, we offer a reflection on the key findings of our study, and in particular, we discuss the type of crowdsourcing tasks that are suitable for postal workers (the what aspect), and how these tasks can be embedded in their daily work routines (the where and when aspects).

*4.4.1 What - Task Dynamics.* Learning from our study and past work [4], we provide here an overview of potential tasks for mobile postal workers. We discuss these tasks from four perspectives:

(1) **Spatial Range:** Mobile crowdsourcing tasks are associated to a specific location. This location may be a certain *point*, i.e., latitude/longitude or an address, or a larger *region* such as a neighbourhood, a city or even

an entire country. A point-level question, for example, may ask the worker to assess the noise level on a public square. While some region-level tasks may require only one response as it is typically homogeneous across the region, e.g., 'How friendly are people in this neighbourhood?', others may be formulated as a collection of point-level questions, e.g., 'How clean are streets in this city?' With a workforce that yields a large spatial coverage, such as postal worker of bpost, a mobile crowdsourcing framework is perfectly situated to collect responses to both point-level and region-level tasks.

(2) **Periodicity:** A crowdsourcing task may be executed *one time*, e.g., check whether a house in a particular address has solar panels, or they may be *repetitive*, i.e., daily, weekly monthly, so that trends of interest can be grasped by interested parties to track a phenomenon of interest, e.g., how the cleanliness of a street changes over time. In addition, a task may be *demand driven*, and required to be carried out at a particular time, e.g., 'Is the train station very crowded at noon?'. Even though the temporal coverage is poor for a mobile workforce, spatial coverage is consistently high across the days that the workforce is in operation. This allows collecting responses to both one-time and repetitive queries. However, the granularity of the repetitive queries cannot be very high since the temporal coverage is poor. A system leveraging such workers cannot expect hourly updates whereas it is possible to collect daily responses to temporal queries. Due to poor temporal coverage on the other hand, demand driven tasks cannot be completed by the workforce since the presence of a worker at a particular location at a certain time cannot be guaranteed.

(3) **Response Characteristics:**Our observations show that, in case of bpost, postal worker sort the letters and parcels they deliver at the beginning of the day, and plan their rounds accordingly. Such a workforce can ensure that a worker is present at a location over the course of a day but cannot indicate the actual time that they are there. Hence, one cannot expect responses to *real-time* tasks that require immediate responses to questions as soon as they are generated. The tasks need to be *delay tolerant* and introduced to the system at least before the workers start their rounds so that the tasks are incorporated to their rounds efficiently.

(4) **Sociality:**We have observed that the workers frequently interact with the public and this presents an opportunity for the workers to act as proxies and direct questions to citizens. Hence, they can carry out tasks that target a *specific* group of people, e.g., 'Is the transportation easy for elderly people?'. This group could be defined by demographic characteristics, such as age, gender, or ethnicity. Workers can draw from their detailed knowledge of the local communities to determine candidates fitting the designated target audience profile as they have a good understanding about the people they serve. Other tasks may be *generic* and directed to the general public to assess how a phenomenon is perceived, e.g., 'Do you find the city centre safe?'. Our observations have shown that the workers establish a trust-based relationship with citizens and as long as the queries they present to citizens do not jeopardize this relationship, they are willing to collect responses from them. Therefore, the tasks regarding the demographics should be prepared well so that the recipients do not feel threatened.

*4.4.2 Where - Spatial Dynamics.* Our study shows that bpost workforce provides widespread spatial coverage. In addition, the work they perform gives them a comfortable and loose schedule during which they can use their mobile devices. Typically, workers have daily routines that they follow along their rounds. They can easily detect any changes in the area they serve. While the workforce provides wide spatial coverage, for complete coverage, the workers need to be steered towards locations of interest, i.e. their rounds need to be augmented task locations. The route augmentation, however, needs to adhere to the schedule of the postal worker.

*4.4.3 When - Temporal Dynamics.* We have observed that the workers are able to engage with mobile devices along their rounds. However, they have situational disadvantages when they carry/hold items for interacting

with devices. Therefore, crowdsourcing tasks need to presented in an *opportune moment* ideally considering the situational context of the worker.

In the next section, we discuss the design and development of the crowdsourcing system informed by these understandings.

## 5 CROWDSOURCING SYSTEM FOR MOBILE WORKFORCE

A crowdsourcing system for a mobile workforce rather than a set of volunteers demands a careful system design due to a number of challenges imposed by the nature of the work this workforce performs. In this section, we first discuss the design decisions we have made to build our system informed by the mixed-method study presented earlier and then we present different components of the system and their operational principles.

### 5.1 Design Decisions

The solution we design for crowdsourcing tasks in an urban setting needs to be tailored to a mobile workforce such as that of bpost. Such a system should consider the behaviour of the workers as well as the types of jobs they can perform. Based on the behavioural study we reported in the last section, we made several design decisions to develop the system. We discuss these design aspects in the following.

**Task Dynamics:** In the previous section, we explained that a mobile workforce is able to collect information in any spatial scope. However, it is not possible to provide a finer temporal granularity than a *day*. While we can decide whether we carry out a task on a day, we cannot define a certain *time* at which a worker needs to answer a query about a particular location. Moreover, a mobile workforce is not able to collect responses to demand-driven and real-time tasks due to poor temporal coverage. In other cases where the workforce has a more flexible route, demand-driven tasks are possible, but a cause of stress [3]. Hence, we assign the crowdsourcing tasks to workers at the beginning of the day, and these tasks do not demand real-time response and can have a longer completion period. In addition, we allow tasks to be skipped at some days. A task requester can provide a deadline for a task that indicates the maximum number of days a task can be skipped.

**Spatial Dynamics:** While we show bpost workforce provides wide spatial coverage, over 50% in terms of the ratio of the streets visited by a worker, the coverage is not complete. In order to achieve full coverage, the worker rounds need to be *augmented* with crowdsourcing tasks. While increasing the number of tasks carried out by the workers is important, it is also essential to avoid long detours for the workers that stretch their rounds. As such, we have designed a cost metric that captures both the completion of a task and the detour it incurs to workers, and then use this metric to augment the spatial route of the workers optimally. We will discuss this component later in detail.

**Temporal Dynamics:** Since we determine the set of tasks to be carried out at different locations on a daily level, we have decided to assign tasks to workers proactively, before the beginning of their daily rounds. Each task has a designated area where the task needs to be performed. However, within that location, it is essential to identify the right and opportune moment when the postal worker can be interrupted to respond to the question or perform the task. We have reported earlier that postal workers have a situational disability due to the nature of their job. Besides, the sociality context as we discussed earlier is also essential to take into account. So, we have designed a purpose-built interruptibility component that identifies the right moment to push a crowdsourcing task to the mobile worker while s/he is within a spatial range of the task location.

**Device Dynamics:** In our study, we explored different device forms with the postal workers for seamless integration of crowdsourcing tasks into their daily routines. Although the in-house scanning device was the preferred device, it introduces additional situational disabilities. Besides, the scanning device was augmented with any sensors except IR and was not suitable for modelling richer context information. As such, we have chosen the second preferred device - a smartwatch for the following reasons:

(1) **Unobtrusive Context Sensing:** For the interruptibility management, we need to model a set of geo-tagged physical and conversational activities. Modern smartwatches offer excellent support for such context sensing due to rich onboard sensors, e.g., accelerometer, gyroscope, microphone, etc..
(2) **Semi-Disconnected Operations:** Since the tasks for bPost workers do not require real-time response, we have decided to develop a semi-disconnected solution. Since the devices need no connectivity during work routine, we do not equip the watches with SIM cards, and we do not pair with any master device such as a smartphone. Instead, they use their WiFi interface to download the tasks at the beginning of the routine and upload responses to those tasks at the end of the routine. This design decision, e.g., removing live networking also help us in maximizing the battery life of the smartwatch.
(3) **Discreet and Faster Interaction:** The tasks for the postal workers that we studied in this work are textual only[13]. We already mentioned situational disabilities. As such we wanted to have these tasks embedded in interactive notifications that are discreet and can be interacted with one hand with minimal engagement. A smartwatch notification is an excellent interactive modality for this.

In our behavioural study, we have noted that the workers often experience situational disabilities associated with carrying parcels, etc. In order to collect responses to tasks more effectively, we need to avoid such disabilities and detect opportune moments.

We opt to use smartwatches to best understand workers' context regarding physical activity. Since the devices need no connectivity, we do not equip the watches with SIM cards and we do not pair them with any master device such as smartphones. Instead, they use their Wi-Fi interface to download the tasks and upload responses to those tasks.

In the next section, we discuss the two main components of our crowdsourcing solution. In particular, we present a route augmentation and interruptibility management techniques to address the spatial and temporal aspects of our purpose-built solution for mobile postal workers.

## 5.2 Route Augmentation for Expanding Spatial Coverage

This component is responsible for augmenting the daily route of a postal worker with crowdsourcing tasks. It collects round information of all workers at the beginning of the day, and augments the worker rounds with crowdsourcing tasks. We use a heuristic that jointly minimizes the total detour and maximizes the number of completed tasks.

*5.2.1 Problem Definition.* We first formally define the notations for the task assignment problem, i.e., allocating two types of tasks - primary and crowdsourcing - to postal workers' daily routine.

Let $N_c$ be the set of crowdsourcing tasks that is available on a day that includes both that are introduced on the current day and those that were skipped in the previous day. Each task $t \in N_c$ has $L_t$ remaining days before it needs to be completed. $N_p$ denotes the set of primary tasks the workforce needs to complete. The set of all tasks $N$ is given by $N_c \cup N_p$. The set of workers is denoted by $M$.

The ordered list of tasks assigned to a worker $m$ defines the trajectory or the round of the worker $R_m = (t_0, t_1, ...t_n)$. The ordered list of tasks are defined by decision variables $x^m(t_i, t_j) \in \{0, 1\}$ where $t_i, t_j \in N$ and $m \in M$ that indicates whether worker $m$ handles task $t_j$ after the task $t_i$. A task $t_j$ is a part of $R_m$ if $x^m(t_i, t_j) = 1$ for any $i$. $R_m$ for each $m$ starts and ends with task 0 that is associated with leaving and arriving the mail center in the beginning and the end of the day, respectively.

Let $\tau^m(t_i, t_j)$ denote the cost for worker $m$ associated with completing tasks $j$ after task $i$. This may take several factors into account including the distance between tasks $i$ and $j$, the type of worker $m$, i.e. whether she is a

[13]We plan to study the multimodal data collection with our solution in our future avenue of our work.

driver or pedestrian, the time it takes to complete task $j$, etc. The cost associated with the entire $R_m$ becomes $T(R_m) = \sum_{i=1}^{n} \tau^m(t_{i-1}, t_i)$.

For convenience, we use $x^m_{i,j}$ instead of $x^m(t_i, t_j)$, $\tau^m_{i,j}$ instead of $\tau^m(t_i, t_j)$. Then, the round cost can also be formulated as

$$T(R_m) = \sum_{i,j \in N} x^m_{i,j} \tau^m_{i,j}. \tag{1}$$

We also consider that postponing a crowdsourcing task incurs a cost if $z_t = 0$ that means task $t$ is not completed by any worker.

We strive to assign tasks to workers in a way that minimizes the cost associated with their rounds and maximizes the the number of tasks that are carried out on a day. This multi-objective optimization problem is defined as follows:

$$\min \left( \sum_{m \in M} \alpha T(R_m) + \sum_{t \in N_c} \beta (1 - z_t) \right). \tag{2}$$

$\alpha$ and $\beta$ values weight the cost associated with the worker round and skipping a task, respectively.

The set of constraints associated with task assignment is following:

$$\sum_i x^m_{0i} = 1 \tag{3}$$

$$\sum_i x^m_{i0} = 1 \tag{4}$$

$$\sum_{i,m} x^m_{ij} = 1 \qquad j \in N_p \tag{5}$$

$$\sum_i x^m_{ij} = \sum_i x^m_{ji} \qquad j \in N \tag{6}$$

$$z_t \leq \sum_{i,m} x^m_{it} \qquad t \in N_c \tag{7}$$

$$(1 - z_t) \leq L_t \qquad t \in N_c \tag{8}$$

(3) and (4) indicates that worker rounds start and end at the central location. With (5), it is ensured that every primary task is handled by a worker. (6) establishes the flow conversation of every worker $m$. (7) states that $z_t$ can be equal to 1 only if $t$ is completed by a worker. Since $z_t$ minimizes the objective, it is equal to 1 whenever possible. (8) maintains that crowdsourcing tasks cannot be delayed beyond their deadlines.

Consider the case in which $L_t = 0$ for each $t$ and every task is assigned without delaying on a day. Then, (2) turns into Vehicular Routing Problem (VRT) [42]. Since VRT is NP-Hard, its generalization is also NP-Hard and the optimal solution cannot be computed effectively.

*5.2.2 Task Augmentation Heuristic.* To solve task assignment problem efficiently, we propose a greedy heuristic that relies on bpost delivery rounds. We assume that these rounds are (near-)optimal for delivering parcels. We build upon package delivery rounds and augment them with crowdsourcing tasks. This round augmentation heuristic is summarized in Algorithm 1.

We start with rounds provided by bpost in the beginning of the day and sort crowdsourcing tasks in ascending order of remaining days before a task must be completed, i.e. $L_t$ values. Then, we iterate through each worker and evaluate the cost of detour in case the task is incorporated into a worker's round. The cost of the round with

**ALGORITHM 1:** Heuristic for augmented round assignment problem

```
foreach m ∈ M do
    R_m ← N_p^m
end
SortedTasks ← Sort(N_c) ;
foreach t ∈ SortedTasks do
    cost ← a large number ;
    u ← ∅;
    foreach m ∈ M do
        cost_t^m ← T(R_m, t) − T(R_m);
        if cost_t^m < cost then
            cost ← cost_t^m;
            u ← m;
        end
    end
    if L_t == 0 then
        R_u ← Insert (R_u, t)
    else
        if αcost < β then
            R_u ← Insert (R_u, t)
        end
    end
end
```

task $t$ added at the best sequence is given by $T(R_m, t)$ where as $T(R_m)$ is calculated as in (1) with the current round of $m$, $R_m$ and the difference yields the added cost of the crowdsourcing task.

We find the worker $u$ that yields the detour with minimum cost, which incurs if the task is carried on the current day. After weighting this cost with $\alpha$ value in (2), we compare it with the weighted cost of postponing the task for a day, i.e. $\beta$. If the cost of the detour is smaller than the cost of postponing the task, we insert task $t$ to $R_u$. In addition, if $L_t = 0$, i.e. the task has no more remaining days, we insert $t$ to $R_u$.

In case a task is postponed, its associated $L_t$ value is decremented for the following day.

This heuristic greedily minimizes the cost of adding tasks to workers' rounds, starting from their delivery rounds. Assuming these starting points are optimal or near-optimal, it provides a greedy approximation to (2).

### 5.3 Interruptibility Management for Identifying Opportune Moments

The principal objective of this component is to identify opportune moments for an interruption for mobile workers. Once the moment is identified, the relevant task in that spatial region is presented to the postal worker. We achieve this by modelling spatiotemporal contextual situations of a postal worker. A worker's spatiotemporal activity trajectory is the primary input to our system. From this trajectory, we seek to identify the right moment in a target spatial area considering worker's physical and conversational activities to interrupt for a crowdsourcing task. To construct this trajectory, we leverage three sensing modalities - location, motion and audio. The overview of the interruptibility management component is shown in Figure 4.

**Modeling Location**: Location is known as the most powerful context for describing the human context. Each crowdsourcing task is associated with a location. To detect the user is near a task location, we track location by sampling GPS once every 30 seconds.

**Modeling Motion**: Motion sensing is constituted by onboard accelerometer and gyroscope to detect physical activity. We are interested in three-movement states: [*stationary*, *walking*, *on-transport*] and these are modeled by processing the raw accelerometer and gyroscope samples from the wearable. We use 5-second window frame with 95% overlap and then extract a set of time domain (mean, median, percentile, and RMS), and frequency domain (spectral energy, information entropy) features borrowing guideline from [11]. We pass these features to

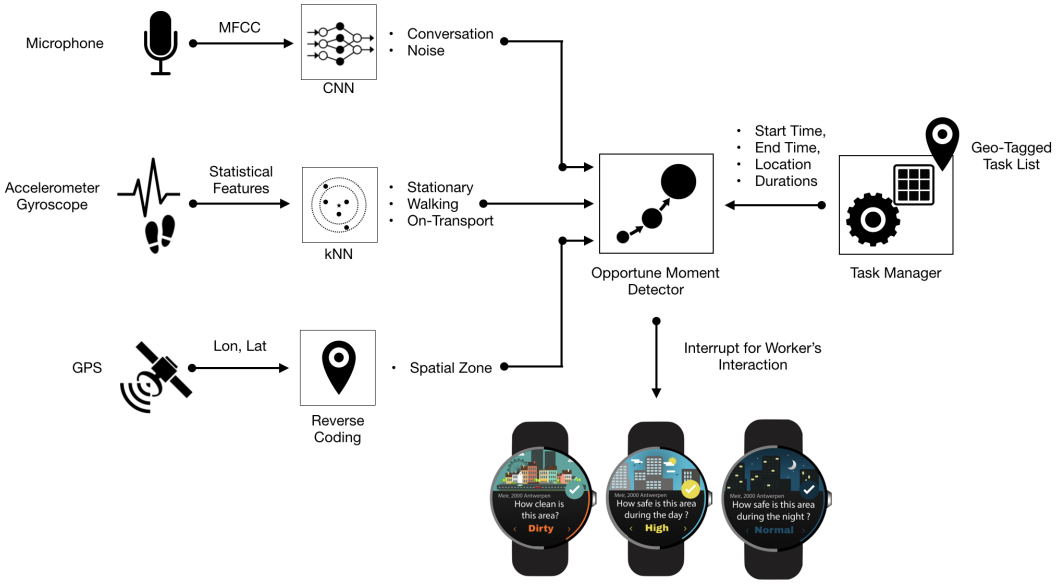



Fig. 4. The overview of the interruptibility management component

a k-nearest neighbours (K-NN)[14] classifier to extract physical activity label. We use 100Hz sampling rate for the sensors, however, produce activity labels at a granularity of 30 seconds.

**Modeling Audio**: Audio is a versatile sensing modality and has been used for various complex tasks in recent literature, e.g., speech recognition, keyword spotting, acoustic scene detection, emotion and stress detection etc. We are interested in a simple task recognition, i.e., if there is a conversation going on or not. Our system listens to the onboard microphone for 3 seconds in every 30 seconds and extracts 13 MFCC features from a 16-bit-PCM audio data following a sliding window approach (25 ms-long window and overlap of 10 ms). MFCC features are then passed to a couple of classifiers - each one composed of a CNN followed by a SoftMax layer - for detecting the presence of human conversation in audio signals.

**Opportune Moment Detector**: We need to identify opportune moments for two kinds of interactions in the form of notifications. The first notifies the worker about a crowdsourcing task that consists of a query and a list of possible answers. We also rely on notifications to steer workers towards crowdsourcing tasks. If the worker round is augmented by a crowdsourcing task that needs to be completed after a parcel is delivered, and the task location is far away from the delivery location, a notification informs the worker that she needs to go towards the task location. Our system treats both types of interruptions identically. We refer to a location where a worker needs to be presented with a notification as *interruptible zone*.

For identifying opportune moments in interruptible zones, we create situational contexts described with the combination of one of two conversational activity labels (*i.e., conversation, noise*), and one of the three physical activity labels (*i.e., stationary, walking, on-transport*).

The objective of the detector is to determine whether a specific time instance is interruptible or not. In particular, out of the six different possible combinations of physical and conversational activity context, the

[14]We have tried a variety of other shallow and deep classifiers (e.g., linear SVM, RBF SVM, decision tree, random forest, multi-layer perceptron, AdaBoost, etc.), and selected the one that yielded the best performance both concerning accuracy and resource footprint.

detector determines whether the current moment has following two context combinations *walking, noise* or *stationary, noise*. This basically identifies the moment when a worker is standing stationary or walking without engaging in a conversation. In both cases, the relevant spatial task is pushed to the worker.

Given the low-resource budget of a wearable device, this component works in a multi-phased way switching between two different operational modes (as illustrated in Figure 5).

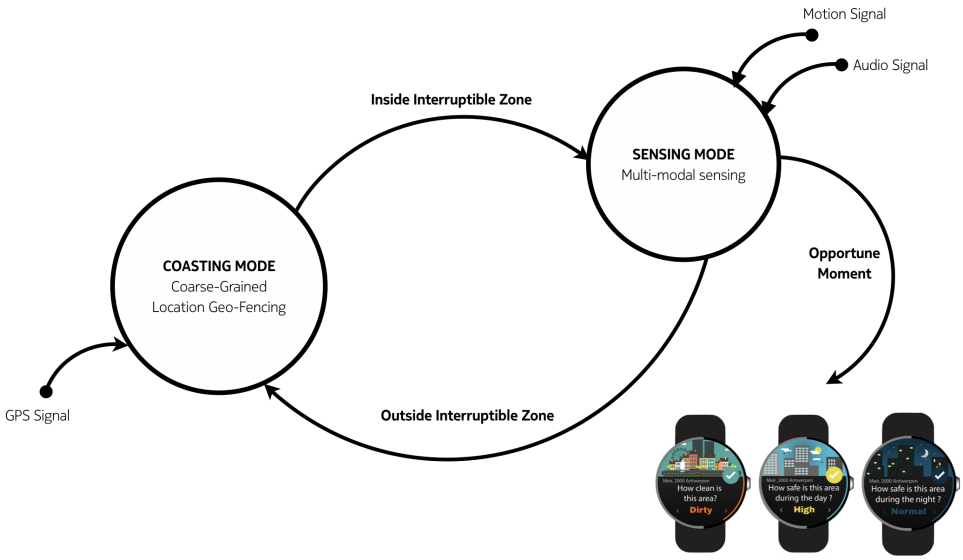



Fig. 5. Multi-phased opportune moment detector

- *Coasting Mode*: In this mode, only coarse-grained location is detected to conserve battery power whenever the worker is not close by any tasks that she needs to carry out.
- *Sensing Mode*: In this mode, a worker is detected to have reached to an interruptible zone - where a crowdsourced task is available. The device activates the sensors to record physical and conversational activities which are then matched against the rule set from the interruptibility detector, and when a moment is found, the notification is presented to the worker.

## 5.4 End-to-End System

In the last two sections, we described two critical components of our solution ensuring enhanced spatial coverage and increased engagement with just-in-time interaction. In this section, we present the end-to-end system solution for mobile crowdsourcing by postal workers. Essentially, our solution is composed of three major components: (i) a *Task Builder* interface in which one can create geo-spatial crowdsourcing tasks, and receive the responses in a variety of forms, (ii) the crowdsourcing *Task Manager* that generates augmented routes by assigning crowdsourcing tasks to worker rounds, pushes tasks to workers' devices, and retrieves task results from workers' devices and (iii)a wearable smartwatch *Task App* that models workers context to engage them with crowdsourcing tasks at the right time at the right place and gathers responses to send back to task manager. The overall architecture of our crowdsourcing solution for a postal worker is illustrated in Figure 6.

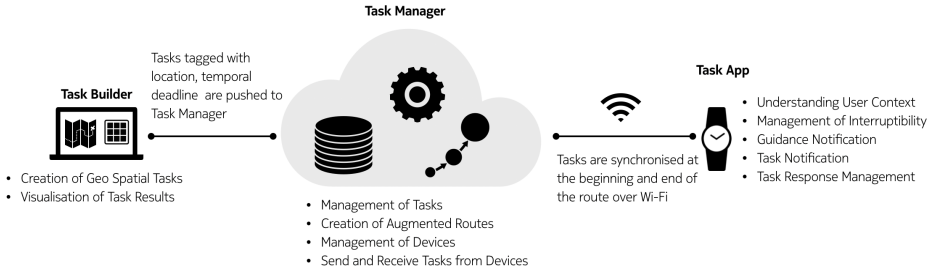



Fig. 6. The overall crowdsourcing system for postal workers

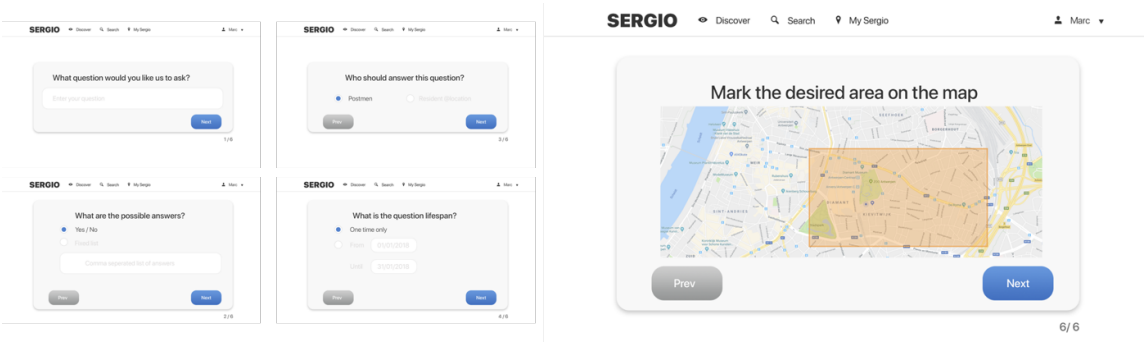



Fig. 7. Steps for creating a task in our Task Builder

*5.4.1 Task Builder.* We provide a web-based interface for task creation and analysis of the task responses. Every task is defined by a question, a number of possible answers to this question, the geographical scope of the tasks, i.e., whether it is a point-level or region-level task, temporal scope, i.e., whether it is one-time or repetitive, if the workers provide responses themselves or collect them from the public, including what segment of the population, the number of responses that need to be collected, and the end of the period over which the task is executed. Figure 7 shows representative screen-shots that demonstrate the several steps in task creation. To avoid overly inefficient detours, we allow tasks to be skipped. Hence, the interface asks the task creator the number of consecutive days that their tasks may be skipped.

In our system, each task is associated with a location, and we translate region-level tasks to one or more point-level queries. For example, if the task owner would like to learn about the friendliness of a neighbourhood, we can create a single task associated with a central location in the neighbourhood. On the other hand, to assess the cleanliness of a city, it creates a task for every street and the city. The feedback on this transformation is provided to the task owner through the Task Builder and the customer can increase/decrease the spatial granularity of the query. The entire Task Builder component is written in HTML5 and Java Scripts with NodeJs as the back-end.

*5.4.2 Task Manager.* The primary objective of this component is to create augmented routes for each postal worker based on the tasks and by implementing the algorithm presented in section 4.2. Once the augmented routes are created, they are pushed to postal workers smartwatches before workers start their rounds. As described earlier, these routes essentially contain geo-tagged crowdsourcing tasks embedded next to their primary delivery


Fig. 8. Task application for Android Wear

tasks. During their rounds, the Task App (see next) directs the workers towards the points of interest using this augmented routes through notifications. At the end of the day, when the workers return to the mail centre, their smartwatches sync again with Task manager to upload a file that contains the responses to the questions. The entire task manager component is written in NodeJs.

*5.4.3 Task App.* This component manifested as a wearable application models workers spatiotemporal context to identify opportune moments implementing the techniques presented in section 4.3. During these moments, notifications are presented to the worker that includes a small cue about the location of interest, the query for the worker to respond and then a list of possible answers. Example screenshots from the app are illustrated in Figure 8.

This component is implemented as a set of Android Services on top of Android Wear (now Wear OS) v1.5, targeting Android Platform 22, potentially working on almost every Wear OS device in circulation. In our case, we have used LG Urbane 2 watch, which features a SnapDragon 400 chip (Quad-core 1.2 GHz Cortex A7 and GPU Adreno 305) with 768 MB RAM and 4GB Flash storage and is equipped with 570 mAh battery. The services communicate with each other and with the Android Notification Manager via Android Intents and deliver the information contextually.

In the next section, we will discuss how this solution was used in a real trial with bPost postal workers together with a set of systematic evaluation.

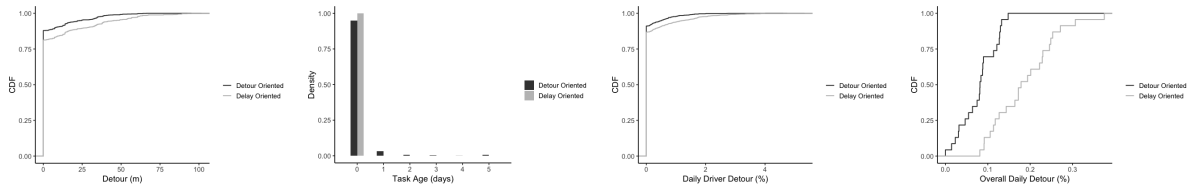
## 6 SYSTEM EVALUATION

In this section, we begin by offering assessments of the two principal components of our system, e.g., the route augmentation and interruptibility management techniques.

### 6.1 Performance of Route Augmentation Technique

Even though we present an end-to-end evaluation of our system later on in this section, we were not able to equip a large number of bpost drivers with smartwatches. Therefore, we evaluate our round augmentation heuristic with trace-driven simulations to assess its efficiency in large scale scenarios. To do this, we use the dataset that we have used for our quantitative behaviour analysis in Section 4.1. We use the delivery information provided by the dataset and augment them with crowdsourcing tasks.

We create crowdsourcing tasks at random addresses in Antwerp city center, an area that drivers from the corresponding mail centre roam around to deliver parcels. Each task needs to be completed within 1 work week,




(a) Detour Distribution (b) Task Age Distribution (c) Driver Detour (d) Overall Detour

Fig. 9. Performance characteristics of the round augmentation heuristic under two scenarios where the priority is on minimizing detour or maximizing the number of crowdsourcing tasks.

i.e. 5 days. Once a task is carried out, a new task is inserted to the system so that we make sure our system evaluates 100 tasks every day and decides whether or not they are incorporated into workers' rounds.

As $\tau$ values that yield the cost of a worker round in (1), we use the driving distance between two task locations since we do not have access to any other information such as task completion times. Since the dataset only provides information regarding the drivers, the worker type is not a discriminating factor in assigning costs between two task locations.

We evaluate our heuristic in terms of the detour incurred by the crowdsourcing tasks added to the worker rounds and the number of days between the day a task is introduced to the system and the day it is completed. If worker rounds are augmented with more tasks, they need to cover a larger distance increasing the cost of a detour while the tasks are completed sooner. This trade-off is governed by $\alpha$ and $\beta$ parameters.

With increasing $\alpha$ values, the heuristic prioritizes limiting the detour. On the other hand, as $\beta$ grows, it becomes essential to carry out the tasks without delay. Evaluating the round augmentation heuristic, we consider two scenarios. In *Detour Oriented* scenario, $\alpha >> \beta$ and we seek to minimize the detour at the cost of delaying tasks to consecutive days. In *Delay Oriented* scenario, on the other hand, the heuristic assigns a task on the current day even if it incurs high detour cost. These are the two extreme scenarios and the heuristic performance is bounded by these two conditions.

Figure 9 presents the performance of the heuristic in both scenarios. Figure 9a shows the distribution of detour each crowdsourcing tasks incurs while Figure 9b presents the distribution of task ages in terms of days when they are completed. A task age of 0 days indicates that the task is added to a worker round the day it is created and a task of age 5 means that the task is completed on its deadline, i.e. the latest possible day.

In Delay Oriented scenario, tasks are never postponed and are always completed on the day they are generated. While this scenario incurs more detour than Detour Oriented scenario, 81% of all tasks handled incur no detour at all. In Detour Oriented scenarios on the other hand, tasks are delayed if they incur detour or no more beneficial delivery is expected within the vicinity of the task. As a result, 87% of the tasks do not incur any cost and while tasks may be postponed, 94% of the tasks are completed when the tasks are created.

Figure 9c shows that drivers are like to have larger detours in case the crowdsourcing tasks are distributed in Delay Oriented scenario. Still, the daily detour of a driver never exceeds 4%. Figure 9d, on the other hand, shows the distribution of accumulated detour experienced by all the drivers. As before, the overall detour in Delay Oriented scenario is larger than that in Detour Oriented scenario. Still, considering all the drivers, the overall detour experienced in a day is below 0.3%.

In either case, since the workforce provides a wide spatiotemporal coverage, the round augmentation causes only a small detour in both scenarios. Even when the tasks may be delayed to the next day to ensure that the detour is even smaller, a vast majority of the tasks are assigned on the day they are created.

## 6.2 Performance and Resource Footprint of Interruptibility Management Technique

We want to reflect on the accuracy, and efficiency of the different inference components that are part of the interruptibility management process. The inference tasks, e.g., detecting physical and conversational activities as learning targets are well studied in the activity recognition literature. As mentioned in section 5.3 we borrowed bleeding-edge principles form the activity recognition community to address our relatively straight-forward learning objectives.

***Apparatus:*** For the benchmark, we have used LG Urbane 2 watch running Wear OS v1.5. LG Urbane 2 watch is built on a SnapDragon 400 chip (Quad-core 1.2 GHz Cortex A7 and GPU Adreno 305) with 768 MB RAM, 4GB Flash storage and 570 mAh battery. It contains multiple onboard sensors including accelerometer, gyroscope, GPS, and microphone.

***Data:*** We recruited ten participants and collected accelerometer, gyroscope, and microphone data for a set of physical and conversational activities. For physical activities, they were asked to perform each activity (walking, stationary, on-transport; car in our case) for a 20-minutes session, totalling 1-hour data for each user. For conversational activity, half of the time in each session, i.e., 10 minutes, the participant was involved in a one-to-one conversation with a researcher, and the rest of the time we ensured no conversation happened within 1m proximity of the watch (i.e., gathering non-conversational noise). Conversations outside this proximate range and other ambient activities were however allowed and were duly recorded by the microphone of the watch.

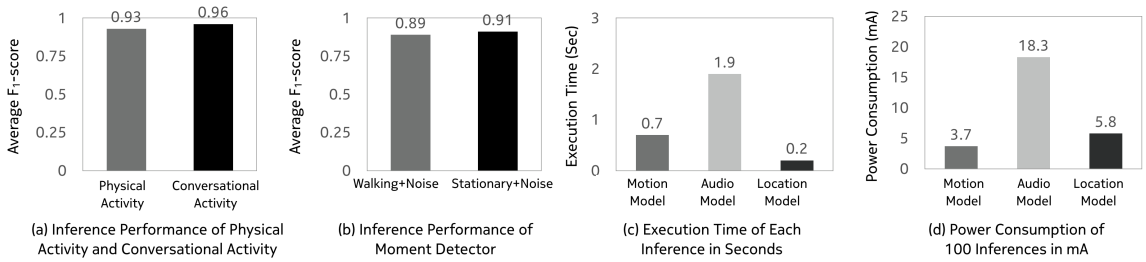



Fig. 10. Inference Accuracy, Execution Performance and Energy Footprint of different models used in the Interruptibility Management

***Results:*** We have assessed the performance of the models in an incremental way. For all the analysis, we conducted 10-fold cross-validation by default. We begin by reporting the performance of the k-nearest neighbours (K-NN) classifier, which was selected over a variety of other shallow and deep classifiers (e.g., linear SVM, RBF SVM, decision tree, random forest, multi-layer perceptron, and AdaBoost) based on the inference accuracy reaching up to an average $F_1$-score of 0.93 as depicted in Figure 10a. We then separately evaluated the performance of the CNN model for conversation detection and achieved an average $F_1$-score of 0.96 (Figure 10a). We then did another offline evaluation of the moment detector component which takes the output of the previous two classifiers and combines them to reach to the inference target. We have achieved $F_1$-scores of 0.89 and 0.91 for the two situations, walking with ambient noise, stationary with ambient noise as illustrated in the Figure 10b. This shows that moment detector demonstrates reasonably high accuracy in capturing moments when postal workers are standing stationary or walking without engaging in a conversation. We will also report in the next section, the performance of this component in the real-world deployment experiment. We then looked at the execution time and resource footprint of different context models. As illustrated in Figure 10c, all three models execute

efficiently with audio models running CNN yields the highest time for performing the inference. Corresponding energy consumption is relatively low for all three models as shown in Figure 10d. The system's overall battery consumption in the worst case is 36mA per hour under the assumptions that 120 inferences are performed continuously, e.g., in the sensing mode. This corresponds to 6.3% battery life for our host device - LG Urbane 2 watch.

## 7 IN-THE-WILD EVALUATION

In this section, first, we report a real-world trial of our purpose-built solution with bpost postal workers to assess different aspects of the system in increasing the spatiotemporal coverage, the accuracy of the information, and the engagement. We end the section by reporting insights from the semi-structured exit interviews with postal workers.

### 7.1 Real-World Deployment with bpost Postal Workers

We had a unique opportunity to work with the bpost workers to evaluate our end-to-end system in an in-the-wild study. In this section, we discuss different facets of this study and report the findings.

***Participants and Task:*** Ten bpost postal workers participated in this phase of our research for two weeks (10 working days). These workers were recruited by bpost management as we described in section 3 and were based in Antwerp. The kind of crowdsourcing tasks that were selected for this experiment is assessing the condition of the traffic signs at the city centre area of Antwerp with a spatial radius of 3.25km including 282 streets. The ten postal workers covered this area during the study. Assessing traffic signs was proposed by bpost as one of their potential domain of application. As illustrated in Figure 11c, the task of assessing the condition of a sign essentially means visiting the sign and recording its condition in of the three possible answers - i) Good Condition, ii) Bad Condition, and iii) Missing. Although simple, this task shares a few properties that we described in section 4. Namely this task is a *point-level*, *one-time* spatial task, does not demand *real-time* response and does not require social engagement, however requires perceptual assessment from the postal workers. We have retrieved the sign locations from the public dataset provided by City of Antwerp[15]. Of the more than 10000 signs in the dataset, we have selected 1000 signs that were part of the area of coverage of our ten participant workers.

***Apparatus and Experiment Settings:*** bpost gave the route information of the ten workers for two weeks. We compile this information together with our route augmentation algorithm to create augmented routes for each postal workers for the study duration. This augmented route includes tasks of assessing road sign condition using *Delay Oriented* approach, i.e. the tasks are always assigned to a driver. If a task is left unanswered on a day, it is not assigned again in order to prevent the accumulation of tasks as we progress in the study. Each worker was handed out the LG Urbane 2 watch pre-loaded with the **Task App**, and the corresponding augmented route at the beginning of the study and were asked to charge the smartwatch at the end of their route daily.

For each assigned task, we provided a notification to the postal worker at the location of the last delivery before the task. The notification indicated the address of the sign and the type of the sign as shown in Figure 11. If a worker engaged with the notification, it remained activated until she inspected the sign and provided an answer. This way, we combined the steering and asking the question in a single notification. To make sure the responses are *accurate and reliable*, we also recorded the coordinates of the location where the worker evaluated the sign.

***Experiment Design:*** We had three objectives in this phase - understanding the impact of our purpose-built solution with route augmentation and interruptibility management in increasing quantitatively i) spatial coverage,

[15] http://datasets.antwerpen.be/v4/gis/verkeersbordpt.json

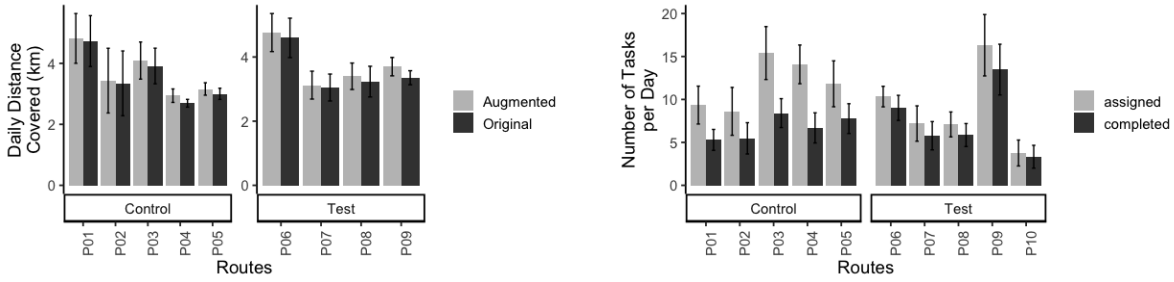

(a) (b) (c)

Fig. 11. Crowd-sourcing tasks in the real-world deployment study and the map showing partial coverage of the tasks. Each red and blue dot represents a task that was completed by the postal workers.

ii) accuracy, and iii) engagement. Route augmentation essentially contributes to the first aspect of spatial coverage. Interruptibility management, however, has contributions to accuracy and engagement aspects. Given the dynamics of these factors, we followed a between-groups study design in which five postal workers in a *Test Group* were exposed to interruptibility management, and the remaining five in a *Control Group* were not. However, both groups were exposed to route augmentation. Our premise was due to route augmentation we would notice an increase of spatial coverage for both groups. However, engagement and accuracy would be higher for the *Test Group* due to the purposeful interruptibility management.

***Results:*** We begin by reporting the coverage. The nature of the task did not allow us the assess the temporal coverage, as such, we look at the spatial coverage here. This aspect is independent of the between-groups study design. The original route distance across the ten workers was 337.1289 km (daily per worker $\mu = 33.71$ km and $\sigma = 4.41$ km) for covering the 137 streets in the target area. Please note that the target area has 282 streets. After applying our route augmentation with 1000 tasks this distance expanded into 353.0982km (daily per worker $\mu = 35.31$ km and $\sigma = 4.11$ km), an increase of 4.7%. This increase in distance essentially leads to a rise in spatial coverage of 40.5% (with 89% absolute coverage). This means, with the original route of all workers, 137 streets were covered out of 282 streets. With our augmentation, the route covered 251 streets. Figure 12a shows how individual worker's routes were affected by augmentation on a day.

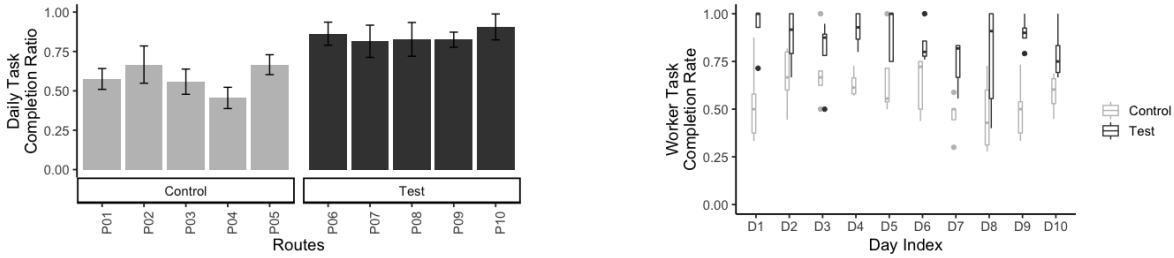
Next, we look at the task dynamics. Our route augmentation algorithm dynamically assigns tasks taking into different factors as explained earlier including their original routes for delivering parcels. As such, the number of tasks assigned to individual postal workers was different. Collectively, 571 tasks were assigned to *Control Group* (daily $\mu = 11$ and $\sigma = 5$) and 429 to *Test Group* (daily $\mu = 9$ and $\sigma = 5$). Please recall that the *Test Group* was exposed to interruptibility management where the *Control Group* was not. Figure 12b shows that workers in *Test Group* consistently provide responds to tasks with a higher percentage on any day. Overall, 323 tasks (56.6%) was completed by the *Control Group* (daily $\mu = 6$ and $\sigma = 3$) whereas 359 tasks (83.6%) were completed by the *Test Group* (daily $\mu = 8$ and $\sigma = 5$) - an increase of 27%. While workers in the *Control Group* are assigned more tasks, workers in *Test Group* complete more tasks than their counterparts. Hence, the increase in task completion ratio in *Test Group* cannot only be attributed a lower number assigned tasks. This is depicted in Figure 13a. We run a Welch's t-test and observe a significant effect between the groups ($p < 0.05$) concerning the factor that the *Test Group* received the notification triggered by interruptibility management. Given the small sample size, we can not




(a) Daily distance covered by each worker

(b) Daily assigned and completed tasks by each worker

Fig. 12. Difference between Control group and Test group in terms of covered distance and task completion




(a) Completion Rate across Worker

(b) Completion Rate across Time

Fig. 13. Response rate per worker and across the duration of the test.

claim that this significance is observed due to our solution. However, we consider this as a promising indicator suggesting the effect of just-in-time trigger for higher engagement of a mobile workforce in crowd-sourcing tasks.

To assess the quality of responses, we have manually inspected all the 682 responses from both groups. Note that, we gathered the coordinates of the location where the tasks were completed. We observed, surprisingly, that the 100% responses from both the groups were accurate. This suggests that the postal workers were sincere in their response and accurately completed their tasks. Again, given the small sample size and short duration, this results cannot be claimed to be conclusive, but certainly, carry a significant indication that mobile workforce can adequately address the lack of accuracy issues observed in conventional crowd-sourcing solutions.

Finally, Figure 13b shows how the task completion rate for workers in both groups changes across the duration of the pilot. Even though the response rate is low initially, for both groups, it increases as the study progressed. This is an interesting observation, which is further qualified in our exit interviews (reported next). But, this observation suggests as the postal workers became familiar with the solution and the nature of the tasks, these tasks became habitual for them, essentially turning into tiny habits [16]. While a more extended study is required to conclusively state that the system leads to sustained engagement over time, the trend in Figure 13b is encouraging for our system's ability to engage with workers.

[16] https://www.tinyhabits.com/

### 7.2 Semi-Structured Interview

We have followed up the two-week test with semi-structured interviews with the ten participants to assess several aspects of their experience concerning our solutions and their overall impression. All of these workers are part of bpost workforce for at least ten years and longer.

The interviews were conducted in Dutch (with audio recordings) by two researchers. One researcher was taking notes, while the other was engaged in the conversation. We structured the questions around four aspects: usability, motivation and incentives, task dynamics and engagement dynamics. We analysed the data by transcribing the audio recordings and by combining with interviewer notes. These were then coded and examined using affinity diagramming to derive conclusions around the three aspects mentioned above.

*On usability* of the overall solution including the wearable application was evaluated positively by ten workers. All of them mentioned that the experience of receiving just-in-time notification, clean and straightforward user interface, and one-touch interaction for task completion was positive, and made their engagement seamless. However, as we discuss later, they remarked on the necessity of a balanced volume and frequency of tasks so that these additional tasks do not keep them busy all the time.

*On motivation and incentives*, all ten postal workers commented positively concerning their participation mentioning they see this as an interesting opportunity for their organisation to generate new business. Essentially, we noticed the presence of an intrinsic motivation across all the workers. One specific comment from worker **P03** qualified this observation well.

> "I've been a mailman for 25 years. Last few years I've seen a decline in volume for letters. So I'm actually quite pleased to see bpost is looking for new business in order to keep my job relevant. I don't want to lose my job..."

However, we also observed that they were curious to understand the implication of such technology on their job-related benefits. In particular, given they were not included in the decision process that resulted in these added responsibilities, they were keen to assess the possible implication on their future compensation and benefits.

*On task dynamics*, workers stated that they were tempted to provide automatic responses to questions without inspecting the signs when they first received the notifications. Two workers indicated that they thought such an action might have been considered as cheating and they were afraid such an action cause an uneasy situation with their supervisors. Overall, they reported their experience with the wearable application was satisfactory and highlighted the efficacy of just-in-time notifications. One particular comment form **P02** was:

> "There is a big moment for these things for deliveries that need a signature. Especially in tall buildings, I have to wait before the owner of the flat comes downstairs to sign for the delivery. So I can answer a question easily..."

*On engagement*, multiple workers ($n = 6$) suggested that a light load of secondary tasks is essential since they want to finish their rounds as soon as possible. If completing a task requires a long detour, they are more likely not to worry about the quality of the answer. They also mentioned that if they receive a large number of notifications, they will ignore them unless the consequence impacts their job performance.

However, one interesting observation emerged through these interviews, if these tasks are offered at the right time (e.g., while they are waiting, or coming back to their car, etc.) over times these tasks turn into habitual routines. Several workers ($n = 7$) affirmed that over the lifetime of the pilot, these tasks already turned into a habit for them along their rounds. One remark from **P09** captured this articulately.

> "Once you get used to it, it's actually quite easy. I feel the watch vibrating, I look at the screen and answer the question. It's easy and doesn't consume a lot of time..."

To summarise, we have observed willingness across all the workers for participation, and they were sincere concerning the quality of the tasks. However, the volume and frequency of tasks are critical for their active engagement. In the next section, we further reflect on the implications of these and the rest of our findings.

## 8 DISCUSSIONS

In the previous section, we carefully assessed the efficacy of our crowd-sourcing solution for a mobile workforce uncovering a set of quantitative and qualitative observations that collectively, we consider, will inform the design of similar systems in related situations. In this section, we further offer a reflection on a variety of aspects concerning such solutions.

### 8.1 Participation Dynamics

We started this research with a premise that a mobile workforce due to the nature of their spatiotemporal activity routines offers plausible opportunities for crowd-sourcing tasks and in particular those that require human perception, and experiential knowledge. While through our multi-staged study we found this premise to be valid, we also learned that this validation comes at the expense of some conditions.

As described in section 3, the decision of engaging mobile workforce for such activities is driven by senior management of Belgian Postal Network as a pilot towards identifying a new business opportunity. We have learned that this decision was not an effect of bottom-up and inclusive discussions. This decision naturally has implications on how the actual workers embraced these additional responsibilities. As mentioned in section 3, the selection of postal workers for all 3 stages, i.e., the initial contextual study, deployment study and the final interviews were done by the bpost management based on the geographical region of interest and with willingness from the postal workers. This created some confusions among the workers concerning their changing job description, and more importantly, the key performance indicator (KPI) of their work. While we have observed positive intention from all of the workers at all stages to participate in the pilot as they recognize the organizational challenges and acknowledge the necessity of their participation, this also meant they need to adapt their daily routine and learn new skills. Through our analysis, we discovered that workers are sincere about these new responsibilities, and would like to participate in their best capacity. However, they lacked meaningful context, and purposeful explanation from the management, and structured training. These observations call attention to both organization management, and solution providers to have inclusive discussions, informed communications and purposeful training for mobile workers so that their participation becomes a strategic extension of their job responsibilities.

### 8.2 Engagement and Incentives

A related discussion point is engagement sustainability and the role of incentives. Existing literature has taught us that purposeful task assignment with some form of rewards and awareness of a variety of contextual attributes can increase users' motivation for more extended engagement [1, 18, 21, 33, 39, 45]. Although related, and in principle, we have applied many lessons reported in these works, our target population is different and as such the dynamics of engagement and incentives needs a slightly different perspective. Perhaps, the work of Bakewell et al. [3] deserves attention in this context, as their study demonstrated that workers feel anxious concerning their performance on these tasks that are beyond their primary activities. In our study, we did not observe any anxiety as such concerning quality. However, we found consistently that workers would like to complete these tasks fast, and preferred the total number of tasks to be reasonably small. In principle, these tasks were completed at the expense of their "free time". Besides, while not explicitly mentioned, we observed from their inquisitive discussions to understand the possibility of receiving benefits for such participation and maintaining them over time as these tasks were beyond their primary job descriptions. Our conversation with the management tells us

that they do not have any intention to introduce any incentives for these tasks as it would require a longer process of compensation review. At the same time, during this pilot, they did not want to modify their job description. There are two takeaways here. Firstly, it is clear that sustained quality participation can only be guaranteed if an incentive mechanism is introduced through a formal review of the compensation structure of the mobile workers, and secondly, if incentives are not offered, alternatively, the job description needs to be adjusted to accommodate these tasks. We hope our study provides data-driven insights and advocate these strategies to be integrated for scaling crowd-sourcing to a mobile workforce.

### 8.3 Trust Dynamics

We reported that postal workers due to the nature of their job develop social relationships with citizens. The significance of this relationship has several implications. On one hand, such a relationship can be leveraged to acquire local and subjective information with certainties, which may not be accessible with other means. For example, in our particular case, postal workers can act as a truthful qualifier of a neighbourhood's perception of safety (which was a big issue during the lifetime of this work due to the socio-political situation in Belgium concerning terrorism). Local citizens can openly converse and share their feelings and concerns with the postal officer because of their long-term relationship. On the other hand, the same relationship dynamics means that specific categories of questions, e.g., the presence of a solar panel (tax implications) or missing notice of household constructions (penalty implications) cannot be pushed to postal workers as it will damage their relationship. Besides, as we have observed, such questions with elements of social stigma act as a strong proponent of resentment from postal workers. Essentially, these observations call the attention of solution providers in 1) acknowledging this trust relationship and 2) designing questions carefully suitable for these workers, and 3) applying other means and multi-modal sensing modalities as necessary instead of mobile workers for a matter that may compromise their relationship with the citizens.

### 8.4 Privacy and Ethical Constraints

We want to draw attention to another sensitive aspect of a solution like ours concerning practical constraints. The study was conducted before the General Data Protection Regulation (GDPR) was applied across Europe. As such, we did not accommodate the principles recommended in this regulation. However, given the nature of data that such a solution can accumulate, it is essential to scrutinize the collection, storage and usage of information acquired using such a solution through a GDPR lens. We consider this as one of the avenues of our future work.

However, we would like to report on relevant issues that we have faced during the study. One of our sensing modality was audio - which was considered severely privacy invasive and not ethical both for the mobile workers and their customers. While we were allowed to run the study given the pilot nature of it and the fact that we only sampled audio signals while a user is in an interruptible zone, we have identified that continuous audio sensing is not a plausible design choice. This issue demands modification of our current interruptibility management component, and in principle, we need to consider the replacement of conversational context or introduce a careful sampling strategy for audio signals. We strive to address this issue in our future work.

### 8.5 Limitations

This study was conducted in Belgium. Certainly, the results presented here must be interpreted in the context of the culture and infrastructure in which they were collected. We expect our results are most appropriate for designers of crowd-sourcing technology for mobile workers in Europe or countries with similar cultures and levels of technology adoption.

Next, our engagement with actual postal workers was limited. Six postal workers were part of our contextual study, and ten postal workers took part in our final deployment trial for two weeks and in the followup subjective

interviews as we were not able to equip a large number of bpost drivers with smartwatches. We sincerely acknowledge that the size of our user group and the scale of our data is limited. Hence, the results reported here should not be considered as general, instead interpreted in the context of the study setup. Running in-the-wild study with postal workers during their real activities was significantly challenging, our methods as such were re-purposed to fit the logistic constraints, e.g., applying simulations to assess route augmentation mechanism or running system assessment in an extremely controlled setting with very simple signage evaluation task. As such, further validation studies with real data in a variety of situations are necessary to assess and widely apply the implications of our route augmentation and interruptibility management mechanisms. We sincerely acknowledge these limitations and however, we hope that our faithful observations and data-driven insights uncovered interesting implications as reported here, and inform the design and development of future crowd-sourcing solutions for mobile workforces.

## 9 CONCLUSION

Traditional approaches to mobile crowdsourcing provide incentives to regular citizens to encourage them to collect data, often incorporating various intelligent strategies to motivate, and sustain their participation. Although interesting, conventional solutions with such pool of citizens suffer from three shortcomings - lack of spatiotemporal coverage, accuracy, and sustained engagement.

In contrast, in this paper, we propose to embed crowdsourcing tasks to the daily routine of a mobile workforce such as postal workers who roam around urban and rural areas across an entire country to deliver letters, packages and parcels.

We first study the mobility of such a mobile workforce, postal workers from Belgium Postal Services bpost, both qualitatively and quantitatively to understand the feasibility of using them for crowd-sourcing tasks. Using a dataset provided by bpost that includes geo-tagged delivery information of packages, we show that postal workers have wide coverage in an urban area. In a qualitative behavioural study, we show that the primary of tasks of postal workers, i.e. delivering letters and parcels, do not prevent them from engaging with mobile devices, however, introduce situational disadvantages that might delay such engagements.

Based on our observations, we developed a first-of-its-kind wearable crowd-sourcing system built on two fundamental techniques - route augmentation, and on-wearable interruptibility management. These mechanisms, collectively, improve the spatial coverage and increase the accurate response rate of crowdsourcing tasks, assessed through a real-world deployment study with ten postal workers for two weeks. We also offer a reflection drawing upon a set of practical aspects critical to the success of such crowd-sourcing solutions for a mobile workforce. We expect, our findings highlight the way of building an efficient and purposeful crowdsourcing solution with an aim to transform the role of a mobile workforce into an intelligent human-sensor network to make significant leaps in our understanding of cities and their citizens.

## ACKNOWLEDGMENTS

This work is part of the SeRGIo project. SeRGIo is an icon project realized in collaboration with imec and with project support from VLAIO (Flanders Innovation & Entrepreneurship).

## REFERENCES

[1] Florian Alt, Alireza Sahami Shirazi, Albrecht Schmidt, Urs Kramer, and Zahid Nawaz. 2010. Location-based Crowdsourcing: Extending Crowdsourcing to the Real World. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries (NordiCHI '10)*. ACM, New York, NY, USA, 13–22. https://doi.org/10.1145/1868914.1868921

[2] E. Aubry, T. Silverston, A. Lahmadi, and O. Festor. 2014. CrowdOut: a Mobile Crowdsourcing Service for Road Safety in Digital Cities. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*. IEEE, 86–91. https://doi.org/10.1109/PerComW.2014.6815170

[3] Lyndsey L. Bakewell, Konstantina Vasileiou, Kiel S. Long, Mark Atkinson, Helen Rice, Manuela Barreto, Julie Barnett, Michael Wilson, Shaun Lawson, and John Vines. 2018. Everything We Do, Everything We Press: Data-Driven Remote Performance Management in a Mobile Workplace. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 371:1–371:14. https://doi.org/10.1145/3173574.3173945

[4] G. Cardone, L. Foschini, P. Bellavista, A. Corradi, C. Borcea, M. Talasila, and R. Curtmola. 2013. Fostering ParticipAction in Smart Cities: A Geo-Social Crowdsensing Platform. *IEEE Communications Magazine* 51, 6 (June 2013), 112–119. https://doi.org/10.1109/MCOM.2013.6525603

[5] G. Chatzimilioudis, A. Konstantinidis, C. Laoudias, and D. Zeinalipour-Yazti. 2012. Crowdsourcing with Smartphones. *IEEE Internet Computing* 16, 5 (Sept 2012), 36–44. https://doi.org/10.1109/MIC.2012.70

[6] Cen Chen, Shih-Fen Cheng, Aldy Gunawan, Archan Misra, Koustuv Dasgupta, and Deepthi Chander. 2014. TRACCS: A Framework for Trajectory-Aware Coordinated Urban Crowd-Sourcing. In *Second AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2014)*. AAAI Press, Palo Alto, CA, USA, 30–40.

[7] Cen Chen, Shih-Fen Cheng, Hoong Chuin Lau, and Archan Misra. 2015. Towards City-scale Mobile Crowdsourcing: Task Recommendations Under Trajectory Uncertainties. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI Press, Palo Alto, CA, USA, 1113–1119. http://dl.acm.org/citation.cfm?id=2832249.2832403

[8] Shih-Fen Cheng, Cen Chen, Thivya Kandappu, Hoong Chuin Lau, Archan Misra, Nikita Jaiman, Randy Tandriansyah, and Desmond Koh. 2017. Scalable Urban Mobile Crowdsourcing: Handling Uncertainty in Worker Movement. *ACM Trans. Intell. Syst. Technol.* 9, 3 (Dec. 2017), 26:1–26:24. https://doi.org/10.1145/3078842

[9] E. Cutrell, M. Czerwinski, and E. Horvitz. 2001. Notification, Disruption and Memory: Effects of Messaging Interruptions on Memory and Performance. In *Proceedings of INTERACT 2001*. IOS, Amsterdam, The Netherlands, 263–269. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.26.418

[10] Jakob Eriksson, Lewis Girod, Bret Hull, Ryan Newton, Samuel Madden, and Hari Balakrishnan. 2008. The Pothole Patrol: Using a Mobile Sensor Network for Road Surface Monitoring. In *Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services (MobiSys '08)*. ACM, New York, NY, USA, 29–39. https://doi.org/10.1145/1378600.1378605

[11] Davide Figo, Pedro C. Diniz, Diogo R. Ferreira, and João M. Cardoso. 2010. Preprocessing Techniques for Context Recognition from Accelerometer Data. *Personal Ubiquitous Comput.* 14, 7 (Oct. 2010), 645–662. https://doi.org/10.1007/s00779-010-0293-9

[12] Joel E. Fischer, Chris Greenhalgh, and Steve Benford. 2011. Investigating Episodes of Mobile Phone Activity As Indicators of Opportune Moments to Deliver Notifications. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11)*. ACM, New York, NY, USA, 181–190. https://doi.org/10.1145/2037373.2037402

[13] Michael J. Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. 2011. CrowdDB: Answering Queries with Crowdsourcing. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data (SIGMOD '11)*. ACM, New York, NY, USA, 61–72. https://doi.org/10.1145/1989323.1989331

[14] Joyce Ho and Stephen S. Intille. 2005. Using Context-aware Computing to Reduce the Perceived Burden of Interruptions from Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. ACM, New York, NY, USA, 909–918. https://doi.org/10.1145/1054972.1055100

[15] B. Hoh, T. Yan, D. Ganesan, K. Tracton, T. Iwuchukwu, and J. S. Lee. 2012. TruCentive: A Game-theoretic Incentive Platform for Trustworthy Mobile Crowdsourcing Parking Services. In *2012 15th International IEEE Conference on Intelligent Transportation Systems*. IEEE, 160–166. https://doi.org/10.1109/ITSC.2012.6338894

[16] Jeff Howe. 2008. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business* (1 ed.). Crown Publishing Group, New York, NY, USA.

[17] Desislava Hristova, Afra Mashhadi, Giovanni Quattrone, and Licia Capra. 2012. Mapping Community Engagement with Urban Crowd-Sourcing. In *Proc. When the City Meets the Citizen Workshop (WCMCW)*. AAAI, Palo Alto, CA, USA, 14–19. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4749/5102

[18] Kuan Lun Huang, Salil S. Kanhere, and Wen Hu. 2010. Are You Contributing Trustworthy Data?: The Case for a Reputation System in Participatory Sensing. In *Proceedings of the 13th ACM International Conference on Modeling, Analysis, and Simulation of Wireless and Mobile Systems (MSWIM '10)*. ACM, New York, NY, USA, 14–22. https://doi.org/10.1145/1868521.1868526

[19] Esther Jang, Mary Claire Barela, Matt Johnson, Philip Martinez, Cedric Festin, Margaret Lynn, Josephine Dionisio, and Kurtis Heimerl. 2018. Crowdsourcing Rural Network Maintenance and Repair via Network Messaging. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 67:1–67:12. https://doi.org/10.1145/3173574.3173641

[20] Hassaan Janjua, Wouter Joosen, Sam Michiels, and Danny Hughes. 2018. Trusted Operations On Mobile Phones. In *14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous 2017)*. ACM, New York, NY, USA, 452–459. https://doi.org/10.4108/eai.7-11-2017.2274952

[21] Thivya Kandappu, Nikita Jaiman, Randy Tandriansyah, Archan Misra, Shih-Fen Cheng, Cen Chen, Hoong Chuin Lau, Deepthi Chander, and Koustuv Dasgupta. 2016. TASKer: Behavioral Insights via Campus-based Experimental Mobile Crowd-sourcing. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 392–402.

https://doi.org/10.1145/2971648.2971690

[22] Thivya Kandappu, Archan Misra, and Randy Tandriansyah. 2017. Collaboration Trumps Homophily in Urban Mobile Crowdsourcing. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 902–915. https://doi.org/10.1145/2998181.2998311

[23] Salil S. Kanhere. 2013. Participatory Sensing: Crowdsourcing Data from Mobile Smartphones in Urban Spaces. In *Distributed Computing and Internet Technology*, Chittaranjan Hota and Pradip K. Srimani (Eds.). Springer, Berlin, Heidelberg, 19–26.

[24] Leyla Kazemi and Cyrus Shahabi. 2012. GeoCrowd: Enabling Query Answering with Spatial Crowdsourcing. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '12)*. ACM, New York, NY, USA, 189–198. https://doi.org/10.1145/2424321.2424346

[25] Leyla Kazemi, Cyrus Shahabi, and Lei Chen. 2013. GeoTruCrowd: Trustworthy Query Answering with Spatial Crowdsourcing. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL'13)*. ACM, New York, NY, USA, 314–323. https://doi.org/10.1145/2525314.2525346

[26] Emmanouil Koukoumidis, Li-Shiuan Peh, and Margaret Rose Martonosi. 2011. SignalGuru: Leveraging Mobile Phones for Collaborative Traffic Signal Schedule Advisory. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services (MobiSys '11)*. ACM, New York, NY, USA, 127–140. https://doi.org/10.1145/1999995.2000008

[27] Thomas Ludwig, Christoph Kotthaus, and Volkmar Pipek. 2016. Situated and Ubiquitous Crowdsourcing with Volunteers During Disasters. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 1441–1447. https://doi.org/10.1145/2968219.2968585

[28] Nicolas Maisonneuve, Matthias Stevens, Maria E. Niessen, and Luc Steels. 2009. NoiseTube: Measuring and mapping noise pollution with mobile phones. In *Information Technologies in Environmental Engineering*, Ioannis N. Athanasiadis, Andrea E. Rizzoli, Pericles A. Mitkas, and Jorge Marx Gómez (Eds.). Springer, Berlin, Heidelberg, 215–228.

[29] B. Martin, B. Hanington, and B.M. Hanington. 2012. *Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions*. Rockport Publishers, Beverly, MA, USA.

[30] Akhil Mathur, Nicholas D. Lane, and Fahim Kawsar. 2016. Engagement-aware Computing: Modelling User Engagement from Mobile Contexts. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 622–633. https://doi.org/10.1145/2971648.2971760

[31] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016. PrefMiner: Mining User's Preferences for Intelligent Mobile Notification Management. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 1223–1234. https://doi.org/10.1145/2971648.2971747

[32] Prashanth Mohan, Venkata N. Padmanabhan, and Ramachandran Ramjee. 2008. Nericell: Rich Monitoring of Road and Traffic Conditions Using Mobile Smartphones. In *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems (SenSys '08)*. ACM, New York, NY, USA, 323–336. https://doi.org/10.1145/1460412.1460444

[33] Mohamed Musthag and Deepak Ganesan. 2013. Labor Dynamics in a Mobile Micro-task Market. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 641–650. https://doi.org/10.1145/2470654.2470745

[34] Maria V. Palacin-Silva, Antti Knutas, Maria Angela Ferrario, Jari Porras, Jouni Ikonen, and Chandara Chea. 2018. The Role of Gamification in Participatory Environmental Sensing: A Study In the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 221:1–221:13. https://doi.org/10.1145/3173574.3173795

[35] Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: Designing Intelligent Prompting Mechanisms for Pervasive Applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. ACM, New York, NY, USA, 897–908. https://doi.org/10.1145/2632048.2632062

[36] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. 2015. When Attention is Not Scarce - Detecting Boredom from Mobile Phone Usage. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 825–836. https://doi.org/10.1145/2750858.2804252

[37] Giovanni Quattrone, Licia Capra, and Pasquale De Meo. 2015. There's No Such Thing As the Perfect Map: Quantifying Bias in Spatial Crowd-sourcing Datasets. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 1021–1032. https://doi.org/10.1145/2675133.2675235

[38] Rajib Kumar Rana, Chun Tung Chou, Salil S. Kanhere, Nirupama Bulusu, and Wen Hu. 2010. Ear-phone: An End-to-end Participatory Urban Noise Mapping System. In *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN '10)*. ACM, New York, NY, USA, 105–116. https://doi.org/10.1145/1791212.1791226

[39] J. Ren, Y. Zhang, K. Zhang, and X. Shen. 2015. Exploiting Mobile Crowdsourcing for Pervasive Cloud Services: Challenges and Solutions. *IEEE Communications Magazine* 53, 3 (March 2015), 98–105. https://doi.org/10.1109/MCOM.2015.7060488

[40] Darshan Santani, Jidraph Njuguna, Tierra Bills, Aisha W. Bryant, Reginald Bryant, Jonathan Ledgard, and Daniel Gatica-Perez. 2015. CommuniSense: Crowdsourcing Road Hazards in Nairobi. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '15)*. ACM, New York, NY, USA, 445–456. https://doi.org/10.1145/2785830.2785837

[41] Jacob Thebault-Spieker, Loren G. Terveen, and Brent Hecht. 2015. Avoiding the South Side and the Suburbs: The Geography of Mobile Crowdsourcing Markets. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 265–275. https://doi.org/10.1145/2675133.2675278

[42] Paolo Toth and Daniele Vigo. 2002. *The Vehicle Routing Problem*. SIAM, Philadelphia, PA, USA. https://doi.org/10.1137/1.9780898718515

[43] Umair ul Hassan and Edward Curry. 2015. Flag-verify-fix: Adaptive Spatial Crowdsourcing Leveraging Location-based Social Networks. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '15)*. ACM, New York, NY, USA, 79:1–79:4. https://doi.org/10.1145/2820783.2820870

[44] Rajan Vaish, Keith Wyngarden, Jingshu Chen, Brandon Cheung, and Michael S. Bernstein. 2014. Twitch Crowdsourcing: Crowd Contributions in Short Bursts of Time. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3645–3654. https://doi.org/10.1145/2556288.2556996

[45] Yufeng Wang, Xueyu Jia, Qun Jin, and Jianhua Ma. 2016. QuaCentive: A Quality-aware Incentive Mechanism in Mobile Crowdsourced Sensing (MCS). *The Journal of Supercomputing* 72, 8 (01 Aug 2016), 2924–2941. https://doi.org/10.1007/s11227-015-1395-y

Received November 2018; revised February 2019; accepted April 2019