Towards User Behavior Forecasting in Mobile Crowdsensing Applications

Luca Bedogni
luca.bedogni@unimore.it
University of Modena and Reggio Emilia
Italy

Matteo Buferli
matteo.buferli@comuni-chiamo.com
Comuni-Chiamo
Italy

Davide Marchi
davide.marchi@comuni-chiamo.com
Comuni-Chiamo
Italy

ABSTRACT

Mobile crowdsensing has rapidly become an interesting and useful methodology to collect data in modern smart cities, thanks to the pervasiveness of users mobile devices. Although there are many different proposals, opportunistic and participatory mobile crowdsensing are the most popular ones. They share a common goal, but require a different effort from the user, which often results in increased costs for the service provider. In this work we forecast user participation in mobile crowdsensing by leveraging a large dataset obtained from a real world application, which is key to understand whether there are areas in a city which need additional data obtained through raised incentives for participants or by other means. We then build a custom regressor trained on the dataset we have, which spans across several years in different cities in Italy, to predict the amount of reports in a given area at a given time. This allows service providers to preventively issue participatory tasks for workers in areas which do not meet a minimum number of measurements. Our results indicate that our model is able to predict the number of reports in an area with an average mean error depending on the precision needed, in the order of 10% for areas with a low number of reports.

CCS CONCEPTS

• **Theory of computation** → *Theory of database privacy and security*; • **Security and privacy** → **Privacy-preserving protocols**; • **Information systems** → *Crowdsourcing*.

KEYWORDS

Crowdsensing, Human behavior, Performance evaluation

ACM Reference Format:
Luca Bedogni, Matteo Buferli, and Davide Marchi. 2023. Towards User Behavior Forecasting in Mobile Crowdsensing Applications. In *ACM International Conference on Information Technology for Social Good (GoodIT '23), September 06–08, 2023, Lisbon, Portugal.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3582515.3609528

GoodIT '23, September 06–08, 2023, Lisbon, Portugal

ACM ISBN 979-8-4007-0116-0/23/09.
https://doi.org/10.1145/3582515.3609528

1 INTRODUCTION

Mobile crowdsensing (MCS) is an emerging paradigm which is focused on collecting data obtained from mobile devices of end users, which in this scenario are referred to as workers. This data is then analyzed to recognize patterns and understand how the area of interest behaves, or if there is any pattern related to the studied phenomenon. Modern mobile devices such as smartphones and wearables now provide a variety of sensors which along with mobile cameras already available on the devices are able to sense and report a wide variety of data or issues such as the air quality, the temperature, photos about environmental problems in the cities and alike. The data collected from various devices is then aggregated, to provide insights about the underlying phenomena and generate information sparked from the variety of such data.

Mobile crowdsensing is rapidly gaining interest since it provides an infrastructure-less platform which do not require to build a static data collection infrastructure with sensors in the area of interest. This opens up promising use-cases for modern smart cities, specifically on fields such as environmental monitoring, smart mobility, reporting, and in general it provides a cost-effective, flexible and scalable solution to a variety of use-cases. Moreover, it can also scale up depending on the service provider needs, which can customize the budget allocated to recruit workers, or to switch on or off certain areas in a city. This allows for greater flexibility, reduced maintenance costs hence a cost-effective and performing solution for different target scenarios and tasks.

However, mobile crowdsensing also presents numerous challenges, which needs to be addressed prior to deploying it in real world scenarios. which range from the energy efficiency of the mobile application to security and privacy, data quality, user engagement and appropriate rewarding mechanism. These include (i) Privacy and Security, since users may be doubtful in sharing personal information or sensor data due to potential misuse of them or unauthorized access; (ii) Reward, since users which are contributing their device resources and data want to be compensated for that; (iii) Quality of the data, since data reported from users may be noisy, and additional filtering and cleaning techniques must be employed; (iv) Data Processing, since the amount of data collected may be huge, and requires appropriate and customized tools to extract information from it. Ultimately, there is a final challenge, which is (v) Spatial and Temporal Coverage, since ensuring adequate spatial and temporal coverage for the data collection process is key to the utility of the MCS system. If areas are undersampled, then no patterns and in general no information can be extracted or forecasted from them.

These challenges are yet to be solved, and constitute one of the major issues in the real deployment of MCS systems. In this paper we mainly focus on the last on, directing our efforts into the forecast of the number of measurements in different areas of a city.

There are two main approaches to mobile crowdsensing: opportunistic and participatory. Opportunistic mobile crowdsensing is a methodology to collect data from workers without their specific intervention, hence whilst they are performing other activities. In other words, it collects data on the background, and it is thus suitable only for data types which do not require explicit actions from the workers, such as environmental data, mobility data or inertial sensors readings. It is an approach which is often used in scenarios such as traffic monitoring or environmental monitoring, since the data needed to fulfill these tasks can be collected from inertial sensors which do not require workers to actively report them. Key challenges in opportunistic mobile crowdsensing are related to user recruitment, energy efficiency and data quality. On the other hand, participatory crowdsensing requires from the workers specific actions, often but not always in response to a task issued by the platform. It is particularly useful when the data collection process requires a specific action from the user, such as taking a photo, or whenever the platform needs data at a particular time or in a particular area of interest. Participatory crowdsensing allows the service manager to ask workers to provide data when or where is most needed, thus allowing for better management of the data collection process. The key challenges of participatory MCS are mainly related to the user participation and user engagement, while also preserving the privacy of workers.

In this paper we provide a preliminary study about the trade-off between different mobile crowdsensing methodologies, studied on real data. Specifically, we focus on understanding whether it is possible to predict reports from workers from a set of parameters, hence to forecast whether on a certain area and at a certain time the crowdsensing platform will be fed up with updated data. This allows the possibility to merge different mobile crowdsensing paradigms, obtaining data from workers without issuing specific tasks, while leveraging participatory crowdsensing whenever data about an area of interest is scarce, thus incentivizing workers to collect it.

We study this problem starting from a real dataset collected over several years with a crowdsensing application in Italy, and we analyze the dataset to find patterns and user behaviors in them. The data we have is about citizens reports about different issues in a city, hence collected through a specific action of the workers, but without the service provider asking for specific data. The dataset is then enriched with other variables about the environment, such as the temperature, clouds, rain and alike, which may have an impact about the number of reports in a certain area. We then fit a regressor model with this data, and we predict the number of measurements in different areas. Our results indicate that our regressor model achieve satisfactory results, with a Root Mean Square Error (RMSE) of around 10% depending on the precision requested. When targeting areas with a low number of measurements, the RMSE is much lower than 1 report.

The rest of this paper is organized as follows: Section 2 describes related works from literature; Section 4 presents our dataset and how we enriched it; Section 3 describes the motivation behind this study, and what is the final objective; Section 5 highlights key results from our analysis; Section 6 concludes this paper and outlines future works on this topic.

2 RELATED WORKS

In recent years there have been some substantial advancements in the field of MCS, including real world applications, rewarding mechanisms, security and privacy and alike.

Regarding rewarding mechanisms, there has been a lot of work in the recent years. Workers of a MCS campaign can be rewarded in different ways, ranging from monetary prizes to immaterial benefits such as the possibility to redeem services from the MCS campaign owner. Clearly the reward also depend on the effort required from the workers: for opportunistic platforms, the specific actions required from the workers are low or nonexistent, as the collection of data is done in the background, while for participatory frameworks the reward is generally higher, given that the platform explicitely ask users to do something to collect data. We can broadly group incentive mechanisms in two different areas: extrinsic rewards and intrinsic rewards. Extrinsic rewards are those related to any kind of incentive which can be redeemed, leveraged or used outside of the platform itself, such as money, services from the MCS campaign owner and alike. Intrinsic rewards are instead those which can be useful within the platform, or are implicit in the sense that they provide a return for the user in terms of personal satisfaction, without necessarily offer anything tangible. Certainly monetary rewards are attractive for workers, and can be an effective boost in user engagement and user participation. However, they may become expensive for prolonged uses of the platform, hence they may not be sustainable.

One of the key challenges for rewarding mechanisms in MCS is how to determine the appropriate reward for users participating in the campaign, as also shown in [17]. For instance for participatory crowdsensing we can observe the work presented in [14], where the authors propose an optimal dynamic programming solution to the problem of the reward, also comparing it against a greedy solution. A similar objective, though tackled through a blockchain system and leveraging game theory is instead presented in [4]. A specific reward is determined in [11], where the authors propose to reward users depending on the sensing cost, thus providing a higher reward for data which is more expensive to get. We also cite [6], since it provides a novel paradigm in which the rewards are given anonymously, and an external aggregation service is leveraged to maintain the anonymity while still providing the desired reward to the workers. We also mention [1], which proposes a novel methodology to dynamically reward users in MCS, by also considering the privacy of the workers.

Other forms of rewards, though less used, are presented in [15] where the authors leverage gamification methodologies, and the more data they share the higher the reward, or [13], where badges and leaderboards are used. A slightly different approach is instead presented in [16], where the amount of reward is proportional to the number of other peers recruited into the campaign, which also fosters user participation. Researchers have also explored the possibility to provide social incentives to workers in MCS, such as social recognition or publicly acknowledging their participation

in the crowdsensing platform. These benefits have been deemed effective, though more for younger workers compared to older ones.

Considering instead the privacy challenges and advances in MCS systems, there have been recent survey papers which explore and discuss this problem [5] [18], which underline how privacy is a key aspect to consider when deploying MCS campaigns. We also cite the work performed by [2], where the authors propose a lightweight mechanism to preserve privacy in MCS campaigns. More recently also different privacy by design frameworks have been proposed, such as the one presented in [9]. When leveraging privacy by design frameworks, the benefit lies in the fact that the system is already built to preserve the privacy of the workers, without the need to employ additional techniques while the platform is running. Finally, there have been also proposals which leverage k-anonymity [12], t-closeness [7], l-diversity [8] and differential privacy [3].

3 MOTIVATION AND OBJECTIVE

In this section we discuss the motivation at the core of our work, and what are the research question we pose to ourselves in this paper.

As we already stated before, MCS is mainly divided into opportunistic architectures or participatory ones. While the former do not require explicit actions from the end workers, the amount of measurements it reports is also difficult to predict [10]. Moreover, it is only suitable for data types which do not require users intervention. The latter instead directly asks to the workers to perform a specific task, hence enabling the service provider to have more control on the area of interests for their service. Moreover, since opportunistic architecture require less effort from the workers, they are typically cheaper for the service provider, compared to participatory activities in which generally there is higher reward granted after fulfilling the given task. However, there can also be mixed architectures, in which both the participatory and the opportunistic methodologies are merged together, and leveraged to obtain the best from both strengths. Thus in this architecture the tradeoff for the service provider is to minimize the amount of explicit tasks asked to the workers, hence reducing the amount of distributed rewards, while still achieving a minimum number of reports in the area of interest. This also goes in the direction of reducing the issue of recruiting large amount of workers, and limiting it only to those really needed.

In this context, it is key to predict the number of reports in a certain area at a certain time, so that in case they are below a desired threshold, it is possible to issue specific tasks for user to fill those gaps. This allows the service provider to collect data in an opportunistic way from the workers, which may be enough in areas well covered by them, and issue participatory tasks is a minimum amount is not achieved. Not all the issued tasks will eventually be completed by the workers, but in this work we are more focused on the need to issue those tasks rather than determining ways of improving the probability of fulfilled participatory tasks, which is left as a future work.

Being totally opportunistic for the service platform translates into less control in the data obtained, while at the same time reducing the overall amount of rewards, since workers which do not have to explicitly complete a task are more willing to contribute to the platform even without an extrinsic reward. On the other hand, a completely participatory platform will present raised costs, and it is not sustainable in the long run. In the latter case, it is also possible to assess whether infrastructure based solutions are more viable.

Therefore the objective of this work, and the main research question we aim to answer, is whether it is possible to forecast the user behavior of workers in MCS platforms, based only on previous history about an area and on environmental parameters. A followup question is also related to the possibility to learn key parameters from areas in which the platform is already deployed, and transfer those insights into newly deployed areas. This will make it possible to forecast the number of measurements in an area, and issue specific tasks for workers so that the service provider can meet their desired quality of service.

4 DATASET ANALYSIS

In this section we present the dataset we have used, the characteristics of it and how we have enriched it to perform our analysis. Comuni-chiamo[1] is an Italian company which provides services to municipalities, specifically in terms of a mobile application which allow citizens to report issues in the area of interest such as broken urban signals, road potholes and alike. Citizens leverage the application to provide a geolocalized report of their findings, helping the municipality to monitor the urban landscape hence to act timely on the issues. For the citizens, the reward is intrinsic in the platform, since what they can report matters to them and they can see the problem solved by the municipality, ideally faster than waiting for the municipality to see the problem, schedule the appropriate actions and perform it. The municipality also has a dashboard, through which it can see the status of the different issues reported by citizens, the area of them, and assign tasks to their employees to tackle the issue.

The data that we have used in this work spans from 2013 to 2020 in different Italian cities, with more than 250.000 reports. These include a variety of report categories, made by citizens in different moments of the day for different purposes, such as reports which address issues on the roads, or regarding the environment like fallen trees, or areas which require cleaning. Users of the application can move freely in their city, and whenever they want to report something, they can simply open the app, select the appropriate categories, insert an adequate description and send the report. For obvious reasons, the data is geolocalized, so that the municipality can act on the precise location of where the report has been made. The application also offers the possibility to interact with the municipality through a chat service, with which users can communicate and possibly provide more information to their report, or can be notified by the municipality whenever the issue they reported has been fixed.

We have excluded data from the first COVID-19 lockdown, which in Italy started in March 2020, to avoid different patterns in the dataset which may be due to the changed people routines. Our data then spans across more than 100 months, hence to the best of our knowledge this makes our study one of the largest performed with real data in this domain.

[1] https://comuni-chiamo.com/

In this work we focus on the urban landscape and on environmental parameters (described in Section 4.3 to predict the number of measurements in a certain area. Therefore we have excluded from our analysis all the chats which have been intertwined between the reporter and the municipality, which we leave as a future work. We have only kept the latitude and longitude, the time of report and the category of each of them.

4.1 Dataset Distribution

In this section we provide more insights on the data we have used. Specifically, we show the distribution of measurements versus different parameters, which highlight the dataset dynamics.

At first we can see a distribution of the number of reports per worker. As it can be seen from Figure 1a, a vast majority of workers reported less than 50 reports. Nevertheless, there are many which are also accountable for hundreds. It is worth to note that in this graph we have hidden workers with more than 1000 reports, since there are few (less than 10 in the whole dataset) and may have shared accounts, thus not reflecting a single behavior. This kind of distribution seems straightforward: there are many users, as also reported by the municipalities, which register to the application only to perform few report to which they particularly care, and then uninstall the application or simply do not use it anymore. On the other hand, there are active reporters which provide several reports per day, which may be people routinely moving across the city hence able to see several different issues.

We then analyzed how much workers spatially move from one measurement to another. This highlights movement patterns, and uncover the real area in which users may actually report their issues. To do so, we have computed, for each user, all the distances traveled from one report to another, and we plot the distribution of the total distance traveled for each user in Figure 1b. We note that this is a lower bound to the distance traveled effectively by each worker, since we do not consider the effective road traveled but only the air distance between them. We also point out that this does not account for the urban landscape. Nevertheless, it gives us an idea on how far each worker moves in a certain area. As it can be seen, workers travel in general few kilometers between one point and another. For sake of readability, we have also removed from the chart few workers which traveled hundreds of kilometers from one point to another, due to the fact that they may be an application error or which may have reported data in another city which runs the application, and in any case they are not representative of the distribution.

Then we also show the total area in which workers move in Figure 1c. It is immediately evident that in general workers tend to share their reports in a smaller area. We can also argue that this is where reports matter more to them, since where it may be where they live or work, although this is not the primary focus of this work. We will also show how this intuition is useful for the clustering analysis, which we perform and show in Section 4.2.

We finally correlate the area and the distance traveled in Figure 1d, where we plot the bivariate distribution of the distance traveled and the area in which the workers reported any report. We can see that although few of them traveled longer distances, a large majority is condensed in the bottom left of the chart, where smaller areas are. Again, this confirms that in general user report locations can be quite well predicted by the location of their previous measurements, or in general by closer areas in which they have already reported something in the past.

Figure 2 shows instead the hourly distribution of reports for different days of the week. An immediate aspect we can observe is the fact that during workdays the reports are more condensed, and peak in the morning when users go to work. During festive days instead the reports are generally more distributed throughout the day, and we have also observed that are generally fewer. This can happen for a number of reason: for instance, workers can travel to other cities during festive days, hence they cannot use the application to report if that it is not present in the city in which they travel to. Even if the application is available, they may be less interested to report an issue in an area in which they do not travel often, hence they are less involved. Moreover, people may also engage in other activities during festive days compared to workdays, in which they may see the same problem day over day, and eventually decide to report it to have ti fixed.

4.2 Clustering analysis

To further investigate the behavior we have discussed in Section 4.1, we now perform a clustering analysis on the data. Specifically, we clusterize the measurements for each user on a growing number of clusters, to confirm the fact that workers tend to report in a small number of areas rather than reporting in more diverse areas of the city.

This behavior is confirmed in Figure 3, where we show the result of our analysis. Specifically, we account for two key metrics: the Within Cluster Sum of Squares (WCSS) which measures the squared average distance from each point in a cluster to its cluster head, and the Between Clusters Sum of Squares (BCSS), which measures the squared average distance between all centroids. More formally the WCSS is defined as follows:

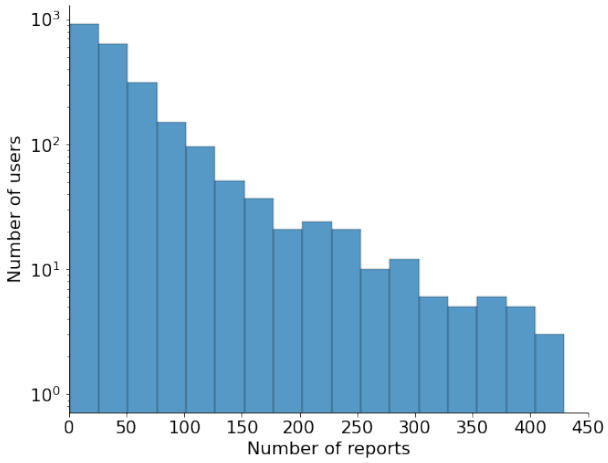
$$WCSS = \sum_{i=1}^{N_c} \sum_{x \in C_i} d(x, x_{C_i})^2, \tag{1}$$

where C_i is the i-th cluster, N_c is the number of clusters and X_{C_i} is the i-th cluster centroid. The BCSS is instead defined as follows:

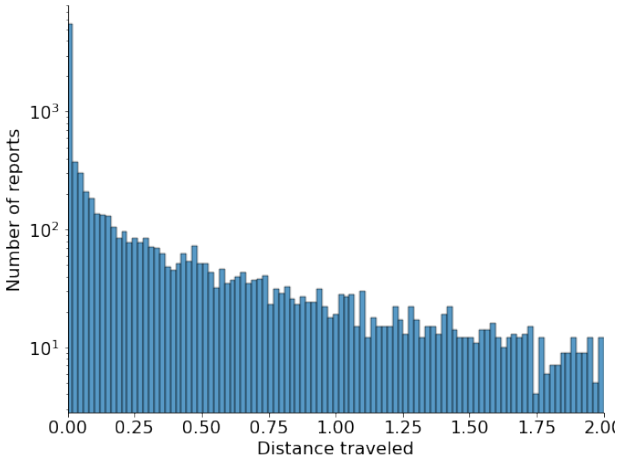
$$BCSS = \sum_{i=1}^{N_c} |C_i| \dot{d}(\mu, x_{C_i})^2, \tag{2}$$

where μ is the sample mean.

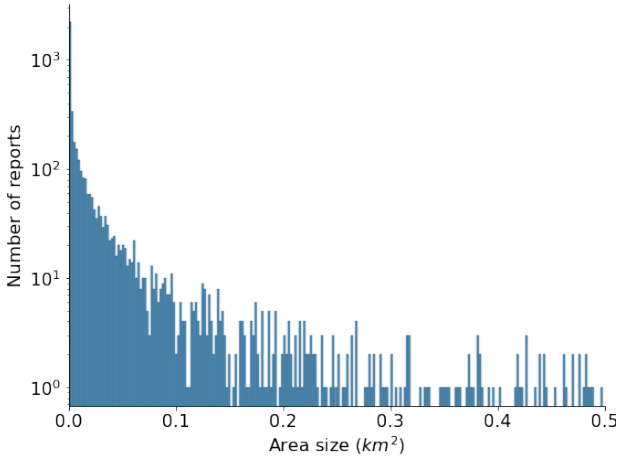
The black line refers to the WCSS, which measures how much a single cluster is spread, while the blue line refers to the BCSS, which measures how much single clusters are spread within an area. A low value on the WCSS means that the i-th cluster is dense, so that individual points are close to each other, while a higher value refers to more spread clusters. For the BCSS the meaning is more on the centroids, with higher values referring to farther centroids, while smaller values indicate closer centroids. We leverage the elbow method to determine that the best number of clusters is 2, since for the WCSS there is no significant advantage in adding more clusters, while for the BCSS the error increases considerably due to farther clusters. What this chart indicates is that in general workers tend to make reports in few zones, generally 2 to 3 key areas of their daily



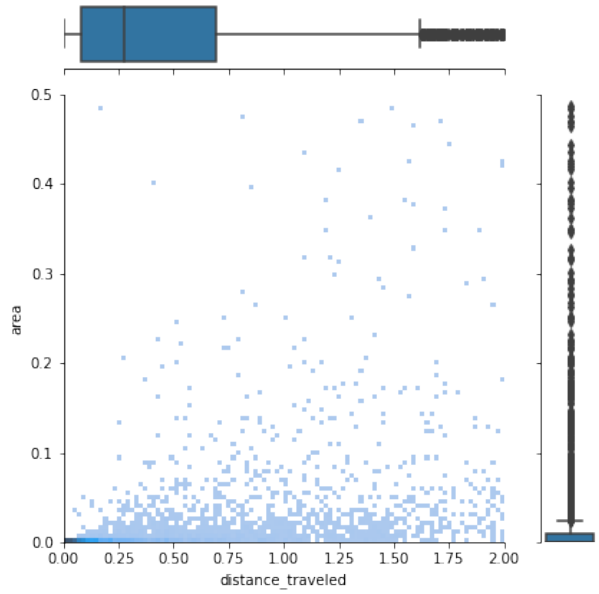
(a) Per-user reports distribution.



(b) Per-user distance distribution.



(c) Per-user area distribution.



(d) Area and distance distribution.

Figure 1: Various distribution on the number of reports. Figure 1a is the distribution of reports for each user, Figure 1b shows the distance between consecutive measurements, Figure 1c shows the overall area covered by each user, and Figure 1d shows the joint distribution between the distance and the area.

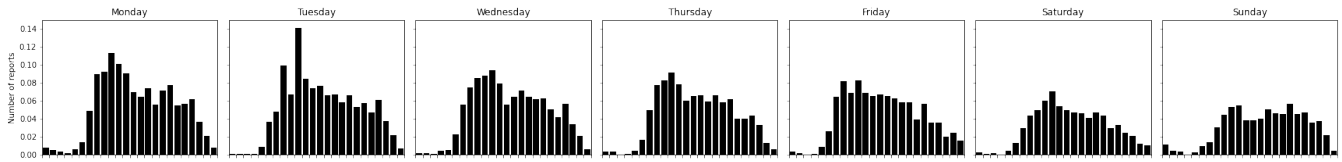


Figure 2: Hourly distribution for the 7 weekdays. It is evident how during festive days, and particularly on Sundays, workers tend to report less compared to the beginning of the week.

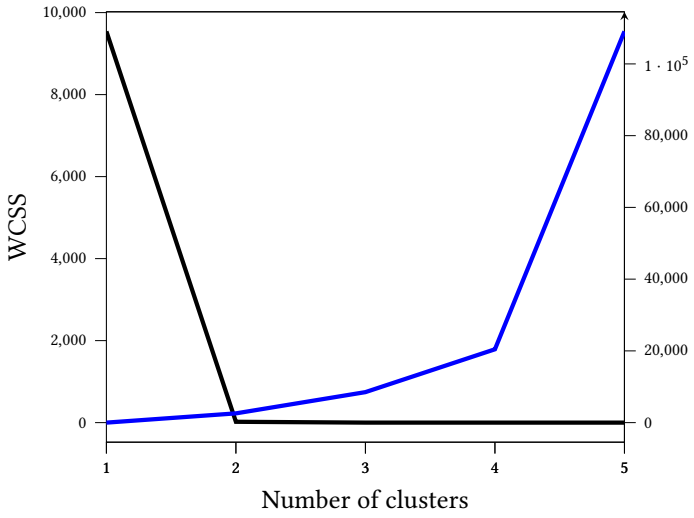


Figure 3: WCSS and BCSS clustering errors. The chart clearly show how the optimal number of clusters is 2, since higher number do not bring any significant advantage with respect to the two metrics considered.

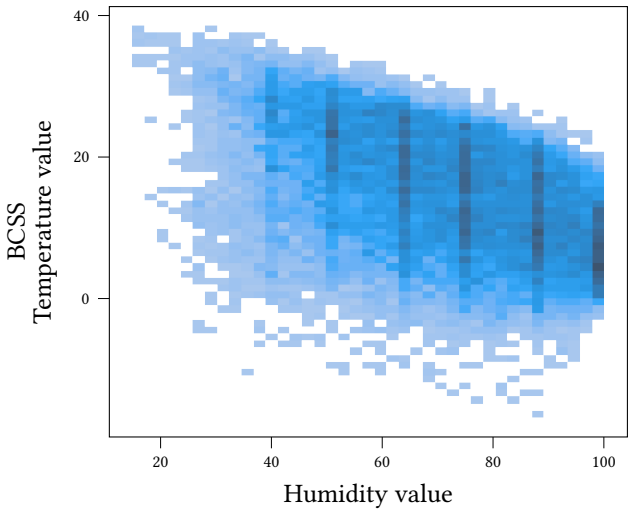


Figure 4: Density of the reports considering the temperature and the humidity in the area. It is evident that both variables play a role in determining the amount of reports, though the temperature is clearly more decisive.

routine, rather than sending reports from more diverse positions in the city. This may suggest that these 2 or 3 locations may be related to the residential area and the work area of the user, in addition to a possibly other interesting area, however more experiments are needed to confirm such hypothesis.

4.3 Environmental parameters

In this section we describe how we enrich our dataset to add environmental parameters, to perform a wider analysis also on variables such as the temperature and the humidity. We argue that these kind of variables may play a role in the user activity in the MCS application, determining whether user want to go outside, or how much they walk in a city during a rainy or sunny day. When the weather is fine, people spend more time outside, hence the opportunity to spot any issue and eventually report it would be higher compared to days spent indoors.

We leverage open APIs available at OpenMeteo[2] to collect historical data about environmental parameters. Given a latitude and a longitude, OpenMeteo provides data about precipitation, clouds, temperature, humidity and other similar weather-related parameters. We query it for each single location in our dataset, approximated to a maximum of 2 decimals, and we add the data to our dataset. We note that approximating the location to 2 decimals does not hinder a precise analysis, since such environmental parameters do not change for such shorter distances. From all the variables returned by OpenMeteo, we decide to keep the humidity, the temperature and the precipitation level, as we consider them the most representatives to determine between a day in which users tend to go outside or days in which they spend more time indoors.

Figure 4 shows the results of the analysis, in which we plot the number of reports accounting for the temperature and the humidity. As it can be clearly seen, there is a pattern in the number of reports, which is highlighted in the chart. We want to note that some combination of temperature and humidity values are more frequent than others, therefore they may provide a higher number of measurements per-se. Nevertheless, Figure 4 shows that there is a correlation between environmental parameters and the possibility for a user to report any data to the platform. As it may be expected, users tend to report more data with mild temperatures and moderate levels of humidity, and tend to avoid too hot or too humid days, in which they probably spend more time indoors hence they are not able to see issues around the city.

[2]https://open-meteo.com

5 NUMERICAL RESULTS

In this section we provide numerical results about the prediction of the number of measurements in a given area. We build upon the knowledge presented in Section 4 and we use a Huber Regressor to fit our initial variables. We note that all the variables that we use, such as the environmental parameters and the time, can be easily gathered from the municipality or in general from the service provider and thus do not constitute a problem in running our model. We remark our aim and main research question, which is to determine whether in some areas there may be a scarcity of measurements, which may be due not to the lack of issues in the city, but rather to the fact that fewer users travel through those zones, hence there are less opportunity for such issues to be reported.

We train two different regressor models, on two different set of data. One is trained on the whole dataset, hence it benefits from a wider set of data but at the same time it also deals with a huge heterogeneity in area types, number of users and diversity in general, and we name this model `COMPLETE`. We also perform a different analysis, in which we train a separate regressor for each single area. Clearly in this case we have less data for each area to be trained

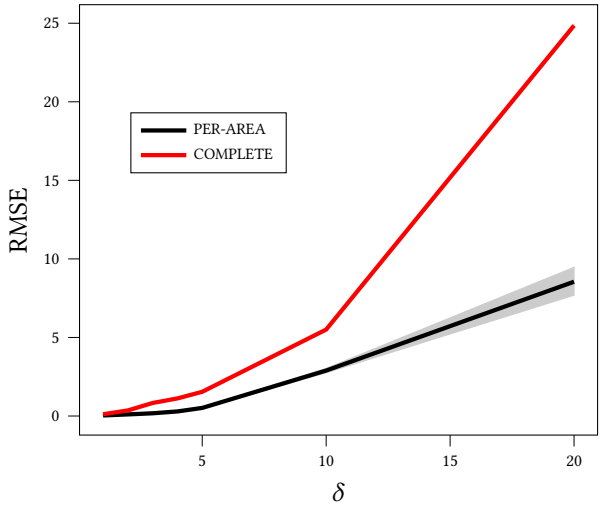


Figure 5: RMSE varying the cutoff. In red we plot the results training on the whole dataset, while in black we plot the results obtained training a separate model for each position.

Feature name	Loss
Day	-0.00850
Month	-0.02281
Year	0.00970
Hour	0.05380
Day type	0
Day of the Week	0.01808
Area ID	0.00078
Temperature	0.04444
Humidity	-0.01127
Latitude	-0.00033
Longitude	0.00064
Rain amount	0.00660
Snow amount	0.00000
Cloud Percentage	0.00988
Wind speed at 10m	-0.00013
Wind speed at 100m	0.00392

Table 1: Feature importance on the PER-AREA model.

on, as it can account only for the reports performed in such a small portion of the city, but it is clearly more tailored to the parameters pertaining to such zone. We name this model PER-AREA.

We present our results in Figure 5, where the red line refers to the COMPLETE model, while the black line is the PER-AREA. To perform a more comprehensive analysis, we also consider a varying cutoff parameter δ. For our purpose, which is to estimate whenever an area may have a low number of measurements, hence specific tasks may be issued, it is not key to estimate the precise number of measurements, but rather to understand if measurements will be less than a certain threshold below which the system needs further reports from the workers, and can issue participatory tasks, as an example. We aggregate and sum measurements which happened in the same area and in the same day, so δ considers all the values above it as $\delta+1$, hence grouping them together with the same value. More formally:

$$N_i(t) = \begin{cases} N_i(t) & \text{if } N_i(t) < \delta \\ \delta & \text{otherwise} \end{cases}, \tag{3}$$

where $N_i(t)$ is the number of measurements for the i-th area at time t.

What it is immediately evident is the fact that as δ increases, also the RMSE does so. This is expected, since higher values of δ requires a higher precision from the model itself. In the extreme case in which $\delta = 1$, the model simply has to forecast whether in such area the number of measurements will be either 0 or 1. For higher values, we are also considering all the intermediate cases, which is clearly a more complex tasks. Still, we can observe that the model predicts the number of measurements precisely up to $\delta = 5$, where both of them increase the RMSE significantly.

In table 1 we show the features for the PER-AREA model, which represent the results which we be obtained by optimizing the L2-regularized Huber loss in the regressor. Among all, it is pretty evident how the most influential features are those related to the hour and the temperature. The hour is straightforward, people tend to report data during daytime much more than during the night, as we have also shown in Figure 2, where it is evident that regardless of the day, people tend to report more from 8AM to around 10PM. The other major feature is the temperature, again if we look at Figure 4 it is clear how the temperature plays a major role in determining the number of reports in an area. Mild temperatures are more favorable to people which tend to stay outside more, hence they have more opportunities to find issues and report them, while too cold or too hot temperatures discourage people from doing so.

6 CONCLUSION

In this paper we have presented a novel study on a real dataset which aims to forecast the number of measurements in a given area in MCS, by leveraging the past history of such area and other parameters such as those related to the environment. This specific task is key to a number of different scenarios, since being able to forecast the amount of data reported in an area may help to better plan participatory tasks for the MCS campaign owner, and in general to have a more homogeneous and complete set of reports over the whole area of interest.

We have trained a custom regressor with such variables, and we have presented it with two initial set of training data, highlighting how a more precise model can be obtained by considering the past history of smaller areas compared to, for instance, the whole city.

Our study contributes to improving MCS deployments in practical applications. The development of an accurate prediction model helps MCS campaign owners make better decisions, allocate resources effectively, and provide comprehensive coverage of the target area. This can also help to better plan resources in cities, hence their development.

Future works on this topic include the sentiment analysis of the chat between the reporter and the municipality. Our hypothesis is that a faster and more appropriate response from the municipality sparks interest in the reporter to actually report more issues in the city, seeing that those are handled and solved efficiently. Moreover,

we are also planning to check whether there have been differences in the user mobility and behavior before, during and after the covid lockdown.

REFERENCES

[1] Luca Bedogni and Federico Montori. 2023. Joint privacy and data quality aware reward in opportunistic Mobile Crowdsensing systems. *Journal of Network and Computer Applications* (2023), 103634. https://doi.org/10.1016/j.jnca.2023.103634

[2] Yudan Cheng, Jianfeng Ma, and Zhiquan Liu. 2022. A Lightweight Privacy-Preserving Participant Selection Scheme for Mobile Crowdsensing. In *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. 1509–1514. https://doi.org/10.1109/WCNC51071.2022.9771871

[3] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.

[4] Jiejun Hu, Kun Yang, Kezhi Wang, and Kai Zhang. 2020. A Blockchain-Based Reward Mechanism for Mobile Crowdsensing. *IEEE Transactions on Computational Social Systems* 7, 1 (2020), 178–191. https://doi.org/10.1109/TCSS.2019.2956629

[5] Jong Wook Kim, Kennedy Edemacu, and Beakcheol Jang. 2022. Privacy-preserving mechanisms for location privacy in mobile crowdsensing: A survey. *Journal of Network and Computer Applications* 200 (2022), 103315. https://doi.org/10.1016/j.jnca.2021.103315

[6] Lorenz Cuno Klopfenstein, Saverio Delpriori, Alessandro Aldini, and Alessandro Bogliolo. 2019. "Worth one minute": An anonymous rewarding platform for crowd-sensing systems. *Journal of Communications and Networks* 21, 5 (2019), 509–520. https://doi.org/10.1109/JCN.2019.000051

[7] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *2007 IEEE 23rd International Conference on Data Engineering*. 106–115. https://doi.org/10.1109/ICDE.2007.367856

[8] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. L-Diversity: Privacy beyond k-Anonymity. *ACM Trans. Knowl. Discov. Data* 1, 1 (mar 2007), 3–es. https://doi.org/10.1145/1217299.1217302

[9] Federico Montori and Luca Bedogni. 2023. Privacy preservation for spatio-temporal data in Mobile Crowdsensing scenarios. *Pervasive and Mobile Computing* 90 (2023), 101755. https://doi.org/10.1016/j.pmcj.2023.101755

[10] F. Montori, L. Bedogni, and L. Bononi. 2017. Distributed data collection control in opportunistic mobile crowdsensing. In *SMARTOBJECTS 2017 - Proceedings of the 3rd Workshop on Experiences with the Design and Implementation of Smart Objects, co-located with MobiCom 2017*. https://doi.org/10.1145/3127502.3127509

[11] M. Pouryazdan, B. Kantarci, T. Soyata, and H. Song. 2016. Anchor-Assisted and Vote-Based Trustworthiness Assurance in Smart City Crowdsensing. *IEEE Access* 4 (2016), 529–541. https://doi.org/10.1109/ACCESS.2016.2519820

[12] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.

[13] Y. Ueyama, M. Tamai, Y. Arakawa, and K. Yasumoto. 2014. Gamification-based incentive mechanism for participatory sensing. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*. IEEE, 98–103. https://doi.org/10.1109/PerComW.2014.6815172

[14] Zhibo Wang, Jiahui Hu, Jing Zhao, Dejun Yang, Honglong Chen, and Qian Wang. 2018. Pay On-Demand: Dynamic Incentive and Task Selection for Location-Dependent Mobile Crowdsensing Systems. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*. 611–621. https://doi.org/10.1109/ICDCS.2018.00066

[15] F. Wu and T. Luo. 2014. WiFiScout: A Crowdsensing WiFi Advisory System with Gamification-Based Incentive. In *2014 IEEE 11th International Conference on Mobile Ad Hoc and Sensor Systems*. IEEE, 533–534.

[16] G. Yang, S. He, Z. Shi, and J. Chen. 2017. Promoting Cooperation by the Social Incentive Mechanism in Mobile Crowdsensing. *IEEE Communications Magazine* 55, 3 (March 2017), 86–92.

[17] Xinglin Zhang, Zheng Yang, Wei Sun, Yunhao Liu, Shaohua Tang, Kai Xing, and Xufei Mao. 2016. Incentives for mobile crowd sensing: A survey. *IEEE Communications Surveys & Tutorials* 18, 1 (2016), 54–67.

[18] Shiting Zhao, Guozi Qi, Tengjiao He, Jinpeng Chen, Zhiquan Liu, and Kaimin Wei. 2022. A Survey of Sparse Mobile Crowdsensing: Developments and Opportunities. *IEEE Open Journal of the Computer Society* 3 (2022), 73–85. https://doi.org/10.1109/OJCS.2022.3177290