MOBILE PHONE DATA TO SUPPORT AIR POLLUTION EXPOSURE ASSESSMENT

Erika Cerasti, Cristina Faricelli, Paolo Mattera, Roberta Radini, Tiziana Tuoto

Abstract. Air pollution is one of the greatest environmental risks to health according to the World Health Organization. Depending on the characteristics of the data sources we can organise the personal exposure assessment methods in two classification criteria: (i) spatial-temporal variations of individuals' activities (point-fixed or trajectory-based) and (ii) characterisation of air quality (variable or uniform).
In line with this approach, the paper presents a study based on Mobile Network Operator (MNO) data to evaluate spatial-temporal variations in human presence along with pollutant measurements to estimate people's exposure to air pollutants. MNO data enable a longitudinal analysis of human presence with high spatial and time resolution. This allows assigning different levels of air pollution exposure to the population at a specific time and in a specific location.
The proposed method can be useful for policymakers to assess the crowdedness of air-polluted areas over time and to find suitable solutions to mitigate the exposure. In addition, our results can be exploited for improved estimation of the risks inherent to the population exposure to air pollution in urban areas and for epidemiological studies.

1. Introduction

Outdoor air pollution is one of the greatest environmental risks to health according to the World Health Organization (WHO), estimated to have caused 4.2 million premature deaths worldwide in 2019[1]. Indeed, exposure to fine particulate matter and nitrogen compounds causes cardiovascular and respiratory diseases both chronic and acute, including asthma (Beelen *et al.* 2014).

Several pollutants are monitored in urban and industrial areas to ensure the air quality is in line with the guidelines provided by the WHO. In order to estimate population exposure levels to different pollutants, data on the air quality measured by the monitoring stations need to be combined with data on population presence in the areas. Usually, Census and administrative data are used for this scope, but of

[1] https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

course, it is a rough approximation: the presence assessment is static and derived only from the place of usual residence (Gray *et al*., 2013).

The spatial-temporal variations of individuals' activities would be preferred to estimate the exposure of the population to pollution in urban areas. However, this information is usually omitted in exposure studies due to the lack of reliable data sources. Mobile Network Operator (MNO) data give the opportunity to fill the gap by assessing the actual frequented places by the population (Dewulf *et al.*, 2016).

The aim of this work is to develop a method for estimating exposure to air pollutants by associating individuals with locations, and thus the corresponding pollution levels, using MNO and air quality data.

MNO data have received increasing attention in recent years to enrich and complement traditional data sources, surveys and administrative data, in Official Statistics. Nevertheless, many of the experiences in exploiting this new data source are often experimental and still require a well-defined standardized methodological framework, which encompasses the requirements for sound methodologies, transparent procedures, coherence and comparability, and in general the commitment to quality that marks the official statistical products. Recently, the Italian National Institute, Istat, has been involved in several initiatives at the national level and at the European level, co-funded by the European Commission. Indeed, the Multi-MNO[2] project (2023-2025) is a first-of-its-kind project in the European Statistical System: a team of industry specialists and NSI experts for co-developing together an open methodological standard. In addition, the research project MNO-MINDS[3] (Methods for Integrating New Data Sources, 2023-2025) focuses on methods for the integration of MNO and non-MNO data (e.g., survey data, census data, other big data sources etc.) with 10 NSIs involved across Europe. The work presented in this paper benefits from both the two projects and it is an example of the potential exploitation of MNO data for innovative official statistics.

2. Data

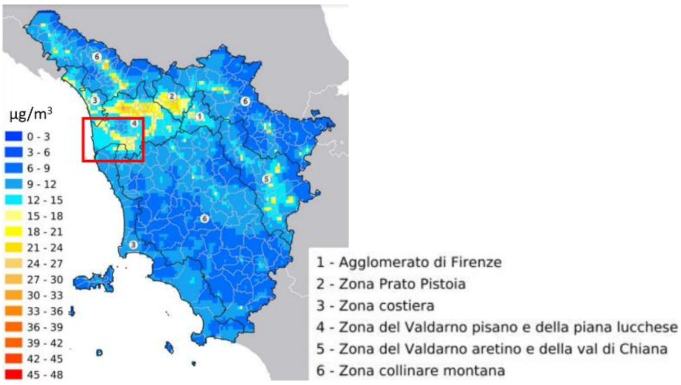
2.1. Study area

The study area is part of "Zona del Valdarno pisano e della piana lucchese" which is the most critical area in terms of air quality in the Tuscany region as it is shown in Figure 1. The map displays the output of the SPARTA modelling system run officially by the LAMMA consortium for ARPAT (Busillo *et al.* 2022) and

[2] https://cros.ec.europa.eu/multi-mno-project
[3] https://cros.ec.europa.eu/mno-minds

represents the average PM10 concentration for the year corresponding to available MNO data, the boundaries are not administrative units but environmental zoning. Valdarno is the focus of this study and is highlighted in the map within the red box.

Figure 1 – *Average PM10 concentration, year 2017.*



2.2. Mobile Network data

The spread of mobile communication technology has provided access to large amounts of data that can be used to analyse human and social behaviour. Usually, Mobile Network Operators (MNOs) store Call Detail Records (CDRs) which contain information about call events, such as time, duration, and location of the used network cell. In addition to CDRs, MNOs also collect additional data for tasks like network management and troubleshooting, e.g. signalling and probe data, which can offer a valuable resource for studying human behaviour and presence. We refer to the totality of these different data as Mobile Network Operator data (MNO data).

Unlike CDRs, signalling data present a higher level of complexity due to the large volume (an event is recorded on average every 2 seconds for each mobile device) and due to the peculiarities of mobile networks for different MNOs. Thanks to the progress in the available computational power and storage capability, many analyses are now focusing on signalling data, rather than CDRs, since in many countries and within some sub-groups phone calls are more and more infrequent due to the spread of online services of instant communication (instant messaging, VoIP calls).

Despite this limitation, in this study we use a sample of pseudo-anonymized CDRs of an Italian province, Pisa, for a study period of 6 weeks starting from January 1 to February 13, 2017. The data are described in detail in De Fausti *et al.*, 2024.

2.3. Air quality data

The air quality data are provided by ARPAT, the regional environmental agency which administers the monitoring network of the Tuscany region. In the study area, we have four monitoring stations. Table 1 specifies the location and type of each station.

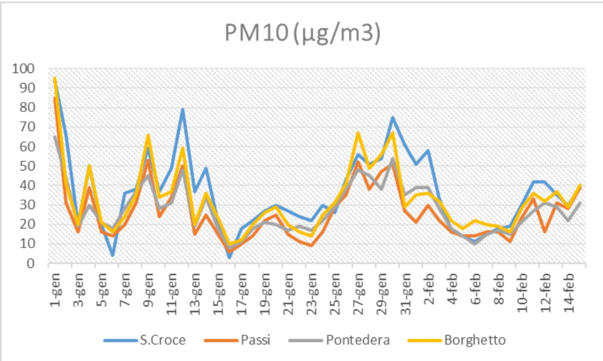
Table 1 – *Monitoring stations in the study zone.*

Station	Municipality	Station type	Zone
Passi	Pisa	Background	Urban
Borghetto	Pisa	Traffic	Urban
Pontedera	Pontedera	Traffic	Urban
Santacroce	Sc Sull'arno	Background	Suburban

Background monitoring stations measure air quality in areas not directly influenced by specific sources of pollution, providing information on regional pollution levels. Traffic monitoring stations, on the other hand, are located near roads and highways to measure pollution from vehicle emissions, reflecting local traffic-related air quality.

The graph in Figure 2 shows the trends in PM10 levels as recorded by the four stations during the study period.

Figure 2 – *PM10 concentration values measured during the study period.*



As mentioned, MNO data available for the current study are CDRs. Since CDR events are rare in a day, they are not appropriate for air pollutants regulated with hourly concentration limit. This is the reason why we use the PM10 pollutant to assess the exposure, for PM10 regulations set daily and yearly average concentration limits. The methodology can be easily extended to other pollutants with hourly concentration limits (O_3, NO_x, CO and SO_2), when other MNO data (e.g. signalling data) become available. The regulation sets for PM10 concentration an annual average limit of 40 µg/m³ and a daily average limit of 50 µg/m³.

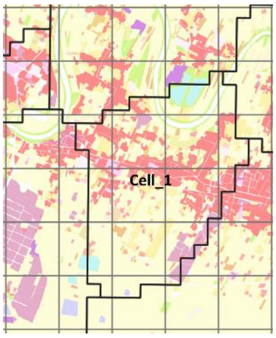
3. Data elaboration and methods

3.1. Positional and temporal features of mobile data and monitoring stations

The geolocation of the SIM associated with each CDR is estimated through the coverage area of the cell to which the device is connected when the call starts and ends. The cell coverage area is defined as the Best Service Area (BSA), which is the area where the connected mobile device has the highest probability of being located and is provided by the MNO under a confidentiality agreement. MNOs estimate the probability by using several parameters, like the quality of the cell signal, the technology of the cell, the coverage orography, the density of potential users, etc. An example of BSA is represented in Figure 3.

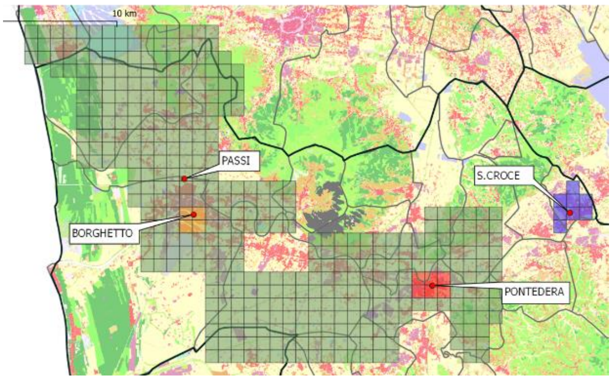
The data provide two events for each call, corresponding to the call start and the call end. Each event is identified by a timestamp and an associated cell ID. In the analysis, the time is discretized into 1-hour slots.

Figure 3– *The Best Service Areas of some cells (black lines) for LTE superimposed to the regular 1x1km grid used in the analysis (grey lines).*



Air monitoring stations' location is fixed in time and space and represents point-based data. Since we are interested in having a measure of the pollution levels for areas around the stations (covering the geographical area of our study), we used two different methods depending on the type of the considered stations. For background stations, we estimated the spatial representativeness of pollution measures by using the output of the SPARTA model system (Busillo *et al.* 2022). For traffic stations, we consider only the urban areas located around the stations. “Passi” station has the largest area of representativeness, extending towards the coast as can be seen in Figure 4.

Figure 4 – *Spatial coverage of the monitoring stations.*



3.2. Spatial and temporal integration: monitoring stations/cells association

CDRs and pollution data from monitoring stations have different spatial and temporal scales, an integration step is then required to get the final indicators. As a reference spatial system we adopted the INSPIRE (Infrastructure for Spatial Information in Europe) grid system provided by the European Commission to improve harmonization[4]. The chosen scale for the reference spatial grid is 1km x 1km.

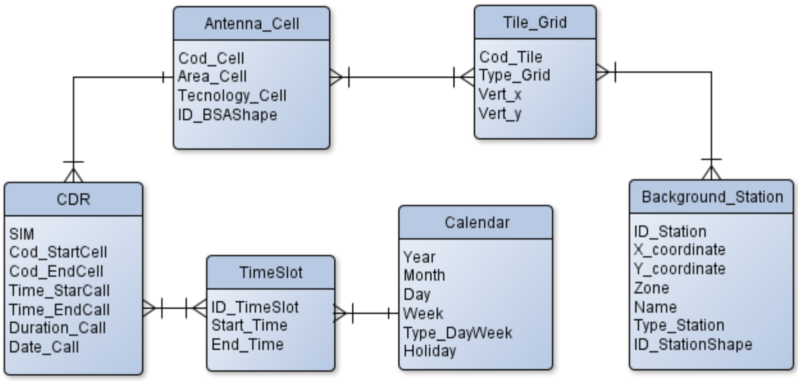
We implemented the following steps:

- the study area was divided into 1km x 1km tiles according to the INSPIRE grid, each tile is uniquely identified by a code;
- the cell coverage areas (BSAs) and the stations' spatial representativeness areas are encoded by shape files mapped to the INSPIRE 1km x 1km through spatial superimposition, as shown in Figure 3 and Figure 4.

After this processing, data were structured according to the entity and relationship scheme shown in Figure 5.

[4] INSPIRE, https://inspire.ec.europa.eu

Figure 5 – *Entity and Relationship Scheme of Data integration: CDR, station and grid*.



As shown in Figure 6, each cell coverage and each station coverage are associated to a set of tiles falling into the relative areas.

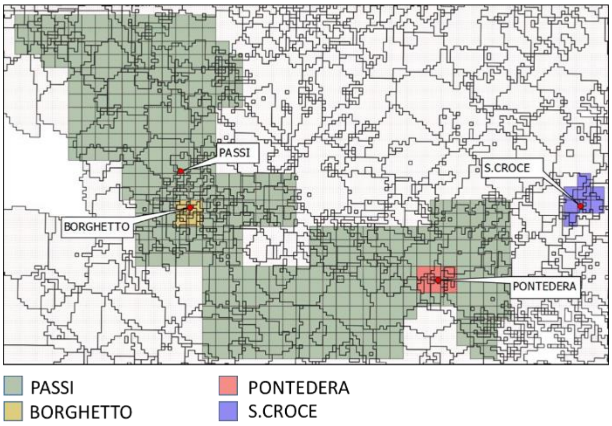
The presence of devices in the study area over the observation period has been estimated via a Permanence Score (PS). As already mentioned, the time dimension was discretized into 1-hour time slots. For each SIM, the PS value is assigned to each time slot of a day. The PS takes value:

- 1 if a call of the SIM occurs at any time in the time slot
- 0 if in the entire 1-hour time slot no calls occur

The permanence score is then a discretized measure of the SIM presence with 1 1-hour time resolution. A set of tiles is also assigned to each time-slot, which is the tiles composing the BSA of the cell to which the device is connected during the call. Note that if during a call the device connects to more than one cell, the time slot will be assigned to PS=1 for all the tiles corresponding to both cells. In this way, for each device and for each tile we calculated the amount of time (measured in number of one-hour time slots) during which a call occurs. Using the permanence score it is possible to carry out different analyses by aggregating the information over the entire period observed. We are able to carry out:

- **longitudinal analyses** for single SIM to obtain a daily exposure time and frequency. This aggregated and discretized information allows us to study the number of SIMs exposed in terms of hours, days or frequency.
- **spatial analyses** to detect, based on the presences and the temporal and spatial trend of pollution, how many SIMs have been affected.

Figure 6– *Overlapping of cells and stations coverage area.*

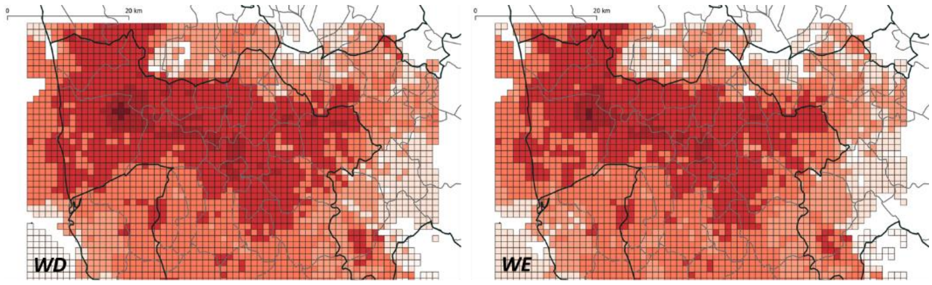


4. Results

4.1. Dynamic human presence

The maps in Figure 7 illustrate the SIMs population estimated using the methodology described in the previous section. It shows the dynamic presence in typical weekday (WD) and weekend (WE) days of January.

Figure 7 – *Dynamic human presence for typical weekday (WD) and weekend (WE) days.*



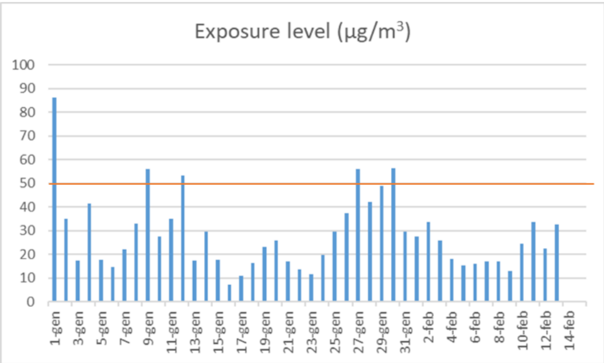
Different presence patterns can be observed with a level of detail that cannot be achieved with any other data sources. It can be observed that during the weekend days there is less dispersion of the population, likely indicating greater mobility for work-related reasons during the weekdays. At the same time Pisa urban centre shows a higher population density during the working day, likely due to the effect of commuting.

Numerical references are intentionally omitted for privacy reasons and because the dataset is not representative of the entire population.

4.2. Longitudinal analysis: daily exposure

Figure 8 shows the average daily exposure of the SIM population, obtained by averaging the values of each SIM separately observed throughout the day and being associated with different pollutant values in function of frequented places and time spent, as described in section 3.2.

Figure 8– *Daily average exposure of the observed SIM population*.



4.3. Longitudinal analysis: long term exposure

Due to the unavailability of a full year's data, we analysed the entire period at our disposal, going from the 1st of January to the 13th of February, as an indicator of long-term exposure.

The average monthly exposure of the observed population is obtained by averaging the values of each SIM. The spatial-temporal variations of each SIM individually contributed to the exposure estimate.

Table 2 shows the PM10 exposure levels of the SIM population. It can be noted that 1% of the SIM population is exposed to a pollution level exceeding the annual limit.

Table 2 – *PM10 exposure levels for the SIM population in the study period.*

Exposure Level (µg/m^3)	Percentage of population (%)
[0,10]	0
(10,20]	1
(20,30]	71
(30,40]	28
(40,50]	1
(50,Inf]	0

5. Concluding remarks

In this paper, we propose a first attempt to improve the estimations of air pollution exposure using MNO data. So far, in fact, the exposure to pollutants levels was calculated considering the usual residence of each inhabitant, disregarding the time spent outside official residence (e.g. at work, school, etc.) that can be significantly large. MNO data give the opportunity to assess the amount of time a user spends in different locations, and, hence, to increase substantially the accuracy of exposure estimations. However, CDRs are quite sparse to exploit the whole potentiality of MNO data. In future, we are going to apply the developed method to signalling data.

In this work, we consider the mobile devices detected in the study area as our target population. Future works will be devoted to the challenge of producing statistics referring to human beings present in the area. This is not a trivial task, since the present population is a composite of different groups: usual residents, commuters, tourists, passing-by, etc. and the mobile devices associated with them may require different estimation factors. Finally, additional data sources like satellites and models can help provide air quality data for areas not covered by monitoring stations.

References

BEELEN R, STAFOGGIA M, RAASCHOU-NIELSEN O, ANDERSEN ZJ, XUN WW, KATSOUYANNI K, DIMAKOPOULOU K, BRUNEKREEF B, WEINMAYR G, HOFFMANN B, WOLF K, SAMOLI E, HOUTHUIJS D, NIEUWENHUIJSEN M, OUDIN A, FORSBERG B, OLSSON D, SALOMAA V, LANKI T, YLI-TUOMI T, OFTEDAL B, AAMODT G, NAFSTAD P, DE FAIRE U, PEDERSEN NL, ÖSTENSON C-G, FRATIGLIONI L, PENELL J, KOREK M, PYKO A, et al. 2014. Long-term exposure to air pollution and cardiovascular mortality: an analysis of 22 European cohorts. *Epidemiology*, Vol. 25, pp. 368-78.

BUSILLO C., CALASTRINI F., GUARNIERI F. 2022, Rappresentatività spaziale delle stazioni della rete di monitoraggio di qualità dell'aria Toscana, Report Tecnico *Consorzio LaMMA – Laboratorio di Monitoraggio e Modellistica Ambientale per lo sviluppo sostenibile*, https://www.regione.toscana.it/-/elenco-pubblicazioni-inerenti-la-rappresentativita-spaziale-delle-stazioni-di-rilevamento-della-qualita-dell-aria-in-toscana

DE FAUSTI F., RADINI R., TUOTO T., VALENTINO L. 2024. Mobile phone data for population estimates and for mobility and commuting pattern analyses, *RIEDS - Rivista Italiana di Economia, Demografia e Statistica*, Vol. 78, No.1, pp. 235-247.

DEWULF, B., NEUTENS, T., LEFEBVRE, W. et al. 2016. Dynamic assessment of exposure to air pollution using mobile phone data. *Int J Health Geogr*, Vol. 15, No.14. https://doi.org/10.1186/s12942-016-0042-z

DIAS D, TCHEPEL O. 2018. Spatial and Temporal Dynamics in Air Pollution Exposure Assessment. *Int J Environ Res Public Health*. Vol. 15, No.3, pp. 558. https://doi.org/10.3390/ijerph15030558

GRAY C., EDWARDS S., LYNN M. 2013. Race, socioeconomic status, and air pollution exposure in North Carolina, *Environmental Research*, Vol. 126, pp. 152-158, ISSN 0013-9351, https://doi.org/10.1016/j.envres.2013.06.005

Erika CERASTI, Istat, erika.cerasti@istat.it
Cristina FARICELLI, Istat cristina.faricelli@istat.it
Paolo MATTERA, Istat, paolo.mattera@istat.it
Roberta RADINI, Istat, roberta.radini@istat.it
Tiziana TUOTO, Istat, tuoto@istat.it