Driving Factors Toward Accurate Mobile Opportunistic Sensing in Urban Environments

Driving Factors Toward Accurate Mobile Opportunistic Sensing in Urban Environments

Marco Fiore, *Member, IEEE,* Alessandro Nordio, *Member, IEEE,* Carla-Fabiana Chiasserini, *Senior Member, IEEE*

Abstract—The dramatic increase in the number and sensing capabilities of mobile devices is fostering opportunistic sensing as a paramount data collection paradigm in smart cities. According to this paradigm, sensing of large-scale phenomena is autonomously performed by mobile devices that provide irregular samples in time and space. The collected data is then transferred to a central controller, and processed so as to obtain a representation of the phenomenon. In this paper, we investigate the factors that impact the accuracy of mobile opportunistic sensing. Specifically, we characterize the accuracy of a phenomenon representation obtained from samples collected by mobile devices and processed through the popular LMMSE filter. We do so by drawing on random matrix theory, which allows us to deal with irregularly spaced samples. Our analytical expressions capture the fundamental relationships existing between the accuracy and the parameters of mobile opportunistic sensing. We apply our analytical results to a realistic scenario where atmospheric pollution samples are collected by vehicular and pedestrian users. We validate the proposed analytical framework, and then exploit the model to investigate the impact on mobile sensing accuracy of a number of parameters. These include the pedestrian and vehicle density, the participation ratio to the sensing application, the type of phenomenon to be sensed, and the level of noise and position errors affecting the collected samples.

1 Introduction

Opportunistic sensing is a specific type of mobile crowdsensing that leverages the ever-growing availability of sensing devices embedded in commodity hardware. It allows widespread, automated information collection by privately-owned smartphones and tablets, as well as cars or even public-service bicycles, by exploiting the trend for all such mobile devices to be increasingly equipped with GPS, cameras and different types of sensors [1]–[4].

The collected data can relate to a number of phenomena, including air quality, noise level, road traffic state, street surface and pavement conditions. This information is typically difficult or expensive to collect via traditional approaches that involve, e.g., air-pollution monitoring stations, induction loop counters, or in-situ inspection by human personnel. Mobile devices can instead put their Internet connectivity in use to upload massive amounts of samples that are fine-grained and cover large geographical areas, and do so at virtually no cost [5], [6] and with no user involvement. However, due to the mobility and lack of synchronization of devices, samples are collected at irregular points in time and space. Thus, an accurate and comprehensive characterization of the original phenomenon requires that a remote Internet-based processing center runs reconstruction techniques accounting for sample irregularity.

The enormous potential yielded by mobile opportunistic sensing is however confronted by a number of challenges, which include that participating devices must (i) be localized, (ii) provide a sufficiently accurate measurements of the monitored phenomenon, (iii) ensure geographical coverage and (iv) collect a substantial amount of measurements. In this paper, we address these precise aspects and aim at characterizing the level of accuracy achieved by a mobile opportunistic sensing process in the estimation of a physical phenomenon (hereinafter also referred to as signal). To this end, we consider a generic system where mobile devices participate in the opportunistic sensing process by collecting samples at irregularly spaced points (i.e., their locations). Mobile devices then transfer wirelessly their samples to an Internet-based processing center. In line with opportunistic sensing principles, data are collected and transmitted in a fully automated manner and with no user involvement. At the processing center, the signal is reconstructed from the collected samples by employing the well-known Linear Minimum Mean Square Error (LMMSE) filter [7]. This filter features a good performance-complexity tradeoff, is general-purpose, and is leveraged for the reconstruction of physical phenomena in practical use cases [8].

In order to analyze the system accuracy and its dependency on the sensing scenario parameters, we leverage recent results that have been derived for the reconstruction of signals from irregularly-spaced samples [9]. We thus develop an analytical method that allows us to evaluate the mean square error (MSE) between the original phenomenon and its estimate at the processing center. Our method accounts for all significant system parameters, i.e., the geographical distribution and the number of the participating sensing devices, their measurement and positioning errors, and the frequency characterization of the phenomenon under study. We apply our signal recon-

- *M. Fiore and A. Nordio are with CNR-IEIIT, Torino, Italy. E-mail: {marco.fiore, alessandro.nordio}@ieiit.cnr.it*
- *C.-F. Chiasserini is with the Department of Electronic and Telecommunications, Politecnico di Torino, Torino, Italy. She is also a research associate with CNR-IEIIT, Torino, Italy. E-mail: chiasserini@polito.it*

struction technique to a realistic urban scenario, featuring a faithful representation of the localization and mobility of the citywide vehicular and pedestrian population, as well as a practical reference phenomenon retrieved from real-world measurements in the region.

In summary, our main contributions are as follows:

- We present a model of the mobile opportunistic sensing process that accounts for all major system factors;
- By assuming that the LMMSE technique is used for signal reconstruction, we provide an expression for the estimation of a multidimensional phenomenon;
- We introduce a novel asymptotic methodology to compute the signal reconstruction accuracy when its bandwidth and the number of samples tend to infinity, while their ratio is constant. Such a technique is computationally efficient. As shown by our validation results, this approach provides a very good performance approximation, even for real-world scenarios where the above parameters take small values.
- We then use our analytical framework to derive results in a realistic scenario, exploiting dependable data sets. Our results show that it is possible to achieve an accurate reconstruction of the considered phenomenon from the samples collected through mobile opportunistic sensing. To that end, it is especially important that the samples collected by mobile devices are not too erroneous, and that a sufficiently high number of devices, in the order of a few tens per square kilometer, can provide coverage of the area. Other system parameters, including the type of geographical area, the daytime, or the positioning error, only play minor roles with respect to the level of accuracy attained by the mobile opportunistic sensing process.

The rest of the paper is organized as follows. After a discussion of the related work in Sec. 2, we introduce in Sec. 3 the model of the opportunistic sensing system and of the phenomenon under study. In Sec. 4, we present our analytical framework for the evaluation of the system performance. Sec. 5 describes the realistic urban population scenario under study, and presents the associated spatial distributions of handheld and vehicular devices. Sec. 6 illustrates the reference phenomenon considered in our performance evaluation, along with results on the reliability of the proposed model and on the impact of a vast range of system parameters on the opportunistic sensing accuracy. Finally, we draw concluding remarks in Sec. 7.

2 Related work

Mobile crowdsensing. Mobile crowdsensing envisages that mobile devices with sensing and communication capabilities collect and share information, so as to monitor some target phenomenon [1], [2]. The interest in mobile crowdsensing has rapidly grown in the last few years, fostered by its potential capability to provide fine-grained sensing at almost no dedicated infrastructure cost.

Mobile crowdsensing paradigms span in between two fundamental approaches, i.e., *opportunistic sensing* and *participatory sensing*. Opportunistic sensing consists in a fully autonomous process, distributely run by the devices without any human intervention. Conversely, in participatory sensing individuals are actively involved in contributing sensor data.

Opportunistic sensing. In our work, we focus on the former approach. The rationale is that opportunistic sensing is in general more acceptable to users, as it reduces the actions they have to undertake; it is thus expected to result in much a wider adoption than participatory sensing applications [1], [2]. The experiments by Cardone *et al.* [10], [11] are especially enlightening here, and support the point above. The authors aim at identifying sensible users depending on the sensing task, and at incentivizing them to participate in the process. To that end, they develop a full-featured experimental platform for mobile sensing, allowing them to profile users and evaluate the incentives in a real-world environment. Lessons learned indicate users' participation and attitude as a major concern.

However, opportunistic sensing is also more complex to implement than participatory sensing, since information that could be easily input by a human user need to be inferred automatically. A typical example is mobile device context information, since a correct sensing often requires to know the specific situation the device is in (e.g., in the users' hand, within his pocket, or laying on the dashboard of his car). Techniques have being developed to address such issues: e.g., Miluzzo *et al.* [12] propose a practical approach to extract mobile phone context using data from the device microphone, camera light sensor, accelerometer, gyroscope, and compass.

Practical applications leveraging the opportunistic sensing principle have already been demonstrated on specific application use cases. As an example, Zhang *et al.* [13] equip a small fleet of buses with sensing and communication interfaces: the goal is the generation of maps of carbon monoxide in Beijing, China, from samples collected through opportunistic sensing. The authors show how this approach enables the identification of significant temporal correlations between air pollution and road traffic levels. Other representative examples are the works by Zhu *et al.* [14] and Du *et al.* [15], [16], who aim at building road traffic maps through opportunistic sensing performed by vehicles. To that end, the authors leverage different approaches, including compressive sensing [14], matrix completion [15], and use of controllable patrol vehicles [16].

Our focus is different from those of the vast majority of previous works on opportunistic sensing. Indeed, we do not develop a technique to perform opportunistic sensing, but we assess the accuracy of the opportunistic sensing process. Moreover, we do not target one specific phenomenon, but we consider a general approach that can be applied to any physical phenomenon. To attain our objective, we model the quality of the sensed phenomenon once it is reconstructed at the data processing center. Our approach let us evaluate the impact of a number of system parameters that are hard (or even impossible) to control in real-world deployments (see Sec. 3).

To the best of our knowledge, the only work to take a perspective comparable to ours is that by Zhao *et al.* [17], who aim at understanding the temporal and spatial frequency of sampling granted by opportunistic sensing. However, their approach is completely different from ours. Specifically, the authors assume that a taxicab fleet is equipped with sensing

capabilities, and study the interval elapsed between two consecutive visits of taxis to each areas of a large conurbation, as a measure of sensing coverage. Our methodology is instead based on signal processing techniques, and allows accounting not only for the spatio-temporal distribution of mobile sensing devices, but also for the data processing phase. We are thus able to assess the actual quality of the reconstructed phenomenon, in terms of its mean square error (MSE), rather than just in terms of a coverage metric. This also makes the outcome of the two methodologies not directly comparable.

Knowledge discovery. As a concluding remark on the related literature, we recall that opportunistic sensing is often considered as a way to collect data that can be later mined for knowledge discovery. In such a context, data sensing is a preliminary (and often irrelevant) step, where information about a number of different phenomena is collected, using multiple sources, into separate databases. Then, the databases are integrated and mined so as to infer the physical fact of interest. There exists a vast literature that builds on this approach, so as to characterize, e.g., air quality dynamics in urban areas starting from road traffic levels, meteorological information, human mobility patterns, and point-of-interest locations [13], [18]. Similarly, databases of cellular data traffic, subway occupancy, and taxicab and bus routes have been leveraged to infer transit patterns in large cities [19].

However, the focus of these works is on the database integration and knowledge discovery phases, and not on the opportunistic sensing. Instead, our work focuses on estimating the accuracy of a pure opportunistic sensing process, where spatiotemporal samples of a target phenomenon are collected and processed to detect the precise phenomenon the samples refer to.

3 System model

Our aim is to evaluate the accuracy of mobile opportunistic sensing, which depends on many aspects. We thus identify a narrow list of factors that are general enough to account for all of the major practical aspects characterising the opportunistic sensing process. The factors are as follows.

- F1 The number of available samples. This depends on the number of mobile devices participating in the sensing process, the device sampling rate and duty cycle: all of these aspects can also be seen as means to control and limit the energy consumption of mobile devices [20], especially in presence of services requiring continuous sensing [1]. In addition, the number of samples received by the processing center also reflects the reliability level of the wireless channel.
- F2 The error in estimating the sampling locations, i.e., the positions of mobile devices detected through GPS or other localization techniques (e.g., via recording of cellular or Wi-Fi signals).
- F3 The spatial distribution of the mobile sensing devices over the geographical area, which depends on the movement patterns of devices and have a varying degree of irregularity over time.
- F4 The varying accuracy level of the mobile sensors. This may depend on which user device the sensors are embedded in (e.g., in-vehicle or smartphone), or on their context (e.g., sensors in smartphones that are carried in a bag rather than handheld).
- F5 The phenomenon spectral characteristics, i.e., the signal bandwidth, which is expressed as number of harmonics used to represent the phenomenon.
- F6 The number of dimensions (spatial coordinates and/or time) over which the signal is defined [21].

Our model accounts for all of the factors listed above. Its formulation can accomodate different communication and localization technologies (F1 and F2), energy management policies (F1), device deployment and mobility (F3), device type and context (F4). Moreover, the model is general, and can be applied to a variety of physical phenomena with diverse complexity and dimensions[1] (F5 and F6). To our knowledge, no previous analytical framework has been proposed for the performance analysis of mobile opportunistic sensing, which accounts for all of the above factors.

Overall, our model allows deriving the level of accuracy of the phenomenon (i.e., signal) reconstruction, as a function of all of the factors above. E.g., it determines the minimum number of samples necessary to achieve a desired MSE for a target phenomenon, when employing specific technologies and a given set of sensing devices. Thus, it provides useful guidelines for the configuration of system parameters.

We stress that the model is applicable to both delay-tolerant and real-time sensing applications. In the first case, the phenomenon timescale is such that one can wait for all devices to upload their data: this is, e.g., the case for the pollution monitoring scenario we consider in our evaluation. In the case of real-time monitoring of, e.g., road safety services, the requirements in terms of latency can be translated in the model by further limiting the number of samples available by the imposed time deadline (e.g., due to the finite sampling rate of mobile devices or to the level of data transfer reliability). Additional requirements, e.g., user privacy preservation during the opportunistic sensing, are orthogonal to our study.

Below we first introduce some notations and definitions that will be used throughout our analysis (Sec. 3.1). We then define a formal representation of the sensed physical phenomenon that is opportunistically sensed (Sec. 3.2).

3.1 Notation and definitions

Throughout the paper, vectors and matrices are denoted by bold lowercase and uppercase letters, respectively. $\mathbf{I}$ represents the identity matrix. The superscripts $^{\mathsf{T}}$ and $^{\mathsf{H}}$ denote matrix transpose and conjugate-transpose, respectively, while $\mathsf{Tr}\{\cdot\}$ represents the matrix-trace operator. The expectation of a random variable a is denoted by $\mathbb{E}[a]$.

1. Although our analysis can accommodate an arbitrary number of dimensions, in this work we focus on phenomena over a bidimensional geographical regions and in a given time period. This case reflects a typical sensing procedure performed in a flat urban environment and that is repeated at different times of the day.

Definition 3.1. *Let us consider an Hermitian random matrix* $\mathbf{A}^{(n)}$ *of size* $n \times n$ *with random eigenvalues* $\lambda_1, \ldots, \lambda_n$*. Its average empirical spectral distribution is defined as* $F_{\mathbf{A}}^{(n)}(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[1\{\lambda_i \leq z\}\right]$ *where* $1\{\cdot\}$ *is the indicator function. The limit* $F_{\mathbf{A}}(z) = \lim_{n\to\infty} F_{\mathbf{A}}^{(n)}(z)$*, if it exists, denotes the Limiting Spectral Distribution of the sequence of matrices* $\{\mathbf{A}^{(n)}\}_{n\in\mathbb{N}}$*. The corresponding* asymptotic probability density function*, if it exists, is denoted by* $f_A(z)$*.*

Definition 3.2. *Let us consider the sequence of random matrices* $\{\mathbf{A}^{(n)}\}_{n\in\mathbb{N}}$ *of size* $n \times n$*. If the limit exists, we define its tracial state [22] as*

$$\phi(\mathbf{A}) = \lim_{n\to\infty} \frac{1}{n} \mathsf{Tr}\left\{\mathbb{E}\left[\mathbf{A}^{(n)}\right]\right\}. \tag{1}$$

Definition 3.3. *Let us consider the sequence of random matrices* $\{\mathbf{A}^{(n)}\}_{n\in\mathbb{N}}$ *of size* $n \times n$ *and a positive real number* γ*. We will denote by*

$$\eta_{\mathbf{A}}(\gamma) = \phi\left((\mathbf{I} + \gamma\mathbf{A})^{-1}\right) \tag{2}$$

the η*-transform of a random variable that follows the Limiting Spectral Distribution of the sequence* $\{\mathbf{A}^{(n)}\}_{n\in\mathbb{N}}$*.*

For simplicity of notation, in the following we drop the superscript (n) and $\mathbf{A}$ denotes the generic element of a sequence of random matrices $\{\mathbf{A}^{(n)}\}_{n\in\mathbb{N}}$.

3.2 Signal model

We consider a two-dimensional signal. Thus, the signal can be described in a general manner by defining a function $s(\mathbf{x})$ over the region $\mathcal{R} \in \mathbb{R}^2$. For simplicity of presentation, we define $\mathcal{R}$ as a square region of side 1, i.e., $\mathcal{R} = [-1/2, 1/2)^2$, although any more general sizes and shapes can be considered.

In our scenario, the non equally spaced samples of the phenomenon are provided by m devices (factor F1), irregularly deployed over $\mathcal{R}$. We assume that sampling devices are equipped with a positioning system (e.g., GPS), so that each sample can be characterized by the location at which it has been taken. Positioning is however affected by errors (factor F2). In general, the position of the q-th sensing device, $q = 1, \ldots, m$, can be described by the vector $\mathbf{p}_q = [p_{1q}, p_{2q}]^{\mathsf{T}} \in \mathcal{R}$ given by

$$\mathbf{p}_q = \hat{\mathbf{p}}_q + \boldsymbol{\delta}_q$$

where $\hat{\mathbf{p}}_q = [\hat{p}_{1q}, \hat{p}_{2q}]^{\mathsf{T}}$ is the estimated position and $\boldsymbol{\delta}_q = [\delta_{1q}, \delta_{2q}]^{\mathsf{T}}$ is the position error (or displacement). Here we assume that $\boldsymbol{\delta}_q$ are i.i.d. Gaussian distributed random variables, with zero mean and covariance $\sigma_\delta^2\mathbf{I}$. We remark that σ_δ^2 is the variance of the position error. Due to the mobility of pedestrian and vehicular users, we can consider the positions $\hat{\mathbf{p}}_q$'s as instances of a random variable with distribution $f_{\hat{\mathbf{P}}}(\mathbf{x})$, $\mathbf{x} \in \mathcal{R}$ (factor F3). Such distribution depends on the specific scenario; we will detail the procedure to derive $f_{\hat{\mathbf{P}}}(\mathbf{x})$ from experimental data in Sec. 5. The sample taken by the q-th sensing device ($q = 1, \ldots, m$) is then given by

$$y_q = s(\mathbf{p}_q) + z_q$$

where z_q is the measurement error due, e.g., to sensing noise and/or quantization inaccuracy (factor F4). The elements $\mathbf{z} = [z_1, \ldots, z_m]^{\mathsf{T}}$ are assumed to be zero-mean uncorrelated random variables with known diagonal covariance matrix $\boldsymbol{\Sigma}$. Furthermore, $\mathbf{z}$ is independent of all other random variables of the system. Note that the diagonal entries of $\boldsymbol{\Sigma}$ can take different values due to the different accuracy of sensing devices (e.g., sensors aboard vehicles or embedded in pedestrians smartphones that can be handheld or carried in a bag). The information on the operational conditions under which sensors operate can be hardcoded in case of technological differences (based, e.g., on the type of device), whereas it can be inferred automatically [12] and communicated along with the sample, in case of context-dependent diversity.

We then observe that any physical phenomenon can be approximated by a band-limited signal, i.e., a finite number of harmonics (factor F5). Thus, it can be written through its Fourier series expansion as:

$$s(\mathbf{x}) \approx \frac{1}{2n+1} \sum_{\ell_1=-n}^{n} \sum_{\ell_2=-n}^{n} a_{\ell_{12}} \mathrm{e}^{\mathrm{j}2\pi(\ell_1 x_1 + \ell_2 x_2)} \tag{3}$$

where $\ell_{12} = \ell_1 + (2n+1)\ell_2$. Here the integer n (i.e., the approximate one-sided bandwidth of the signal) is chosen so that most part of the signal energy falls in the first $2n+1$ harmonics per dimension. In Eq. (3), the terms $a_{\ell_{12}}$, $-2n(n+1) \leq \ell_{12} \leq 2n(n+1)$, denote the signal spectrum coefficients, while $\mathbf{x} = [x_1, x_2]^{\mathsf{T}}$ with $\mathbf{x} \in \mathcal{R}$. We remark that the above signal expression is very general and can represent different phenomena.

The vector of signal samples at the true sampling points $\mathbf{s} = [s_1, \ldots, s_m]^{\mathsf{T}}$ ($s_q = s(\mathbf{p}_q)$) can be approximated as $\mathbf{s} \approx \mathbf{V}_{\mathbf{P}}^{\mathsf{H}}\mathbf{a}$ where $\mathbf{a} = [a_{-2n(n+1)}, \ldots, a_{2n(n+1)}]^{\mathsf{T}}$, $\mathbf{V}_{\mathbf{P}}$ is an $n^2 \times m$ multifold Vandermonde matrix [9] with entries:

$$(\mathbf{V}_{\mathbf{P}})_{\ell_{12},q} \doteq (2n+1)^{-1} \exp\left(-2\pi\mathrm{j}(\ell_1 p_{1q} + \ell_2 p_{2q})\right) \tag{4}$$

and $p_{iq} = (\mathbf{P})_{iq}$. The subscript $\mathbf{P}$ indicates that the matrix $\mathbf{V}_{\mathbf{P}}$ is a function of the true positions of the sensing devices $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_m]$. About spectrum $\mathbf{a}$, since in general we do not have any *a priori* information about its statistic, we assume that its elements are uncorrelated with zero mean and variance σ_a^2. Without loss of generality and for normalization reasons, we assume $\sigma_a^2 = 1$. We then define

$$\beta_{n,m} \doteq (2n+1)^2/m$$

as the ratio of the total number of the signal harmonics, $(2n+1)^2$, to the number of sensing devices, m. This parameter also corresponds to the *aspect ratio* (ratio of the number of rows to the number of columns) of matrix $\mathbf{V}_{\mathbf{P}}$ and plays an important role in our analysis. Overall, the vector of samples, $\mathbf{y} = [y_1, \ldots, y_m]^{\mathsf{T}}$, taken by the m devices, can be written as

$$\mathbf{y} = \mathbf{s} + \mathbf{z} = \mathbf{V}_{\mathbf{P}}^{\mathsf{H}}\mathbf{a} + \mathbf{z} \tag{5}$$

where $\mathbf{s} = [s_1, \ldots, s_m]^{\mathsf{T}}$ is the vector of signal values at the true sampling points and $\mathbf{z}$ is the vector of measurement errors at the sensing devices.

4 ACCURACY OF SIGNAL RECONSTRUCTION IN MOBILE OPPORTUNISTIC SENSING

We now investigate the accuracy of the signal that is reconstructed from the irregular samples collected through mobile opportunistic sensing. In particular, we present the MSE metric representing the reconstruction accuracy and its analytical expression, which accounts for all of the factors listed in Sec. 3.

As detailed in Sec. 4.1, we assume that the well-known LMMSE filter is used for signal reconstruction, and we tailor our analysis to such reconstruction technique. The rationale behind this choice is that linear estimators are commonly employed due to their simplicity and, among these, the LMMSE filter is known to provide the best performance in terms of MSE. Also, the LMMSE is a general-purpose filter that can be adopted for the reconstruction of different phenomena, and is used in practical applications [8]. This fact, along with the LMMSE mathematical tractability, makes this reconstruction methodology suitable for the study of the impact of the major factors characterizing mobile opportunistic sensing and for deriving guidelines for the system design. We remark that (i) although the reconstruction technique is a standard one, the expression of the LMMSE filter in a multidimensional scenario is original, and (ii) our methodology could be extended to other linear estimators as well.

In Sec. 4.2 we introduce our novel methodology to compute the accuracy of the phenomenon reconstruction. Our approach is based on the observation that the locations where the phenomenon is sampled, i.e., the positions of the mobile devices participating in the opportunistic sensing, are instances of random variables. It follows that the analysis is based on random – rather than deterministic – matrices. By leveraging random matrix theory we are able to derive asymptotic expressions for the MSE, i.e., considering that the phenomenon bandwidth and number of samples tend to infinity, while their ratio is constant. These expressions represent a computationally efficient way to characterize the MSE achieved through opportunistic sensing, which proves to hold also for practical finite scenarios. Additionally, in the following we highlight how our expressions reflect the impact of the factors that play a major role on the performance of opportunistic sensing, as outlined in Sec. 3.

4.1 MSE performance metric

Once samples $\mathbf{y}$ and position estimates $\widehat{\mathbf{P}} = [\hat{\mathbf{p}}_1, \ldots, \hat{\mathbf{p}}_m]$ are acquired, an estimate $\hat{s}(\mathbf{x})$ of $s(\mathbf{x})$, for $\mathbf{x} \in \mathcal{R}$, can be obtained by applying a suitable signal reconstruction algorithm.

The reconstructed signal can as well be approximated by its Fourier series and thus written as

$$\hat{s}(\mathbf{x}) \approx \frac{1}{2n+1} \sum_{\ell_1=-n}^{n} \sum_{\ell_2=-n}^{n} \hat{a}_{\ell_{12}} \mathrm{e}^{\mathrm{j}2\pi(\ell_1 x_1 + \ell_2 x_2)} \tag{6}$$

where $\hat{\mathbf{a}} = [\hat{a}_{-2n(n+1)}, \ldots, \hat{a}_{2n(n+1)}]^{\mathsf{T}}$ denotes the estimated spectrum. The reconstruction accuracy can be measured in terms of MSE and is a function of positions $\widehat{\mathbf{P}}$

$$\mathrm{MSE}(\widehat{\mathbf{P}}) = \mathbb{E}\left[\int_{\mathcal{R}} |\hat{s}(\mathbf{x}) - s(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x}\right] \tag{7}$$

where the operator $\mathbb{E}[\cdot]$ averages over the randomness contained in $\hat{s}(\mathbf{x})$ and $s(\mathbf{x})$. Replacing Eq. (3) and Eq. (6) in Eq. (7), we obtain

$$\begin{aligned} \mathrm{MSE}(\widehat{\mathbf{P}}) &= \mathbb{E} \int_{\mathcal{R}} \left| \sum_{\ell_1,\ell_2=-n}^{n} \frac{\hat{a}_{\ell_{12}} - a_{\ell_{12}}}{(2n+1)\mathrm{e}^{-\mathrm{j}2\pi(\ell_1 x_1 + \ell_2 x_2)}} \right|^2 \mathrm{d}\mathbf{x} \\ &= \mathbb{E} \sum_{\ell_1,\ell_2,\ell_3,\ell_4=-n}^{n} \frac{(\hat{a}_{\ell_{12}} - a_{\ell_{12}})(\hat{a}_{\ell_{34}} - a_{\ell_{34}})}{(2n+1)^2} \\ &\quad \cdot \int_{\mathcal{R}} \mathrm{e}^{\mathrm{j}2\pi(\ell_1 x_1 + \ell_2 x_2 - \ell_3 x_1 - \ell_4 x_2)} \, \mathrm{d}\mathbf{x} \\ &= \frac{\mathbb{E}\left[\|\hat{\mathbf{a}} - \mathbf{a}\|^2\right]}{(2n+1)^2} \end{aligned} \tag{8}$$

where $\ell_{34} = \ell_3 + (2n+1)\ell_4$. The last line of Eq. (8) shows that the MSE depends on the estimate of the signal spectrum. This is a general expression; in order to further proceed in the MSE computation, we need to explicit the reconstructed signal, or, equivalently, $\hat{\mathbf{a}}$.

In the literature, many estimators for $\mathbf{a}$ have been proposed. As mentioned, we consider a linear estimator, i.e.,

$$\hat{\mathbf{a}} = \mathbf{B}^{\mathsf{H}} \mathbf{y}$$

where matrix $\mathbf{B}$ is called *linear filter*. Among linear estimators, the LMMSE filter is derived by minimizing the MSE in Eq. (8) over the possible choices of matrix $\mathbf{B}$. In general, matrix $\mathbf{B}$ should depend (among other system parameters) on the true sampling positions $\mathbf{P}$ and on the measurement error covariance matrix $\mathbf{\Sigma}$. For simplicity in the following we assume $\mathbf{\Sigma} = \sigma_z^2 \mathbf{I}$, i.e., all mobile devices exhibit the same level of accuracy. Since the true sampling positions, $\mathbf{P}$, are unavailable, the estimated positions $\widehat{\mathbf{P}}$ are to be used instead. This results in a mismatch between the filter and the samples $\mathbf{y}$, which depends on $\mathbf{P}$ (see Eq. (5)). In particular, such mismatch depends on the statistics of the displacements $\mathbf{\Delta} = [\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_q]$ and affects the MSE by reducing the reconstruction accuracy.

Lemma 4.1. *For the system model under consideration and for any given realization of the random sampling positions $\widehat{\mathbf{P}}$, the expression of the LMMSE filter, $\mathbf{B}$, is given by*

$$\mathbf{B} = \mathbf{V}_{\widehat{\mathbf{P}}}^{\mathsf{H}} \mathbf{C} (\mathbf{C} \mathbf{V}_{\widehat{\mathbf{P}}} \mathbf{V}_{\widehat{\mathbf{P}}}^{\mathsf{H}} \mathbf{C} + \gamma \mathbf{I})^{-1} \tag{9}$$

where

$$\gamma = \sigma_z^2 + 1 - \frac{\mathsf{Tr}\{\mathbf{C}^2\}}{(2n+1)^2} \tag{10}$$

is a signal-to-noise ratio which takes into account the penalty introduced by the measurement error (σ_z^2) and the position error ($\mathbf{\Delta}$). Indeed, $\mathbf{C}$ is a $(2n+1)^2 \times (2n+1)^2$ diagonal matrix and its elements depend on the characteristic function of the displacements. In particular, when $\boldsymbol{\delta}_q$, $q = 1, \ldots, m$, are i.i.d. Gaussian with zero mean and covariance $\sigma_\delta^2 \mathbf{I}$, the elements of $\mathbf{C}$ are given by

$$(\mathbf{C})_{\ell_{12},\ell_{12}} = \exp\left(-2\pi^2 \sigma_\delta^2 (\ell_1^2 + \ell_2^2)\right). \tag{11}$$

In general, $\gamma \geq \sigma_z^2$ and, in the special case $\sigma_\delta^2 = 0$ (i.e., no position errors), we have $\mathbf{C} = \mathbf{I}$, $\mathsf{Tr}\{\mathbf{C}^2\} = (2n+1)^2$ and $\gamma = \sigma_z^2$.

Proof: The proof can be found in the Supplemental Material. □

Using the LMMSE filter Eq. (9) in Eq. (8), the achieved MSE is given by

$$\mathrm{MSE}(\widehat{\mathbf{P}}) = \frac{\mathsf{Tr}\left\{\left(\mathbf{I}+\frac{1}{\hat{\sigma}_z^2}\mathbf{C}\mathbf{V}_{\widehat{\mathbf{P}}}\mathbf{V}_{\widehat{\mathbf{P}}}^{\mathsf{H}}\mathbf{C}\right)^{-1}\right\}}{(2n+1)^2}. \quad (12)$$

Remark. The MSE in Eq. (12) corresponds to that achieved by a system whose output signal

$$\hat{\mathbf{y}} = \mathbf{V}_{\widehat{\mathbf{P}}}^{\mathsf{H}}\mathbf{C}\mathbf{a} + \hat{\mathbf{z}} \quad (13)$$

is filtered by the LMMSE filter (in Eq. (9)), and where the noise $\hat{\mathbf{z}}$ has covariance $\gamma\mathbf{I}$. Note that the LMMSE filter depends on matrix $\mathbf{V}_{\widehat{\mathbf{P}}}^{\mathsf{H}}\mathbf{C}$ and it is matched to $\hat{\mathbf{y}}$. By comparing the signals in Eqs. (5) and (13), we observe that the effect of the uncertainty in measuring positions $\mathbf{P}$ is two-fold: (i) it increases the noise variance from σ_z^2 to γ, and (ii) it modifies the system transfer function through the weight matrix $\mathbf{C}$.

4.2 Asymptotic analysis

In order to evaluate the performance of mobile opportunistic sensing in large-scale scenarios, we resort to asymptotic analysis. The idea is to compute the MSE in the case where the number of harmonics, $(2n+1)^2$, and the number of samples, m, grow to infinity, while their ratio $\beta_{n,m}$ is kept constant. The rationale behind this choice is that the asymptotic MSE can be handled much more easily than the MSE for finite values of m and n, and that, as shown by our validation results, it is an excellent approximation of the MSE already for small values of m and n.

The asymptotic MSE is defined as

$$\begin{aligned}
\mathrm{MSE}^{(\infty)} &= \lim_{n,m\to\infty} \mathop{\mathbb{E}}_{\widehat{\mathbf{P}}}\left[\mathrm{MSE}(\widehat{\mathbf{P}})\right] \\
&= \lim_{n,m\to\infty} \mathop{\mathbb{E}}_{\widehat{\mathbf{P}}}\left[\frac{1}{n^2}\mathsf{Tr}\left\{\left(\mathbf{I}+\frac{\mathbf{C}\mathbf{V}_{\widehat{\mathbf{P}}}\mathbf{V}_{\widehat{\mathbf{P}}}^{\mathsf{H}}\mathbf{C}}{\gamma}\right)^{-1}\right\}\right] \\
&= \phi\left(\left(\mathbf{I}+\frac{1}{\gamma^{(\infty)}}\mathbf{C}\mathbf{R}\mathbf{C}\right)^{-1}\right) \\
&= \eta_{\mathbf{CRC}}\left(\frac{1}{\gamma^{(\infty)}}\right) \quad (14)
\end{aligned}$$

where the tracial state $\phi(\cdot)$ and the η-transform have been defined in Eq. (1) and Eq. (2), respectively. Furthermore, $\mathbf{R} = \mathbf{V}_{\widehat{\mathbf{P}}}\mathbf{V}_{\widehat{\mathbf{P}}}^{\mathsf{H}}$ and $\gamma^{(\infty)} = \lim_{n,m\to\infty}\gamma$.

Note that, using Eq. (10), the asymptotic SNR is given by:

$$\begin{aligned}
\gamma^{(\infty)} &= \lim_{n,m\to\infty}\left(\sigma_z^2 + 1 - \frac{\mathsf{Tr}\{\mathbf{C}^2\}}{(2n+1)^2}\right) \\
&= \sigma_z^2 + 1 - \phi(\mathbf{C}^2). \quad (15)
\end{aligned}$$

In general, for an arbitrary integer h, the tracial state $\phi(\mathbf{C}^h)$ can be written as

$$\begin{aligned}
\phi(\mathbf{C}^h) &= \lim_{n\to\infty}\frac{1}{(2n+1)^2}\sum_{\ell_1,\ell_2}\exp\left(-2h\pi^2\sigma_\delta^2(\ell_1^2+\ell_2^2)\right) \\
&= \lim_{n\to\infty}\prod_{j=1}^{2}\frac{1}{2n+1}\sum_{\ell=-n}^{n}\exp\left(-2h\pi^2\sigma_\delta^2\ell^2\right)
\end{aligned}$$

Now, by switching the limit and the product operator we get

$$\begin{aligned}
\phi(\mathbf{C}^h) &= \left(\lim_{n\to\infty}\frac{1}{2n+1}\sum_{\ell=-n}^{n}\exp\left(-2h\pi^2\sigma_\delta^2\ell^2\right)\right)^2 \\
&= \left(\int_{-1/2}^{+1/2}\exp\left(-2h\pi^2\omega^2\beta w^2\right)\,\mathrm{d}w\right)^2 \\
&= \frac{\left(\mathrm{erf}\left(\sqrt{\frac{h\beta}{2}}\pi\omega\right)\right)^2}{2\pi h\beta\omega^2} \quad (16)
\end{aligned}$$

where $\omega^2 = \frac{\sigma_\delta^2}{(1/\sqrt{m})^2} = m\sigma_\delta^2$ is the ratio between the variance of the position error σ_δ^2 and the average device separation, $1/\sqrt{m}$ and $\beta = \lim_{n,m\to\infty}\beta_{n,m}$. Replacing Eq. (16) in Eq. (15), we obtain:

$$\gamma^{(\infty)} = \sigma_z^2 + 1 - \frac{\left(\mathrm{erf}(\sqrt{\beta}\pi\omega)\right)^2}{4\pi\beta\omega^2}.$$

We are now interested in computing the asymptotic MSE, i.e., $\eta_{\mathbf{CRC}}(1/\gamma^{(\infty)})$ in Eq. (14). We first observe that $\eta_{\mathbf{CRC}}(1/\gamma^{(\infty)}) = \eta_{\mathbf{DR}}(1/\gamma^{(\infty)})$ where $\mathbf{D} = \mathbf{C}^2$. This is due to the properties of the matrix trace appearing in the definition of the tracial state $\phi(\cdot)$. Indeed,

$$\begin{aligned}
\mathsf{Tr}\left\{\left(\mathbf{I}+\frac{\mathbf{CRC}}{\gamma}\right)^{-1}\right\} &= \mathsf{Tr}\left\{\mathbf{C}^{-1}\left(\mathbf{C}^{-2}+\frac{\mathbf{R}}{\gamma}\right)^{-1}\mathbf{C}^{-1}\right\} \\
&= \mathsf{Tr}\left\{\mathbf{C}^{-2}\left(\mathbf{C}^{-2}+\frac{\mathbf{R}}{\gamma}\right)^{-1}\right\} \\
&= \mathsf{Tr}\left\{\left(\mathbf{I}+\frac{\mathbf{DR}}{\gamma}\right)^{-1}\right\}. \quad (17)
\end{aligned}$$

By assuming that $\mathbf{D}$ and $\mathbf{R}$ are *asymptotically free* [22], we can use the property in [22, Theorem 2.68, p.86], which relates the η-transform of a matrix product to the η-transform of each single matrix. In our case, we can write

$$\eta_{\mathbf{DR}}^{-1}(\zeta) = \eta_{\mathbf{D}}^{-1}(\zeta)\eta_{\mathbf{R}}^{-1}(\zeta)\frac{\zeta}{1-\zeta}. \quad (18)$$

As detailed below, the term $\eta_{\mathbf{R}}^{-1}(\zeta)$ depends only on the measurement error and on the spatial distribution of the sensors while the term $\eta_{\mathbf{D}}^{-1}(\zeta)$ accounts for both the measurement and position errors.

A simple expression for $\eta_{\mathbf{D}}(\cdot)$ can be obtained by exploiting Eq. (16), as follows:

$$\begin{aligned}
\eta_{\mathbf{D}}\left(\frac{1}{\sigma_z^2}\right) &= \phi\left(\left(\mathbf{I}+\frac{\mathbf{D}}{\sigma_z^2}\right)^{-1}\right) \\
&= \sum_{h=0}^{\infty}(-\sigma_z^2)^{-h}\phi(\mathbf{D}^h) \\
&= \sum_{h=0}^{\infty}(-\sigma_z^2)^{-h}\left[\int_{-\frac{1}{2}}^{\frac{1}{2}}\mathrm{e}^{-4h\pi^2\omega^2\beta w^2}\,\mathrm{d}w\right]^2
\end{aligned}$$

By switching the integral and sum operators we then get

$$\eta_{\mathbf{D}}\left(\frac{1}{\sigma_z^2}\right) = \int_{[-\frac{1}{2},\frac{1}{2}]^2} \sum_{h=0}^{\infty} \frac{\mathrm{e}^{-4h\pi^2\omega^2\beta(w_1^2+w_2^2)}}{(-\sigma_z^2)^h}\,\mathrm{d}w_1\,\mathrm{d}w_2 = \int_{[-\frac{1}{2},\frac{1}{2}]^2} \frac{1}{1+\frac{1}{\sigma_z^2}\mathrm{e}^{-4\pi^2\omega^2\beta\|\mathbf{w}\|^2}}\,\mathrm{d}\mathbf{w} \quad (19)$$

where $\mathbf{w} = [w_1, w_2]^{\mathsf{T}}$ and we recall that $\omega^2 = m\sigma_\delta^2$. Then, as it can be seen from the definition of the η-transform given in Eq. (2), the function $\eta_{\mathbf{R}}$ depends on the distribution of the random matrix $\mathbf{R}$. In turns, $\mathbf{R} = \mathbf{V}_{\widehat{\mathbf{P}}}\mathbf{V}_{\widehat{\mathbf{P}}}^{\mathsf{H}}$ is a Hermitian matrix whose distribution depends both on the distribution of $\mathbf{V}_{\widehat{\mathbf{P}}}$ and on its aspect ratio β. The entries of $\mathbf{V}_{\widehat{\mathbf{P}}}$ are driven by the distribution of the estimates $\widehat{\mathbf{P}}$, denoted by $f_{\hat{\mathbf{p}}}(\mathbf{x})$. In conclusion, $\eta_{\mathbf{R}}$ depends on the distribution $f_{\hat{\mathbf{p}}}(\mathbf{x})$, on the aspect ratio β, and on the parameter σ_z^2. In the following, it will be denoted by $\eta_{\mathbf{R}}\left(\beta, \frac{1}{\sigma_z^2}, f_{\hat{\mathbf{p}}}\right)$. Such function can be computed numerically by using the result in [23, Corollary 4.2]. This result links $\eta_{\mathbf{R}}\left(\beta, \frac{1}{\sigma_z^2}, f_{\hat{\mathbf{p}}}\right)$ to the η-transform $\eta_{\mathbf{R}}\left(\beta, \frac{1}{\sigma_z^2}, f_u\right)$, computed in the case where the distribution of the estimates $\widehat{\mathbf{P}}$ is uniform over the entire sampling area. Specifically, we have

$$\eta_{\mathbf{R}}\left(\beta, \frac{1}{\sigma_z^2}, f_{\hat{\mathbf{p}}}\right) = 1 - |\mathcal{A}| + |\mathcal{A}| \int_0^{\infty} g(y)\eta_{\mathbf{R}}\left(\frac{\beta}{y}, \frac{y}{\sigma_z^2}, f_u\right)\,\mathrm{d}y \quad (20)$$

where $g(y)$ is the first derivative of the cumulative density function $G(y)$, defined as

$$G(y) = |\mathcal{A}|^{-1}\left|\{\mathbf{x} \in \mathcal{R}\,|f_{\hat{\mathbf{p}}}(\mathbf{x}) \leq y\}\right|.$$

In Eq. (20), $|\mathcal{A}|$ denotes the Lebesgue measure of the set $\mathcal{A}$, and $\mathcal{A} = \{\mathbf{x} \in \mathcal{R} | f_{\hat{\mathbf{p}}}(\mathbf{x}) > 0\}$. More simply, $g(y)$ represents the spatial density distribution of the sensing devices.

Lastly, using Eqs. (18), (19) and (20), we can compute the asymptotic MSE through Eq. (14). Note that the advantage of using Eq. (20) to obtain the reconstruction accuracy (i.e., the MSE_∞) is that function $\eta_{\mathbf{R}}\left(\beta, \frac{1}{\sigma_z^2}, f_u\right)$ can be computed numerically very easily. In the next section, we exemplify how $f_{\hat{\mathbf{p}}}(\mathbf{x})$ and $g(y)$ can be derived from experimental data.

5 Description and Characterization of the Opportunistic Sensing Scenario

We assume that sensing devices can be either located onboard vehicles or embedded into handheld appliance, and we assess the validity of our technique to estimate the quality mobile opportunistic sensing in a real-world scenario. We focus on the region of Cologne, Germany, and employ information on the daily dynamics of the local population to infer realistic distributions of the mobile devices participating in the distributed mobile sensing.

Our methodology is detailed in the remainder of this section. We first present the datasets we leverage to characterize the sensing device mobility within the Cologne region, in Sec. 5.1. Then, in Sec. 5.2, we describe the process through which we obtain the devices spatial density distributions, $f_{\hat{\mathbf{p}}}(\mathbf{x})$ and $g(y)$, required for signal reconstruction.

5.1 Device mobility datasets

As mentioned, we consider both onboard-vehicle sensing devices and sensors embedded in, e.g., smartphones. We are thus interested in both vehicular and pedestrian mobility dynamics, as they drive the spatio-temporal presence of sensing devices in the Cologne region.

We infer such dynamics mainly by leveraging results of the Travel and Activity PAtterns Simulation (TAPAS) methodology [24], which allows computing the movements of individuals in a large-scale population. To that end, TAPAS exploits information on (i) home locations and socio-demographic characteristics of the actual population whose mobility is to be modeled, (ii) land use in the target region, and (iii) the time use patterns, i.e., habits of the locals in organizing their daily schedule.

The TAPAS methodology was applied on real-world data collected by the German Federal Statistical Office, including 30,700 daily activity reports from more than 7,000 households in the Cologne region [25], [26]. The result is a faithful and detailed representation of the local population daily activities [27]. We exploit such data for the characterization of the vehicular and pedestrian mobility dynamics during a typical weekday. The two representations are discussed separately in the following.

On-vehicle devices. The movement of individual vehicles in the Cologne region is extracted from a synthetic dataset generated by blending different state-of-art tools. We provide a brief description of the dataset below, while more details are available in [28].

The vehicular mobility dataset combines three key components that specify (i) the road topology and infrastructure, (ii) the microscopic-level driver behavior, and (iii) the macroscopic-level traffic flows. The OpenStreetMap (OSM) database is queried for the *road network layout* and infrastructure information (including, e.g., per-street speed limits, lane capacity, and intersection signalization). The open-source OSM database is contributed by a vast user community leveraging satellite imagery and GPS logs as sources of reference, and it is commonly regarded as the highest-quality map database publicly available to date. The *microscopic mobility* of vehicles is simulated with the Simulation of Urban Mobility (SUMO) software. SUMO implements validated car-following and lane-changing models and faithfully reproduces drivers' behavior in presence of complex road structures and signalization. As a result, it is today the de-facto standard among open-source microscopic vehicular mobility generators. At the *macroscopic road traffic* level, vehicular flows in the Cologne region are computed by coupling a traffic demand model with a traffic assignment model. The former is used to determine the locations at which each vehicle starts and ends its trip: we inferred such information from the TAPAS dataset introduced before. The latter computes the exact path followed by each driver, and we implemented it via Gawron's relaxation technique [29]. Such a technique models the road topology as a graph and iterates over a weighted shortest path algorithm, re-assigning edge costs based on traffic congestion levels. Gawron's scheme is known to achieve a so-called dynamic

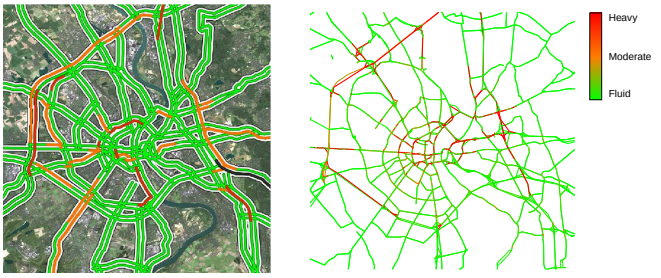


Fig. 1. Road traffic at 5 pm of a typical work day in the Cologne region, as recorded in the real-world by the ViaMichelin live traffic information service (left) and in the synthetic vehicular mobility trace we used (right).

user equilibrium after a sufficient number of iterations.

Overall, the dataset describes 24 hours of road traffic over an area of 400 km^2, and includes more than 700,000 car trips. The mix of tools employed to generate it allows for an unprecedented combination of scale and realism – as proven by the good match between the road traffic observed in the synthetic dataset and that provided by live traffic information services, in Fig. 1.

Handheld devices. We generated the mobility dynamics of handheld appliance, such as sensing-enabled smartphones and tablets, by merging different data sources.

First, we retrieved data from a recent demographic survey information on the population density and age distribution in each district (*Stadtteile*) of the Cologne region [30]. We then coupled such data with global statistics on the usage of smartphones for different age groups [31]. That way, we could estimate the number of smartphones owners in the different districts of the Cologne region.

Next, we leveraged again the TAPAS dataset, and extracted the non-vehicular (mainly, pedestrian) trips, which amount to around 800,000 individual source-destination descriptions. An analysis of such trips allowed us to determine the volume of non-vehicular movements between each pair of districts during 24 hours of a typical work day. By mapping the inter-district mobility flows to the aforementioned per-district smartphone user population, we could finally estimate the daily dynamics of smartphone presence in the whole Cologne region.

An intuitive representation of the above dynamics is displayed in Fig. 2. There, each district is assigned a color reflecting the smartphone user population variation during a 30-minute interval, expressed in users/km^2. Lighter colors indicate that users are leaving a district, i.e., that there is an out-flow of users from the district. Darker colors indicate that users are instead moving into the district, generating an in-flow of handheld devices. We can easily observe the realistic population dynamics obtained via our methodology. While no appreciable variations are found at night (4 am), the morning hours are characterized by significant flows from the outer regions towards the city center (e.g., around 8 am). Reverse flows mark instead the mid afternoon hours (e.g., starting from 4 pm). Detailed phenomena are reproduced as well, such as flows of users returning home for lunch (out-flow from the city center at 12 pm).

5.2 Device density distributions

Our study focuses on four different areas within the larger Cologne region, highlighted by the light grey squares in Fig. 3(a). Such areas cover 25 km^2 each, and were selected so as to consider environments of diverse nature. More precisely:

- *area A* maps to downtown Cologne, whose road layout is detailed in Fig. 3(b); since Cologne is a typical mid-sized European city of medieval origins, its center is a dense web of minor urban roads inlaid in a sparser network of arterial primary roads;
- *area B* represents a work/industrial area close to the city center, in Fig. 3(c); the area is crossed by highways and characterized by day-long intense car traffic over arterial roadways;
- *area C* consists of the suburban area in Fig. 3(d); the vast majority of road traffic passes by the highway junctions;
- *area D* is portrayed in Fig. 3(d), and represents a residential area in the outskirts of Cologne.

For each of such areas, we computed the time-varying densities, $f_{\hat{\mathbf{p}}}(\mathbf{x})$ and $g(y)$, of on-vehicle and handheld devices that are required by our model. To that end, we processed the datasets presented in Sec. 5.1, as follows.

On-vehicle devices. As far as on-vehicle sensing devices are concerned, we extracted from the vehicular mobility dataset information about the density $f_{\hat{\mathbf{p}}}(\mathbf{x})$ of cars, in each region and at several times of the day. It is to be noted that we performed such a process separately on three different road categories:

- *Highway* roads include high-capacity highways and motorways, as well as high-speed bypass and orbital roads;
- *Primary* roads are major traffic arteries that cover the whole urban region and link it to the suburban areas;
- *Urban* roads represent the finer portion of the road network mesh, interconnecting primary roads and granting access to every location of interest in the region.

The rationale is that such heterogeneous road categories are characterized by very dissimilar road traffic intensities. Aggregating them would thus cause loss of information about the actual density of on-vehicle sensing devices in the area. Considering them separately allows instead for a more reliable description of the on-vehicle sensing devices. Colors and line widths in the maps of Fig. 3 outline the road classification in each target area.

For every combination of area and daytime, we measured the geographic car density (expressed in vehicles/km^2), as a function of the road category. Clearly, the density also depends on the fraction of vehicles equipped with sensing devices and participating in the system. Thus, we also considered different participation ratios r, i.e., the fraction of mobile devices present in the area that take part to the opportunistic sensing of the atmospheric pollution.

Examples of the geographic on-vehicle sensing device density $f_{\hat{\mathbf{p}}}(\mathbf{x})$, observed over primary and urban roads at 8 am in area A, are shown in Fig. 4, for a participation ratio $r = 1$.

The spatial densities $f_{\hat{\mathbf{p}}}(\mathbf{x})$ of on-vehicle sensing devices were leveraged to derive the experimental distributions of the same. We then employed the nonlinear least-squares (NLLS)

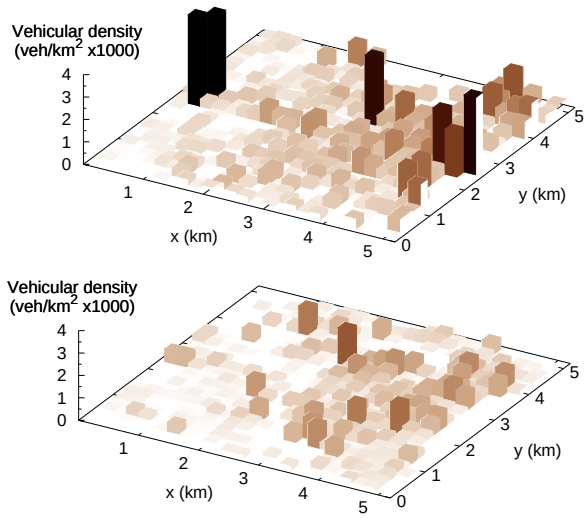


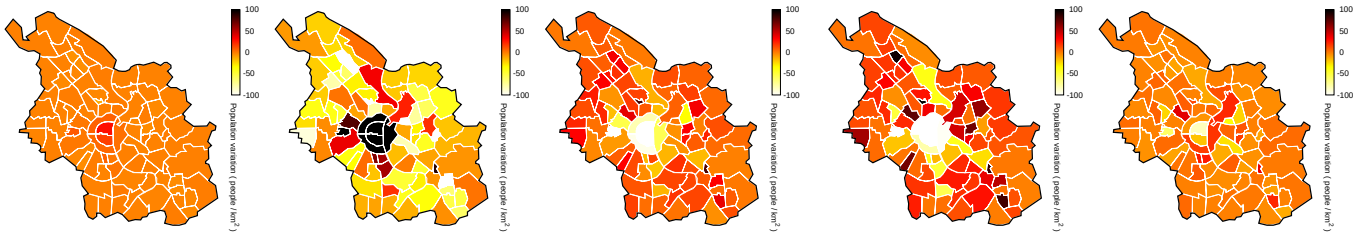
Fig. 4. On-vehicle sensing device density ($f_{\hat{\mathbf{p}}}(\mathbf{x})$) in area A (downtown Cologne) at 5 pm, over primary (top) and urban (bottom) roads. The participation ratio is r = 1.

Marquardt-Levenberg algorithm to fit a set of candidate theoretical distributions onto the experimental ones. This allowed us to finally retrieve analytical expressions for the device density $g(y)$ in Eq. (20), as required by the signal reconstruction methodology presented in Sec. 3.

A representative sample of the fitting process is shown in Fig. 5. The plots present fittings of the candidate theoretical distributions $g(y)$ to the experimental complementary cumulative distribution functions (CCDF) of the vehicular densities, previously shown in Fig. 4. The top plot evidences the exponential tail of the experimental distribution, appearing linear in a linear-logarithmic plot. Therefore, the data is best represented by exponentially tailed distributions, whereas heavy-tailed distributions provide a poor fit. However, the bottom linear-linear plot shows how the probability mass next to the origin does not follow an exponential law. As a result, these experimental distributions are best fitted by the Exponentially Modified Gaussian (EMG) distribution that characterizes the sum of two independent normal and exponential random variables. The EMG distribution has indeed an exponential tail, but is a tunable normal distribution around the origin.

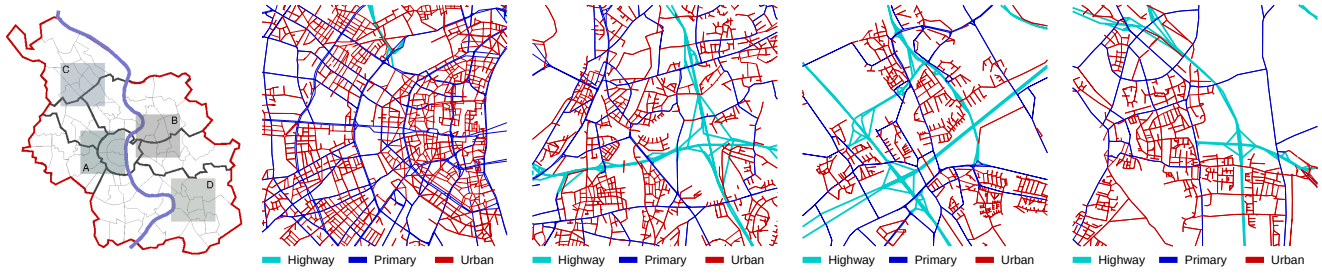
A more complete summary of the fitting results is provided in Tab. 1 in the Supplemental Material, for all combinations of area, daytime and road category, under varying participation ratios, r. The table allows comparing the quality of fittings obtained through different candidate theoretical distributions, in terms of the residual sum of squares (RSS) with respect to the experimental data.

Interestingly, the aforementioned EMG distribution provides a best fit in most situations. When it does not, it yields a negligible RSS distance from the best fit. This allows us to model, for on-vehicle sensing devices, the generic analytical expression of $g(y)$ in Eq. (20) as a set of EMG distributions (one per road category), each to be weighted by the corresponding road type surface.



(a) 4 am (b) 8 am (c) 12 am (d) 4 pm (e) 8 pm

Fig. 2. Variation of handheld device population in each district (*Stadtteile*) of the Cologne region during a typical working day, measured in users/km^2. Darker colors indicate stronger in-flows of users, while lighter colors indicate stronger out-flows of users. This figure is best viewed in colors.



(a) Cologne districts (b) Area A (c) Area B (d) Area C (e) Area D

Fig. 3. Geographical areas considered in our study. (a) Cologne region districts, with the surfaces of the four target areas highlighted in light grey. (b) Area A: city downtown. (c) Area B: industrial/transit. (d) Area C: suburban highways. (e) Area D: residential outskirts. This figure is best viewed in colors; in the Areas plots light blue, blue and red denote highway, primary and urban roads, respectively.

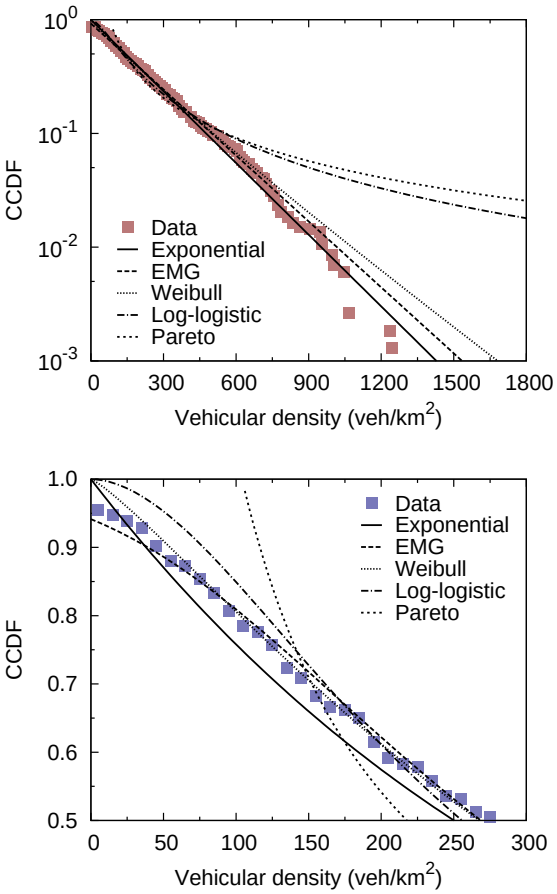


Fig. 5. Nonlinear least-squares fittings of theoretical probability distributions ($g(y)$) on two sample experimental distributions. The latter are derived from the vehicular densities ($f_{\hat{\mathbf{p}}}(\mathbf{x})$) in Fig. 4, for urban (top) and primary (bottom) roads.

Handheld devices. Following a similar procedure, the density distribution $g(y)$ of handheld devices participating in the sensing process was inferred directly from the smartphone user population dynamics presented in Sec. 5.1. In this case, however, the spatial granularity of the data about the device density is at the district level, i.e., too coarse for an analytical distribution fitting. Therefore, we simply collected the information about the time-varying user density in the districts that (partially) fall within each of the four geographical areas, and assumed a uniform distribution $f_{\hat{\mathbf{p}}}(\mathbf{x})$ of smartphone users in each of such districts. We then modeled $g(y)$ in Eq. (20) for handheld devices as a set of Dirac delta functions at the density values recorded in the districts within the target area. As it happened for the on-vehicle device density distribution, the probability mass of each Dirac delta is also weighted. This time, weights are assigned according to the area surface ratio occupied by the district corresponding to the delta function.

6 Evaluation

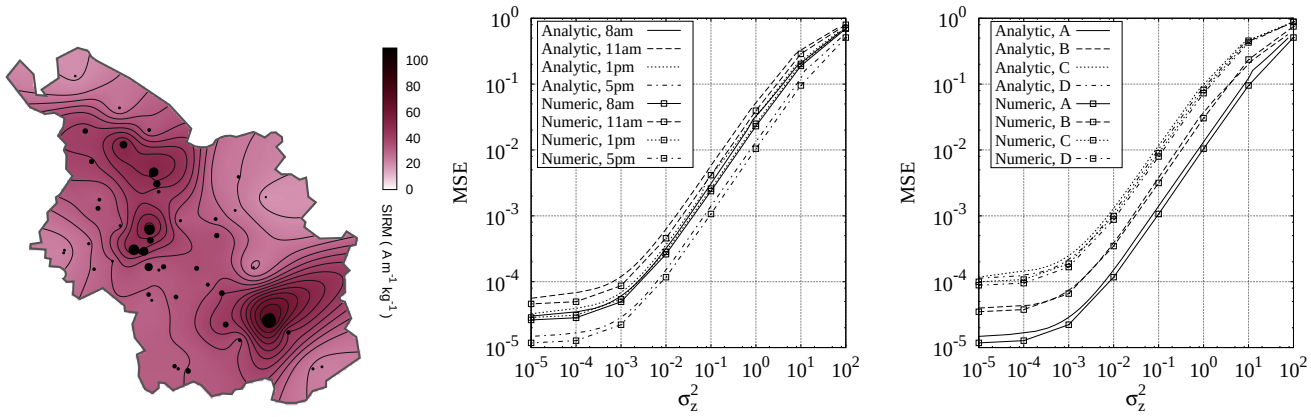
The application use case of mobile opportunistic sensing we consider for evaluation is that of atmospheric pollution monitoring. Pollution thus represents the phenomenon that is sensed by mobile devices, or, equivalently, the original signal whose reconstruction accuracy we want to assess through our proposed technique. To this end, accurately modeling pollution in the Cologne region is paramount to the reliability of our study. Unfortunately, traditional fixed stations for the measurement of atmospheric pollutants are expensive to deploy and maintain, and their number is typically limited to a few units per city. As a result, such data does not allow building the fine-grained map we need for our analysis.

However, techniques have been proposed that enable gathering high-resolution pollution information through biomonitoring of natural vegetation. We thus retrieved data obtained via magnetic analysis of pine needles within the Cologne conurbation [32], and use it to build a more precise model of the average long-term presence of atmospheric pollution in the region. Among the measures available from that study, we employed the Saturation Isothermal Remanent Magnetization (SIRM), which has been shown to be an excellent proxy for biomonitoring of combustion pollutants. The signal (i.e., pollution) map in the Cologne region, resulting from SIRM data collected at 63 locations, is portrayed in Fig. 6(a). There, dots represent the measurement locations, whereas colors and isolines identify different levels of SIRM presence.

In order to generate samples at each device, we therefore link the dataset above with those describing the spatial distribution of sensing devices in the area (see Sec. 5.1). Specifically, from the mobility data set we obtain the estimated device positions, $\hat{\mathbf{p}}$. Then, we remove the position error, i.e., the instances of a zero-mean Gaussian distributed random variable with variance σ_δ^2, which allows us to retrieve the true device locations. We then associate the phenomenon samples to each device depending on its true position.

Next, we proceed to the validation and exploitation of our proposed approach, presented in Sec. 3, in the opportunistic sensing and application use case scenarios detailed in Sec. 5.1 and above. Specifically, we compute the asymptotic MSE obtained by evaluating Eq. (18), where $\eta_{\mathbf{D}}$ is given by Eq. (19) and $\eta_{\mathbf{R}}$ is computed by using Eq. (20). The spatial density distribution $g(y)$ of mobile (on-vehicle and handheld) devices that appears in Eq. (20) is obtained through distribution fittings, as explained in Sec. 5.2. Ultimately, our model allows obtaining a measure of the MSE of the atmospheric pollution in the Cologne region, as estimated from samples collected by devices in the area.

Concerning the system parametrization, unless otherwise specified, we set $\sigma_\delta^2 = 25\,\text{m}$, $r = 1$, $\sigma_z^2 = 0.01$, $n = 13$, and a handheld fraction of 0.8. The latter is the fraction of mobile devices participating in the sensing process that is handheld, as opposed to that of on-vehicle devices (whose default fraction is thus 0.2). Also, we denote by ρ the spatial density of mobile devices participating to opportunistic sensing. As an example, when $r = 1$ and and the handheld fraction is 0.8, in the data sets corresponding to 5 pm we have $\rho = 667$, $\rho = 230$, $\rho = 85$, and $\rho = 98$ samples/km^2, in areas, A, B, C, and D, respectively. Finally, note that, in our settings we assume the same value for the variance of the position error at vehicular and handheld devices. Indeed, on-board commercial GPS receivers are typically combined with an error correction system that



(a) Pollution heatmap (b) Validation, daytime (c) Validation, regions

Fig. 6. Validation. (a) Heatmap of SIRM in the Cologne region. (b,c) Analytic and numeric MSE versus the variance of the measurement error σ_z^2, computed at different times of the day and over different geographical regions, respectively.

leverages other information provided by the vehicle itself (e.g., vehicle speed). Thus, in spite of the higher speed of vehicles, on-board GPS devices provide performance similar to that of GPS receivers in handheld devices.

6.1 Analytical framework validation

In order to validate our approach, we compare the MSE indicated by our analytic model against that computed via a numerical approach, in presence of multiple system settings. The numerical MSE is obtained by computing Eq. (12), i.e., by averaging over many instances of random variables distributed as the density of sensing devices.

Fig. 6(b) depicts the dynamics of the MSE versus σ_z^2, i.e., the variance of the measurement error. In other words, σ_z^2 represents the quality of the data collected by mobile devices. Values of σ_z^2 larger than 1 indicate that mobile devices collect low-quality, error-prone records that are poorly representative of the actual atmospheric pollution in their proximity. Conversely, values of σ_z^2 below 0.1 indicate that dependable measures of the phenomenon are gathered by sensors embedded in the mobile devices. It is thus natural that all the curves in Fig. 6(b) have a monotonic, decreasing trend with respect to decrementing values of σ_z^2: the accuracy of the pollution map reconstructed at the data fusion center cannot but improve (and thus its MSE decrease) as the mobile devices provide more reliable samples. Notably, values of σ_z^2 below 1 are already sufficient to reduce the MSE below 0.1 in all cases, and σ_z^2's below 0.1 guarantee a MSE below 0.01.

What is especially interesting for us is the comparison of the asymptotic MSE determined by our model (denoted by *analytic* in the plot) with the MSE computed through the *numerical* approach. We can observe that there is a very good match between the curves referring to the two methodologies, for any value of σ_z^2. Interestingly, the match is consistent when considering different hours of the day, which are characterized by a diverse presence of on-vehicle and handheld devices, i.e., values of ρ.

The results in Fig. 6(b) refer to the case of the geographical area denoted as A in Sec. 5.2, and portrayed in Fig. 3(b). In fact, focusing on other areas of the Cologne conurbation does not vary the outcome. Fig. 6(c) shows that the match between the asymptotic MSE and that computed by the numerical approach remains good throughout geographical areas with distinctive and heterogeneous road layouts, such as those depicted in Fig. 3. The diversity of such areas emerges when observing the quality of the pollutant presence estimation, in terms of absolute MSE, for a same value of σ_z^2. Indeed, the reconstructed information is significantly less accurate in scarcely populated areas (low ρ) crossed by a limited number of roads, such as areas C and D, than in crowded, highly trafficked areas (high ρ) such as A.

Overall, no matter the topological features of the geographical area considered, nor the time at which the analysis is performed, we remark that our model always provides a reliable indication of the MSE of the atmospheric pollution estimated from the samples collected by mobile devices. We conclude that the proposed model can be safely employed for the characterization of the phenomenon reconstruction accuracy in the realistic opportunistic sensing and application use case scenarios we consider.

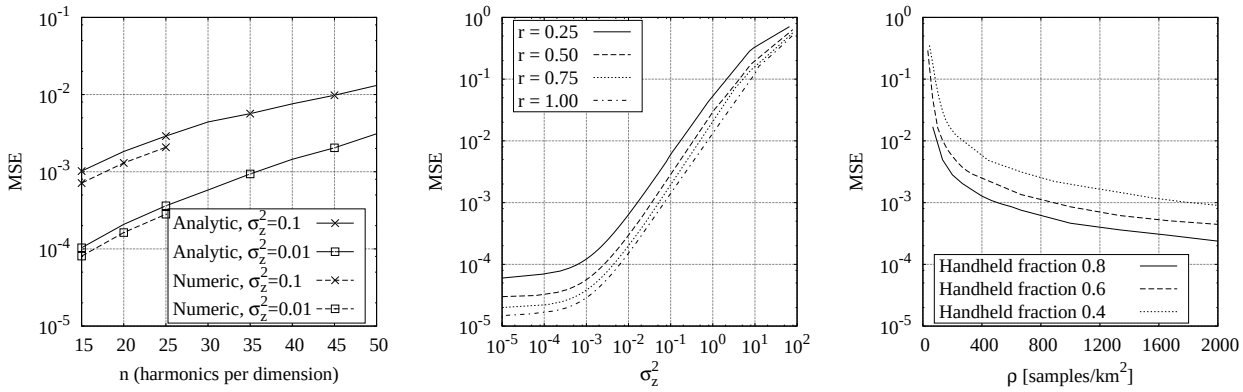
The model becomes crucial to better investigate the performance, in terms of signal reconstruction accuracy, of the mobile opportunistic sensing process. Indeed, as also highlighted by the following results, the computational cost of the numerical approach grows rapidly with increasing values of the system parameters, and soon becomes unmanageable. Instead, the model allows exploring the full parameter space, as done in the remainder of the section.

Takeaways. *Our model provides an excellent approximation of the accuracy achieved by a mobile opportunistic sensing system. Its scalability with respect to a wide range of parameters makes it suitable to comprehensive performance evaluations of such systems.*

6.2 Accuracy of opportunistic sensing

In our performance evaluation, we focus on the densely populated area A at 5 pm, representing an ideal scenario for a participatory approach, as outlined by the previous results.

First, we study the impact of the desired quality of the reconstructed atmospheric pollution signal at the data fusion



(a) Quality of samples and signal representation (b) Participation ratio (c) Device density and type

Fig. 7. Exploitation. (a) Impact on the MSE of the measurement error (σ_z^2) and of the number of harmonics of the reconstructed signal, (n). (b) Impact of the participation ratio on the MSE of the pollution estimate. (c) Impact of different densities ρ and of on-vehicle/handheld ratios for $\sigma_z^2 = 0.01$.

center. Fig. 7(a) shows the MSE as a function of the number of harmonics per dimension n of the final pollution map estimated from the collected samples. As explained in Sec. 3, n is a measure of the precision with which we try to reconstruct the original phenomenon, and higher values lead to a more detailed representation. Therefore, the MSE tends to increase with n. However, the good news is that the growth is not particularly rapid, i.e., mobile sensing can support a high-detail estimation without reducing too much the accuracy of the result. Moreover, disposing of higher quality samples allows obtaining estimates that are both very accurate and precise, as observed when comparing the curves for different values of σ_z^2. For the sake of completeness, Fig. 7(a) also includes equivalent curves obtained with the numerical approach. We stress how (i) the numerical curves are again very close to the analytic ones obtained with our proposed model, which further proves the quality of the latter, and (ii) the numerical curves are interrupted at $n = 25$ harmonics per dimension due to their computational cost, which demonstrates how our model can be leveraged to explore portions of the parameter space that cannot be studied otherwise.

The impact of the participation ratio, r, is presented in Fig. 7(b). Different curves denote participation ratios of 0.25, 0.5, 0.75, and 1, respectively, and are plotted versus the sample quality represented by the variance of the measurement error σ_z^2. As one could expect, an increased participation of mobile users results in a lower MSE, i.e., a higher accuracy of the estimated pollution map. Interestingly, the difference among the curves remains constant in a logarithmic scale. This implies that, for any value of σ_z^2, a participation ratio $r = 1$ can compensate for a difference of sample quality of around one order of magnitude with respect to a participation ratio $r = 0.25$. The same result also leads to the consideration that the impact of a higher participation ratio is much more important when the quality of samples is low. As an example, the MSE drops from 0.40 to 0.10 for $\sigma_z^2 = 10$ when all mobile devices take part in the sensing process with respect to the case where one every four does so. The decrement between the same two ρ scenarios is instead of just 0.04 for $\sigma_z^2 = 1$.

We further delve into the analysis, by assessing the impact of the density and type of participating devices for $\sigma_z^2 = 0.01$, in Fig. 7(c). There, the abscissa denotes the total density of devices, indicated as ρ and measured in samples collected per square kilometer. We recall that in area A at 5 pm, with a handheld fraction of 0.8, the density corresponding to a participation ratio $r = 1$ is $\rho = 667$ samples/km^2, thus higher values of ρ correspond to future scenarios where the pervasiveness of sensing-enabled devices will be even larger. Different curves represent instead diverse handheld fractions.

The main observation here is the super-exponential decay of the MSE with ρ, which indicates that the total density of sampling devices is the key factor towards an accurate sensing process. In particular, ρ is especially critical at low densities, where a difference of a few tens of devices per km^2 can result in a MSE reduction of two orders of magnitude. The effect is instead attenuated once ρ grows beyond a few hundreds of devices per km^2. Concerning the type of mobile devices involved in the process, we note that handhelds prove to be better samplers than on-vehicle ones. The reason is that vehicles are constrained to roads in their movement, and thus tend to collect data on the atmospheric pollution that always refer to the same portions of the area. Thus, increasing the presence on-vehicle devices does not bring a significant advantage, as it only leads to over-sampling at a limited number of locations. Instead, handheld devices can move more or less freely around the area, and thus provide a much better coverage of the original phenomenon. This translates into a decreased MSE when their presence grows.

Takeaways. *The accuracy of mobile sensing is mainly driven by participation. A minimum critical threshold (e.g., ∼100 samples/km^2 in our scenario) of uniformly distributed (e.g., handheld in our scenario) mobile devices that provide good-quality measurements (e.g., error variance below 1% in our scenario) is required to faithfully reconstruct the original phenomenon.*

6.3 Impact of position error

We also investigate the impact on the accuracy of the phenomenon estimation of the position and measurement errors affecting the collected samples. Results are shown in Fig. 8.

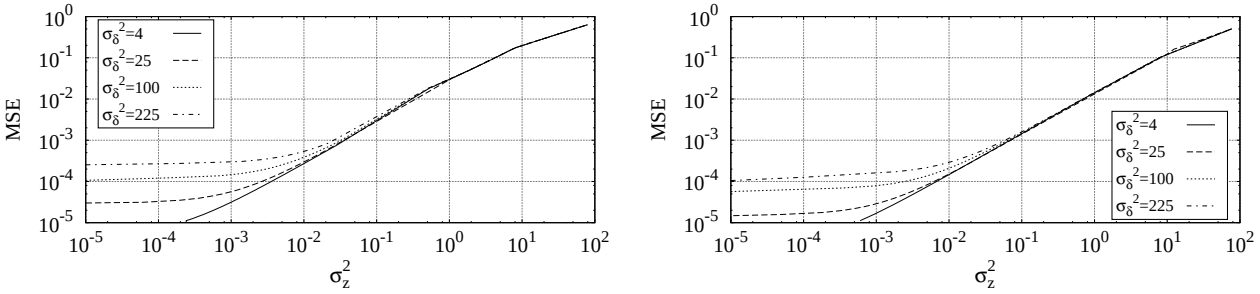


Fig. 8. Exploitation: MSE as a function of the measurement error variance, σ_z^2, for different values of the variance of the position error affecting the collected samples. The scenario under study is area A at 5 pm with a handheld fraction of 0.8, under $r = 0.5$ (left) and $r = 1$ (right).

The scenario under study is still area A at 5 pm with 80% handheld devices and $r = 0.5$ (left) and $r = 1$ (right). Note that values of position error variance, σ_δ^2, from 4 to 225 m^2 correspond to position errors ranging between 2 and 15 m, which are typical values for GPS receivers. Interestingly, the effect of the position error becomes noticeable only when the impact of measurement errors is marginal, i.e, for values of σ_z^2 smaller than 0.01 for $r = 0.5$ and 0.005 for $r = 1$. This suggests that the measurement noise drives the system performance and δ plays a role only when the measured signal samples are very accurate. We further observe that the position error contributes to determining the MSE floor, i.e., the asymptotic value that we obtain as the measurement error (or, equivalently, σ_z^2) tends to zero. The reason for this behavior is that the LMMSE filter used for the phenomenon reconstruction cannot be optimized with respect to δ (see Sec. 4.1 for details). Specifically, the MSE floor increases by almost one order of magnitude as the variance of the positioning error σ_δ^2 varies from 25 to 225 m^2. Furthermore, in the region where σ_z^2 dominates, the accuracy decreases as a power law function of the measurement error variance.

Takeaways. *Typical GPS position errors do not affect the accuracy of mobile opportunistic sensing in a significant way.*

6.4 Impact of device deployment

In order to complete our analysis, we compare the performance of the mobile sensing system to a traditional approach using fixed monitoring stations [8]. The latter are uniformly distributed over area A with two different densities, namely, 10 and 400 stations/km^2. The first value is representative of quite extensive real-world station deployment, while the second coincides with the considered density of the mobile sensing devices. Results are shown in Fig. 9.

Let us first focus on the three curves that have been obtained for the same number of samples per km^2 (i.e., $\rho = 400$). Recall that in our scenario pedestrians are uniformly distributed in the non-road zones, thus a higher handheld fraction implies a larger number of uniformly distributed samples. We note that the spatial distribution of samples has a significant impact on the MSE: the more uniform the distribution, the better the performance. This suggests that a massive monitoring infrastructure would lead to a highly accurate reconstruction of the phenomenon of interest. However, mobile opportunistic sensing yields performance that is very close to that of a pervasive sensing infrastructure deployment, without the associated costs.

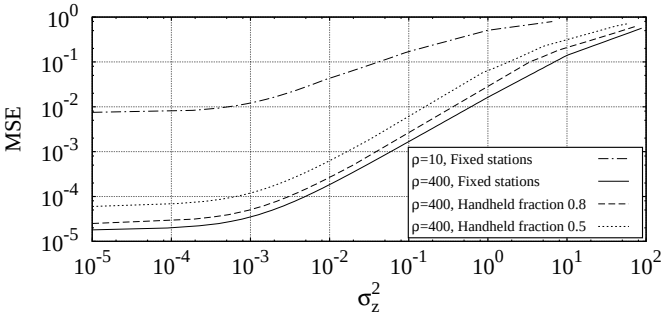


Fig. 9. Exploitation: Impact of the spatial deployment of sensing devices over the geographical area (area A). Different densities (10 and 400 stations/km^2) of fixed monitoring stations are compared to a mobile sensing process ($\rho = 400$ samples/km^2) with 0.5 and 0.8 handheld fraction.

Also, assuming large but more realistic values of fixed stations density (e.g., 10 stations/km^2), we clearly see that the reconstruction accuracy achieved by the fixed infrastructure is severely reduced. It follows that mobile opportunistic sensing can represent an excellent alternative to monitoring infrastructures that are expensive to deploy and maintain.

Takeaways. *Mobile opportunistic sensing has the potential to provide phenomena representations that are much more accurate (e.g., at least two orders of magnitude smaller MSE in our scenario) than those achievable by quite extensive (e.g., 10 stations/km^2 in our scenario) sensing infrastructures, at a much lower cost.*

7 Conclusions

We addressed the problem of evaluating the accuracy of mobile opportunistic sensing in urban environments. By using a signal processing approach, we developed an analytical framework that describes the relationship between the accuracy of the phenomenon reconstruction and the mobile sensing parameters. Our framework assumes that the well-known LMMSE filter is used for signal reconstruction, and accounts for major factors

such as position and measurement errors affecting the collected samples, as well as the density and spatial distribution of sensing devices. We validated our approach through numerical results in a realistic scenario where both on-vehicle and handheld mobile devices participate in the sensing process. We then exploited the analytical expressions we derived to investigate the impact of the mobile sensing parameters on the accuracy of air pollution sensing. Our results highlight that the noise level affecting the measurements collected by the users is more critical than the sheer number of users, and that pedestrian users are paramount to the quality of the urban sensing process. Position errors instead play a role only in presence of very accurate measurement of the sensed phenomenon. Finally, given the type of phenomenon under study, the number of samples to be collected can be modulated according to the required level of accuracy.

ACKNOWLEDGMENTS

We would like to thank Dr. Eva Lehndorff of Bonn University for providing us with data of the average long-term atmospheric pollution in the Cologne area. This work has been partially supported by the LIMPID project (POR FESR 2007/2013) funded by Regione Piemonte (Italy), and by funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement n.630211 ReFleX.

REFERENCES

[1] N.D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, A.T. Campbell, "A Survey of Mobile Phone Sensing," *IEEE Comm. Mag.,* Vol. 48, No. 9, pp. 140-150, 2010.

[2] R.K. Ganti, Y. Fan Ye, L. Hui, "Mobile Crowdsensing: Current State and Future Challenges," *IEEE Comm. Mag.,* Vol. 49, No. 11, pp. 32-39, 2011.

[3] Copenhagen Wheel, `http://senseable.mit.edu/copenhagenwheel`.

[4] R.N. Charette, "This Car Runs on Code," *IEEE Spectrum*, 2009.

[5] Telefónica, "Connected Car Industry Report," *White Paper*, 2013.

[6] Telematics Update, "The Connected Car in Europe: Gaining Market Share with Telematics," *White Paper*, 2013.

[7] D. Stranneby, W. Walker *Digital Signal Processing and Applications Elsevier,* Elsevier, 2nd Ed., 2004.

[8] A. Tilloy, V. Mallet, D. Poulet, C. Pesin, F. Brocheton, "BLUE-based NO2 Data Assimilation at Urban Scale," *Journal of Geophysical Research,* Vol. 118, No. 4, 2013.

[9] A. Nordio, C.-F. Chiasserini, E. Viterbo, "Performance of Linear Field Reconstruction Techniques with Noise and Uncertain Sensor Locations," *IEEE Trans. on Signal Proc.,* Vol. 56, No. 8, pp. 3535–3547, 2008.

[10] G. Cardone, L. Foschini, P. Bellavista, A. Corradi, C. Borcea, M. Talasila, R. Curtmola, "Fostering ParticipAction in Smart Cities: A Geo-Social Crowdsensing Platform," *IEEE Communications Magazine*, Vol. 51, No. 6, pp. 112–119, 2013.

[11] G. Cardone, A. Cirri, A. Corradi, L. Foschini, "The ParticipAct Mobile Crowd Sensing Living Lab: The Testbed for Smart Cities," *IEEE Communications Magazine*, Vol. 52, No. 10, pp. 78–85, 2014.

[12] E. Miluzzo, M. Papandrea, N.D. Lane, H. Lu, A.T. Campbell, "Pocket, Bag, Hand, etc. - Automatically Detecting Phone Context through Discovery," *First International Workshop on Sensing for App Phones (PhoneSense10)*, 2010.

[13] W. Zhang, B. Zhu, L. Zhang, J. Yuan, I. You, "Exploring Urban Dynamics based on Pervasive Sensing: Correlation Analysis of Traffic Density and Air Quality," *IEEE IMIS*, 2012.

[14] Y. Zhu, Z. Li, H. Zhu, M. Li, Q. Zhang, "A Compressive Sensing Approach to Urban Traffic Estimation with Probe Vehicles," *IEEE Transactions on Mobile Computing*, Vol. 12, No. 11, pp. 2289–2302, 2013.

[15] R. Du, C. Chen, B. Yang, X. Guan, "Vanet Based Traffic Estimation: A Matrix Completion Approach," *IEEE Globecom*, 2013.

[16] R. Du, C. Chen, B. Yang, N. Lu, X. Guan, X. Shen, "Effective Urban Traffic Monitoring by Vehicular Sensor Networks," *IEEE Transactions on Vehicular Technology*, Vol. 64, No. 1, pp. 273–286, 2015.

[17] D. Zhao, H. Ma, L. Liu, J. Zhao, "On Opportunistic Coverage for Urban Sensing," *IEEE MASS*, 2013.

[18] Y. Zheng, F. Liu, H.-P. Hsieh, "U-Air: When Urban Air Quality Inference Meets Big Data," *ACM SIGKDD*, 2013.

[19] D. Zhang, F. Zhang, J. Huang, C. Xu, Y. Li, T. He, "Exploring Human Mobility with Multi-Source Data at Extremely Large Metropolitan Scales," *AMC MobiCom*, 2014.

[20] K. Rachuri, C. Efstatiou, I. Leontiadis, C. Mascolo, P. Rentfrow, "METIS: Exploring Mobile Phone Sensing Offloading for Efficiently Supporting Social Sensing Applications," *IEEE PerCom,* 2013.

[21] A. Nordio, C.-F. Chiasserini, E. Viterbo, "Reconstruction of Multidimensional Signals From Irregular Noisy Samples" *IEEE Trans. on Signal Proc.,* Vol. 56, No. 9, pp. 4274–4285, September 2008.

[22] A. Tulino and S. Verdú, *Random Matrix Theory and Wireless Communication,* Now Publishers, 2004.

[23] A. Nordio, C.-F. Chiasserini, "Field Reconstruction in Sensor Networks with Coverage Holes and Packet Losses," *IEEE Transactions on Signal Processing,* Vol. 59, No. 8, pp. 3943-3953, August 2011.

[24] C. Varschen, P. Wagner, "Mikroskopische Modellierung der Personenverkehrsnachfrage auf Basis von Zeitverwendungstagebüchern", *Stadt Region Land*, Vol. 81, pp. 63–69, 2006.

[25] G. Rindsfüser, J. Ansorge, H. Mühlhans, "Aktivitätenvorhaben", in K.J. Beckmann (editor), *SimVV Mobilität verstehen und lenken – zu einer integrierten quantitativen Gesamtsicht und Mikrosimulation von Verkehr*, Final report, Ministry of School, Science and Research of Nordrhein-Westfalen, Düsseldorf, Germany, 2002.

[26] M. Ehling, W. Bihler, "Zeit im Blickfeld. Ergebnisse einer repräsentativen Zeitbudgeterhebung", in K. Blanke, M. Ehling, N. Schwarz (editors), *Schriftenreihe des Bundesministeriums für Familie, Senioren, Frauen und Jugend*, Vol. 121, pp. 237–274, Kohlhammer, Stuttgart, Germany, 1996.

[27] G. Hertkorn, P. Wagner, "The Application of Microscopic Activity Based Travel Demand Modelling in Large Scale Simulations", *World Conference on Transport Research*, Istanbul, Turkey, Jul. 2004.

[28] S. Uppoor, O. Trullols-Cruces, M. Fiore, J.M. Barcelo-Ordinas, "Generation and Analysis of a Large-scale Urban Vehicular Mobility Dataset," *IEEE Transactions on Mobile Computing*, Vol. 13, No. 5, 2014.

[29] C. Gawron, "An Iterative Algorithm to Determine the Dynamic User Equilibrium in a Traffic Simulation Model", *International Journal of Modern Physics C*.

[30] Stadt Köln, "Die Kölner Stadtteile in Zahlen", 2010.

[31] GoGulf, "Smartphone Users Statistics and Facts", 2012. `http://www.go-gulf.com/blog/smartphone`.

[32] M. Urbat, E. Lehndorff, L. Schwark, "Biomonitoring of Air Quality in the Cologne Conurbation Using Pine Needles as a Passive SamplerPart I: Magnetic Properties", *Atmospheric Environment*, Vol. 38, pp. 3781–3792, 2004.

Marco Fiore (S'05, M'09) is a researcher at CNR–IEIIT, Italy. He is a co-founder and collaborator of the Inria UrbaNet team hosted by the CITI Lab, France. He received a PhD degree from Politecnico di Torino, Italy, and an Habilitation à Diriger des Recherches (HDR) from INSA Lyon, France, as well as MSc degrees from University of Illinois at Chicago and Politecnico di Torino. He held positions as Associate Professor at INSA Lyon, and visiting researcher at Rice University, TX, USA and Universitat Politecnica de Catalunya, Spain. His research interests are in the fields of vehicular networking and mobile traffic analysis.

Alessandro Nordio (S'00, M'03) is a researcher at the CNR-IEIIT. In 2002 he received the Ph.D. from "Ecole Polytechnique Federale de Lausanne", Switzerland. From 1999 to 2002, he performed active research with Eurecom Institute, Sophia Antipolis (France). From 2002 to 2009 he was a post-doc researcher with the Electronic Department of Politecnico di Torino, Italy. His research interests are in the field of signal processing, space-time coding, wireless sensor networks and theory of random matrices.

Carla-Fabiana Chiasserini (M'98, SM'09) received her Ph.D. in 2000 from Politecnico di Torino, where she is currently an Associate Professor. She is also an Associate Researcher at CNR-IEIIT. Her research interests include protocols and performance analysis of wireless networks. Dr. Chiasserini has published over 250 papers at major venues.