Estimating Spatial Averages of Environmental Parameters based on Mobile Crowdsensing

Ioannis Koukoutsidis
Hellenic Telecommunications & Post Commission, 60 Kifissias Avenue, 151 25 Maroussi, Greece

Keywords: Mobile Crowdsensing, Spatial Average, Stratified Sampling.

Abstract: Mobile crowdsensing can facilitate environmental surveys by leveraging sensor-equipped mobile devices that carry out measurements covering a wide area in a short time, without bearing the costs of traditional field work. In this paper, we examine statistical methods to perform an accurate estimate of the mean value of an environmental parameter in a region, based on such measurements. The main focus is on estimates produced by considering the mobile device readings at a random instant in time. We compare stratified sampling with different stratification weights to sampling without stratification, as well as an appropriately modified version of systematic sampling. Our main result is that stratification with weights proportional to stratum areas can produce significantly smaller bias, and gets arbitrarily close to the true area average as the number of mobiles increases, for a moderate number of strata. The performance of the methods is evaluated for an application scenario where we estimate the mean area temperature in a linear region that exhibits the so-called *Urban Heat Island* effect, with mobile users moving in the region according to the Random Waypoint Model.

1 INTRODUCTION

Sensor-equipped mobile devices (e.g. smartphones or connected car devices) bring new possibilities for environmental surveys, as they enable data collection remotely through crowdsensing, without conducting traditional field work. Compared to the deployment of static sensor nodes, mobile crowdsensing is an attractive low-cost alternative for sensing of the environment; it takes advantage of the ubuiquitous presence of mobile users in practically all areas and can exploit the more advanced memory, processing and communication capabilities of mobile devices for conducting and transmitting complex measurements. In recent examples, specially-equipped mobile devices have been used to measure temperature, relative humidity, air-quality and other environmental parameters in large cities (van der Hoeven et al., 2014; Antonic et al., 2014).

Aggregating all measurements and producing an average value that correctly estimates the mean parameter value in an area[1] is a complex task. The researcher must decide for the number of mobile devices collecting measurements, the number of measurements, the method and the time at which they are taken, as well as the estimator formula. The complexity arises from the movement of the mobiles, and is increased by spatial autocorrelation (observations in nearby locations are more likely to be similar than observations further apart) and heterogeneity (observations vary systematically from place to place) of the measurement values.

If we assume that there are measurements of the mobile users that densely cover the whole area (so that, if we partition the area into a large number of subareas, the probability that a subarea is not measured approaches zero), then a good method to approach the true mean value is to split the area into a very large number of subareas, take the average of measurements in each subarea, and then average over the subareas. This will be later shown in Sect. 3. In cases where no dense set of measurements is available and a quick estimate is in order, it would be desirable to get the current readings of the mobile sensing devices, so as to have a "snapshot"-sensing of the measurement area, with a single measurement from each mobile. The caveat is that many mobility models effect a higher concentration of mobile devices near the center of the area, therefore taking a random sample of the devices will produce a biased estimate. Statistical methods to perform an accurate estimate in this setting are the main topic of this paper.

[1] Throughput the paper, we use the term *area* more broadly to refer to a region, and not strictly its size.

Koukoutsidis I.
Estimating Spatial Averages of Environmental Parameters based on Mobile Crowdsensing.
DOI: 10.5220/0006110900150026
In *Proceedings of the 6th International Conference on Sensor Networks (SENSORNETS 2017)*, pages 15-26
ISBN: 421065/17

Related lines of work concern the estimation of aggregate quantities in peer-to-peer (P2P) ((Massoulié et al., 2006; Kempe et al., 2003; Datta and Kargupta, 2007; Stutzbach et al., 2009; Kurant et al., 2011)) and sensor networks ((Considine et al., 2004; Ganesan et al., 2004; Bash et al., 2004; Shrivastava et al., 2004)). The research in P2P networks does not explicitly address the estimation of spatial statistics, although the techniques themselves could be used for such purpose (i.e. population values could be values in some locations recorded by the nodes). Some of the existing research in sensor networks is closer to our context and tries to deal with the bias that can be introduced in the computation of aggregates due to non-uniform device locations. However, all these techniques are conceived for fixed networks and cannot be directly imported in a mobile scenario, as node mobility is an important factor that complicates the situation. For example, uniform random sampling is harder to achieve when we introduce node mobility and opportunistic encounters between nodes. Furthermore, the location at which the measurements are taken must be taken into consideration when estimating spatial averages of environmental values.

We attempt to tackle the problem by using spatial sampling techniques. We examine the case where mobile devices move in an area according to a mobility model with a stationary location distribution, and take measurements of an environmental parameter at a random instant in time. The goal is to estimate the average of the environmental parameter in the area as accurately as possible. Measurements from all devices are assumed to be sent to a central processor which carries out the estimation. We compare stratified sampling with different stratification weights to sampling without stratification. Our main result is that a method for estimating the average based on stratifying the measurement area with weights proportional to stratum areas significantly outperforms other methods in terms of bias, and can get arbitrarily close to the true average as the number of mobiles and the number of strata increases. We also show that systematic sampling, which is known to usually be more accurate than other spatial sampling techniques (Quenouille, 1949), would rather not perform well in this setting.

We evaluate the methods in an application scenario where mobile nodes move in a linear region according to a Random Waypoint Model (RWP) – for which analytical expressions for the stationary location distribution have been derived in (Bettstetter and Wagner, 2002; Hyytiä et al., 2006) – and take temperature measurements. A phenomenon that occurs in large urban areas is the so-called *Urban Heat Island* (UHI) effect, in which temperatures rise considerably as we move towards the center of the city within relatively small distances (Unger et al., 2001). Such a phenomenon cannot be captured satisfactorily by sparsely located metereological stations, so it is presumed that the use of crowdsensing can produce area temperature estimates with much more accuracy (Muller et al., 2015). In Sect. 5 we construct a simple model of the UHI effect and evaluate the examined sampling techniques when estimating the average temperature in the area.

In remaining parts of the paper, Section 2 provides some basic results on spatial sampling. In Section 4, we focus on estimates at a random instant in time and demonstrate the properties of the stratification method with weights proportional to stratum areas. Section 6 presents numerical results for the bias reduction that can be achieved with the stratification method for a wide range of test cases and configuration parameters. The paper ends in Section 7 with a summary of the most important conclusions and a discussion of open research issues.

2 SPATIAL SAMPLING BASICS

For a continuous[2] parameter $T(\mathbf{x})$ the mean value within an area A of size a is

$$\tilde{T}(A) = \int_A T(\mathbf{x}) d(\mathbf{x})/a \,. \tag{1}$$

If we are at liberty to sample anywhere within the area, then both uniform random sampling and stratified random sampling will produce an unbiased estimate of the mean area value. Indeed, suppose there are n sample points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. In uniform random sampling, each point is selected uniformly independently within A. The expected value of the sample average $\bar{T} = \frac{1}{n}\sum_{i=1}^{n} T(\mathbf{x}_i)$ is

$$E[\bar{T}] = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{a}\int_A T(\mathbf{x})d\mathbf{x} = \frac{1}{a}\int_A T(\mathbf{x})d\mathbf{x} = \tilde{T}(A) \,.$$

In stratified random sampling, the area A is partitioned into L strata or subareas $A'_1, \ldots, A'_L$, each of area s. A uniform random sample of size k is taken from each of the L strata, so that $kL = n$. Suppose the measurement values in stratum i are $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ik}$. First the average value of measurements within each stratum is taken, $\bar{T}_i = \frac{1}{k}\sum_{j=1}^{k}\mathbf{x}_{ij}$. The overall sample average is then calculated as: $\bar{T} = \frac{1}{L}\sum_{i=1}^{L}\bar{T}_i$.

[2]The results also apply to the case of a parameter taking discrete values in the area.

Its expected value is

$$E[\bar{T}] = \frac{1}{L}\sum_{i=1}^{L} E[\bar{T}_i] = \frac{1}{L}\sum_{i=1}^{L}\tilde{T}_i = \frac{1}{sL}\sum_{i=1}^{L}\int_{A'_i} T(\mathbf{x})d(\mathbf{x}) = \tilde{T}(A) \; .$$

Systematic sampling, in which the area is again split into subareas, and the sample values are taken at locations following a deterministic pattern,[3] with some initial randomization, does not in general produce an unbiased estimate. An exception is the case where T is a realization of a homogeneous stochastic process with average value $\mu = E[T(\mathbf{x})]\ \forall \mathbf{x}$.

Despite the difficulty in producing unbiased estimates in systematic sampling, Ripley (Ripley, 2004) showed that systematic sampling outperforms other sampling schemes when $T(\mathbf{x})$ is random and no prior knowledge is available, except when there is periodicity in the measured parameter; similar results where known from (Cochran, 1946) and (Quenouille, 1949).

3 ESTIMATE BASED ON A DENSE MEASUREMENTS SET

In mobile crowdsensing we face a complex situation. The locations of the sample values are determined from the mobiles, so we are not at liberty of selecting any locations we wish. If we indeed had a large number of measurement values that densely covered the whole area, we could derive an accurate estimate based on all available measurements, irrespective of the number and distribution of these measurements in the area. We show that in Proposition 1. But before let us introduce the considered setup and some necessary notation.

We consider a discrete-valued parameter T, for which we want to estimate the mean value over an area A. The parameter is modeled by a step function $T(A_c)$, $c = 1 \ldots C$, where A_c is a subarea of A in which the parameter value remains constant. The parameter value in subarea A_c is also denoted as T_c for brevity. The average value in the area is equal to

$$\tilde{T} = \sum_c T_c\, a(A_c)/a(A) \; , \qquad (2)$$

where $a(\cdot)$ is the measure of the size of the area (length in $\mathbb{R}$, surface in $\mathbb{R}^2$, volume in $\mathbb{R}^3$).

The estimation method is as follows: The area A is split into strata $A'_1,\ldots,A'_L$ of equal size (generally different from subareas $A_1,\ldots,A_C$) and the estimate over each stratum A'_i equals $\hat{T}_{A'_i} = (T_{m_1} + \cdots + T_{m_{n_{A'_i}}})/n_{A'_i}$, where $n_{A'_i}$ is the number of collected measurements in this stratum, with values T_{m_i}, $i = 1,\ldots,n_{A'_i}$. The sampling average over the whole area equals $\hat{T} = (\hat{T}_{A'_i} + \cdots + \hat{T}_{A'_L})/L$.

Proposition 1. *Provided that each stratum is non-empty (i.e. has at least one measurement) w.h.p.(with high probability), then as the number of strata tends to infinity, the estimate $\hat{T}$ tends to the true average $\tilde{T}$ w.h.p.*

Proof. As L increases, there will be a point where each subarea A_c will be greater than each stratum A'_i, $i = 1,\ldots,L$, so that $a(A_c)$ can be decomposed as $a(A_c) = a(A_c^1) + \cdots + a(A_c^{m_c}) - a(\varepsilon_c)$, where $\{A_c^i\}_{i=1\ldots m_c}$ is the subset of $\{A'_i\}_{i=1\ldots L}$ that is the minimum cover set of A_c and ε_c is the excess area that exceeds area A_c when we add the areas in the cover set ($a(\varepsilon) < a(A_c^i)$).

As $L \to \infty$, $a(\varepsilon_c) \to 0$ so that $\{A_c^i\}_{i=1\ldots m_c}$ tends to cover exactly the area A_c. But, since each stratum is non-empty w.h.p., the estimates in $\{A_c^i\}$ are constant and equal to T_c, except for a number of border subareas which cross-over area A_c. Denote the number of border subareas by b_c, and the sum of the estimates in those subareas by $\hat{T}_{b_c}$.

The sampling average over the whole area can then be rewritten as

$$\hat{T} = (\sum_c T_c(m_c - b_c) + \sum_c \hat{T}_{b_c})/L \; .$$

As $L \to \infty$, the excess areas tend to zero, $m_c/L \to a(A_c)/a(A)$ (since the subareas $A'_1,\ldots,A'_L$ are of equal size) while $b_c/L \to 0$, $\sum_c \hat{T}_{b_c}/L \to 0$. Hence $\hat{T} \to \tilde{T}$ w.h.p. □

Intuitively, the condition that, as $L \to \infty$, each subarea is non-empty w.h.p. holds when the number of measurements is greater or increases faster than the number of subareas and the distribution of measurement locations is close-to-uniform.[4] For example, if n measurements are uniformly distributed in the area, this becomes a *balls and bins* problem with n balls into L bins. If, as $L \to \infty$, $n \to \infty$ with $\lambda = n/L$, a well-known result in combinatorics is that the distribution of balls into bins approaches a Poisson distribution with rate λ (e.g. see (Mitzenmacher and Upfal, 2005, Section 5.3.1.)). Therefore, the probability that

[3] There are many variations of systematic sampling, depending on the area dimensions and the alignment or non-alignment of sampling locations in each direction. For examples, the reader is referred to (Ripley, 2004, Section 3.1)

[4] If mobile measurements are uniformly distributed over the area, one need only take the sample average for a finite number of measurements without any stratification to produce an unbiased estimate of the area average. However, stratification may still be useful in reducing the sample variance.

each subarea is non-empty is, in the limit, $1-e^{-\lambda}$ and for $n > L$ each subarea is non-empty w.h.p.

However, if the distribution of measurement locations is not uniform (for instance, as a result of a non-uniform movement distribution of the mobile users), it may well happen that some subareas are left empty. In a practical algorithm, given a number of available measurements in an area, one could increase the number of strata L successively to produce a more accurate estimate, and the increase could stop when an empty area is found.

4 ESTIMATE AT A RANDOM INSTANT

We now consider that there are n mobile users roaming in the area following the same mobility model. Suppose that each mobile user's location X in area A, if sampled at a random instant in time, is described by a distribution with pdf $f_X(x)$.[5]

The probability that a single mobile user is in a subarea $A_c \subset A$ is then $P(A_c) = \int_{A_c} f_X(x)dx$. Assuming that the movements of the mobile users are independent, we can derive the expected number of mobile users in the area as $nP(A_c)$.

Note that we do not demand that the random sampling instants are the same for each mobile, as this could face synchronization difficulties (especially in the absence of a GPS service). For independent movement processes of the mobiles the analysis still holds, as long as each mobile's process is sampled independently (in the sense explained in footnote 5). Further, although we assume a single measurement from each mobile, the analysis that follows also holds when we can afford to take more than measurements from each mobile at random instants in time, since this would be equivalent to increasing the number of mobiles in the area.

[5] This distribution could be the limiting distribution of a stationary ergodic stochastic process describing user movements, whereas the sampling process could be an independent Poisson process; then the time average distribution (of the stochastic process describing user movements) is the same as the distribution obtained when averaging over the sampling times. This also holds under weaker assumptions on the observed process, as well as more generic sampling processes (such as when the observed process only has a constant finite time average and the sampling process is an independent renewal process with a non-lattice cycle length distribution, where the cycle length ℓ satisfies $\mathrm{E}\ell^{1+\varepsilon} < \infty$, for some $\varepsilon > 0$). For more details readers are referred to (Glynn and Sigman, 1998).

4.1 Estimate without Stratification

We first examine the case where we estimate the mean parameter value in A by sampling all mobiles in the area without any stratification. The measurement of each mobile i, T_{m_i}, is supposed to be the one at the point where the mobile is found when it is sampled. Given that the movements of mobile users are independent and that sampling is performed at a random instant in time, the parameter readings of the mobile users become i.i.d. random variables.

The expected value $E[T_m] := E[T_{m_i}]$ of the parameter reading of each mobile i ($i = 1, \ldots, n$) is

$$E[T_m] = \sum_c T_c P(A_c) \; . \qquad (3)$$

To derive an estimate of the mean area value, denoted by $\hat{T}_w$, we simply take the average of these measurements. Since we have a set of i.i.d. random variables, the expectation of their average equals the expected value of each of these variables. Hence $E[\hat{T}_w] = E[T_m]$.

As anticipated, this expectation is independent of the number of mobiles n. Therefore, the estimate does not change if we randomly select a subset of the mobile users rather than the whole population.

Clearly this estimate is biased. The bias $E[\hat{T}_w] - \tilde{T}$ reflects the extent to which the location distribution of each mobile deviates from the uniform distribution.

Denoting the variance of the measurement value of each mobile by $Var(T_m)$, the variance of the average is

$$Var(\hat{T}_w) = \frac{Var(T_m)}{n} = \frac{1}{n}\left(E[T_m^2] - E^2[T_m]\right) \; , \qquad (4)$$

that is, it is $1/n$ times the variance of the parameter reading of a single mobile in the area. (This is also straightforward since we take the variance of an average of i.i.d random variables.) It is also readily derived that if we randomly select a subset of the mobile population of size $k < n$, the variance of the sample average will be $Var(T_m)/k$.

4.2 Estimate with Stratification

Consider now partitioning the area into subareas or strata $A'_1, \ldots, A'_L$, taking the average of measurements in each stratum and combining these into a single estimate. Stratification can be done as part of the processing of the values recorded by the mobiles; it is not necessary to sample all mobiles in a certain stratum separately. Provided that each mobile also records the location at which the measurement is taken, the processing unit can subsequently discern which measurements are taken at each stratum.

We will consider two different types of weights of the stratum averages: a) based on the number of mobiles found in each stratum, and b) based on the area of each stratum:

$$\text{(a)} \quad \hat{T}_{st}^{n} = \sum_{h=1}^{L} \frac{n_h \hat{T}_{w,h}}{n} \tag{5a}$$

$$\text{(b)} \quad \hat{T}_{st}^{s} = \sum_{h=1}^{L} \frac{a(A'_h)\mathbb{1}_{A'_h}\hat{T}_{w,h}}{\sum_{j=1}^{L} a(A'_j)\mathbb{1}_{A'_j}}\,, \tag{5b}$$

where n_h is the number of users from stratum h, $h = 1\ldots L$, and $\hat{T}_{w,h} = (T_{m_1} + \cdots + T_{m_{n_h}})/(n_h)$ is the temperature estimate based on the users in this stratum. $\mathbb{1}_{A'_h}$ is the indicator function which equals 1 if A'_h is non-empty, and zero otherwise.

In the special case where the strata are of equal size, the estimate (b) becomes

$$\hat{T}_{st}^{s} = \sum_{h=1}^{L} \frac{\mathbb{1}_{A'_h}\hat{T}_{w,h}}{\sum_{j=1}^{L} \mathbb{1}_{A'_j}}\,. \tag{6}$$

Only non-empty subareas are considered in the estimate, that is if no mobile is found in a subarea, then this subarea is omitted. This is reflected with the indicator function in (5b). (No indicator function is needed in (5a), since n_h will be zero if A'_h is empty.)

If the strata are always non-empty (i.e. $n_h \neq 0$ $\forall\, h$), then as the number of strata increases, the estimate will approach the true average from Proposition 1. However, as the number of strata increases, so does the probability of a stratum being empty, in which case the error is expected to increase. We will investigate this trade-off.

4.2.1 Weighting Proportionally to the Number of Mobiles in Each Stratum

Interestingly, the expected value of the estimate in (5a) is the same as in the non-stratification case. To show this, we begin by noting that the parameter readings of mobile users in each stratum are i.i.d. random variables. Therefore, by applying Wald's equation,

$$E[\hat{T}_{st}^{n}] = \frac{1}{n}\sum_{h=1}^{L} E[n_h]E[T_{m|h}]\,, \tag{7}$$

where $E[T_{m|h}]$ is the expected parameter reading of a mobile user in stratum h.[6]

The expected number of users in stratum h is $E[n_h] = n\int_{A_h} f_X(x)dx$. Denoting by $A_{h,c}$ the subarea formed by the intersection of A'_h, A_c, we have that

$$E[T_{m|h}] = \sum_c T_c \int_{A_{h,c}} f_{X|h}(x)dx\,, \tag{8}$$

where $f_{X|h}(x)$ is the conditional distribution of the mobile user position confined in A'_h:

$$f_{X|h}(x) = \frac{f_X(x)}{\int_{A'_h} f_X(x)dx}\,. \tag{9}$$

Hence from (7),(8),(9) the mean value of the estimate is

$$\begin{aligned} E[\hat{T}_{st}^{n}] &= \frac{1}{n}\left(\sum_{h=1}^{L} n\int_{A'_h} f_X(x)dx\left(\sum_c T_c \int_{A_{h,c}} f_{X|h}(x)dx\right)\right) \\ &= \sum_{h=1}^{L}\int_{A'_h} f_X(x)dx\left(\sum_c T_c \int_{A_{h,c}} f_{X|h}(x)dx\right) \\ &= \sum_{h=1}^{L}\sum_c T_c \int_{A_{h,c}} f_X(x)dx \\ &= \sum_c T_c \int_{A_c} f_X(x)dx = E[T_m]\,. \end{aligned} \tag{10}$$

Therefore, however we may stratify the area, the expected value of the estimate is the same as in the non-stratification case.

4.2.2 Weighting Proportionally to Stratum Areas

We will proceed to derive the expected value of the average in the case of stratification with weights proportional to the area of each stratum. From (5b) we have:

$$E[\hat{T}_{st}^{s}] = \sum_{h=1}^{L} E\left[\frac{a(A'_h)\mathbb{1}_{A'_h}\hat{T}_{w,h}}{\sum_{j=1}^{L} a(A'_j)\mathbb{1}_{A'_j}}\right] \tag{11}$$

In order to proceed with the analysis, we assume that the total non-empty area under a certain partition (which is in the denominator of the fraction in (11)) is approximately independent of $a(A'_h)\mathbb{1}_{A'_h}\hat{T}_{w,h}$ in any of the strata.[7] Furthermore, using the first-degree Taylor series approximation[8]

$$E\left[\frac{1}{\sum_{j=1}^{L} a(A'_j)\mathbb{1}_{A'_j}}\right] \approx \frac{1}{E[\sum_{j=1}^{L} a(A'_j)\mathbb{1}_{A'_j}]} \tag{12}$$

we have that

$$E[\hat{T}_{st}^{s}] \approx \sum_{h=1}^{L} \frac{a(A'_h)P_{ne}(A'_h)}{\sum_{j=1}^{L} a(A'_j)P_{ne}(A'_j)} E[\hat{T}_{w,h}]\,, \tag{13}$$

[6]Note that Wald's equation, and therefore (7) also holds when $n_h = 0$ in some stratum h.

[7]In reality, a very weak dependence is expected between these two terms.

[8]For a random variable x, a more accurate approximation is $E[1/x] \approx 1/E[x] + 1/E[x]^3 Var(x)$. In our case, x is the total non-empty area; further, as the number of mobiles increases, the second term of the approximation decreases and the first-degree approximation is tighter.

where $P_{ne}(A'_h)$ is the probability of A'_h being non-empty, $P_{ne}(A'_h) = 1 - (1 - \int_{A'_h} f_X(x)dx)^n$.

If $T_{m_{i|h}}$ is the parameter reading of a mobile user i ($i = 1 \ldots n_h$) in stratum h, with common expectation $E[T_{m|h}]$, then

$$E[\hat{T}_{w,h}] = E\left[\frac{T_{m_{1|h}} + \cdots + T_{m_{n_h|h}}}{n_h}\right] = E[n_h]E\left[\frac{T_{m_{i|h}}}{n_h}\right] \approx E[T_{m|h}] \,, \tag{14}$$

by applying Wald's equation and subsequently using the approximation $E[1/n_h] \approx 1/E[n_h]$. Combining (13),(14) and (8) gives us an approximate value of the expectation.

4.3 Estimate with Systematic Sampling

We consider that the area is partitioned into kL contiguous subareas of equal size. Initially, a random subarea is selected from $1 \ldots k$ and then sampling continues by selecting every k_{th} consecutive subarea until L subareas have been chosen. Similarly to stratified sampling, an estimate of the environmental parameter value in a subarea is produced by averaging the mobile measurements in this subarea. To produce the overall estimate, the estimate in each selected non-empty subarea is weighted by the fraction of the subarea size relative to the sum of the sizes of all selected non-empty subareas.

The formal expression of the estimate is thus similar to (5b):

$$\hat{T}^s_{sy} = \sum_{h=1}^{kL} \frac{a(A'_h)\mathbb{1}'_{A'_h}\hat{T}_{w,h}}{\sum_{j=1}^{kL} a(A'_j)\mathbb{1}'_{A'_j}} \,, \tag{15}$$

where $\mathbb{1}'_{A'_h}$ now equals one if A'_h is included in the sample *and* it is non-empty. For each subarea A'_h we now have that $E[\mathbb{1}'_{A'_h}] = P_{ne}(A'_h)/k$.

Using the same approximations that led us to (13), we have for the expected value of the estimate:

$$E[\hat{T}^s_{sy}] \approx \sum_{h=1}^{kL} \frac{a(A'_h)P_{ne}(A'_h)/k}{\sum_{j=1}^{kL} a(A'_j)P_{ne}(A'_j)/k} E[\hat{T}_{w,h}] \,. \tag{16}$$

Hence we derive the following conclusion:

Corollary 1. *The expectation of the estimate with systematic sampling and L selected strata is approximately the same as the expectation of the stratification estimate with weights proportional to stratum areas, and a total of kL strata.*

4.4 Properties of the Stratification Estimate with Weights Proportional to Stratum Areas

At this point, it is worth elaborating on some properties of the stratification estimate with weights proportional to stratum areas, which help to illuminate the worthiness of the method and provide insight for the results that follow. We consider equal-sized strata, i.e. that the estimate (6) is used.

First, we show in the following proposition that when n is finite, the two estimates (6) and (5a) coincide as $L \to \infty$.

Proposition 2. *Consider equal-sized strata and a finite mobile population, where each mobile has a continuous location pdf f_X. Then as the number of strata tends to infinity, the stratification estimate with weights proportional to stratum areas and the stratification estimate with weights proportional to the number of mobiles in each stratum coincide with probability 1.*

Proof. Suppose $X_1, \ldots, X_n$ are the random variables representing the mobiles' positions in area A. Then since the location pdf of each mobile is a continuous function, the probability that any two mobiles are infinitesimally close is zero. Hence, as $L \to \infty$, after some value of L only a single mobile will reside in each stratum and all variables $\mathbb{1}_{A'_j}$ in (6) become zero except for some areas $A''_1, \ldots, A''_n$ around the mobiles. Therefore, $\lim_{L\to\infty} \sum_{j=1}^{L} \mathbb{1}_{A'_j} = \sum_{i=1}^{n} \mathbb{1}_{A''_i} = n$. For the same reason, $\lim_{L\to\infty} \sum_{h=1}^{L} n_h \hat{T}_{w,h} = \sum_{i=1}^{n} T_{m_i} = \lim_{L\to\infty} \sum_{h=1}^{L} \mathbb{1}_{A'_h} \hat{T}_{w,h}$. Hence the two estimates coincide with probability 1. □

Since, as we saw in Section 4.1 the expected value of the stratification estimate with weights proportional to the number of mobiles in each stratum coincides with the expected value of the non-stratification estimate, we also have the following:

Corollary 2. *Under the setting of Proposition 2, the expected value of the stratification estimate with weights proportional to stratum areas coincides with the expected value of the non-stratification estimate.*

Additionally, if the mobile population is so large that the strata are non-empty w.h.p as L increases, $\hat{T}^s_{st}$ will tend to the true average from Proposition 1. For finite L this does not hold. But what is challenging is to show that even for finite L, $\hat{T}^s_{st}$ produces a smaller bias than the non-stratification estimate.

Intuitively, the explanation for this goes as follows. The bias is mainly caused by the larger concentration of mobiles in one or more areas. (If mo-

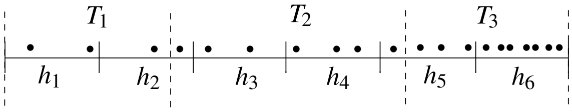


Figure 1: A realization of the mobiles' positions in a region with three discrete environmental parameter values and a stratification into 6 strata.

biles were uniformly distributed in the region the estimate would be unbiased, as this is equivalent to uniform random sampling.) Stratification serves to create a virtual sample of measurement locations, which is closer to a uniform distribution.

An illustration of this is shown in Fig. 1. The region is divided in three areas, with environmental parameter values T_1, T_2 and T_3. We stratify into 6 equal subareas (strata), $h_1, \ldots, h_6$. Suppose we have 20 mobiles, whose location distribution is concentrated in the right-most areas. The filled dots represent a realization of the mobiles' positions at a random instant of time. The estimate without stratification will produce an average much closer to T_3, since half of the mobiles are located in this subarea. On the other hand, stratification produces the same effect as if we had a virtual sample composed of a single location in each stratum. Hence the stratification estimate will smooth out the skewness caused by the concentration of mobiles, producing a value much closer to the true area average.

More formally, let us denote the bias of the estimate in a stratum h_i by B^{h_i}. If we assume that strata are always non-empty, the bias of the stratification estimate equals the bias of a randomly chosen stratum:

$$E[\hat{T}^s_{st}] - \tilde{T} = \frac{1}{L}(B^{h_1} + \cdots + B^{h_L}) . \quad (17)$$

By considering the same partition into strata in the non-stratification case, the bias can be written as[9]

$$E[\hat{T}_w] - \tilde{T} = E[T_m] - \tilde{T} = B^{h_1}P(h_1) + \cdots + B^{h_L}P(h_L) . \quad (18)$$

The concentration of mobiles in some areas, which largely causes the bias, is only reflected in (18), and not in (17). Therefore, in practical cases we can expect that the stratification estimate with weights proportional to stratum areas will produce a much smaller bias than the non-stratification estimate.

5 APPLICATION SCENARIO

We consider a linear area that exhibits the so-called *Urban Heat Island* (UHI) effect. Mobile nodes (either human users or vehicles) are roaming in the area, equipped with devices able to conduct temperature measurements. Our goal is to estimate the area mean in (2) from a sample of the mobile measurements, as accurately as possible. For simplicity, we assume that there are no errors in individual measurements.

We employ the RWP model, which is one of the most widely used models in mobile and ad-hoc networks. In the general version of the model, a mobile user chooses a random destination and moves to it at a randomly chosen speed. Once at the destination, the user stops for a pause time, then picks another destination at random and repeats the same process. Parameters of the model include the movement area, the number of mobile users, speed and pause time, as well as the resolution of the destination points (may range from a single point to a bounded area).

The main reason for choosing the RWP model in this paper is the fact that analytic formulas for the limiting spatial distribution of a mobile node exist. For a node moving according to the RWP model in a restricted one-dimensional area $[-x_m, x_m]$ with constant speed, uniformly distributed destination points and equal pause times at those points, the probability density function of its location X is (Bettstetter and Wagner, 2002):

$$f_X(x) = -\frac{3}{4x_m^3}x^2 + \frac{3}{4x_m} \quad \text{for } -x_m \le x \le x_m . \quad (19)$$

Under this model, a node is more likely to be found in the center of the area, while the probability that is it is located at the border tends to zero (solid blue curve in Fig. 2).

The UHI effect is quantified by the difference between the temperature at a certain point and the lowest temperature observed in the area. Usually an area is split into subareas and normalized UHI values are taken in each subarea by dividing with the largest UHI value. Climate studies have shown that normalized values are largely independent of the seasonal climatological conditions and are determined to a high degree by urban factors (buildings, roads, population density, traffic, etc.) (Unger et al., 2001).

The general cross-section of the typical UHI effect described in (Oke, 2002) consists of a cliff, plateau and peak, corresponding to rural, suburban, and urban areas. In each one of these areas the temperature may fluctuate, but on average clear level shifts can be observed when we move from one area to the other.

A simplified model of the UHI effect consisting of a 3-step function – without normalizing temperature values – is depicted in Fig. 2 (red densely dotted line). Each step corresponds to a subarea (rural, suburban and urban). The corresponding temperatures are $T_r < T_{sub} < T_u$ and x_u, x_r mark the limits of the

[9] Since the bias of a single mobile measurement in each stratum is equal to the bias of the average of n mobile measurements in the same stratum.

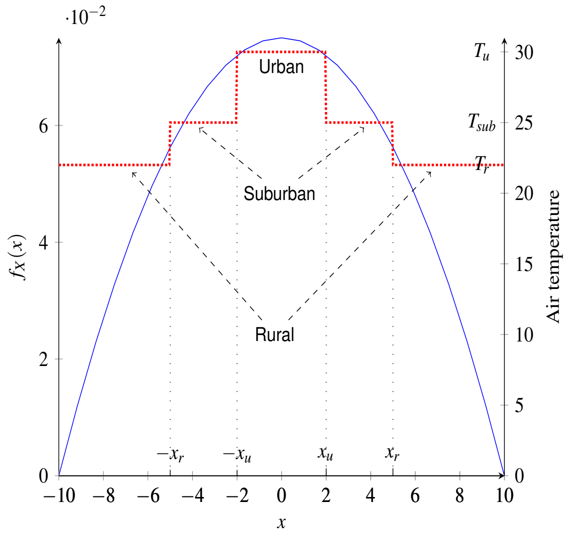


Figure 2: Plot of the pdf of the one-dimensional RWP model (x_m=10), together with a simple model of an UHI (red densely dotted line).

urban and rural areas respectively. The mean value of the temperature in the area, which we are trying to estimate, is

$$\tilde{T} = T_u \frac{x_u}{x_m} + T_{sub} \frac{x_r - x_u}{x_m} + T_r \frac{x_m - x_r}{x_m} \,. \tag{20}$$

6 NUMERICAL RESULTS

Supposing each mobile user follows a one-dimensional RWP mobility model with pdf given in (19), the probability of a mobile to be in a subarea $[a,b]$ corresponding to a step of the environmental parameter function is

$$\int_a^b \left(-\frac{3}{4x_m^3} x^2 + \frac{3}{4x_m} \right) dx = -\frac{1}{4} \frac{b^3 - a^3}{x_m^3} + \frac{3}{4} \frac{b-a}{x_m} \,. \tag{21}$$

We first present some basic performance evaluation results based on the model in Fig. 2. We examine the bias reduction of the stratification method, when weighting proportionally to stratum areas, compared to the bias of the estimate without stratification. Equal sized-strata are considered. The bias reduction is zero for $L = 1$, $L = 2$: The first case is evident since it amounts to no-stratification. The second is because in the setting of Fig. 2, for $L = 2$ we divide into two symmetric subareas, each of which produces the same estimate.

As shown in Fig. 3, the stratification estimate with weights proportional to stratum areas is much better than the non-stratification estimate, and approaches its value as the number of strata increases. We can also get arbitrarily close to the true value of the average (which yields a bias reduction of almost 100%), as the number of mobiles increases. All these results were anticipated from the analysis in Section 4.4.

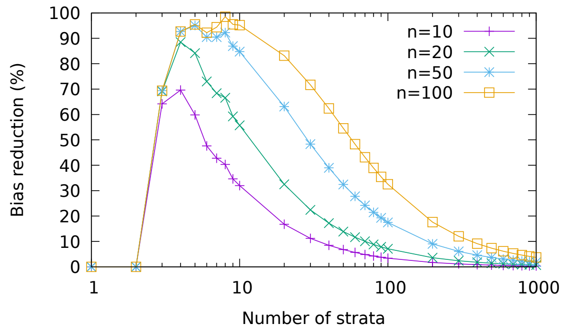


Figure 3: Bias reduction of the stratification method when weighting proportionally to stratum areas. The 3-step function shown in Fig. 2 is used for the environmental parameter.

Notice also that a maximum reduction exists for some intermediate value of L, which reflects the trade-off discussed in Section 4.2 between attempting to improve the accuracy of the estimate by introducing more strata and the possibility of finding these strata empty. In the results of Fig. 3 the maximum is achieved for only a few strata, and the optimal value of L increases with the number of mobiles. (The optimal value is $L = 4$ for $n = 10, 20, 50$, and $L = 8$ for $n = 100$.) Empty strata modify the weights in the estimate so that the subareas that contain the most mobiles have non-zero weights with higher probability, thus skewing the estimate.

We also conclude that the version of systematic sampling studied in Section 4.3 would only show smaller bias than the stratification estimate for very small values of the parameters L and k, where the product kL would be kept relatively small. Hence systematic sampling would be less efficient than stratified sampling, unless the savings (in messaging and processing cost) by sampling only the selected areas can outweigh the performance deterioration.

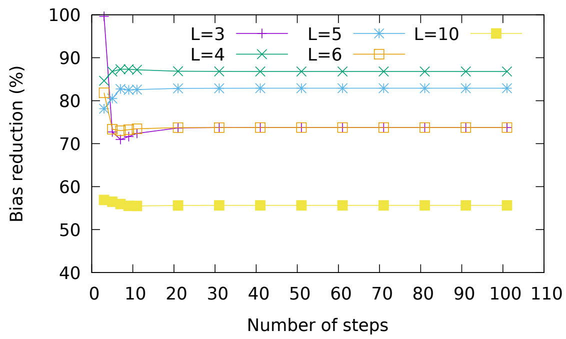
Performance is always improved when increasing the number of mobiles. However, even a small number of mobiles suffices to get a significant bias reduction. Additionally, as the number of mobiles increases, there may also be local maxima in the bias reduction (notice the cases $n = 50$, $n = 100$). Nevertheless, these local maxima are close to each other and their respective values do not differ very much.

Next we proceed to a more systematic assessment of the performance of the stratification method, compared to the estimate without stratification. This assessment serves to provide guidance into how the number of strata L should be selected for different change patterns of the environmental parameter. We will vary both the number of steps, as well as the relative lengths of the steps in the environmental parameter function. A larger number of steps represents the more realistic case where the environmental parameter changes less abruptly in the area. The relative lengths will be defined by the ratio of a geometric series, which can provide us with different patterns, from a steep decrease of the inner subarea lengths to a more uniform distribution.

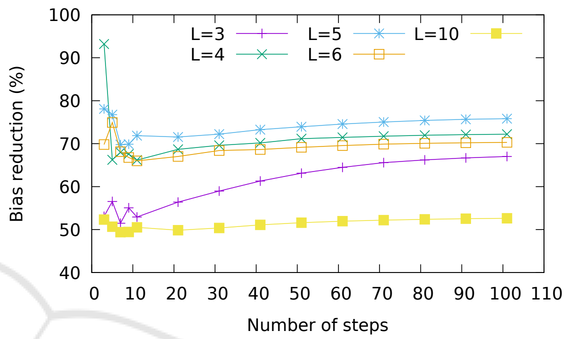
We consider C steps of the environmental parameter step function (in total). The subareas corresponding to steps which are symmetric with respect to the center of the area are equal (hence C is always an odd number). Larger subareas appear toward the edges, similarly to the function in Fig. 2. To this effect, subarea lengths are defined by a geometric series with ratio r. This results into the length of the two edge subareas being equal to $x_m(1-r)/(1-r^{(C+1)/2})$; the subsequent inner subarea lengths are defined by multiplying successively by r. As r increases, the lengths of the different subareas become more uniform. The values of the environmental parameter are also symmetric with respect to the center and gradually increase from T_{min} in the edge subareas to T_{max} in the center subarea, with a fixed increment equal to $2(T_{max}-T_{min})/(C-1)$.

An important property that follows from this setup is that the probability of a mobile to be in a subarea $[a,b]$ (corresponding to a step of the environmental parameter function) is dependent only on the parameters C, r, and independent of the actual value of x_m. This follows directly from (21) since in the considered setup all subarea lenghts are defined as multiples of x_m; hence all points a, b inside $[-x_m, x_m]$ that delimit the subareas are also multiples of x_m and the location probability (21) remains the same. Similarly, since the strata are derived by splitting the entire area into equal parts, the length of the strata, as well as the points inside $[-x_m, x_m]$ that delimit the strata are proportional to x_m. Therefore, all probabilities in (13) are also independent of x_m.

Results for the bias reduction under the stratification method with varying C and r are shown in Fig. 4. It can be observed that a larger bias reduction occurs for decreasing r. On the other hand, the bias reduction is approximately constant as the number of steps increases, except when there is a very small number of steps. Indeed, for all but very small values of C, the bias when using stratification is approximately a constant fraction of the bias without stratification. The fluctuations for small C depend on the match between the set of subareas corresponding to the steps



(a) r=0.5



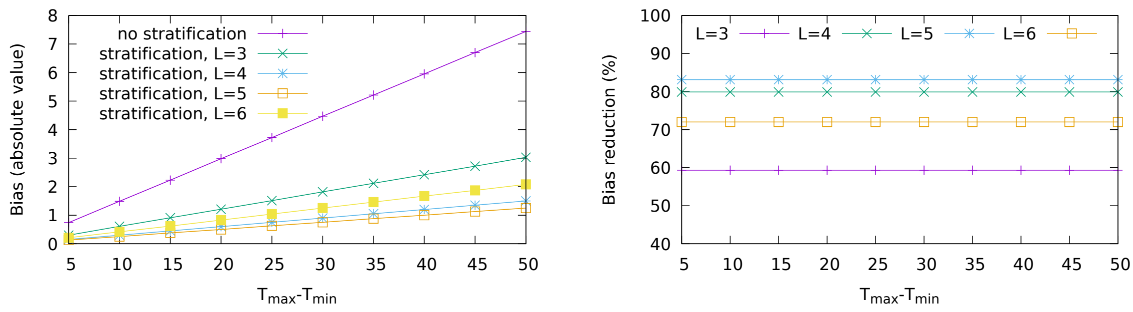
(b) r=0.9

Figure 4: Bias reduction of the stratification method for different numbers of steps of the environmental parameter function and different number of strata ($n = 20$, T_{min}=22, T_{max}=30).

of the environmental parameter function and the set of strata. For example, in subfigure 4a, for $C = 3$, $L = 3$, the two sets almost coincided, and the bias reduction approaches 100%. On the other hand, in subfigure 4b, for $C = 3$, $L = 3$, the points where the environmental parameter changes were {-10, -4.74, 4.74, 10}, while the end points of the strata were {-10, -3.33, 3.33, 10} and there is a much lower reduction. As C increases and the size of steps becomes smaller than the size of the strata, this effect is mitigated.

The optimal number of strata is for all examined cases small and does not depend significantly on the value of r (in Fig. 4, the optimal number is $L = 4$ for $r = 0.5$ and $L = 5$ for $r = 0.9$); amongst the shown values of L, the highest value yields the lowest bias reduction. One might have anticipated that, as C increases, a larger L would bring more benefits. This however is not true and it seems that the possibility of a stratum being empty weighs more in the performance of the algorithm, not allowing to achieve more gains. Overall, we observe that the number of steps is not a significant factor in the performance of the stratification method.

Results for larger temperature intervals are shown in Fig. 5. Notice that, since both the average esti-



(a) (b)

Figure 5: Absolute values of the bias (a) and bias reduction (b) of the stratification method for different temperature intervals ($n = 20$, C=5, r=0.7).

mates and the actual mean value of the environmental parameter are obtained as normalized weighted sums, the bias values depend only on the difference $T_{max} - T_{min}$ and the bias increases linearly with greater temperature ranges. (See Fig. 5a. The bias increase results from the larger concentration of mobiles towards the center, whose higher temperature readings increase the estimate value.) As a result, the relative bias reduction remains constant as the parameter range increases (Fig. 5b).

7 CONCLUSIONS AND OPEN RESEARCH ISSUES

The theoretical and numerical results in this paper manifest that the stratification method for estimating the area average of an environmental parameter achieves significant bias reduction over a naive estimate without stratification, even by sampling a small number of mobiles and having a small number of strata. Moreover, the method can get arbitrarily close to the true average as the number of mobiles increases, for a moderate number of strata. The optimal strata number increases as the number of mobiles increases, but still remains relatively small in all our test cases. (This is convenient, as increasing the number of strata would increase the processing cost, and thus the overall cost of the method.) In addition to the above, the number of strata could be chosen independently of the difference in range in the environmental parameter values within the region of interest and of the change pattern of this parameter. Although the setting for the evaluation was tailored to the pattern of temperature change in an area that exhibits the Urban Heat Island effect, these conclusions should have wider applicability, since they reflect fundamental properties of the estimate.

Despite these encouraging results, there are still open issues that require further research, both to advance the theory and to proceed to a real implementation of the method. A first issue is the calculation of the variance of the stratification estimates. The calculation of the variance is important, especially since we would like to estimate the average with a single measurement from each mobile. Another challenge would be to evaluate the performance of the methods in two-dimensional space. An exact expression for the pdf of the mobile locations generated by the RWP model was given in (Hyytiä et al., 2006) for a general convex area, and simpler approximate expressions for square and circular areas were given in (Bettstetter and Wagner, 2002). The theoretical analysis extends to two dimensions in a straightforward manner, although the computation becomes much more difficult. Extending the results to the case of continuous varying parameters in space would also be interesting when there are relevant models available. Otherwise, the discrete analysis suffices since the true parameter value would only be known by measurements on a set of discrete subareas. Finally, a fundamental problem is to examine the accuracy of an estimate obtained by periodically sampling the mobiles within a certain time period. This is of great interest, since this kind of sampling is more likely applicable in practice. The issue that arises is whether an estimate based on measurements that were collected periodically would be better than an estimate at a random snapshot.

Additionally, we mention a few open issues regarding a potential implementation of the method. We intentionally left out the details about the communication process for acquiring the measurement results from the mobile devices, as we consider that there are a lot of solutions available, each of which would deserve a thorough analysis. For example, the mobiles could have a software installed that executes to have measurements taken at random time instants and transmit the results, along with their geographical coordinates, to a central unit for deriving the estimate. Other solutions could involve the sending of query messages, e.g. from cellular base stations or WLAN access points. The mobile devices would receive the messages, execute the measurement and transmit the result and their position to the sender.

The method should also be checked for its efficiency when the movements of the mobile devices are described by other mobility models or are not independent, as in the cases where they move in groups, or move towards popular places. The RWP model has been shown to be a good approximation for modeling the motion of vehicles in a road (Saha and Johnson, 2004). Models for better approximating human mobility have been described in (Gonzalez et al., 2008; Rhee et al., 2011), while non-independence of mobile movements through group mobility models has been studied in (Musolesi et al., 2004),(Lee et al., 2009). The RWP model is a super-diffusive model, which means that there is a higher probability of longer displacements; hence the mobile locations are likely to be even more concentrated towards the center in a more realistic model. Clustered user movements will skew the location distribution, so that the estimated average is farther from the true value. Nevertheless, the method with stratification has the effect of smoothing out the skewness and therefore could still produce a significant improvement. As a matter of fact, the higher the skewness, the higher the expected bias reduction by the stratification method. Therefore we expect even more gains by applying the stratification method under more realistic mobility models, than under the RWP model.

Furthermore, we have not been concerned with the accuracy of single user measurements, or the effect of noise in such measurements. The authors in (Fiore et al., 2013) showed that the accuracy of the measurements collected by the users plays a more critical role than the number of users participating in crowdsensing, and an accurate overall estimate can be obtained with a relatively small number of accurate user measurements. Hence it is important to examine the accuracy of user measurements, and filter out measurements that are suspected to be inaccurate. In the application scenario examined in this paper, devices in vehicles could more accurately measure ambient temperature than devices carried by humans, as direct contact with ambient air is always achieved. In both cases, filtering of measurements would be required to eliminate possible sources of bias: indoor environments (detection of indoor/outdoor environment as in (Krumm and Hariharan, 2004)), human contact with the sensor, exhaustion gas from other vehicles, etc.

Finally, there exist many other challenges for conducting crowdsensing measurements, such as providing participation incentives to the users, or protecting from malicious users who may “pollute” the data. Interested readers are referred to the surveys (Ma et al., 2014; Ganti et al., 2011) for basic information.

REFERENCES

Antonic, A., Bilas, V., Marjanovic, M., Matijasevic, M., Oletic, D., Pavelic, M., Zarko, I. P., Pripuzic, K., and Skorin-Kapov, L. (2014). Urban crowd sensing demonstrator: Sense the Zagreb air. In *Software, Telecommunications and Computer Networks (SoftCOM), 2014 22nd International Conference on*, pages 423–424. IEEE.

Bash, B. A., Byers, J. W., and Considine, J. (2004). Approximately uniform random sampling in sensor networks. In *Proceeedings of the 1st international workshop on Data management for sensor networks: in conjunction with VLDB 2004*, pages 32–39. ACM.

Bettstetter, C. and Wagner, C. (2002). The spatial node distribution of the random waypoint mobility model. *WMAN*, 11:41–58.

Cochran, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *The Annals of Mathematical Statistics*, pages 164–177.

Considine, J., Li, F., Kollios, G., and Byers, J. (2004). Approximate aggregation techniques for sensor databases. In *Data Engineering, 2004. Proceedings. 20th International Conference on*, pages 449–460. IEEE.

Datta, S. and Kargupta, H. (2007). Uniform data sampling from a peer-to-peer network. In *Distributed Computing Systems, 2007. ICDCS'07. 27th International Conference on*, pages 1–8. IEEE.

Fiore, M., Nordio, A., and Chiasserini, C.-F. (2013). Investigating the accuracy of mobile urban sensing. In *Wireless On-demand Network Systems and Services (WONS), 2013 10th Annual Conference on*, pages 25–28. IEEE.

Ganesan, D., Ratnasamy, S., Wang, H., and Estrin, D. (2004). Coping with irregular spatio-temporal sampling in sensor networks. *ACM SIGCOMM Computer Communication Review*, 34(1):125–130.

Ganti, R. K., Ye, F., and Lei, H. (2011). Mobile crowdsensing: current state and future challenges. *Communications Magazine, IEEE*, 49(11):32–39.

Glynn, P. and Sigman, K. (1998). Independent sampling of a stochastic process. *Stochastic processes and their applications*, 74(2):151–164.

Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Hyytiä, E., Lassila, P., and Virtamo, J. (2006). Spatial node distribution of the random waypoint mobility model with applications. *IEEE Transactions on Mobile Computing*, 5(6):680–694.

Kempe, D., Dobra, A., and Gehrke, J. (2003). Gossip-based computation of aggregate information. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 482–491. IEEE.

Krumm, J. and Hariharan, R. (2004). Tempio: inside/outside classification with temperature. In *Second International Workshop on Man-Machine Symbiotic Systems*.

Kurant, M., Gjoka, M., Butts, C. T., and Markopoulou, A. (2011). Walking on a graph with a magnifying glass: stratified sampling via weighted random walks. In *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pages 281–292. ACM.

Lee, K., Hong, S., Kim, S. J., Rhee, I., and Chong, S. (2009). Slaw: A new mobility model for human walks. In *INFOCOM 2009, IEEE*, pages 855–863. IEEE.

Ma, H., Zhao, D., and Yuan, P. (2014). Opportunities in mobile crowd sensing. *Communications Magazine, IEEE*, 52(8):29–35.

Massoulié, L., Le Merrer, E., Kermarrec, A.-M., and Ganesh, A. (2006). Peer counting and sampling in overlay networks: random walk methods. In *Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing*, pages 123–132. ACM.

Mitzenmacher, M. and Upfal, E. (2005). *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press.

Muller, C., Chapman, L., Johnston, S., Kidd, C., Illingworth, S., Foody, G., Overeem, A., and Leigh, R. (2015). Crowdsourcing for climate and atmospheric sciences: current status and future potential. *International Journal of Climatology*.

Musolesi, M., Hailes, S., and Mascolo, C. (2004). An ad hoc mobility model founded on social network theory. In *Proceedings of the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*, pages 20–24. ACM.

Oke, T. R. (2002). *Boundary layer climates*. Routledge.

Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, pages 355–375.

Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S. J., and Chong, S. (2011). On the levy-walk nature of human mobility. *IEEE/ACM transactions on networking (TON)*, 19(3):630–643.

Ripley, B. D. (2004). *Spatial statistics*, volume 575. John Wiley & Sons.

Saha, A. K. and Johnson, D. B. (2004). Modeling mobility for vehicular ad-hoc networks. In *Proceedings of the 1st ACM international workshop on Vehicular ad hoc networks*, pages 91–92. ACM.

Shrivastava, N., Buragohain, C., Agrawal, D., and Suri, S. (2004). Medians and beyond: new aggregation techniques for sensor networks. In *Proceedings of the 2nd international conference on Embedded networked sensor systems*, pages 239–249. ACM.

Stutzbach, D., Rejaie, R., Duffield, N., Sen, S., and Willinger, W. (2009). On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Transactions on Networking (TON)*, 17(2):377–390.

Unger, J., Sümeghy, Z., and Zoboki, J. (2001). Temperature cross-section features in an urban area. *Atmospheric Research*, 58(2):117–127.

van der Hoeven, F., Wandl, A., Demir, B., Dikmans, S., Hagoort, J., Moretto, M., Sefkatli, P., Snijder, F., Songsri, S., Stijger, P., et al. (2014). Sensing hotterdam: Crowd sensing the rotterdam urban heat island. *SPOOL*, 1(2):43–58.