Estimating spatial averages of environmental parameters based on mobile crowdsensing

Ioannis Koukoutsidis

February 25, 2017

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 ubiquitous 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 [1, 32].

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

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

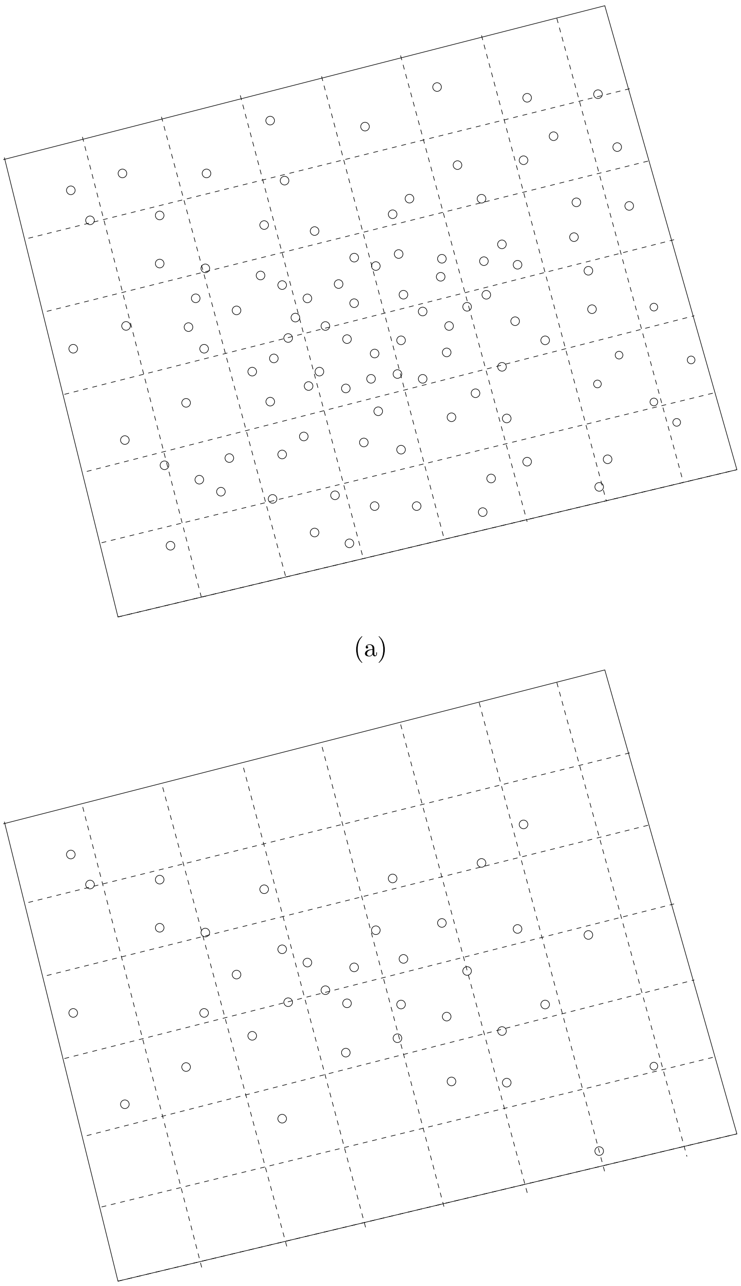
decide for the number of mobile devices collecting measurements, the number of measurements, the method and the time at which they are taken, as well as the estimator formula. The complexity arises from the movement of the mobiles, and is increased by spatial autocorrelation (observations in nearby locations 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. 4. 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 (Fig. 1). Statistical methods to perform an accurate estimate in this setting are the main topic of this paper.

We attempt to tackle this difficult 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 increases, for a moderate number of strata. We also show that systematic sampling, which is known to usually be more accurate than other spatial sampling techniques [26], 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 [3, 14] – 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 and vary significantly over small distances [31]. 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 [23]. In Sect. 6 we construct a simple model of the UHI effect and evaluate the examined sampling techniques when estimating the average temperature in the area.

The remaining parts of the paper are as follows. In Section 2 we discuss works in



(a)

(b)

Figure 1: Hypothetical examples of mobile measurements in a closed area under a certain partition, with a higher concentration of measurements towards the center: (a) Dense measurements covering the whole area, (b) Sparse measurements, leaving some subareas empty.

peer-to-peer and sensor networks which have similar statistical objectives, although they cannot be directly imported in mobile crowdsensing scenarios. Section 3 provides some basic results on spatial sampling that we use in the paper, covering uniform, stratified and systematic sampling. Section 4 discusses an appropriate estimate when there exists a measurement dataset that densely covers the area, and shows when this estimate can approach the true average. Estimates at a random instant of mobile locations are discussed in Section 5, including non-stratified sampling, stratified sampling with different stratification weights, and systematic sampling. Properties of the stratification method with weights proportional to stratum areas, which corroborate the worthiness of the method, are presented in Section 5.4. The application scenario that is used for evaluating the methods is explained in Section 6, along with the definition and spatial node distribution of the RWP Model. Section 7 presents numerical results for the absolute bias, as well as the bias reduction that can be achieved with the stratification method for a wide range of test cases and configuration parameters, such as the number of mobiles, the number of strata, and the pattern of change of the environmental parameter. In Section 8 a summary of the most important conclusions is presented. The paper ends in Section 9 with an extended discussion of open research issues, that are worth investigating both to advance the theory and to proceed to a real implementation.

2 Related work

Related lines of work concern the estimation of aggregate quantities in peer-to-peer and sensor networks. In [21], a random walk method was employed to gradually construct an estimate of a sum function, by exploiting peer-to-peer communication. In [15], Kempe et al. analyzed simple gossip-based protocols for the computation of sums, averages, random samples and quantiles. When using uniform gossip, the protocols converge exponentially fast to the true values. Uniform random sampling from a peer-to-peer network was also considered in [7] using the Metropolis-Hastings (M-H) algorithm. A modified version of the M-H algorithm was used in [30], that keeps a record of active neighbors to cope with churn (peers randomly joining and leaving the network).

A weighted random walk was used in [17] for estimating population averages. Population units are seen as nodes in a graph, which are partitioned into a set of non-overlapping categories. The appropriate sample size from each category is decided based either on proportional or on Neyman allocation. Then a weighted random walk is performed, with the weights defined in such a way so as to visit each category with a frequency roughly proportional to its appropriate sample size. Thus an accurate estimate of the population average can be obtained following the paradigm of the stratified sampling technique.

The above research does not explicitly address the estimation of area statistics, although the techniques themselves could be used for such purpose (i.e. population values could be values of environmental parameters recorded by the nodes). On this respect, some of the existing research in sensor networks is closer to our context. Roughly, the approach for computing aggregates in sensor networks is to connect to one of the sensors (called the root node), propagate a query through this node to the whole network and process the responses to estimate the desired aggregate quantity. The work in [6] focused on algorithms for dealing with node and link failures, as well as with duplicate values returned from a query that can seriously disturb the computation of sums and averages. The bias that can be introduced in the computation of aggregates due to non-uniform sensor locations was recognized in [10], where the authors proposed the use of spatial interpolation of data to resample data more uniformly in an area. In [2], the authors explored the problem of uniformly sampling sensor nodes in an area, when the location of the sensors is unknown. They considered partitioning the area into Voronoi cells, each cell corresponding to a sensor. The idea was to sample an area uniformly and get the required environmental parameter information from the sensor that is closest to that point. They also proposed to use rejection-based sampling to under-sample large areas and over-sample smaller ones, in order to approximate uniform sampling. Finally, the authors in [29] proposed aggregation techniques for computing more complex functions, such as quantiles, most frequent data values, range queries, as well as approximate distribution values.

The above works on sensor networks focused on general statistics of node values, and not spatial averages. The only work we know that focused explicitly on calculating spatial averages is [9]. The proposed method employs a Voronoi tessellation of the sensor area and performs a weighted nodal average, weighting each node by the area assigned to it as a Voronoi cell. The authors considered both centralized and distributed implementations, corresponding to the way Voronoi cells are constructed. In the centralized implementation, they considered two versions of their method, one with periodic, and one with one-time queries (which they also call “snapshots”). They showed that the accuracy in the estimation of the spatial average is significantly improved, compared to the simple nodal average.

One could imagine the application of distributed query techniques in a mobile or ad-hoc network, by exploiting opportunistic encounters or short-range communication between mobiles. However, all these techniques 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 then harder to achieve due to node mobility. Further, the location at which the measurements are taken must be taken into consideration when estimating spatial averages of environmental values. This paper addresses such important factors by introducing spatial sampling techniques.

3 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

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

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: $\bar{T} = \frac{1}{L}\sum_{i=1}^{L} \bar{T}_i$.

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 $\mu = E[T(\mathbf{x})]\ \forall \mathbf{x}$.

In uniform random sampling, an unbiased estimate of the sampling error variance $E[\bar{T} - \tilde{T}(A)]^2$ is given by s^2/n, where s^2 is the sample variance. In stratified random sampling, when the number of elements in each stratum is greater or equal to 2, an unbiased estimate of error is derived as $\bar{s}^2/n$, where $\bar{s}^2$ is the average within stratum variance. On the contrary, no unbiased estimate is available in the case of systematic sampling, or stratified random sampling with one element per stratum.

Despite the difficulty in producing unbiased estimates in systematic sampling, Ripley [28] 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 [5] and [26].

4 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

[3] Note that 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 [28, Section 3.1]

that in Proposition 1. But before we 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

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

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,

[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.

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 [22, Section 5.3.1.]). Denoting by X_i the number of balls into bin i, $i = 1 \dots L$, we have:

$$\Pr\{X_i = k\} \approx \frac{1}{k!} \binom{n}{L}^k e^{-n/L} = \frac{1}{k!} \lambda^k e^{-\lambda} \ .$$

Therefore, the probability that each subarea is non-empty is, in the limit, $1 - e^{-\lambda}$; hence 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.

5 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 [12].

5.1 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

$$E[T_m] = \sum_c T_c P(A_c) \ . \tag{3}$$

To derive an estimate of the area temperature, 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

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

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 $Var(T_m)/k$.

5.2 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:

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

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

where n_h is the number of users from each stratum h, $h = 1 \dots L$, and $\hat{T}_{w,h} = (T_{m_1} + \dots + T_{m_{n_h}})/(n_h)$ 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

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

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.

5.2.1 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,

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

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

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

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

$$f_{X|h}(x) = \frac{f_X(x)}{\int_{A'_h} f_X(x)dx} \ . \tag{8}$$

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

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

$$\begin{aligned}E[\hat{T}^n_{st}] &= \frac{1}{n}\left(\sum_{h=1}^{L} n \int_{A'_h} f_X(x)dx \left(\sum_c T_c \int_{A_{h,c}} f_{X|h}(x)dx\right)\right)\\ &= \sum_{h=1}^{L} \int_{A'_h} f_X(x)dx \left(\sum_c T_c \int_{A_{h,c}} f_{X|h}(x)dx\right)\\ &= \sum_{h=1}^{L}\sum_c T_c \int_{A_{h,c}} f_X(x)dx\\ &= \sum_c T_c \int_{A_c} f_X(x)dx = E[T_m]\ . \qquad (9)\end{aligned}$$

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

5.2.2 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:

$$E[\hat{T}^s_{st}] = \sum_{h=1}^{L} E\left[\frac{a(A'_h)\mathbb{1}_{A'_h}\hat{T}_{w,h}}{\sum_{j=1}^{L} a(A'_j)\mathbb{1}_{A'_j}}\right] \qquad (10)$$

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 (10)) is approximately independent of the estimate $a(A'_h)\mathbb{1}_{A'_h}\hat{T}_{w,h}$ in any of the strata.[7] Furthermore, using the first-degree Taylor series approximation[8]

$$E\left[\frac{1}{\sum_{j=1}^{L} a(A'_j)\mathbb{1}_{A'_j}}\right] \approx \frac{1}{E[\sum_{j=1}^{L} a(A'_j)\mathbb{1}_{A'_j}]} \qquad (11)$$

we have that

$$E[\hat{T}^s_{st}] \approx \sum_{h=1}^{L} \frac{a(A'_h)P_{ne}(A'_h)}{\sum_{j=1}^{L} a(A'_j)P_{ne}(A'_j)} E[\hat{T}_{w,h}]\ , \qquad (12)$$

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 $(i = 1 \dots n_h)$ in stratum h, with common expectation $E[T_{m|h}]$, then

$$E[\hat{T}_{w,h}] = E\left[\frac{T_{m_{1|h}} + \cdots + T_{m_{n_h|h}}}{n_h}\right] = E[n_h]E\left[\frac{T_{m_{i|h}}}{n_h}\right] \approx E[T_{m|h}] \qquad (13)$$

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

[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.

by applying Wald's equation and subsequently using the approximation $E[1/n_h] \approx 1/E[n_h]$.

$$\begin{aligned} E[T_{m|h}] &= \sum_c T_c \int_{A_{h,c}} f_{X|h}(x)dx \\ &= \sum_c T_c \int_{A_{h,c}} \frac{f_X(x)}{P(A'_h)} dx \\ &= \sum_c T_c \frac{P(A_{h,c})}{P(A'_h)} \ . \end{aligned} \tag{14}$$

From (12),(13),(14):

$$E[\hat{T}^s_{st}] \approx \sum_{h=1}^{L} \frac{a(A'_h)P_{ne}(A'_h)}{\sum_{j=1}^{L} a(A'_j)P_{ne}(A'_j)} \sum_c T_c \frac{P(A_{h,c})}{P(A'_h)} \ . \tag{15}$$

5.3 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 $1 \dots k$ 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):

$$\hat{T}^s_{sy} = \sum_{h=1}^{kL} \frac{a(A'_h)\mathbb{1}'_{A'_h}\hat{T}_{w,h}}{\sum_{j=1}^{kL} a(A'_j)\mathbb{1}'_{A'_j}} \ , \tag{16}$$

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 (12), we have for the expected value of the estimate:

$$E[\hat{T}^s_{sy}] \approx \sum_{h=1}^{kL} \frac{a(A'_h)P_{ne}(A'_h)/k}{\sum_{j=1}^{kL} a(A'_j)P_{ne}(A'_j)/k} E[\hat{T}_{w,h}] \ . \tag{17}$$

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.*

5.4 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 to provide insight for the results that follow. We consider that all strata are of equal size, 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. Then, 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 5.1 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 mobiles 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. 2. 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

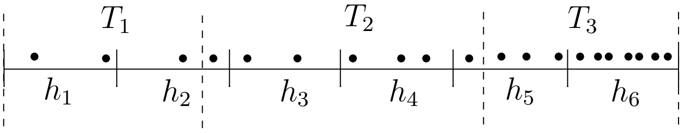


Figure 2: A realization of the mobiles' positions in a region with three discrete environmental parameter values and a stratification into 6 strata

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 close 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:[9]

$$E[\hat{T}^s_{st}] - \tilde{T} = \frac{1}{L}(B^{h_1} + \cdots + B^{h_L}) \ . \tag{18}$$

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

$$E[\hat{T}_w] - \tilde{T} = E[T_m] - \tilde{T} = B^{h_1}P(h_1) + \cdots + B^{h_L}P(h_L) \ , \tag{19}$$

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.

The concentration of mobiles in some areas, which largely causes the bias, is only reflected in (19), and not in (18). 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.

6 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 surface 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.

[9] A more general approximate expression for the bias would be produced by replacing L with the mean number of non-empty areas and B^{h_i} by a r.v. that equals B^{h_i} if the stratum is non-empty, and zero otherwise.

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 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 [3]:

$$f_X(x) = -\frac{3}{4x_m^3}x^2 + \frac{3}{4x_m} \quad \text{for } -x_m \leq x \leq x_m \ . \tag{20}$$

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. 3).

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.) [31].

The general cross-section of the typical UHI effect described in [25] 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 is depicted in Fig. 3 (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 urban and rural areas respectively. The mean value of the temperature in the area, which we are trying to estimate, is

$$\tilde{T} = T_u\frac{x_u}{x_m} + T_{sub}\frac{x_r - x_u}{x_m} + T_r\frac{x_m - x_r}{x_m} \ . \tag{21}$$

7 Numerical results

Supposing each mobile user follows a one-dimensional RWP mobility model with pdf given in (20), the probability of a mobile to be in a subarea $[a, b]$ corresponding to a step

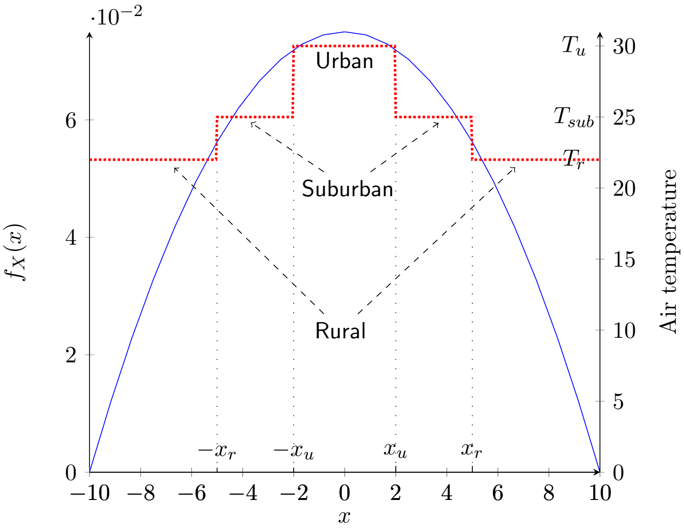


Figure 3: Plot of the pdf of the one-dimensional RWP model (x_m=10), together with a simple model of an UHI (red densely dotted line).

of the environmental parameter function is

$$P([a,b]) = \int_a^b \left(-\frac{3}{4x_m^3}x^2 + \frac{3}{4x_m}\right) dx$$
$$= -\frac{1}{4}\frac{b^3 - a^3}{x_m^3} + \frac{3}{4}\frac{b-a}{x_m} \,. \tag{22}$$

We first present some basic performance evaluation results based on the model in Fig. 3. A plot of the relative bias $(E[\hat{T}_w] - \tilde{T})/\tilde{T}$, for different values of x_r, x_u ($x_r = 0 \ldots x_m$, $x_u = 0 \ldots x_r$) and $x_m = 10$ is shown as a percentage value in Fig. 4. An interesting observation is that the bias is maximized when $x_u = x_r = 5.77$, that is when there is only a single change in temperature (and thus the temperature change in the area is more abrupt). The bias becomes zero at the extreme points, that is when there is no change at all in the temperature.

Continuing the scenario shown in Fig. 3, 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. 3, for $L = 2$ we divide into two symmetric subareas, each of which produces the same estimate.

As shown in Fig. 5, 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 5.4.

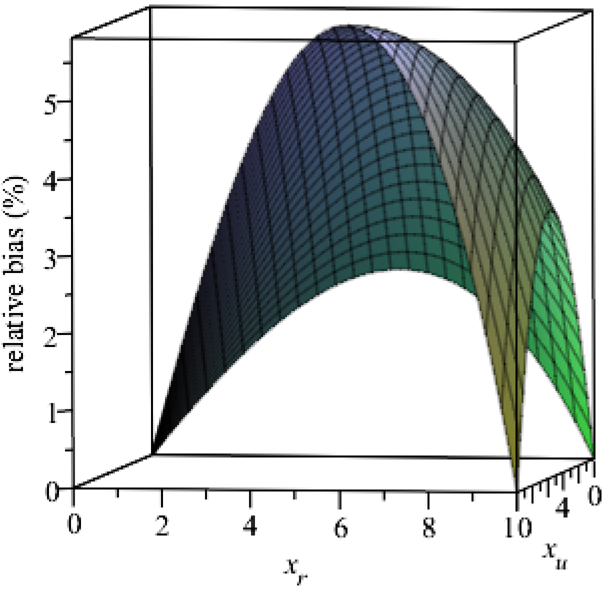


Figure 4: Example plot of the bias function with variables x_r, x_u (T_u=30, T_{sub}=25, T_r=22, x_m=10). The maximum value is observed for $x_u = x_r = 5.77$

Notice also that a maximum reduction exists for some intermediate value of L, which reflects the trade-off discussed in Section 5.2 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. 5 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 5.3 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. Thus, systematic sampling is not so appropriate for this setting because of the probability of selecting empty strata that distort the estimate.

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.

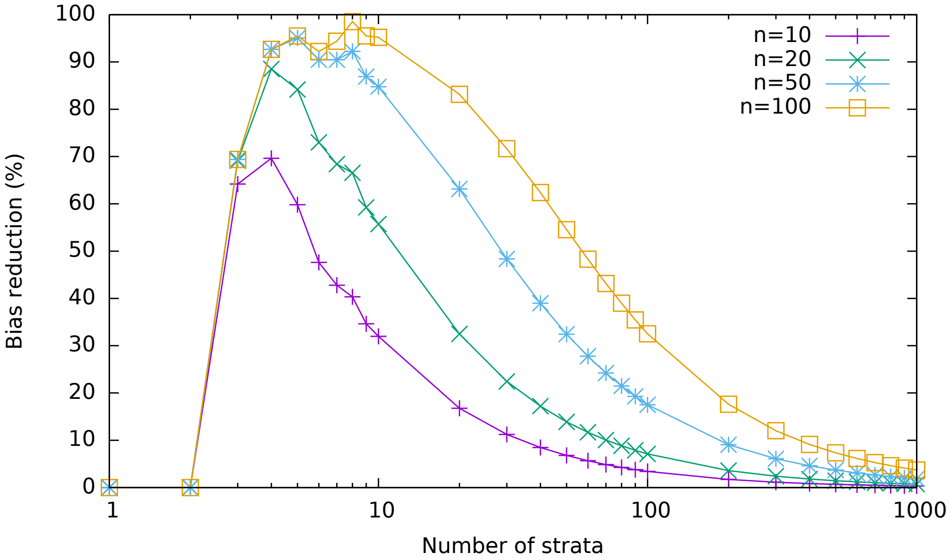


Figure 5: Bias reduction of the stratification method when weighting proportionally to stratum areas. The 3-step function shown in Fig. 3 is used for the environmental parameter.

In reality, the temperature in a region exhibiting the UHI effect changes much less abruptly than the three-step function we examined so far. A hypothetical example of a more realistic change pattern is shown in Fig. 6. As shown in the figure, a stepwise function can still be used to model such a change pattern.

In an attempt to model this pattern, we consider a generalized step function as shown in Fig. 7. We consider C unequal steps of the environmental parameter step function. The relative lengths of the steps 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. The subareas corresponding to each step are symmetric with respect to the center of the area (hence C is always an odd number) and larger subareas appear toward the edges, similarly to the function in Fig. 3. 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 subareas, with a fixed increment equal to $2(T_{max} - T_{min})/(C-1)$.[10] Notice that even for a real change pattern where we fit a step function, a fixed increment can always be

[10] Consider numbering the subareas as $1 \dots C$ from left to right. Then if $T(i)$ denotes the parameter value of subarea $T(i)$, $T(1) = T_{min}$ and $T(i) = T(i-1) + 2(T_{max} - T_{min})/(C-1)$ for $i = 2 \dots (C-1)/2$. Further, $T((C+1)/2) = T_{max}$ and $T(i) = T(C-i+1)$ for $i = (C+3)/2 \dots C$.

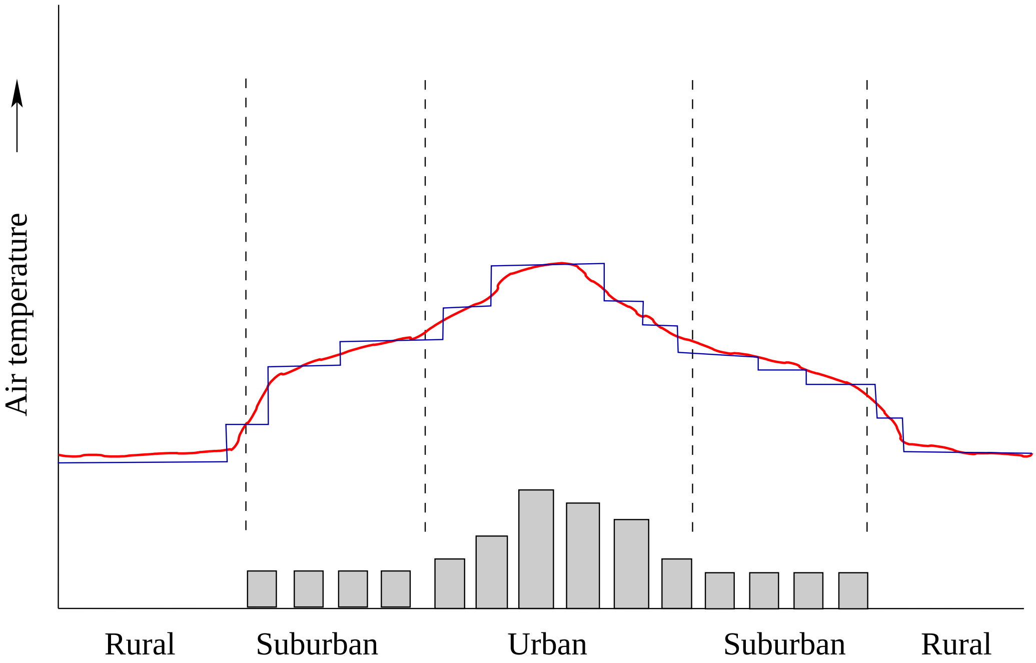


Figure 6: Hypothetical example (based on [31]) of a more realistic change pattern in a region exhibiting the UHI effect and fitting step function.

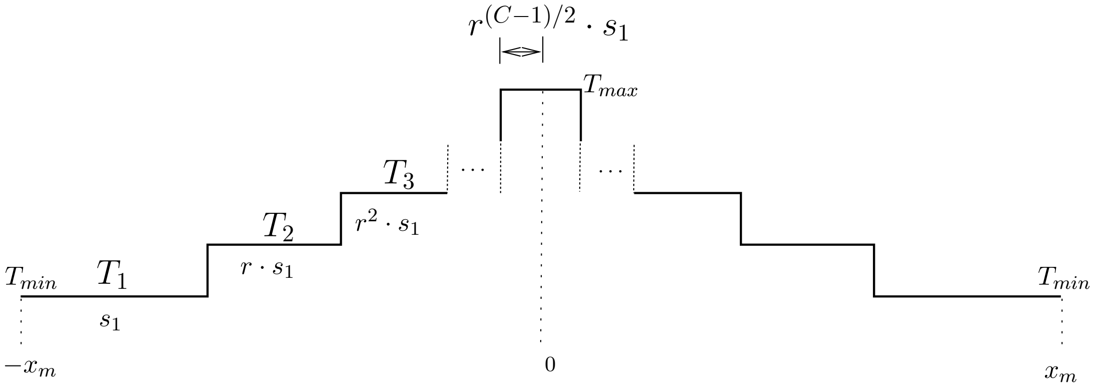


Figure 7: Generalized step function for modeling the UHI effect (C: number of steps, r: ratio of geometric series).

achieved, by appropriately modifying the step lengths.

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 independent of the actual value of x_m. This follows directly from (22) 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 (22) 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 (15) are also independent of x_m.

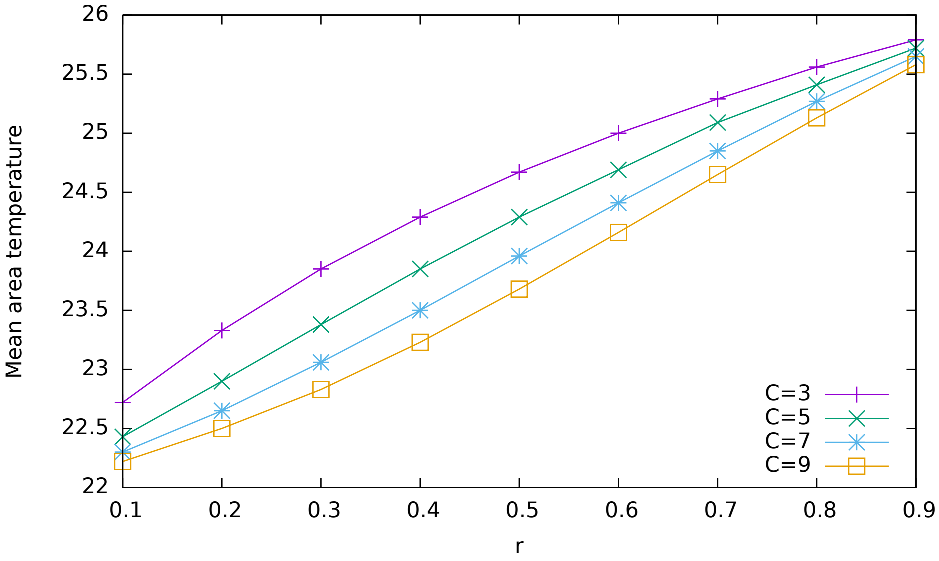
In our evaluation, we will vary both the number of steps, as well as the relative lengths of the steps in the environmental parameter function. First, to better understand this setup, we show in Fig. 8 the actual mean area temperature and the absolute value of the bias of the method without stratification, for different values of r and C. The number of mobiles is $n = 20$, while $T_{min} = 22$ and $T_{max} = 30$. As the subarea lengths become more uniform (r increases), the mean area temperature increases in Fig. 8a. This happens because the length of the inner subareas – which have the highest values – increases and they get to weigh more in the mean. At the same time, while higher temperatures exist farther from the center, there are less mobiles there to record them. Thus we observe in Fig. 8b that the bias of the non-stratification method also increases. On the other hand, for the same value of r the mean area temperature decreases as the number of steps C increases, as a result of the decreasing lengths of the steps towards the center (which have the highest values). Higher than average temperatures are then recorded by a smaller number of mobiles and the bias decreases.

Results for the bias reduction under the stratification method with varying C and r are shown in Fig. 9. Different values for the number of strata L have been examined. The remaining parameter values are as in Fig. 8 ($n = 20$, T_{min}=22, T_{max}=30).

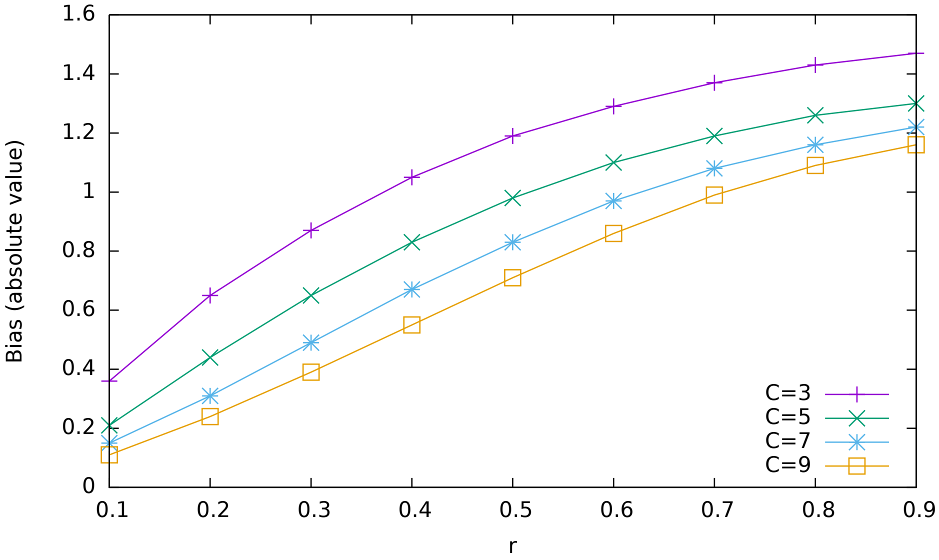
It can be observed that a larger bias reduction occurs for decreasing r. Hence, stratification helps to reduce the bias even more than the reduction we witnessed in Fig. 8b. Furthermore, the results show that the bias reduction is approximately constant as the number of steps increases, except when there is a very small number of steps. Therefore, for all but very small values of C, the bias when using stratification is approximately a constant fraction of the bias without stratification. An exception to this can be seen for a close to uniform distribution and very small number of strata (r=0.9 and L=3 in subfigure 9c), where as C increases the bias with the stratification method decreases more rapidly that the bias without stratification.

The fluctuations for small C depend on the match between the set of subareas corresponding to the steps of the environmental parameter function and the set of strata. For example, in subfigure 9a, for $C = 3$, $L = 3$, the two sets almost coincide, and the bias reduction approaches 100%. On the other hand, in subfigure 9c, for $C = 3$, $L = 3$, the points where the environmental parameter changes are {-10, -4.74, 4.74, 10}, while the end points of the strata are {-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 strata, this effect is mitigated. For illustration, we also show in Fig. 10 the absolute values of the bias with and without stratification, for $r = 0.7$.

The optimal number of strata is for all examined cases small and does not depend significantly on the relative lengths of subareas. As we see in Fig. 9, the optimal number is $L = 4$ for $r = 0.5$ and $L = 5$ for higher r, while the highest examined value of L yields the lowest bias reduction. One might have anticipated that, as C increases, a larger L

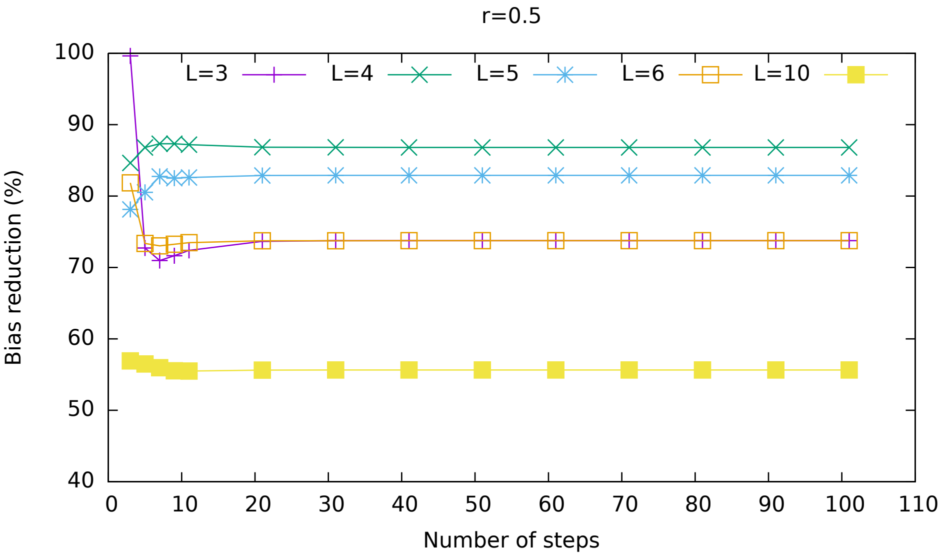


(a)

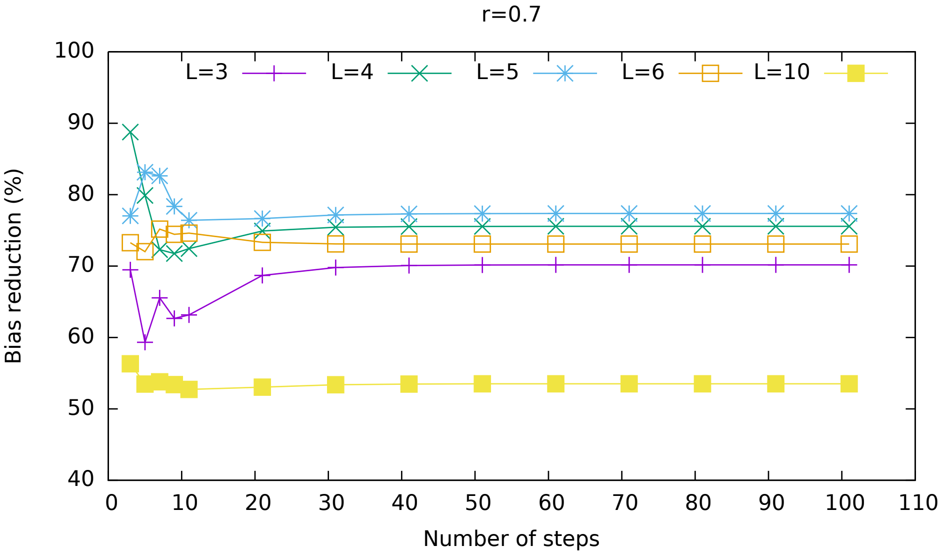


(b)

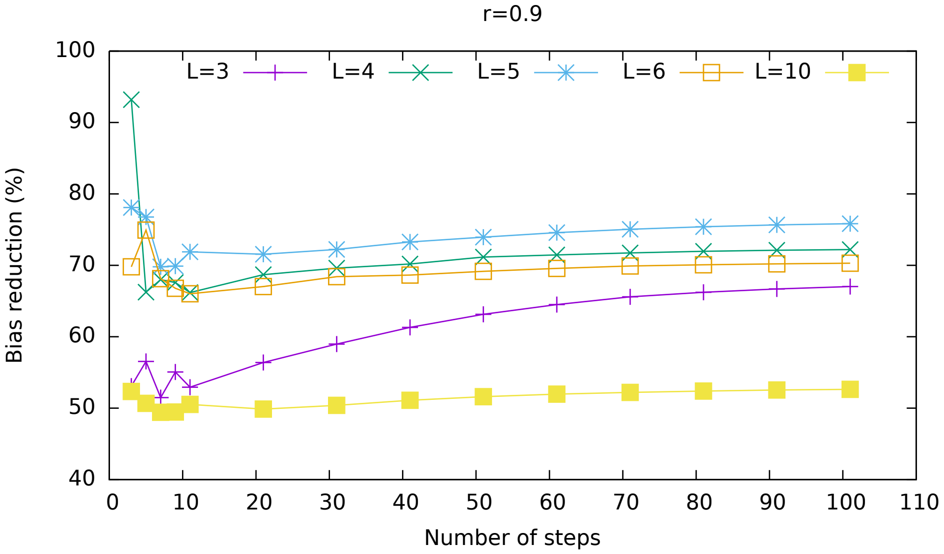
Figure 8: Actual mean area temperature (a) and bias of the non-stratification method (b) as the subarea lengths become more uniform (r increases), for different number of steps C of the environmental parameter function ($n = 20$, T_{min}=22, T_{max}=30)


r=0.5
Bias reduction (%)
Number of steps
L=3
L=4
L=5
L=6
L=10
40
50
60
70
80
90
100
0
10
20
30
40
50
60
70
80
90
100
110

(a)


r=0.7
Bias reduction (%)
Number of steps
L=3
L=4
L=5
L=6
L=10
40
50
60
70
80
90
100
0
10
20
30
40
50
60
70
80
90
100
110

(b)



(c)

Figure 9: Percentage of bias reduction with respect to the non-stratification case for different numbers of steps of the environmental parameter function and different number of strata ($n = 20$, T_{min}=22, T_{max}=30).

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. 11. We consider the same setup as previously, i.e. the subarea limits are defined by a geometric series and temperatures in each subarea are increasing with a fixed increment. Notice that, since both the average estimates and the actual mean value of the environmental parameter are obtained as normalized weighted sums, the bias values depend only on the difference $T_{max} - T_{min}$. Moreover, the temperature at each step is a linear function of $T_{max} - T_{min}$ and the bias increases linearly with greater temperature ranges. (See Fig. 11a. 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. 11b).

8 Conclusions

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

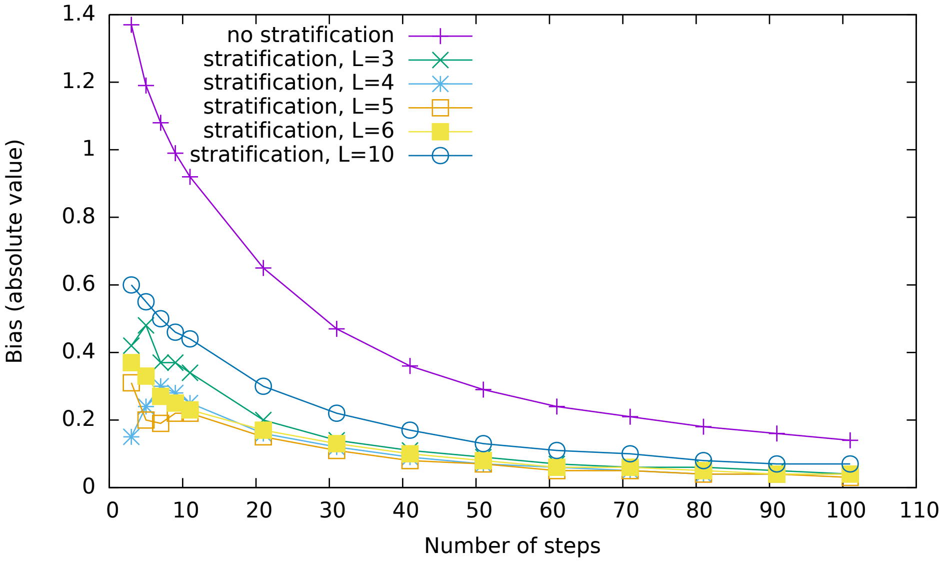


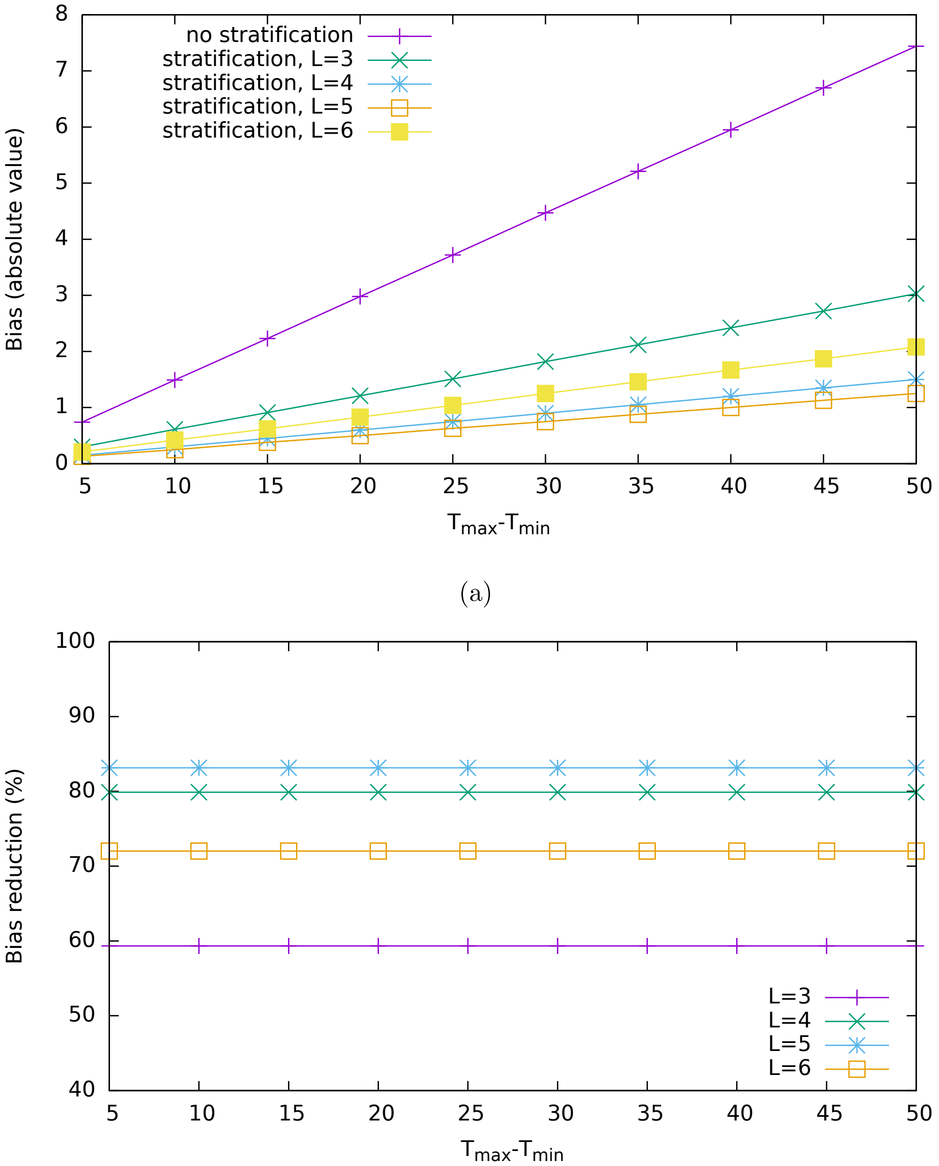
Figure 10: Absolute values of the bias with and without stratification for increasing number of steps of the environmental parameter function ($n = 20$, $r = 0.7$, $T_{min} = 22$, $T_{max} = 30$)

small number of mobiles. Furthermore, the method can get arbitrarily close to the true average as the number of mobiles increases.

The greatest bias reduction by the method is achieved for small 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 very convenient, as increasing the number of strata would increase the processing cost, and thus the overall cost of the method.

Furthermore, the number of strata could be chosen fairly independent of the change pattern of the environmental parameter within the region of interest. We derived numerical results for a wide range of patterns of the step function of the environmental parameter, including the number of steps and their relative length, as well as the range of values of the parameter in the area, and they all achieved the maximum bias reduction for a small number of strata. Although the setting for the evaluation was tailored to the pattern of temperature change in an area that exhibits the Urban Heat Island effect, this should have wider applicability, since it reflects the property that for a large number of strata the stratification estimate deteriorates and approaches the estimate without stratification.

We have also shown that systematically sampling every kth stratum starting from a random stratum, until L strata have been chosen, is equivalent, on the average, to stratified sampling with kL contiguous strata. Therefore, since the bias tends to increase for a larger number of strata, systematic sampling is less efficient than randomized sampling, unless the savings (in messaging and processing cost) by sampling only the selected areas can outweigh the performance deterioration.



(a)

(b)

Figure 11: 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)

Other conclusions, although not backed up by a theoretical analysis, seem to also be far-reaching. For example, as was shown in Fig. 8b, it can be intuitively anticipated that when a large number of mobiles is concentrated in the extremes of the area where the parameter also achieves its extreme values, then smaller bias occurs if these extreme steps (in the stepwise function) are smaller. Furthermore, that as the step lengths are less uniform (decreasing r in Fig. 9) or, equivalently, when the change in the parameter function is more abrupt, there are more gains by stratifying the area. These conclusions provide additional insight on the relationship between the environmental parameter pattern and the derived estimate with or without stratification.

Overall, the evaluation with different patterns of the environmental parameter function showed significant bias reduction in all cases. Furthermore, the achieved bias reduction is independent of the difference range in the parameter values and is stable for a large number of steps of the stepwise function.

9 Open research issues

In this last section, we discuss some open issues that would require further research, both to advance the theory and to proceed to a real implementation of the method.

9.1 Theoretical issues

An open 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. It would be interesting to see, for example, if the variance of such estimates is always smaller than that of the estimate without stratification. Unfortunately, it has not been possible to derive accurate results for the variance of the stratification estimates and compare them with the variance of the non-stratification estimate in (4). Attempting to calculate the variance of $\hat{T}^n_{st}$, $\hat{T}^s_{st}$ leads us to a complex analysis which, if simplified through approximations similar to (12) results in large inaccuracies.

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 [14] for a general convex area, and simpler approximate expressions for square and circular areas in [3]. The theoretical analysis extends to two dimensions in a straightforward manner for different shapes of strata and subareas of constant parameter value, although the computation becomes much more difficult.

Extending the results to the case of continuous varying parameters in space would be interesting when there are relevant functions available that model how the parameters vary. Otherwise, the discrete analysis suffices since the true parameter value would only be known by measurements on a set of discrete subareas. The estimate formulas and the analysis for calculating the expected estimate values extends easily using integral formulas, however the proofs for the asymptotic performance are more intricate.

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 based on the random snapshot. But this is a very complicated issue that depends on the number of collected measurements, their temporal pattern, the mobility model, and the sampling technique.

9.2 Implementation issues

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 and sending them to a central unit for processing. We consider that there are a lot of solutions available, each of which would deserve a thorough analysis. A first solution we envisage would be for the mobiles to have a software installed that executes to have measurements taken at random time instants and send the results to the central unit. It is assumed that the mobile devices would be equipped with a GPS receiver or other positioning technology, so as to send their geographical coordinates at the time of measurement (along with the measurement result) to the central unit for deriving the estimate. Other solutions could involve the sending of query messages. For example, the sampling of the mobiles could be performed by sending broadcast messages from cellular base stations or WLAN access points. The mobile devices would receive the messages, execute the required measurement and transmit the result, along with their geographical coordinates to the central unit. Another solution is to exploit near-field communication capabilities of mobile devices and flood a message from a source node to other nodes in the network, so as to cover the desired area.

The important issue for the accuracy of the sampling algorithm would be the degree to which each solution results in independent sampling of the stochastic process describing the movement of the mobiles. This would be easier accomplished in the first and second solutions since the sampling instant can be programmed with more accuracy, but might also be satisfied by queries sent in a hop-by-hop fashion. Apart from the sampling accuracy, each method should also be examined regarding its applicability, efficiency and cost in order to make an appropriate choice.

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 the mobile users move in groups, or move towards attraction points or popular places. Models for better approximating human mobility have been described in [13,27], where the travel patterns of individual users have been approximated by Lévy Walks up to a certain distance. A Lévy Walk is a random walk for which the flight time and pause times follow a power law distribution. Although the distribution parameters may change depending on whether a user travels long or short distances, or due to constraints in the area that would lead to truncation phenomena, the requirements for independent sampling of the observed stochastic process that we discussed in Section 5 may still hold, since a Lévy Walk is a stationary process that is known to have ergodicity properties in some cases [20].

Non-independence of mobile movements is usually taken into account through group mobility models. Several group mobility models have been developed, for example to model cases where the mobile users organize in groups or communities according to their social relationships and move collectively in an area (e.g. [24]). Another way to represent collective movements is to consider that the mobiles move towards most popular places (see [18] for a paper that models waypoints as popular places while also incorporating realistic features of human mobility regarding flight and pause times), which also breaks the independence assumption down. Clustered user movements will skew the location distribution, so that the estimated average is farther from the true value. Nevertheless, the method with stratification can still produce a significant improvement compared to the non-stratification method. This is because, as shown in this paper, stratification has the effect of smoothing out the skewness. In fact, 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. As we concluded in the paper, the higher the skewness, the higher the expected bias reduction by the stratification method. In this way it can be justified to expect even more gains by applying the stratification method under more realistic mobility models, than under the RWP model.

Since closed-form expressions for the limiting location distribution for such complex models are very hard to derive, simulations would be required to evaluate performance. The location probabilities could be calculated as the empirical probabilities produced by real mobility traces. The degree of resolution of the traces should be decided, based on how the desired environmental parameter changes in the area.

In addition, we have not been concerned with the accuracy of single user measurements, or the effect of noise in such measurements. A recent paper by Fiore et al. [8] discusses the impact of the accuracy of single user measurements in aggregate statistics. The authors of that paper took a signal processing approach and leveraged results from signal reconstruction from sets of irregularly spaced samples. They also considered the level of noise affecting the sample. The aim there was to predict the values of the physical phenomenon at each point in an area, and not just an aggregate statistic. One major conclusion was 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. Therefore, it would be important to also 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 [16], human contact with the sensor, exhaustion gas from other vehicles, etc.

We should also note that the temperature estimation, which was shown here as an application, was chosen mostly because of the good structural properties of the UHI model, and the intuitive understanding that it offers. However, as one may conclude from some absolute bias results (e.g. in Fig. 11a), for realistic temperature ranges the actual bias may be deemed rather small (about 1-2 °C) to necessitate a very accurate measurement. Thus, using such a measurement would probably be more appropriate in cases where absolute range differences can be large over small areas, as in measurement of air pollution indicators (e.g. particle concentrations [4]), or in cases where higher precision is indeed necessary.

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 [11, 19] for basic information.

References

[1] Aleksandar Antonic, Vedran Bilas, Martina Marjanovic, Maja Matijasevic, Dinko Oletic, Marko Pavelic, Ivana Podnar Zarko, Kresimir Pripuzic, and Lea Skorin-Kapov. Urban crowd sensing demonstrator: Sense the Zagreb air. In *Software, Telecommunications and Computer Networks (SoftCOM), 2014 22nd International Conference on*, pages 423–424. IEEE, 2014.

[2] Boulat A Bash, John W Byers, and Jeffrey Considine. 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, 2004.

[3] Christian Bettstetter and Christian Wagner. The spatial node distribution of the random waypoint mobility model. *WMAN*, 11:41–58, 2002.

[4] A Chaloulakou, P Kassomenos, N Spyrellis, Philip Demokritou, and P Koutrakis. Measurements of pm 10 and pm 2.5 particle concentrations in athens, greece. *Atmospheric Environment*, 37(5):649–660, 2003.

[5] William G Cochran. Relative accuracy of systematic and stratified random samples for a certain class of populations. *The Annals of Mathematical Statistics*, pages 164–177, 1946.

[6] Jeffrey Considine, Feifei Li, George Kollios, and John Byers. Approximate aggregation techniques for sensor databases. In *Data Engineering, 2004. Proceedings. 20th International Conference on*, pages 449–460. IEEE, 2004.

[7] Soupayan Datta and Hillol Kargupta. Uniform data sampling from a peer-to-peer network. In *Distributed Computing Systems, 2007. ICDCS'07. 27th International Conference on*, pages 1–8. IEEE, 2007.

[8] Marco Fiore, Alessandro Nordio, and Carla-Fabiana Chiasserini. 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, 2013.

[9] Saurabh Ganeriwal, Chih Chieh Han, and Mani B Srivastava. Spatial average of a continuous physical process in sensor networks. In *Proceedings of the 1st international conference on Embedded networked sensor systems*, pages 298–299. ACM, 2003.

[10] Deepak Ganesan, Sylvia Ratnasamy, Hanbiao Wang, and Deborah Estrin. Coping with irregular spatio-temporal sampling in sensor networks. *ACM SIGCOMM Computer Communication Review*, 34(1):125–130, 2004.

[11] Raghu K Ganti, Fan Ye, and Hui Lei. Mobile crowdsensing: current state and future challenges. *Communications Magazine, IEEE*, 49(11):32–39, 2011.

[12] Peter Glynn and Karl Sigman. Independent sampling of a stochastic process. *Stochastic processes and their applications*, 74(2):151–164, 1998.

[13] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[14] Esa Hyytiä, Pasi Lassila, and Jorma Virtamo. Spatial node distribution of the random waypoint mobility model with applications. *IEEE Transactions on Mobile Computing*, 5(6):680–694, 2006.

[15] David Kempe, Alin Dobra, and Johannes Gehrke. Gossip-based computation of aggregate information. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 482–491. IEEE, 2003.

[16] John Krumm and Ramaswamy Hariharan. Tempio: inside/outside classification with temperature. In *Second International Workshop on Man-Machine Symbiotic Systems*, 2004.

[17] Maciej Kurant, Minas Gjoka, Carter T Butts, and Athina Markopoulou. 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, 2011.

[18] Kyunghan Lee, Seongik Hong, Seong Joon Kim, Injong Rhee, and Song Chong. Slaw: A new mobility model for human walks. In *INFOCOM 2009, IEEE*, pages 855–863. IEEE, 2009.

[19] Huadong Ma, Dong Zhao, and Peiyan Yuan. Opportunities in mobile crowd sensing. *Communications Magazine, IEEE*, 52(8):29–35, 2014.

[20] Marcin Magdziarz and Aleksander Weron. Ergodic properties of anomalous diffusion processes. *Annals of Physics*, 326(9):2431–2443, 2011.

[21] Laurent Massoulié, Erwan Le Merrer, Anne-Marie Kermarrec, and Ayalvadi Ganesh. 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, 2006.

[22] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis.* Cambridge University Press, 2005.

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

[24] Mirco Musolesi, Stephen Hailes, and Cecilia Mascolo. 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, 2004.

[25] Timothy R Oke. *Boundary layer climates.* Routledge, 2002.

[26] Maurice H Quenouille. Problems in plane sampling. *The Annals of Mathematical Statistics*, pages 355–375, 1949.

[27] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, Seong Joon Kim, and Song Chong. On the levy-walk nature of human mobility. *IEEE/ACM transactions on networking (TON)*, 19(3):630–643, 2011.

[28] Brian D Ripley. *Spatial statistics*, volume 575. John Wiley & Sons, 2004.

[29] Nisheeth Shrivastava, Chiranjeeb Buragohain, Divyakant Agrawal, and Subhash Suri. 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, 2004.

[30] Daniel Stutzbach, Reza Rejaie, Nick Duffield, Subhabrata Sen, and Walter Willinger. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Transactions on Networking (TON)*, 17(2):377–390, 2009.

[31] Janos Unger, Zoltán Sümeghy, and Judit Zoboki. Temperature cross-section features in an urban area. *Atmospheric Research*, 58(2):117–127, 2001.

[32] Frank van der Hoeven, Alexander Wandl, Betul Demir, Sophie Dikmans, Jafeth Hagoort, Marco Moretto, Pinar Sefkatli, Frans Snijder, Siriluck Songsri, Patrick Stijger, et al. Sensing hotterdam: Crowd sensing the rotterdam urban heat island. *SPOOL*, 1(2):43–58, 2014.