Reducing Multi-Hop Calibration Errors in Large-Scale Mobile Sensor Networks

Olga Saukh[*], David Hasenfratz[*], and Lothar Thiele
Computer Engineering and Networks Laboratory, ETH Zurich, Switzerland
{saukh, hasenfratz, thiele}@tik.ee.ethz.ch

ABSTRACT

Frequent sensor calibration is essential in sensor networks with low-cost sensors. We exploit the fact that temporally and spatially close measurements of different sensors measuring the same phenomenon are similar. Hence, when calibrating a sensor, we adjust its calibration parameters to minimize the differences between co-located measurements of previously calibrated sensors. In turn, freshly calibrated sensors can now be used to calibrate other sensors in the network, referred to as multi-hop calibration.

We are the first to study multi-hop calibration with respect to a reference signal (micro-calibration) in detail. We show that ordinary least squares regression—commonly used to calibrate noisy sensors—suffers from significant error accumulation over multiple hops. In this paper, we propose a novel multi-hop calibration algorithm using geometric mean regression, which *(i)* highly reduces error propagation in the network, *(ii)* distinctly outperforms ordinary least squares in the multi-hop scenario, and *(iii)* requires considerably fewer ground truth measurements compared to existing network calibration algorithms. The proposed algorithm is especially valuable when calibrating large networks of heterogeneous sensors with different noise characteristics. We provide theoretical justifications for our claims. Then, we conduct a detailed analysis with artificial data to study calibration accuracy under various settings and to identify different error sources. Finally, we use our algorithm to accurately calibrate 13 million temperature, ground ozone (O_3), and carbon monoxide (CO) measurements gathered by our mobile air pollution monitoring network.

1. INTRODUCTION

Wireless sensor networks (WSNs) are increasingly used in a wide range of application domains to gather information about the physical world. Nowadays, WSNs are embedded in real-world deployments that go beyond research prototypes. Examples include the monitoring of permafrost in high-alpine regions [5], surveillance of a heritage building in Northern Italy [10], and air pollution monitoring in urban regions [28]. All these installations have been in operation for multiple years. Such long-term deployments need to function correctly over long time periods without requiring frequent maintenance phases. In addition, in many of these deployments, good quality data is vital. Data collected by WSNs are used for adaptive lighting in road tunnels [9], data center monitoring [29], and clinical patient surveillance [11]. In these application scenarios, wrong or inaccurate sensing may lead to wrong decisions with significant societal and economic impact.

Challenges. Achieving good data quality and preserving it during the whole system lifetime is essential. However, in most application scenarios this is a very challenging task. Deployment specifics and limited budgets often constrain the choice of sensing hardware. Typical limits apply to sensor size, price, and energy consumption. Advances in sensor technology constantly widen the spectrum of phenomena that can be captured with WSNs by bringing small, cheap, and portable sensors onto the market. The downside of this trend is a reduced accuracy, precision, and reliability of many available sensors [23, 36]. One common example are sensors produced for air quality monitoring. In recent years, low-cost gas sensors ($\sim$100 dollars) appeared on the market, which are compact in size and suitable for mobile air pollution measurements. However, compared to traditional instruments, many gas sensors have a very limited accuracy, *e.g.*, [2, 43]. Furthermore, many low-cost sensors suffer from sensor drift, have limited stability, and are sensitive to changing environmental conditions [38]. Hence, low-cost sensors need to be frequently calibrated to preserve a good data quality [4, 7, 31, 34]. Manual calibration is an elaborate and time-consuming task [38, 47]. Automatic sensor calibration is essential but challenging, particularly in networks with uncontrolled mobility.

Problem statement. In the context of this work, a fundamental observation is that two measurements taken in the spatial and temporal vicinity of each other are similar. The required spatio-temporal closeness between two measurements depends on the locality of the process monitored. If two sensors fulfill the vicinity requirements, we term their meeting point a *rendezvous*. Rendezvous between calibrated and uncalibrated sensors can be used to calibrate uncalibrated sensors. These sensors in turn, can calibrate other uncalibrated sensors. We refer to this as *multi-hop sensor calibration*.

One-hop calibration, *i.e.*, the calibration between a high-quality reference signal and an uncalibrated sensor, is common practice. State-of-the-art network calibration algorithms designed for mobile sensing networks only perform one-hop calibration [34, 35]. Hence, these algorithms can only calibrate sensors, which *directly* pass by a high-quality reference sensor. This introduces a strong constraint, requiring a high density of reference sensors, which is not feasible for many application scenarios. Moreover, many large-scale monitoring systems have constrained and uncontrollable mobility patterns preventing sensors to frequently pass by a reference sensor, *e.g.*, sensors installed in private smartphones [33].

The only two pieces of work we know of that study multi-hop calibration by leveraging meeting points between uncalibrated sensors are [7, 22]. In our previous work, initial simulation results reveal a linear error accumulation with increasing network size [22]. Both previous work use ordinary least squares regression (*OLS*) to calibrate a network of sensors. *OLS* is typically used to map uncalibrated raw sensor readings to calibrated measurements if the dependency between target signal and sensor response is linear [4,30,35]. In this paper, we show that despite the linear dependency between target signal and sensor response, *OLS* is not appropriate for multi-hop calibration. We propose a novel multi-hop calibration algorithm, which distinctly outperforms existing work by considerably reducing error propagation in the network.

Contributions and road-map. We are the first to study multi-hop calibration with respect to a reference signal in detail. This work presents a thorough theoretical analysis based on the assumptions and models introduced in Sec. 2 and Sec. 3. We analyze the sources of error propagation and propose a network calibration scheme, which *(i)* considerably reduces error propagation in the multi-hop setting, *(ii)* distinctly outperforms multi-hop calibration with *OLS*, and *(iii)* compared to existing work does not require high density of reference sensors. The algorithm proposed leverages geometric mean regression (*GMR*) and one of its main properties: absence of regression dilution bias, caused by noise in sensor readings [49].

In summary, this paper makes the following contributions:

- We analyze in Sec. 4 the *OLS* line fitting method—commonly used to calibrate sensors—and introduce *GMR* line fitting, the centerpiece of our multi-hop calibration algorithm.
- In Sec. 5, we present a detailed theoretical analysis of calibration error accumulation in a network of uncalibrated sensors. We prove that multi-hop calibration with *GMR* is optimal under specific assumptions. Further, we show with simulations that *GMR* outperforms *OLS* in many scenarios.
- We show in Sec. 6 how to best choose calibration paths in dense networks, where uncalibrated sensors can calibrate themselves to possibly multiple calibrated sensors.
- We run in Sec. 7 the proposed network calibration algorithm on multiple real-world data sets, containing more than 13 million air pollution measurements. We show the benefits of calibrating real, noisy, low-cost sensors with our algorithm.

The algorithm proposed enables the multi-hop calibration of large, heterogeneous, and mobile sensor networks, *e.g.*, as found in participatory sensing scenarios [6]. Until now, this has not been possible with existing network calibration algorithms. We survey related work in Sec. 8, and conclude in Sec. 9.

2. ASSUMPTIONS AND MODELS

Let a monitoring system consist of a set of sensor nodes (or sensors) U measuring a process P. A sensor node can be either static or mobile. Each node knows its position and current time. A sensor $u \in U$ takes a time-ordered sequence of measurements and meets other sensors from time to time. There are no constraints on the sensor sampling rate and on the frequency of rendezvous between any two sensors.

Process model. Let a process of interest P exhibit a continuous measurable signal $\eta: T \times L \to D$ with time domain $T \subseteq \mathbb{R}^+$, location domain $L \subseteq \mathbb{R} \times \mathbb{R}$ defined by latitude and longitude coordinates, and domain of measurable values $D \subseteq \mathbb{R}$. We assume that the process change is upper bounded in both time and space, *i.e.*, for any two locations $(t, l), (t', l') \in T \times L$ in time and in space, it holds that

$$|\eta(t,l) - \eta(t',l')| \le \gamma(|t - t'|, |l - l'|), \qquad (1)$$

where γ is a *monotonically non-decreasing* function of temporal and spatial distances $|t - t'|$ and $|l - l'|$. The slower γ grows and the smaller the distances $|t - t'|$ and $|l - l'|$ are, the more *similar* are the values of η at the two locations (t, l) and (t', l').

Sensor model. A sensor $u \in U$ takes a sequence of measurements $\{m_u(t_i, l_i)\} \in D_u$ of a process of interest P from the domain of values D_u possibly aperiodically at discrete time-space locations $(t_i, l_i) \in T \times L, i \in \mathbb{N}$. We consider a measurement as point measurement, that is, it has no duration. A sensor u is *perfect* if at any point $(t, l) \in T \times L$ it holds that $m_u(t, l) = \eta(t, l)$. We call η the *phenomenon signal*, as opposed to the sensor measured value m_u.

Many low-cost sensors show a close to *linear* dependence between sensor values m_u and phenomenon signal η, *e.g.*, [2, 3, 43]. Thus, for all sensors $\forall u \in U$, we assume

$$\eta = \alpha_u + \beta_u m_u + e_u, \qquad (2)$$

where α_u and β_u are *calibration coefficients* of the first order polynomial and e_u is a *noise* component (or *sensor error*) with zero mean [4, 7]. It includes all variations not explained by the calibration curve [45]. We assume that the errors of any two sensors are independent. The calibration parameters α_u and β_u describe intercept (offset) and slope (gain) of the calibration curve used to map $\mathbb{R}^2 \times D_u \to D$. We *calibrate* a sensor $u \in U$ by assigning it calibration parameters. A *calibrated measurement* is then $\widehat{m}_u = \alpha_u + \beta_u m_u$.

We distinguish two types of sensors: *reference sensors* $u^* \in U^*$, which are calibrated at all times, and *low-cost sensors* $v \in U \setminus U^*$, which require periodic calibration. We require (infrequent) rendezvous between at least one reference sensor and one low-cost sensor to calibrate a network of low-cost sensors.

3. RENDEZVOUS BETWEEN SENSORS

We perform sensor calibration by leveraging rendezvous between sensors, defined as spatially and temporally close measurements of two sensors. This section introduces the main concepts.

Definition of a rendezvous. Let $u, v \in U$ be two sensors. Both sensors take measurement sequences $\{m_u\} \subset D_u$, $\{m_v\} \subset D_v$ and eventually meet, *i.e.*, some measurements are in each other's vicinity. We define the set of spatially and temporally close pairs of measurements $\Phi^{(u,v)}$ between sensors u and v within a time interval s as

$$\Phi^{(u,v)} = \{ (m_u(t_i, l_i), m_v(t_j, l_j)) \mid$$
$$(t_i, t_j \in s) \wedge (|t_i - t_j| \le \Delta t) \wedge (|l_i - l_j| \le \Delta d) \}. \quad (3)$$

The parameters Δd and Δt set temporal and spatial constraints on the required closeness of the measurements and, thus, limit the possible change of the monitored process (see Eq. (1)). The choice of Δd and Δt depends on the process of interest. For example, the change of ozone concentration over short distances is insignificant [41] in contrast to the concentration variability of fine particles [37]. We detail in Sec. 7 our choice of Δd and Δt for temperature, ozone (O_3), and carbon monoxide (CO) measurements.

The set $\Phi^{(u,v)}$ represents a *rendezvous* between sensors u and v. We refer to the measurement pairs $(m_u, m_v) \in \Phi^{(u,v)}$ as *calibration pairs* (CPs) and use them to improve sensor calibration. We use the *projection operator* "↓" to split the measurements belonging to CPs according to the sensors by which the measurements were made. For example, the measurements of rendezvous $\Phi^{(u,v)}$ can be split into the two sets $\Phi^{(u,v)}_{\downarrow u} = \{m_u\}$ and $\Phi^{(u,v)}_{\downarrow v} = \{m_v\}$ denoting the measurements of sensor u and sensor v, respectively.

A rendezvous $\Phi^{(u,v)}$ is characterized by the parameters Δd and Δt, which steer the number of CPs $|\Phi^{(u,v)}|$ in the set. There is

an important trade-off when choosing the values of the rendezvous parameters. On the one hand, high values of Δd and Δt result in a high number of CPs $|\Phi^{(u,v)}|$, potentially giving calibration possibility to a larger set of uncalibrated sensors. On the other hand, large Δd and Δt may lead to high values of γ in Eq. (1) and, thus, to poor correlation of the measurements belonging to a rendezvous. In this case, a rendezvous $\Phi^{(u,v)}$ may include many dissimilar pairs of measurements and result in high calibration errors.

Rendezvous connection graph. A *rendezvous connection graph* Γ is an undirected graph with sensors U as set of nodes and set of edges $E = \{(u,v) \mid u, v \in U, u \neq v, \Phi^{(u,v)} \neq \emptyset\}$ between nodes with co-located measurements in time interval s. We assign to each edge in E the corresponding set of measurements. We introduce set $N(u) = \{v \mid (v, u) \in E\}$ to represent the *direct neighbors* of sensor u in graph Γ, *i.e.*, comprises sensors that have rendezvous with sensor u.

Let sensor u be calibrated and consider a rendezvous connection graph Γ with edge $(u, v) \in E$. Sensor u can calibrate sensor v (denoted as $u \rightarrow v$) by leveraging rendezvous $\Phi^{(u,v)}$ to compute calibration coefficients α_v and β_v for sensor v. We denote the path in graph Γ used to calibrate a sensor as *calibration path*, *e.g.*, $u \rightarrow v \rightarrow \ldots \rightarrow p \rightarrow q$, and refer to sensor p, the immediate parent of sensor q, as *calibration parent* of q.

Let Γ be an arbitrary rendezvous connection graph constructed for a time interval s. A *multi-hop calibration algorithm* accepts Γ as input and assigns calibration coefficients to every uncalibrated sensor in Γ if the sensor is part of a connected component of the graph including at least one reference sensor.

Our multi-hop network calibration algorithm consists of two parts: a calibration method and a calibration parent selection strategy. Given a calibrated and an uncalibrated sensor connected with an edge in Γ, the *calibration method* defines how to compute calibration parameters for the uncalibrated sensor, which best map the uncalibrated raw sensor readings to calibrated measurements. The *parent selection* strategy defines how to best choose the calibration path between a reference sensor and an uncalibrated sensor in Γ. In the following, we present an in-depth theoretical analysis of multi-hop calibration, identify calibration error sources, and propose a calibration method and a calibration parent selection strategy.

4. CALIBRATION METHODS

We start by introducing the calibration metric used throughout this paper to *(i)* compare line fitting methods (Sec. 4.2), *(ii)* evaluate their ability to suppress multi-hop error propagation (Sec. 5), and *(iii)* quantify the goodness of sensor calibration (Sec. 7).

4.1 Calibration Metric

We use the root-mean-square error (*RMSE*)—a standard metric to quantify calibration errors [4, 7]—to evaluate the calibration accuracy. We compute the *RMSE* between calibrated measurements $\widehat{m}_v$ of sensor v and the corresponding phenomenon signal η as

$$RMSE_v = \left(\frac{1}{|\{\widehat{m}_v\}|} \sum (\widehat{m}_v - \eta)^2 \right)^{\frac{1}{2}}. \qquad (4)$$

We know the true phenomenon signal for every generated sensor reading, when evaluating calibration methods with artificial data. For real data sets, we derive this from data collected with high-quality instruments in the temporal and spatial vicinity of $\widehat{m}_v$.

4.2 Line Fitting Methods

Let sensor $u \in U$ be calibrated. Consider an uncalibrated sensor $v \in U$, which meets u and forms a rendezvous $\Phi^{(u,v)}$. In the
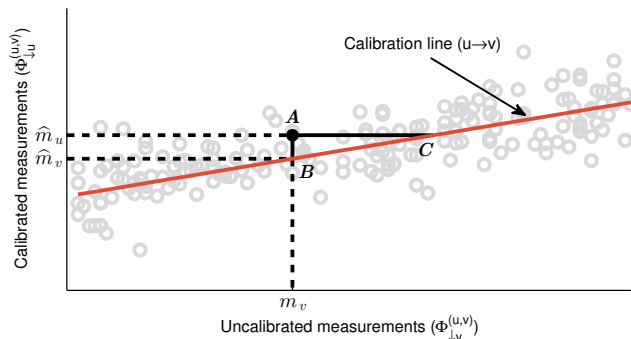

Figure 1: *OLS* minimizes the sum of squared vertical residuals *AB*. *GMR* minimizes the sum of the areas of triangles *ABC* formed by the vertical and horizontal residuals *AB* and *AC*.

following, we discuss line fitting methods that can be used to find the calibration parameters α_v and β_v of uncalibrated sensor v.

Ordinary least squares regression. Ordinary least squares (*OLS*) is a well-known standard method for line fitting, *e.g.*, [4,22,30,35]. It is *optimal* with respect to *RMSE* as error metric. *OLS* chooses the calibration parameters α_v and β_v such that the sum of squared differences between calibrated measurements $\widehat{m}_v = \alpha_v + \beta_v m_v$ of sensor v and measurements of calibrated sensor u is minimized:

$$\sum_{(\widehat{m}_u, m_v) \in \Phi^{(u,v)}} \left((\alpha_v + \beta_v m_v) - \widehat{m}_u \right)^2$$

$$= \sum_{(\widehat{m}_u, m_v) \in \Phi^{(u,v)}} \left(\widehat{m}_v - \widehat{m}_u \right)^2 \rightarrow \min. \qquad (5)$$

Hence, *OLS* minimizes the sum of squared vertical residuals, as depicted with distance AB in Fig. 1.

OLS treats the measurements of uncalibrated sensor v and the calibrated sensor u as independent and dependent variables, respectively. Due to this distinction, two different lines regress v on u (denoted as $u \rightarrow v$) and u on v (denoted as $v \rightarrow u$). Fig. 2 shows both regression lines $OLS(u \rightarrow v)$ and $OLS(v \rightarrow u)$ for a sample set of measurements in $\Phi^{(u,v)}$. The *true parameters* denote the used calibration parameters for sensor v to generate the artificial data. As stated in Eq. (5), we regress v on u to compute the slope $\beta_{OLS}^{u \rightarrow v}$ of the calibration curve (above denoted in short form as β_v):

$$\beta_{OLS}^{u \rightarrow v} = \frac{cov(\widehat{\Phi}_{\downarrow u}^{(u,v)}, \Phi_{\downarrow v}^{(u,v)})}{var(\Phi_{\downarrow v}^{(u,v)})}. \qquad (6)$$

The slope depends on $cov(\widehat{\Phi}_{\downarrow u}^{(u,v)}, \Phi_{\downarrow v}^{(u,v)})$, which is the covariance between calibrated measurements of u and uncalibrated measurements of v at rendezvous $\Phi^{(u,v)}$ and on $var(\Phi_{\downarrow v}^{(u,v)})$, which is the variance of the latter. The covariance is independent of sensor noise e_u and e_v as they are independent variables (*i.e.*, $cov(e_u, e_v) = 0$, $cov(\Phi_{\downarrow u}^{(u,v)}, e_v) = 0$, and $cov(\Phi_{\downarrow v}^{(u,v)}, e_u) = 0$). However, the variance in the denominator in Eq. (6) depends on sensor noise e_v of the uncalibrated sensor v because

$$var(\Phi_{\downarrow v}^{(u,v)}) = var(\Phi_{\downarrow v'}^{(u,v')}) + var(e_v), \qquad (7)$$

where $\Phi_{\downarrow v'}^{(u,v')}$ denotes the noise-free measurements of sensor v. Hence, sensor noise e_v introduces a bias towards zero in the computed slope $\beta_{OLS}^{u \rightarrow v}$. The greater the noise of sensor v, the stronger the estimated slope approaches zero instead of the true slope. This is known as *regression attenuation* or *regression dilution* [8, 18].
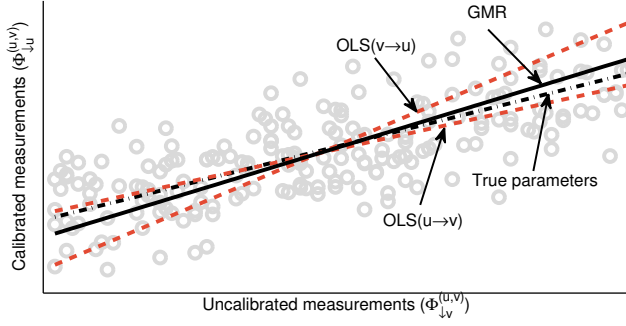
Figure 2: Using *OLS*, two different lines regress *u* on *v* and *v* on *u*. With *GMR* there is only one line as it is symmetric. The regression *OLS* ($u \to v$) is optimal with respect to the *RMSE* when calibrating sensor *v* using sensor *u*.

It is in particular a problem if sensors are calibrated over multiple hops as the bias towards zero increases with every hop. We discuss this in detail in Sec. 5.

A number of methods tackle the regression dilution problem by compensating the bias in slope estimates [18, 21]. Many of these methods assume that the variance of the sensor noise is known in advance, *e.g.*, Deming regression [13] requires that the ratio of the sensors' noise variances is known. This is particularly difficult with real sensors, which can change their characteristics over time, *e.g.*, due to changing environmental conditions and aging effects. Using imprecise sensor noise estimates for compensating the bias may lead to worse results than not compensating at all [44].

Below, we present geometric mean regression, a line fitting method that does not suffer from regression dilution and does not require any knowledge on the sensor noise. It is frequently used in natural sciences (*e.g.*, astronomy and biology) and is known under different names, such as reduced major axis [45], line of organic correlation [26], Strömberg's impartial line [15], and geometric mean regression [39]. The latter name is used throughout this work.

Geometric mean regression. Geometric mean regression (*GMR*) minimizes the sum of the areas of triangles formed by the deviation of a point from the regression line in both axis directions. We depict such a triangle in Fig. 1, encapsulated by the lines AB, AC, and BC. The point A has coordinates $(\widehat{m}_u, m_v) \in \Phi^{(u,v)}$. The calibration $\widehat{m}_v = \alpha_v + \beta_v m_v$ maps raw measurements m_v to calibrated values $\widehat{m}_v$. Due to noise in sensor readings $\widehat{m}_v \neq \widehat{m}_u$. We can convert the calibrated values $\widehat{m}_u$ of sensor u to the domain of raw measurements of sensor v using $-\frac{\alpha_v}{\beta_v} + \frac{1}{\beta_v}\widehat{m}_u$. Thus, the sum of areas minimized by *GMR* is

$$\sum_{(\widehat{m}_u, m_v) \in \Phi^{(u,v)}} \left((\alpha_v + \beta_v m_v) - \widehat{m}_u \right) \left(m_v - (\frac{1}{\beta_v}\widehat{m}_u - \frac{\alpha_v}{\beta_v}) \right)$$

$$= \frac{1}{\beta_v} \sum_{(\widehat{m}_u, m_v) \in \Phi^{(u,v)}} \left(\widehat{m}_v - \widehat{m}_u \right)^2 \to \min. \quad (8)$$

The slope $\beta_{GMR}^{u \to v}$ of the calibration curve (above denoted in short form as β_v) is

$$\beta_{GMR}^{u \to v} = \left(\frac{\beta_{OLS}^{v \to u}}{\beta_{OLS}^{u \to v}} \right)^{\frac{1}{2}} = \left(\frac{var(\widehat{\Phi}_{\downarrow u}^{(u,v)})}{var(\Phi_{\downarrow v}^{(u,v)})} \right)^{\frac{1}{2}}. \quad (9)$$

The proof of this result is given in [49]. We show in Fig. 2 the *GMR* line for a sample set of measurements. It is always located between the regression lines $OLS(v \to u)$ and $OLS(u \to v)$.
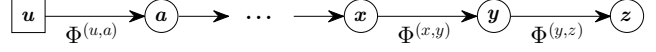

Figure 3: Calibration path $u \to a \to \ldots \to x \to y \to z$. The sensors $a, \ldots, x, y, z$ are calibrated hop-by-hop starting from the reference sensor u.

Both *OLS* and *GMR* are least squares methods, in that they optimize the sum of squared residuals using different definitions for residuals. However, the following three properties make *GMR* more attractive in the context of multi-hop sensor calibration:

- *Symmetry.* Regressing $v \to u$ and $u \to v$ yields the same regression line (*i.e.*, $\beta_{GMR}^{u \to v} = \frac{1}{\beta_{GMR}^{v \to u}}$), since switching the axes does not affect the areas of triangles to be minimized. Thus, *GMR* does not make a distinction between dependent and independent variables.
- *Scale-invariance.* The *GMR* line is invariant to any linear transformation of any of the variables, *i.e.*, scaling measurements of either sensor leads to an equivalent regression equation [27, 40].
- *No bias towards zero.* In contrast to *OLS*, *GMR* does not suffer from regression dilution. This is essential to reduce calibration error propagation in the network.

Next, we investigate the error accumulation of the two line fitting methods discussed above. We investigate the worst-case scenario when noisy sensors are calibrated along a line topology with one reference sensor at one end of the line. We show that under the given assumptions *GMR* does not suffer from error accumulation and leads to accurate network calibration.

5. REDUCING ERROR ACCUMULATION

In the following, we provide a theoretical analysis of calibration error accumulation over multiple hops using *OLS* and *GMR*. We emphasize our findings with detailed simulation results. Our main findings are:

- *OLS* suffers from significant multi-hop error accumulation due to sensor noise.
- In contrast to *OLS*, *GMR* is optimal and does not suffer from any multi-hop error accumulation if the variances $var(\Phi_{\downarrow p}^{(p,q)})$, $\forall p \in U$ are accurately estimated.
- In the general case, *GMR* suffers from error accumulation but at a distinctly lower rate than *OLS*.

5.1 Theoretical Analysis

We continue with a detailed theoretical analysis.

Setting. Assume a calibration path $u \to a \to \ldots \to x \to y \to z$ starting with a reference sensor u, as depicted in Fig. 3. The line topology presented in the plot enforces sensor calibration hop-by-hop from sensor u to sensor z using the available intermediate rendezvous $\Phi^{(u,a)}, \ldots, \Phi^{(x,y)}, \Phi^{(y,z)}$. All sensors but u and z have two direct neighbors. For example, sensor y has two neighbors $N(y) = \{x, z\}$, and the projections $\Phi_{\downarrow y}^{(x,y)}$ and $\Phi_{\downarrow y}^{(y,z)}$ are sets of measurements of sensor y taken in the vicinity of its direct neighbors x and z, respectively. In the general case, we can not control sensor mobility patterns and can not choose when sensors meet. Thus, the measurements in both sets have different means and variances. We use this setup to analyze multi-hop error accumulation of the previously discussed line fitting methods.

OLS multi-hop error accumulation. We analyze the calibration error of sensor z, which receives its calibration over multiple hops starting from reference sensor u. We derive the slope of the calibration curve $\beta_{OLS}^{u\to z}$ using Eq. (2) and Eq. (6) as

$$
\beta_{OLS}^{u\to z} = \frac{cov(\widehat{\Phi}_{\downarrow y}^{(y,z)}, \Phi_{\downarrow z}^{(y,z)})}{var(\Phi_{\downarrow z}^{(y,z)})}
$$

$$
= \frac{cov(\alpha_{OLS}^{u\to y} + \beta_{OLS}^{u\to y} \cdot \Phi_{\downarrow y}^{(y,z)}, \Phi_{\downarrow z}^{(y,z)})}{var(\Phi_{\downarrow z}^{(y,z)})}
$$

$$
= \beta_{OLS}^{u\to y} \cdot \frac{cov(\Phi_{\downarrow y}^{(y,z)}, \Phi_{\downarrow z}^{(y,z)})}{var(\Phi_{\downarrow z}^{(y,z)})} = \frac{cov(\Phi_{\downarrow u}^{(u,a)}, \Phi_{\downarrow a}^{(u,a)})}{var(\Phi_{\downarrow a}^{(u,a)})} \cdots
$$

$$
\cdots \frac{cov(\Phi_{\downarrow x}^{(x,y)}, \Phi_{\downarrow y}^{(x,y)})}{var(\Phi_{\downarrow y}^{(x,y)})} \cdot \frac{cov(\Phi_{\downarrow y}^{(y,z)}, \Phi_{\downarrow z}^{(y,z)})}{var(\Phi_{\downarrow z}^{(y,z)})}. \tag{10}
$$

The covariances are independent of sensor noise, as discussed in the previous section. However, the denominators with the variances of all sensors along the calibration path depend on sensor noise, as denoted in Eq. (7). Every additional sensor in the calibration path, introduces an additional term in the denominator, which increases the bias towards zero. The total introduced bias is given by

$$
\frac{var(\Phi_{\downarrow a'}^{(u,a')})}{var(\Phi_{\downarrow a}^{(u,a)})} \cdots \frac{var(\Phi_{\downarrow y'}^{(x,y')})}{var(\Phi_{\downarrow y}^{(x,y)})} \cdot \frac{var(\Phi_{\downarrow z'}^{(y,z')})}{var(\Phi_{\downarrow z}^{(y,z)})}, \tag{11}
$$

where $\Phi_{\downarrow a'}^{(u,a')}$ denotes the noise-free measurements of sensor a. In the following, we show that *GMR*, in contrast to *OLS*, does not accumulate errors over multiple hops under certain conditions.

GMR multi-hop error accumulation. We use the line topology above to show that, in contrast to *OLS* fitting, *GMR* does not suffer from error accumulation if all rendezvous sets of a sensor have the same variance. For example, we require that the measurements of sensor y in Fig. 3 have the same variance in the rendezvous involving sensor x and sensor z, i.e., $var(\Phi_{\downarrow y}^{(x,y)}) = var(\Phi_{\downarrow y}^{(y,z)})$. We prove our statement by showing that the *GMR* calibration of a sensor is independent of the calibration path. Later, we release the above assumption of equal variances and analyze error propagation in the general case. We show that the error obtained is distinctly lower than with *OLS*.

THEOREM 1. *Let us consider a rendezvous connection graph Γ with a calibration path $u \to a \to \ldots \to x \to y \to z$. Let u be a reference sensor. If $var(\Phi_{\downarrow r}^{(p,r)}) = var(\Phi_{\downarrow r}^{(r,q)})$, $\forall p, q \in N(r)$, $\forall r \in U$, then the calibration of sensor z is independent of the path between reference sensor u and sensor z.*

PROOF. It suffices to show that the statement of the theorem holds for the calibration parameter β_z, as this uniquely determines the value of α_z. We use Eq. (2) and Eq. (9) to derive the calibration slope $\beta_{GMR}^{u\to z}$ of sensor z, calibrated over multiple hops starting from reference sensor u:

$$
\beta_{GMR}^{u\to z} = \left(\frac{var(\widehat{\Phi}_{\downarrow y}^{(y,z)})}{var(\Phi_{\downarrow z}^{(y,z)})} \right)^{\frac{1}{2}}
$$

$$
= \left(\frac{var(\alpha_{GMR}^{u\to y} + \beta_{GMR}^{u\to y} \cdot \Phi_{\downarrow y}^{(y,z)})}{var(\Phi_{\downarrow z}^{(y,z)})} \right)^{\frac{1}{2}}
$$

$$
= \beta_{GMR}^{u\to y} \cdot \left(\frac{var(\Phi_{\downarrow y}^{(y,z)})}{var(\Phi_{\downarrow z}^{(y,z)})} \right)^{\frac{1}{2}}
$$

$$
= \left(\frac{var(\Phi_{\downarrow u}^{(u,a)})}{var(\Phi_{\downarrow a}^{(u,a)})} \cdots \frac{var(\Phi_{\downarrow x}^{(x,y)})}{var(\Phi_{\downarrow y}^{(x,y)})} \cdot \frac{var(\Phi_{\downarrow y}^{(y,z)})}{var(\Phi_{\downarrow z}^{(y,z)})} \right)^{\frac{1}{2}} \tag{12}
$$

$$
= \left(\frac{var(\Phi_{\downarrow u}^{(u,a)})}{var(\Phi_{\downarrow z}^{(y,z)})} \right)^{\frac{1}{2}}. \tag{13}
$$

We leverage in Eq. (13) the assumption of equal variances, i.e., $var(\Phi_{\downarrow y}^{(x,y)}) = var(\Phi_{\downarrow y}^{(y,z)})$. Since the calibration slope of z only depends on the measurements of sensor u and z, i.e., on $var(\Phi_{\downarrow u}^{(u,a)})$ and $var(\Phi_{\downarrow z}^{(y,z)})$, it is independent of the choice of intermediate sensors along the calibration path. $\square$

If the assumption of the theorem holds, calibration parameters calculated with *GMR* are independent of the choice of calibration paths and are *optimal*, since the minimization function of *GMR* defined in Eq. (8) achieves its minimum.

Relaxing the assumption. The theorem assumption implies that the variance of the phenomenon signal at rendezvous between any pair of sensors must be the same. In reality this is hard to achieve, since meeting points between sensors cannot be enforced. Therefore, in many real systems the reduction used from Eq. (12) to Eq. (13) does not apply and, thus,

$$
\beta_{GMR}^{u\to z} = \left(\frac{var(\Phi_{\downarrow u}^{(u,a)})}{var(\Phi_{\downarrow z}^{(y,z)})} \cdot \prod_{\substack{\forall p \to r \to q \\ p,q \in N(r)}} \frac{var(\Phi_{\downarrow r}^{(r,q)})}{var(\Phi_{\downarrow r}^{(p,r)})} \right)^{\frac{1}{2}}. \tag{14}
$$

In multi-hop calibration, *GMR* does not continuously introduce a bias in one direction (like *OLS* towards zero), since both numerator and denominator contain the sensor noise term (see Eq. (7)):

$$
\frac{var(\Phi_{\downarrow r}^{(r,q)})}{var(\Phi_{\downarrow r}^{(p,r)})} = \frac{var(\Phi_{\downarrow r'}^{(r',q)}) + var(e_r)}{var(\Phi_{\downarrow r'}^{(p,r')}) + var(e_r)}. \tag{15}
$$

This observation is essential to understand the small error propagation property of *GMR*. *OLS* continuously underestimates the sensor slopes, and, hence, introduces an ever-increasing bias towards zero. In contrast, *GMR* closely approaches the true slope (sometimes underestimating and other times overestimating it) and, thus, does not accumulate a bias towards one direction.

Next, we show through extensive simulations that *GMR* achieves significantly lower calibration errors than *OLS*.

5.2 Simulation Results

In the following, we use a line topology graph Γ and artificially generated rendezvous Φ between pairs of adjacent sensors to highlight the differences between *OLS* and *GMR*. Generated data correspond to typical carbon monoxide (CO) concentrations in urban environments. Artificial data generation gives us three advantages: we *(i)* can freely choose the topology of the rendezvous connection graph, *(ii)* can precisely control sensor characteristics (e.g., noise level, calibration parameters), and *(iii)* know for every generated sensor reading the true phenomenon signal value, which allows us to accurately evaluate calibration errors.

Setup. All generated line topologies comprise 21 sensors, with a noise-free reference sensor at the beginning of the line, followed by 20 noisy sensors. The sensor IDs denote the hop distance to the reference sensor. A rendezvous $\Phi^{(u,u+1)}$ between a pair of sensors u and $(u+1)$ in Γ is constructed by uniformly sampling 1000 phenomenon signal values in the range $[0.2, 2]$ ppm. This range is typical for carbon monoxide concentrations in urban areas.
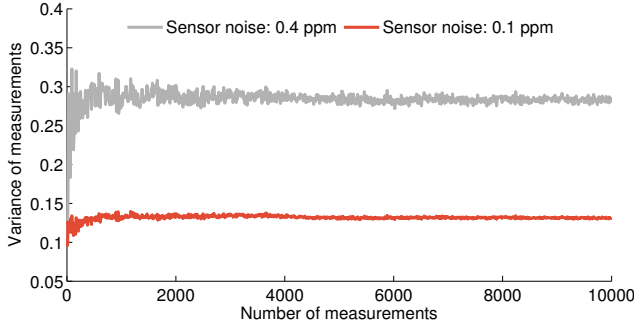
Figure 4: Impact of measurement quantity and sensor noise on the variance estimates. High noise and low number of measurements impede the accurate estimation of the variance.

Figure 6: Despite the 2x higher noise of sensor 8, *GMR* accurately calibrates sensors 9–20. With *GMR* a sensor's calibration is mostly independent of the calibration parent's sensor noise.
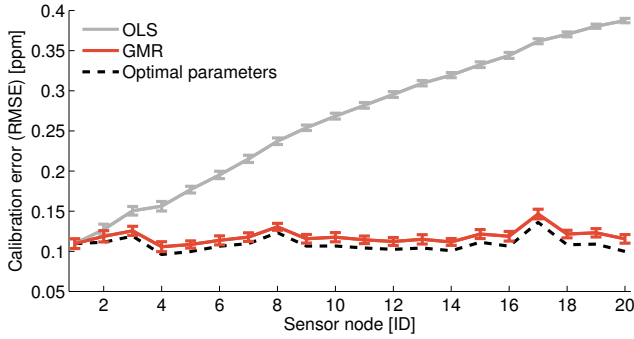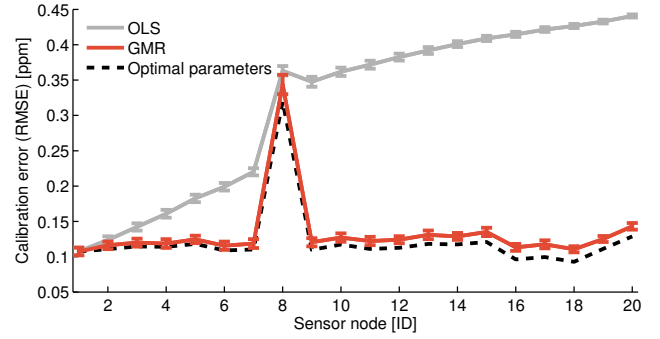

Figure 5: Fixed phenomenon range (Theorem 1 holds). *GMR* can accurately calibrate all sensor nodes of a 21-node line topology with very small error propagation.
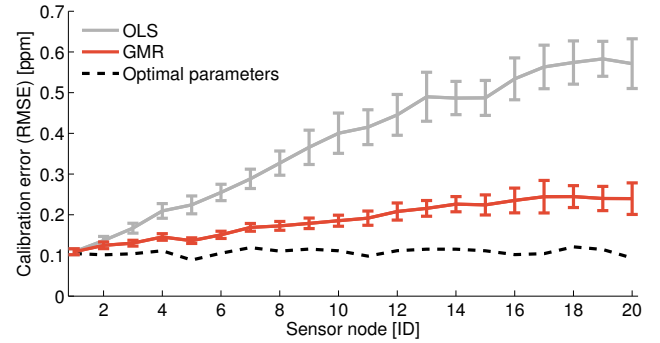

Figure 7: Variable phenomenon range (Theorem 1 does not hold). *GMR* has distinctly slower error accumulation than *OLS* when calibrating sensors on a 21-node line topology.

We randomly sample for each sensor u the calibration parameters $\alpha_u \in [0, 1]$ and $\beta_u \in [0.3, 2]$. We use these parameters to convert the generated phenomenon signal values to sensor-specific raw measurements and add sensor noise e_u to get uncalibrated sensor measurements m_u. The variance of the sensor noise is randomly sampled from $[0, 0.2]$ ppm for each sensor. The *RMSE* between phenomenon signal and calibrated sensor measurements is used to quantify calibration errors. All plots show average *RMSEs* over 100 randomly generated line topologies. For each sensor, *optimal parameters* are obtained by computing *OLS* regression between raw sensor data (including noise) and the known phenomenon signal. The optimal parameters represent the best calibration parameters achieving the minimum possible *RMSE* between phenomenon signal and calibrated sensor measurements.

Impact of sensor noise. We need to know the variances of the sensor measurements to compute the sensors' calibration slopes, as described in the previous section. We estimate the variances based on rendezvous Φ between pairs of sensors. However, both *(i)* low number of measurements in rendezvous and *(ii)* sensor noise make an accurate variance estimation difficult. We exemplify this in Fig. 4 using artificial data. We show the fluctuation of the variance estimates for two noisy measurement sets depending on the number of measurements in the set. As expected, variance estimation exhibits less fluctuation as the number of measurements increases. Moreover, we observe that higher sensor noise leads to higher uncertainties in estimating the variance.

In Fig. 5, we compare the calibration errors of *OLS* and *GMR* assuming that the values of the phenomenon signal for all rendezvous are sampled from the same fixed interval (the statement of Theorem 1 holds). Since the assumption of Theorem 1 holds, in theory *GMR* should not accumulate any calibration errors. With noisy sensor data, however, we observe an insignificant error accumulation due to small inaccuracies in estimating the variance of the measurements (for node 20 there is a small difference between optimal parameters and *GMR*). Nevertheless, the *RMSE* of *GMR* is close to the *RMSE* achieved with optimal parameters. In contrast, *OLS* suffers from rapid error accumulation. This is the result of regression dilution: due to sensor noise, the calibration slope of a sensor gets a stronger bias towards zero with every additional calibration hop.

We make the impact of noise on the calibration error more explicit by increasing the noise of sensor 8 by a factor two. Naturally, sensor 8 gets worse *RMSE* for any parameter setting, as depicted in Fig. 6. We observe that *OLS* suffers from high error increase for nodes 9–20, due to the high bias towards zero of the calibration slope of sensor 8. In contrast, the nodes calibrated by *GMR* are mostly independent of the high noise of sensor 8. There is a minor error increase due to an increased inaccuracy in estimating the variance $var(\Phi_{\downarrow 8}^{(8,9)})$ of the measurements of sensor 8.

The above analysis shows that for *GMR (i)* calibration error of a sensor mostly depends on the sensor's noise, but the noise of the calibration parent has minor impact, *(ii)* calibration error of a sensor can be smaller than the calibration error of its calibration parent,
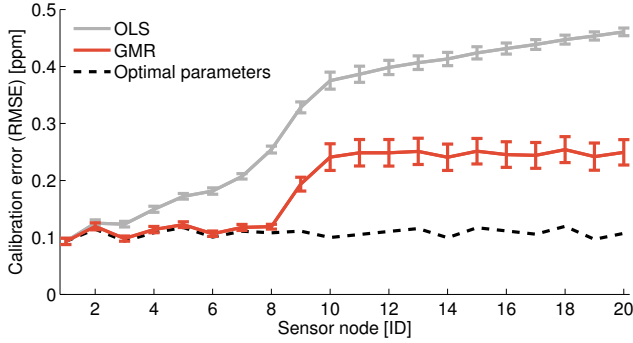
Figure 8: The 3x smaller measurement range between sensor 8 and 9 results in lower range-to-noise ratios of the two sensors, which influences the calibration error of all subsequent sensors.
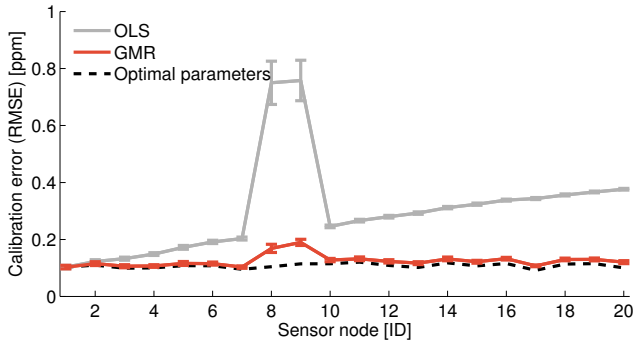

Figure 9: The 3x larger measurement range between sensor 8 and 9 results in higher range-to-noise ratios of the two sensors. This does not influence the calibration of sensors 10–20.

and *(iii)* there is almost no multi-hop accumulation of errors. The last two findings are crucial when calibrating a network of heterogeneous sensors with different noise characteristics.

Impact of phenomenon signal range. The assumption of Theorem 1 rarely holds for real data sets, because rendezvous between sensors might occur at different ranges of the phenomenon signal. In the following, we use the line topology above but randomly choose for each rendezvous the maximum signal range in $[1, 4]$ ppm. The resulting *RMSE* of each sensor is plotted in Fig. 7. Both fitting methods suffer from error accumulation as showed by the theoretical analysis, although *GMR* is less affected than *OLS*.

To better understand the source of error accumulation, we use equal phenomenon signal ranges for all rendezvous, except between sensor 8 and 9, *i.e.*, $\widehat{\Phi}^{(8,9)}$. The latter range is decreased in Fig. 8 and increased in Fig. 9 by factor three compared to the rendezvous of all other sensors. This changes the respective *range-to-noise ratio* (*RNR*). Decreasing the range lowers *RNR*, which means that the noise has an increasing impact on the sensor measurements. This, in turn, makes it more difficult to estimate the variance of the measurements (see Fig. 4) and, hence, results in a higher calibration error of *OLS* and *GMR*. Moreover, the introduced error affects any sensor that has sensor 9 in its calibration path, as Eq. (10) and Eq. (14) suggest and Fig. 8 shows. By contrast, increasing the range, raises the *RNR*. The calibration error of both fitting methods increases because the calibration of sensor 8 is based on measure-

ments from the standard (small) range. However, in the rendezvous with sensor 9, sensor 8 has to measure in a much larger (factor three) range. This introduces with *GMR* only a small error, because *GMR* can accurately estimate the true slope of sensor 8, even if it only has a small range of measurements available. However, the bias introduced by *OLS* has a large impact on the calibration error when the range is increased. Hence, increasing the range has a much larger impact on its calibration accuracy, as shown in Fig. 9.

We conclude that *(i)* varying the phenomenon signal range between rendezvous degrades the performance of both *GMR* and *OLS*, although *OLS* is significantly more affected than *GMR*, and *(ii)* a low *RNR* is particularly problematic, since the introduced calibration error propagates over multiple hops. Our findings suggest that calibration parents need to be carefully selected to minimize error propagation in the network.

6. NETWORK CALIBRATION

In this section, we present our calibration parent selection algorithm and analyze it in combination with the *OLS* and *GMR* line fitting methods on artificial topologies. Later, we evaluate their performance on real data sets.

6.1 Selecting the Calibration Parent

We showed with our previous analysis that an efficient parent selection algorithm must build on the following properties:

- Avoid calibration parents with low *RNR* to reduce multi-hop error propagation.
- Maximize number of checkpoints with calibration parents to reduce inaccuracies in estimating the variance.

Optimizing each of the above properties is not straightforward for the following reasons: Given a noisy measurement, it is impossible to split it into its noise-free and noise components, as done in Eq. (7), without additional knowledge about the sensor noise. Therefore, it is difficult to reliably identify parent candidates with low *RNR*. Similarly, despite high number of calibration pairs, we may have low correlation between paired values leading to a poor regression line. In the worst case, the phenomenon signal has the same value for all calibration pairs making them useless for sensor calibration.

We propose to *(i)* use a combination of multiple calibrated sensors as calibration parent to increase the number of measurements involved in sensor calibration, and *(ii)* analyze the Pearson correlation with the calibration parent to avoid sensors with low *RNR*.

Combining rendezvous of multiple calibrated sensors $u, \ldots, v$ and using the combination as calibration parent results in the following slope computation:

$$\beta_{GMR}^{u,\ldots,v \to z} = \left(\frac{var(\widehat{\Phi}_{\downarrow u}^{(u,z)} \cup \ldots \cup \widehat{\Phi}_{\downarrow v}^{(v,z)})}{var(\Phi_{\downarrow z}^{(u,z)} \cup \ldots \cup \Phi_{\downarrow z}^{(v,z)})} \right)^{\frac{1}{2}}. \quad (16)$$

Choosing multiple parents *(i)* maximizes the number of measurements used to compute the calibration slope, and *(ii)* potentially increases the phenomenon signal range and, thus, increases the *RNR*. The latter holds with high probability if rendezvous occur randomly at distinct values of the phenomenon signal. *RNR* is then maximized as the result of the increased phenomenon range.

To minimize the chance of having a parent with low *RNR*, we compute the Pearson correlation between measurements of different sensors, *i.e.*, between measurements in $\Phi_{\downarrow z}^{(u,z)} \cup \ldots \cup \Phi_{\downarrow z}^{(v,z)}$ and $\widehat{\Phi}_{\downarrow u}^{(u,z)} \cup \ldots \cup \widehat{\Phi}_{\downarrow v}^{(v,z)}$ of Eq. (16), and use it as noise estimate.
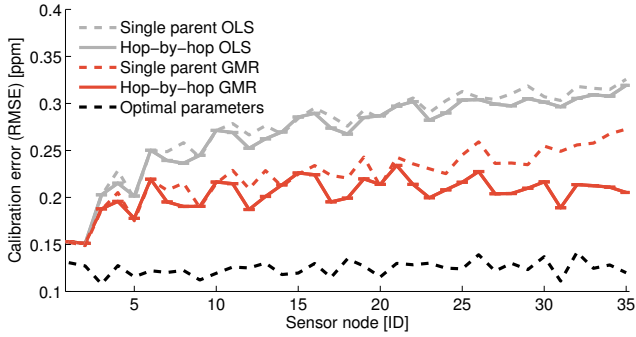
Figure 10: Comparison of *OLS* and *GMR* on a 35-node grid topology. The hop-by-hop calibration algorithms use multiple calibration parents and, thus, achieve a lower calibration error than the single parent version.

We *blacklist* rendezvous between sensors if the correlation between their measurements is below a given threshold.

Putting everything together. The last key piece is to tie all parts together.

Calibration parameters may vary over time, since low-cost sensors typically lose sensitivity over time and are influenced by environmental conditions. This has significant influence on the sensor measurements and results in the fluctuation of the calibration coefficients α_u and β_u. We introduce a periodic calibration approach by partitioning the time domain T into a sequence of equally-spaced, non-overlapping time intervals $\{s_k\} = S \subseteq T$. We assume that the calibration parameters of all sensors do not change within s_k. For each s_k, we consider all meeting points that occurred between pairs of sensors and construct a rendezvous connection graph Γ_{s_k}, which is used as input for the calibration algorithm.

Using the rendezvous connection graph Γ_{s_k}, the calibration algorithm calibrates the network of noisy sensors hop-by-hop starting from a reference sensor. The calibration algorithm starts by calibrating sensors that directly pass by reference sensors (first-hop sensors). Once calibrated, the sensors can be used as references by the second-hop sensors. The algorithm stops as soon as all sensors in the network are calibrated. We use for the calibration of a sensor $u \in U \setminus U^*$, *all* non-blacklisted previously calibrated sensors in $N(u)$ as calibration parent.

6.2 Simulation Results

We end this section by showing a comparison between hop-by-hop calibration with *OLS* and *GMR* and demonstrating the benefit of our parent selection strategy. We generate artificial data as described in the previous section, but use a grid topology in order to have multiple possible calibration parents for every sensor node. We arrange 36 nodes on a grid with 35 uncalibrated sensors and one reference node in the top left corner of the grid. The node IDs increase with increasing hop-distance from the reference sensor, *i.e.*, sensor 1 and 2 are the direct neighbors of the reference sensor.

Fig. 10 depicts the comparison between two versions of *OLS* and *GMR*: *hop-by-hop*, using the parent selection scheme described above and *single parent*, selecting as calibration parent the sensor with the maximum number of measurement pairs. We see again that *OLS* accumulates errors with a much faster rate than *GMR* and that multiple parents help to reduce the calibration error.

Figure 11: Air quality sensor node installed on top of a streetcar of the local public transport network.

Data set	Measurements [in millions]	Sampling interval [s]
Temperature	2.7	20
Ozone (O_3)	2.1	20
Carbon monoxide (CO)	8.5	10

Table 1: Evaluated data sets of the mobile air pollution monitoring network collected from March to August 2014.

7. REAL-WORLD CASE STUDY: AIR POLLUTION MONITORING

In this section, we use the proposed network calibration algorithm to calibrate a large set of air pollution measurements collected with mobile sensor nodes traversing a large urban area. Further, we demonstrate on real measurements that Hop-by-hop *GMR* considerably reduces error propagation and is able to more accurately calibrate a set of low-cost gas sensors than Hop-by-hop *OLS*.

7.1 Mobile Air Pollution Monitoring Network

Starting in 2012, we gradually equipped ten streetcars of a public transport network with air quality measurement stations, as shown in Fig. 11. The sensor nodes cover a $100 \, km^2$ urban area on a regular schedule, depicted in Fig. 12, and monitor a wide range of air pollutants and environmental parameters. The measurements are stored locally in a database and transmitted in real-time over GSM (cellular network) to the back-end server running Global Sensor Network (GSN) [1], a software middleware that facilitates data collection in sensor networks. The sensor readings are removed from the local database once their reception is acknowledged by the back-end server. All measurements are publicly available.[1]

Air quality sensor node. The core of the sensor node is a Gumstix embedded computer with a 600 MHz CPU running an embedded Linux operating system. A GPS receiver supplies the station with precise geospatial information. The station supports bidirectional communication over GSM and WiFi. GSM is used under regular system operation while WiFi is used during debugging and maintenance phases. While the streetcars are in operation, on average 20 h per day, they supply the nodes with power. During the night, typically from 1:00 AM to 5:00 AM, the streetcars are in their depots and the nodes are turned off.

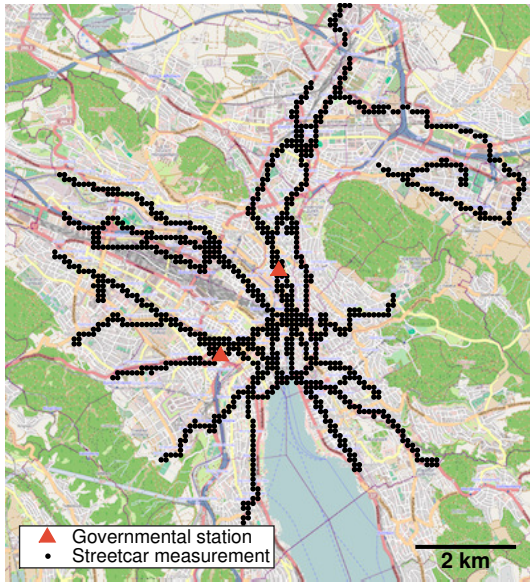[1]OpenSense data: `http://data.opensense.ethz.ch`

Figure 12: Measurement coverage of the ten sensor nodes deployed on top of ten streetcars of the public transport network. The two governmental stations provide highly accurate reference measurements.
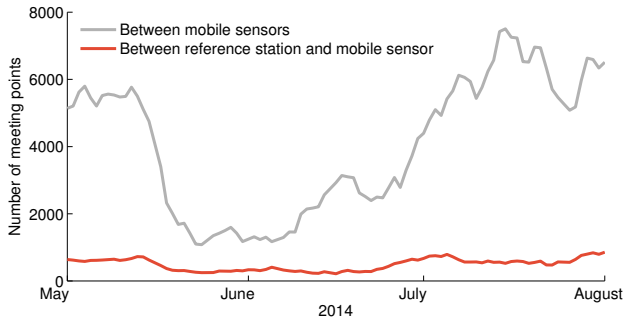

Figure 13: Number of meeting points among mobile sensors and reference stations accumulated every ten days during three months of operation.

To monitor air pollution, the sensor nodes are equipped with a semiconductor ozone (O_3) sensor [43], electrochemical carbon monoxide (CO) and nitrogen dioxide (NO_2) sensors [2, 3], and a novel compact device to measure ultrafine particle concentrations [17]. Additionally, the nodes monitor environmental parameters, such as temperature and humidity [42]. In this work, we focus on the calibration of the temperature, ozone, and carbon monoxide data sets. We use 13 million measurements collected in six months between March and August 2014, as summarized in Table 1.

7.2 Sensor Calibration

Traditionally, air pollutants are monitored by networks of fixed measurement stations, which are highly reliable and able to accurately measure many different pollutants. Two stations of the local governmental measurement network are located in the city center, as depicted in Fig. 12. The two stations are 4 m and 16 m apart from the streetcar tracks. We use their high-quality measurements to calibrate our low-cost sensors deployed on top of the streetcars. **Setup**. We use all measurements to create every ten days a rendezvous connection graph. We assume that the measurements from

Data set	Average calibration error [RMSE]	
	Single calibration	Periodic calibration
Temperature	2.3 °C	1.6 °C
Ozone (O_3)	12.9 ppb	9.8 ppb
Carbon monoxide (CO)	0.23 ppm	0.08 ppm

Table 2: Calibration errors with GMR. Single calibration, only calibrates sensors once at the beginning of the deployment, compared to periodic calibration as described in this paper.
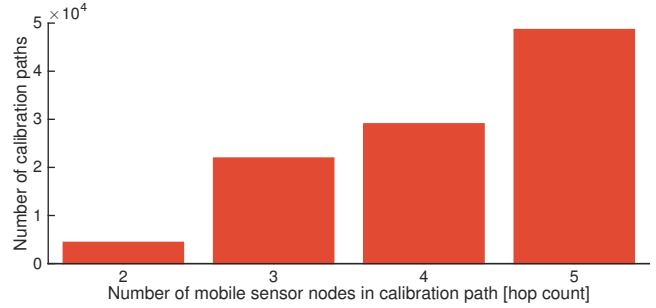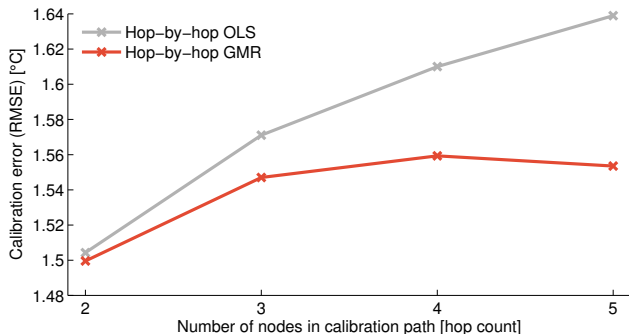

Figure 14: Number of calibration paths with 2–5 sensor nodes.

two different sensors are similar if their spatial and temporal distances are within $\Delta d = 50$ m and $\Delta t = 5$ min. We found through extensive evaluations that this parameterization achieves a good trade-off between number of meeting points and similarity of the measured phenomena. On average, within a slot size of ten days, the rendezvous connection graph of the ozone measurements comprises 4000 meeting points among mobile sensors and 500 meeting points between reference stations and mobile sensors, as shown in Fig. 13. We blacklist a rendezvous if the correlation of the sensors' measurements is below 0.5.
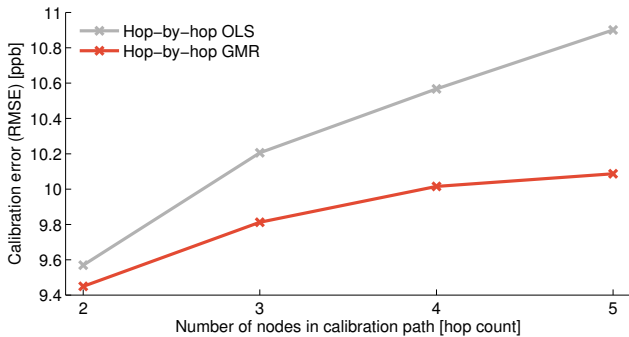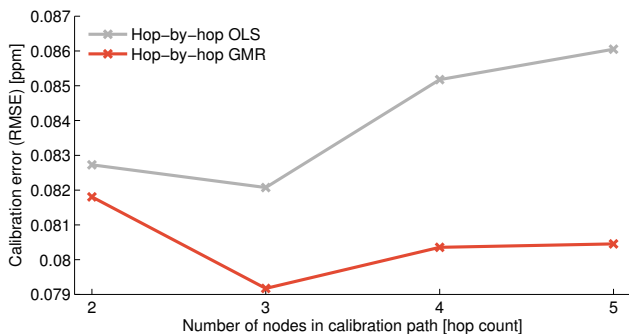
Evaluation. We evaluate the calibration accuracy with a two-fold cross-validation approach. We calibrate the network using one reference station and evaluate the calibration accuracy with the second reference station. In every time slot, we repeat this approach with both reference stations acting once as reference station. Table 2 shows the average RMSE between calibrated measurements of the mobile sensor nodes and the reference station using GMR. The achieved accuracy is reasonable for the deployed low-cost sensors and is considerably better than what is obtained if the sensors are only calibrated once at the beginning of the deployment, also listed in Table 2.

We illustrate the advantages of *GMR* by extracting specific calibration path lengths from the rendezvous connection graph. We compare the calibration errors of *GMR* and *OLS* by analyzing more than 100,000 different calibration paths composed of 2–5 sensor nodes, as shown in Fig. 14. We show in Fig. 15 that on all three pollution data sets the error accumulation of *GMR* is considerably than with *OLS*. The difference between *OLS* and *GMR* increases for longer path lengths. These results are in line with our previous findings based on theoretical analysis and simulations. However, the differences between *OLS* and *GMR* are lower than what has been seen in simulations. This is due to the following two reasons: *(i) GMR* reaches its full potential on large networks, while our network is relatively small and *(ii)* the sensor data is very noisy, hence, the RMSE is dominated by errors introduced due to noise and to a lesser extent due to regression dilution.

Finally, we compare in Table 3 the number of sensors, which can be calibrated with existing one-hop algorithms (*i.e.*, only calibrating sensors directly passing by a reference sensor, *e.g.*, [34]) to the

(a) Temperature.

(b) Ozone (O$_3$).

(c) Carbon monoxide (CO).

Figure 15: Dependency between calibration error and length of the calibration path. Hop-by-hop *GMR* has a considerably lower error accumulation than *OLS* for all three pollutants.

Data set	Sensors calibrated		Total sensors
	one-hop	multi-hop	
Temperature	4.7	8.2	9
Ozone (O$_3$)	3.8	6.2	7
Carbon monoxide (CO)	4.5	8.2	10

Table 3: Average number of sensors calibrated over a 10-day period using one-hop and multi-hop calibration algorithms.

8.1 Line Fitting

Linear regression is a common statistical operation in many sciences. Besides *OLS* and *GMR*, both discussed in detail in this work, there is a broad range of line fitting methods, such as Theil-Sen estimator [20], which chooses the median slope among all lines through pairs of points, total least squares regression [32], which optimizes the sum of squared orthogonal residuals to the regression line, and Deming regression [13], which is the maximum likelihood estimator. Each method was designed for a specific regression model, that must match application requirements. In the context of sensor network calibration, the following requirements are important: we need to *(i)* account for errors in both variables, *(ii)* handle both variables symmetrically, and *(iii)* be invariant to changes in the scale. Our choice of line fitting method and its alternatives are detailed below.

When calibrating a sensor using another noisy sensor, the chosen regression method must account for measurement errors in both variables. Standard *OLS* ignores the problem and thus suffers a noise-driven bias in estimating the slope, known as regression dilution [8]. The bias can be compensated [18, 21], but requires additional knowledge about the errors, *e.g.*, their variances or the ratio thereof [13]. This knowledge is often not available, especially if sensor noise depends on environmental conditions.

Many regression methods make a distinction between the predictor (independent) variable and the response (dependent) variable, because their optimization functions are asymmetric with respect to the two variables, *e.g.*, *OLS* and Theil-Sen estimator. Thus, these regression methods derive two different regression lines depending on the direction of regression. Using asymmetric regression methods poses a dilemma in many applications, when trying to identify a cause-effect relationship between the variables, *e.g.*, if one would want to find the dependency between body length and body mass [45]. Thus, it is often desirable to have one regression line describing the relationship rather then two. A similar problem arises in the context of sensor calibration: It is hard to fix the direction of calibration for a pair of noisy sensors of the same type. In this case, symmetric regression models, such as *GMR* and total least squares, produce a more comprehensible fit.

Lack of scale invariance makes the regression dependent on the range of raw measurements limiting its applicability for sensor calibration. For example, total least squares is not scale-invariant.

GMR is known under different names such as reduced major axis [45], line of organic correlation [26], and Strömberg's impartial line [15]. It is the only linear regression method in the two-dimensional space, which is simple (*i.e.*, only based on the ratio of the variables' variances), symmetric, and scale-invariant [40].

8.2 Sensor Calibration

A great body of work is dedicated to calibrating low-cost sensors, *e.g.*, [12,19,31,34]. The existing literature on sensor calibration can be classified into micro- and macro-calibration approaches [46,48].

number of nodes reached with our multi-hop calibration approach. Note that not all ten sensor nodes are equipped with all sensor types. We observe that our multi-hop algorithm calibrates at least 34 % more sensors than its one-hop alternative. This is remarkable, since our air pollution monitoring network is well-connected with two reference stations located in the city center. We expect an even higher benefit when using multi-hop calibration algorithms for large-scale systems, *e.g.*, in participatory sensing applications.

8. RELATED WORK

This section summarizes existing knowledge related to this work. We overview line fitting methods in the context of sensor calibration, list state-of-the-art sensor calibration algorithms, and discuss their differences to this work.

Micro-calibration. Micro-calibration algorithms calibrate every sensor in the network according to a high-quality reference signal [22, 34, 38, 47]. The goal is to have accurate absolute sensor measurements of the phenomenon monitored.

Tolle *et al.* [47] and Ramanathan *et al.* [38] calibrate their sensors in the laboratory before the deployment phase. Manual calibration is an elaborate and time-consuming task and many sensors need periodic re-calibrations to deliver measurements with a good data quality during the whole system lifetime (see Table 2).

Miluzzo *et al.* [34] propose CaliBree, a self-calibration system for mobile sensor nodes. CaliBree makes use of meeting points between low-cost sensors and high-quality reference sensors to update the calibration of low-cost sensors. Compared to the algorithm proposed in this paper, it ignores all meeting points between low-cost sensors and, thus, requires in general a high density of reference sensors to calibrate all sensors in a network (see Table 3).

In our previous work [22], we introduced a multi-hop calibration algorithm applying *OLS* regression to calibrate a sensor network. We showed with simulations a linear increase of calibration error with increasing number of nodes in the calibration path. In this paper, we show with our theoretical analysis, simulations, and real pollution measurements a considerably smaller error accumulation.

Macro-calibration. Macro-calibration algorithms try to maximize the similarity among the measurements of all sensor nodes in the network [4, 7, 16, 31, 46]. In macro-calibration, the main goal is not to adjust sensor calibration according to a reference signal but to achieve a high similarity between sensors in the network. Hence, most macro-calibration algorithms do not require access to reference measurements.

Bychkovskiy *et al.* [7] propose a two phase multi-hop calibration algorithm. In the first phase, the algorithm derives relationships between co-located sensors and in the second phase it maximizes the consistency among groups of sensor nodes. The authors assume that sensor noise can be filtered out. In general, this is very difficult to achieve in mobile sensing networks.

Balzano and Nowak [4] and Lipor and Balzano [31] introduce blind calibration algorithms to determine sensor gains by oversampling the phenomenon signal. The authors compare different methods to derive calibration parameters: singular value decomposition and *OLS* in [4], and total least squares in [31]. Blind calibration methods depend on the assumption that the mean of the monitored signal is zero (or the same for all sensors). This is in most real-world deployments not the case, *e.g.*, air pollution monitoring.

Xiang *et al.* [50] propose a collaborative sensor calibration and placement approach using meeting points between mobile sensors to adjust their calibration parameters and error estimates. The goal is to compensate for sensor drift errors. The measurements from our long-term, mobile air quality monitoring system indicate that sensor drift has a minor effect on the fluctuation of calibration parameters. In contrast, dynamically changing environmental conditions highly affect sensor performance. This observation is in line with the findings published in [38].

Other macro-calibration algorithms improve target detection [46], localization [48], and the calibration of light sensors [16].

In contrast to all above works, the distributed monitoring of physical phenomena (*e.g.*, urban noise monitoring [33], air pollution measurements [24,25], and earth-quake detection [14]) requires accurate absolute sensor measurements. Hence, in these scenarios a micro-calibration approach is needed, such as the calibration algorithm proposed in this paper. Moreover, a good micro-calibration algorithm also accomplishes the goals of macro-calibration, namely a high consistency among measurements of co-located sensor nodes, whereas the opposite is not true.

9. CONCLUSIONS

Wireless sensor networks (WSNs) are used in an increasing number of applications enabling the dense monitoring of the environment. Many sensors used in WSNs need to be frequently calibrated to constantly deliver accurate measurements.

Calibrating a whole network of sensors based on a few high-quality reference measurements is challenging. Error propagation in the network hinders the accurate calibration of all sensor nodes. In this paper, we propose a novel multi-hop calibration algorithm, which distinctly reduces error accumulation in the network. Instead of using ordinary least squares—typically employed to calibrate noisy sensors—we use geometric mean regression and exploit its robustness against regression dilution, which is caused by noisy sensor readings. This is essential in order to reduce error propagation in multi-hop calibration systems.

We show the benefits of our algorithm by performing a detailed theoretical analysis, conducting simulations under various settings, and calibrating measurements from our mobile air pollution monitoring network. We demonstrate the following key features of our calibration algorithm: *(i)* Very low hop-by-hop error propagation in the network. Hence, the algorithm can handle long calibration paths incorporating many sensors without accumulating large errors. *(ii)* The calibration quality of a sensor is largely independent of the calibration error of its calibration parent. A low-noise sensor can be accurately calibrated despite having a high-noise calibration parent. *(iii)* Due to multi-hop calibration, only a low number of reference sensors are required since meeting points between low-cost sensors are also exploited as calibration opportunities.

Hence, the algorithm proposed is suitable for the accurate calibration of large, mobile, and heterogeneous sensor networks, such as found in participatory sensing scenarios.

10. REFERENCES

[1] K. Aberer, M. Hauswirth, and A. Salehi. A middleware for fast and flexible sensor network deployment. In *VLDB*, 2006.

[2] Alphasense. CO-B4 4-Electrode carbon monoxide sensor (datasheet). `http://goo.gl/ba2dnV`, 2014.

[3] Alphasense. NO2-B4 4-Electrode nitrogen dioxide sensor (datasheet). `http://goo.gl/IpmhPb`, 2014.

[4] L. Balzano and R. Nowak. Blind calibration of sensor networks. In *ACM/IEEE IPSN*, pages 79–88, 2007.

[5] J. Beutel, S. Gruber, A. Hasler, R. Lim, A. Meier, C. Plessl, I. Talzi, L. Thiele, C. Tschudin, M. Woehrle, and M. Yuecel. PermaDAQ: A scientific instrument for precision sensing and data recovery under extreme conditions. In *ACM/IEE IPSN*, pages 265–276, 2009.

[6] J. A. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. Participatory sensing. In *ACM WSW*, 2006.

[7] V. Bychkovskiy, S. Megerian, D. Estrin, and M. Potkonjak. A collaborative approach to in-place sensor calibration. In *ACM/IEEE IPSN*, pages 301–316, 2003.

[8] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement error in nonlinear models: A modern perspective*. Chapman & Hall/CRC, 2006.

[9] M. Ceriotti, M. Corrà, L. D'Orazio, R. Doriguzzi, D. Facchin, S. Guna, G. P. Jesi, R. L. Cigno, L. Mottola, A. L. Murphy, et al. Is there light at the ends of the tunnel? Wireless sensor networks for adaptive lighting in road tunnels. In *ACM/IEEE IPSN*, pages 187–198, 2011.

[10] M. Ceriotti, L. Mottola, G. P. Picco, A. L. Murphy, S. Guna, M. Corra, M. Pozzi, D. Zonta, and P. Zanon. Monitoring heritage buildings with wireless sensor networks: The Torre Aquila deployment. In *ACM/IEEE IPSN*, pages 277–288, 2009.

[11] O. Chipara, C. Lu, T. C. Bailey, and G.-C. Roman. Reliable clinical monitoring using wireless sensor networks: Experiences in a step-down hospital unit. In *ACM SenSys*, pages 155–168, 2010.

[12] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Elsevier Sensors and Actuators B: Chemical*, 129:750–757, 2007.

[13] W. E. Deming. *Statistical adjustment of data*. 1944.

[14] M. Faulkner, M. Olson, R. Chandy, J. Krause, K. M. Chandy, and A. Krause. The next big one: Detecting earthquakes and other rare events from community-based sensors. In *ACM/IEEE IPSN*, pages 13–24, 2011.

[15] E. D. Feigelson and G. J. Babu. Linear regression in astronomy. II. *The Astrophysical Journal*, 397:55–67, 1992.

[16] J. Feng, S. Megerian, and M. Potkonjak. Model-based calibration for sensor networks. In *IEEE Sensors*, pages 737–742, 2003.

[17] M. Fierz, C. Houle, P. Steigmeier, and H. Burtscher. Design, calibration, and field performance of a miniature diffusion size classifier. *Taylor & Francis AS&T*, 45(1):1–10, 2011.

[18] C. Frost and S. G. Thompson. Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society Series A*, 163(2):173–189, 2000.

[19] J. G. Monroy, A. Lilienthal, J.-L. Blanco, J. González-Jiménez, and M. Trincavelli. Calibration of MOX gas sensors in open sampling systems based on Gaussian Processes. In *IEEE Sensors*, pages 1–4, 2012.

[20] R. O. Gilbert. *Statistical Methods for Environmental Pollution Monitoring*. John Wiley and Sons, 1987.

[21] J. Gillard. An historical overview of linear regression with errors in both variables. *TR Cardiff University*, 2006.

[22] D. Hasenfratz, O. Saukh, and L. Thiele. On-the-fly calibration of low-cost gas sensors. In *Springer EWSN*, pages 228–244, 2012.

[23] D. Hasenfratz, O. Saukh, and L. Thiele. Model-driven accuracy bounds for noisy sensor readings. In *IEEE DCOSS*, pages 165–174, 2013.

[24] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, J. Beutel, and L. Thiele. Deriving high-resolution urban air pollution maps using mobile sensor nodes. *Elsevier Journal of Pervasive and Mobile Computing*, 2015.

[25] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, and L. Thiele. Pushing the spatio-temporal resolution limit of urban air pollution maps. In *IEEE PerCom*, pages 69–77, 2014.

[26] K. A. Kermack and J. B. S. Haldane. Organic correlation and allometry. *JSTOR Biometrika*, 37:30–41, 1950.

[27] W. H. Kruskal. On the uniqueness of the line of organic correlation. *JSTOR Biometrics*, 9:47–58, 1953.

[28] J. J. Li, B. Faltings, O. Saukh, D. Hasenfratz, and J. Beutel. Sensing the air we breathe–the OpenSense Zurich dataset. In *AAAI*, pages 323–325, 2012.

[29] C.-J. M. Liang, J. Liu, L. Luo, A. Terzis, and F. Zhao. RACNet: A high-fidelity data center sensing network. In *ACM SenSys*, pages 15–28, 2009.

[30] S. Lindenthal, V. Ussyshkin, J. Wang, and M. Pokorny. Airborne LIDAR: A fully-automated self-calibration procedure. In *ISPRS*, pages 73–78, 2011.

[31] J. Lipor and L. Balzano. Robust blind calibration via total least squares. In *IEEE ICASSP*, pages 4244–4248, 2014.

[32] I. Markovsky and S. V. Huffel. Overview of total least-squares methods. *Elsevier Signal Processing*, 87:2283–2302, 2007.

[33] C. Meurisch, K. Planz, D. Schäfer, and I. Schweizer. Noisemap: Discussing scalability in participatory sensing. In *ACM SenseMine*, 2013.

[34] E. Miluzzo, N. D. Lane, A. T. Campbell, and R. Olfati-Saber. CaliBree: A self-calibration system for mobile sensor networks. In *IEEE DCOSS*, pages 314–331, 2008.

[35] S. Moltchanov, I. Levy, Y. Etzion, U. Lerner, D. M. Broday, and B. Fishbain. On the feasibility of measuring urban air pollution by wireless distributed sensor networks. *Science of The Total Environment*, 502:537–547, 2015.

[36] S. Mukhopadhyay, D. Panigrahi, and S. Dey. Model-based techniques for data reliability in wireless sensor networks. In *IEEE TMC*, pages 528–543, 2009.

[37] J. Peters, J. Theunis, M. V. Poppel, and P. Berghmans. Monitoring PM10 and ultrafine particles in urban environments using mobile measurements. *Aerosol and Air Quality Research*, 13:509–522, 2013.

[38] N. Ramanathan, L. Balzano, M. Burt, D. Estrin, T. Harmon, C. Harvey, J. Jay, E. Kohler, S. Rothenberg, and M. Srivastava. Rapid deployment with confidence: Calibration and fault detection in environmental sensor networks. *TR 62 CENS*, 2006.

[39] W. E. Ricker. Linear regressions in fishery research. *Fisheries Research Board of Canada*, 30(3):409–434, 1973.

[40] P. A. Samuelson. A note on alternative regressions. *Econometrica*, 10(1):80–83, 1942.

[41] O. Saukh, D. Hasenfratz, C. Walser, and L. Thiele. On rendezvous in mobile sensing networks. In *Springer RealWSN*, pages 29–42, 2013.

[42] Sensirion. SHT10 humidity and temperature sensor (datasheet). `http://goo.gl/LHBas`.

[43] SGX Sensortech. MiCS-OZ-47 ozone sensor (datasheet). `http://goo.gl/hT1AoM`.

[44] G. D. Smith and A. N. Phillips. Inflation in epidemiology: The proof and measurement of association between two things revisited. *BMJ*, 312(7047):1659–1661, 6 1996.

[45] R. J. Smith. Use and misuse of the reduced major axis for line-fitting. *Wiley AJPA*, 140(3):476–86, 2009.

[46] R. Tan, G. Xing, Z. Yuan, X. Liu, and J. Yao. System-level calibration for data fusion in wireless sensor networks. *ACM TOSN*, 9(3):1–27, 2013.

[47] G. Tolle et al. A macroscope in the redwoods. In *ACM SenSys*, pages 51–63, 2005.

[48] K. Whitehouse and D. Culler. Calibration as parameter estimation in sensor networks. In *ACM WSNA*, pages 59–67, 2002.

[49] E. B. Woolley. The method of minimized areas as a basis for correlation analysis. *Econometrica*, 9(1):38–62, 1941.

[50] Y. Xiang, L. Bai, R. Piedrahita, R. P. Dick, Q. Lv, M. Hannigan, and L. Shang. Collaborative calibration and sensor placement for mobile sensor networks. In *ACM/IEEE IPSN*, pages 73–84, 2012.