

ICT: In-field Calibration Transfer for Air Quality Sensor Deployments

YUN CHENG, ETH Zurich, Switzerland

XIAOXI HE, ETH Zurich, Switzerland

ZIMU ZHOU, ETH Zurich, Switzerland

LOTHAR THIELE, ETH Zurich, Switzerland

Recent years have witnessed a growing interest in urban air pollution monitoring, where hundreds of low-cost air quality sensors are deployed city-wide. To guarantee data accuracy and consistency, these sensors need periodic calibration after deployment. Since access to ground truth references is often limited in large-scale deployments, it is difficult to conduct city-wide post-deployment sensor calibration. In this work we propose In-field Calibration Transfer (ICT), a calibration scheme that transfers the calibration parameters of source sensors (with access to references) to target sensors (without access to references). On observing that (i) the distributions of ground truth in both source and target locations are similar and (ii) the transformation is approximately linear, ICT derives the transformation based on the similarity of distributions with a novel optimization formulation. The performance of ICT is further improved by exploiting spatial prediction of air quality levels and multi-source fusion. Experiments show that ICT is able to calibrate the target sensors as if they had direct access to the references.

CCS Concepts: • **Human-centered computing** → *Ubiquitous and mobile computing systems and tools*; • **Hardware** → *Sensor applications and deployments*.

Additional Key Words and Phrases: Air Pollution; Sensor Calibration Transfer

ACM Reference Format:

Yun Cheng, Xiaoxi He, Zimu Zhou, and Lothar Thiele. 2019. ICT: In-field Calibration Transfer for Air Quality Sensor Deployments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 6 (March 2019), 19 pages. <https://doi.org/10.1145/3314393>

1 INTRODUCTION

Motivation. Urban air pollution is a serious problem for public health. For example, high concentration of particulate matter (PM) with diameters less than 2.5 micron, or $PM_{2.5}$, can cause respiratory or cardiovascular diseases [3]. According to the World Health Organization (WHO), air pollution accounts for 4.2 million premature deaths per year. 91% of the world's population lives in areas where air quality levels exceed WHO limits [25].

To raise awareness and for quantitative studies, many large-scale air pollution monitoring systems have been deployed, where tens to hundreds of low-cost air quality sensors are installed across the city to measure air

Corresponding Author: Zimu Zhou.

Authors' addresses: Yun Cheng, ETH Zurich, Computer Engineering and Networks Laboratory, Zurich, Switzerland, chengyu@ethz.ch; Xiaoxi He, ETH Zurich, Computer Engineering and Networks Laboratory, Zurich, Switzerland, hex@ethz.ch; Zimu Zhou, ETH Zurich, Computer Engineering and Networks Laboratory, Zurich, Switzerland, zzhou@tik.ee.ethz.ch; Lothar Thiele, ETH Zurich, Computer Engineering and Networks Laboratory, Zurich, Switzerland, thiele@tik.ee.ethz.ch.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2019 Association for Computing Machinery.

2474-9567/2019/3-ART6 \$15.00

<https://doi.org/10.1145/3314393>

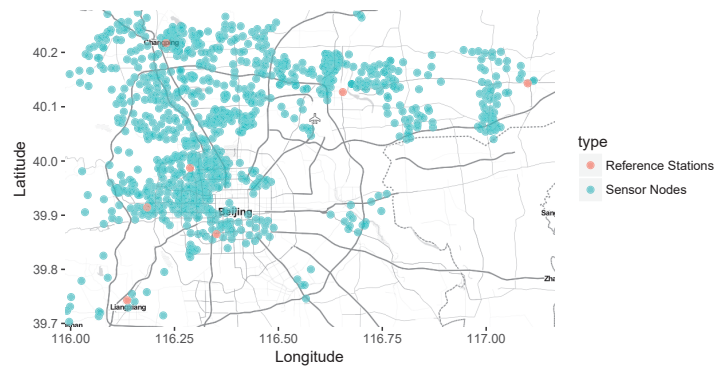


Fig. 1. An illustration of sensors deployed in Beijing, China for $PM_{2.5}$ monitoring. Among the 1000 sensors deployed, only a few are installed close to the public environment monitoring stations, which are used as reference stations.

pollution concentrations in real time [4, 6, 21, 27]. However, the raw measurements of these deployments often lack sufficient accuracy due to sensor noise, inter-device differences or environmental interference [10, 15].

An effective approach to improve the data quality of air quality sensors is calibration [12, 15, 21, 26]. To calibrate a low-cost sensor, its measurements are transformed in a way that the calibrated measurements agree with the measurements of a highly accurate reference. While air quality sensors are usually calibrated before deployment (*pre-deployment calibration*), the calibration parameters still need to be frequently adjusted in the field after deployment (*post-deployment calibration*) [15]. It is reported that the calibration parameters may drift within one month after sensor deployment without re-calibration [16].

Challenges. Post-deployment calibration is challenging particularly for large-scale static air pollution monitoring deployments. This is because once deployed, these sensors tend to have irregular or even no access to references. Fig. 1 shows a real sensor deployment for $PM_{2.5}$ monitoring in Beijing, China. Among the 1,000 $PM_{2.5}$ sensors deployed, only 7 are installed next to highly accurate reference stations. Most existing post-deployment calibration schemes focus on *mobile* deployments, where virtual references are created when sensors meet in space and time, *i.e.*, sensor rendezvous [14, 21, 26, 27]. However, since the sensors do not physically meet in a static deployment, rendezvous-based calibration does not apply. A few pioneer proposals [16, 23] leverage special situations when pollution concentrations are expected to be uniform in certain regions to calibrate sensors in a static deployment. This approach offers calibration opportunities of near-zero concentrations and is only useful for simple offset and gain calibration [15]. Yet the calibration model for $PM_{2.5}$ can be complex [4, 12] and needs to be derived with measurements covering a wide concentration range. It remains open how to calibrate a $PM_{2.5}$ sensor without access to a reference, a common problem faced in urban-scale static deployments.

Our Approach. To conduct *post-deployment calibration* for static sensor deployments, we take an approach inspired by *calibration transfer* [29, 31] in *pre-deployment calibration*. Calibration transfer is a calibration paradigm for sensors without access to references (*target* sensors) leveraging those with access to references (*source* sensors). It calibrates a target sensor by transferring the calibration parameters of a source sensor to a target sensor. The method has been adopted to reduce the *pre-deployment* calibration overhead in mass sensor production [28, 29, 31]. A pre-requisite of conventional calibration transfer is that measurements of the source and target sensors should be *synchronized*, *i.e.*, the two sets of measurements from both sensors can be organized into pairs, in which

both measurements are made upon *the same ground truth*. Synchronized measurements are guaranteed in pre-deployment calibration by putting both the source and the target sensors in the same testing environment. However, for post-deployment calibration, there is often limited, if any, prior knowledge on which pair of measurements from the source and the target sensors are made upon the same ground truth. That is, the measurements are largely *unsynchronized*. Hence conventional calibration transfer for pre-deployment calibration cannot be directly applied to post-deployment calibration.

In this paper, we ask the question: *can we transfer the calibration parameters of source sensors to a target sensor, when no synchronized measurements are available?* We formulate the question as an *unsynchronized calibration transfer* problem, which aims to learn a transformation of the calibration parameters of the source sensors, and applies the transferred calibration on the target sensor to achieve high accuracy, even if the measurements of the source and the target sensors are unsynchronized. Note that unsynchronized measurements are not comparable, and it can be erroneous to directly learn a transformation using unsynchronized measurements. Although it is difficult to solve the generic unsynchronized calibration transfer problem, we make a key observation that helps to solve the unsynchronized calibration transfer problem for urban air pollution monitoring deployments. Specifically, we observe that the $PM_{2.5}$ concentrations at two separate yet sufficiently close locations during the same period of time exhibit similar distributions (see Sec. 4.1.2). It implies that for a source sensor and a target sensor deployed at different locations, the ground truth concentrations of their measurements during the same period of time conform to similar distributions. Using this similarity between distributions of ground truth as a common reference, we develop a solution called *statistical calibration transfer* to this special unsynchronized calibration transfer problem.

On this basis, we propose In-field Calibration Transfer (ICT), an optimization based solution to the unsynchronized calibration transfer problem for static air quality sensor deployments. ICT has three technical novelties.

- We introduce statistical calibration transfer, which makes use of the similarity in distributions of the ground truth at different locations as common references rather than rely on synchronized measurements. Statistical calibration transfer learns the transformation from the estimated distribution of measurements using a novel optimization objective, which can be solved via Bayesian optimization.
- We reduce the search space in statistical calibration transfer by assuming a linear transformation between the calibration parameters between the source and target sensors. This assumption has been tested in labs [28, 29, 31] and we extend it into in-field scenarios.
- We further improve the accuracy of statistical calibration transfer by using an extra air pollution inference engine to generate $PM_{2.5}$ concentration level estimates for the target location. We augment the original optimization objective of statistical calibration transfer with an additional term. We empirically show that even coarse-grained $PM_{2.5}$ concentration levels suffice to improve the calibration accuracy.

Contributions and Roadmap. The main contributions of this work are summarized as follows. (i) To the best of our knowledge, ICT is the first solution to the unsynchronized calibration transfer problem for low-cost air quality sensors. It offers a practical solution to conduct post-deployment calibration for large-scale static urban air pollution monitoring deployments. (ii) We evaluate the performance of ICT on real deployment data and experimental results show that ICT is able to provide approximately equally good calibration performance as if the target sensors have direct access to references, which could potentially increase the usability of large-scale air pollution monitoring sensor deployments.

In the rest of the paper, we first review relevant literature (Sec. 2) and present the background and the problem (Sec. 3). Then we elaborate on the ICT (Sec. 4) and its evaluation (Sec. 5). Finally we conclude this work (Sec. 6).

2 RELATED WORK

Our work is a post-deployment sensor calibration scheme for static air pollution monitoring deployments. It is inspired by applications of transfer learning in sensor calibration. We review the closely related literature below.

2.1 Post-deployment Calibration for Air Quality Sensors

Although low-cost air quality sensors are usually calibrated before installation, periodic post-deployment calibration is still necessary to ensure long-term data accuracy of urban air pollution monitoring systems. Unlike pre-deployment calibration, where every sensor has a reference *e.g.*, in labs, a unique challenge in post-deployment calibration is the lack of references. Virtual references can be created if the sensors are *mobile* and meet in space and time, *i.e.*, sensor rendezvous [26]. Sensors in a rendezvous are supposed to sense the same phenomenon and can be utilized as references for calibration [14, 21]. However, rendezvous-based calibration only applies to mobile sensors and a sensor with no rendezvous cannot be calibrated [6].

For static sensors, post-deployment calibration is viable by exploiting situations where all sensors measure the same pollution concentrations so that they can share the same reference for calibration. Tsujita *et al.* [23] the NO₂ concentrations are almost uniform within the city if the concentrations are low. Thus they propose to calibrate the offset of NO₂ sensors deployed in the city to four references once a NO₂ concentration below 10 ppb is reported. Mueller *et al.* [16] assume that O₃ and NO₂ concentrations are uniform during night at inner city locations and during the afternoon at outer city locations. Correspondingly, the sensors in the inner/outer city can be calibrated to a remote reference in the inner/outer city during night/afternoon.

Our work is also a post-deployment calibration scheme for static sensors, but differs from existing efforts in two-fold. (i) Previous studies on gas sensors [16, 23] are built upon *linear* calibration models. As we will show in Sec. 3.2.1, linear models are insufficient for dust sensors *e.g.*, PM_{2.5} in our case. (ii) The calibration opportunities in [16, 23] only provide near-zero concentrations, which will yield large calibration errors if they are used in complex non-linear calibration models (see the NZ-ICT baseline in Sec. 5). Therefore these two prior studies are not directly applicable to in-field PM_{2.5} calibration transfer. In contrast, our work applies to both simple and complex calibration models.

2.2 Transfer Learning in Sensor Calibration

Transfer learning is a machine learning paradigm aims to improve the learning of the target predictive function in the target domain using the knowledge in a source domain and a source learning task [20]. It has broad applications in text mining [19], computer vision [18], urban computing [8, 24], etc.

In the sensor and measurement research, the concept of transfer learning has been mainly applied in calibrating electronic noses (e-noses). E-noses are sensor arrays for hazardous odor detection. Due to their significant inter-device differences, per-instrument calibration is necessary, and transfer learning is utilized to reduce the calibration overhead in mass production [5, 28, 29, 31]. Assume a source e-nose and a target e-nose. The raw measurements of the target are first standardized to those of the source e-nose. Then the source e-nose is calibrated to a reference and finally the calibration parameters can be directly adopted on the target e-nose.

Our work is inspired by the concept of calibration transfer in e-noses. However, most e-nose calibration transfer studies are performed *in labs* for *pre-deployment calibration* while we focus on *in-field* calibration transfer for *post-deployment calibration*. The former assumes the source and the target sensors are measuring the same phenomenon in the same lab setting, *i.e.*, synchronized. Yet the latter is more challenging because the source and the target sensors are installed at different locations and their measurements are largely unsynchronized.

3 BACKGROUND AND PROBLEM

In this section, we first introduce the basics of sensor calibration (Sec. 3.1) and then conduct a measurement study on a $PM_{2.5}$ monitoring deployment to motivate the need for calibration transfer (Sec. 3.2). Finally we formally define the unsynchronized calibration transfer problem (Sec. 3.3).

3.1 Primer on Air Pollution Sensor Calibration

Calibration is an efficient approach to improve the data quality of low-cost sensors. It finds a *calibration model* that maps the measurements of a *low-cost sensor* to those of an accurate *reference sensor* [15]. Given a set of measurements $X = \{x_1, x_2, \dots, x_N\}$ of a low-cost sensor and a set of measurements $Y = \{y_1, y_2, \dots, y_N\}$ of a reference sensor, a calibration model C establishes a relationship between X and Y such that certain error metric between the calibrated measurements $C(X)$ and the reference measurements Y is minimized. There has been extensive research on how to derive calibration models suited for different air pollution sensors and error sources. We refer interested readers to [15] for a comprehensive review.

For air pollution sensors, it is crucial to conduct both *pre-deployment* and *post-deployment* calibration. Pre-deployment calibration identifies the proper calibration model, while periodic post-deployment calibration is important to maintain the data quality of long-term deployment. One major challenge in post-deployment calibration is the lack of reference sensors to re-calibrate the low-cost sensors. This is particularly the case for large-scale static air pollution sensor deployments, which our work focuses on.

3.2 Measurement Study

This subsection presents a measurement study on a real-world $PM_{2.5}$ sensor deployment to motivate the need for calibration transfer. Specifically, the measurement study aims to answer three questions: (i) Is a linear calibration model sufficient for $PM_{2.5}$ sensor calibration? (ii) Is it necessary to periodically re-calibrate $PM_{2.5}$ sensors? (iii) Is it feasible to directly apply calibration parameters of one sensor to sensors at other locations?

Sensor Deployment and Dataset. We collect measurements from a large-scale $PM_{2.5}$ monitoring system deployed in Beijing, China. It consists of 1,000 low-cost sensors measuring $PM_{2.5}$, temperature and humidity (see Fig. 1). In addition to $PM_{2.5}$ concentration, the PM sensor [30] in each sensor box (see Fig. 2-(b)) also reports 12 low-level features. Each sensor uploads its readings to a back end server every minute. Among the 1,000 low-cost sensors, only 7 (denoted as $S1$ to $S7$ in Fig. 2-(a)) are installed next to highly accurate air pollution monitoring stations as references (see Fig. 2-(c)). The remaining sensors have no access to the reference stations. The low-cost $PM_{2.5}$ sensors are based on light scattering principles [30], while the reference stations are based on beta-attenuation or tapered element oscillating microbalance method [17].

For the measurement study, we collect readings ($PM_{2.5}$ concentration and the 12 low-level features) from the 7 sensors as well as the $PM_{2.5}$ readings from the corresponding 7 reference stations as ground truth [1]. The dataset collected covers a time period of 10 months from October 1, 2017 to July 31, 2018.

3.2.1 Is a Linear Calibration Model Sufficient for $PM_{2.5}$ Calibration? While linear calibration models are prevalent in gas sensor calibration [16, 23], non-linear models are often needed for dust sensor calibration such as $PM_{2.5}$ [12]. Table 1 shows the mean absolute errors (MAE) of applying the popular linear (multiple linear regression [16, 23]) and non-linear (random forest [12]) models to calibrate the raw measurements of the low-cost $PM_{2.5}$ sensor with reference to their co-located highly accurate reference station. MAE is a widely used metric to evaluate the data accuracy of air pollution sensors [15]. For $PM_{2.5}$ concentration, a MAE below 10 is considered accurate for data mining applications [2]. The evaluation is conducted on the sensor node $S6$ and the MAEs are averaged over 10 months (October 1, 2017 to July 31, 2018). For each month, 70% of the data are used for training and 30% for testing to calculate the MAEs of different calibration methods. The results show that linear calibration models

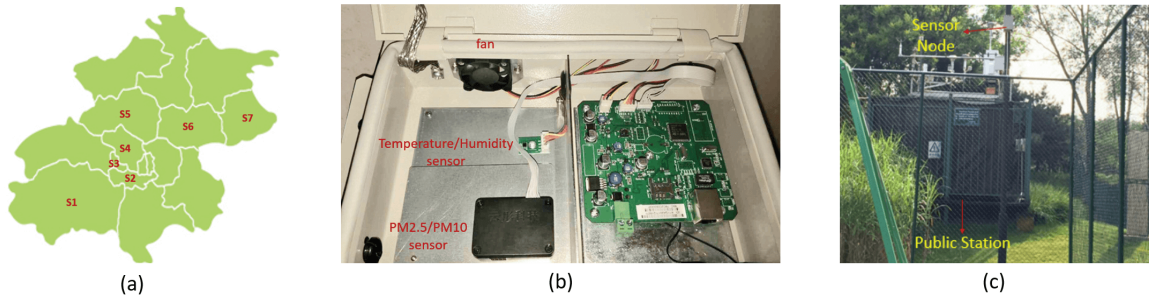


Fig. 2. Illustration of sensor deployment. (a) Locations of sensors (S_1 to S_7) with access to public reference stations (R_1 to R_7). (b) Hardware of sensor. (c) Installation of a sensor next to a reference station.

Table 1. MAEs of different calibration models.

Calibration Model	Raw Data	Multiple Regression (Linear)	Random Forest (Non-linear)
MAE	30	22	9

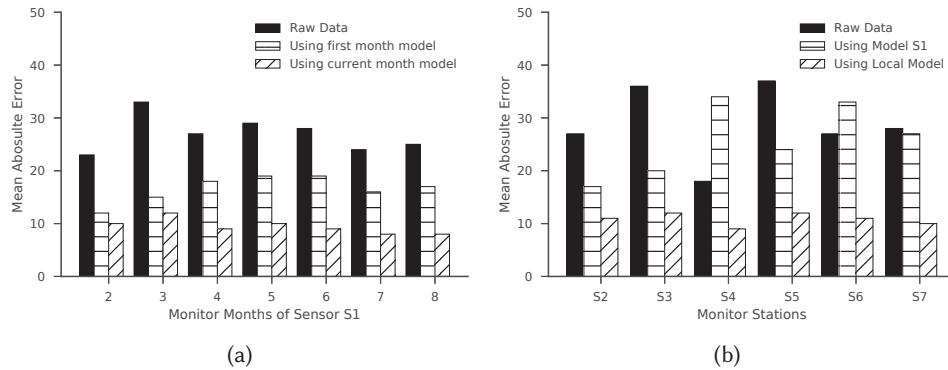


Fig. 3. MAEs of directly applying the calibration model learned from measurements of S_1 collected in first month to calibrate (a) measurements of S_1 collected from second month to the eighth month and (b) measurements of the other 6 sensors (S_2 to S_7) collected in the first month.

fail to yield satisfactory accuracy of calibration on $PM_{2.5}$ sensors. This suggests that previous calibration transfer studies [16, 23], which are built upon linear calibration models, are not directly applicable. In the rest of this paper, we take the random forest described in [12] as the calibration model for $PM_{2.5}$ sensors.

3.2.2 Is Periodic Calibration Necessary? Fig. 3a shows the MAEs of the uncalibrated raw measurements of Sensor S_1 , and two different calibration approach applied on it. The first approach is to train the calibration model on the 70% training set in the first month, and then test it on the 30% testing set in the next 7 months. The second approach is to directly train the calibration model on the current month's training set and test on the testing set. As shown in the figure, the second approach provides lower MAEs. This result indicates that the optimal

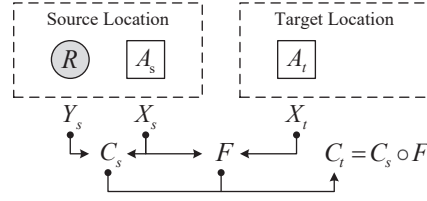


Fig. 4. An illustration of calibration transfer problem. A source sensor A_s is co-located with a reference sensor R at the source location, while a target sensor A_t has no access to any reference at the target location. Their measurements are X_s , Y_s , and X_t , respectively. A calibration model C_s for A_s can be learned from X_s and Y_s . The calibration transfer problem tries to derive a function F using X_s and X_t such that the calibration model C_s can be transferred to A_t , where the calibration model for A_t can be calculated as $C_t = C_s \circ F$.

calibration parameters for the same sensor do vary over time. It may induce large errors by directly adopting a previously trained calibration model to calibrate even the same sensor after a long period of time.

3.2.3 Is One Set of Calibration Parameters Applicable to Sensors at Different Locations? Fig. 3b plots the MAEs by applying the calibration model of S1 to calibrate the $PM_{2.5}$ measurements of the other 6 sensors collected in the same month. The results show that the MAEs can be even larger than those of the raw measurements without calibration, which indicates that the optimal calibration parameters for sensors at different locations can differ significantly. Therefore, the calibration model learned for one sensor requires to be adapted (transferred) to be used on other sensors.

Summary. $PM_{2.5}$ sensor calibration needs non-linear models *e.g.*, random forests (Sec. 3.2.1). It is necessary to conduct periodic re-calibration for each deployed sensor (Sec. 3.2.2 and Sec. 3.2.3), which can be expensive and labour-intensive. This is particularly the case when large numbers of sensors are static and have no access to the references. To reduce the overhead of post-deployment calibration, we explore to *transfer* the calibration results from source sensors (with access to references) to target sensors (without access to references).

3.3 Unsynchronized Calibration Transfer Problem

For ease of presentation, we explain our problem by using one source sensor and one target sensor (Fig. 4). We discuss extensions to multi-source scenarios in Sec. 4.4. The **calibration transfer problem** is defined as follows.

Denote A_s as a source sensor, which is co-located with a highly accurate reference station R_s . We use $X_s = \{x_s^{(i)}\}_{i=1}^{N_s}$ to represent the measurements of A_s , where $x_s^{(i)} \in \mathbb{R}^d$ is the i^{th} measurement, *i.e.*, a d -dimension feature vector, and N_s is the number of measurements of A_s . Similarly, $Y_s = \{y_s^{(i)}\}_{i=1}^{N_s}$ represents the measurements of R_s , where $y_s^{(i)} \in \mathbb{R}$ is the i^{th} measurement, *i.e.*, the ground truth $PM_{2.5}$ concentration corresponding to the measurement $x_s^{(i)}$. Then a calibration model $C_s : \mathbb{R}^d \rightarrow \mathbb{R}$ can be learned for the source sensor A_s from X_s and Y_s , as discussed in Sec. 3.1 by minimizing $\|Y_s - C_s(X_s)\|_F^2$, where $\|\cdot\|_F$ is the Frobenius Norm. Finally, denote A_t as a target sensor, and $X_t = \{x_t^{(i)}\}_{i=1}^{N_t}$ as its measurements ($x_t^{(i)} \in \mathbb{R}^d$, and N_t is the number of measurements of A_t). Y_t is used to denote the corresponding ground truth $PM_{2.5}$ concentration at the target location. The calibration transfer problem aims to find a transformation function $F : \mathbb{R}^d \rightarrow \mathbb{R}^d$, such that $\|Y_t - C_s(F(X_t))\|_F^2$ is minimized. In other words, the calibration model C_s for A_s is transferred to $C_t = C_s \circ F$ for A_t .

There are two types of calibration transfer problems, **synchronized calibration transfer problem** and **unsynchronized calibration transfer problem**. The former assumes that the measurement set X_s and X_t are synchronized, *i.e.*, $N_s = N_t = N$ and for each $i = 1 \dots N$, and we have $y_s^{(i)} = y_t^{(i)}$. This type of calibration transfer can be solved by direct standardization [5], which assumes the transformation function F to be linear. It has been

applied to calibrate large numbers of instruments *in labs* when it is time-consuming to learn a (often complex) calibration model for each instrument [5, 28, 29, 31]. In our particular interest is the latter, *i.e.*, **unsynchronized calibration transfer problem**, where Y_t is not known, and X_t *cannot be synchronized* to X_s . This is the common case for static air pollution sensor deployments. In this case, it remains open how to learn F from X_t and X_s , which is the focus of this work.

4 IN-FIELD CALIBRATION TRANSFER METHOD

To solve the unsynchronized calibration transfer problem, we propose ICT (in-field calibration transfer). We elaborate each technique in ICT for single-source calibration transfer, including statistical calibration transfer (Sec. 4.1), exploiting linearity of the transformation (Sec. 4.2), and exploiting results from spatial predictions (Sec. 4.3). Then we extend ICT to the multi-source scenario (Sec. 4.4).

4.1 Statistical Calibration Transfer

4.1.1 Main Idea. One fundamental challenge in the unsynchronized calibration transfer problem is that Y_t is unknown, so there is no common reference to synchronize X_t and X_s . The key idea of our solution is based on the following assumption: for the same period of time, and when the distance between A_s and A_t are small enough, we have $p(Y_s) \approx p(Y_t)$, where $p(\cdot)$ denotes the probability distribution. Based on this assumption, it is possible to find the transformation function F by solving the following optimization problem:

$$\operatorname{argmin}_F d_{KL}[\hat{p}(Y_s), \hat{p}(C_s(F(X_t)))] \quad (1)$$

where \hat{p} is a histogram density estimator, and $d_{KL}[\cdot, \cdot]$ is the Kullback-Leibler (KL) divergence. Instead of synchronizing individual measurements, we learn the transformation F by minimizing the difference between the *estimated distribution of calibrated target measurement* and the *ground truth at source location*.

While conventional calibration transfer requires explicit pre-knowledge of Y_t , such that both measurement sets X_s and X_t can be synchronized accordingly, our statistical calibration transfer loosens this requirement: when the distributions of ground truth in both location, $p(Y_s)$ and $p(Y_t)$, are known to be similar, it is enough to transfer the calibration.

4.1.2 Empirical Validation of Key Assumption. The effectiveness of statistical calibration transfer relies on the key assumption that for the same period of time, and when the distance between A_s and A_t are small enough, we have $p(Y_s) \approx p(Y_t)$. This assumption is built upon our observation that *during the same period of time*, when the source location and target location *both locate near to each other* (*e.g.*, in the same city), the distributions of the true $PM_{2.5}$ concentrations in both location are similar. Below we empirically demonstrate this observation.

While the $PM_{2.5}$ concentration usually varies in space and time, its distributions over a certain period of time may be similar at different locations because of *e.g.*, similar land-use and pollution sources. Fig. 5 plots the ground truth $PM_{2.5}$ concentrations measured by reference stations $R6$ and $R7$, co-located next to the low-cost sensors $S6$ and $S7$ shown in Fig. 2 over two different months. Fig. 5a and Fig. 5b shows the histogram and density distributions of reference sensor $R6$ and $R7$ during May, 2018, which are quite similar to each other. The same phenomenon can also be observed in Fig. 5c and Fig. 5d. However, on the other hand, the distribution in different months of the same reference sensors varies greatly (*e.g.*, Fig. 5a vs. Fig. 5c).

To explore whether this observation is an artefact of the Beijing dataset, we collect $PM_{2.5}$ measurements from public stations in three other major cities in China, Tianjin, Shanghai and Shenzhen, which are 130, 1200 and 1900 kilometres away from Beijing, respectively. From each of these cities, 9 public stations are selected. KL divergence is calculated between the monthly $PM_{2.5}$ concentration distributions of two sensors, and averaged over the 12 months from October, 2017 to October, 2018. The results are shown in Fig. 6 and Fig. 7.

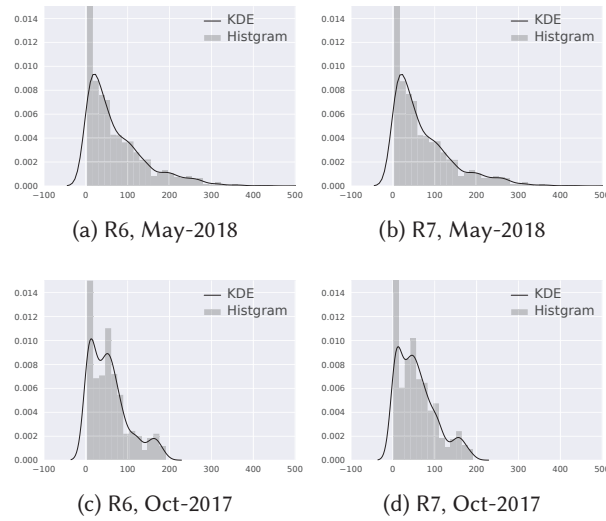


Fig. 5. The ground truth distribution of reference sensor $R6$ and $R7$, co-located near the low cost sensor $S6$ and $S7$, respectively, as shown in Fig. 2. The distribution of different reference sensors in the same month is similar, while the distribution of even the same sensor among different months varies a lot.

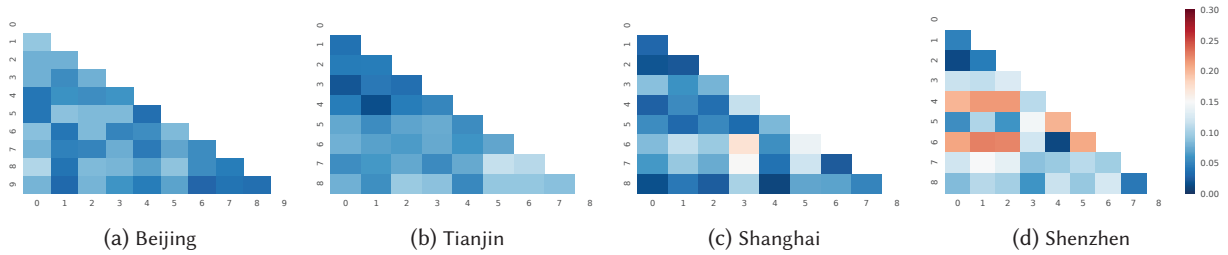


Fig. 6. $PM_{2.5}$ distribution differences (measured by KL divergence) between stations in (a) Beijing; (b) Tianjin; (c) Shanghai and (d) Shenzhen.

In Fig. 6, we calculate the KL divergence among $PM_{2.5}$ measurement distributions of different public stations located within each of the four cities. The overall KL divergence of the other three cities is at the same low level as Beijing. As shown in Sec. 5, the similarity among the measurement distributions within Beijing suffices for our ICT to provide a solid calibration performance. Since the same level of similarity is also observed in other cities, we believe that ICT is generally applicable to intracity post-deployment calibration problems.

To further investigate the similarity of $PM_{2.5}$ measurement distributions from sensors in different locations, we compare sensors located in different cities with each other and summarise the results in Fig. 7. As shown in the figure, the $PM_{2.5}$ measurement distributions in Beijing & Tianjin have higher similarity than Beijing & Shanghai, which then have higher similarity than Beijing & Shenzhen. In general, we observed that the similarity of $PM_{2.5}$ measurement distributions between two cities is negatively correlated to the geological distance between them.

We make the following comments on the usage of the observation.

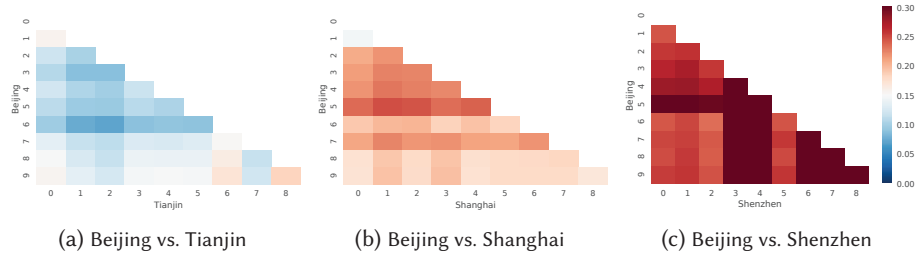


Fig. 7. $PM_{2.5}$ distribution differences (measured by KL divergence) between stations in (a) Beijing and Tianjin; (b) Beijing and Shanghai and (c) Beijing and Shenzhen.

- We only consider mapping the measurements between the source and the target collected during the same period of time in statistical calibration transfer. We do not consider transferring measurements of different months because their distributions are likely to differ due to seasonal changes.
- We limit the spatial range between the source and the target within a city for statistical calibration transfer. It is also possible that the observation may not hold at certain locations within the city. Nevertheless, as we will introduce in Sec. 4.4, our method still works because it can regard these locations as negative transfer samples with the help of multi-source calibration transfer.

4.2 Exploiting Linearity of Transformation Function F

As discussed earlier, direct standardization [5] learns F directly from X_s and X_t . Previous studies [5, 28, 29, 31] have assumed that the transformation function F to be linear and have shown that this assumption works well for gas sensors in labs. We extend this assumption to in-field scenario, *i.e.*, we assume transformation function F is also a linear function in in-field situation.

The experiment results Sec. 5 shows that this assumption allows ICT to provide decent calibration accuracy, while notably reducing the searching space of F and increases the efficiency of the algorithm.

4.3 Exploiting Spatial Prediction Results

In practice, sometimes there are inferences of the ground truth at the target area provided by other methods, *e.g.*, air quality map. The performance of ICT can be further improved by making use of these inferred target ground truth, denoted as Y'_t . Specifically, a new term can be added to Eq. (1):

$$\operatorname{argmin}_F d_{KL}[\hat{p}(Y_s), \hat{p}(C_s(F(X_t)))] + \lambda \cdot d_c[Y'_t, C_s(F(X'_t))] \quad (2)$$

where $d_c[\cdot, \cdot]$ is a measure of distance between two sample sets and λ is a parameter used to adjust the influence of the inferred target ground truth.

In this work, we take $PM_{2.5}$ concentration levels inferred by an air quality map as Y'_t at the location of the target sensor. Instead of accurate $PM_{2.5}$ concentration, Y'_t consists of integer value ranging from 1 to 6, which represent different $PM_{2.5}$ concentration levels. Hence $d_c[\cdot, \cdot]$ in Eq. (3) is defined as the classification error rate. As we will show in Sec. 5.3.4, a small fraction of inferences with high confidence suffices to provide considerable improvement in calibration transfer.

In ICT, we apply Gaussian process regression [4] for inferred ground truth generation. While other models for air quality inference also apply, we choose Gaussian process regression for its simplicity and efficiency. We take sensor readings from high quality public stations, GPS location information, eight categories of POI data

Algorithm 1: In-field calibration transfer

Input: m source measurement and ground truth pairs (X_{s_j}, Y_{s_j}) , target measurements X_t , certainty threshold τ , readings from other high quality public stations, GPS location information, POI data, and meteorological data

Output: Ensembled calibrated measurements of target sensor Y_t^e

- 1 Conduct spatial prediction and forms the data tuple (X_t^τ, Y_t^τ)
- 2 **for** each source sensor $j \in 1 \dots m$ **do**
- 3 Get calibration model C_{s_j} of source sensor A_{s_j} using the data pairs (X_{s_j}, Y_{s_j})
- 4 Using the Bayesian optimization to solve the objective function and get transformation matrix F_j
- 5 Calculate the similarity weight $\Psi(s_j)$
- 6 $Y_t^e = \sum_{j=1}^m \Psi(s_j) (C_{s_j}(F_j(X_t)))$
- 7 **return** Y_t^e

(culture & education, parks, sports, hotels, shopping malls & supermarkets, entertainment, decoration & furniture markets, and vehicle services) and meteorological data (temperature, humidity, wind speed, wind direction) as the input of the spatial predictor and the output is $y_t^{(i)}$ with an according variance $\sigma_t^{(i)}$, which indicates the confidence of the prediction. We only use the prediction data $y_t^{(i)}$ with $\sigma_t^{(i)}$ smaller than some threshold τ . We form a prediction set $Y_t^\tau = \{y_t^{(i)} | \sigma_t^{(i)} < \tau\}$ and its corresponding measurement set from target sensor X_t^τ , and Eq. (2) then becomes

$$\underset{F}{\operatorname{argmin}} d_{KL}[\hat{p}(Y_s), \hat{p}(C_s(F(X_t)))] + \lambda \cdot d_c[Y_t^\tau, C_s(F(X_t^\tau))] \quad (3)$$

4.4 Extension to Multi-source Calibration Transfer

This subsection extends ICT to support calibration transfer from multiple source sensors. The main challenge is to select the most promising sources to avoid negative transfer [7].

To support multi-source selection and transfer, we first need to quantify the differences between the environment of the source sensors and that of the target sensor. We use a classifier induced divergence measure called \mathcal{H} distance [11] which measures the divergence that only affects the classification accuracy. We use \mathcal{D}_s to represent the source domains, which has reference sensors and calibration models, while \mathcal{D}_t is used to represent the target domain. We want to transfer the models learned in \mathcal{D}_s to the target domain \mathcal{D}_t . Then, we use $d_{\mathcal{H}}^y(\mathcal{D}_s, \mathcal{D}_t)$ to represent the distance between the source environment and the target environment. The smaller the distance, the more similar the two environments are. We define the similarity between a source and the target as below.

$$\Phi^y(s, t) = 1 - d_{\mathcal{H}}^y(\mathcal{D}_s, \mathcal{D}_t) \quad (4)$$

Then we select the most promising sources according to their relative similarity weight, which is defined as:

$$\Psi(s_j) = \frac{\Phi(\mathcal{D}_{s_j}, \mathcal{D}_t)}{\sum_{s=1}^m \Phi(\mathcal{D}_s, \mathcal{D}_t)} \quad (5)$$

where s_j is the j^{th} source domain among m source domains, and $\Psi(s_j)$ is the weight used for ensemble. We can use these similarity weights to calculate the ensemble calibration transfer result. The similarity between feature values of the source and target domains reflects the similarity of domains. If a pair of domains are more similar, we can rely more on the calibration transfer result between them and put more weight on it.

Algorithm 1 shows the entire process of ICT for multi-source in-field calibration transfer.

5 EVALUATION

This section presents the evaluations of ICT. We introduce the experiment setups (Sec. 5.1), present the overall performance (Sec. 5.2), and then conduct micro-benchmark evaluations to understand the performance of each module in ICT (Sec. 5.3). Finally we conduct a case study on pollution source localization as an application of calibration transfer (Sec. 5.4).

5.1 Experiment Setup

Datasets. We mainly evaluate the performance of ICT using measurements from the 7 low-cost sensors, denoted as $S1 \dots S7$, which are installed next to public stations (Fig. 1). These public stations are also called *reference stations*, denoted as $R1 \dots R7$. For each sensor, we collect the 12 low-level features as well as its $PM_{2.5}$ readings, which is called a *measurement*. The $PM_{2.5}$ concentration recorded by each reference station is collected as *ground truth*. Each measurement and ground truth at the same time are formed as a *tuple*. We collect one tuple per hour for 10 months (from October 1, 2017 to July 31, 2018), which covers various weather conditions and $PM_{2.5}$ concentration range. In total, there are more than 50,210 tuples. In our experiments, 70% of the collected data are used as training set, and 30% as test set.

Note that ICT also needs data to infer ground truth via spatial prediction (see Sec. 4.3). To built up the spatial prediction model, we collect $PM_{2.5}$ concentration data from all the 35 public stations in Beijing [1] (including $R1 \dots R7$), meteorological data [22] (including temperature, humidity, wind speed, wind direction), GPS location information, as well as POI data [9] (see Sec. 4.3 for details) for the same time period.

Metrics. Mean absolute error (MAE) after calibration is the main metric used to assess the performance of ICT. To make the results more semantically useful for end users, we also calculate the classification accuracy on the officially defined 6 discrete $PM_{2.5}$ levels [4], which is indicated as $L1 \dots L6$.

Baselines. We compare the performance of ICT with the following baselines.

- **Direct Transfer:** It directly use the calibration parameters of a source sensor to a target sensor without any transformation.
- **TCA:** It is a popular transfer learning method in computer vision. Some research [15] suggests it is also a potential solution to calibration transfer. TCA [19] calculates a common transfer Φ which applies on both X_s and X_t . The source calibration model C_s is learned by minimizing $\|Y_s - C_s(\Phi(X_s))\|_F^2$, rather than $\|Y_s - C_s(X_s)\|_F^2$. This source calibration is then applied on the transformed target measurement, *i.e.*, $C_s(\Phi(X_t))$, which is evaluated with the ground truth Y_t . We use TCA with a polynomial kernel to reduce the feature space to 10 dimensions, *i.e.*, $\Phi(x_s^{(i)}) \in \mathbb{R}^{10}$.
- **NZ-ICT:** It represents ICT using only near-zero measurements as references. Some studies on gas sensors [16, 23] use near-zero measurements for calibration transfer. We consider $PM_{2.5}$ concentrations under 35, *i.e.*, *level 1* as near zero data, and feed them into the standard ICT for calibration transfer.

When comparing the performance of the above baselines in multi-source calibration transfer scenario, we use the corresponding ensemble solution of these methods.

Implementation. All our experiments are conducted in a PC with Intel(R) Core(TM) i7-7600U CPU. All the code is implemented in python and the source code will be published online soon.

5.2 Overall Performance

Table 2, Table 3, Table 4 and Table 5 show the results of single-source calibration transfer using different methods for all of the seven sensor pairs. The MAEs greater than 20 are marked in red. As is shown, direct transfer has the worst MAEs. TCA and NZ-ICT yields smaller MAEs but there are still some large MAEs marked in red. In contrast, all the MAEs of ICT are below 18. Worth mentioning is that the emboldened diagonal elements of these

Table 2. Direct transfer result

		target sensor						
		S1	S2	S3	S4	S5	S6	S7
source sensor	S1	10	14	15	21	36	28	17
	S2	13	12	12	12	31	22	15
	S3	18	18	9	9	28	17	12
	S4	25	21	13	7	25	14	14
	S5	53	49	42	34	11	26	43
	S6	39	36	24	15	22	10	24
	S7	19	19	12	15	30	18	9

Table 3. TCA transfer result

		target sensor						
		S1	S2	S3	S4	S5	S6	S7
source sensor	S1	10	19	24	23	19	28	15
	S2	19	12	18	13	22	27	18
	S3	21	15	9	12	18	17	21
	S4	30	19	10	7	12	18	19
	S5	18	18	29	12	11	17	18
	S6	21	32	15	25	22	10	31
	S7	29	15	14	15	18	22	9

Table 4. NZ-ICT transfer result

		target sensor						
		S1	S2	S3	S4	S5	S6	S7
source sensor	S1	10	17	18	22	24	23	16
	S2	17	12	19	15	26	31	20
	S3	17	16	9	15	22	19	18
	S4	36	22	13	7	13	21	17
	S5	31	29	22	24	11	19	25
	S6	31	22	17	21	17	10	26
	S7	22	17	12	18	23	17	9

Table 5. ICT transfer result

		target sensor						
		S1	S2	S3	S4	S5	S6	S7
source sensor	S1	10	15	12	13	14	13	13
	S2	12	12	12	9	14	10	13
	S3	13	14	9	10	15	10	12
	S4	13	13	11	7	15	9	11
	S5	13	15	12	11	11	14	16
	S6	13	13	11	7	14	10	11
	S7	13	16	11	11	17	10	9

Table 6. Cross-validation results using multi-source calibration transfer

	S1	S2	S3	S4	S5	S6	S7
Ensemble Direct Transfer	23	20	19	16	24	28	17
Ensemble TCA	18	17	18	15	14	14	15
Ensemble NZ-ICT	16	18	17	14	18	21	14
Ensemble ICT	11	12	10	8	11	8	11
Direct Calibration	10	12	9	7	11	10	9

three tables represent the calibration transfer of the 7 sensors with themselves. In other words, they are *directly calibrated* with their co-located reference stations $R1 \dots R7$. As we can see from Table 5, the performance of ICT (non-diagonal elements) is already close to that of direct calibration (diagonal elements). This can be seen as a proof of the effectiveness of ICT.

Table 6 shows the cross-validation results using multi-source calibration transfer. Each sensor is set as target sensor and we transfer the calibration model from all other sensors. Again, *direct calibration* represents the performance of the calibration model learned directly using target sensor data and the the ground truths, which can be seen as the best possible performance for any calibration transfer method. The result shows that ICT in the multi-source scenario achieves the best performance among all the methods, and it provides an almost equal performance as the direct calibration.

5.3 Micro-benchmarks

In this series of experiments, we set S_6 as the target sensor and the other six sensors ($S_1 \dots S_5, S_7$) as the source sensors, and evaluate the impact of different techniques in ICT on the overall performance.

Table 7. Direct transfer matrix
MAE=36, Acc=0.48

Ground Truth	Predictions						Recall
	L1	L2	L3	L4	L5	L6	
L1	444	265	16	2	3	0	0.61
L2	26	228	169	15	3	3	0.51
L3	1	4	89	120	75	8	0.30
L4	1	0	4	22	109	17	0.14
L5	0	0	6	5	47	119	0.27
L6	0	0	0	0	0	62	1.00
	0.94	0.46	0.31	0.13	0.20	0.30	
	Precision						

Table 8. TCA transfer matrix
MAE=17, Acc=0.68

Ground Truth	Predictions						Recall
	L1	L2	L3	L4	L5	L6	
L1	516	211	1	2	1	0	0.71
L2	25	391	15	1	2	0	0.90
L3	0	120	137	33	6	1	0.46
L4	1	9	42	81	20	0	0.53
L5	0	8	8	29	119	13	0.67
L6	0	0	0	0	39	23	0.37
	0.95	0.53	0.67	0.55	0.64	0.62	
	Precision						

Table 9. NZ-ICT transfer matrix
MAE=24, Acc=0.58

Ground Truth	Predictions						Recall
	L1	L2	L3	L4	L5	L6	
L1	685	43	3	0	0	0	0.94
L2	147	275	11	0	1	0	0.63
L3	2	208	83	2	2	0	0.28
L4	1	13	130	9	0	0	0.06
L5	0	10	57	86	24	0	0.14
L6	0	0	0	12	48	2	0.03
	0.82	0.50	0.29	0.08	0.32	1.0	
	Precision						

Table 10. ICT transfer matrix
MAE=13, Acc=0.78

Ground Truth	Predictions						Recall
	L1	L2	L3	L4	L5	L6	
L1	577	146	5	2	1	0	0.79
L2	27	351	53	1	2	0	0.81
L3	1	34	239	17	5	1	0.80
L4	1	2	38	99	12	1	0.65
L5	0	1	14	25	135	2	0.76
L6	0	0	0	0	26	36	0.58
	0.95	0.66	0.68	0.69	0.75	0.9	
	Precision						

5.3.1 Performance of Single-source Calibration Transfer. Here we show the calibration transfer results from S_7 to S_6 . Table 7 shows the performance of direct transfer. The MAE is 36 and the classification accuracy of the 6 $PM_{2.5}$ levels ($L_1 \dots L_6$) is only 0.48, which are almost unusable. Table 8 shows the results of TCA. The MAE decreases to 17 and the overall accuracy improves to 0.68. Table 9 shows the results of NZ-ICT. Using only near-zero measurements as references, NZ-ICT provides only MAE of 24 and accuracy of 0.58, which is even not as good as TCA. Finally, Table 10 shows the results of our ICT, with λ in Eq. (3) set to 0.3. The MAE and classification accuracy is improved to 13 and 0.78, respectively. Moreover, the recall of each level is generally better than the other three methods.

5.3.2 Performance of Multi-source Calibration Transfer. Table 11 shows the resulting ensemble weights. The weights of S_5 and S_7 are relatively large because they are close to S_6 and may have a similar environment. Table 12 shows the results using ensemble direct transfer. The MAE decreases from 36 in the single-source scenario to 28 in the multi-source scenario, and the accuracy improves from 0.48 to 0.55. Ensemble TCA also achieves a better result than the single-source scenario, with $MAE = 14$ and $Acc = 0.76$, as shown in Table 13. Ensemble NZ-ICT acquires only a slightly better transfer result when compared with ensemble direct transfer, with $MAE = 21$ and $ACC = 0.621$, as shown in Table 14. Using the full ensemble ICT approach, however, the final MAE is improved to 8 and accuracy to 0.86, as shown in Table 15, which is significantly better than the other methods. The results show that in general, multi-source calibration transfer outperforms the corresponding single-source calibration transfer, and our ensemble ICT also outperforms the other methods.

Table 11. Ensemble weights for different source sensors.

Source Sensor	S1	S2	S3	S4	S5	S7
Ensemble Weight	0.07	0.16	0.15	0.13	0.22	0.27

 Table 12. Ensemble direct transfer
MAE=28, Acc=0.55

Ground Truth	Predictions						Recall
	L1	L2	L3	L4	L5	L6	
L1	470	242	15	1	3	0	0.64
L2	18	264	136	12	2	2	0.61
L3	1	7	120	119	46	4	0.40
L4	1	0	5	34	109	4	0.22
L5	0	0	7	7	73	90	0.41
L6	0	0	0	0	0	62	1.00
	0.96	0.51	0.42	0.20	0.31	0.38	
	Precision						

 Table 13. ensemble TCA transfer
MAE=14, Acc=0.76

Ground Truth	Predictions						Recall
	L1	L2	L3	L4	L5	L6	
L1	568	153	7	2	1	0	0.78
L2	24	337	68	3	2	0	0.78
L3	1	25	231	31	6	3	0.78
L4	1	0	27	93	30	2	0.61
L5	0	2	13	17	133	12	0.75
L6	0	0	0	0	10	52	0.84
	0.96	0.65	0.67	0.64	0.73	0.75	
	Precision						

 Table 14. ensemble NZ-ICT transfer
MAE=21, Acc=0.61

Ground Truth	Predictions						Recall
	L1	L2	L3	L4	L5	L6	
L1	678	49	4	0	0	0	0.93
L2	131	289	14	0	0	0	0.67
L3	2	183	106	4	2	0	0.35
L4	1	7	124	20	1	0	0.13
L5	0	9	41	88	39	0	0.22
L6	0	0	0	3	56	3	0.05
	0.83	0.54	0.37	0.17	0.39	1.0	
	Precision						

 Table 15. ensemble ICT transfer
MAE=8, Acc=0.86

Ground Truth	Predictions						Recall
	L1	L2	L3	L4	L5	L6	
L1	666	65	0	0	0	0	0.91
L2	33	372	29	0	0	0	0.86
L3	1	33	243	20	0	0	0.82
L4	0	2	18	115	18	0	0.75
L5	0	1	9	17	142	8	0.80
L6	0	0	0	2	5	55	0.89
	0.95	0.79	0.81	0.75	0.86	0.87	
	Precision						

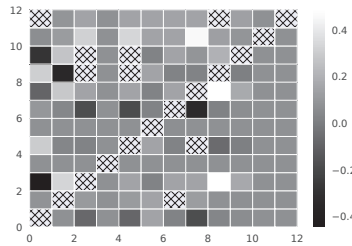


Fig. 8. Transformation matrix derived from direct standardization. Blocks marked with nets are weights greater than 0.5. The diagonal elements are more dominant than the non-diagonal ones.

5.3.3 *Validation of Linear Transformation Function F .* To validate the linearity of the transformation function, we deploy an extra low-cost sensor S'_6 near the reference sensor $R6$. Fig. 8 illustrates the transformation function F derived from direct standardization [5]. The diagonal elements represent the linear relationship between the same features of the two sensors, while the non-diagonal elements can be seen as the linear relationship between different features, *i.e.*, cross-features or cross-sensitivity [13]. As is shown, the diagonal elements are more dominant than the non-diagonal ones, indicating that F can be approximated as a linear function.

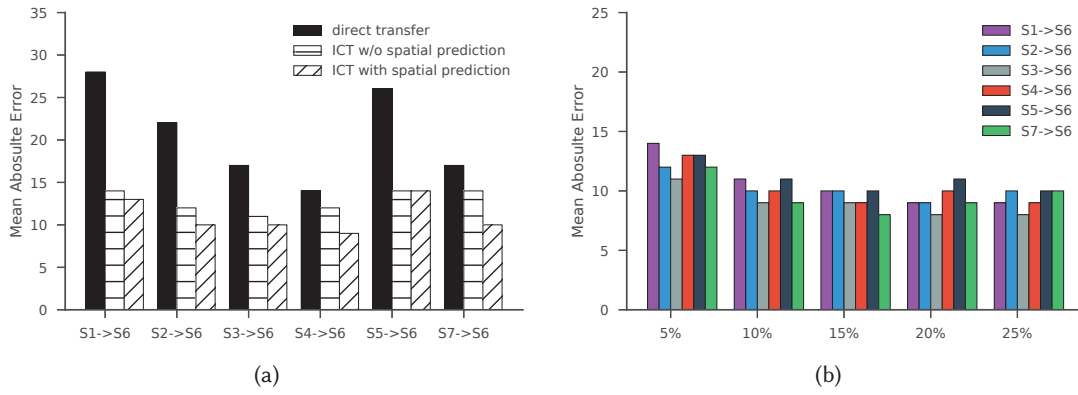


Fig. 9. Impact of spatial prediction. (a) Performance of ICT with and without spatial prediction (d_c). (b) Performance using different size of Y_t^τ . The percentage represents $\|Y_t^\tau\|/\|Y_t'\|$.

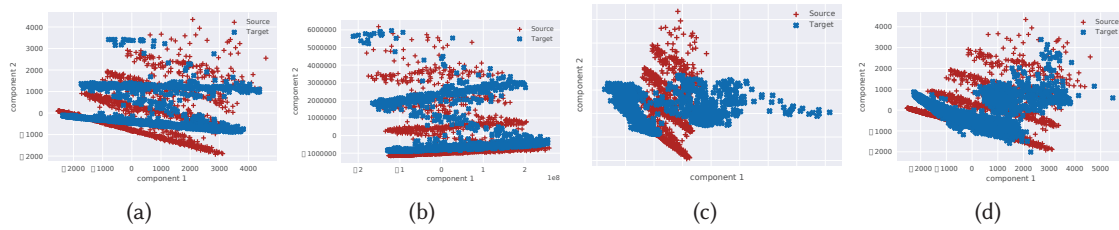


Fig. 10. PCA visualization between the transformed measurements of source sensor $S7$ and target sensor $S6$. (a) Blue dots represent the largest two PCA components of X_t from $S6$, while red dots represent X_s from $S7$; (b) TCA transfer method; (c) NZ-ICT transfer method; (d) ICT transfer method.

5.3.4 Impact of Spatial Prediction. In Eq. (3), we add the second term $d_c[Y_t^\tau, C_s(F(X_t^\tau))]$ to the optimization objective to improve the performance of ICT via spatial prediction. Fig. 9a shows the MAEs of ICT with and without the help of spatial prediction. For comparison, we also plot the MAEs using direct transfer. ICT without spatial prediction already decreases the MAEs from 28 to 11 compared with direct transfer. ICT with the help of spatial prediction can further reduce the MAEs by 12%.

We are also interested in how many spatial prediction results are necessary to improve the performance of ICT. We select different fractions of predicted $PM_{2.5}$ concentration level Y_t' by changing the threshold τ and forming Y_t^τ . Note that the smaller τ is, the higher average accuracy of Y_t^τ is. By setting $\tau = 100$, we select around 25% of all predicted $PM_{2.5}$ concentration level in the location where $S6$ is installed, which yields an overall accuracy of 0.96. Then we select different sizes of Y_t^τ and evaluate the MAEs of ICT.

Fig. 9b shows the MAEs to transfer calibration parameters from different sources to $S6$ using ICT with different sizes of Y_t^τ , ranging from 5% to 25% (the percentage represents $\|Y_t^\tau\|/\|Y_t'\|$). When the size of Y_t^τ increases from 5% to 15%, the MAE decreases. However, the decreasing of MAE stops after that. This suggests 15% of the most accurate predictions from Y_t' is sufficient.

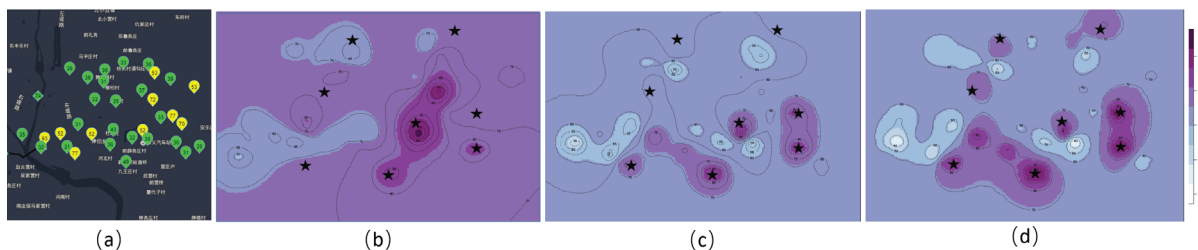


Fig. 11. (a) Locations of 30 low-cost sensors within a $5\text{km} \times 5\text{km}$ area around S6 and heat maps generated with measurements of the 30 sensors calibrated by using (b) direct transfer, (c) TCA and (d) ICT. The ground truth pollution source locations are marked by stars.

5.3.5 Visualization of Transferred Measurements. To get a deeper understand of the cause behind the varying results of four approaches, we use principal component analysis (PCA) to illustrate the difference between the transformed measurements of source sensor S_7 and target sensor S_6 , *i.e.*, X_s and $F(X_t)$. In order to enable a visualization of the results, we choose the two largest components and show them in 2 dimension figures in Fig. 10. Since we have already empirically proved that the distribution of ground truth in both source and target locations are similar in Fig. 5, *i.e.*, $p(Y_s) \approx p(Y_t)$, and the same calibration model is applied on both the transformed measurements of source and target sensor, we can reasonably assume that the overlapping area of the largest two components is positively correlated to the calibration transfer accuracy.

In direct transfer, since no transformation are made, *i.e.*, $F(X_t) = X_t$, the PCA results directly represents the original measurements X_s and X_t , as shown in Fig. 10a. There are obvious shift and misalignment shown in the figure, which could explain the reason why the performance of direct transfer is limited.

Fig. 10b shows the PCA components visualization of TCA transfer method. TCA tries to correct the shift and misalignment between the measurement features by mapping original features into a new reduced feature space. In the new feature space, the two largest PCA components of $\Phi(X_s)$ and $\Phi(X_t)$ have more overlapping area, which could explain the reason of accuracy improvement. However, there are still large non-overlapping area.

Fig. 10c and Fig. 10d shows the PCA components visualization of NZ-ICT and ICT, which do not reduce dimensions and use the original measurement features to find the linear transformation. We can see that in Fig. 10d, the overlapping area between the components of X_s and $F(X_t)$ is much larger compared to direct transfer and TCA method. Notice that if only the near zero data is used (NZ-ICT), the overlapping area is smaller and has a visible shift compared with the full ICT.

5.4 Case Study: Pollution Source Location Inference

Due to lack of ground truth $PM_{2.5}$ concentrations from co-located reference stations, it is difficult to evaluate the performance of ICT on a sensor deployed at an arbitrary location in Beijing. Alternatively, this subsection aims to *indirectly* assess the performance of different algorithms via a case study, in which calibration parameters are transferred from a single source sensor to tens of target sensors in arbitrary locations within a certain range. Specifically, we apply different calibration transfer methods on the raw sensor readings, and compare their performance to infer the locations of pollution sources from the calibrated sensor readings. The rationale is that pollution concentrations should be high at locations close to the pollution sources. Therefore an intuitive way to locate pollution sources is to firstly generate a heat map of air pollution concentration from the sensor measurements, then find locations/areas where the concentration peaks as potential pollution source locations.

The accuracy of pollution source localization is correlated to the accuracy of the calibrated sensor measurements, thus an indirect assessment of the effectiveness of different sensor calibration algorithms.

We conduct a case study in a $5\text{km} \times 5\text{km}$ area around S6, where 30 low-cost sensors (including S6) are deployed. Fig. 11-(a) shows the locations of the 30 low-cost sensors. The squared spot is the location of the high-cost reference station co-located with S6. We focus on this area because we have access to the ground truth locations of the pollution sources of $PM_{2.5}$ within this area, which is generally inaccessible for other areas in Beijing. We use the pollution source locations as the ground truth for pollution source location inference. We perform calibration transfer using three methods: (i) direct transfer, (ii) TCA and (iii) ICT. Then we average the calibrated sensor measurements over one month and use the gaussian process regression model in [4] to generate the heat map of the area. The locations of the ground truth pollution sources within this area are marked by stars. Ideally, the peaks (high concentration locations) in the heat map should match with the pollution source locations.

Fig. 11-(b), Fig. 11-(c) and Fig. 11-(d) show the heat maps generated by applying direct transfer, TCA and ICT for sensor calibration. By comparing the highly polluted locations in the heat maps to the ground truth pollution source locations, we observe that the heat map generated by sensor measurements calibrated by direct transfer is able to correctly locate three out of the eight pollution sources, while five are located when applying TCA. With ICT, however, all eight pollution sources are correctly located. These results indicate that ICT outperforms direct transfer and TCA in calibration accuracy on these 30 sensors.

6 CONCLUSION

In this work we propose In-field Calibration Transfer (ICT), a calibration scheme that transfers the calibration parameters of sensors with access to references (source sensors) to those without access to references (target sensors). It is challenging to derive such a transformation between the source and target sensors installed at different locations because their measurements are unsynchronized. On observing that (i) the distributions of ground truth in both source and target locations are similar and (ii) the transformation is approximately linear, ICT learns the transformation based on the similarity of distributions with a novel optimization formulation. The performance of ICT is further improved by using spatial prediction of air quality level as an aid for calibration transfer task, and using ensemble techniques to enable multi-source calibration transfer. Experiments show that ICT is able to provide approximately equal calibration performance as if the target sensors have direct access to references. We believe ICT notably increases the usability of large-scale air pollution monitoring deployments.

In the future, we plan to explore the possibility of conduct ICT on larger scale scenarios *e.g.*, inter-city calibration transfer. We also plan to extend ICT for moving sensors by augmenting the measurements of the moving sensor collected from multiple locations to perform multi-source multi-target calibration transfer.

ACKNOWLEDGMENTS

This work was funded by the Swiss National Science Foundation (SNSF) under the FLAG-ERA CONVERGENCE project.

REFERENCES

- [1] Beijing Municipal Environmental Monitoring Center. 2018. Beijing Municipal Environmental Monitoring Center Official Website. <http://www.bjmecm.com.cn/>.
- [2] Colin Bellinger, Mohamed Shazan Mohamed Jabbar, Osmar Zaiane, and Alvaro Osornio-Vargas. 2017. A Systematic Review of Data Mining and Machine Learning for Air Pollution Epidemiology. *BMC Public Health* 17, 1 (2017), 907.
- [3] Elena Boldo, Sylvia Medina, Alain Le Tertre, Fintan Hurley, Hans-Guido Mücke, Ferrán Ballester, Inmaculada Aguilera, et al. 2006. Aphis: Health Impact assessment of Long-term Exposure to PM 2.5 in 23 European Cities. *European Journal of Epidemiology* 21, 6 (2006), 449–458.
- [4] Yun Cheng, Xiucheng Li, Zhijun Li, Shouxu Jiang, Yilong Li, Ji Jia, and Xiaofan Jiang. 2014. AirCloud: a Cloud-based Air-Quality Monitoring System for Everyone. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems*. 251–265.

- [5] J Fonollosa, L Fernández, Agustín Gutiérrez-Gálvez, Ramón Huerta, and S Marco. 2016. Calibration Transfer and Drift Counteraction in Chemical Sensor Arrays using Direct Standardization. *Sensors and Actuators B: Chemical* 236 (2016), 1044–1053.
- [6] Kaibo Fu, Wei Ren, and Wei Dong. 2017. Multihop Calibration for Mobile Sensing: k-hop Calibratability and Reference Sensor Deployment. In *Proceedings of the IEEE International Conference on Computer Communications*.
- [7] Liang Ge, Jing Gao, Hung Ngo, Kang Li, and Aidong Zhang. 2014. On Handling Negative Transfer and Imbalanced Distributions in Multiple Source Transfer Learning. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 7, 4 (2014), 254–271.
- [8] Bin Guo, Jing Li, Vincent W Zheng, Zhu Wang, and Zhiwen Yu. 2018. CityTransfer: Transferring Inter-and Intra-City Knowledge for Chain Store Site Recommendation based on Multi-Source Urban Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 135:1–135:22.
- [9] Baidu Inc. 2018. Baidu POI. <http://lbsyun.baidu.com/index.php?title=androidsdk/guide/search/poi>.
- [10] Wan Jiao, Gayle Hagler, Ronald Williams, Robert Sharpe, Ryan Brown, Daniel Garver, Robert Judge, Motria Caudill, Joshua Rickard, Michael Davis, et al. 2016. Community Air Sensor Network (CAIRSENSE) Project: Evaluation of Low-cost Sensor Performance in a Suburban Environment in the Southeastern United States. *Atmospheric Measurement Techniques* 9, 11 (2016), 5281–5292.
- [11] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. 2004. Detecting Change in Data Streams. In *Proceedings of the International Conference on Very Large Databases*. 180–191.
- [12] Yuxiang Lin, Wei Dong, and Yuan Chen. 2018. Calibrating Low-Cost Sensors by a Two-Phase Learning Approach for Urban Air Quality Measurement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 18:1–18:18.
- [13] Xiao Liu, Sitian Cheng, Hong Liu, Sha Hu, Daqiang Zhang, and Huansheng Ning. 2012. A Survey on Gas Sensing Technology. *Sensors* 12, 7 (2012), 9635–9665.
- [14] Balz Maag, Zimu Zhou, Olga Saukh, and Lothar Thiele. 2017. SCAN: Multi-Hop Calibration for Mobile Sensor Arrays. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 19:1–19:21.
- [15] Balz Maag, Zimu Zhou, and Lothar Thiele. 2018. A Survey on Sensor Calibration in Air Pollution Monitoring Deployments. *IEEE Internet of Things Journal* (2018).
- [16] Michael Mueller, Jonas Meyer, and Christoph Hueglin. 2017. Design of an Ozone and Nitrogen Dioxide Sensor Unit and Its Long-term Operation within a Sensor Network in the City of Zurich. *Atmospheric Measurement Techniques* 10, 10 (2017), 3783–3799.
- [17] Ministry of Ecology and Environment. 2018. Ambient air quality standards. <https://tinyurl.com/ybqnswnv>.
- [18] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. Learning and Transferring Mid-level Image Representations using Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1717–1724.
- [19] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. 2011. Domain Adaptation via Transfer Component Analysis. *IEEE Transactions on Neural Networks* 22, 2 (2011), 199–210.
- [20] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2010), 1345–1359.
- [21] Olga Saukh, David Hasenfratz, and Lothar Thiele. 2015. Reducing Multi-hop Calibration Errors in Large-scale Mobile Sensor Networks. In *Proceedings of the ACM International Conference on Information Processing in Sensor Networks*. 274–285.
- [22] The Public Weather Service Center of CMA. 2018. Beijing Weather. <http://bj.weather.com.cn/>.
- [23] Wataru Tsujita, Akihito Yoshino, Hiroshi Ishida, and Toyosaka Morizumi. 2005. Gas Sensor Network for Air-Pollution Monitoring. *Sensors and Actuators B: Chemical* 110, 2 (2005), 304 – 311.
- [24] Ying Wei, Yu Zheng, and Qiang Yang. 2016. Transfer Knowledge Between Cities. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1905–1914.
- [25] World Health Organization. 2018. Ambient Air Pollution - a Major Threat to Health and Climate. <http://www.who.int/airpollution/ambient/en/>.
- [26] Yun Xiang, Lan Bai, Ricardo Piedrahita, Robert P. Dick, Qin Lv, Michael Hannigan, and Li Shang. 2012. Collaborative Calibration and Sensor Placement for Mobile Sensor Networks. In *Proceedings of the ACM International Conference on Information Processing in Sensor Networks*. 73–84.
- [27] Xiangxiang Xu, Xinlei Chen, Xinyu Liu, Hae Young Noh, Pei Zhang, and Lin Zhang. 2016. Gotcha II: Deployment of a Vehicle-based Environmental Sensing System. In *Proceedings of the ACM Conference on Embedded Network Sensor Systems*. 376–377.
- [28] Ke Yan, Lu Kou, and David Zhang. 2018. Learning Domain-Invariant Subspace Using Domain Features and Independence Maximization. *IEEE Transactions on Cybernetics* 48, 1 (2018), 288–299.
- [29] Ke Yan and David Zhang. 2016. Calibration Transfer and Drift Compensation of E-noses via Coupled Task Learning. *Sensors and Actuators B: Chemical* 225 (2016), 288 – 297.
- [30] YunTong. 2018. YunTong. <http://www.yuntongkeji.com/uploadfile/2017128174038670.pdf>.
- [31] Lei Zhang, Fengchun Tian, Chaibou Kadri, Bo Xiao, Hongjuan Li, Lina Pan, and Hongwei Zhou. 2011. Online Sensor Calibration Transfer among Electronic Nose Instruments for Monitoring Volatile Organic Chemicals in Indoor Air Quality. *Sensors and Actuators B: Chemical* 160, 1 (2011), 899 – 909.

Received August 2018; revised November 2018; accepted January 2019