Title

Automating the interpretation of PM2.5 time-resolved measurements using a data-driven approach

Permalink

https://escholarship.org/uc/item/9vc7k29j

Journal

Indoor Air, 31(3)

ISSN

0905-6947

Authors

Tang, Hao
Chan, Wanyu Rengie
Sohn, Michael D

Publication Date

2021-05-01

DOI

10.1111/ina.12780

Peer reviewed

Automating the interpretation of $PM_{2.5}$ time-resolved measurements using a data-driven approach

Running title: Interpreting $PM_{2.5}$ data with machine learning

Hao Tang [a], Wanyu Rengie Chan [b], Michael Sohn [b*]

[a] Joint International Research Laboratory of Green Buildings and Built Environments,
Chongqing University, Chongqing 400045, China
[b] Indoor Environment Group, Energy Technologies Area, Lawrence Berkeley National
Laboratory, Berkeley CA 94720, USA
[*] *Corresponding email: mdsohn@lbl.gov*
Corresponding phone: 510-486-7620

Acknowledgement

This research was supported by the California Energy Commission via Contract No. PIR-16-012. Funding was also provided by the Building America Program and the U.S.-China Clean Energy Research Center, both within the U.S. Department of Energy Building Technologies Program, Office of Energy Efficiency and Renewable Energy under DOE Contract No. DE-AC02-05CH11231. Involvement of Hao Tang was financially supported by China Scholarship Council.

Abstract

The rapid development of automated measurement equipment enables researchers to collect greater quantities of time-resolved data from indoor and outdoor environments. While significant, the interpretation of the resulting data can be a time-consuming effort. This paper introduces an automated process of interpreting $PM_{2.5}$ time-resolved data and differentiating $PM_{2.5}$ emissions resulting from indoor and outdoor sources. We use Random Forest (RF), a machine learning approach, to study a dataset of 836 indoor emission events that occurred over a two-week period in 18 apartments in California. In this paper, we show model development and evaluate its performance as the sample size and source vary. We discuss the characteristics of dataset that tended to help the source identification and why. For example, we show that data from many events and from different apartments are essential for the model to be suitable for analyzing the new separate dataset. We also show that longitudinal data appears to be more helpful than the time frequency of measurements in a given apartment. We use the resulting RF model to analyze $PM_{2.5}$ data of an entirely separate dataset collected from 65 new homes in California. The RF model identifies 442 indoor emission events, with a few misidentifications.

Keywords: Machine learning; $PM_{2.5}$; time-resolved measurement; indoor emission; random forest; residential

Practical implications: This paper illustrates the development and use of a rapid and automatable approach to (i) differentiate between indoor and outdoor sources of $PM_{2.5}$ emissions, (ii) determine the start and end of indoor sources, and (iii) compute emissions from indoor events. We demonstrate the use of random forest, a machine learning method to interpret data from two datasets containing $PM_{2.5}$ measurements from a total of 18 apartments and 65 homes. We also discuss the particular characteristics of the time-series data that are most helpful in the event reconstruction. This method can be applied to other datasets when both indoor and outdoor $PM_{2.5}$ were monitored concurrently, and preferably from a large number of buildings.

Conflict of Interest or Acknowledgments: The authors have no conflict of interest to declare.

Data Availability Statement: The data that support the findings of this study are openly available in Dryad at https://doi.org/10.7941/D1HG9J.

1 Introduction

Fine particles, especially for particulate matter with diameters 2.5 μm and smaller ($PM_{2.5}$), are one of the most studied indoor air contaminates due to their association with health effects, such as lung function reduction[1,2], cardiovascular disease[3–5] and respiratory symptoms[6,7]. Fine penetration particles indoors can originate from both indoor and outdoor sources. Indoor sources include smoking, cooking, cleaning activities, and burning incenses and candles[8]. Outdoor sources include vehicle emissions and stationary sources.[9] The relative contributions of ambient and indoor-generated $PM_{2.5}$ vary widely across homes.[10–12] It is important to study and understand PM emission events resulting from indoor and outdoor sources in order (1) characterize the exposure of occupants to PM from indoor versus outdoor sources, (2) weigh the benefits of mitigation strategies to reduce outdoor infiltration versus indoor sources, for example improving filtration versus increasing ventilation at the source, (3) improve the interpretation of high volumes of experimental PM data during a typical experiment.

Several efforts have been made to differentiate the impact of outdoor sources on indoor $PM_{2.5}$ levels. Experimental methods typically involve studying the chemical characterization of trace elements to differentiate the typical characteristics from indoor and outdoor sources.[13,14] While significant, these methods are laborious and the results are coarse in time resolution. Computational methods for differentiating indoor emission and outdoor contributions involve characterizing the time-resolved behavior of typical indoor and outdoor pollutants. For example, $PM_{2.5}$ that originated from indoor processes disperse into the indoor air directly and thus lead to an acute rise in the indoor concentrations, whereas outdoor $PM_{2.5}$ infiltrates into indoor spaces via ventilation and infiltration and thus tend to raise indoor concentration more gradually.[15–17]

Chan et al. (2017) characterized indoor emission events from indoor and outdoor time-resolved $PM_{2.5}$ concentrations collected from 18 low-income apartments in California using a rule-based approach.[18] Their rules included identification of indoor $PM_{2.5}$ peaks, functional characteristics of the start and end times, and the shapes of indoor concentration decays. They identified indoor events of sufficient emissions that resulted in a rise of indoor $PM_{2.5}$ of at least 5 $\mu g/m^3$. Successions of indoor emissions were determined if they belong to the same event or not by rules – a comparison of the decay in $PM_{2.5}$ with respect to the prior peak concentration. Chan et al., note that the process was laborious because many of the rules required manual adjustments that could not be easily automated.

In this study, we developed an automated approach that can be used to interpret indoor $PM_{2.5}$ data from a broad number of experimental studies, including understanding the important characteristics of measurements that tends to improve source attribution and characterization. The scientific contributions of this paper are to:

1. Present a comprehensive first use of an existing method for interpreting $PM_{2.5}$ data, and to demonstrate the degree that the approach, Random Forest, improves upon the existing methods for source characterization, such as the rule-based method.
2. Identify features of the data that appear to be particularly helpful for source apportionment and characterization.
3. Discuss the effect of sample size and sample source on the performance of the machine learning model and how that reflects our knowledge of buildings.

2 Method

The goal of this study is to demonstrate an automated approach to identify emission events resulting from indoor sources by interpreting indoor and outdoor $PM_{2.5}$ time-resolved measurements. Based on the previously-analyzed data of a study in low-income apartments, a data-driven method was applied to simplify the process by constructing a machine learning model to substitute a rule-based technique. The proposed model is then applied to a separate set of data collected from 65 new single-family homes. We identified emission events using outputs from the machine learning model, and visually inspected results to determine if the model developed from low-income apartments sufficiently captures emission events that likely occurred in the new single-family homes. Finally, we used a simple mass balance equation to calculate emission rates, and compared the simulated concentrations with measured data to show how outputs from the machine learning model can be used to analyze $PM_{2.5}$ data and generate emission event characteristics. The raw data and code used in this study can be freely accessed through Dryad. [19]

2.1 Description of Data

Two sets of time-resolved $PM_{2.5}$ data were gathered for the model development, performance assessment, and application. Dataset one, which we call the training dataset, was collected from two-weeks of monitoring data in 18 low-income apartments in California.[18] This data is used to develop a source apportionment model and to test its performance. Dataset two is collected from 65 new California single-family homes[20], and is used for the application. **Table 1** describes the broad content of the two datasets.

In dataset one, indoor and outdoor $PM_{2.5}$ concentration was measured with light-scattering monitors (DustTrak II 8530; TSI Inc., Shoreview, MN, USA) at a 2-min interval. The indoor monitor was placed in the center of the apartment – typically the living room or dining room. The measured concentrations were multiplied by a factor of 0.4 to correct for DustTrak scaling adjustment, then smoothed using the MALDIquant Savitzky Golay method, following the same method listed by Chan et al.[18] **Figure 1** presents an example of a typical time series consisting of emissions and decay periods identified using a rule-based method.[18]

In dataset two, indoor and outdoor $PM_{2.5}$ concentrations were measured at 1-min intervals using a different set of monitors: MetOne ES-642 and BT-645 instruments (Met One Instruments Inc., Grants Pass, OR, USA).[20] Both instruments are light-scattering monitors that work on a similar operation principle as the DustTrak. The indoor concentrations were measured centrally in a large open room (e.g. dining or living room). Following the same method described in Singer et al.[20], indoor $PM_{2.5}$ concentrations were adjusted by a multiplier of 1.23 based on gravimetric filter measurements collected in a subset of the homes, and outdoor $PM_{2.5}$ concentrations were not adjusted. Measurements were smoothed using the same approach described for dataset one, so that results could be comparable to a rule-based method.[18]

2.2 Method of analysis

Our objective is to classify whether a given data point was taken during an emission event. We chose to use a Random Forest (RF), due to its potential for good prediction accuracy and resistance to overfitting, compared to single-decision trees. Since it was introduced in 2001[21], RF has become one of the more popular decision tree-based approach for ensemble learning. However, only a few applications have been reported in IAQ studies of predicting indoor $PM_{2.5}$ concentration[22,23] and indoor radon[24]. While significant, these applications do not explore how to identity periods of emissions or differentiate between indoor and outdoor sources.[25]

We refer the reader to [21,26,27] for a broad description of the approach. Here, we provide an overview. The two basic concepts of a RF model are classification decision trees and characteristics of the data, called features, that can be used for classification. Features are the information about a measurement that helps to explain the data, in our case the concentrations of the $PM_{2.5}$. A classification decision tree is a classifying approach consisting of a number of nodes and branches. Each node consists of a characterization of a feature used to split the data into two smaller groups in order to minimize the within-group variance. For example, a decision tree could classify a group of data points into two groups of which one has higher indoor $PM_{2.5}$ concentration and one has lower indoor $PM_{2.5}$ concentration, where the indoor $PM_{2.5}$ concentration is a "feature" of the data. The goal of a decision tree is to describe each data point using a sequence of feature descriptions. Because the sequence of selecting features conditions is not unique, the RF model generates a large number of decision trees in order to express the full possible sequences of features to explain a datapoint. The predominant classification of all the decision trees becomes the final prediction of the RF model. In the demonstration that follows, we describe the possible features that are germane to $PM_{2.5}$ interpretation.

We use the Mtry package 'randomForest' in R[26] for selecting the number of features to consider for each decision node. For the applications here we used four features. We tested a range of quantities and found that anywhere from 3 to 6 features produced the same results.

Another tuned parameter is the number of classification trees used in the RF model. A large number of trees helps to explore the many possible sequences of features that could classify a data point, while recognizing that too many trees is computationally expensive with no added benefit. A common metric for assessing model performance is the "out of bag error rate (OOB)".[21] The OOB is a measure of performance of the classification for data that are not used in a particular decision tree. High OOB means that the model performed poorly in classifying the unused data for a given tree. **Figure 2** illustrates the OOB error rate versus the number of trees in the forest for dataset 1. It shows that 200 trees are sufficient for classification. However, we used 400 trees in order to assuage any concerns with insufficient event characterization. The time needed to run 200 versus 400 trees is negligible; approximately 5 min of additional computational time.

3 Feature selection for RF model

The first step of the model development is to compute all of the possible features that might be helpful in describing whether a measurement was taken during an emission event. For example, **Table 2** presents the possible features which may contain information for the classification process, and **Figure 3** explains the features.

The backward- and forward-difference features are the difference between the concentration of the datapoint in relation to the value right before and after it. For example, if the previous measurement is less than the present data point, it might be expressive of an event since concentrations tend to increase during an emission event. Conversely, when the next measurement is less than the present data point, it might be expressive of a source not being present since concentrations tend to decrease after a source event due to loses.

A feature for an extreme point is a local maximum or minimum in time-resolved measurements. Any two adjacent extreme points are the start and end (or end and start) of an increment or decrement in measurement, and thus can be the possible start and end (or end and start) points of an emission or decay process. The concentrations at two adjacent extreme points provide information on whether the increment or decrement is significant enough to be an emission event or decay. For example, a small increment in indoor concentration less than 5 μg/m^3 may not be significant enough to be classified as an indoor emission event. Features of adjacent extreme points may help the model reduce false positive errors by excluding minor changes in the concentrations.

Features that are included based on outdoor concentrations provide information for assessing the potential impact of outdoor conditions on indoor concentrations, and thus may help differentiate the impact of indoor emissions from outdoor air. Indoor concentrations which peak at concentrations higher than outdoor concentrations are more likely to result from indoor emission sources.

Table 2 shows the list of features we considered in our training dataset. Although any number or type of features could be considered, one must use judgement to avoid repeating the information content across the features. Too many features could cloud the interpretation of the importance of a feature during the performance assessment by diluting the information across many related features. Additionally, an excess number of the same correlated features can result in an improved fit to training data but would not improve the predictive power of the model.

To assess the importance of each feature in the classification, we computed a metric called the mean decrease in accuracy (MDA), as shown in Eq. 1. The MDA is the average decrease in accuracy with and without a feature divided by the standard deviation of the accuracy. The MDA shows the benefit of a feature relative to the overall performance of the model. Features that have higher MDA are expected to be more important for the model.

$$MDA(Feature\ A) = \frac{Mean(Decreases\ in\ accuracy\ of\ trees\ by\ permuting\ A)}{StandardDeviation(Decreases\ in\ accuracy\ of\ trees\ by\ permuting\ A)} \qquad (1)$$

Another common index for assessing importance of features is the mean decrease in “Gini”. However we did not use here because it has been reported to favor features which are continuous over categorical data.[27]

The procedure of selecting features is summarized as following:

1. Construct the RF model with all features and calculate the MDA;
2. Remove the features with lowest MDA one by one and output OOB error for the new model.

3. If the OOB error increases significantly after being removed, we added it back; otherwise it was removed.
4. Repeat steps 3 and 4 until no feature could be removed without weakening the prediction accuracy of model.

This procedure yielded two sets of features for the prediction of emission and decay periods, as shown in **Table 2**. The identification of decay periods utilizes the model results of whether a data point is classified as part of an emission event or not. 14 features were used for emission identification and 9 features were used for decay identification. We discuss these features, and their importance in Section 4.4.

4 Result and discussion

4.1 Performance of RF model

Typically, in RF applications, all of the data are used to construct the model, and the primary measure of model performance is the OOB error rate. **Table 3** shows the performance of the model using this approach. In this table, 'Yes' and 'No' refers to if a data point was identified as during an emission or decay period. The error rate of each category was calculated by dividing the number of incorrect predictions by the total number of prediction opportunities. Similarly, the overall error rate of the RF model was calculated by dividing the sum of incorrect predictions by the total number of predictions. The overall model accuracy is approximately 99.5% for identifying emission and 99.3% for identifying decay. These results indicate a high prediction accuracy from the RF model.

We also compared the class error rates in **Table 3** from one another. Class error rates that differ greatly from one another would indicate, for example, a tendency for a model to classify a data point as the major category to reduce the overall error rate. This is a common concern for machine learning methods and can have detrimental effects on convergence in training and application.[28] In this case, judging from the fact that all class errors are small (5% or lower), we did not observe large differences among them.

Other tree-like classifier such as gradient boosting [29] can be another option. Here, we focus on exploring, and demonstrating, the potential of classification-type machine learning approaches. This is in part because the choice in classification method is largely to improve computational speed. Here, although the number of datapoints is quite high for indoor experimental work, these existing methods were designed to manage several thousands of data points, hence they are all quite fast for our intended applications. In this paper we also focus on discussing the importance of feature selection, sample size and sample source which provide important insights that apply to all machine learning methods employed in their practice.

4.2 Sample size and model performance

In order to further review the performance of the model, including evaluating the influence of sample size on model performance, we chose to construct models which were trained on a fraction of dataset 1 and then tested it against the remainder of the data. The performance of the RF models built on different percentages of the training dataset (50%-50%, 70%-30%, 90%-10% and 100%-0%) are shown in **Figure 4**. We see that the error on both training data set and test dataset decreases with increasing amount of training data, indicating that a large training dataset yields a model with the higher prediction accuracy. Notice also that there is no significant difference between the OOB estimated error on training dataset and the error based on the test dataset. This suggests that the OOB error rate estimated using the training dataset is not biased when we randomly assigned as training data or test data.

4.3 Sample source and model performance

In assessing the performance of the RF model, we wanted to test the ability of the model to be trained using data from one or more buildings and then to use the resulting model to interpret data from an entirely different building. The data used for the training came from 18 low-income apartments inside 3 different buildings that were retrofitted for energy efficiency. The 3 buildings differ by location and the energy retrofit measures implemented. We constructed three different RF models by training the model on data from two buildings. We then tested its performance using data from the third building. Unsurprisingly, we found that the performance of the model depended heavily on whether the training data included similar patterns in the data features as in the test data (**Figure 5**). For example, in general the errors are higher for all cases than the errors shown in **Figure 4**. Having a restricted amount of data, but from a broad range of buildings, always performed better than having a lot of data from a few buildings to interpret data from another building. Nonetheless, the errors are not necessary unsatisfactorily high. In Section 5, we test an RF model developed from all of dataset 1 and test it against another set of buildings (dataset 2) that are very different in characteristics, to illustrate an application of the RF model.

4.4 Relative importance of features

An inherent procedure in RF model is estimating the relative importance of input features. **Figure 6** shows the mean decrease in accuracy with each of the features omitted. As mentioned previously, two features could be highly correlated. The correlated features would share the importance and thus reduce the relative importance of each of them.[30] We tried to minimize the number of these features in the model since they generally did not offer much benefit to the models. On the other hand, there are closely related features that we retained in the model because they each contain important information. For example, the features DE (indoor concentration difference between next extreme point and last extreme point), NE (indoor concentration at next extreme point) and LE (indoor concentration at last extreme point) are closely related, but together still appear to be helpful for model performance.

DE was found to be the most important feature in identification of emissions because it can provide information on whether increment of concentration is significant. For identification of decays, E (A binary index determined by if the point is (0) or is not (1) included in an emission event) was found to have the most importance as a point included in an emission should not be included in a decay.

5 Application using Dataset 2

The established RF model was applied to the study of single-family homes (dataset 2). Constructing a model using one dataset and applying it to study a completely different dataset and study area helps us assess the overall viability of this approach.

A total of 442 indoor $PM_{2.5}$ emissions was identified by the RF model in 65 California new homes, each monitored for a weeklong period. The model execution was completed on a common personal computer in less than 5 minutes. **Figure S1** presents the cumulative percentage of number of indoor emission events identified in each home during the weeklong period. The mean number of identified indoor emission events per home was seven and the median was four.

5.1 Identify performance of RF model on dataset 2 with visually review

We visually reviewed identified emission events on dataset 2 to evaluate how well the RF model performed. We present some patterns of identified emission and decay periods below to illustrate.

Figure 7 (a) presents a typical emission event with a significant rise in concentration and a smooth decay. This pattern can be found in most of the identified emission events. Overall, the RF model was found to perform well in identifying this type of emission events. Occasionally, there are some cases which the concentrations tended to have very sharp increases where the identified start and end of emission period did not quite align with the rise and fall in concentration, caused by the smoothing of $PM_{2.5}$ data. **Figure 7 (b)** shows an example where the RF model determined an early start in emission occurring before the rise in indoor concentration, and a delay in the end of the emission event.

We found that the RF model successfully identified linked emission events, defined as consecutive emission periods that shared a common decay period. This is an important advantage of using the RF model over the rule-based approach that requires additional parameters to specify how to link emission events. An example is shown in **Figure 7 (c)**. In this case, the RF model identified two emission periods rather than identifying them as one combined emission period. The linked emission events allow subsequent determination of emission characteristics, such as the emission rate, to more accurately describe the measured indoor concentration. In comparison, treating the linked emission events as one continuous event with a constant source would agree poorly with measurements.

We introduced features computed based on outdoor concentration to enable the RF model to differentiate emission events resulting from indoor and outdoor sources. The RF model performed well in terms of identifying and excluding emissions from outdoor air that consisted of high $PM_{2.5}$ concentrations. An example is shown in **Figure 7 (d)**, where the high indoor concentrations followed the same trend as the high outdoor concentrations. The RF model correctly excluded this period as an indoor emission event.

A key limitation of any machine learning approach is that it can only identify patterns that share similar features contained in the training dataset. There are a small number of emission events in dataset 2 that were not present in dataset 1. For example, **Figure 8 (a)** shows a sudden increase in outdoor concentrations, which is due to the use of an outdoor grill as reported by the occupants. There is a subsequent "indoor" emission event identified by the RF model, even though the source was originated outdoors. The identification of this "indoor" emission event is still useful, but one could risk misinterpretation of the results if the differences between dataset 1 and 2 were not carefully taken into consideration.

Figure 8 (b) shows another example of emission events that occurred in dataset 1 but only very rarely. The event is characterized by a small increase in PM that lasted many hours (from 7 μg/m^3 to 23 μg/m^3 in about 5 hours), after that the indoor concentration sustained at that level for many hours (23 to 25 μg/m^3 for another 5 hours), and finally the indoor concentration decreased. Throughout this period, the outdoor concentration remained constant at a low level. Because there were too few of these types of events present in the training dataset, the RF model identified sporadic periods as emission events in **Figure 8 (b).**

5.2 Characterization of indoor emission identified in dataset 2

We used a first-order mass balance model, the same as used in the rule-based paper [18], to estimate the loss rate and emission rate. **Figure 9** illustrates the process (see Supporting Information for more detail).

We selected events with an emission period that occurred long enough (at least 8 minutes) for estimating emission rates. Decay rates were computed for ones that lasted at least 20 minutes, and occurring within two hours after emission event concluded. 398 out of 442 identified emission periods were used to estimate emission rates. It should be noted that one decay period could be corresponding to multiple emissions because of linked emission events (see **Figure 7 (c)** for example).

Table 4 summarizes the resulting estimates. In general, the average R square is 0.92 for fitting the model to estimate the loss rate and 0.88 for fitting the model to estimate the emission rate. The average RSME (root mean square error) between measured and estimated $PM_{2.5}$ concentration is 5.2 for decay periods and 16.6 for emission periods. We note that in some cases, the R square is lower than 0.7, which we attribute to irregular patterns in the time series. Such incidences may be due to imperfect mixing in the room or, intermittent emissions behavior, or unstable sensor measurement. As a check, we compared the simulated concentrations using the fitted emission and decay rates with measured concentrations. For the 53 homes with identified emission events, we found good agreement between the simulated and measured mean concentrations during the emission periods (see Supporting Information). However, an important weakness of this analysis is that we lack ground truth of when indoor emissions occurred in these homes, so model performance is not precisely evaluated.

The characteristic of the identified indoor emission event in 18 low-income apartments is presented as comparison in **Table 4**. We found that the identified indoor emission events in 65 new homes have significantly lower emission rates and total emitted mass compared to the 18 low-income apartments, even though their durations are similar. It should be noted that the apartments (23.2 events per home week) had approximately four times the number of identified indoor emission events than the homes (6.8 events per home week). While it is beyond of scope of this paper, further assessments of additional data collected for the 65 homes, such as self-reported activities by occupants including cooking and window uses, will be helpful to verify these results.

6 Conclusions

This study demonstrated the construction of a supervised machine learning model to represent time-resolved measurements of indoor and outdoor $PM_{2.5}$ in 18 low-income apartments for the identification of indoor emission events. We found that the machine learning method is a promising substitution of the more common rule-based methods and manual handling because it may produce highly consistent results and requires much less time and ad-hoc human interpretation of a large dataset. The overall accuracy of resulting model was 99.5% for identification of emission and 99.3% for identification of decay. The high prediction accuracy benefits from the feature selection by describing the position of each data point in the time-resolved measurement with difference values, extreme points and outdoor conditions. By exploring the performance of the model on different sample sizes and sample sources, we found that the model could benefit most from a larger training dataset consisting of multiple sample sources, where the diversity of the sample source is more significant. We applied the proposed model to a dataset collected from 65 new single-family homes and characterized the 442 identified emission events using a first-order mass balance model. The resulting predictions of the number of emission events in the new single-family homes were found to be three quarters less frequent compared to identified events in the low-income apartments; the mean emitted mass of the identified emission events was only half as many. In future work, machine learning model performance can be evaluated by applying to studies that measure indoor and outdoor $PM_{2.5}$ using scripted activities. Once a model has been well validated, the predicted emission events can be cross-referenced with occupant activities to better characterize the wide range of $PM_{2.5}$ sources in buildings.

Reference

1. Isiugo K, Jandarov R, Cox J, Ryan P, Newman N, Grinshpun SA, Indugula R, Vesper S, Reponen T. Indoor particulate matter and lung function in children. *Sci Total Environ*. 2019;663:408-417.
2. Tomczak A, Miller AB, Weichenthal SA, To T, Wall C, van Donkelaar A, Martin R V., Crouse DL, Villeneuve PJ. Long-term exposure to fine particulate matter air pollution and the risk of lung cancer among participants of the Canadian National Breast Screening Study. *Int J Cancer*. 2016;139:1958-1966.
3. Rich DQ, Zhang W, Lin S, Squizzato S, Thurston SW, van Wijngaarden E, Croft D, Masiol M, Hopke PK. Triggering of cardiovascular hospital admissions by source specific fine particle concentrations in urban centers of New York State. *Environ Int*. 2019;126:387-394.
4. Brook RD, Rajagopalan S, Pope CA, Brook JR, Bhatnagar A, Diez-Roux A V., Holguin F, Hong Y, Luepker R V., Mittleman MA, Peters A, Siscovick D, Smith SC, Whitsel L, Kaufman JD. Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the american heart association. *Circulation*. 2010;121:2331-2378.
5. Brook RD, Franklin B, Cascio W, Hong Y, Howard G, Lipsett M, Luepker R, Mittleman M, Samet J, Smith SC, Tager I. Air pollution and cardiovascular disease: A statement for healthcare professionals from the expert panel on population and prevention science of the American Heart Association. *Circulation*. 2004;109:2655-2671.
6. Gold DR, Damokosh AI, Pope CA, Dockery DW, McDonnell WF, Serrano P, Retama A, Castillejos M. Particulate and ozone pollutant effects on the respiratory function of children in Southwest Mexico City. *Epidemiology*. 1999;10:470.

7. Sofer T, Baccarelli A, Cantone L, Coull B, Maity A, Lin X, Schwartz J. Exposure to airborne particulate matter is associated with methylation pattern in the asthma pathway. *Epigenomics*. 2013;5:147-154.

8. Li Z, Wen Q, Zhang R. Sources, health effects and control strategies of indoor fine particulate matter (PM2.5): A review. *Sci Total Environ*. 2017;586:610-622.

9. Wang Y, Li L, Chen C, Huang C, Huang H, Feng J, Wang S, Wang H, Zhang G, Zhou M, Cheng P, Wu M, Sheng G, Fu J, Hu Y, Russell AG, Wumaer A. Source apportionment of fine particulate matter during autumn haze episodes in Shanghai, China. *J Geophys Res*. 2014;119:1903-1914.

10. Morawska L, Ayoko GA, Bae GN, Buonanno G, Chao CYH, Clifford S, Fu SC, Hänninen O, He C, Isaxon C, Mazaheri M, Salthammer T, Waring MS, Wierzbicka A. Airborne particles in indoor environment of homes, schools, offices and aged care facilities: The main routes of exposure. *Environ Int*. 2017;108:75-83.

11. Meng QY, Spector D, Colome S, Turpin B. Determinants of indoor and personal exposure to PM2.5 of indoor and outdoor origin during the RIOPA study. *Atmos Environ*. 2009;43:5750-5758.

12. Habre R, Coull B, Moshier E, Godbold J, Grunin A, Nath A, Castro W, Schachter N, Rohr A, Kattan M, Spengler J, Koutrakis P. Sources of indoor air pollution in New York City residences of asthmatic children. *J Expo Sci Environ Epidemiol*. 2014;24:269-278.

13. Ji W, Li H, Zhao B, Deng F. Tracer element for indoor PM2.5 in China migrated from outdoor. *Atmos Environ*. 2018;176:171-178.

14. Jeong C-H, Salehi S, Wu J, North ML, Kim JS, Chow C-W, Evans GJ. Indoor measurements of air pollutants in residential houses in urban and suburban areas: Indoor versus ambient concentrations. *Sci Total Environ*. 2019;693:133446.

15. Xia T, Chen C. Differentiating between indoor exposure to PM2.5 of indoor and outdoor origin using time-resolved monitoring data. *Build Environ*. 2019;147:528-539.

16. Allen R, Larson T, Sheppard L, Wallace L, Liu LJS. Use of real-time light scattering data to estimate the contribution of infiltrated and indoor-generated particles to indoor air. *Environ Sci Technol*. 2003;37:3484-3492.

17. Wallace L, Williams R, Rea A, Croghan C. Continuous weeklong measurements of personal exposures and indoor concentrations of fine particles for 37 health-impaired North Carolina residents for up to four seasons. *Atmos Environ*. 2006;40:399-414.

18. Chan WR, Logue JM, Wu X, Klepeis NE, Fisk WJ, Noris F, Singer BC. Quantifying fine particle emission events from time-resolved measurements: Method description and application to 18 California low-income apartments. *Indoor Air*. 2018;28:89-101.

19. [dataset] Chan, Wanyu; Tang, Hao; Sohn, Michael (2020), Automating the interpretation of PM2.5 time-resolved measurements using a data-driven approach, Dryad, Dataset, https://doi.org/10.7941/D1HG9J.

20. Singer BC, Chan WR, Kim YS, Offermann FJ, Walker IS. Indoor air quality in California homes with code-required mechanical ventilation. *Indoor Air*. 2020:1-15.

21. Breiman L. Random Forests. *Mach Learn*. 2001;45:5-32.

22. Yuchi W, Gombojav E, Boldbaatar B, Galsuren J, Enkhmaa S, Beejin B, Naidan G, Ochir C, Legtseg B, Byambaa T, Barn P, Henderson SB, Janes CR, Lanphear BP, McCandless LC, Takaro TK, Venners SA, Webster GM, Allen RW. Evaluation of random forest

regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. *Environ Pollut*. 2019;245:746-753.

23. Xu C, Xu D, Liu Z, Li Y, Li N, Chartier R, Chang J, Wang Q, Wu Y, Li N. Estimating hourly average indoor PM2.5 using the random forest approach in two megacities, China. *Build Environ*. 2020;180.
24. Kropat G, Bochud F, Jaboyedoff M, Laedermann JP, Murith C, Palacios M, Baechler S. Improved predictive mapping of indoor radon concentrations using ensemble regression trees based on automatic clustering of geological units. *J Environ Radioact*. 2015;147:51-62.
25. Wei W, Ramalho O, Malingre L, Sivanantham S, Little JC, Mandin C. Machine learning and statistical models for predicting indoor air quality. *Indoor Air*. 2019;29:704-726.
26. Liaw A, Wiener M. Classification and Regression by randomForest. *R News*. 2002;2:18-22.
27. Archer KJ, Kimes R V. Empirical characterization of random forest variable importance measures. *Comput Stat Data Anal*. 2008;52:2249-2260.
28. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*. 2018;106:249-259.
29. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29:1189-1232.
30. Grömping U. Variable importance assessment in regression: Linear regression versus random forest. *Am Stat*. 2009;63:308-319.

Figure legends

Figure 1 An example of emission (red) and decay (green) periods identified by the rule-based method.

Figure 2 Example of the OOB error rate as a function of the number of trees.

Figure 3 Illustration of some of the input features considered in the RF model.

Figure 4 Overall error and class error for different percentages of data used for training and testing. For example, a 50%-50% split means 50% of the data was used for training and the other half was used for testing.

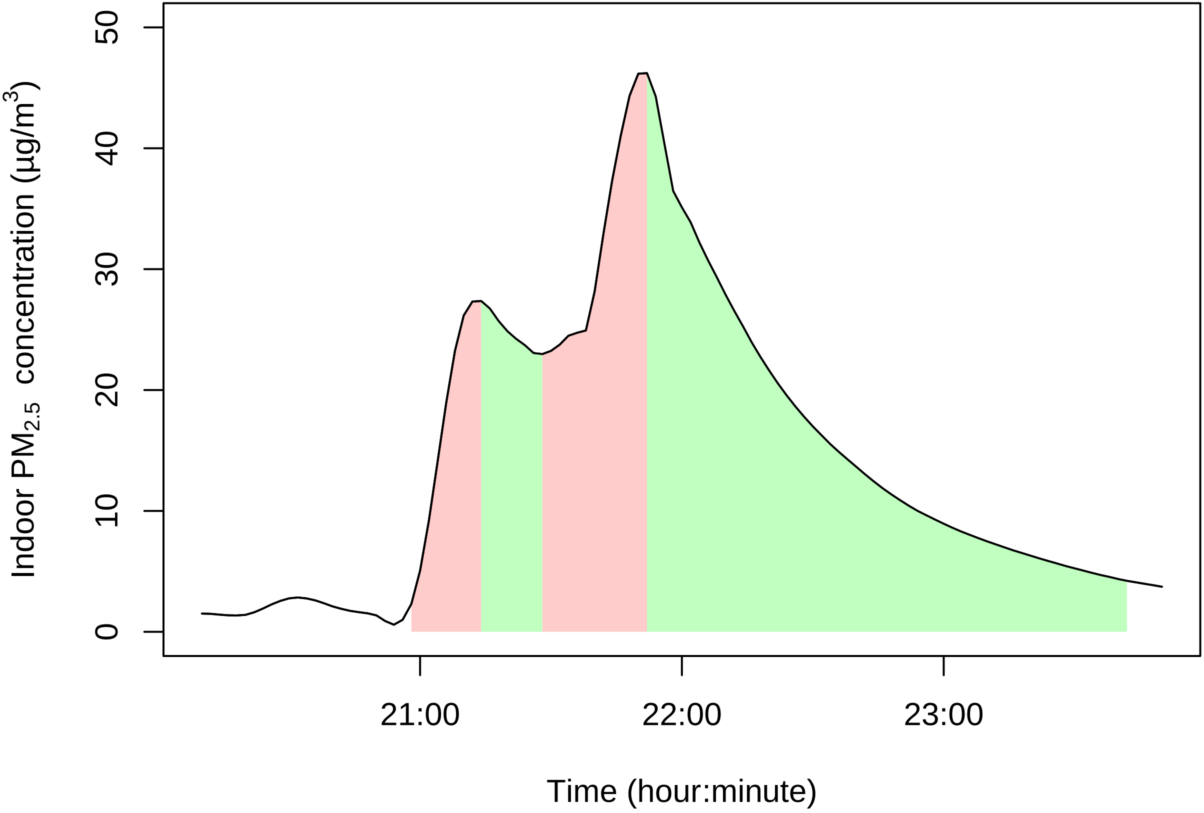
Figure 5 Overall and class error rate for the identification of emission and decay when predict one building with other two buildings. The error rate on training data set was estimated based on OOB data.

Figure 6 Importance of input features for identification of emission (a) and decay (b), sorted by mean decrease accuracy.

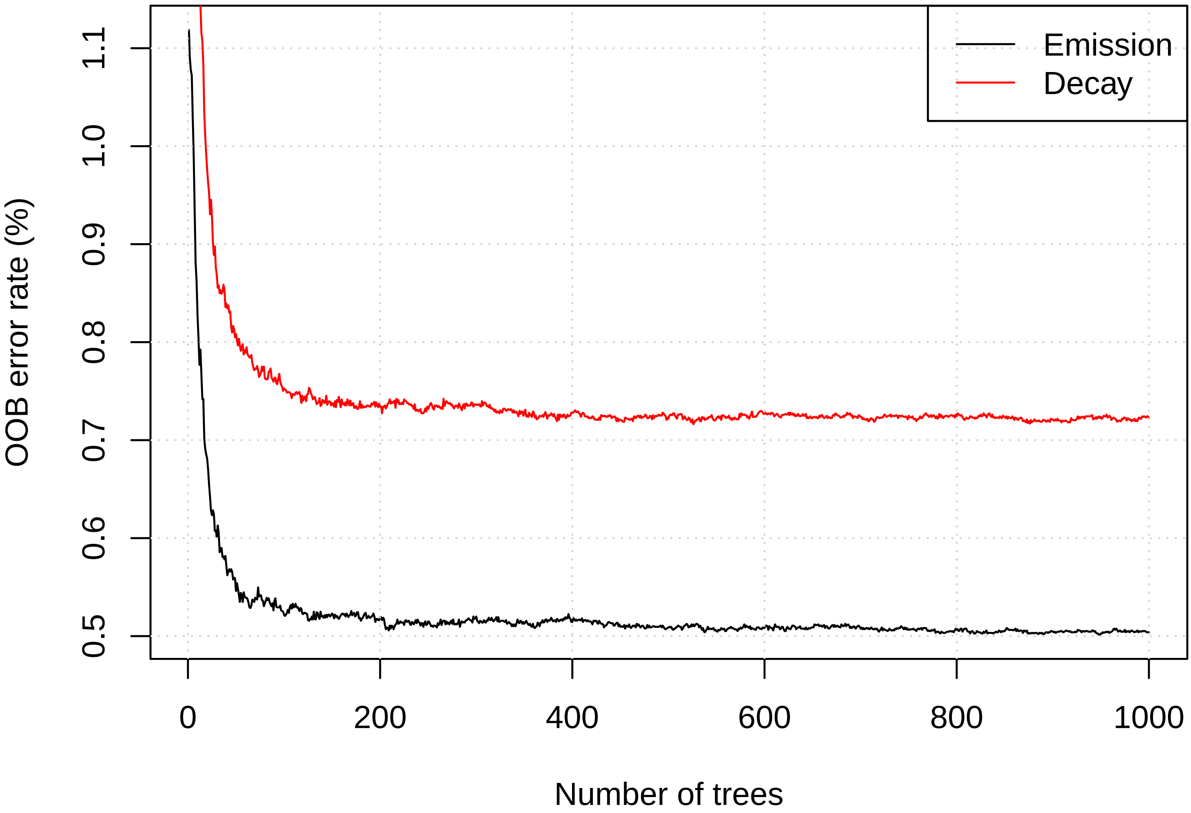
Figure 7 Selected patterns of identified emission events: (a) a typical emission event, (b) an emission event with inaccurate identified start and end time, (c) two linked emission events, (d) a case that RF model excluded events resulted by outdoor air.

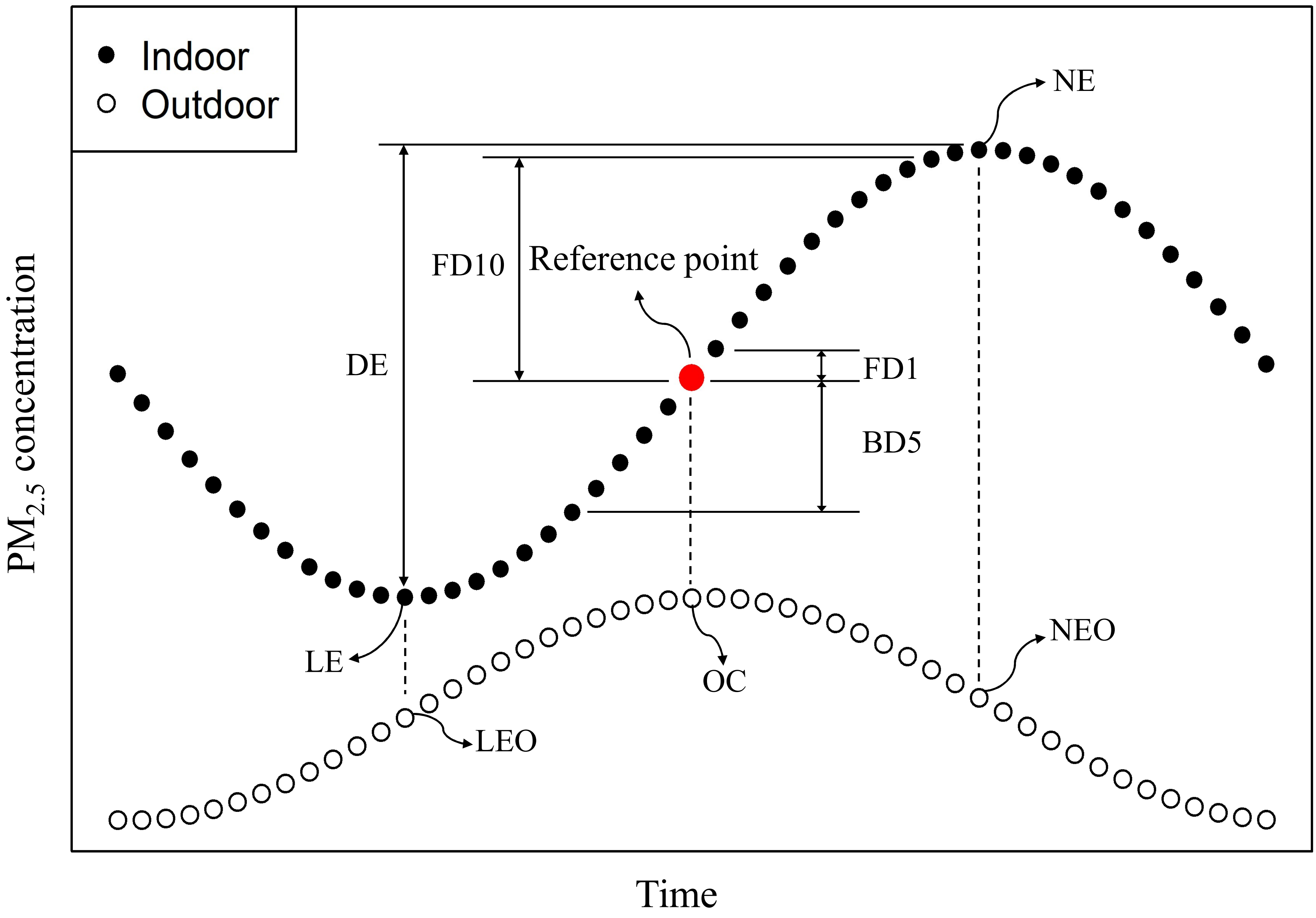
Figure 8 Misidentified emission events: (a) an emission event resulted by outdoor source and (b) an emission event with a moderate increase in indoor concentration lasting a very long period.

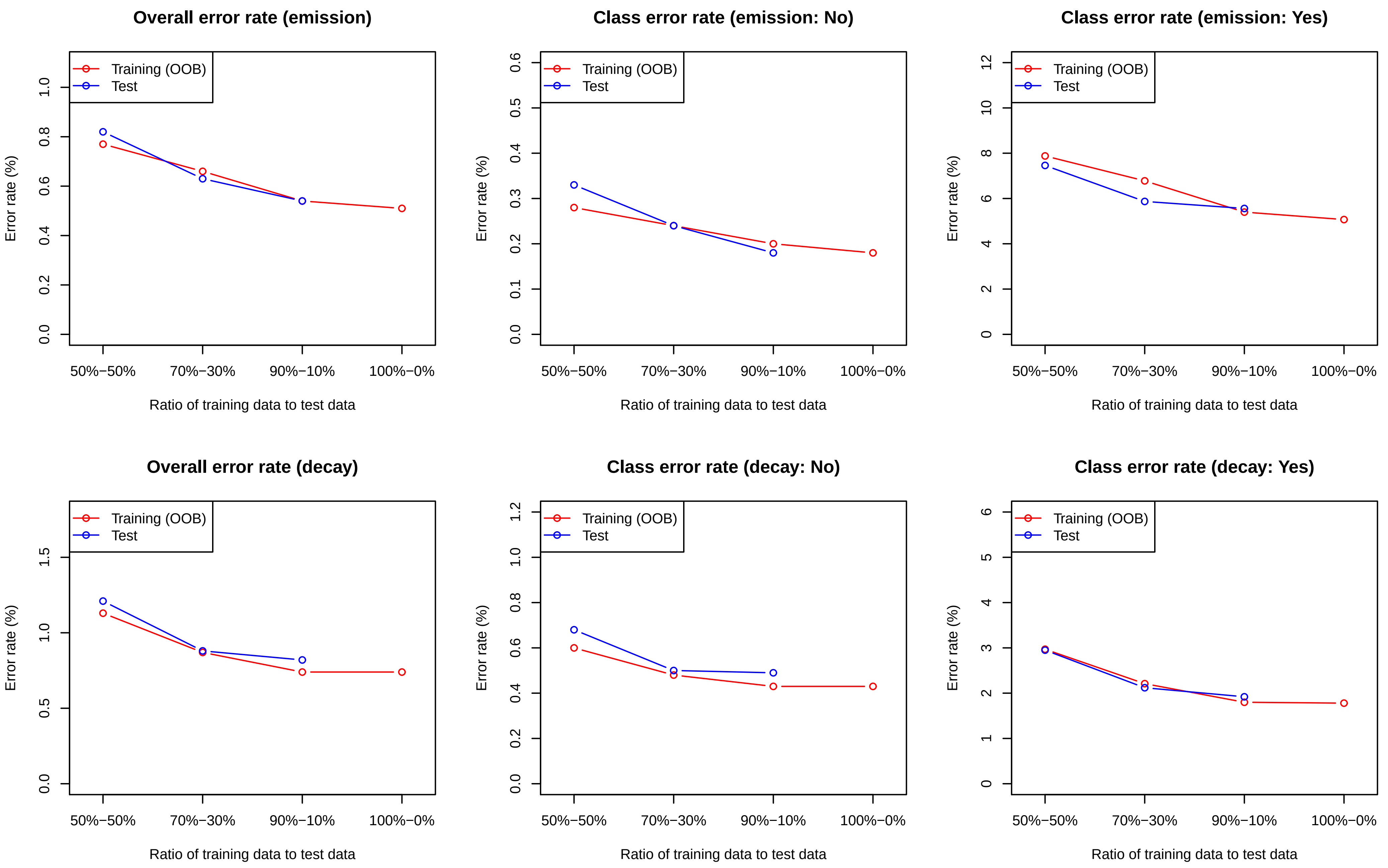
Figure 9 Estimated concentration during the emission and decay period based on first-order mass balance method



Indoor PM$_{2.5}$ concentration (μg/m^3)
0
10
20
30
40
50
21:00
22:00
23:00
Time (hour:minute)


OOB error rate (%)
0.5
0.6
0.7
0.8
0.9
1.0
1.1
0
200
400
600
800
1000
Number of trees
Emission
Decay


● Indoor
○ Outdoor
NE
FD10
Reference point
DE
FD1
BD5
LE
OC
NEO
LEO
$PM_{2.5}$ concentration
Time



Overall error rate (emission)

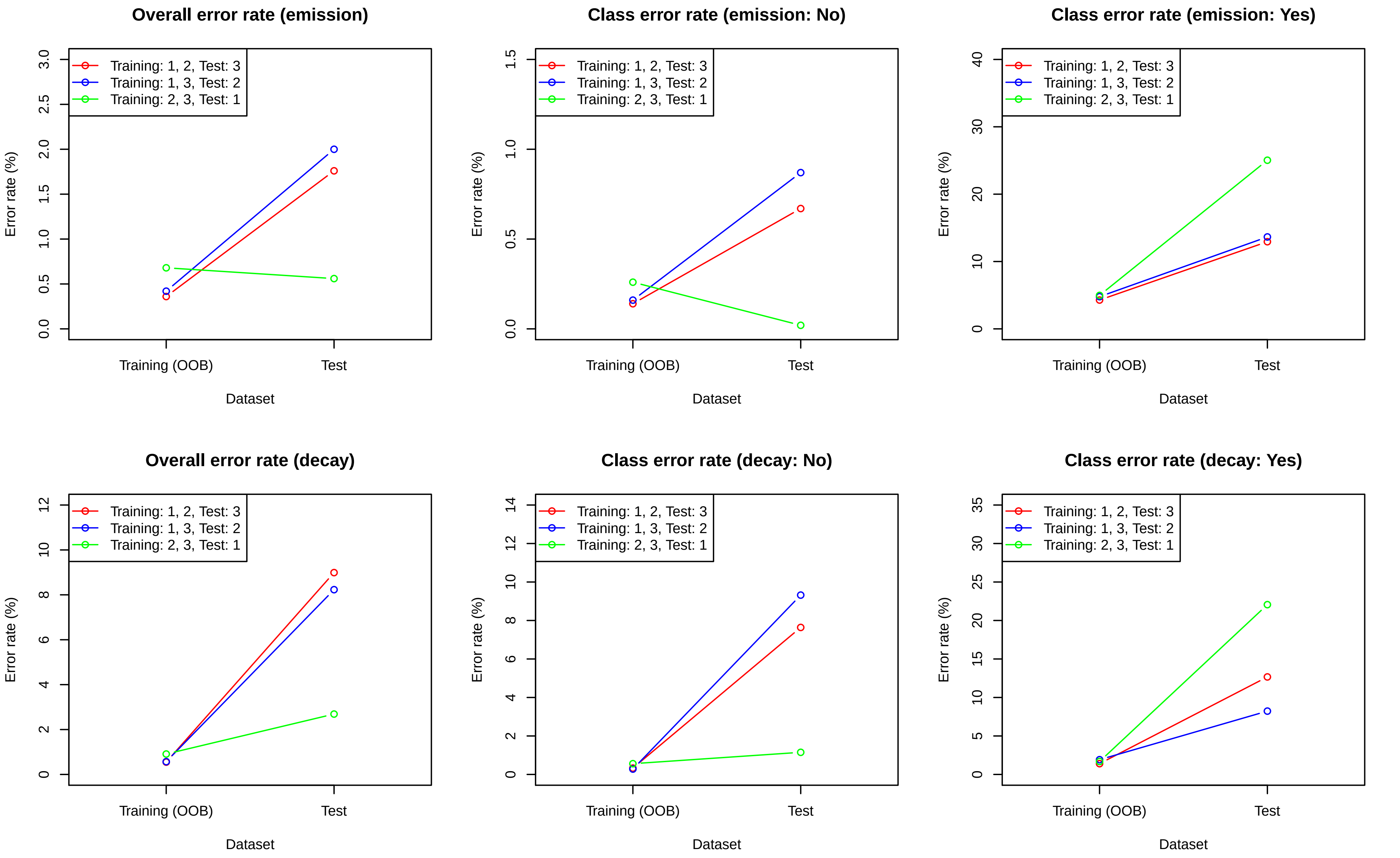
Class error rate (emission: No)

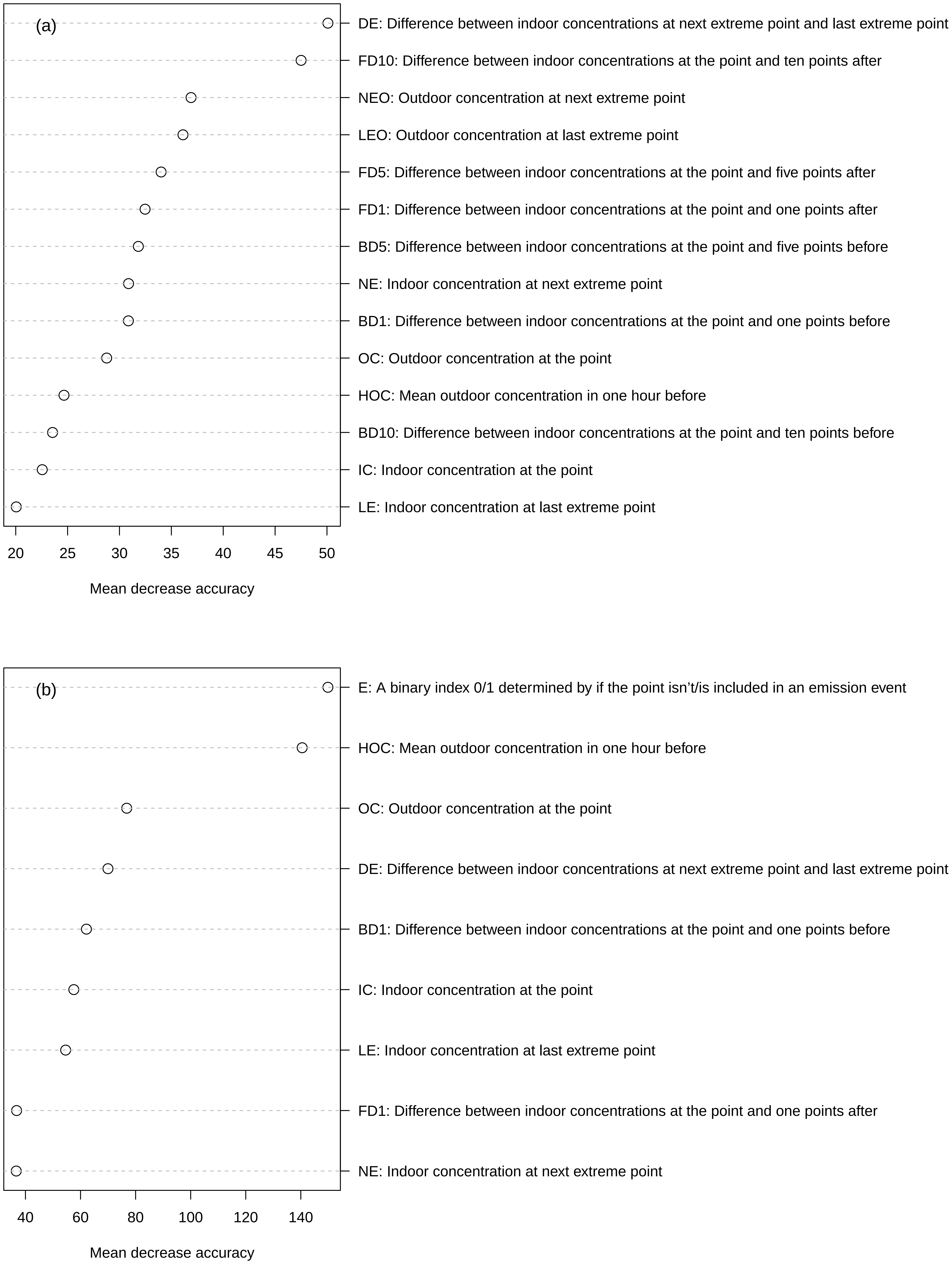
Class error rate (emission: Yes)

Overall error rate (decay)

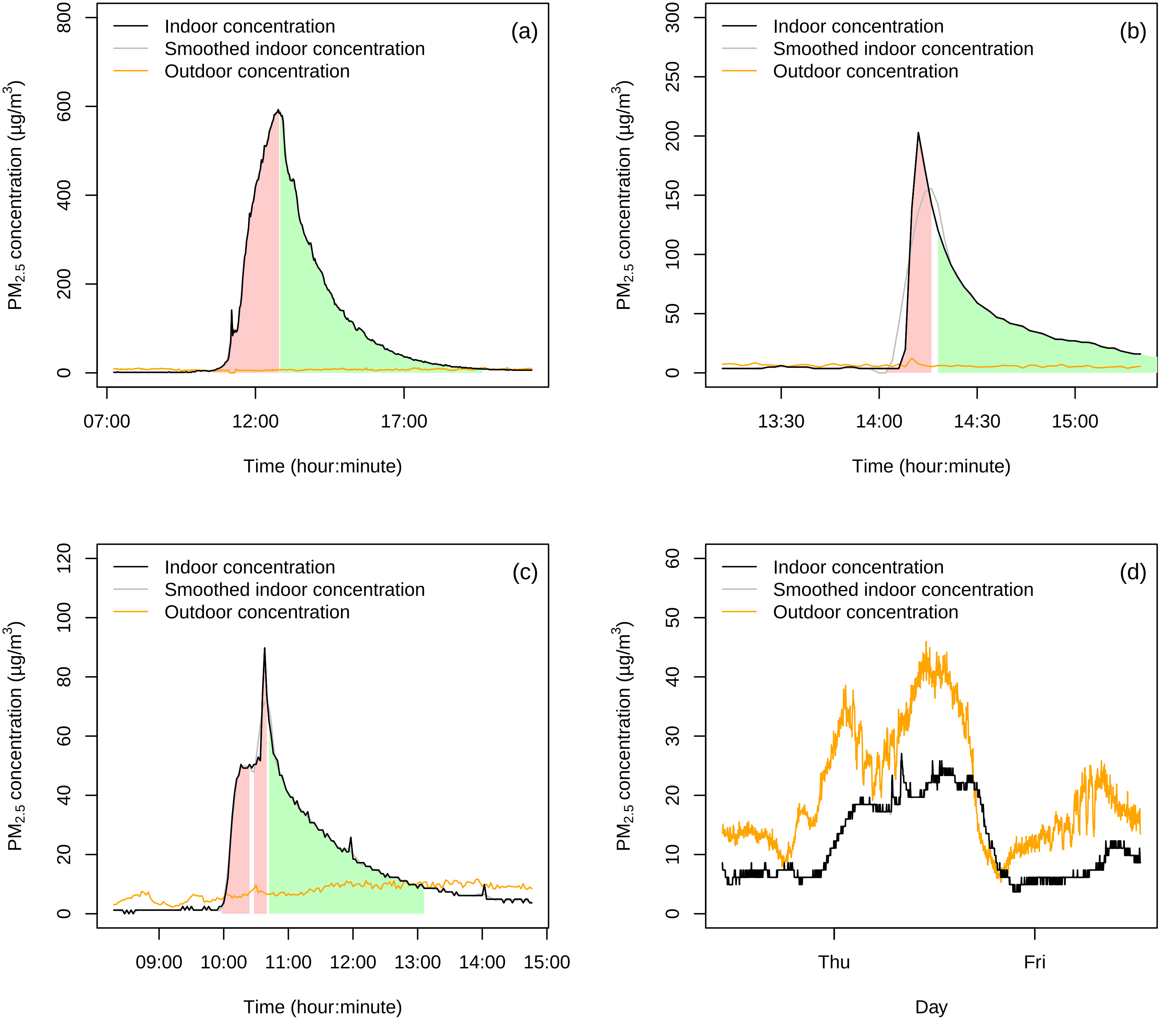
Class error rate (decay: No)

Class error rate (decay: Yes)

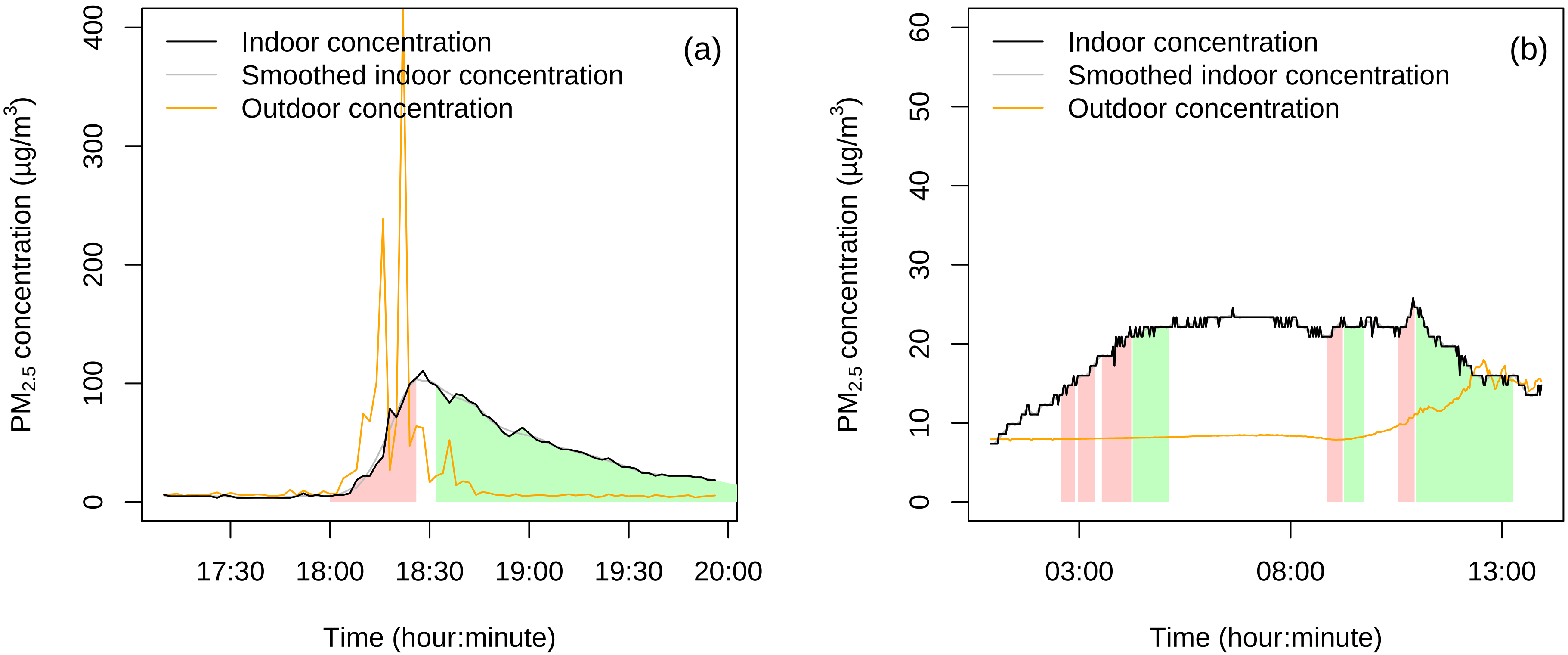

Overall error rate (emission)
Class error rate (emission: No)
Class error rate (emission: Yes)
Overall error rate (decay)
Class error rate (decay: No)
Class error rate (decay: Yes)
Training: 1, 2, Test: 3
Training: 1, 3, Test: 2
Training: 2, 3, Test: 1
Error rate (%)
Dataset
Training (OOB)
Test
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
10
20
30
40
2
4
6
8
10
12
14
5
15
25
35



(a)
DE: Difference between indoor concentrations at next extreme point and last extreme point
FD10: Difference between indoor concentrations at the point and ten points after
NEO: Outdoor concentration at next extreme point
LEO: Outdoor concentration at last extreme point
FD5: Difference between indoor concentrations at the point and five points after
FD1: Difference between indoor concentrations at the point and one points after
BD5: Difference between indoor concentrations at the point and five points before
NE: Indoor concentration at next extreme point
BD1: Difference between indoor concentrations at the point and one points before
OC: Outdoor concentration at the point
HOC: Mean outdoor concentration in one hour before
BD10: Difference between indoor concentrations at the point and ten points before
IC: Indoor concentration at the point
LE: Indoor concentration at last extreme point
20
25
30
35
40
45
50
Mean decrease accuracy
(b)
E: A binary index 0/1 determined by if the point isn’t/is included in an emission event
HOC: Mean outdoor concentration in one hour before
OC: Outdoor concentration at the point
DE: Difference between indoor concentrations at next extreme point and last extreme point
BD1: Difference between indoor concentrations at the point and one points before
IC: Indoor concentration at the point
LE: Indoor concentration at last extreme point
FD1: Difference between indoor concentrations at the point and one points after
NE: Indoor concentration at next extreme point
40
60
80
100
120
140
Mean decrease accuracy



(a)
Indoor concentration
Smoothed indoor concentration
Outdoor concentration
$PM_{2.5}$ concentration (µg/m^3)
0
200
400
600
800
07:00
12:00
17:00
Time (hour:minute)
(b)
Indoor concentration
Smoothed indoor concentration
Outdoor concentration
$PM_{2.5}$ concentration (µg/m^3)
0
50
100
150
200
250
300
13:30
14:00
14:30
15:00
Time (hour:minute)
(c)
Indoor concentration
Smoothed indoor concentration
Outdoor concentration
$PM_{2.5}$ concentration (µg/m^3)
0
20
40
60
80
100
120
09:00
10:00
11:00
12:00
13:00
14:00
15:00
Time (hour:minute)
(d)
Indoor concentration
Smoothed indoor concentration
Outdoor concentration
$PM_{2.5}$ concentration (µg/m^3)
0
10
20
30
40
50
60
Thu
Fri
Day


Indoor concentration
Smoothed indoor concentration
Outdoor concentration
(a)
(b)
PM$_{2.5}$ concentration (μg/m^3)
0
100
200
300
400
17:30
18:00
18:30
19:00
19:30
20:00
Time (hour:minute)
0
10
20
30
40
50
60
03:00
08:00
13:00
Time (hour:minute)

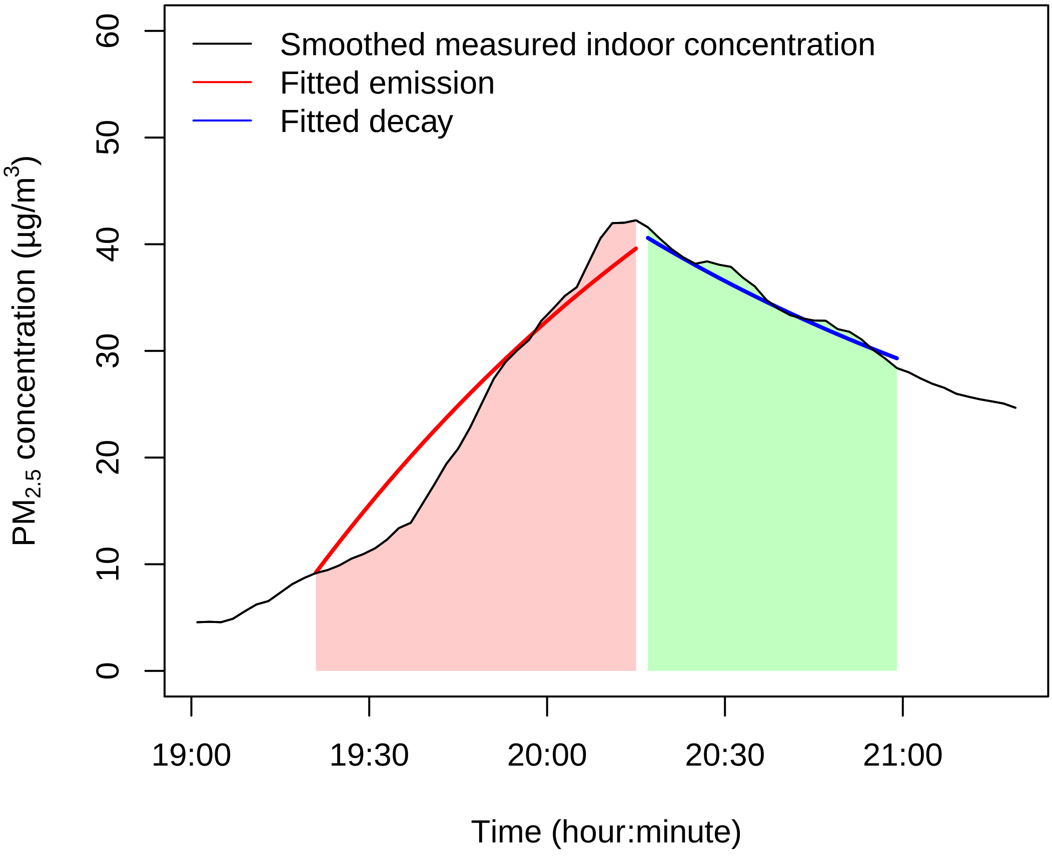

Smoothed measured indoor concentration
Fitted emission
Fitted decay
PM$_{2.5}$ concentration (μg/m^3)
0
10
20
30
40
50
60
19:00
19:30
20:00
20:30
21:00
Time (hour:minute)

Table 1 Home characteristic of the two datasets

Dataset	Numbers of residence	Stove type	Mean floor area [a] (m^2)	Home Type
Dataset one	18	Natural gas (12) and electric (6)	94.7 (19.75)	Low-income apartment
Dataset two	65	Natural gas	241.5 (79.7)	New (2011-2017) single-family home

a: Standard deviation in parentheses

Table 1 Features tested in RF model for identification of emission and decay

Abbreviation	Interpretation	Emission[*]	Decay[*]
IC	Indoor concentration at the point	√	√
OC	Outdoor concentration at the point	√	√
HOC	Mean outdoor concentration in one hour before	√	√
FD1/FD5/FD10	Difference between indoor concentrations at the point and one/five/ten points after	√	Only FD1
BD1/BD2/BD10	Difference between indoor concentrations at the point and one/five/ten points before	√	Only BD1
NE	Indoor concentration at next extreme point	√	√
LE	Indoor concentration at last extreme point	√	√
NEO	Outdoor concentration at next extreme point	√	
LEO	Outdoor concentration at last extreme point	√	
DE	Difference between indoor concentrations at next extreme points and last extreme points	√	√
E	A binary index determined by if the point is (0) or is not (1) included in an emission event		√
BD2/BD3/BD4	Difference between indoor concentrations at the point and two/three/four points before		
FD2/FD3/FD4	Difference between indoor concentrations at the point and two/three/four points after		
V4/V8/V12/V16	Variance of indoor concentration in a four/eight/twelve/sixteen minutes interval		
IO	Ratio of indoor concentration to outdoor concentration		

[*] Check marks indicate features used in the final model.

Table 3 RF model performance determined based on OOB data

Category	Group	Predicted: No	Predicted: Yes	Class error rate	Overall error rate
Emission	Actual: No	150593	296	0.20%	0.51%
	Actual: Yes	532	10116	5.00%	
Decay	Actual: No	124095	493	0.40%	0.71%
	Actual: Yes	651	36298	1.76%	

Table 4 Summary statistics of identified PM emission events for datasets 1 and 2

	Mean		Median		Geometric Mean		Geometric Std. Deviation		5 to 95 Percentile Range	
Dataset (D)	D1	D2	D1	D2	D1	D2	D1	D2	D1	D2
Emitted mass (mg)	30	14	12	5.4	13	6.3	3.7	3.3	1.4-154	1.1-53
Event duration (min)	23	21	16	16	19	18	1.7	1.6	8-66	10-44
Emission rate (mg/h)	103	42	37	18	40	20	3.9	3.3	3-582	3.2-146
Loss rate (1/h)	2.0	1.3	1.3	1.0	1.3	1.0	2.5	2.0	0.2-7.5	0.4-3.0

SUPPORTING INFORMATION

Automating the interpretation of PM2.5 time-resolved measurements using a data-driven approach

Hao Tang [a], Wanyu Rengie Chan [b], Michael Sohn [b*]

[a] Joint International Research Laboratory of Green Buildings and Built Environments, Ministry of Education, Chongqing University, Chongqing 400045, China

[b] Indoor Environment Group, Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley CA 94720, USA

S1. Identified emission events for dataset 2

Figure S1 presents the cumulative percentage of the number of indoor emission events identified in each home during a weeklong period. 94% of homes were found to have 20 or less identified indoor emission events during a weeklong period. Home 123 was found to have the most number of identified indoor emission events, 36. No $PM_{2.5}$ emission event was identified in 11 of the homes. Time-resolved indoor and outdoor concentrations in homes 123 and 13, which had no identified indoor emission events are presented in **Figure S2**.

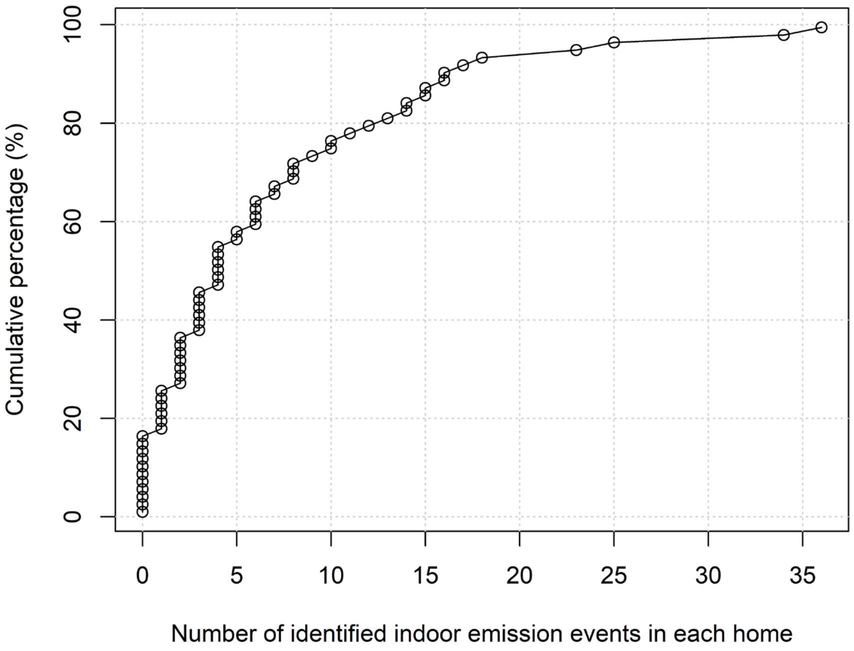


Figure S1 Cumulative percentage of number of identified emission events per home during each weeklong period

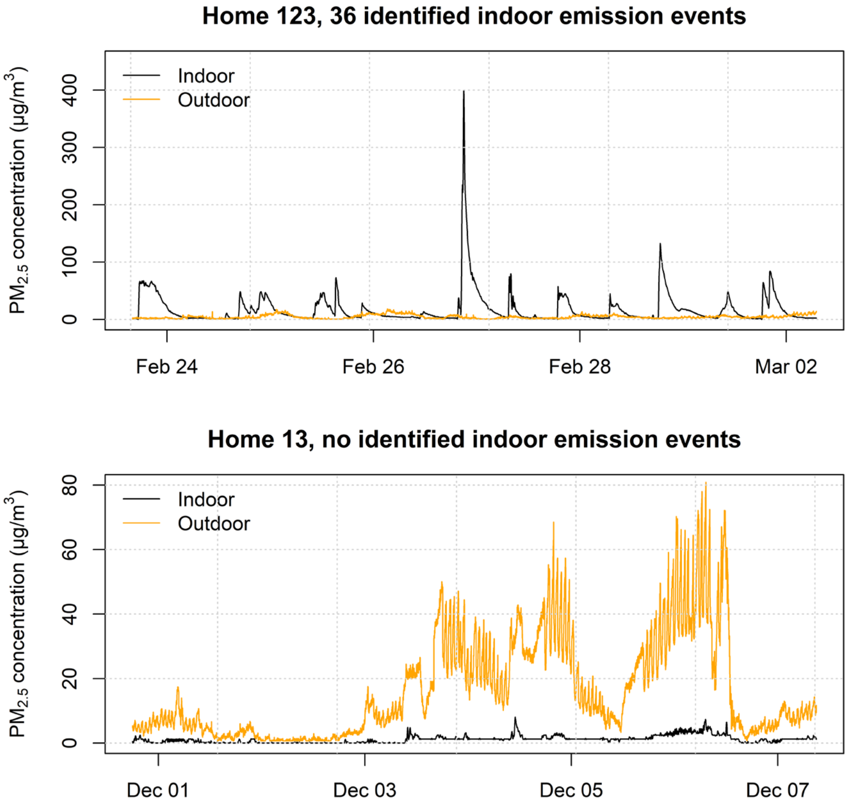


Figure S2 Time-resolved indoor and outdoor $PM_{2.5}$ concentrations in home 116 and home 13

S2. Estimating emission rates for dataset 2

We estimated the emission rate from an indoor source using a method based on first-order mass balance. We refer readers to [1] for details of this method. There are other methods to estimate emission rate (e.g., [2]). We used this method as it was the same method used in the rule-based paper.

A linear regression was first conducted to estimate composite loss rate L based on the identified decay period, as shown in Eq. S1. Where L is the pseudo-first-order loss rate and represents an overall effect of ventilation, filtration and deposition, t is time, C_{in} is the indoor concentration, $C_{in}(t_d)$ is the indoor concentration at the beginning of decay, C_{in_O} is the indoor concentration originated from outdoor, taken as the minimum of indoor concentration and mean outdoor concentration during the emission and decay period.

$$\ln\left(\frac{C_{in}(t) - C_{in\_O}}{C_{in}(t_d) - C_{in\_O}}\right) = -L(t - t_d) \tag{S1}$$

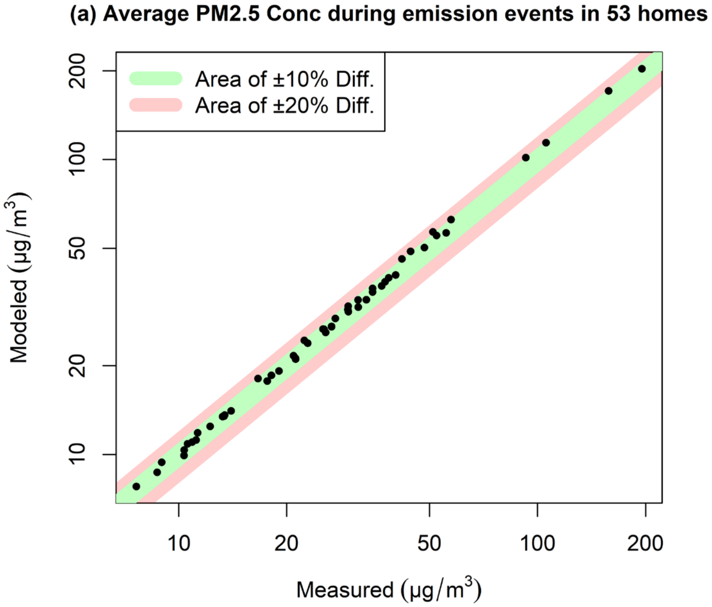
After the composite loss rate was determined, emission rate E was estimated using another linear regression, as shown in Eq. S2, where the $C_{in}(t_0)$ is the indoor concentration at the beginning of identified emission, V is the mixing volume.

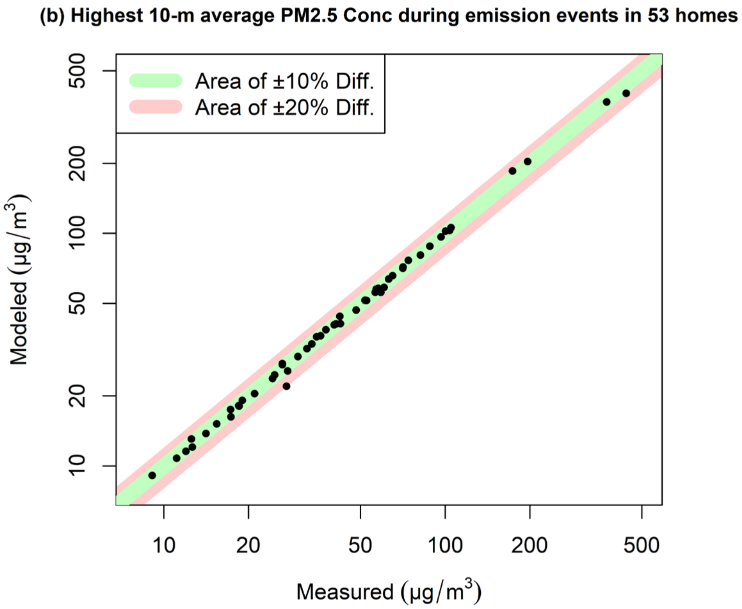
$$\left(C_{in}(t) - C_{in_O}\right) - (C_{in}(t_0) - C_{in_O})\exp(-\mathrm{L}(\mathrm{t} - t_0) = E\left[\frac{1}{LV}\left(1 - \exp\left(-L(t - t_0)\right)\right)\right] \tag{S2}$$

S3. Comparisons between modeled and measured results for dataset 2

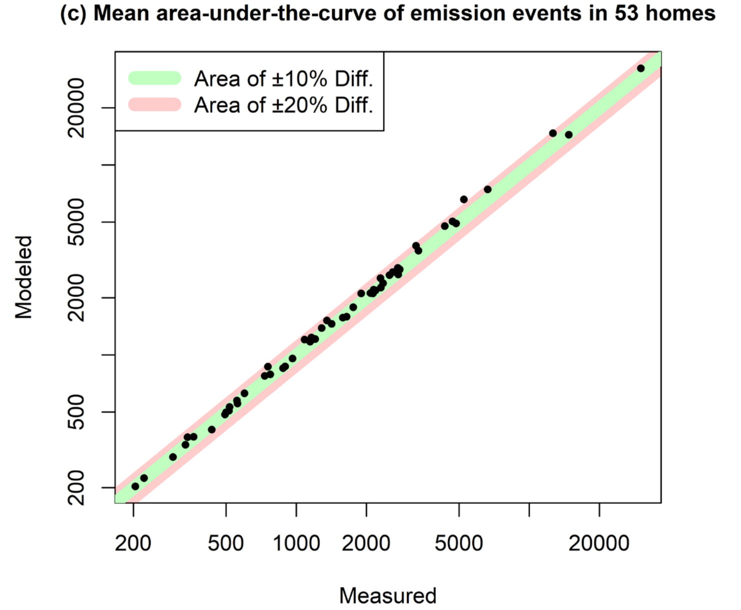
In **Figure S3**, we presented the modeled and measured result for each home using: (a) average $PM_{2.5}$ concentration during emission events, (b) peak concentration (highest 10-minute average $PM_{2.5}$ concentration) during emission events and (c) mean area-under-the-curve (AUC) of emission events. Results of 53 homes were plotted in figures, excluding 11 homes which had no identified emission events and 1 home which only had one identified emission but no eligible decay. Areas of ±10% and ±20% difference between modeled and measured results are shaded in green and red, respectively.

Figure S3 shows a good consistency between measured and modeled result. Average % of absolute difference between measured and modeled result are 4% (SD = 3%) for average concentration, 3% (SD = 3%) for peak concentration and 5% (SD = 5%) for AUC. 96% of average concentration, 98% of peak concentration and 81% of AUC are within the area of ±10% difference. Except 2% of AUC, all data points of average concentration and peak concentration is within the area of ±20% difference.





(b) Highest 10-m average PM2.5 Conc during emission events in 53 homes



(c) Mean area-under-the-curve of emission events in 53 homes

Reference

1. Chan WR, Logue JM, Wu X, Klepeis NE, Fisk WJ, Noris F, Singer BC. Quantifying fine particle emission events from time-resolved measurements: Method description and application to 18 California low-income apartments. *Indoor Air*. 2018;28:89-101.
2. O'Leary C, de Kluizenaar Y, Jacobs P, Borsboom W, Hall I, Jones B. Investigating measurements of fine particle (PM 2.5) emissions from the cooking of meals and mitigating exposure using a cooker hood. *Indoor Air*. 2019;29:423-438.