Estimating indoor pollutant loss using mass balances and unsupervised clustering to recognize decays

--- Supporting Information (SI) ---

Bowen Du[1,2], Jeffrey A. Siegel[1,3]*

[1] Department of Civil and Mineral Engineering, University of Toronto, Toronto, Canada, M5S 1A4
[2] School of Architecture, Civil and Environmental Engineering, École Polytechnique Fedérale de Lausanne, 1015 Lausanne, Switzerland
[3] Dalla Lana School of Public Health, University of Toronto, Toronto, Canada, M5T 1R4

This supporting information contains 11 pages, 18 figures, and one table.

Figure S1 is an illustrative example of elevation detection, decay extraction, and grouping of individual decay episodes.

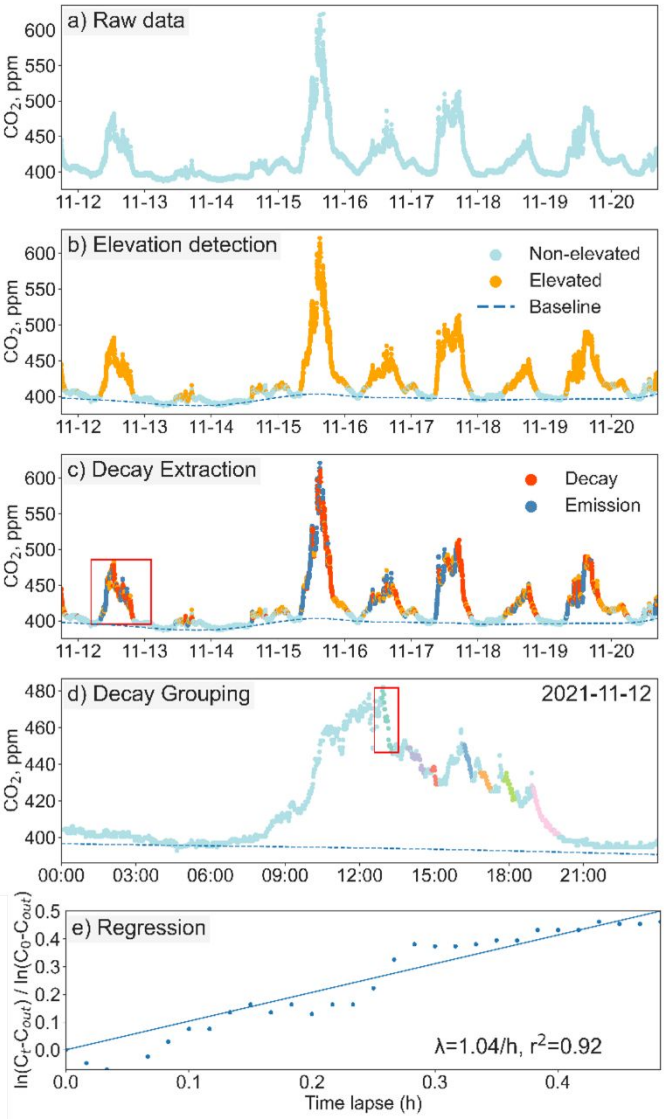

Figure S1. An illustrative example of the key steps of the model based on CO_2 data collected from an office for nine days. Data in Figure S1.d) are extracted from the red box in S1.c), and data in Figure S1.e) are further extracted from the red box in S1.d).

Figures S2-S4 show the clustering results of elevation detection (k-means), emission/decay separation (k-means), and decay segmentation (DBSCAN) with regard to the corresponding data features based on the office CO_2 data. As is shown, elevated and non-elevated periods are well separated based on the concentration difference from baseline and the absolute value of concentration gradient. Emissions and decays can be distinguished according to the positive or negative sign of concentration gradient and the relative high-low position. Plateau periods (i.e., concentration higher than baseline and relatively stable) are separated from emissions and decays too (Figure S1.c). Further, individual decay episodes can be segmented based on the increase in the number of previous non-decay samples.

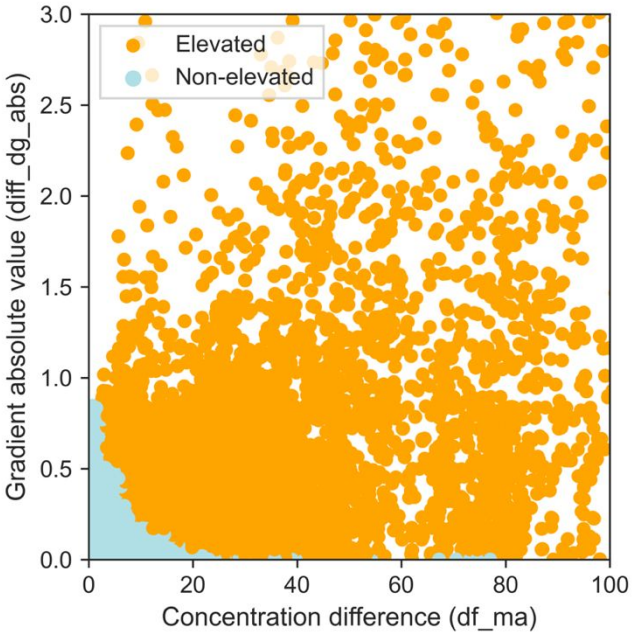

Figure S2. Elevation detection clustering result based on office CO_2 data.

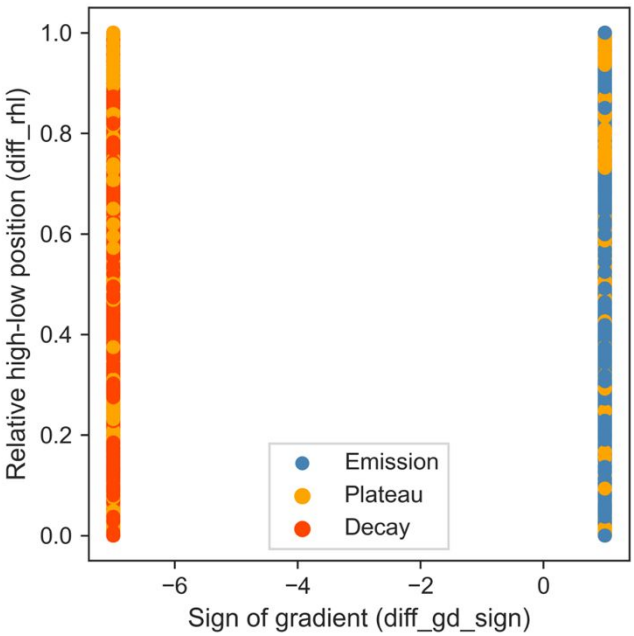

Figure S3. Emission/decay separation clustering result based on office CO_2 data.

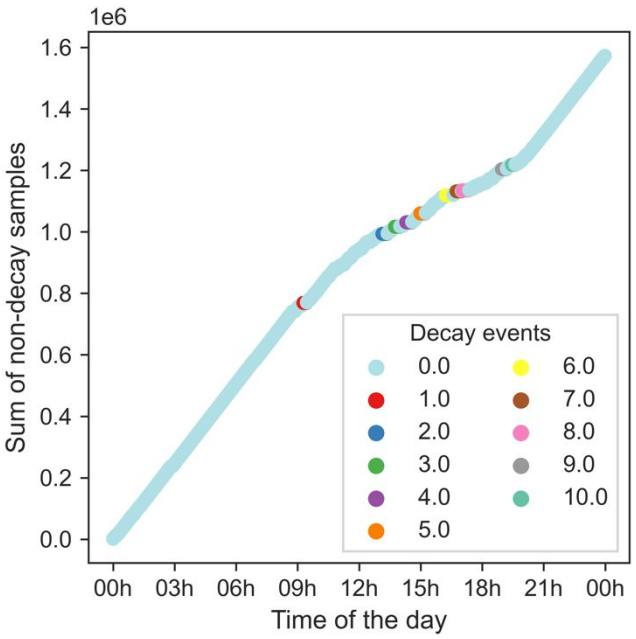

Figure S4. Decay events segmentation clustering result based on 1-day office CO_2 data.

Figures S5 and S6 demonstrate the hyperparameter selection approach, while Figures S7-S9 demonstrate the result filtration process. Figure S5 shows an example of the grid search process for selecting the optimal moving average window length and relative high-low position window length using CO_2 data from an office (continuous CO_2 concentration time series at 1-min intervals for 8 consecutive days). In this example, the relative high-low window length has a small impact when it is greater than 3 min (also 3 data points), while a moving average window length of 5 min (5 data points) leads to the highest Calinski-Harabasz score and the lowest Davies-Bouldin score, indicating better defined clusters. Therefore, this combination of hyperparameters was used in the following analysis for this dataset.

As is shown in Figure S6, a high minimum sample requirement (e.g., 30 data points) results in significantly higher regression R^2 values at the cost of the number of decays. Its impact on the results is negligible within the range of 2 – 10 samples, so we used the default value of 5. The maximum distance governs whether temporally disconnected data points will be considered as in the sample group. The smaller this parameter is, the higher the R^2 values in regression analysis are. We selected a conservative value of 0.001 for this parameter to avoid joining disconnected decay periods into a longer one, although consequently fewer decays remained available for characterizing temporal variations.

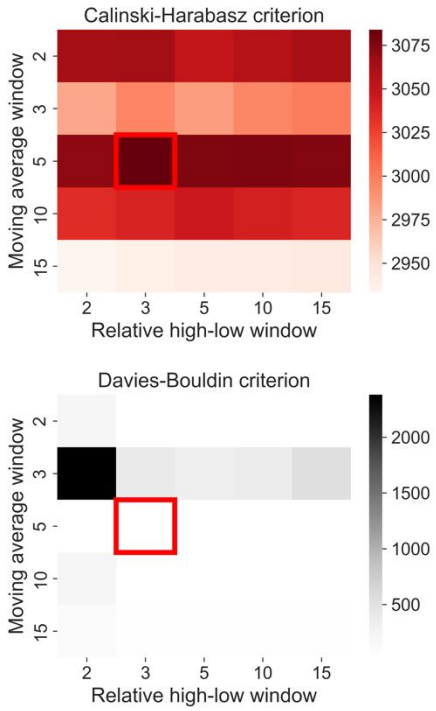

Figure S5. k-means clustering performance with different hyperparameters based on CO_2 data in the office.

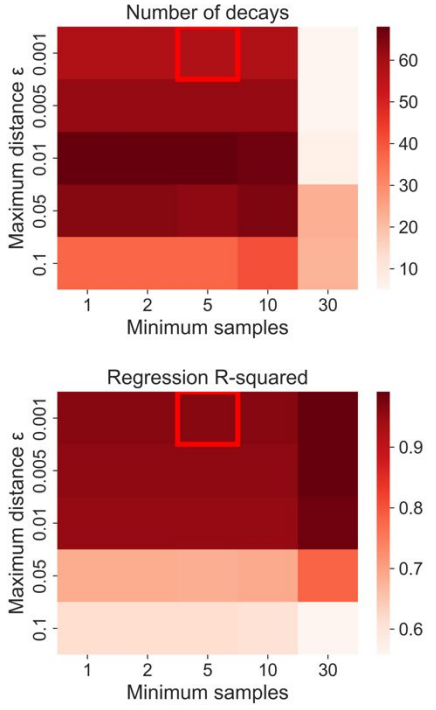

Figure S6. DBSCAN performance with different hyperparameters based on CO_2 data in the office.

Based on CO_2 and $PM_{2.5}$ concentration data collected from an office, Figure S7 shows the trade-off between the duration threshold (decays shorter than the corresponding threshold were excluded) and the number of decays left as well as the consequent impact on the median of the estimated decay rate. Without post-regression result filtration, 283 and 267 decays were

recognized (approximately 6.3 and 5.6 per sensor per day) for CO_2 and $PM_{2.5}$, respectively. Applying a duration threshold increases the average R^2 value, especially for $PM_{2.5}$ where the raw data are noisier. However, in this case, this increase is moderate and the median loss rate remains relatively consistent, while the number of remaining decays significantly drops. Therefore, we applied a duration threshold of 5 min to remove those extremely short decays. Further, Figure S8 shows the impact of setting a threshold for the regression r-squared value on the number of remaining decays and the median decay rate. In this case, most of the remaining decays after duration thresholding have a relatively high r-squared value already, and an r-squared threshold of 0.7 was selected to keep most of the decays while excluding those with substantial noise. For decays consisting of concentration data that are close to the estimated baseline, the uncertainties resulting from the baseline estimation process can be magnified. Therefore, setting a threshold for the median concentration difference from the baseline may remove those decays to improve the accuracy of loss rate estimation. Figure S9 shows the estimated decay rate and the corresponding concentration difference from baseline after the duration and r-squared thresholding. It is found that those decays with a smaller median concentration difference from the corresponding baselines indeed show higher variations. We decided to use a concentration difference threshold of 25 ppm for CO_2 and 5 $\mu g/m^3$ for $PM_{2.5}$ in this example.

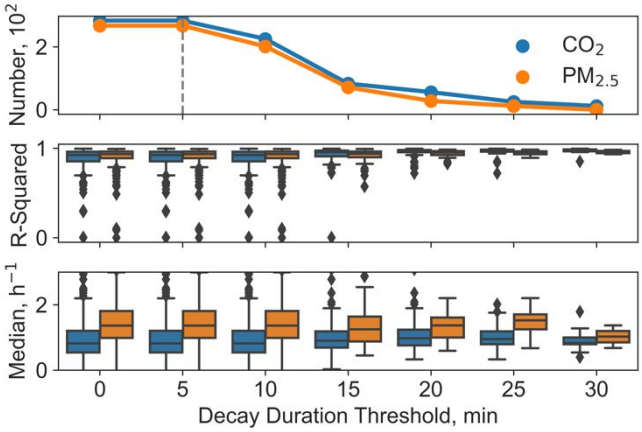

Figure S7. The number of remaining decays, average r-squared value, and median decay rate vs. duration thresholds in the office.

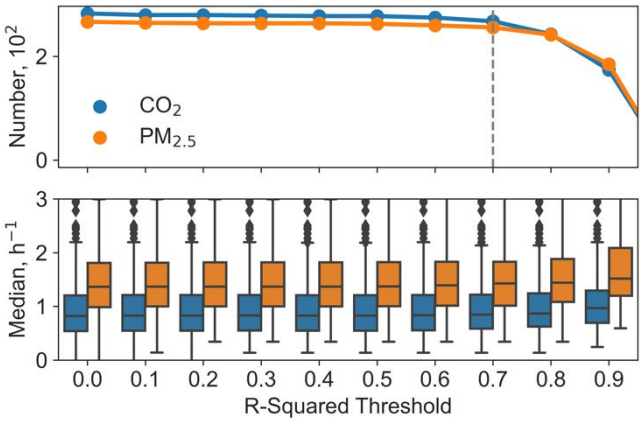

Figure S8. The number of remaining decays and median decay rate vs. r-squared thresholds in the office.

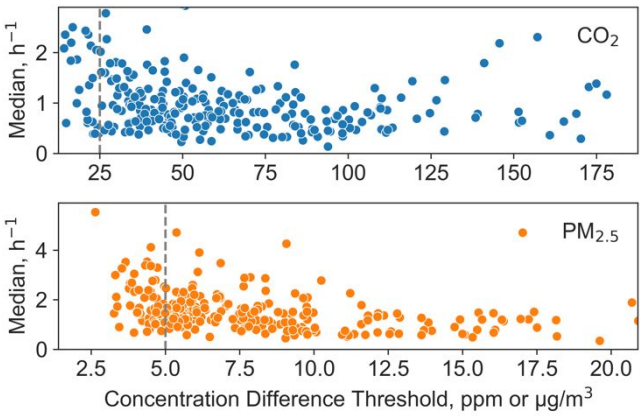

Figure S9. The estimated decay rate vs. the concentration difference from the baseline threshold in the office.

Figures S10, S12, and S15 provide examples of daily CO2 and PM2.5 concentrations with recognized decays in the office, classroom, and house, respectively. Figure S11 is a schematic of the laboratory chamber's HVAC system.

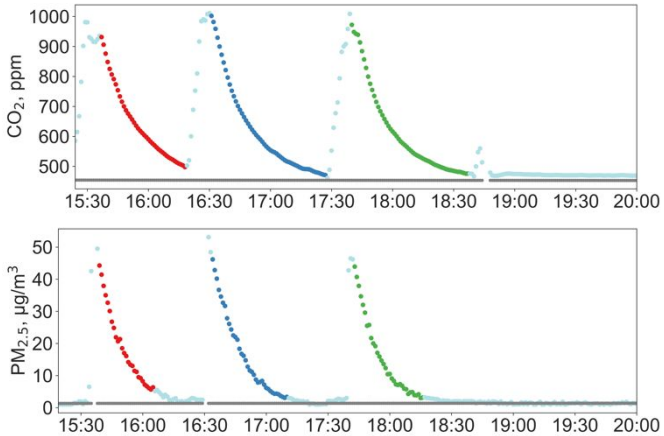

Figure S10. An example of daily CO_2 and $PM_{2.5}$ concentrations with recognized decays in the environmental chamber (center location).

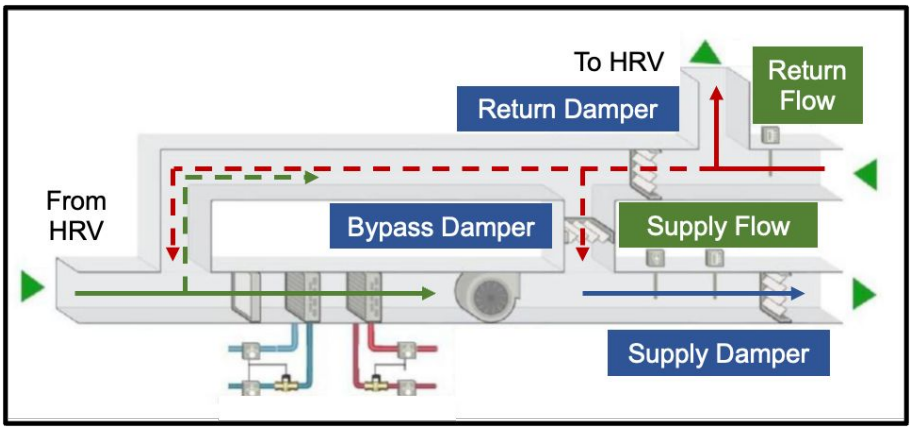

Figure S11. A schematic of the laboratory chamber's HVAC system

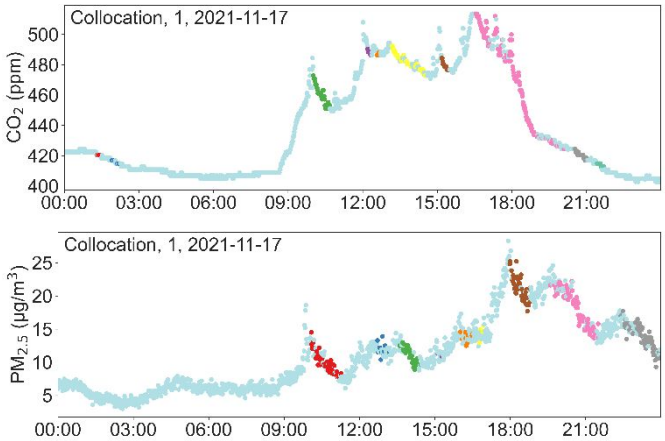

Figure S12. An example of daily CO_2 and $PM_{2.5}$ concentrations with recognized decays in the office.

Figures S13 and S14 compare baseline estimation methods and sensor consistency results based on the office CO2 data, as an addition to the PM2.5 data shown in the main document.

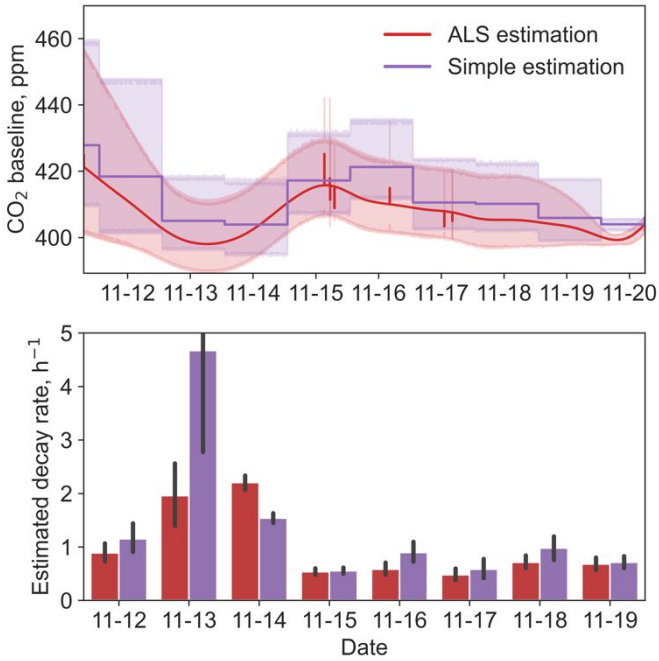

Figure S13. top) Different CO_2 baseline estimates (the band shows the standard deviation across five different sensors) and bottom) the corresponding estimated decay rate in the office.

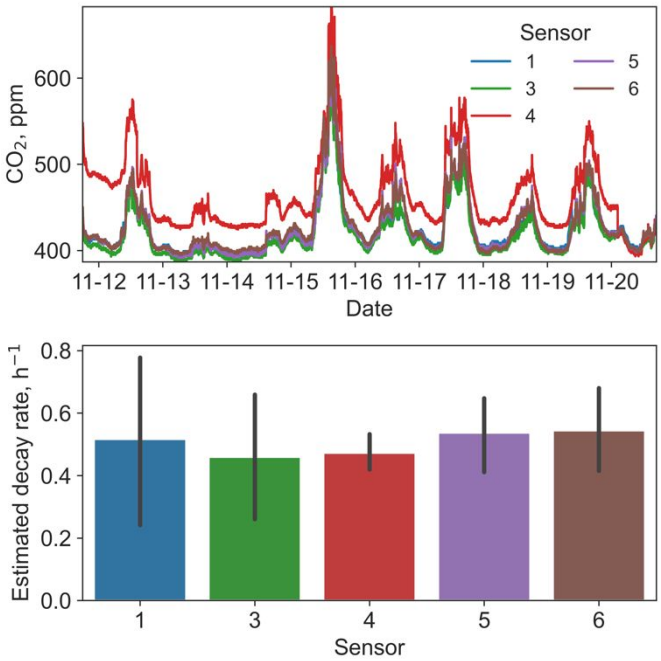

Figure S14. top) CO_2 concentration data from five collocated low-cost CO_2 sensors and bottom) the corresponding decay rate estimates in the office.

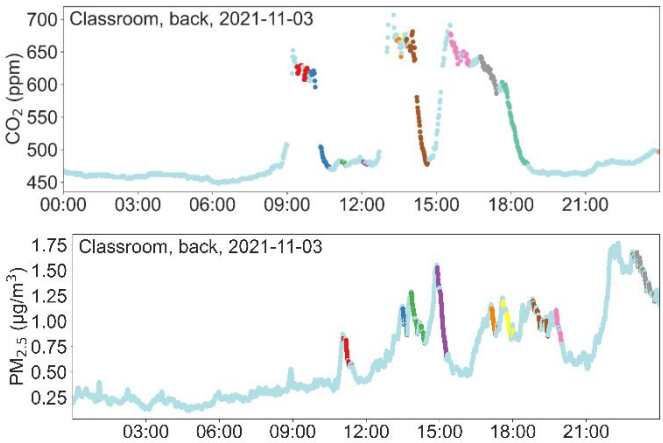

Figure S15. An example of daily CO_2 and $PM_{2.5}$ concentrations with recognized decays in the classroom (back location).

Figure S16 is a photo of the classroom showing the location of air supply and return.

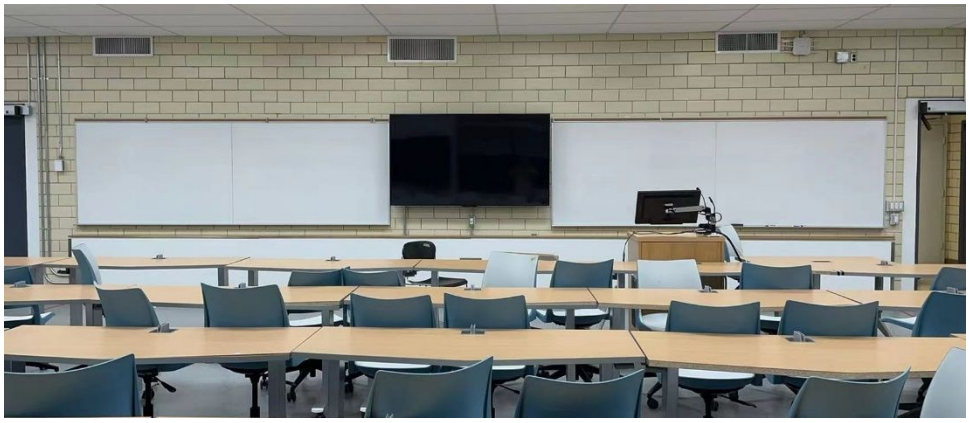

Figure S16. A photo of the classroom

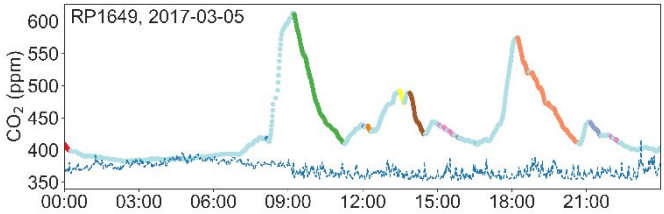

Figure S17. An example of daily CO_2 concentration with recognized decays in the home (with the outdoor concentration in navy dashed line).

Figure S18 shows the seasonal variation of the estimated air change rate based on the CO2 decay rate in the home throughout the year.

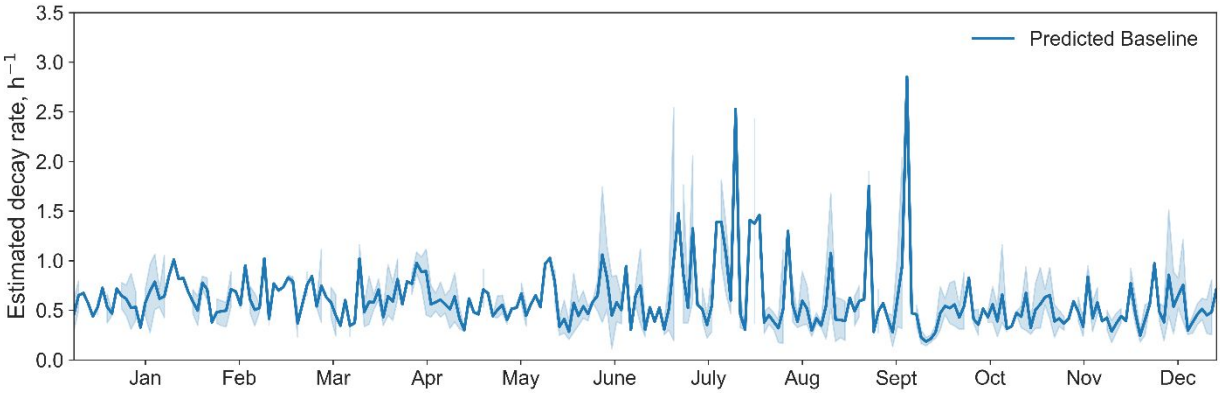

Figure S18. Seasonal variation of the estimated air change rate based on CO_2 decay rate in the home.

Table S1 shows the results of a sensitivity analysis on all hyperparameters used in this work. The hyperparameters in the base scenario are $\lambda = 10^9$, p = 0.001, ma = 5, rhl = 5, eps = 0.01, ms = 5. Before result filtration, baseline estimation and DBSACN parameters have a large impact on the mean CO_2 loss rate. However, after result filtration, only the maximum distance (eps) in DBSCAN has an impact of greater than 10% on the mean loss rate. This parameter determines whether adjacent decay data points are considered single or separate decay events,

and it is easy to recognize an unreasonable eps value by visually examining the decay recognition results.

Table S1. The impact of hyperparameter selection on the estimated CO_2 loss rate in the office

	Before filtration				**After filtration**			
	mean	% change	std	% change	mean	% change	std	% change
base	0.86		0.51		0.88		0.41	
Baseline estimation								
$\lambda = 10^8$	0.95	10%	0.72	41%	0.89	1%	0.41	1%
$\lambda = 10^{10}$	0.81	-6%	0.50	-1%	0.79	-10%	0.39	-3%
p = 0.0001	0.75	-13%	0.40	-22%	0.79	-10%	0.40	-1%
p = 0.01	1.12	30%	1.02	101%	0.93	5%	0.44	7%
Feature extraction								
ma = 3	0.86	0%	0.51	0%	0.88	0%	0.41	0%
ma = 10	0.86	0%	0.69	37%	0.82	-7%	0.34	-16%
rhl = 3	0.86	0%	0.51	0%	0.88	0%	0.41	0%
rhl = 10	0.86	0%	0.51	0%	0.88	0%	0.41	0%
DBSCAN								
eps = 0.001	1.12	29%	0.53	4%	1.03	17%	0.45	11%
eps = 0.1	0.15	-82%	0.49	-4%	0.29	-67%	0.15	-63%
ms = 3	1.20	39%	1.00	96%	0.97	10%	0.47	15%
ms = 10	0.95	10%	0.47	-8%	0.94	6%	0.33	-20%