Overlapping indoor air quality windows in five low income households

A reproducible exploratory analysis of PM2.5 excursions and environmental profiles

Ramyar Tajik, Arvin Ebrahimkhanlou and Simi Hoque
Drexel University, Philadelphia, USA

Research manuscript | Version 1 | 14 October 2026

Abstract

This study develops an auditable description of indoor air quality records from five Philadelphia households associated with a community weatherization study. Original AirVisual exports were reconciled, screened and aggregated to ten-minute medians. A sustained positive PM2.5 slope detector identified six-hour contextual windows, with intentional overlap and explicit coverage requirements. Twelve physical features were standardized, reduced by principal component analysis and grouped using k-means. The archived analysis reproduced exactly: 1,404 windows, six components explaining 95.79% of standardized feature variance, and three groups containing 1,029, 261 and 114 windows. H01 supplied 79.49% of the library. Of 1,399 adjacent within-household pairs, 811 overlapped; the windows covered 35,935 unique ten-minute bins. Group 2 was 91.57% H01 and group 3 was 69.30% H05. With three groups held fixed, agreement with the primary partition remained high for a disjoint-window subset (adjusted Rand index 0.990) but fell under equal household weighting (0.344), robust scaling (0.342) and logarithmic particle/carbon-dioxide features (0.295). A seven-day block resampling analysis gave median agreement of 0.930 for the original weighting. Five archived rising signal references in four selected day-long blocks matched five retained triggers. The contribution is a reproducible exploratory characterization of observed excursions, with household composition, overlap and preprocessing sensitivity made explicit. The profiles provide a basis for designing subsequent household-level validation.

Keywords: indoor air quality; low-income housing; overlapping windows; PM2.5; principal component analysis; clustering; reproducibility

1 Introduction

Community monitoring can connect housing conditions with the indoor environment, provided that the measurement record is interpreted in the context of recruitment, deployment and data continuity. The associated Philadelphia field study described a pathway from 54 targeted homes to 11 installations and five usable datasets [1]. That progression makes data availability part of the scientific question. The present study analyzes the available records from those five household folders and asks what environmental structure can be described reproducibly within them.

Short windows offer a practical representation of transient changes because they retain conditions before and after a trigger. Overlap is intentional: successive excursions can occur close enough that their contextual records share observations. The corresponding feature vectors therefore have two levels of dependence, through shared bins and through repeated observation of the same household. A large window library does not increase the number of households represented. Structured validation must respect the target of inference and the dependence of the source data [3].

This analysis has three objectives: reproduce the archived event-window and clustering results directly from exports; quantify coverage, household imbalance and window dependence; and examine whether the descriptive partition persists under defensible changes to the analysis. The primary estimand is the distribution of profiles among the available, eligible, PM2.5-triggered contextual windows. Equal-household weighting is examined as a distinct descriptive target. The quantitative work was performed in October 2026 as an analytical extension of the published 2024 fieldwork account.

2 Materials and methods

2 1 Study context and source records

The fieldwork involved Drexel University, the Philadelphia Energy Authority, Energy Coordinating Agency and Esperanza. The published account describes partner-assisted household access, installation and retrieval, together with surveys and participant communication [1]. The present analysis uses the four named AirVisual export channels: PM2.5 in µg/m³, carbon dioxide in ppm, temperature in °C and relative humidity in percent. These channels agree with the field-era AirVisual Pro guide [2]. Awair and the survey material provide context for the broader field study; neither contributes variables to this event-window analysis.

Forty-nine semicolon-delimited export files were identified across five source household folders. The inherited H01–H05 pseudonym mapping was preserved. The available retained timestamps extend from September 2022 to June 2023 for H01 and H03, and from March to June 2023 for H02, H04 and H05. These are export spans. All timestamp calculations use the exported epoch interpreted as UTC, and no hour-of-day behavioral inference is made. Source hashes and exact private paths were recorded before analysis. Original exports, historical scripts and previous results were retained without alteration. An additional unnamed folder contained four files with 2,296 valid rows; every timestamp and four-channel tuple matched H01 exactly. Those derived copies therefore added no independent observations and remained excluded.

2 2 Ingestion and missingness

Trailing empty export fields were handled explicitly and the named numeric columns were parsed. The reproduced date screen retained epochs from 1 January 2022 through 31 December 2023. Exact duplicates across timestamp and the four channels were removed, then all rows at conflicting timestamps were excluded. The primary plausibility screens were PM2.5 0–1,000 µg/m³, CO_2 250–10,000 ppm, temperature −10 to 60 °C and relative humidity 0–100%. Values outside these bounds became missing rather than zero. These are inherited analysis screens; they are not laboratory calibration criteria.

Each channel was aggregated to left-labelled ten-minute bins using the median of available valid readings. A measured bin required all four channels to be present. Missing runs of one or two bins could be interpolated in time only inside observed endpoints; longer runs remained missing. Complete four-channel windows were retained only when at least 80% of their 36 bins were measured, equivalent to at least 29 measured bins. The measured flag was preserved after interpolation. Coverage was calculated within each export span rather than against an assumed common deployment duration.

An independent parsing audit verified the measured-bin mask and measured values against the replayed grids. It counted excluded dates, duplicate and conflicting rows, channel-specific screening, and observation density per bin. A strict sensitivity removed every non-measured complete-channel bin before rerunning detection and extraction. A second required at least 90% measured bins. After the raw audit identified readings above the inherited PM2.5 ceiling, two additional stress tests retained all nonnegative PM2.5 readings or capped them at 1,000 µg/m³ before aggregation. The rationale for this addition was recorded separately before those analyses were run.

2 3 Excursion detection and contextual windows

For each ending bin t, ordinary least squares was applied to six successive PM2.5 values with indices i = 0, …, 5. The slope was $b(t) = \Sigma[(i - 2.5)y(t - 5 + i)] / 17.5$. Its unit is µg/m³ per ten-minute index step. Six bin labels span 50 minutes. A potential trigger required $b(t) > 4$, equivalent to 0.4 µg/m³ per minute, and a six-point range of at least 20 µg/m³. The local minimum of these six values defined the baseline. At least two values among the trigger and next six bins also had to exceed baseline + 20 µg/m³. The code permits these values to be nonconsecutive and interpolated; describing this check as two consecutive measured rises would be inaccurate.

After a qualifying trigger, repeat triggers were suppressed until two consecutive measured bins recovered to at most baseline + 20 µg/m³ following an elevation, a PM2.5 outage intervened, or the 72-hour suppression cap was reached. A six-hour contextual window started six bins before the trigger and comprised 36 ten-minute bins. With left-bin labels, its domain is [−1, 5) hours relative to the trigger. The last plotted label is 4 hours 50 minutes, while the final bin ends at 5 hours. The window is a fixed context, not an estimate of the excursion duration. Close triggers can generate overlapping context after the preceding episode has recovered.

The primary threshold and recovery definition were frozen from the archived working analysis. Slopes 1, 2 and 5 were examined as sensitivity choices alongside the primary 4. A new implementation reconstructed the detector and twelve features independently, comparing household, start, trigger and suppression-end indices with the preserved registry. Because persistence checks look ahead, interpolation can use later observations and complete context extends five hours beyond the trigger, this is a retrospective extraction procedure.

2 4 Physical features and grouping

For each window, the mean, sample standard deviation and range were calculated for each of the four channels, producing twelve features. The stored order is four means, four standard deviations and four ranges. Timestamps, household labels and derived slope columns were metadata only. StandardScaler centered each feature and divided it by its population standard deviation across windows. Principal component analysis retained the smallest number of components reaching at least 90% cumulative standardized variance, using the full singular-value decomposition implementation [5,6].

K-means was fitted for k = 2 through 10, using 30 initializations and random seed 42. The saved selection rule required the smallest group to contain at least max(5, 0.01N) windows and median adjusted Rand index (ARI) of at least 0.75 against five additional fits using seeds 3, 11, 23, 67 and 91 with ten initializations each. Among eligible candidates, the lowest Davies–Bouldin index was selected. Silhouette values were reported alongside this index. ARI measures partition agreement after allowing arbitrary label permutations; it is not classification accuracy. Internal separation and repeated-seed reproducibility address different questions from sampling stability [4,6].

Representative plots use actual windows nearest the corresponding centroid in retained PCA space. A centroid in feature space does not uniquely determine a time series: mean, standard deviation and range do not encode temporal order. For this reason, no reverse-engineered centroid waveform is presented. Group descriptions use empirical within-group summaries, and group numbering remains that of the reproduced primary run.

2 5 Dependence and units of inference

Windows within each home were ordered by start index. Overlap was defined as sharing at least one ten-minute bin. We counted adjacent overlapping pairs, all intersecting pairs, connected overlap components and unique covered bins. A greedy earliest-start procedure retained the first window and then the next starting at or after its end, producing a disjoint sensitivity subset. Equal window lengths make this a maximum-cardinality disjoint selection for the ordered intervals, but it is only one possible subset. Neither components nor disjoint windows are assumed independent, because environmental conditions can persist beyond six hours within the same home.

For claims about new homes, the household is the relevant sampling unit. Five observed homes do not support a precise population-level estimate from 1,404 rows of derived windows. Therefore, the analysis reports descriptive counts, agreement and distance diagnostics without window-level significance tests, population confidence intervals or an estimated effective independent sample size. Missingness is described rather than assumed random, and counts are not treated as comparable household activity rates.

2 6 Sensitivity and stability analyses

Each sensitivity refitted scaling, PCA and k-means. Fixed-k = 3 agreement with primary labels was calculated only on exactly shared household/trigger keys; windows with changed trigger times were not forced into a

correspondence. Where the original unweighted rule applied, candidate k values were selected again and reported separately. Scenarios included the disjoint subset, the stricter coverage rules, alternative slopes, each leave-one-household-out fit, robust median/interquartile scaling, and log1p transformation of the six PM2.5/CO_2 mean, standard-deviation and range features. A within-household-centered sensitivity subtracted each home's feature means before pooled standardization and PCA, changing the target to within-home deviations.

For equal-household weighting, every window received weight proportional to 1/n(h), so each of the five homes contributed the same total weight. The weighted mean and variance, weighted covariance PCA and weighted k-means were all fitted under these weights. PCA again retained at least 90% variance and k was held at three. The original unweighted Davies–Bouldin selection rule was not reused as a weighted model-selection criterion. This is a sensitivity to the chosen descriptive target, not a claim that the sparse homes supply equally reliable information.

Conditional sampling stability was assessed in 100 resamples for each weighting scheme. Each household export span was divided into fixed seven-day bins anchored to its grid origin. Bins, including empty bins, were sampled with replacement within each home; a window belonged to the bin containing its trigger. The entire scaling–PCA–k-means pipeline was refitted at k = 3, then assigned the fixed original library. Agreement was measured against the corresponding original weighted or unweighted fit. Empirical 2.5th and 97.5th percentiles summarize perturbation, with no population confidence-interval interpretation. Windows crossing bin boundaries and residual dependence longer than seven days remain limitations of this diagnostic.

2 7 Leakage and reference checks

To quantify the consequence of randomly splitting windows, 100 random 80/20 partitions were generated and the fraction of test windows sharing at least one bin with a training window was calculated. An alternative chronological diagnostic split each home at its 70th percentile window start, retaining training windows ending at least 72 hours before the cut and test windows starting at least 72 hours after it. Scaling, PCA and k-means at k = 3 were fitted only to training data. The ratio of median test to training nearest-centroid distance was used as a descriptive domain-shift measure. Leave-one-household-out fits provided a complementary distance diagnostic for omitted homes. These procedures evaluate representation behavior rather than correctness of activity labels.

The preserved reference material consists of thirteen earlier development/checking blocks and four subsequently selected 24-hour blocks from H01, H03, H04 and H05. The archived protocol describes detector-blind review of observed signal traces; annotations were not recreated or amended in this audit. The fresh set includes eight directional changes: five rises and three falls. A one-to-one assignment matched retained trigger times to annotated intervals, initially allowing 20 minutes on either side. We replayed this check and repeated it with ten-minute and zero tolerance. The rise-only subset followed the archived description-based rule. The reference periods were not used to optimize the new sensitivity scenarios.

3 Results

3 1 Reproduction and available coverage

All five replayed grids matched the saved grids exactly. The independent detector reproduced all primary window indices, and the twelve reconstructed features had maximum absolute difference zero from the preserved feature matrix. The replayed cluster partition had ARI 1.000 against the original. Forty-nine export files contained 2,488,573 rows: 199 rows with 1970 reset-era dates, 49,684 exact duplicate rows and two conflicting rows were removed, leaving 2,438,688 unique timestamped rows. No malformed rows were found. The primary channel screen marked 2,062 PM2.5 values above 1,000 µg/m³ and two already-missing PM2.5 values; no values in the other channels failed the inherited bounds.

Measured coverage varied from 2.10% for H02 to 95.09% for H01 (Table 1 and Figure 1). H01 and H03 each spanned about ten months, whereas the remaining records spanned about 113 days. The median count of source observations in a nonempty ten-minute bin was 60 for H01 and two for each other home. Thus, a common bin

duration does not imply equal within-bin sampling density. H03 gained 6,242 complete bins through short-gap interpolation, compared with 294 for H01. The number of eligible six-hour start positions varied from 192 to 40,245 across homes.

Table 1. Coverage and window support within each recorded span. Filled bins are additional complete bins after interpolation.

Home	Span days	Measured bins	Measured %	Filled bins	Windows	Disjoint
H01	298.8	40,919	95.09	294	1,116	586
H02	112.8	341	2.10	53	6	3
H03	295.2	34,907	82.10	6,242	133	111
H04	112.9	2,815	17.32	505	24	18
H05	112.8	12,333	75.96	2,192	125	81

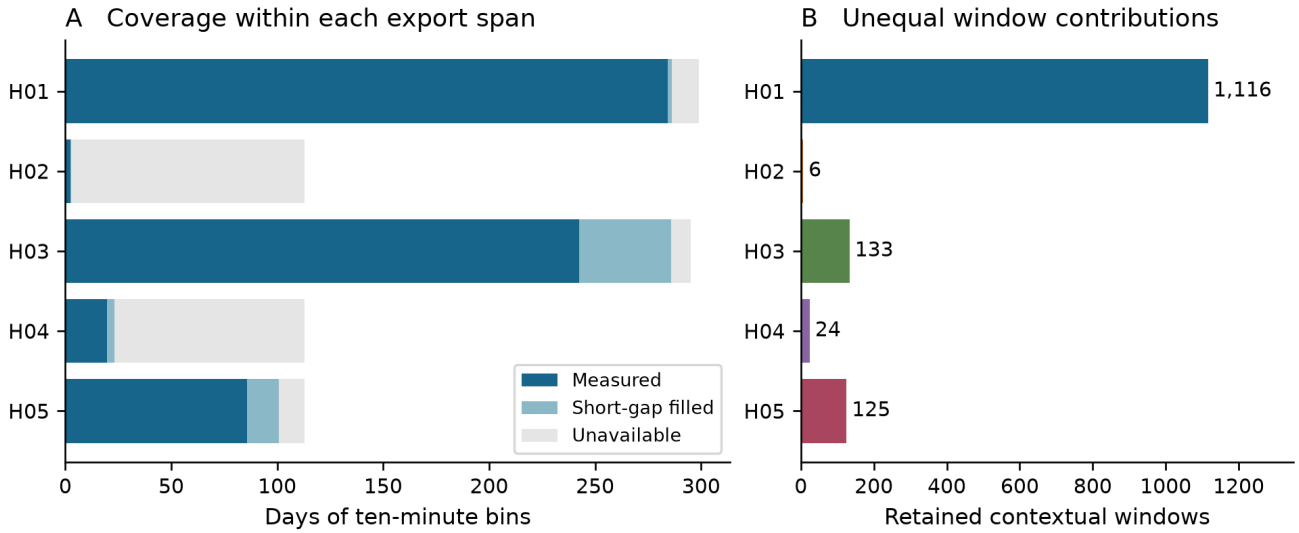


Figure 1. Measured, short-gap-filled and unavailable ten-minute bins within each household export span, and retained six-hour contextual windows. Span includes outages and is not verified occupancy or deployment duration. Household labels are pseudonyms.

3 2 Window dependence

The detector retained 1,404 contextual windows: H01 1,116; H02 6; H03 133; H04 24; and H05 125. H01 therefore contributed 79.49%. Of 1,399 adjacent within-home pairs, 811 (57.97%) overlapped. Counting nonadjacent intersections as well gave 1,125 overlapping pairs. The library contained 50,544 window-bin slots drawn from 35,935 unique bins, a mean multiplicity of 1.407 among used bins. At most five windows shared an individual bin. There were 593 connected overlap components and 799 windows in the greedy disjoint subset. These counts characterize reuse of measurements rather than independent replication.

3 3 Primary environmental profiles

Six components retained 95.793% of standardized feature variance. The first two accounted for 53.941%, so a two-dimensional PCA view alone would omit substantial retained structure. The primary rule selected three groups, with Davies–Bouldin index 1.3055, silhouette 0.3295 and median repeated-seed ARI 1.000. The two-group silhouette was higher (0.3642), while the four-group Davies–Bouldin index was lower (1.0748). The latter candidate was excluded because its smallest group contained 14 windows, just below the 14.04 size criterion.

These differences show that three groups follows the stated eligibility and selection rule rather than a unique optimum agreed upon by every criterion (Figure 3).

The three groups contained 1,029, 261 and 114 windows. Group 2 had the largest median PM2.5 window mean and range; group 3 had the largest median CO_2 window mean (Table 2). H01 contributed 239 of 261 group-2 windows (91.57%), and H05 contributed 79 of 114 group-3 windows (69.30%). All five homes appeared in groups 1 and 3, while H02 supplied no group-2 windows. The joint composition in Figure 2 supports both within-household variation and strong differences in household representation.

Table 2. Primary group summaries. Values are medians of six-hour window means, except PM2.5 range, which is the median within-window maximum minus minimum.

Group	Windows	PM2.5 mean µg/m³	PM2.5 range µg/m³	CO_2 mean ppm	Temp mean °C	RH mean %
1	1029	31.80	135.55	473.64	23.63	34.42
2	261	111.82	484.90	522.82	23.82	42.86
3	114	26.42	115.32	639.07	24.88	39.66

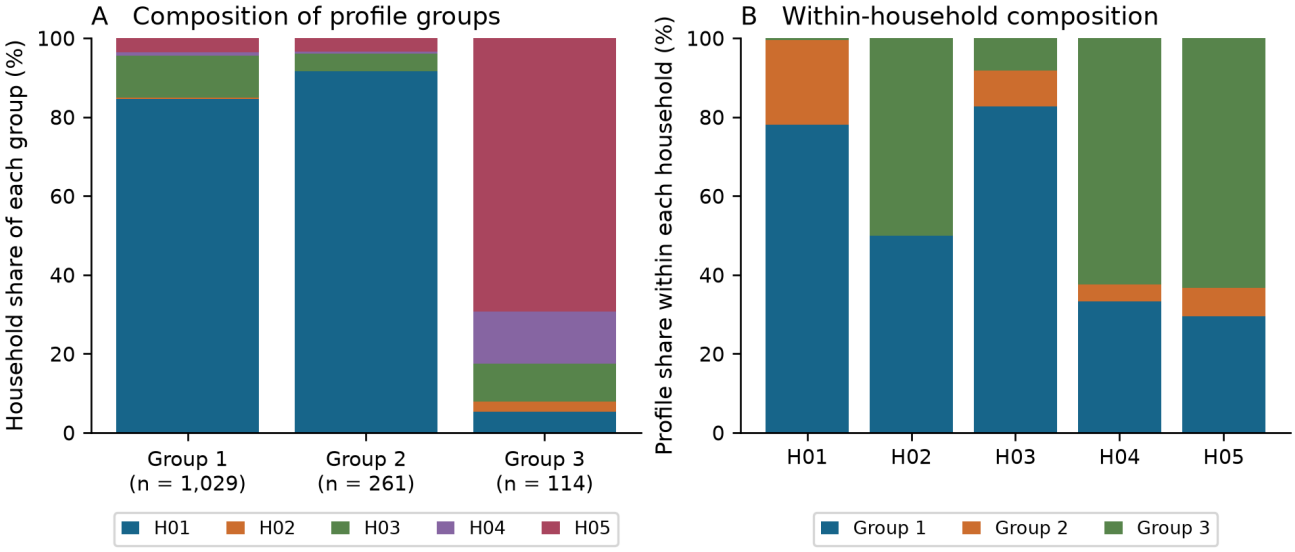


Figure 2. Profile composition under the reproduced window-weighted three-group solution. H01 contributes 239/261 windows in group 2 and H05 contributes 79/114 in group 3. Panel B uses different denominators and shows that profiles also vary within homes.

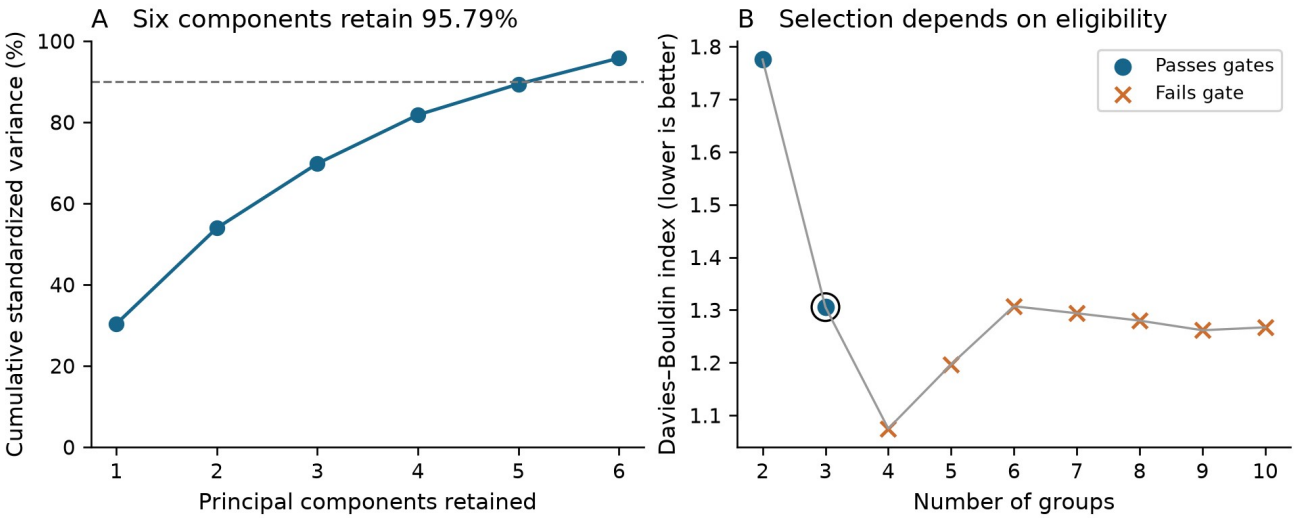


Figure 3. StandardScaler and PCA are fitted to the full exploratory window library. Candidate groups use the saved size and random-seed gates. The four-group solution has a lower Davies–Bouldin index but its smallest group has 14 windows, below the required 14.04 (at least 15 integer windows). Selection of three groups is conditional on this rule.

The PCA coefficients also distinguish correlated feature families. In the first component, CO_2 mean, standard deviation and range had coefficients of approximately 0.406, 0.427 and 0.430. In the second, PM2.5 mean, standard deviation and range had coefficients of approximately 0.513, 0.527 and 0.506. Coefficient signs are arbitrary up to component orientation. These coefficients describe the standardized representation, not causal importance. Actual nearest-centroid traces are shown in Figure 4; they preserve variation within the six-hour context that summary features alone cannot reconstruct.

3 4 Sensitivity to representation and household contributions

At fixed k = 3, the disjoint subset agreed closely with primary labels (ARI 0.990 on 799 windows), although the selection rule chose four groups for that subset. Alternative slope thresholds retained 1,348–1,455 windows and agreed at ARI 0.989–0.993 on shared trigger keys, but reselected k ranged from four to eight. A 90% measured-coverage requirement retained 1,362 windows (ARI 0.990), and the strict no-interpolation pipeline retained 1,315 (ARI 0.916 on 1,301 common windows). Removing interpolation can create additional triggers by changing suppression around outages, so this sensitivity is not simply deletion from the primary registry.

Partition agreement fell to 0.344 under equal-household weighting, 0.342 under robust scaling and 0.295 after logarithmic PM2.5/CO_2 feature transformation. Within-household centering gave ARI 0.621. Omitting H01 left 288 windows and yielded ARI 0.575 on those windows, with five groups selected; omitting H05 gave ARI 0.742, while omission of H02 left the common primary partition unchanged. Table 3 and Figure 5 distinguish the fixed-three-group comparison from the number of groups reselected in each scenario.

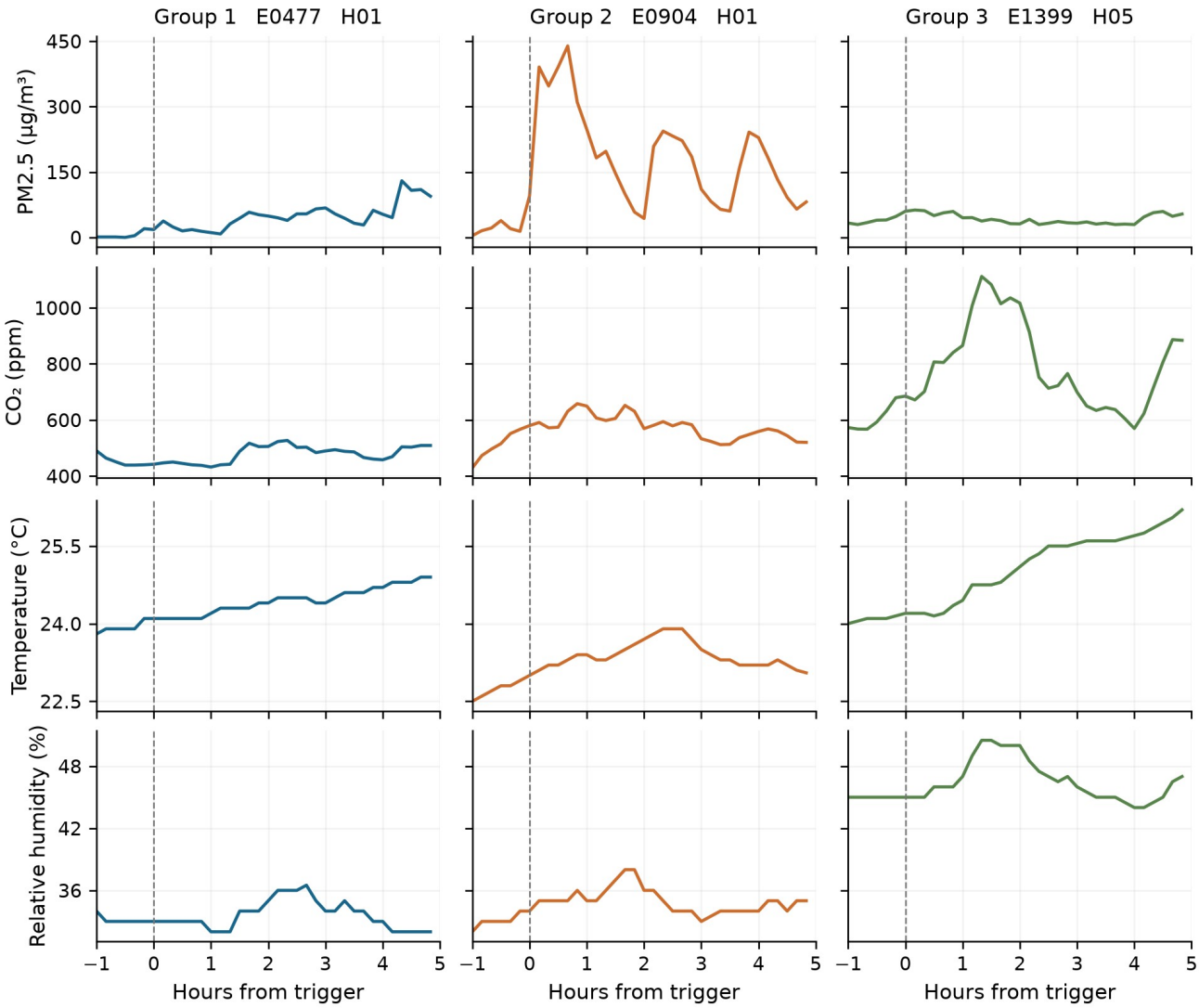


Figure 4. Actual windows nearest each fitted group centroid in retained PCA space. Axes are shared within channels, with separate physical units. Trigger is at zero; the 36 left-labelled ten-minute bins cover [−1, 5) hours. These are contextual measured/short-gap-filled traces, not reconstructed time series or identified household activities.

Table 3. Full sensitivity results. ARI always uses fixed k = 3 on exact shared windows. Fixed means k = 3 for the weighted target without reselection.

Scenario	N	Common N	PCs	Selected k	Fixed 3 ARI
Primary	1404	1404	6	3	1.000
Disjoint earliest	799	799	6	4	0.990
Slope 1	1455	1125	6	8	0.993
Slope 2	1449	1176	6	7	0.990
Slope 5	1348	1155	6	4	0.989
90% measured	1362	1362	6	3	0.990
No interpolation	1315	1301	6	3	0.916

Scenario	N	Common N	PCs	Selected k	Fixed 3 ARI
Equal household weight	1404	1404	5	Fixed	0.344
Robust scaling	1404	1404	4	3	0.342
Log PM/CO_2 features	1404	1404	6	5	0.295
Within-home centered	1404	1404	6	5	0.621
Omit H01	288	288	5	5	0.575
Omit H02	1398	1398	6	3	1.000
Omit H03	1271	1271	6	4	0.991
Omit H04	1380	1380	6	5	0.995
Omit H05	1279	1279	6	3	0.742
Retain PM above 1000	1409	1403	6	3	0.686
Cap PM at 1000	1409	1403	6	3	0.920

The above-ceiling PM2.5 values reached 3,601.8 µg/m³ in H01, 4,488.4 in H03 and 4,200.7 in H05. Retaining them produced 1,409 windows and ARI 0.686 on 1,403 common windows; capping them at 1,000 also produced 1,409 windows, with ARI 0.920. Both reselected three groups. The modest change in window count therefore coexisted with appreciable changes in the fitted partition. The validity of these extreme sensor outputs remains unresolved.

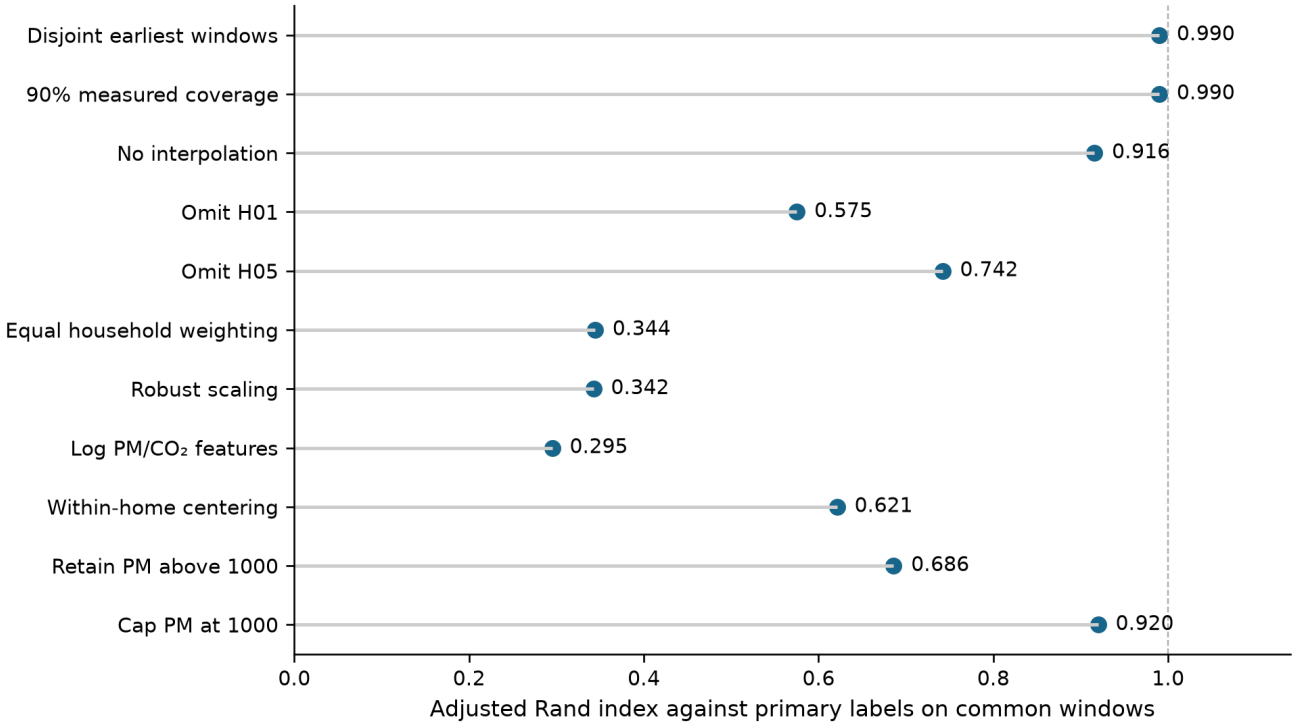


Figure 5. Partition agreement with the primary fit while holding k = 3. Every sensitivity refits scaling, PCA and k-means. Comparisons use exact shared household/trigger keys. Equal weighting and centering change the descriptive target; they are stress tests, not automatically better estimates. Selected k varies separately in Table 3.

3 5 Conditional stability and leakage diagnostics

Seven-day block resampling yielded median ARI 0.930, with empirical 2.5th–97.5th percentiles 0.835–0.973, for the primary weighting. For equal-household weighting, the corresponding values were 0.921 and 0.279–0.977 against that scheme's own original fit (Figure 6). Some resamples represented only four homes because a sparse home's occupied blocks were not drawn. The broader lower tail under equal weighting exposes how heavily that target depends on limited observations from sparse homes. Repeating k-means seeds on unchanged data did not reveal this sensitivity.

In the random 80/20 split diagnostic, an average 72.89% of test windows shared source bins with at least one training window; the empirical range between the 2.5th and 97.5th percentiles was 67.26–77.99%. The chronological purged split retained 951 training and 396 test windows from four homes, with no eligible H02 windows remaining. The median test/training nearest-centroid distance ratio was 1.007. Omitted-home ratios ranged from 1.089 for H01 to 3.990 for H04. These distances describe how the fitted representation encounters held-out data; no distance threshold was calibrated as a generalization criterion.

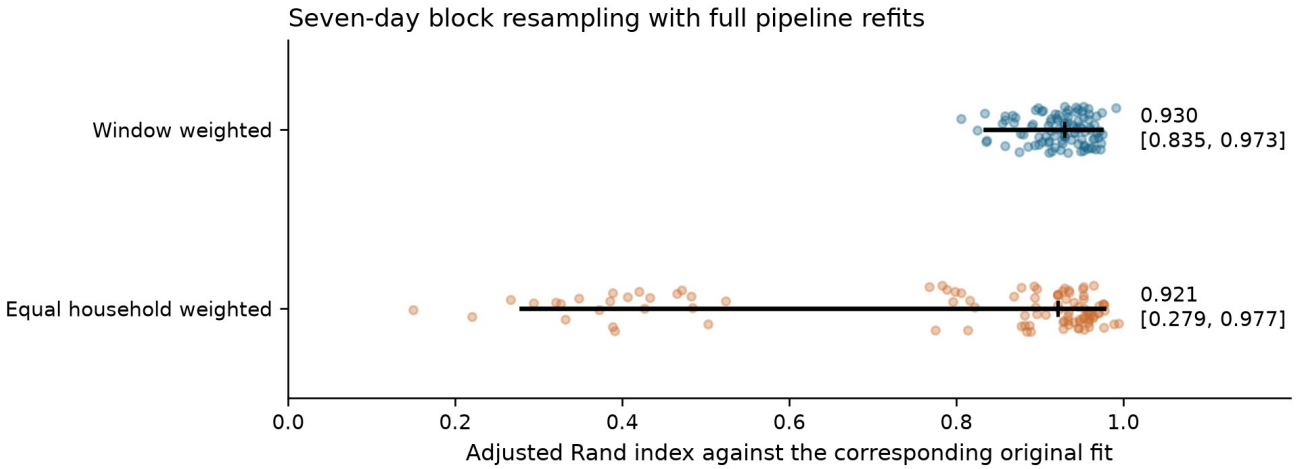


Figure 6. One hundred resamples per weighting scheme, resampling seven-day bins separately within households, including empty bins. Each resample refits scaling, PCA and fixed-k = 3 k-means and predicts the fixed library. Lines show empirical 2.5th–97.5th percentiles; ticks show medians. These are conditional perturbation ranges, not confidence intervals or validation in new homes. Some resamples contain only four homes.

3 6 Archived signal reference check

All five rising references matched five retained triggers in the four archived 24-hour blocks. Four rising references were in H01 and one was in H05. H03 and H04 each supplied a block with no qualifying annotated change and no retained candidate. Considering all eight directional references gave five matches and three unmatched falls, as expected for a positive-rise target. Counts were unchanged when the interval tolerance was reduced from 20 minutes to ten minutes or zero. The reference check is reported through its counts and target definition; the five-positive result is not used as a population detection-rate estimate.

4 Discussion

4 1 Scientific interpretation of the profiles

The available data support a reproducible library of PM2.5-triggered contexts and interpretable associations among their physical summaries. A particle-variable group and a group with higher CO_2 window means can be identified under the primary representation. The household composition and sensitivity results, however, constrain the meaning of those groups. In particular, stable optimization from different seeds does not establish stable environmental classes across homes. The analysis reaches stronger conclusions about how the observed library is organized than about how a new household would behave.

Intentional overlap is compatible with exploratory description. It becomes problematic when shared observations are treated as independent replication or placed on opposing sides of a validation split. Here, the nearly unchanged fixed-three-group partition in the disjoint subset indicates that direct overlap alone does not account for the primary partition. The substantial changes after household reweighting, omission and feature transformation show that unequal representation and the metric used to define similarity also matter. This distinction avoids attributing every weakness to the windowing design.

Equal weighting emphasizes a different question: what profile distribution would describe these five observed homes if each were given the same total contribution? Because H02 has only six windows, this weighting amplifies a very small and incomplete record. It does not recover the missing observations, correct nonrandom participation or turn the sample into a representative housing survey. Neither weighting scheme is a universal correction. Reporting both reveals which findings depend on the target being summarized.

4 2 Limitations of measurement and selection

The household folders constitute a small, operationally selected sample. The fieldwork progression from 54 targets to five usable datasets and the unequal measured coverage limit generalization to other homes. Export spans may include time outside verified in-home deployment; installation, relocation and removal logs were not linked sufficiently to restrict every retained timestamp. Household identity was inferred from the preserved source-folder mapping, not independently revalidated against device movement records. Weatherization dates, occupancy records, outdoor concentrations and contemporaneous source diaries were not incorporated, so before/after retrofit effects and specific household activities cannot be inferred.

The primary plausibility bounds provide reproducibility but do not establish measurement accuracy. No traceable reference-instrument calibration, humidity correction, detection-limit assessment or quantitative sensor-drift evaluation was available in the audited material. H01’s sixty readings per nonempty bin and the other homes’ median of two can yield different noise attenuation after median aggregation. Screening PM2.5 above 1,000 µg/m³ changes the feature geometry, while retaining those values assumes an accuracy that has not been verified. The device guide lists an operating temperature upper bound of 40 °C [2], whereas the inherited code accepts values through 60 °C and H04 includes a raw maximum of 41 °C. Operating specifications, data plausibility and validated concentration accuracy are separate issues.

Missingness can depend on power, connectivity, placement, participation or conditions related to the exposure itself. No missing-at-random mechanism was established. Short interpolation smooths excursions and may affect slope, range, recovery and window eligibility. The strict sensitivity exposes part of this dependence, but cannot reveal unrecorded events. A measured bin requires channel completeness rather than a minimum count of observations per channel, so coverage percentages do not quantify equal sensor information. The PM2.5 trigger also selects a conditional subset of IAQ conditions and omits steady elevated concentrations, isolated changes in other channels and downward-only transitions.

4 3 Limits of clustering and validation

The correlated mean, standard-deviation and range features can give a channel several contributions to distance. Standardization does not remove this redundancy, and mean centering preserves between-home differences in the primary fit. Robust scaling and logarithmic transformations change the geometry, with large consequences here. PCA variance retention is a representation criterion; it does not establish relevance to health, occupancy or source categories. Euclidean k-means can subdivide continuous variation, and the chosen group count depends on the small-group gate. A 14-window fourth group was excluded by a rule only slightly above that size. The apparent precision of k = 3 should therefore not be interpreted as evidence for exactly three physical processes.

The seven-day resampling range describes perturbations conditional on the observed homes, chosen block length and fitted feature system. Trigger-based block allocation allows some contextual windows to cross block boundaries. Longer environmental dependence may also persist, and only five households are available for higher-level inference. No confidence interval or p-value is justified by treating these resamples as a replacement

for new household sampling. The purged chronological split reduces direct shared-bin leakage, but the detector and its conventions had already been developed within these same source homes. Its distance ratio and the leave-home-out diagnostics are stress checks, not validated classification performance.

The reference annotations are concentration-pattern judgments, not participant-confirmed activity ground truth. Their claimed detector-blind provenance is recorded in the archived protocol but cannot be established by re-executing the scoring code. The earlier thirteen blocks belong to development/checking, and the four fresh blocks were selected for sufficient continuous coverage rather than randomly sampled from deployment time. The rise-only result was derived from the broader annotation set. Four of five positive references came from one home, matching intervals can be long, and no independent adjudication or inter-rater agreement was documented. Zero unmatched candidates in these selected blocks does not establish zero false alarms in the full record. New labeled data are needed to estimate sensitivity, false alarms, timing error and activity-identification performance.

4 4 Implications for further research

A prospective study should define the validation target before deployment: within-home future excursions, transfer to unseen homes, or attribution to a recorded activity. For an activity target, synchronized diaries or direct reference observations should distinguish source type, start and end, with blinded independent review and disagreement resolution. Calibration and deployment logs should accompany device identifiers, while coverage and retrieval outcomes should be recorded for all recruited homes. With a larger household sample, evaluation can reserve entire homes and contiguous periods, fit every preprocessing step on training data, and report uncertainty at the household level. Predeclared detector thresholds, coverage criteria and cluster-selection rules would reduce retrospective flexibility.

The useful output of the present work is an auditable empirical starting point for that design. The preserved primary workflow, alternative estimands and stress tests make it possible to identify which conclusions need new measurements rather than further optimization on the same library. The findings support environmental-profile description and targeted follow-up; they do not establish health effects, retrofit benefits, household activity prevalence or scalable detection accuracy.

5 Conclusions

The archived five-home analysis was reproduced exactly from 49 original export files, and an independent implementation confirmed the retained windows and physical features. Intentional six-hour overlap preserved excursion context while creating substantial shared measurement support. The primary window-weighted representation yielded three interpretable profiles, but household weighting, scaling and treatment of extreme PM2.5 readings materially changed the partition. The defensible conclusion is an exploratory characterization of the available records, supported by explicit dependence diagnostics and reproducible sensitivity analyses. Generalization requires additional households and independent reference observations.

Data code and study documentation

The accompanying versioned package contains the scripts, environment versions, aggregate result tables, figure provenance, source hashes, archived baseline results and scoring records needed to reproduce this analysis. Original household exports and source-path mappings remain private. The local owner can rerun the workflow against the original source root; public figures use pseudonyms and relative time. No missing concentrations or activity labels were fabricated. The full file inventory and verification record are supplied outside this manuscript.

This manuscript extends the previously published fieldwork account [1] and does not replace its original results. Formal ethics protocol identifiers, authorization for secondary public release of participant-level records, funding details and current author declarations were not established from the audited analysis files. Those administrative

details require confirmation by the study team before external submission. The source manuscript author list is retained for review; this version does not imply coauthor approval.

References

[1] Tajik R, Hoque S. Challenges in Mixed Methods Research: Data Collection and Assessment of Indoor Air Quality (IAQ) in Low-Income Households. In: Proceedings of the 7th Residential Building Design and Construction Conference. Pennsylvania Housing Research Center; 2024:471–487. Official proceedings: https://www.phrc.psu.edu/assets/docs/Publications/2024RBDCCPapers/2024-RBDCC-Proceedings.pdf. The archived extract was inspected directly; PDF pages 480–496 correspond to printed pages 471–487.

[2] IQAir. AirVisual Pro User Guide. Version 3.3, 2021. https://www2.iqair.com/sites/default/files/documents/AVPro_Indoor_User_Manual_v3.3en_092721.pdf. Accessed 14 October 2026.

[3] Roberts DR, Bahn V, Ciuti S, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40:913–929. https://doi.org/10.1111/ecog.02881.

[4] von Luxburg U. Clustering Stability: An Overview. Foundations and Trends in Machine Learning. 2010;2(3):235–274. https://doi.org/10.1561/2200000008. Author manuscript: https://arxiv.org/abs/1007.1075.

[5] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html.

[6] Scikit-learn developers. Clustering and clustering performance evaluation. Version 1.9.1 documentation. https://scikit-learn.org/stable/modules/clustering.html. Accessed 14 October 2026.

Appendix A Reproduction and inference checks

Table A1. Verification checks and interpretation.

Check	Verified result
Raw ingestion	Five replayed grids match saved grids exactly
Independent extraction	All 1,404 window keys match; maximum feature difference 0
Primary clustering	ARI 1.000 against saved labels; six PCs retain 95.793%
Signal check	5/5 rising references matched; 5/8 all-direction changes matched
Reference tolerance	Same counts at 0, 10 and 20 minutes
Physical reconstruction	Representatives are actual windows, not inverse-PCA waveforms
Unit of inference	Five homes; overlap counts are not independent sample sizes
Generalization status	Exploratory; no externally validated activity classifier

Appendix B Source and method reconciliation

The historical course scripts and manuscript contain multiple development states, including absolute-slope triggers, additional numeric channels, time-of-day metadata entering PCA, feature-name ordering inconsistent with array construction, and interpolation across the full source span. The manuscript also describes both use and later removal of PCA. The present report follows the explicitly retained PCA workflow and reproduces the October 2026 four-channel implementation. Original source versions are preserved so these methodological

differences are inspectable. The analytical extension is not presented as an exact rerun of every historical course result.

The published field paper’s sensor discussion is inconsistent with the field-era AirVisual guide regarding VOC channels. The present analysis uses only the four verified export channels and does not rank monitors by an unsupported sensitivity or calibration comparison. The official conference page identifies the event as March 27–28, 2024, whereas a recurring footer in the archived proceedings says May. The bibliographic reference uses the year and page range to avoid propagating that footer inconsistency. These reconciliations concern documentation; no original publication or raw source file was edited.