atmosphere

MDPI

Article

Adaptive Data-Driven Framework for Unsupervised Learning of Air Pollution in Urban Micro-Environments

Abdelrahman Eid [1,2,*], Shehdeh Jodeh [3,*], Raghad Eid [4], Ghadir Hanbali [3], Abdelkhaleq Chakir [5] and Estelle Roth [5]

[1] Department of Mathematics, An-Najah National University, Nablus P.O. Box 7, Palestine
[2] Data Science Unit, An-Najah Innovation Park, Nablus P.O. Box 7, Palestine
[3] Department of Chemistry, An-Najah National University, Nablus P.O. Box 7, Palestine
[4] Department of Mathematics, Palestine Technical University, Tulkarm P.O. Box 7, Palestine
[5] Groupe de Spectrométrie Moléculaire et Atmosphérique (GSMA), UMR CNRS 7331, Université de Reims, Moulin de la Housse B.P. 1039, CEDEX 02, 51687 Reims, France
* Correspondence: abed.eid@najah.edu (A.E.); sjodeh@najah.edu (S.J.); Tel.: +970-598-117416 (A.E.); +970-599-590498 (S.J.)

Abstract

(1) Background: Urban traffic micro-environments show strong spatial and temporal variability. Short and intensive campaigns remain a practical approach for understanding exposure patterns in complex environments, but they need clear and interpretable summaries that are not limited to simple site or time segmentation. (2) Methods: We carried out a multi-site campaign across five traffic-affected micro-environments, where measurements covered several pollutants, gases, and meteorological variables. A machine learning framework was introduced to learn interpretable operational regimes as recurring multivariate states using clustering with stability checks, and then we evaluated their added explanatory value and cross-site transfer using a strict site hold-out design to avoid information leakage. (3) Results: Five regimes were identified, representing combinations of emission intensity and ventilation strength. Incorporating regime information increased the explanatory power of simple NO_2 models and allowed the imputation of missing H_2S day using regime-aware random forest with an R^2 near 0.97. Regime labels remained identifiable using reduced sensor sets, while cross-site forecasting transferred well for NO_2 but was limited for PM, indicating stronger local effects for particles. (4) Conclusions: Operational-regime learning can transform short multivariate campaigns into practical and interpretable summaries of urban air pollution, while supporting data recovery and cautious model transfer.

Keywords: traffic-related air pollution; urban micro-environments; cross-site generalization air quality; interpretable machine learning; transfer learning for air quality prediction; pollution regime classification; unsupervised clustering environmental data; sensor data imputation; data-driven air quality regimes; multi-pollutant modeling; SDG 11 sustainable cities and communities

Academic Editor: William R. Stockwell

Received: 22 November 2025
Revised: 14 January 2026
Accepted: 22 January 2026
Published: 24 January 2026

1. Introduction

A micro-environment is a physical or functional space where contaminant concentrations are relatively homogeneous and well mixed while a person is present, e.g., an automobile cabin, a garage, a forecourt, a classroom, or a workplace zone [1]. Urban residents spend substantial amounts of time in micro-environments shaped by traffic

flow, building geometry, and fuel handling, where pollutant mixtures and exposure intensities can differ markedly from values at regulatory background monitors. Near-road corridors, street canyons, and queue-prone intersections often show sharp spatial concentration gradients over tens of meters and strong diurnal variability tied to congestion and boundary-layer dynamics. These features are recognized in guidance and siting criteria for dedicated near-road monitoring [2]. Beyond the curbside, activities such as vehicle parking and refueling create semi-enclosed or source-proximate spaces that mix primary emissions with ventilation constraints. In many mid-sized cities, permanent monitoring networks are sparse, so short, intensive campaigns are often used to understand exposure patterns in these settings. A key challenge is to summarize a few days of multi-pollutant measurements into clear conditions that allow fair comparison across micro-environments and support exposure interpretation and follow-up decisions like targeted monitoring, ventilation checks, or operational guidance during high-risk regimes.

Vehicle exhaust emissions are a major contributor to air pollution in semi-enclosed traffic environments such as parking garages, where pollutants accumulate due to limited air exchange. These emissions contain a complex mixture of carbon monoxide (CO), nitrogen oxides (NO_x), hydrocarbons (HCs), aldehydes, and particulate matter (PM) [3,4]. In practice, CO is often used as an operational indicator in garages, yet fine particles remain a major health concern, and short exposure peaks can matter [5].

Traffic-related pollutants are linked to respiratory and cardiopulmonary risks, and guideline values emphasize the importance of both short and long exposures [6]. Some gases, such as hydrogen sulfide (H_2S), may remain low in ambient urban air but can rise episodically in poorly ventilated settings and cause acute irritation and nuisance odors [7]. When a campaign covers several micro-environments in only a few days, there is a practical need to summarize multi-pollutant conditions in a consistent way. This helps compare sites and time periods, link observed patterns to plausible drivers such as traffic activity and ventilation, and apply simple analyses without mixing very different conditions.

In this work, an *operational regime* refers to a recurring multi-pollutant condition that reflects the joint behavior of emissions and ventilation, expressed through a characteristic combination of pollutant levels, diurnal timing, and simple meteorological indicators [8–11]. We define regimes by grouping observations with similar multi-pollutant and meteorological patterns, and we then use the regime label as a compact description of conditions that matter for exposure interpretation and simple analyses. Because regimes are defined from the measured variables rather than fixed labels, a given regime can occur at different sites and at different times when conditions are similar. Practically, the regime approach turns a short multi-pollutant campaign into (i) a time series of regime labels at each site and (ii) per-regime summaries of pollutant levels and co-variation. This supports like-for-like comparison across micro-environments (same regime), identification of regime-specific peak exposure periods, and use of the regime label as an additional categorical predictor for simple analyses and gap-filling.

Compared with common segmentation approaches such as site-based grouping, fixed time-slot grouping, or single-pollutant thresholding, the regime approach aims to identify recurring multi-pollutant situations that can reappear across-sites and times [8,10,11]. This is useful in short campaigns because it provides a small set of interpretable conditions that can be summarized and tested by keeping some days and some sites completely separate for evaluation. It also supports simple analyses and gap-filling by using the regime label as a practical indicator of conditions, together with the other pollutants and meteorological variables measured at the same time.

Short multi-site urban campaigns reveal strong diurnal and micro-environmental structure, yet few studies formalize compact and interpretable multi-pollutant regimes,

test them using separate days and separate sites, and examine whether patterns identified at one location also appear at another [8–10,12,13]. Recent reviews confirm the growing use of machine learning in air quality studies, and they also highlight the importance of interpretability and careful evaluation when data are limited or heterogeneous [14,15]. We address this gap with three aims: (1) to define operational regimes from simple and physically grounded features across five contrasting micro-environments, (2) to examine whether the regime label helps explain key tracers beyond wind, time of day, and activity by testing on separate days and separate sites, and (3) to examine how far results can be used across locations by identifying when cross-site use is reasonable and when local effects dominate. Because the campaign duration is short and sources differ by location, we treat cross-site use cautiously and present it as limited to comparable settings rather than a universal rule.

Our contribution is a practical data-driven framework that links the regime approach to concrete uses in short campaigns. These uses include summarizing multi-pollutant conditions into a small number of interpretable situations, checking whether the regime label improves simple exposure-relevant models when tested on separate days and sites, and evaluating when cross-site use is reasonable and when site dependence dominates.

Hyperlocal monitoring has shown sharp concentration contrasts over short distances, which motivates multi-site designs that compare micro-environments within limited sampling periods [12,13]. Meteorological normalization further emphasizes that diurnal timing and mixing conditions are central for interpreting observed variability in traffic-related pollution [10]. Together, these findings motivate the regime approach as a structured way to summarize short campaign observations and to support cautious interpretation and limited use across comparable settings [9].

2. Materials and Methods

To identify and interpret recurring multi-pollutant conditions that govern exposure in traffic-affected urban micro-environments, we implemented a pipeline that links high-cadence, multi-site measurements to operational regime identification, robustness checks, and task-based evaluation. The pipeline is designed for short campaigns, where the practical goal is to summarize complex pollutant–meteorology mixtures into a small number of interpretable states (regimes) and to test whether the regime label is useful for downstream analyses. We evaluate performance using designs that keep whole days and whole sites separate, so the reported results reflect out-of-sample behavior rather than reuse of information from the held-out blocks.

Figure 1 shows the full analytical pipeline, from measurement to regime discovery, stability testing, and evaluation under day and site-blocked designs.

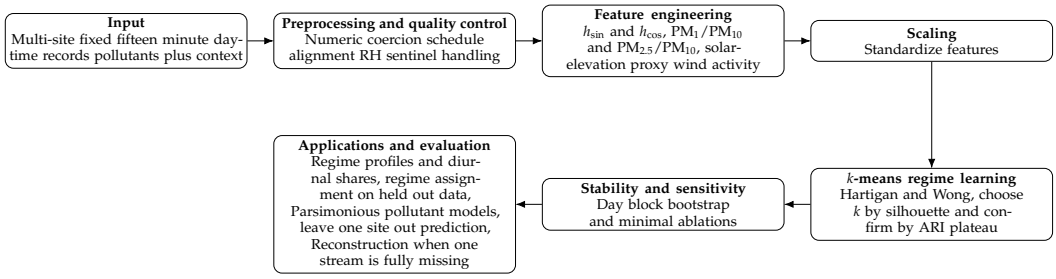

Figure 1. Analysis workflow.

As shown in Figure 1, we start from quality-controlled multi-site observations and derive physically grounded features that reflect traffic activity, ventilation, and diurnal timing. We then identify operational regimes by grouping similar multivariate feature

patterns, and we summarize each regime by its typical pollutant profile and its occurrence across micro-environments and hours. Finally, we evaluate three practical uses under day- and site-blocked designs: compact condition summarization, explanatory modeling for key tracers, and reconstruction when one pollutant stream is fully missing. In each evaluation, every data-driven step (scaling, regime identification, and model fitting) is learned on training blocks only and then applied unchanged to the held-out blocks.

2.1. Study Design, Instrumentation, and Data Quality Control

We conducted a short fixed-schedule campaign that covered five urban micro-environments typical of traffic-affected settings. The five sites were an open garage, a large roundabout with street canyon form, a closed municipal garage, a gasoline forecourt, and a campus entry that served as an urban background. Sampling followed a constant fifteen-minute cadence from 08:00 to 18:45 local time on four consecutive weekdays during August 2024. This window aligns with peak activity and human presence at these locations and is therefore the period of primary interest for exposure and operations.

We recorded PM_1, $PM_{2.5}$, PM_{10}, NO_2, CO, H_2S, and total VOCs. We also recorded relative humidity, wind speed, geographic coordinates, and a vehicle activity proxy that captures near field traffic intensity. The proxy was derived from an on site infrared traffic counter that produced fifteen minute counts. Site-wise summary statistics appear in Table 1.

For this work, we group the variables into three input sets: (i) measured inputs, pollutant concentrations and contextual variables recorded during the campaign; (ii) regime identification features, engineered variables that reflect emissions, ventilation, and diurnal structure, used to define the regimes; (iii) task predictors, the variables used in each analysis task, such as regression, cross-site testing, and imputation, which may include the regime label when we test its added value. This organization makes it clear what each step uses, and it helps keep test days and test sites separate from model fitting.

Using this organization, we applied a common preprocessing pipeline before any analysis. All channels were synchronized to a fixed 15 min time grid and checked for invalid readings and logging artifacts. Relative humidity was retained and inspected because optical PM can be biased upward during humid periods. Day and site identifiers were kept so that any step that learns from the data, such as scaling, regime identification, and model fitting, was fitted using only the calibration days and sites and then applied unchanged to the test days and sites.

Additionally, instruments were selected to balance robustness, portability, and traceability under field conditions. Particulate matter (PM_1, $PM_{2.5}$, PM_{10}) was monitored with a Casella Dust Detective optical instrument operating on the light-scattering principle, where a laser beam passes through the sampled air and scattered light intensity is measured to determine particle concentration. The device includes an internal pump that maintains a stable flow rate, provides real-time output for fine and coarse fractions, and is well suited for urban air quality studies. Trace gases (NO_2, CO, H_2S, total VOCs) were measured with Aeroqual Series 500 modular heads, and meteorological variables were recorded using a Kestrel 5500 vane and cup anemometer. All sensors were co-located on a tripod at breathing height ≈ 1.5 m, and powered by field batteries. Table 2 summarizes the instruments, principles, and ranges, where PID refers to a Photoionization Detector Head used for total VOCs.

Table 1. Site-wise statistics of pollutants and context variables over four campaign days.

		PM_1 (µg/m^3)		$PM_{2.5}$ (µg/m^3)		PM_{10} (µg/m^3)		NO_2 (ppb)		CO (ppm)		H_2S (ppb)		VOC (ppb)		RH (%)		Wind Speed (m/s)	
Site	**Description**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Site 1	Open air taxi park with high idling and turnover	136.9	12.7	175.1	6.8	195.7	6.7	66.8	7.1	46.3	4.1	59.5	7.6	91.4	26.6	56.9	2.8	0.7	0.1
Site 2	Busy roundabout in a street canyon	55.4	4.8	99.6	5.8	111.6	4.8	59.2	8.5	26.7	5.1	51.5	5.2	56.3	2.2	49.1	4.0	0.9	0.1
Site 3	Multi storey semi-enclosed taxi hub	132.4	15.5	160.6	7.0	175.0	6.8	64.3	7.7	49.8	2.7	68.1	6.4	78.1	6.9	58.4	2.3	0.7	0.2
Site 4	Fuel forecourt with near source emissions	67.4	3.1	88.9	4.0	97.2	3.4	58.3	6.7	52.2	10.7	52.1	5.0	91.4	25.7	56.0	5.4	1.0	0.1
Site 5	University campus entry as urban background	37.7	3.5	88.4	3.6	93.2	4.0	53.0	36.9	22.2	2.2	55.1	5.6	85.6	24.5	55.0	4.5	1.2	0.6

Table 2. Instruments, principles, and ranges.

Variable	**Instrument and Model**	**Measurement Principle**	**Range and Resolution**
PM_1, $PM_{2.5}$, PM_{10}	Casella Dust Detective (Casella CEL Ltd., Bedford, UK)	Optical light scattering (photometric)	0–150 mg m^{-3}, 1 µg/m^3
NO_2	Aeroqual Series 500 (NO_2 head; Aeroqual Ltd., Auckland, New Zealand)	Electrochemical	0–1 ppm, 0.001 ppm
CO	Aeroqual Series 500 (CO head)	Electrochemical	0–100 ppm, 0.1 ppm
H_2S	Aeroqual Series 500 (H_2S head)	Electrochemical	0–5 ppm, 0.001 ppm
Total VOCs	Aeroqual Series 500 (PID head)	Photoionization detector	0–30 ppm, 0.01 ppm
Relative humidity, wind speed	Kestrel 5500 Weather Meter (Nielsen-Kellerman Co., Boothwyn, PA, USA)	Capacitive humidity and cup anemometer	10–90%, 0.1%, 0–20 m s^{-1}, 0.1 m s^{-1}

Moreover, quality assurance and quality control (QA/QC) procedures were applied to ensure precision, accuracy, and full traceability. Prior to deployment, the Casella Dust Detective and Aeroqual analyzers were calibrated using certified manufacturer standards, with calibration certificates archived. Daily zero and span checks were performed in the field using High-Efficiency Particulate Air (HEPA)-filtered air for the particulate instrument and certified calibration gases for the gas analyzers. Instrument response and flow-rate stability were verified within 2% of nominal values before each sampling session. Device clocks were synchronized to local time within 1 min and logged at fixed 15 min intervals aligned to wall time. Each day's measurement sequence included start- and end-of-day bump checks, and any deviation beyond acceptance limits triggered corrective action and notation in the field log. Sampling conditions (duration, location, temperature, RH, and airflow) were recorded concurrently and cross-checked with calibration records. Data validation involved cross-comparison between the Casella Dust Detective and a secondary colocated handheld monitor (HAZ-Scanner) used as a consistency check, inspection for outliers, and exclusion of points attributable to instrument fault or procedure anomalies. Optical PM data were screened with a relative humidity sentinel rule to limit hygroscopic bias. Intervals with RH $\geq$ 85% were flagged and removed prior to summary analyses [16]. All serial numbers, maintenance logs, and calibration records were archived for traceability.

A simple site-level calibration protocol defined threshold criteria to maintain consistency across all locations. Zero and span verifications for each pollutant analyzer and flow-rate stability within $\pm$2% were confirmed before and after daily measurements. This standardized procedure ensured that observed spatial differences reflected genuine environmental variability rather than calibration drift or bias.

The campaign emphasizes hours that dominate activity and human presence around the sites, and it spans contrasting micro-environments within a compact window. This supports the discovery of recurring daytime conditions and provides a shared context across locations [9]. Hyperlocal studies that mapped street-scale variability concentrated their measurements in active daytime periods and showed consistent spatial patterns across repeats. This aligns with our focus on practical regime identification under real constraints [12,13,17,18]. Because temporal coverage is short, we explicitly evaluate the robustness of the learned regimes using block-resampling stability and sensitivity checks as seen in Section 2.4. Overall, the combined design delivers the spatial contrasts needed for regime learning while documenting stability of the resulting structure.

2.2. Physically Grounded Feature Engineering

Regime learning is driven by a feature vector that combines pollutant levels with simple indicators of timing and mixing. The goal is to represent recurring emission and ventilation conditions in a way that remains interpretable and comparable across sites.

Short roadside campaigns show strong diurnal structure driven by activity timing, mixed-layer growth, and ventilation in semi-enclosed spaces. To represent these rhythms smoothly without arbitrary cutoffs, we encode local time of day $t \in [0, 24)$ with two circular terms

$$h_{\sin} = \sin\left(\frac{2\pi t}{24}\right), \qquad h_{\cos} = \cos\left(\frac{2\pi t}{24}\right). \tag{1}$$

We include two composition ratios that reflect shifts in the particle-size mixture linked to emission and processing pathways

$$R_{\frac{1}{10}} = \frac{\mathrm{PM}_1}{\mathrm{PM}_{10}}, \qquad R_{\frac{2.5}{10}} = \frac{\mathrm{PM}_{2.5}}{\mathrm{PM}_{10}}. \tag{2}$$

The ratios increase when the fine mode dominates (e.g., fresh exhaust or confined microenvironments) and decrease when coarse mechanisms (e.g., resuspension or road dust) dominate. These ratios provide a compact unitless indicator of shifts in size mix that is comparable across sites in short campaigns.

We also include a simple daytime mixing proxy based on the sine of the solar-elevation angle α, used here as a practical indicator of convective mixing strength. The angle is computed from site latitude ϕ, solar declination δ from day of year, and hour angle H using the NREL Solar Position Algorithm and then truncated at zero

$$\sin \alpha \;=\; \sin \phi \, \sin \delta \;+\; \cos \phi \, \cos \delta \, \cos H. \tag{3}$$

The truncation is appropriate because the window is daytime only; negative solar elevation corresponds to night, when this proxy would not represent convective mixing. More details about this algorithm are found in [19].

2.3. Learning Recurring Conditions as Regimes

In this work, clustering is used as a regime-construction step. The objective is to obtain a small set of stable and interpretable operational prototypes, "centroids," that can be summarized physically and used consistently in subsequent regime-conditioned modeling and transport checks across sites.

Operational regimes summarize recurring multivariate conditions that matter for exposure and for simple forecasting. Let $\mathbf{x}_i \in \mathbb{R}^p$ denote the feature vector with pollutant levels, basic meteorology, the diurnal terms, and the composition ratios.

To place heterogeneous variables on a comparable geometry and to avoid information leakage, each column j is standardized using training only moments $(\mu_j, \sigma_j > 0)$

$$x_{ij}^{\text{std}} \;=\; \frac{x_{ij} - \mu_j}{\sigma_j}. \tag{4}$$

k-means is used in this study because the regimes are intended to act as operational prototypes: each record must receive a single, reproducible label, and the same regime definition must be assignable to unseen days and unseen sites using training information only. In this setting, centroid-based regimes provide (i) direct regime-wise profiles via cluster centroids/medians and (ii) a deterministic out-of-sample rule (nearest-centroid assignment) that fits naturally with the day- and site-blocked evaluation design. Other clustering families such as model-based mixtures or density-based methods are useful when the scientific goal is to represent non-spherical structure or variable-density groups. However, for short campaigns and for the specific downstream tasks studied here (regime recognition from reduced sensors, regression with a regime factor, and fold-safe reconstruction), the key requirement is a stable partition with complete assignment and a transparent out-of-sample mapping under leakage control. We therefore use k-means as a controlled, stability-documented baseline and quantify robustness via block-resampling stability, perturbation checks, and feature-group ablations (Section 2.4), alongside alternative partitions in Appendix A.3 [20–27].

Furthermore, we learn k regimes with k-means by minimizing within cluster distortion using the Hartigan and Wong algorithm with multiple random starts [22]. The number of regimes k is chosen from a small grid by maximizing mean silhouette and then confirmed by plateaus in adjusted Rand index under day block bootstrap resampling [20,26]. This stability-aware selection reduces sensitivity to unequal densities and supports reproducibility. Let $\{\mathbf{c}_j\}_{j=1}^{k}$ denote the centroids of the k clusters in the standardized feature space, where each $\mathbf{c}_j$ is a p-dimensional mean vector representing the typical conditions

of regime j. For new observations $\tilde{\mathbf{x}}$ we assign regimes by nearest centroid in the training standardized space

$$\hat{r}(\tilde{\mathbf{x}}) \;=\; \arg\min_{1\le j\le k} \left\| \frac{\tilde{\mathbf{x}} - \boldsymbol{\mu}_{\mathrm{tr}}}{\boldsymbol{\sigma}_{\mathrm{tr}}} \;-\; \mathbf{c}_j \right\|_2, \tag{5}$$

with $(\boldsymbol{\mu}_{\mathrm{tr}}, \boldsymbol{\sigma}_{\mathrm{tr}})$ and $\{\mathbf{c}_j\}$ learned on training data only, and $\|.\|_2$ is the Euclidean norm. Alternative views that included Ward linkage on Euclidean distances, k-means in principal component analysis (PCA) space, and adjusted Rand index (ARI) agreement checks were explored. As reported in the Appendix A, candidate k was scanned over a small grid using average silhouette, and stability was evaluated by day block bootstrap and ARI [27–29].

2.4. Stability, Sensitivity, and Interpretability

To evaluate whether the discovered regimes are robust under a short duration and limited sample size, we assessed stability and sensitivity using blocked resampling and controlled input perturbations. Stability was assessed by resampling entire site–day blocks with replacement, refitting k-means with the selected k, and summarizing agreement with the reference partition using the adjusted Rand index (ARI) [20,27]. Sensitivity was assessed in two complementary ways. First, we applied small multiplicative perturbations (jitter) to VOC and RH and refit the clustering, following stability-based clustering validation practice [30]. Second, we performed feature-group ablations by refitting the clustering after removing one feature group at a time (diurnal terms, PM ratios, VOC, or RH) and quantifying partition agreement using ARI. These checks are used to confirm that the main regime structure is not driven by a single channel and that it persists under small changes to the input representation; they are not used to claim causal importance.

For each regime, pollutant and meteorological distributions were summarized using the median and interquartile range, and hourly regime occurrence was tabulated over the daytime window to characterize diurnal representation. Robust summaries, such as median and interquartile range (IQR), are used because short roadside campaigns often produce right-skewed distributions with episodic peaks.

2.5. Imputing One Missing Sensor Stream

One site–day record of H_2S was unavailable in the collected dataset. Device logs indicated a transient sensor fault, while all other channels remained stable. We treated this gap as missing at random conditional on observed covariates and reconstructed the missing values using a regime-aware random forest [31,32]. Random forest was selected because it captures nonlinear relationships among co-pollutants and meteorology, and it is also robust to multicollinearity. The operational regime label was included as an auxiliary predictor to encourage coherence with the multivariate pollutant–meteorology state, linking the regime framework to a practical sensor-recovery task. This imputation example is presented as a proof of concept for feasibility in short campaigns; generalization to longer gaps or substantially different source mechanisms is discussed in Section 4.

The input features used as predictors for the imputation task included site/location covariates (site name, latitude, longitude), time-of-day features (hour and sin/cos encoding), meteorology/proxies (wind speed, relative humidity, and a mixing proxy), a traffic proxy (car-parking), co-pollutants (PM_1, $PM_{2.5}$, PM_{10}, NO_2, CO, total VOCs), and the regime label. This use of co-pollutants as predictors follows the general idea that co-occurring air-pollution variables can carry strong predictive information, which has been shown in multi-output VOC prediction settings [33]. Regimes were learned on the training data only; the resulting cluster ID $(1, \ldots, k)$ was then assigned to held-out and new records (including the missing day) using the trained centroids and included in the random forest model as a categorical factor. Random forest models were fit without hyperparameter tuning,

keeping the default number of candidate predictors considered at each split, because the goal was a robust reconstruction baseline for a short campaign rather than optimizing performance. We used ensembles of 600 trees in cross-validation and 1000 trees for the final imputation fit to stabilize the ensemble predictions [31,32]. Because H_2S is unobserved on the missing day, performance cannot be computed for that day. We therefore evaluated the reconstruction model on non-missing days using day-blocked splits, then refit the final model on all available non-missing days and applied it to the missing site–day. This avoids using information from the target day during fitting.

Predictive dispersion across trees was used as an exploratory proxy for variability in the imputed estimates. We report the spread of predictions across the random forest ensemble only as a heuristic indicator of prediction stability, not a statistically calibrated predictive uncertainty interval.

2.6. Do Regimes Add Explanatory Value Beyond Basic Covariates?

The analysis tests whether the learned regimes capture exposure-relevant multivariate conditions that are not already explained by common covariates used in air-quality interpretation (wind speed, solar mixing proxy, diurnal timing, and traffic activity). Regime membership represents recurring states in which pollutants and meteorology co-vary (e.g., traffic build-up under weak ventilation versus well-mixed periods). We quantify the added explanatory value of regimes by comparing a baseline model to an augmented model that includes the regime label.

The baseline linear model for pollutant Y_i at time i is

$$Y_i = \beta_0 + \beta_1 \mathrm{WS}_i + \beta_2 \mathrm{SE}_i + \beta_3 \sin\left(\frac{2\pi t_i}{24}\right) + \beta_4 \cos\left(\frac{2\pi t_i}{24}\right) + \beta_5 \mathrm{Cars}_i + \varepsilon_i, \quad (6)$$

where wind speed is denoted WS, the truncated solar-elevation proxy is SE, the cyclic hour terms describe diurnal timing, and the vehicle activity indicator is Cars, which is recorded by the IR traffic counter (vehicles per 15 min) at all sites.

Let $R_i \in \{1, \ldots, k\}$ denote the regime label at time i, with regime 1 as the reference level. The augmented model adds the regime label as a categorical fixed effect to estimate the mean pollutant shift associated with each regime:

$$Y_i = \beta_0 + \beta_1 \mathrm{WS}_i + \beta_2 \mathrm{SE}_i + \beta_3 \sin\left(\frac{2\pi t_i}{24}\right) + \beta_4 \cos\left(\frac{2\pi t_i}{24}\right) + \beta_5 \mathrm{Cars}_i + \sum_{r=2}^{k} \gamma_r \mathbb{1}\{R_i = r\} + \varepsilon_i. \quad (7)$$

Treating regime as a fixed effect allows a direct estimation of the average pollutant difference across the identified multivariate conditions and enables formal testing of whether regime membership adds independent explanatory value. Models were estimated by ordinary least squares [34].

Residuals were inspected (residuals versus fitted and Q–Q plots). For skewed targets, we repeated the analysis on the log scale as a sensitivity check, as presented in Appendix A.4; results are reported on the natural scale.

2.7. Generalization Across Sites and Leakage Control

To evaluate whether models trained in some locations can predict pollutant levels in a new but comparable micro-environment(site), we assessed cross-site generalization using a Leave-One-Site-Out (LOSO) design and trained identical models with and without the regime factor. This reflects a common operational setting where monitoring exists at only a subset of locations.

In each LOSO round, one site served as the test set and models were trained on the remaining sites. Any step that learns from the data distribution (standardization and regime identification) was fit on the training sites only and applied unchanged to the held-out site to prevent information leakage. Predictive performance was summarized by R^2 and RMSE, and we compared models with and without the regime factor.

Within-site analyses used day-blocked cross-validation [35], where entire days were held out together to respect temporal dependence and avoid optimistic splits within the same day. Together, LOSO and day-blocked validation address (i) generalization to a new day at the same site and (ii) generalization to a new site.

These validation schemes provide a rigorous test of robustness under strict leakage control [36] and support the use of the framework for short multivariate campaigns in operational and regulatory contexts.

3. Results

3.1. Operational Regimes: Selection, Profiles, and Stability

A grid search over the number of clusters k gave the highest mean silhouette 0.379 at $k = 5$. Stability diagnostics show high ARI stability k = 3–5 with a plateau at k = 4–5, and a clear deterioration for $k \geq 6$ (Appendix A.1, Table A1). Considering these values supports selecting $k = 5$ as a stable and well-separated solution.

Per-regime pollutant profiles (median and IQR) are summarized in Table 3 and visualized in Figure 2. The regimes mainly differ by (i) particulate loading versus ventilation (PM size fractions: PM_1, $PM_{2.5}$, PM_{10} versus wind speed) and (ii) near-source gaseous signatures (CO and VOC). Resampling whole site–days confirms high stability of the partition (ARI mean 0.945, Table A1). Alternative partitions, like Ward's linkage and PCA-space k-means, recover a comparable structure as presented in Table A3 and Figure A1 in Appendix A.3, indicating that the main regime patterns are not specific to one clustering choice.

With the regime labels fixed at $k = 5$, Table 3 and Figure 2 provide the five reference profiles used in the remaining analyses. In practical terms, Regime 1 is characterized by elevated CO/VOC (near-source influence). Regime 2 captures the highest PM levels under the weakest winds and Regime 3 remains PM-elevated but with slightly stronger mixing. Regimes 4–5 represent cleaner and/or better-ventilated states; in particular, Regime 5 shows the lowest PM and CO under the strongest winds (background/ventilated conditions), while VOC may remain elevated.

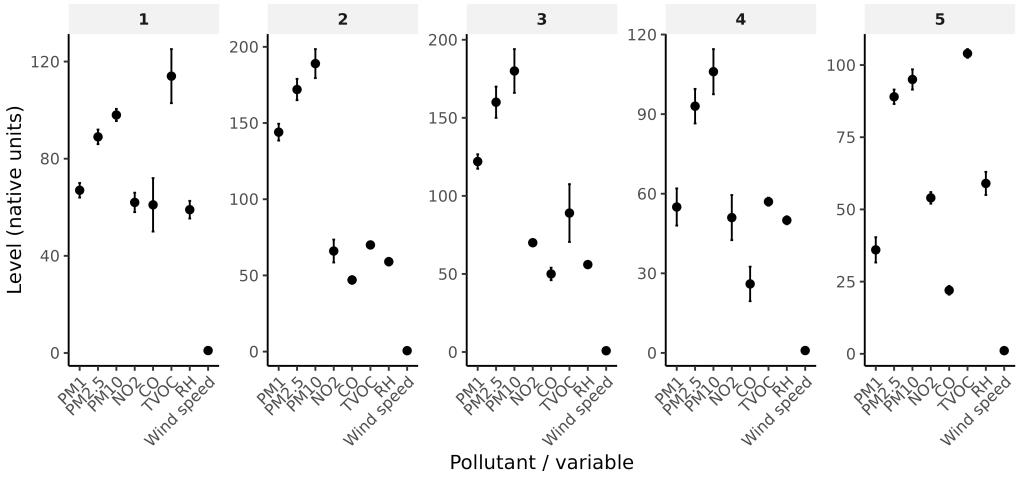

Figure 2. Regime-wise pollutant and meteorology profiles (Regimes 1–5 correspond to the five reference regimes from the $k = 5$ clustering), where points represent central tendency and error bars show spread.

Table 3. Regime-wise pollutant profiles across all sites and daytime campaign hours (median, IQR).

	PM_1 (µg/m^3)		$PM_{2.5}$ (µg/m^3)		PM_{10} (µg/m^3)		NO_2 (ppb)		CO (ppm)		VOC (ppb)		RH (%)		Wind Speed (m/s)	
Regime	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**
1	67	6	89	6	98	5	62	8	61	22	114	22.25	59	7.25	1	0.2
2	144	11	172	14	189	19	66	15	47	5	70	3	59	5	0.6	0.14
3	122	9.25	160	20	180	28	70	4	50	8	89	37	56	1	0.8	0.06
4	55	14	93	13	106	17	51	17	26	13	57	3	50	3	0.9	0.2
5	36	8.75	89	5	95	7	54	4	22	3	104	3	59	8	1.1	0.1

3.2. Recognition from Partial Sensor Sets

Table 4 shows that regime labels are highly classifiable under day-blocked cross-validation (i.e., all samples from the same day are kept in the same fold), where accuracy is 0.928 using PM only, 0.989 using gases and meteorology, and 0.993 using all variables. This indicates that a lean gases and meteorology suite can recover the learned regimes with near-perfect fidelity. Indeed, this task does not re-learn regimes; it quantifies how well a reduced sensor set can reproduce the fixed regime labels learned from the full feature representation under day-block testing.

Table 4. Day-block cross-validated accuracy for regime classification by sensor set.

Set	Accuracy
PM only	0.928
Gases+Met	0.989
All (mix variables)	0.993

Moreover, perturbation ("jitter") checks indicate that labels are insensitive to small VOC/RH noise (Appendix A.2, Table A2). These results summarize the day-blocked regime-recognition performance for the three sensor sets considered.

3.3. Diurnal Distribution of Regimes

Figure 3 presents, for each daytime hour, the proportion of all concurrent observations from the five monitored micro-environments that were assigned to each pollution regime. Because all sites were sampled simultaneously, each stacked segment represents the distribution of regimes across the entire network at a given hour rather than a temporal evolution within any single site. The figure therefore illustrates how the prevalence of each pollution regime changes through the day when combining all sites and campaign days.

The results reveal clear transitions between contrasting conditions. Early-morning hours (08:00–10:00) are shared mainly between Regimes 2 and 4, reflecting the coexistence of stagnant, high-PM conditions typical of garage-like sites (Regime 2) and cleaner, better-ventilated settings (Regime 4). From around 10:00 onward, Regime 5, the cleanest and most ventilated state, emerges and remains steady at roughly one-quarter of observations. Around midday, Regime 1 becomes more frequent, while after 14:00 the mix shifts toward Regime 3, which shows elevated PM and NO_2 but lower humidity, indicating stronger dispersion yet persistent local emissions. Overall, no single regime dominates the daytime period: cleaner (4–5) and more polluted (2–3) states coexist at the same hours as each site responds to its own micro-environmental drivers such as ventilation, traffic activity, and enclosure.

Figure 4 presents the same information as a compact hour-by-regime matrix. Since each hour contains five simultaneous site observations, a regime share near to 40% corresponds roughly to two sites out of five at that hour.

To make the site–regime links explicit, Table 5 summarizes the regimes together with their dominant site types and typical active hours. In line with Figure 5, Regimes 2–3 are driven mainly by the two garage-like sites (Sites 1 and 3), Regime 1 is largely associated with the forecourt site (Site 4), Regime 4 reflects the more ventilated outdoor settings (often including Site 2), and Regime 5 is almost exclusively associated with the campus/background site (Site 5).

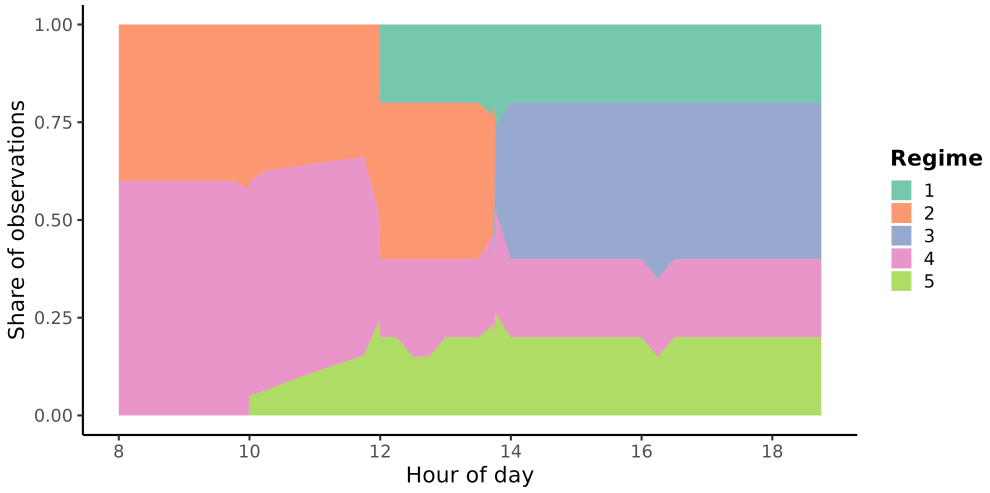

Figure 3. Hourly shares of regimes across all sites and days (daytime 08:00–18:45).

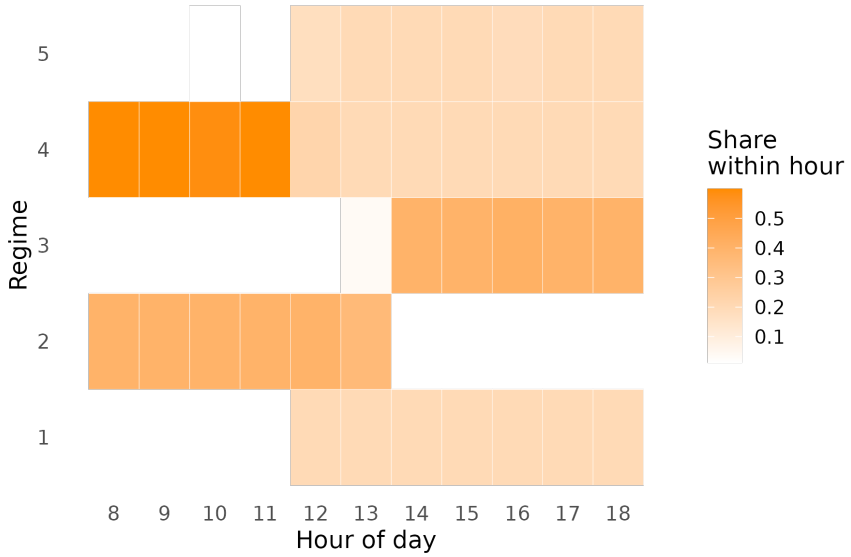

Figure 4. Fraction of all observations at each hour that belong to each regime.

Table 5. Link between learned regimes, site types, and dominant daytime hours.

Regime	Description (From Profiles)	Main Contributing Sites	Dominant Hours
1	Forecourt/near-source, CO and VOC elevated, moderate PM	Site 4: fuel forecourt	12:00–18:45
2	High PM (1/2.5/10), weak wind, traffic/garage conditions	Site 1: open taxi park, Site 3: semi-enclosed taxi hub	08:00–12:45
3	High PM, later-day version (more mixing, similar gases)	Sites 1 and 3: same garages as Regime 2	14:00–18:45
4	Outdoor/street-canyon/mixed ventilation	Site 2: roundabout, Site 4, Site 5: background	08:00, 09:00, 11:00
5	Cleanest/windiest/background	Site 5: campus entry	13:00–15:00, 17:00–18:45

Figure 5 shows the fractional contribution of each site to each regime. It confirms that Regimes 1 and 5 are almost entirely associated with single-site micro-environments (forecourt and campus, respectively), while Regimes 2 and 3 are jointly shared between the two garage-like sites. Together, Figures 3–5 show that different micro-environments can occupy different regimes at the same hour, so the observed diurnal pattern is driven by spatial heterogeneity and local forcing rather than a single synchronized daily cycle.

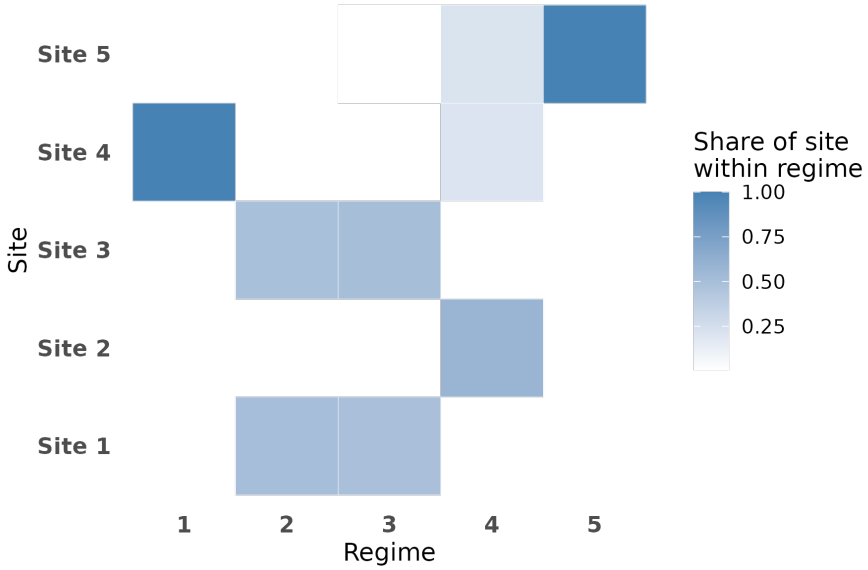

Figure 5. Share of each site within each regime. Darker cells indicate higher contribution of that site to the regime.

3.4. Effect of Regimes on NO_2 Model Performance

A baseline OLS model for NO_2 (wind speed, solar-elevation proxy, cyclic hour, vehicle activity) achieved $R^2 = 0.194$. Adding the five-level regime factor increased R^2 to 0.251 ($\Delta R^2 = 0.057$, ANOVA $p = 4.78 \times 10^{-13}$; Figure 6). In other words, the regime label explains an additional 5.7 percentage points of variance in hour-to-hour NO_2 beyond wind, diurnal timing, solar mixing proxy, and activity, using only one categorical descriptor of the multipollutant state. Practically, regimes provide a compact label for recurring "states" (like build-up under weaker ventilation versus better-mixed periods), which helps summarize and compare exposure conditions across hours and micro-environments. This indicates that regime membership captures joint pollutant–meteorology structure that is not fully represented by the individual covariates. The gain is modest, as expected in short near-road campaigns with unmeasured drivers, but it is consistent and obtained with only one additional categorical factor, improving interpretability without increasing model complexity. Residual diagnostics and a natural-vs.-log robustness check are provided in Appendix A.4 (Figure A2 and Table A4).

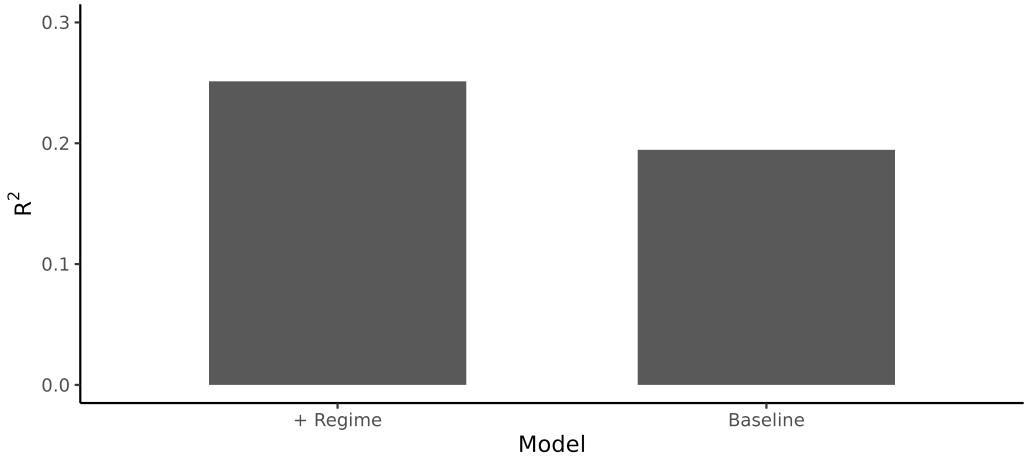

Figure 6. Incremental explanatory power of the regime factor for NO_2 hourly variability.

Figure 6 reports the NO_2 model-fit change associated with including the regime factor. These results support that incorporating regime information enhances the explanatory capacity of pollutant models by capturing complex mixture-dispersion patterns. Building on this added interpretive value, next, we assess the practical utility of the framework in reconstructing a fully missing pollutant channel through a fold-safe, regime-aware imputation approach.

3.5. Fold-Safe H_2S Reconstruction and Predictive Uncertainty

The regime-aware random forest achieved $R^2 = 0.972$ and $RMSE = 1.467$ under day-blocked cross-validation ($n = 836$). Applied to the single missing site–day, the model produced 44 imputed values (Figure 7) with low predictive dispersion (mean predictive standard deviation ≈ 0.067), supporting stable reconstruction for this missing-channel case.

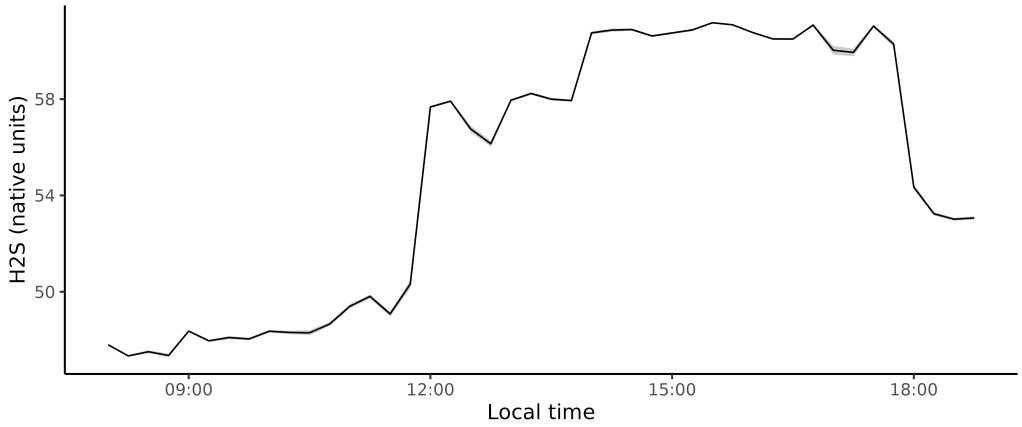

Figure 7. Reconstruction of a fully missing H_2S site–day using a fold-safe, regime-aware random forest.

3.6. Cross-Site Generalization

Leave-one-site-out (LOSO) results in Table 6 show the strongest portability for NO_2 (mean site-wise $R^2 = 0.402$). By contrast, PM fractions, CO, VOC, RH, and H_2S often show weak or negative R^2, reflecting strong local heterogeneity where negative R^2 values indicate a performance that is worse than predicting the mean concentration of the held-out observations (the R^2 baseline), so it provides a direct signal that transportability is limited in that setting [37]. Several LOSO R^2 values are strongly negative for some targets/sites. This can occur when the held-out site has a different mean/scale or a smaller variance than the pooled training sites: even moderate absolute errors can exceed the test-site variance, producing large negative R^2 despite RMSE values that remain interpretable on the original scale. For this reason, we report both R^2 and RMSE and interpret strongly negative R^2 primarily as clear evidence of limited transportability under cross-site extrapolation in this campaign. Paired LOSO runs with and without the regime factor are summarized in Appendix A.5 (Table A5).

Table 6. Leave-one-site-out performance by site and target pollutant.

Method	Site	Target	R^2	RMSE	n_{test}
RF	1	CO	−11.411	14.256	176
RF	2	CO	−3.29	10.491	176
RF	3	CO	−2.522	4.971	176
RF	4	CO	−0.21	11.751	176

Table 6. *Cont.*

Method	Site	Target	R^2	RMSE	n_{test}
RF	5	CO	−44.03	14.444	176
RF	1	H_2S	0.521	5.277	176
RF	2	H_2S	0.101	4.961	176
RF	3	H_2S	−3.513	13.643	176
RF	4	H_2S	−6.247	13.349	176
RF	5	H_2S	0.491	4.010	132
RF	1	NO_2	0.554	4.745	176
RF	2	NO_2	0.377	6.664	176
RF	3	NO_2	0.546	5.152	176
RF	4	NO_2	0.489	4.749	176
RF	5	NO_2	0.046	35.935	176
RF	1	PM_1	−12.723	46.810	176
RF	2	PM_1	−30.676	26.685	176
RF	3	PM_1	−4.823	37.327	176
RF	4	PM_1	−149.927	37.676	176
RF	5	PM_1	−69.908	29.759	176
RF	1	PM_{10}	−47.683	46.428	176
RF	2	PM_{10}	−28.854	26.141	176
RF	3	PM_{10}	−15.883	28.065	176
RF	4	PM_{10}	−262.094	55.263	176
RF	5	PM_{10}	−30.192	22.404	176
RF	1	$PM_{2.5}$	−32.934	39.381	176
RF	2	$PM_{2.5}$	−18.134	25.379	176
RF	3	$PM_{2.5}$	−20.048	31.893	176
RF	4	$PM_{2.5}$	−182.216	53.315	176
RF	5	$PM_{2.5}$	−17.864	15.751	176
RF	1	RH	0.236	2.405	176
RF	2	RH	−1.157	5.919	176
RF	3	RH	−1.327	3.459	176
RF	4	RH	0.327	4.402	176
RF	5	RH	−1.298	6.864	176
RF	1	VOC	−0.612	33.718	176
RF	2	VOC	−189.356	30.241	176
RF	3	VOC	−12.712	25.411	176
RF	4	VOC	0.599	16.259	176
RF	5	VOC	−0.251	27.322	176

A key implication is that models trained on pooled data may generalize for NO_2 (a more spatially transferable signal in this campaign), while several other channels remain highly site-specific and benefit more from local calibration.

4. Discussion

This section interprets the results in light of the study's three main objectives: (i) learning operational regimes that summarize daytime conditions across sites, (ii) testing whether these regimes add explanatory value beyond basic covariates, and (iii) assessing cautious model transport across comparable micro-environments. In practical terms, we focus on a short-campaign problem: how to summarize concurrent multi-pollutant measurements across different micro-environments in a way that remains usable for modeling and data-quality tasks.

The operational regimes derived in this study summarize the joint pollutant–meteorology structure observed across the five monitored micro-environments during daytime hours. Compared with conventional segmentation (by site, by time slot, or by

single-pollutant thresholds), the regime labels are assigned from multiple variables jointly, so they can represent recurring multi-pollutant states that appear across different sites and hours within the same campaign. At the same time, the results also show site-dominant regimes (e.g., regimes largely associated with one site type), which helps separate site-specific conditions from those shared across similar micro-environments. Because the regime factor remained interpretable and was highly classifiable from reduced sensor subsets, it provides a compact way to report multivariate conditions using a small number of labels.

In a short campaign, a practitioner can use the regime labels to (i) summarize exposure conditions by reporting the fraction of time each micro-environment spends in higher-burden regimes (by hour and by site), (ii) compare locations on a common basis beyond simple averages (e.g., whether a site is dominated by accumulation-type regimes or by well-mixed regimes), and (iii) support operational actions such as prioritizing ventilation or traffic-management measures during the hours when higher-burden regimes occur most frequently. The same labels can also support data-quality workflows by flagging records that are inconsistent with the typical regime profile and by providing context for reconstructing short sensor outages.

We used *k*-means because the regime labels are intended to function as operational prototypes: they must be (i) easy to summarize via regime-wise profiles and (ii) assignable to new observations using only training information (nearest-centroid assignment) under the day- and site-blocked evaluation design. This supports a clear separation between learning and testing and keeps the regime definition reproducible across resamples. Model-based mixtures and density-based clustering can be valuable when regimes are expected to be strongly non-spherical, overlapping, or multi-density; however, in short campaigns they can require additional tuning choices that affect reproducibility and the stability of downstream comparisons. Further, methods that are more sensitive to model specification or density/hierarchy hyperparameters may change the effective number and definition of clusters across folds or environments, which complicates the same-regime interpretation required by our framework. Additionally, because our regime label is used as a factor/predictor and as a target in regime-recognition, we require a complete partition where every record is assigned a regime, and we require a deterministic out-of-sample assignment rule under site/day hold-out. We therefore focus on prototype-based regimes via k-means and outline a leakage-safe comparison protocol for alternative clustering families in Appendix A.3. Accordingly, this study emphasizes a controlled, stability-documented baseline that can be extended using the same leakage-safe validation principles when richer datasets or additional context variables are available [20,24,25,38].

In the NO_2 regression, adding the five-level regime factor increased R^2 from 0.194 to 0.251 ($\Delta R^2 = 0.057$). This change is reported as an incremental increase in explained variability for a model that remains parsimonious (one additional categorical factor). The result is consistent with near-road work highlighting the role of local dispersion and mixing conditions in shaping concentration variability [2,8]. We report the effect as a model-level change in explained variance and do not translate it into a ppb-scale improvement without additional concentration-error summaries.

Moreover, the regime-aware random forest imputation reconstructed a completely missing H_2S site–day with strong cross-validated performance and low predictive dispersion, using a fold-safe design in which regimes are learned in training folds and assigned in held-out folds. In this manuscript, the imputation case study is limited to one pollutant, one site, and one missing day, so it is presented as a proof-of-concept for methodological feasibility under short-campaign constraints rather than a general statement about performance for longer missing periods or for sites with different source mechanisms.

Concerning cross-site transferability and practical boundaries, leave-one-site-out testing revealed substantial heterogeneity in pollutant behavior. NO_2 showed the strongest portability among the tested targets in this dataset, while several other channels showed weak or negative R^2. Here, a negative R^2 indicates performance worse than a site-mean baseline for the held-out site, which provides a clear marker of limited transportability. Overall, these results highlight that some targets support pooled modeling across sites in this campaign, whereas others remain strongly site-dependent and would require additional contextual descriptors (e.g., geometry, land use, or source mix) and/or site-specific calibration for deployment beyond the training micro-environments.

The present campaign focused on daytime hours over a small number of representative sites, so the learned regimes are specific to the observed daytime conditions and may not represent nocturnal chemistry, other seasons, or more complex urban morphologies. The solar-elevation proxy is a simplified representation of boundary-layer dynamics and does not replace direct PBL or mixing-height measurements. Although portable analyzers were factory-calibrated, instrument bias and channel noise remain possible, particularly for VOC and H_2S sensors. Future work should extend regime learning across longer timeframes and incorporate physical and contextual covariates such as mixing height, canopy geometry, and land-use class to strengthen interpretation and transport across dissimilar micro-environments. The same stability and leakage-safe validation design used here is compatible with other clustering families when the data support additional complexity. Combining machine learning with mechanistic dispersion modeling is also a promising direction for improving transport across dissimilar micro-environments.

5. Conclusions

This study develops and tests an operational-regime workflow for short, synchronized multi-site monitoring in traffic-related urban micro-environments (open and semi-enclosed garages, a fuel forecourt, a street setting, and a campus/background location). The regimes are learned from the joint pollutant–meteorology space and therefore provide an alternative to conventional segmentation by site, hour, or single-pollutant thresholds. A five-regime solution was selected by the silhouette criterion and supported by day-block stability diagnostics and sensitivity checks (Appendices A.1 and A.3), and the regime profiles summarize consistent contrasts in PM loading, gaseous signatures, and ventilation conditions across the monitored network.

Regime labels were recoverable from reduced sensor configurations under day-blocked validation, with accuracy 0.989 using gases and meteorology and 0.993 using all variables, supporting regime recognition when only a subset of channels is available. In the NO_2 regression, adding the five-level regime factor increased R^2 from 0.194 to 0.251 ($\Delta R^2 = 0.057$), providing a quantified gain in explained variability within the same parsimonious model form. For missing data reconstruction, the fold-safe, regime-aware random forest approach reproduced a fully missing H_2S site–day with strong cross-validated performance in this proof-of-concept case, while explicitly avoiding leakage by learning regimes within training folds.

In leave-one-site-out tests, NO_2 showed the strongest cross-site portability in this dataset, whereas several other channels were strongly site-dependent and included negative R^2 values, indicating performance worse than a site-mean baseline. These results summarize both what transfers across the monitored micro-environments in this campaign and what remains local, and they define the practical operating range of the proposed workflow for short field deployments.

Author Contributions: Conceptualization, A.E., S.J., A.C. and E.R. ; Methodology, A.E. and R.E.; Formal analysis, A.E. and R.E.; Data curation, S.J. and G.H.; Investigation, S.J., G.H., A.C. and E.R.; Writing—review and editing, A.E., R.E., S.J., G.H., A.C. and E.R. Funding acquisition, A.C. All authors have read and agreed to the published version of the manuscript.

Funding: The Al Maqdesi project (Chargée de mission universitaire et scientifique, Consulat Général de France à Jérusalem-SCAC) and the Palestinian Ministry of Higher Education (Grant No. 01/2023) provided financial support for this work.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Acknowledgments: The authors gratefully acknowledge financial support from the PHC Al-Maqdissi program and the Palestinian Ministry of Higher Education. We also thank An-Najah National University (Nablus, Palestine), and the University of Reims Champagne-Ardenne (Reims, France) for their valuable logistical and technical assistance.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Appendix A.1. k Selection: Silhouette and ARI Stability

For each $k \in \{3, \ldots, 7\}$ we fitted k-means (Hartigan–Wong) on standardized features, then performed day-block bootstrap resampling ($B = 80$) and computed the Adjusted Rand Index (ARI) between the reference labels and each bootstrap refit.

Silhouette refers to the average silhouette width (computed on the same standardized feature space) and is used as an internal separation/compactness diagnostic, while ARI stability provides an external reproducibility check under resampling.

We selected k using a joint criterion that balances (i) cluster separation (mean silhouette) [26], (ii) label stability under day-block bootstrap (Adjusted Rand Index, ARI) [27], and (iii) interpretability/operational usefulness of the resulting regimes. The silhouette score was highest at $k = 5$, indicating the strongest separation at this resolution. The ARI distribution showed a high-stability plateau for $k = 3$–5, with a clear deterioration for larger k, suggesting that further splitting produces less reproducible partitions. We therefore chose $k = 5$ as the smallest value that (a) maximizes separation while (b) remaining on the stability plateau. In contrast, smaller k values tend to merge qualitatively distinct pollutant–meteorology states observed in the data, reducing interpretability and weakening downstream analyses that rely on regime-specific behavior. This multi-criterion selection supports $k = 5$ as a practical balance between separation, stability, and interpretability.

Table A1. Stability across k via day-block bootstrap (ARI distribution).

k	Mean	Median	P_5	P_{95}
3	0.987	1.000	0.926	1.000
4	0.966	0.997	0.808	1.000
5	0.945	0.997	0.796	1.000
6	0.852	0.847	0.655	1.000
7	0.788	0.801	0.682	0.871

Appendix A.2. Sensitivity to Small Multiplicative Perturbations in VOC and RH

We applied log-normal multiplicative noise to VOC and RH (sd = 0.05) and recomputed the partition $B = 100$ times; agreement with the reference labels was summarized by ARI.

Table A2. Partition sensitivity to VOC/RH "jitter".

	Mean	Median	P_5	P_{95}
ARI	0.991	0.991	0.982	0.997

The partition is essentially unchanged by small perturbations, indicating robustness to minor sensor-level variation.

Appendix A.3. Alternative Partitions: Ward.D2 and PCA-Space k-Means

We compared the reference solution presented in Section 2 to Ward.D2 hierarchical clustering using Euclidean distance, and *k*-means applied in principal-component space retaining $\geq 90\%$ variance. As seen in Table A3, agreement with the reference labels was measured by ARI.

Table A3. Agreement (ARI) between alternative partitions and the reference ($k = 5$).

Alternative Partition	ARI vs. Reference
Ward.D2 (Euclidean)	0.86
k-means in PCA space (90% var.)	0.90

From Figure A1 below, profiles align closely, supporting algorithmic robustness.

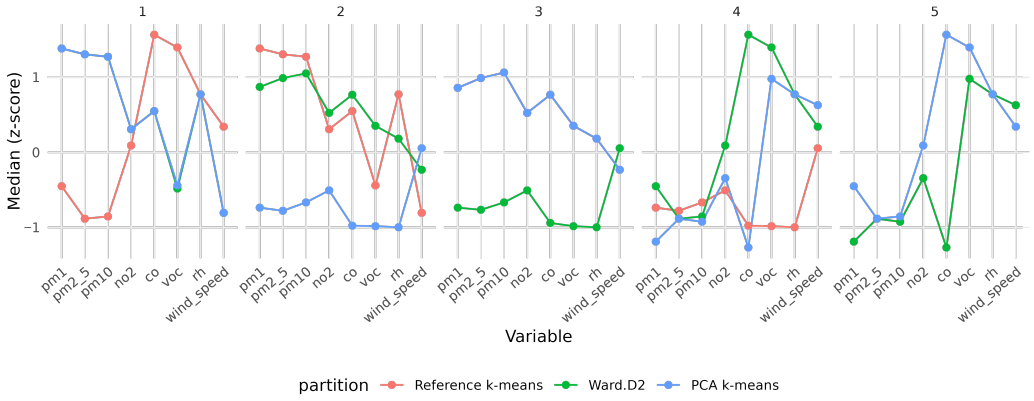

Figure A1. Per-regime pollutant medians under the three partitions (Regimes 1–5 correspond to the five reference regimes from the $k = 5$ clustering).

Appendix A.4. Residual Diagnostics and Log-Scale Robustness for Section 3.6

As shown in Figure A2, residuals are approximately homoscedastic around zero across the fitted range, with a single extreme positive residual producing a modest upper-tail departure in the Q–Q plot. The bulk of residuals align with normality, indicating that inference is not driven by systematic variance patterns. For interpretability, we use these diagnostics as a sanity check for the linear-model inference; the predictive comparisons are additionally supported by the hold-out design described in the main text.

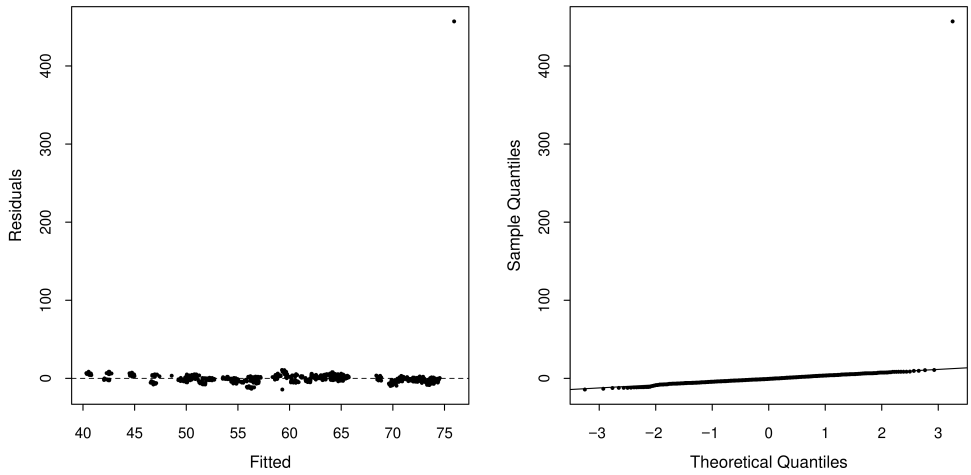

Figure A2. Left: NO_2 residuals vs. fitted. **Right**: Normal Q–Q plot (augmented model with regime).

According to Figure A2 and Table A4, conclusions on the regime effect are consistent on both scales with the results obtained in Section 3.

Table A4. Natural- vs. log-scale models for NO_2 (R^2 and ΔR^2 from adding regime).

Scale	R^2 (Base)	R^2 (+Regime)	ΔR^2
Natural	0.194	0.251	0.057
Log	0.617	0.717	0.100

Appendix A.5. LOSO Portability: With vs. Without the Regime Factor

We fitted paired LOSO Random Forest models using identical predictors, once without and once with the regime factor, where regimes were trained on the training sites and applied unchanged to the held-out site.

Negative R^2 values can occur under LOSO when the model performs worse than a simple baseline (e.g., predicting the the test-site mean baseline); in our setting, strong cross-site heterogeneity can lead to severe extrapolation errors for some targets/sites.

As seen in Table A5, gains are most apparent for NO_2. For PM and some gases, cross-site heterogeneity limits portability despite modest RMSE reductions.

Table A5. Paired LOSO performance (excerpt). Δ denotes (withReg − noReg).

Method	Site	Target	R^2 (noReg)	R^2 (withReg)	ΔR^2	ΔRMSE
RF	1	CO	−18.50	−11.41	7.09	−3.61
RF	2	CO	−7.81	−3.29	4.52	−4.54
RF	3	CO	−3.96	−2.52	1.43	−0.93
RF	4	CO	−0.29	−0.21	0.08	−0.38
RF	5	CO	−48.40	−44.03	4.41	−0.69
(full table for all targets/sites available on request)						

References

1. U.S. Environmental Protection Agency (EPA). Exposure Assessment Tools by Routes–Inhalation (EPA ExpoBox). Last Updated on 1 April 2025. Available online: https://www.epa.gov/expobox/exposure-assessment-tools-routes-inhalation (accessed on 21 January 2026).
2. U.S. Environmental Protection Agency (EPA). Near-Road NO_2 Monitoring Technical Assistance Document (TAD). EPA-454/B-12-002. June 2012. Available online: https://www.epa.gov/sites/default/files/2020-09/documents/nearroadtad.pdf (accessed on 21 January 2026).

3. Hwa, M.-Y.; Hsieh, C.-C.; Wu, T.-C.; Chang, L.-F.W. Real-world vehicle emissions and VOCs profile in the Taipei tunnel located at Taiwan Taipei area. *Atmos. Environ.* **2002**, *36*, 1993–2002. [CrossRef]
4. Kim, S.R.; Dominici, F.; Buckley, T.J. Concentrations of vehicle-related air pollutants in an urban parking garage. *Environ. Res.* **2007**, *105*, 291–299. [CrossRef]
5. Braniš, M.; Šafránek, J.; Hytychová, A. Exposure of children to airborne particulate matter of different size fractions during indoor physical education at school. *Build. Environ.* **2009**, *44*, 1246–1252. [CrossRef]
6. World Health Organization. *WHO Global Air Quality Guidelines: Particulate Matter ($PM_{2.5}$ and PM_{10}), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*; World Health Organization: Geneva, Switzerland, 2021; ISBN 978-92-4-003422-8. Available online: https://iris.who.int/bitstream/handle/10665/345329/9789240034228-eng.pdf (accessed on 21 January 2026).
7. Agency for Toxic Substances and Disease Registry (ATSDR). *Toxicological Profile for Hydrogen Sulfide and Carbonyl Sulfide*; U.S. Department of Health and Human Services: Atlanta, GA, USA, 2016. Available online: https://www.atsdr.cdc.gov/toxprofiles/tp114.pdf (accessed on 21 January 2026).
8. Karner, A.A.; Eisinger, D.S.; Niemeier, D.A. Near-roadway air quality: Synthesizing the findings from real-world data. *Environ. Sci. Technol.* **2010**, *44*, 5334–5344. [CrossRef]
9. Health Effects Institute. *Traffic-Related Air Pollution: A Critical Review of the Literature on Emissions, Exposure, and Health Effects*; HEI Special Report 17; Health Effects Institute: Boston, MA, USA, 2010. Available online: https://www.healtheffects.org/publication/traffic-related-air-pollution-critical-review-literature-emissions-exposure-and-health (accessed on 21 January 2026).
10. Grange, S.K.; Carslaw, D.C. Using meteorological normalisation to detect interventions in air quality time series. *Sci. Total Environ.* **2019**, *653*, 578–588. [CrossRef]
11. Austin, E.; Coull, B.A.; Thomas, D.; Koutrakis, P. A framework for identifying distinct multipollutant profiles in air pollution data. *Environ. Int.* **2012**, *45*, 112–121. [CrossRef]
12. Apte, J.S.; Messier, K.P.; Gani, S.; Brauer, M.; Kirchstetter, T.W.; Lunden, M.M.; Marshall, J.D.; Portier, C.J.; Vermeulen, R.C.H.; Hamburg, S.P. High-resolution air pollution mapping with Google Street View cars: Exploiting big data. *Environ. Sci. Technol.* **2017**, *51*, 6999–7008. [CrossRef] [PubMed]
13. Messier, K.P.; Matthews, J.L.; Rubinstein, S.; Alvarez, R.; Brauer, M.; Choi, J.J.; Hamburg, S.P.; Kerckhoffs, J.; LaFranchi, B.; Lunden, M.M.; et al. Mapping air pollution with Google Street View cars: Efficient approaches with mobile monitoring and land-use regression. *Environ. Sci. Technol.* **2018**, *52*, 12563–12572. [CrossRef]
14. Agbehadji, I.E.; Obagbuwa, I.C. Systematic review of machine learning and deep learning techniques for spatiotemporal air quality prediction. *Atmosphere* **2024**, *15*, 1352. [CrossRef]
15. Houdou, A.; ElBadisy, I.; Khomsi, K.; Abdala, S.A.; Abdulla, F.; Najmi, H.; Obtel, M.; Belyamani, L.; Ibrahimi, A.; Khalis, M. Interpretable machine learning approaches for forecasting and predicting air pollution: A systematic review. *Aerosol Air Qual. Res.* **2024**, *24*, 230151. [CrossRef]
16. Jayaratne, R.; Liu, X.; Ahn, K.; Asumadu-Sakyi, A.B.; Fisher, G.; Gao, J.; Mabon, A.; Mazaheri, M.; Mullins, B.; Nyaku, M.; et al. Low-cost PM2.5 sensors: An assessment of their suitability for various applications. *Aerosol Air Qual. Res.* **2020**, *20*, 520–532. [CrossRef]
17. Hankey, S.; Marshall, J.D. On-bicycle exposure to particulate air pollution: Particle number, black carbon, $PM_{2.5}$ and particle size. *Atmos. Environ.* **2015**, *122*, 65–73. [CrossRef]
18. Hofman, J.; Samson, R.; Joosen, S.; Blust, R.; Lenaerts, S. Cyclist exposure to black carbon, ultrafine particles and $PM_{2.5}$ in different traffic conditions: A real-time mobile monitoring study. *Environ. Res.* **2018**, *164*, 530–538. [CrossRef]
19. Reda, I.; Andreas, A. Solar position algorithm for solar radiation applications. *Sol. Energy* **2004**, *76*, 577–589. [CrossRef]
20. Hennig, C. Cluster-wise assessment of cluster stability. *Comput. Stat. Data Anal.* **2007**, *52*, 258–271. [CrossRef]
21. MacQueen, J.B. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*; University of California Press: Berkeley, CA, USA, 1967; pp. 281–297. Available online: https://api.semanticscholar.org/CorpusID:6278891 (accessed on 21 January 2026).
22. Hartigan, J.A.; Wong, M.A.C A *K*-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1979**, *28*, 100–108. [CrossRef]
23. Lloyd, S.P. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]
24. McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical Density Based Clustering. *J. Open Source Softw.* **2017**, *2*, 205. [CrossRef]
25. McLachlan, G.; Peel, D. *Finite Mixture Models*; Wiley: New York, NY, USA, 2000. [CrossRef]
26. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
27. Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2*, 193–218. [CrossRef]
28. Ward, J.H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
29. Jolliffe, I.T.; Cadima, J. Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. A* **2016**, *374*, 20150202. [CrossRef] [PubMed]

30. Lange, T.; Roth, V.; Braun, M.L.; Buhmann, J.M. Stability-based validation of clustering solutions. *Neural Comput.* **2004**, *16*, 1299–1323. [CrossRef] [PubMed]
31. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
32. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22. Available online: https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf (accessed on 21 January 2026).
33. Eid, A.; Jodeh, S.; Hanbali, G.; Hawawreh, M.; Chakir, A.; Roth, E. Multi-Output Machine-Learning Prediction of Volatile Organic Compounds (VOCs): Learning from Co-Emitted VOCs. *Environments* **2025**, *12*, 216. [CrossRef]
34. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*, 6th ed.; Wiley: Hoboken, NJ, USA, 2021.
35. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
36. Cawley, G.C.; Talbot, N.L.C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107. Available online: https://www.jmlr.org/papers/v11/cawley10a.html (accessed on 21 January 2026).
37. Scikit-Learn Developers. *Model Evaluation: Quantifying the Quality of Predictions (R^2 Score)*; Scikit-Learn Documentation. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html (accessed on 21 January 2026).
38. Castro, G.J.; Zimek, A.; Sander, J.; Campello, R.J.G.B. A unified view of density-based methods for semi-supervised clustering and classification. *Data Min. Knowl. Discov.* **2019**, *33*, 1894–1952. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC7410108/ (accessed on 21 January 2026). [CrossRef]