

Tracking Urban Atmospheric Pollutants using Sentinel-5P Satellite Data

To be added

¹To be added

²To be added

To be added

Keywords: Geospatial analytics, Sentinel-5P, Remote sensing, Air quality.

Abstract: Urban nitrogen dioxide (NO_2) is a key indicator of combustion-related air pollution and exhibits strong spatial and temporal variability in cities. This study presents a satellite-based framework for tracking urban NO_2 pollution using tropospheric column observations from Sentinel-5P/TROPOMI over Guayas Province, Ecuador. Rather than estimating surface concentrations, the methodology emphasizes robust distributional metrics, including the median and upper-tail percentiles (P_{90} , P_{95} , and P_{99}), to characterize background conditions and localized pollution extremes at the canton scale. Multi-year satellite observations are aggregated annually and analyzed using unsupervised K-means clustering to identify characteristic pollution regimes without predefined thresholds. Results show that highly urbanized cantons consistently exhibit elevated extreme NO_2 values and greater variability, while less urbanized areas display lower and more homogeneous patterns. The proposed approach provides an interpretable and scalable tool for urban air-quality assessment in data-scarce regions using satellite observations alone. The implementation is publicly available on GitHub <https://anonymous.4open.science/w/sentinel-5P-clustering-81DE/>.

1 INTRODUCTION

Urban nitrogen dioxide (NO_2) is a key air-quality pollutant because it is tightly coupled to high-temperature combustion sources (e.g., road traffic, power generation, and industrial activity) and therefore exhibits strong spatial heterogeneity over cities (Fenger, 1999; Judd et al., 2020). As a short-lived species, NO_2 responds rapidly to emission changes and meteorological transport, which makes satellite observations particularly valuable for tracking urban-scale variability (Andreae, 2019). In addition to its direct health relevance, NO_2 plays an important role in atmospheric chemistry: in the presence of sunlight and volatile organic compounds, NO_2 contributes to photochemical ozone production and to broader processes affecting air-quality and oxidation capacity (Lorente et al., 2021; Mejía et al., 2023).

Satellite remote sensing complements in situ monitoring by providing spatially continuous coverage. The Sentinel-5 Precursor mission (Sentinel-5P) carrying the TROPospheric Monitoring Instrument (TROPOMI) provides near-daily global observations of NO_2 at kilometer-scale resolution, enabling the identification of emission hotspots and the characterization of urban pollution gradients (Tian et al.,

2022; Fiore et al., 2015). However, the operational TROPOMI NO_2 product is a column quantity rather than a direct near-surface concentration, so urban interpretation is best framed in terms of spatiotemporal contrasts, anomalies, and trends in retrieved columns (Davybida, 2023; Sha et al., 2021).

Robust use of TROPOMI NO_2 for city-scale analyses requires careful spatiotemporal processing because satellite observations are high-dimensional, noisy, and irregularly sampled across space and time (Verhoelst et al., 2021; Douros et al., 2023; Judd et al., 2020). In this context, machine learning (ML) provides a practical framework to transform multi-year NO_2 columns into interpretable urban indicators by (i) learning dominant spatiotemporal structures, (ii) separating persistent patterns from short-term variability, and (iii) grouping areas with similar pollution behavior using unsupervised methods (e.g., clustering). Building on these foundations, this work integrates machine learning with Sentinel-5P/TROPOMI NO_2 to track urban atmospheric pollution patterns and to derive robust summary statistics that enable comparison of temporal changes and spatial heterogeneity across the study region.

The manuscript is organized as follows. Section 2 presents works related to tracking urban atmospheric

pollutants using satellite data. Section 3 presents the proposed methodology. Experimental results and comparisons are given in Section 4. Finally, conclusions are presented in Section 5.

2 BACKGROUND

Recent studies show an increasing use of Sentinel-5P/TROPOMI data for urban air-quality assessment, with diverse analytical strategies and complementary strengths. In data-scarce contexts, Mejía et al. (Mejía et al., 2023) demonstrate how spatial interpolation of Sentinel-5P NO_2 can improve intra-urban interpretability in Guayaquil, highlighting empirical Bayesian kriging as an effective approach when ground monitoring is limited. However, interpolation-based methods remain constrained by the native satellite resolution and cannot generate new sub-pixel information.

At the city scale, Shah et al. (Shah et al., 2024) illustrate the operational feasibility of satellite-based urban air-quality monitoring through spatiotemporal analysis over Pune City. Their work shows that repeated TROPOMI observations can be transformed into interpretable urban indicators, although results remain sensitive to processing choices and to the column-based nature of the satellite product.

More recently, Prajesh et al. (Prajesh et al., 2025) propose a signal-isolation framework that exploits seasonal structure and reference-region contrasts to extract differential pollution signals from Sentinel-5P time series. While this approach strengthens attribution-like interpretations, it also highlights challenges related to non-stationarity and evolving background conditions in atmospheric columns.

Overall, existing literature suggests that Sentinel-5P/TROPOMI enables robust characterization of urban NO_2 patterns when analyses focus on spatiotemporal contrasts and distributional behavior rather than absolute surface concentrations. These insights motivate the present work, which emphasizes robust statistical metrics and unsupervised learning to track urban pollution dynamics under the constraints of column-based satellite observations.

3 MATERIAL & METHODS

The proposed methodology integrates satellite-based remote sensing data from Sentinel-5P/TROPOMI with robust statistical analysis to characterize the spatiotemporal behavior of urban NO_2 pollution in Guayas Province, Ecuador. Rather than attempting

to infer surface-level concentrations, the methodology focuses on relative contrasts, temporal evolution, and upper-tail behavior of tropospheric NO_2 columns at the canton scale. The workflow is designed to be reproducible, data-driven, and suitable for regions with limited ground-based air-quality monitoring.

Figure 1 illustrates the geographical context of the study area, highlighting the spatial distribution of cantons within Guayas Province. The overall methodology consists of four main stages: (i) data acquisition and filtering, (ii) spatial aggregation at the canton level, (iii) temporal compositing and statistical characterization, and (iv) exploratory pattern analysis using unsupervised learning.

3.1 Study Area

Guayas Province is one of the most densely populated and economically active regions in Ecuador, hosting major urban centers such as Guayaquil, Durán, and Samborombón. According to official population projections, the province concentrates a large share of national industrial, commercial, and transportation activities, making it a relevant case study for satellite-based urban air-quality assessment. The administrative division into cantons provides a natural spatial unit for aggregation and comparison of NO_2 pollution patterns.

3.2 Dataset

This study uses tropospheric nitrogen dioxide (NO_2) column data from the Sentinel-5 Precursor (Sentinel-5P) mission, acquired by the TROPospheric Monitoring Instrument (TROPOMI). The Level-2 NO_2 product provides near-daily global coverage with a spatial resolution on the order of a few kilometers, enabling consistent monitoring of urban-scale pollution dynamics.

Satellite observations are filtered using standard quality assurance criteria to remove unreliable retrievals affected by clouds, surface albedo, or retrieval convergence issues. Valid pixels are spatially intersected with canton boundaries, and only pixels whose centers fall within a given canton are retained. This spatial masking ensures administrative consistency while preserving the native satellite sampling characteristics.

Temporal coverage spans multiple consecutive years, allowing the construction of annual composites for each canton. For each year, all valid daily observations are pooled, forming a distribution of NO_2 column values that reflects both persistent emission patterns and short-term variability due to meteorology

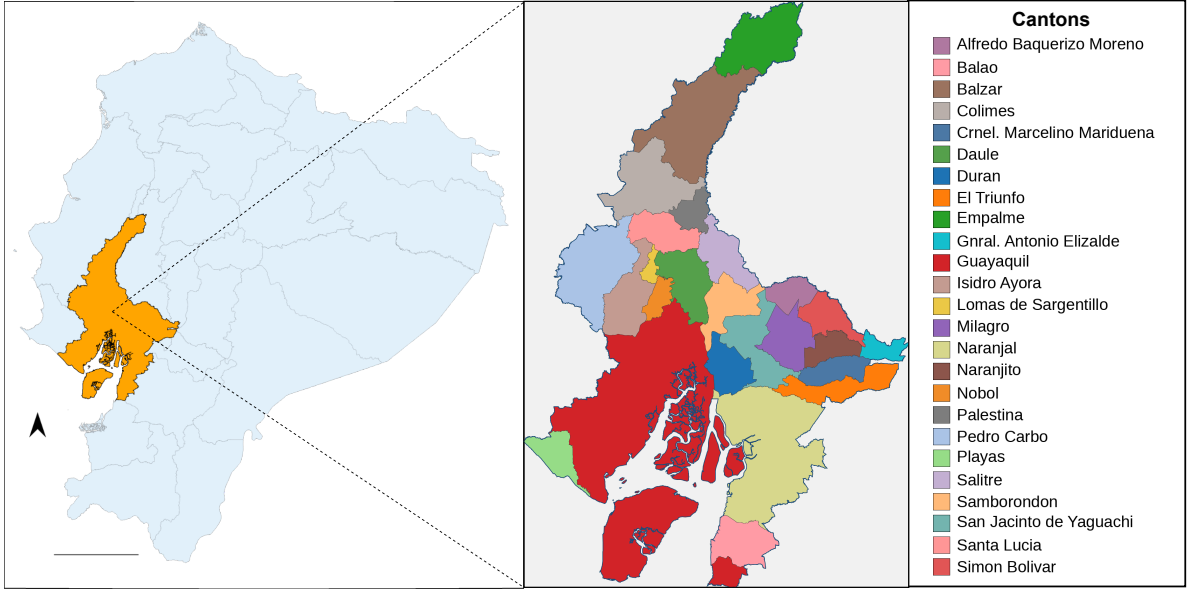


Figure 1: Map of Guayas Province, Ecuador, illustrating the spatial distribution of cantons across the province.

and activity changes.

3.3 Statistical Characterization

Instead of relying on mean values, which can be strongly influenced by clean background areas or isolated outliers, the methodology emphasizes robust distributional metrics derived from the empirical pixel-value distribution within each canton and year. Five summary statistics are considered: the median (\tilde{x}), and the 90th, 95th, and 99th percentiles (P_{90} , P_{95} , and P_{99}).

The median represents a typical background pollution level, while upper-tail percentiles characterize increasingly severe pollution conditions affecting a smaller fraction of the urban area. In particular, P_{90} captures conditions experienced by the most polluted decile of pixels, P_{95} highlights more persistent extreme values, and P_{99} emphasizes severe hotspots while remaining more stable than the absolute maximum.

Let a sample of pixel values be x_1, x_2, \dots, x_n , and let the order statistics be

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

Median.

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{if } n \text{ is odd,} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & \text{if } n \text{ is even.} \end{cases}$$

Percentile (discrete definition). For $p \in (0, 100)$,

$$P_p = x_{(\lceil \frac{p}{100} n \rceil)}.$$

Percentile (interpolated definition). Define

$$h = \frac{p}{100}(n+1), \quad j = \lfloor h \rfloor, \quad \gamma = h - j.$$

Then (with endpoint handling),

$$P_p = \begin{cases} x_{(1)}, & h \leq 1, \\ (1-\gamma)x_{(j)} + \gamma x_{(j+1)}, & 1 < h < n, \\ x_{(n)}, & h \geq n. \end{cases}$$

Specific upper-tail percentiles. Using the interpolated definition:

$$h_{90} = \frac{90}{100}(n+1), \quad j_{90} = \lfloor h_{90} \rfloor, \quad \gamma_{90} = h_{90} - j_{90}, \\ P_{90} = (1-\gamma_{90})x_{(j_{90})} + \gamma_{90}x_{(j_{90}+1)}.$$

$$h_{95} = \frac{95}{100}(n+1), \quad j_{95} = \lfloor h_{95} \rfloor, \quad \gamma_{95} = h_{95} - j_{95}, \\ P_{95} = (1-\gamma_{95})x_{(j_{95})} + \gamma_{95}x_{(j_{95}+1)}.$$

$$h_{99} = \frac{99}{100}(n+1), \quad j_{99} = \lfloor h_{99} \rfloor, \quad \gamma_{99} = h_{99} - j_{99}, \\ P_{99} = (1-\gamma_{99})x_{(j_{99})} + \gamma_{99}x_{(j_{99}+1)}.$$

By jointly analyzing these metrics, the methodology distinguishes between broad urban background pollution and localized high-emission areas, enabling a nuanced interpretation of spatial heterogeneity and temporal evolution.

3.4 Temporal and Spatial Analysis

For each canton, the selected metrics are computed annually, producing multi-year time series that allow the assessment of trends, interannual variability, and potential regime shifts. Comparisons across cantons highlight spatial contrasts within the province and support the identification of consistently high-pollution areas versus more stable or cleaner regions.

To further explore spatial structure, unsupervised clustering (K-means) is applied to the multi-metric feature space composed of median and upper-tail percentiles. This step groups cantons or pixels with similar pollution behavior, facilitating the identification of characteristic pollution regimes without imposing predefined thresholds. The resulting cluster maps provide an intuitive spatial summary of urban NO_2 patterns across Guayas Province.

Overall, the proposed methodology offers a statistically robust and interpretable framework for tracking urban atmospheric pollution using Sentinel-5P data, particularly suited to data-scarce regions where ground-based monitoring is limited or unevenly distributed.

4 RESULTS

This section presents the experimental results obtained by applying the proposed methodology to Sentinel-5P/TROPOMI NO_2 data over Guayas Province. Results are organized into quantitative and qualitative analyses, focusing on temporal evolution, spatial contrasts among cantons, and the identification of characteristic pollution patterns.

4.1 Quantitative Evaluation

The quantitative evaluation is based on the annual distributional metrics described in Section 3, computed for each canton over the study period. The cantons shown in the Table 1 are based on the number of inhabitants according to official data¹. Table 1 summarizes the \bar{x} and upper-tail percentiles (P_{90} , P_{95} , and P_{99}) of tropospheric NO_2 columns for the most representative cantons in Guayas Province.

Across all cantons, the \bar{x} values exhibit relatively moderate interannual variability compared to upper-tail metrics, indicating that background NO_2 levels remain comparatively stable over time. In contrast, P_{90} , P_{95} , and especially P_{99} show substantially larger

fluctuations, reflecting the sensitivity of extreme pollution conditions to changes in anthropogenic activity, meteorology, and episodic events.

Highly urbanized and industrialized cantons such as Guayaquil, Durán, and Samborombón consistently present the highest upper-tail values. For example, Guayaquil shows pronounced increases in P_{95} and P_{99} in recent years, suggesting the persistence and, in some cases, intensification of localized NO_2 hotspots even when median levels remain relatively stable. This divergence between central and extreme metrics highlights the importance of analyzing the full distribution rather than relying on a single summary statistic.

Less densely populated cantons, including Salitre, El Triunfo, and Empalme, exhibit lower median and percentile values overall, as well as reduced variability in extreme metrics. These patterns are consistent with lower traffic density and industrial activity, supporting the physical interpretability of the satellite-derived indicators.

4.2 Qualitative Evaluation

The qualitative evaluation focuses on the spatial distribution of NO_2 pollution patterns and their evolution across the study area. Figure 2 presents spatial maps of the median, P_{90} , P_{95} , and P_{99} NO_2 columns, revealing clear spatial gradients and localized hotspots that are not apparent when using average values alone.

Median maps emphasize broad urban–rural contrasts, with elevated background NO_2 levels over major urban centers. In contrast, P_{95} and P_{99} maps highlight compact, spatially coherent hotspots associated with dense traffic corridors, industrial zones, and port-related activities. These results confirm that extreme percentiles are effective for isolating persistent high-emission areas within heterogeneous urban environments.

Figure 3 illustrates the temporal evolution of the median and upper-tail percentiles aggregated at the provincial level. The figure shows that extreme percentiles respond more strongly to interannual changes than the median, reinforcing their utility for tracking changes in urban pollution severity.

To further synthesize spatial patterns, an intensity-based K-means clustering is applied to canton-level NO_2 statistics. For each canton, three features are extracted from the annual NO_2 distributions: the mean, maximum, and standard deviation of the selected NO_2 metric across the study period. Prior to clustering, features are standardized using z-score normalization to ensure equal contribution of all variables.

The optimal number of clusters (K) is determined

¹ <https://www.ecuadorencifras.gob.ec/proyecciones-poblacionales/>

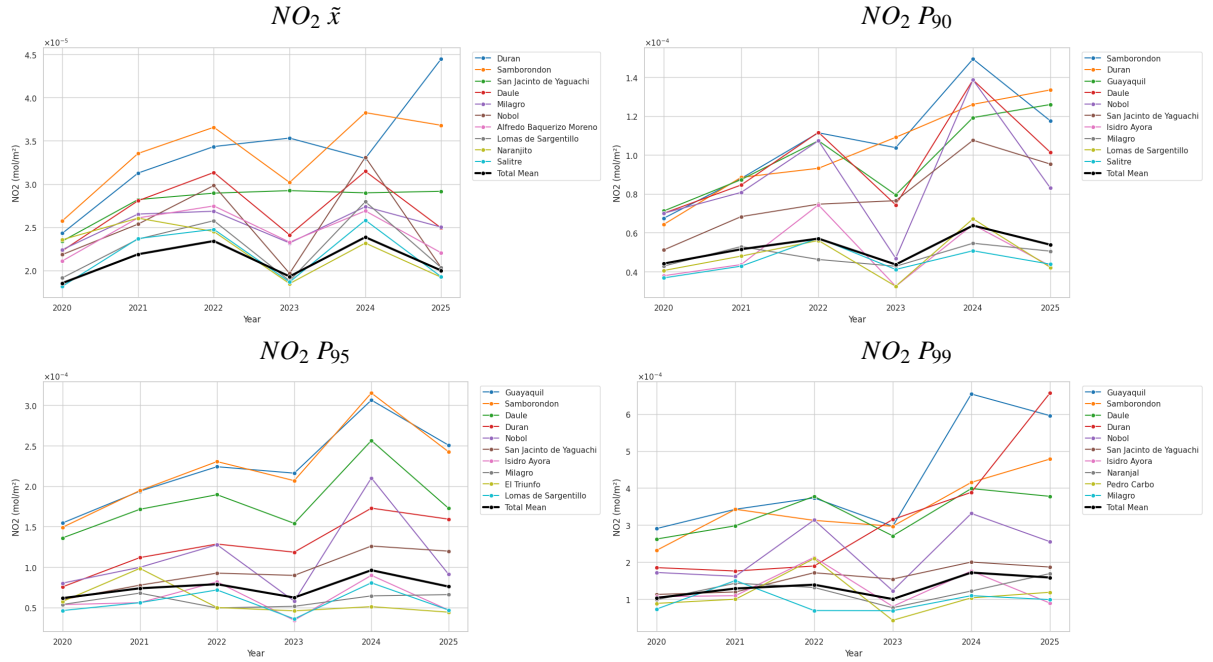


Figure 2: Spatial distribution of annual NO_2 median (\bar{x}), P_{90} , P_{95} , and P_{99} tropospheric column values over Guayas Province, illustrating background levels and localized pollution hotspots.

automatically using the silhouette score, evaluated for values of K ranging from 2 to 7 (see Fig. 4). The silhouette analysis consistently indicated an optimal solution at $K = 2$, reflecting a clear separation between cantons characterized by higher NO_2 intensity and variability and those exhibiting lower and more homogeneous pollution levels.

Figure 5 shows the K-means clustering results obtained from canton-level NO_2 distributional metrics derived from Sentinel-5P/TROPOMI data. The clustering, based on standardized mean, maximum, and standard deviation of NO_2 columns, identifies two clearly distinct pollution regimes across Guayas Province.

The cluster highlighted in red corresponds to cantons with higher NO_2 intensity and variability, characterized by elevated median values and pronounced upper-tail percentiles. This group is dominated by highly urbanized and industrial cantons, including Guayaquil, Durán, Daule, and Samborondón, where dense traffic networks, port activities, and concentrated economic activity contribute to persistent NO_2 hotspots.

In contrast, the remaining cantons form a low-intensity cluster with lower NO_2 levels and reduced variability, consistent with less urbanized or predominantly rural environments. The strong spatial coherence of the clusters and their agreement with known urbanization patterns support the interpretability and

robustness of the proposed unsupervised classification approach.

Overall, the experimental results demonstrate that the proposed methodology successfully captures both temporal dynamics and spatial heterogeneity of urban NO_2 pollution using Sentinel-5P data. The combined use of robust statistical metrics and unsupervised learning provides an interpretable and scalable framework for urban air-quality assessment in data-scarce regions.

5 CONCLUSIONS

This study shows that Sentinel-5P/TROPOMI NO_2 data, combined with robust statistics and unsupervised machine learning, can effectively describe urban pollution patterns in regions with scarce monitoring. Instead of relying on averages, it uses distribution metrics (e.g., percentiles) to capture both typical conditions and extreme NO_2 events.

In Guayas Province, the most urbanized cantons present consistently higher upper-tail NO_2 percentiles even when median values are stable, suggesting persistent localized emission hotspots. K-means clustering based on intensity separates the area into two clear pollution regimes that match known differences in urbanization, traffic, and economic activity.

Overall, the framework is scalable and repro-

Canton	Population	Metric	$NO_2 (\times 10^{-5})$					
			2020	2021	2022	2023	2024	2025
Guayaquil	2,924,038	\tilde{x}	1.41	1.64	1.69	1.60	1.67	1.60
		P_{90}	7.12	8.73	10.73	7.95	11.92	12.59
		P_{95}	15.46	19.38	22.41	21.61	30.64	25.07
		P_{99}	29.09	34.28	37.36	29.70	65.40	59.56
Durán	319,622	\tilde{x}	2.43	3.13	3.43	3.53	3.30	4.45
		P_{90}	6.43	8.85	9.31	10.91	12.60	13.35
		P_{95}	7.56	11.18	12.85	11.85	17.28	15.93
		P_{99}	18.52	17.61	18.96	31.52	38.87	65.67
Daule	228,833	\tilde{x}	2.23	2.81	3.13	2.41	3.15	2.49
		P_{90}	7.00	8.46	11.17	7.43	13.86	10.14
		P_{95}	13.61	17.14	18.95	15.41	25.64	17.27
		P_{99}	26.25	29.82	37.76	27.11	39.88	37.77
Milagro	212,982	\tilde{x}	2.24	2.65	2.69	2.32	2.74	2.50
		P_{90}	4.29	5.29	4.63	4.28	5.46	5.05
		P_{95}	5.38	6.78	4.95	5.15	6.44	6.60
		P_{99}	7.37	15.04	6.94	6.89	10.96	9.89
Samborondon	103,266	\tilde{x}	2.57	3.35	3.66	3.02	3.83	3.68
		P_{90}	6.75	8.79	11.14	10.38	14.94	11.76
		P_{95}	14.93	19.46	23.05	20.68	31.52	24.26
		P_{99}	23.23	34.28	31.29	29.70	41.54	47.84
Naranjal	90,511	\tilde{x}	1.62	2.07	2.29	2.01	2.05	1.93
		P_{90}	3.47	4.61	4.00	4.00	3.95	4.74
		P_{95}	4.67	6.34	4.89	5.21	4.99	6.63
		P_{99}	9.86	14.32	13.17	7.70	12.22	16.92
Empalme	87,296	\tilde{x}	1.50	1.73	1.96	1.65	2.02	1.52
		P_{90}	3.36	4.23	3.94	2.97	3.98	3.36
		P_{95}	4.29	4.72	4.79	3.61	5.21	4.11
		P_{99}	8.57	11.31	6.55	8.13	13.02	5.13
San Jacinto de Yaguachi	76,761	\tilde{x}	2.34	2.82	2.89	2.92	2.90	2.92
		P_{90}	5.12	6.83	7.47	7.65	10.76	9.53
		P_{95}	5.98	7.77	9.25	8.97	12.60	11.97
		P_{99}	11.29	11.96	17.14	15.45	20.03	18.68
Salitre	65,470	\tilde{x}	1.82	2.37	2.48	1.87	2.58	1.93
		P_{90}	3.68	4.28	5.73	4.11	5.07	4.38
		P_{95}	4.24	4.87	7.16	4.76	6.19	5.89
		P_{99}	5.22	6.52	9.92	5.47	9.41	10.20
El Triunfo	63,924	\tilde{x}	1.99	2.16	2.42	1.93	2.18	1.81
		P_{90}	4.68	5.34	4.35	3.37	4.11	3.55
		P_{95}	5.76	9.84	5.01	4.59	5.11	4.43
		P_{99}	8.21	14.32	8.10	5.93	11.13	8.25

Table 1: Annual median (\tilde{x}) and upper-tail percentiles (P_{90} , P_{95} , P_{99}) of tropospheric NO_2 columns derived from Sentinel-5P/TROPOMI for the ten most populated cantons of Guayas Province during the period 2020–2025. Cantons are ordered by population size.

ducible for satellite-based urban air-quality assessment. Although it cannot directly estimate surface concentrations, focusing on relative contrasts, ex-

tremes, and spatiotemporal variability makes it useful for long-term monitoring and comparisons where ground data are limited.

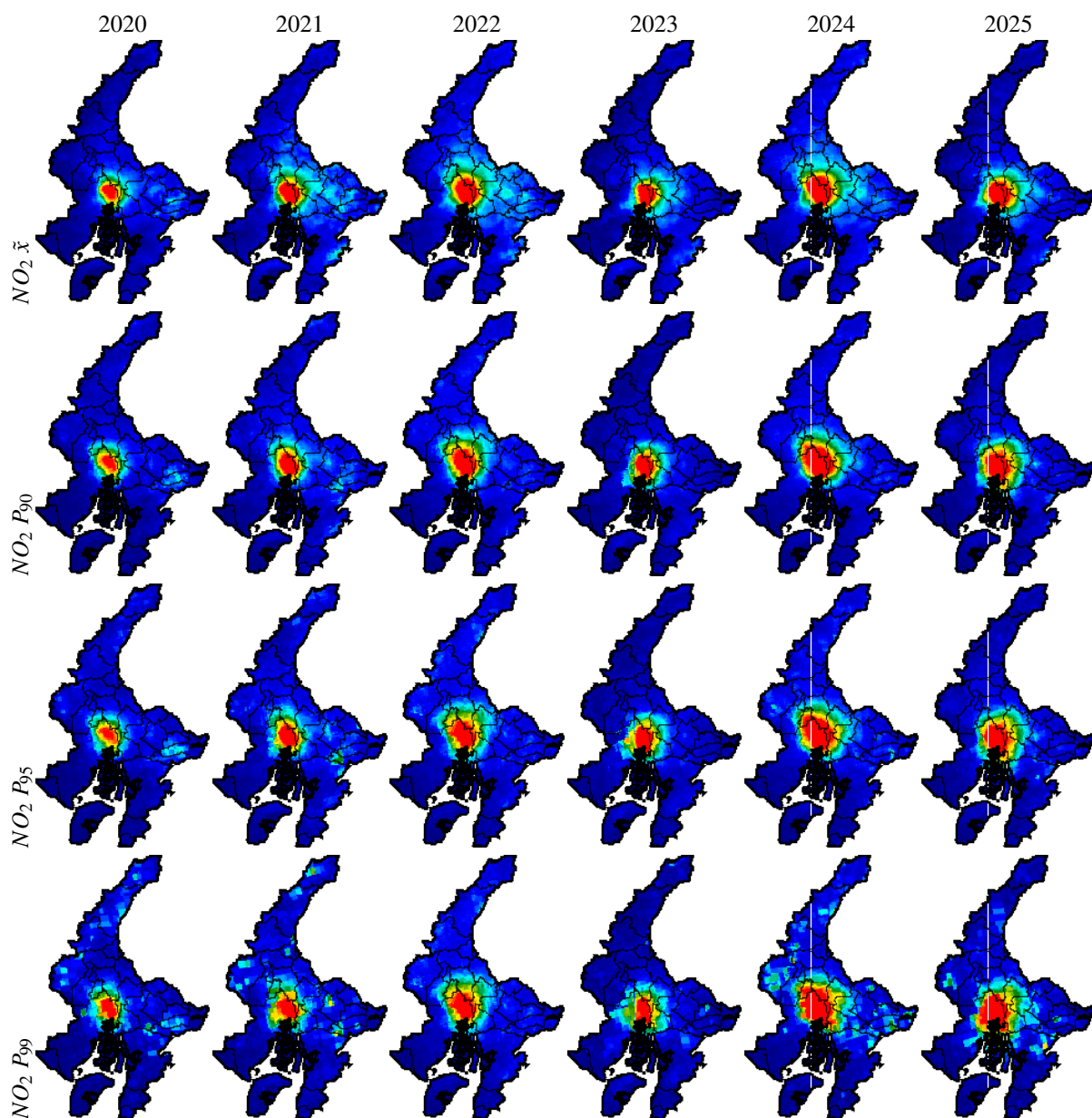


Figure 3: Interannual evolution (2020–2025) of provincial-scale NO_2 distributional metrics (\bar{x} , P_{90} , P_{95} , and P_{99}), highlighting the stronger variability of extreme percentiles relative to background conditions.

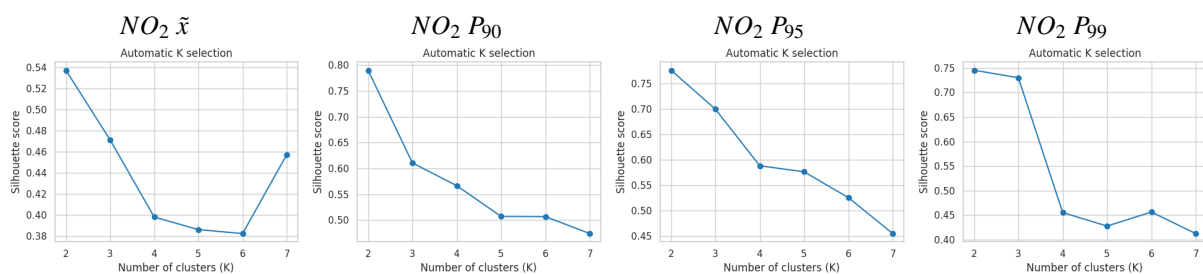


Figure 4: Silhouette-based automatic K selection for K-means clustering using NO_2 distributional metrics, showing the optimal number of clusters ($K = 2$) across median and upper-tail percentiles.

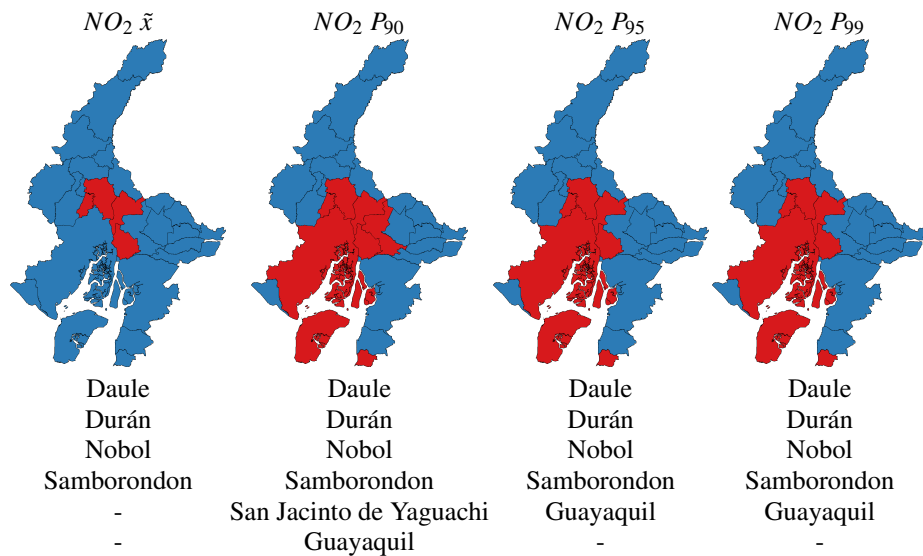


Figure 5: K-means clustering map of Guayas Province based on NO_2 distributional metrics, grouping cantons with similar pollution intensity and variability patterns. The names of the cantons highlighted in red (cluster with high value of NO_2) are listed below the map.

ACKNOWLEDGEMENTS

To be added.

REFERENCES

- Andreae, M. O. (2019). Emission of trace gases and aerosols from biomass burning—an updated assessment. *Atmospheric Chemistry and Physics*, 19(13):8523–8546.
- Davybida, L. (2023). Air quality impacts of war detected from the sentinel-5p satellite over ukraine. In *Conf. Series: Earth and Environmental Science*, volume 1254, page 012112. IOP Publishing.
- Douros, J., Eskes, H., van Geffen, J., Boersma, K. F., Compornolle, S., Pinardi, G., Blechschmidt, A.-M., Peuch, V.-H., Colette, A., and Veeffkind, J. P. (2023). Comparing sentinel-5p tropomi no_2 column observations with the cams regional air quality ensemble. *Geoscientific Model Development*, 16:509–534.
- Fenger, J. (1999). Urban air quality. *Atmospheric environment*, 33(29):4877–4900.
- Fiore, A. M., Naik, V., and Leibensperger, E. M. (2015). Air quality and climate connections. *Journal of the Air & Waste Management Association*, 65(6):645–685.
- Judd, L. M., Al-Saadi, J. A., Szykman, J. J., Valin, L. C., Janz, S. J., Kowalewski, M. G., Eskes, H. J., Veeffkind, J. P., Cede, A., et al. (2020). Evaluating sentinel-5p tropomi tropospheric no_2 column densities with airborne and pandora spectrometers near new york city and long island sound. *Atmospheric Measurement Techniques*, 13(11):6113–6140.
- Lorente, A., Borsdorff, T., Butz, A., Hasekamp, O., aan de Brugh, J., Schneider, A., Wu, L., Hase, F., Kivi, R., Wunch, D., et al. (2021). Methane retrieved from tropomi: improvement of the data product and validation of the first 2 years of measurements. *Atmospheric Measurement Techniques*, 14(1):665–684.
- Mejía, D., Alvarez, H., Zalakeviciute, R., Macancela, D., Sanchez, C., and Bonilla, S. (2023). Sentinel satellite data monitoring of air pollutants with interpolation methods in guayaquil, ecuador. *Remote Sensing Applications: Society and Environment*, 31:100990.
- Prajesh, P. J., Ragunath, K., Gordon, M., and Neethirajan, S. (2025). Satellite-based seasonal fingerprinting of methane emissions from canadian dairy farms using sentinel-5p. *Climate*, 13(7):135.
- Sha, M. K., Langerock, B., Blavier, J.-F. L., Blumenstock, T., Borsdorff, T., Buschmann, M., Dehn, A., De Mazière, M., Deutscher, N. M., Feist, D. G., et al. (2021). Validation of methane and carbon monoxide from sentinel-5 precursor using tccon and ndacc-irwg stations. *Atmospheric Measurement Techniques Discussions*, 2021:1–84.
- Shah, S. V., Gaikwad, S. V., and Vibhute, A. D. (2024). Air quality monitoring using sentinel-5p tropomi—a case study of pune city. *SN Computer Science*, 5(8):1125.
- Tian, Y., Hong, X., Shan, C., Sun, Y., Wang, W., Zhou, M., Wang, P., Lin, P., and Liu, C. (2022). Investigating the performance of carbon monoxide and methane observations from sentinel-5 precursor in china. *Remote Sensing*, 14(23):6045.
- Verhoelst, T., Compornolle, S., Pinardi, G., Lambert, J.-C., Eskes, H. J., Eichmann, K.-U., Fjærraa, A. M., Granville, J., Niemeijer, S., Cede, A., et al. (2021). Ground-based validation of the copernicus sentinel-5p tropomi no_2 measurements with the ndacc zsl-doas, max-doas and pandonia global networks. *Atmospheric Measurement Techniques*, 14:481–510.