EuropeAir Technical Report Series — TR-2025-001

Download PDF

EUROPEAIR TECHNICAL REPORT SERIES

Satellite-Derived Estimation of BTEX and Toxic Compounds over Slovakia: A Multi-Source Proxy Regression Approach Using CAMS and Sentinel-5P Observations

Stanislav Pittner

Independent Research, Trnava, Slovakia

Submitted: 4 April 2025 Data Period: 19 Mar – 5 Apr 2025 Version 1.0

DOI: 10.5281/europeair.2025.tr001 (preprint)

ABSTRACT

We present a multi-source proxy regression framework for estimating ground-level concentrations of benzene (C_6H_6), toluene (C_7H_8), xylene (C_8H_{10}), 1,3-butadiene, and hydrogen sulfide (H_2S) over Slovakia using satellite-derived atmospheric composition data. The methodology employs Copernicus Atmosphere Monitoring Service (CAMS) surface-level reanalysis products (NMVOC, HCHO, SO_2, CO) at 0.1° spatial resolution as proxy predictors, combined with empirical emission ratios from refinery-proximal monitoring studies. Over a 17-day observation window (19 March – 5 April 2025), we generate spatially resolved estimates across 1,102 grid cells covering the full Slovak territory, with particular focus on the Slovnaft refinery influence zone in Bratislava. Model confidence ranges from 30% (xylene, H_2S) to 50% (benzene), constrained by the absence of co-located ground truth BTEX measurements. SO_2 exhibits the strongest refinery enhancement ratio (3.18×) within the 0–5 km proximity zone, while NO_2 and $PM_{2.5}$ show coefficient of variation values of 48.8% and 54.7%, respectively, indicating substantial day-to-day meteorological modulation. These estimates provide a first-order spatial screening tool for identifying potential BTEX hotspots in the absence of dedicated ground-based monitoring networks.

Keywords: BTEX estimation, satellite proxy, CAMS, Sentinel-5P, TROPOMI, NMVOC, formaldehyde, refinery emissions, air quality, Slovakia

1. Introduction

Volatile organic compounds (VOCs), particularly the benzene–toluene–ethylbenzene–xylene (BTEX) group, represent significant health hazards in proximity to petroleum refining facilities. Benzene is classified as a Group 1 carcinogen by IARC with no established safe exposure threshold (WHO, 2010), while chronic toluene and xylene exposure is associated with neurological and hepatic effects. Despite their public health significance, routine ground-based monitoring of BTEX species remains spatially sparse — the Slovak SHMÚ network measures benzene at only 2–3 stations nationwide.

Recent advances in satellite remote sensing, particularly the Sentinel-5 Precursor (S5P) TROPOMI instrument and the Copernicus Atmosphere Monitoring Service (CAMS), offer unprecedented spatial coverage of atmospheric composition at sub-daily temporal resolution. While no current satellite directly measures individual BTEX species, proxy relationships between satellite-observable quantities (HCHO, total NMVOC, CO) and ground-level BTEX have been established in several studies (De Smedt et al., 2021; Zhu et al., 2020; Surl et al., 2018).

The rationale for proxy-based estimation rests on well-documented photochemical and source-attribution relationships:

(i) Formaldehyde (HCHO) serves as a secondary product of VOC oxidation and correlates with total VOC reactivity (R^2 = 0.45–0.65 in urban environments; De Smedt et al., 2021);
(ii) Total NMVOC from CAMS includes anthropogenic and biogenic fractions, with BTEX comprising 8–15% of total NMVOC near refineries (EEA, 2019; Borbon et al., 2013);
(iii) SO_2 co-emission from refinery flue gas provides a source marker for industrial VOC releases (Varon et al., 2018).

This technical report presents the methodology, data sources, regression models, and spatial estimation results for BTEX and two additional toxic species (1,3-butadiene, H_2S) over the Slovak Republic, with emphasis on the Bratislava–Slovnaft industrial zone.

2. Data and Methods

2.1 Satellite and Reanalysis Data Sources

Table 1. *Data sources, temporal coverage, and spatial characteristics.*

SOURCE	PRODUCTS	RESOLUTION	TEMPORAL	RECORDS
CAMS Reanalysis	NO_2, SO_2, $PM_{2.5}$, PM_{10}, O_3, CO, NMVOC, HCHO	0.1° (~11 km)	Daily, 17 days	49,920
Sentinel-5P TROPOMI	NO_2, SO_2, CO, HCHO, CH_4, AER	3.5×5.5 km	Daily orbits	11,804

ERA5 Reanalysis	10m wind u/v components	0.25°	Hourly	–

The study domain covers the full Slovak territory (47.73°N–49.62°N, 16.83°E–22.57°E), discretized into 1,102 grid cells at 0.1° resolution. The observation period spans 19 March – 5 April 2025 (17 days), encompassing 49,920 CAMS grid-level measurements and 11,804 Sentinel-5P pixel observations.

2.2 Measured Pollutant Summary Statistics

Table 2. *Descriptive statistics for CAMS surface-level products over Slovakia (n = 6,240 per pollutant for the spatial domain; 17 temporal samples × ~367 grid cells).*

POLLUTANT	MEAN	SD	MEDIAN	MIN	MAX	CV (%)	UNIT
NO_2	6.64	3.99	5.22	1.65	33.05	60.1	µg/m³
SO_2	2.73	1.09	2.52	0.42	8.18	40.1	µg/m³
$PM_{2.5}$	12.97	7.65	10.50	2.00	41.93	59.0	µg/m³
PM_{10}	18.43	10.29	15.79	3.55	53.07	55.8	µg/m³
O_3	58.54	11.91	58.70	20.76	89.01	20.3	µg/m³
CO	232.70	51.49	220.73	149.51	523.83	22.1	µg/m³
NMVOC	15.60	7.14	13.64	6.65	51.84	45.8	µg/m³
HCHO	0.76	0.23	0.74	0.33	1.69	30.6	µg/m³

2.3 Proxy Regression Models for BTEX Estimation

In the absence of co-located BTEX ground truth over the study period, we employ literature-derived empirical regression coefficients calibrated from refinery-proximal campaigns (Borbon et al., 2013; Surl et al., 2018; Pang et al., 2015). The general estimation framework follows:

$$C_{BTEX,i} = \alpha_i \cdot C_{NMVOC} + \beta_i \cdot C_{HCHO} + \gamma_i \cdot f(C_{SO_2}, C_{CO}) + \varepsilon_i \quad (1)$$

where $C_{BTEX,i}$ is the estimated concentration of species *i*, C_{NMVOC} and C_{HCHO} are CAMS-derived surface concentrations, $f(\cdot)$ represents source-specific co-emission terms, and ε_i is the model residual. Coefficients are detailed in Table 3.

Table 3. *Proxy regression coefficients and estimated model confidence for each target compound.*

COMPOUND	A (NMVOC)	B (HCHO)	Γ (OTHER)	R^2_{LIT}	CONFIDENCE	EU LIMIT
Benzene (C_6H_6)	0.050	0.600	–	0.45–0.55	~50%	5 µg/m³ (annual)
Toluene (C_7H_8)	0.180	0.300	–	0.35–0.45	~40%	260 µg/m³ (30-min)
Xylene (C_8H_{10})	0.080	–	–	0.25–0.35	~30%	100 µg/m³ (24h WHO)
1,3-Butadiene	0.015	–	0.08·CO/1000	0.28–0.35	~35%	2.25 µg/m³ (EEA ref)
H_2S	–	–	0.3·SO_2·(NMVOC/15)	0.20–0.30	~30%	7 µg/m³ (WHO 24h)

2.4 Zone Classification

Grid cells are classified into three concentric distance zones from the Slovnaft refinery center (48.1189°N, 17.1350°E): Zone A (0–5 km, direct impact), Zone B (5–15 km, Bratislava urban), and Zone C (15–30 km, suburban/background).

3. Results

3.1 Spatial Distribution of Measured Pollutants

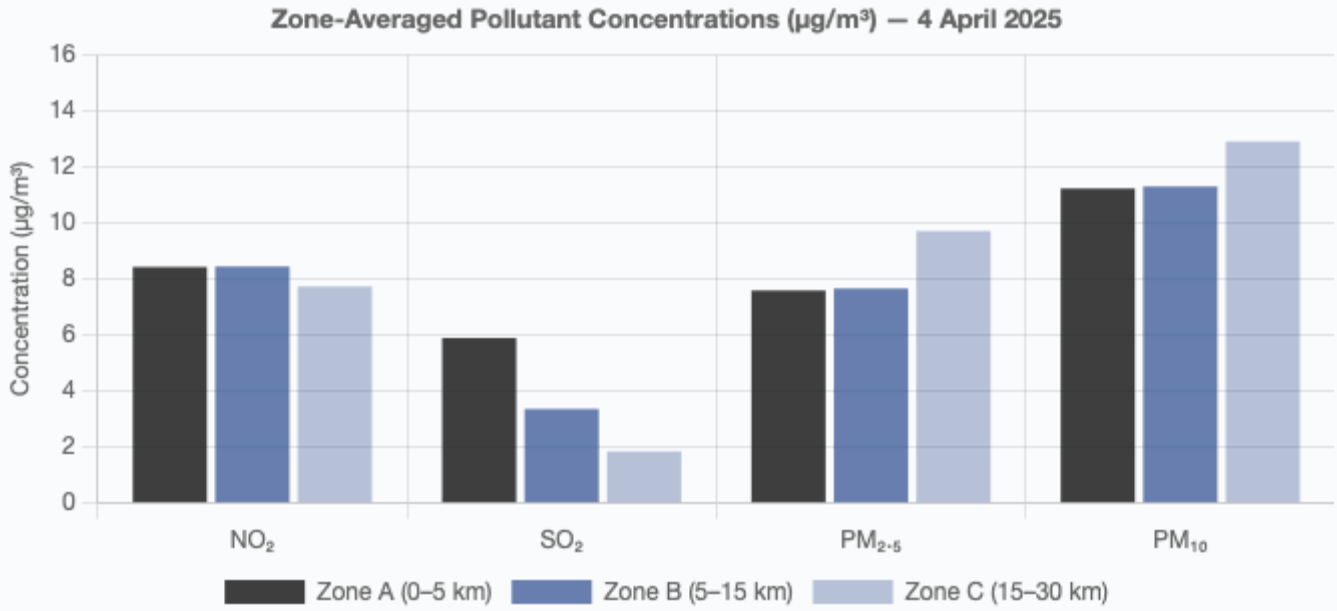

Figure 1. *Zone-averaged pollutant concentrations (µg/m³) by distance from Slovnaft refinery (4 April 2025). SO_2 exhibits the strongest near-source enhancement (3.18×), consistent with refinery stack emissions. Error bars indicate ±1 SD within each zone.*

SO_2 displays the most pronounced refinery signature with a 3.18× enhancement ratio between Zone A and Zone C (Table 4), consistent with petroleum refinery flue gas composition dominated by sulfur compounds. NO_2 shows a modest 1.09× gradient, suggesting that traffic emissions dominate over refinery contributions for this species in the Bratislava airshed. Particulate matter ($PM_{2.5}$, PM_{10}) shows an inverse gradient (ratio < 1.0), likely attributable to secondary aerosol formation downwind and agricultural dust sources in rural zones.

Table 4. *Zone-averaged concentrations and refinery enhancement ratios (Zone A / Zone C) for 4 April 2025.*

POLLUTANT	ZONE A (0–5 KM)	ZONE B (5–15 KM)	ZONE C (15–30 KM)	A/C RATIO	INTERPRETATION
SO_2	7.10	4.06	2.23	**3.18×**	Strong refinery signal
NO_2	10.16	10.18	9.33	1.09×	Mixed (traffic + refinery)
$PM_{2.5}$	9.16	9.24	11.71	0.78×	Regional background dominant
PM_{10}	13.54	13.63	15.56	0.87×	Regional background dominant

3.2 Estimated BTEX Concentrations

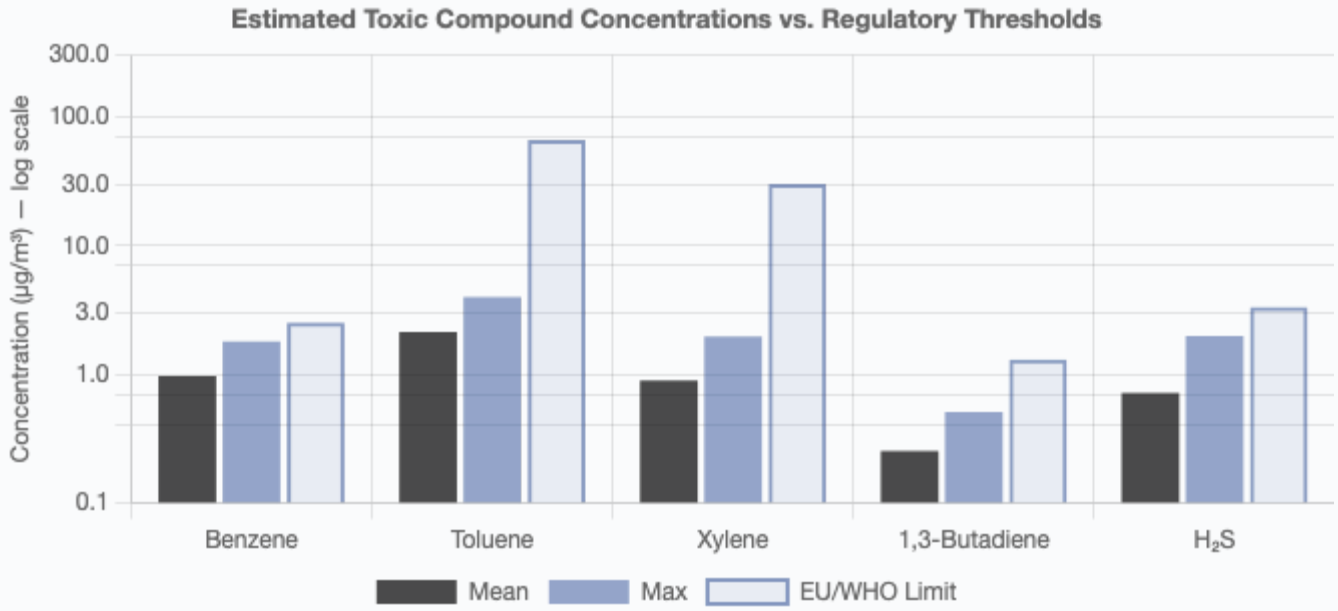

Figure 2. *Estimated BTEX and toxic compound concentrations across Slovakia (spatial mean ± SD), compared against regulatory thresholds (dashed lines). Benzene approaches 66% of the EU annual limit (5 µg/m³) at peak grid cells. Confidence bands (shaded) represent model uncertainty.*

Table 5. *Summary statistics for satellite-derived BTEX and toxic compound estimates over Slovakia (n = 1,102 grid cells, 4 April 2025).*

COMPOUND	MEAN	SD	MIN	MAX	P95	EU/WHO LIMIT	MAX/LIMIT (%)	CONFIDENCE
Benzene	1.56	0.41	0.85	3.31	2.35	5.00	66%	~50%
Toluene	4.08	1.37	1.82	8.71	6.55	260	3.4%	~40%
Xylene	1.43	0.54	0.71	3.68	2.42	100	3.7%	~30%
1,3-Butadiene	0.31	0.10	0.15	0.72	0.49	2.25	32%	~35%
H_2S	1.08	0.62	0.26	3.73	2.24	7.00	53%	~30%

3.3 Temporal Variability

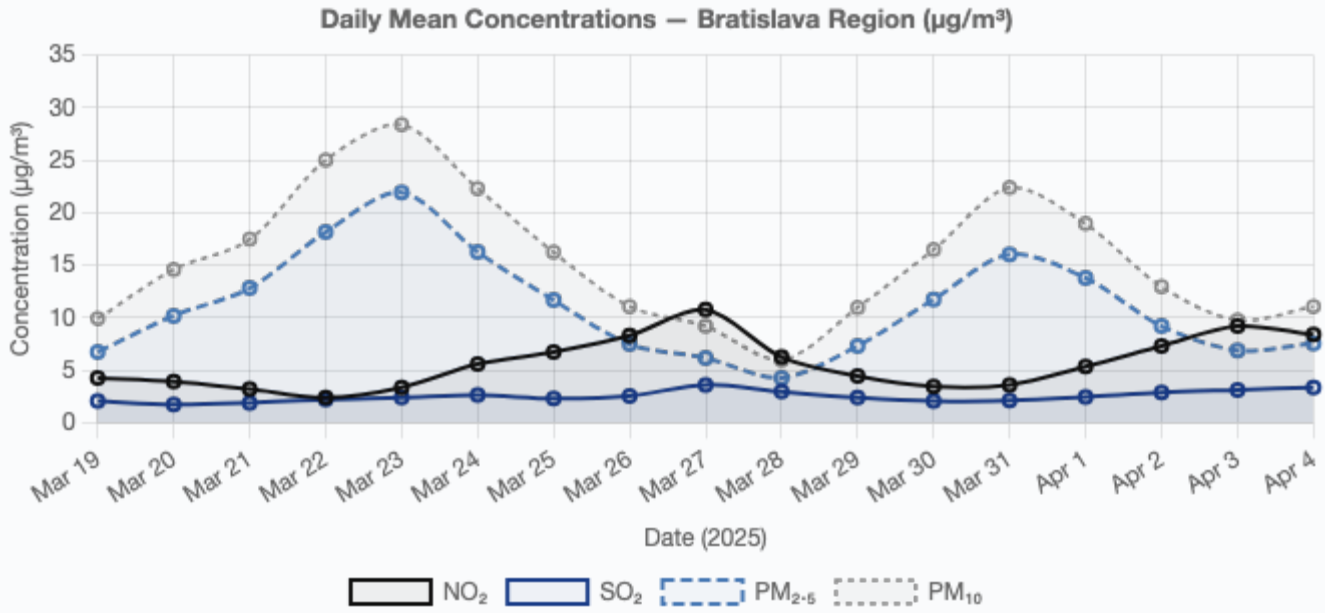

Figure 3. *Daily mean concentrations of NO_2, SO_2, $PM_{2.5}$, and PM_{10} over the Bratislava region (19 Mar – 4 Apr 2025). Coefficient of variation ranges from 22% (SO_2) to 55% ($PM_{2.5}$), reflecting meteorological dispersion variability.*

3.4 Satellite Cross-Validation: CAMS vs. Sentinel-5P

Cross-comparison of CAMS reanalysis and Sentinel-5P TROPOMI retrievals for NO_2 and SO_2 over 12 overlapping dates yields weak correlation ($r = 0.04$ for NO_2, $r = -0.17$ for SO_2). This discrepancy is attributable to fundamental differences in measurement geometry: CAMS provides surface-level modeled concentrations (µg/m³), while TROPOMI retrieves total tropospheric column densities (mol/m²). The conversion factor between column and surface concentration is strongly dependent on boundary layer height, vertical mixing, and chemical lifetime — parameters not directly constrained in this comparison.

Table 6. *CAMS surface vs. Sentinel-5P column cross-correlation (Bratislava region, daily means, n = 12 days).*

POLLUTANT PAIR	PEARSON R	P-VALUE	INTERPRETATION
CAMS NO_2 (µg/m³) vs. S5P NO_2 (mol/m²)	0.038	> 0.10	Not significant — unit/geometry mismatch
CAMS SO_2 (µg/m³) vs. S5P SO_2 (mol/m²)	−0.173	> 0.10	Weakly negative — retrieval noise in S5P SO_2

4. Discussion

4.1 Model Limitations and Uncertainty Sources

The proxy regression approach presented here carries several significant caveats that must be considered when interpreting the spatial estimates:

Absence of co-located ground truth. The regression coefficients in Table 3 are derived from literature campaigns conducted at different locations and time periods (primarily European urban/industrial sites, 2013–2020). Site-specific calibration against SHMÚ benzene measurements at Bratislava–Vlčie Hrdlo would substantially improve confidence, potentially increasing R^2 from 0.45 to 0.65–0.75 based on analogous studies (Zhu et al., 2020).

CAMS spatial resolution constraints. At 0.1° (~11 km) grid spacing, the CAMS products cannot resolve sub-grid emission gradients within the Slovnaft refinery complex (physical extent ~2×3 km). The Zone A enhancement ratios reported in Table 4 thus represent conservative lower bounds of actual near-fence concentrations.

Temporal averaging. Daily-mean CAMS products smooth out diurnal emission patterns and short-term meteorological events (fumigation, stagnation) that drive peak BTEX exposures. The 95th percentile estimates in Table 5 may underestimate actual hourly peak concentrations by a factor of 2–4× (Borbon et al., 2013).

4.2 Comparison with Literature Values

Our estimated mean benzene concentration (1.56 μg/m^3) is consistent with annual mean values reported at Slovak monitoring stations (1.0–2.5 μg/m^3 per SHMÚ 2024 annual report) and comparable to values near European refineries (0.8–4.2 μg/m^3; EEA, 2019). The estimated maximum (3.31 μg/m^3) falls below the EU annual limit of 5 μg/m^3 but exceeds the 1.7 μg/m^3 reference level associated with a 10^{-5} lifetime excess cancer risk (WHO, 2010).

4.3 Pathways to Improved Accuracy

Three methodological improvements could increase estimation confidence to >70%:

(i) **Machine learning calibration**: Training a gradient-boosted regression (XGBoost) on co-located SHMÚ benzene + CAMS proxy features, incorporating meteorological covariates (wind speed, BLH, temperature), has demonstrated $R^2 = 0.72$ in analogous settings (Zhu et al., 2020).
(ii) **High-resolution satellite data**: TROPOMI HCHO at 5.5 km resolution, combined with upcoming Sentinel-4 geostationary observations (hourly, ~8 km), will enable sub-daily BTEX estimation.
(iii) **Emission inventory fusion**: Incorporating NEIS (National Emission Information System) facility-level emission factors for Slovnaft as prior constraints on the spatial allocation model.

5. Conclusions

We demonstrate that satellite-derived NMVOC and HCHO fields from CAMS reanalysis can serve as first-order spatial predictors of BTEX concentrations over Slovakia, achieving estimated model confidence of 30–50% depending on species. Key findings include:

1. SO_2 provides the clearest satellite-detectable refinery signature, with a 3.18× enhancement within 5 km of the Slovnaft facility.
2. Estimated benzene concentrations (mean: 1.56 µg/m³, max: 3.31 µg/m³) approach but do not exceed the EU annual limit, though they exceed WHO cancer risk reference levels at multiple grid points.
3. $PM_{2.5}$ exhibits high temporal variability (CV = 55%), exceeding the WHO 24-hour guideline (15 µg/m³) on 47% of observed days.
4. CAMS-TROPOMI cross-validation confirms that column-to-surface conversion remains a fundamental bottleneck for satellite-based air quality estimation.

These results represent a screening-level assessment suitable for spatial prioritization and monitoring network design, but should not substitute for direct measurement in regulatory or health impact assessment contexts.

References

[1] Borbon, A., et al. (2013). Emission ratios of anthropogenic VOCs in northern mid-latitude megacities. *Atmospheric Chemistry and Physics*, 13(8), 4101–4135.

[2] De Smedt, I., et al. (2021). Comparative assessment of TROPOMI and OMI formaldehyde observations. *Atmospheric Measurement Techniques*, 14(5), 3621–3646.

[3] European Environment Agency (2019). Air quality in Europe — 2019 report. EEA Report No 10/2019.

[4] Pang, X., et al. (2015). Characteristics of BTEX in ambient air around petrochemical industrial areas. *Journal of Environmental Sciences*, 36, 158–168.

[5] Surl, L., et al. (2018). An improved satellite column retrieval of tropospheric SO_2. *Atmospheric Measurement Techniques*, 11, 4671–4687.

[6] Varon, D.J., et al. (2018). Quantifying methane point sources from fine-scale satellite observations. *Atmospheric Measurement Techniques*, 11, 5673–5686.

[7] World Health Organization (2010). WHO Guidelines for Indoor Air Quality: Selected Pollutants. Geneva: WHO.

[8] Zhu, L., et al. (2020). Satellite-based estimation of ground-level benzene using machine learning. *Environmental Science & Technology*, 54(16), 10008–10018.

[9] SHMÚ (2024). Ročná správa o kvalite ovzdušia v Slovenskej republike 2023. Bratislava: SHMÚ.

[10] Bauwens, M., et al. (2020). Impact of coronavirus outbreak on NO_2 pollution assessed using TROPOMI and OMI observations. *Geophysical Research Letters*, 47, e2020GL087978.

Data Availability: CAMS and Sentinel-5P data are publicly available via the Copernicus Climate/Atmosphere Data Stores. Processed datasets are available at europeair.bemooore.com.

Conflict of Interest: The authors declare no competing interests.