SASYR Symposium of Applied Science for Young Researchers

5th Symposium of Applied Science for Young Researchers

PROCEEDINGS 2025

July 2 , 2025

5th Symposium of Applied Science for Young Researchers

Proceedings

SASYR 2025

2 July 2025

CeDRI
Research Centre in Digitalization and Intelligent Robotics

2Ai APPLIED ARTIFICIAL INTELLIGENCE LABORATORY

gecad
Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development

ipvc adit-lab
Applied Digital Transformation Laboratory
Instituto Politécnico de Viana do Castelo

ipb INSTITUTO POLITÉCNICO DE BRAGANÇA

IPCA POLITÉCNICO DO CÁVADO E DO AVE

P.PORTO

ipvc Instituto Politécnico de Viana do Castelo

Editors

Florbela P. Fernandes
Research Centre in Digitalization and Intelligent Robotics (CeDRI)
Instituto Politécnico de Bragança

Helena Torres
Applied Artificial Intelligence Laboratory (2Ai)
Instituto Politécnico do Cávado e do Ave

Pedro Pinto
Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development (GECAD)
Instituto Politécnico do Porto

Silvestre Malta
Applied Digital Transformation Laboratory (ADiT-LAB)
Instituto Politécnico de Viana do Castelo

Instituto Politécnico de Bragança – 2025
Campus de Santa Apolónia
5300-253 Bragança – Portugal
ISBN: 978-972-745-360-3

Book cover: Natália Santos, Instituto Politécnico do Cávado e do Ave

Exploratory Data Analysis and Insights on Volatile Organic Compounds for Hazardous Waste Detection

Mahdia Ahmadi[1], Natalia Méndez Pérez[1,2], Helena Cristina Almeida da Cruz[3], Getúlio Igrejas[1], Pedro João Rodrigues[1], and Rui Pedro Lopes[1]

[1] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança (IPB), Bragança, Portugal
{mahdia, igrejas, pjsr, rlopes}@ipb.pt
[2] Univerdidad de Lalaguna, Tenerife, Spain
lu0101487038@ull.edu.es
[3] Instituto Politécnico de Coimbra, Coimbra, Portugal
cac.helena12@gmail.com

Abstract. Volatile Organic Compounds (VOCs) are critical indicators of environmental contamination, particularly in hazardous waste contexts. While gas chromatography–mass spectrometry (GC–MS) provides high specificity, it struggles with scalability and pattern discovery in large, complex datasets. This study presents a data-driven framework integrating Exploratory Data Analysis (EDA) techniques — including principal component analysis (PCA), hierarchical clustering, and correlation mapping — to uncover emission patterns in compost-derived VOC data.

Using the LCSC VOC 2022 Compost Dataset (141 variables, 90 samples), we identified strong co-emission clusters (e.g., D-Limonene and α-Pinene) and a temperature-dependent ethanol emission pattern unique to food-and-yard waste samples. Pearson correlation analysis revealed shared emission behavior, and regression confirmed a positive slope (25.6) for ethanol versus temperature.

These findings highlight EDA's potential to enhance VOC dataset interpretability and source identification. The proposed framework supports practical applications such as early-warning systems, sensor deployment, and data-informed environmental policy.

Keywords: Hazardous Waste, Pollution Source Identification, Data-Driven Decision Making, Machine Learning, Sensors.

1 Introduction

Volatile Organic Compounds (VOCs) are a broad class of carbon-based chemicals that readily vaporize at ambient temperatures. They are emitted from a wide range of natural and anthropogenic sources, including industrial activities, landfills, and composting systems. The presence of VOCs in the environment is of increasing concern due to their adverse impacts on both ecological systems and human health. Studies have established links between long-term exposure to VOCs and serious health conditions such as chronic respiratory illnesses, skin cancer, and neurological disorders [6].

To monitor VOCs, gas chromatography-mass spectrometry (GC–MS) has long been considered the gold standard, offering high specificity and sensitivity. However, the growing complexity and dimensionality of environmental data pose serious challenges for traditional techniques. Specifically, GC-MS struggles with scaling to

large datasets, lacks real-time analytical capacity, and is limited in its ability to differentiate between complex emission sources in mixed environments [1]. These shortcomings are particularly evident in hazardous waste contexts, where data is often heterogeneous, poorly labeled, and influenced by dynamic environmental variables.

In response to these limitations, data-driven analytical frameworks are gaining traction in environmental monitoring. Among these, Exploratory Data Analysis (EDA) offers a compelling approach. EDA is a flexible, assumption-free method that leverages statistical visualization and unsupervised learning to reveal underlying structures, correlations, and outliers within datasets [5]. Its non-parametric nature makes it especially useful for VOC analysis, where datasets are multidimensional, noisy, and lack uniform labels. In biomedical and environmental contexts, EDA has shown promise in applications ranging from air quality modeling to disease biomarker detection via VOCs [10].

In this work, we apply EDA techniques — including principal component analysis (PCA), hierarchical clustering, and correlation mapping — to a real-world VOC dataset obtained from composting systems. Unlike prior studies that focus solely on compound identification or source quantification, our approach integrates chemical analysis with pattern discovery to uncover emission trends, temperature-driven behavior, and co-emission clusters. Our main contribution lies in demonstrating how EDA can supplement traditional chemical monitoring, enabling scalable, interpretable, and actionable insights that are critical for hazardous waste management.

The rest of this paper is structured as follows. Section 2 discusses relevant background literature and related work. Section 3 outlines the methodological framework and dataset selection process. Section 4 presents key findings from the data analysis, while Section 5 concludes the study with insights and future research directions.

2 Background and related work

Volatile Organic Compounds (VOCs) are a well-established environmental and public health concern due to their role in air pollution and potential toxic effects on humans. VOCs originate from various anthropogenic sources such as petroleum refineries, landfills, and industrial waste processes, as well as from natural biological emissions [3] [10].

In recent years, there has been growing attention on the health implications of chronic VOC exposure. For instance, large-scale studies have revealed associations between long-term VOC exposure and increased risks of respiratory diseases and skin cancer, especially in vulnerable populations. These findings underscore the importance of precise and scalable VOC detection techniques, particularly in sensitive environments like hazardous waste sites.

Conventional detection methods such as gas chromatography-mass spectrometry (GC-MS) offer high analytical specificity, but they are resource-intensive, slow, and poorly suited for high-throughput or real-time applications. More importantly, they fall short when interpreting complex, heterogeneous, and multidimensional datasets,

which are often the norm in environmental monitoring contexts involving mixed waste types, variable emission profiles, and fluctuating environmental conditions [1].

To address these limitations, data-driven methods, including machine learning and statistical modeling, have been increasingly explored. For example, VOC-based classification has been used for food quality inspection [7], tracing coffee origin through network analysis [8], and disease detection via exhaled breath [5]. However, these methods often require labeled data, clear training objectives, or domain-specific features, which may not always be available in hazardous waste settings.

This creates an important gap — the limited use of Exploratory Data Analysis (EDA) in VOC research, especially in unstructured or poorly labeled datasets. Unlike predictive models, EDA does not assume prior distributions or require labels, making it well-suited for uncovering hidden structures, correlations, and co-emission patterns in environmental VOC data. Additionally, EDA allows for early insight extraction and hypothesis generation without committing to predefined models, an advantage when dealing with novel or dynamic emission environments like composting systems and hazardous waste facilities.

The main challenges include: **data heterogeneity** (varying VOC profiles by source and interaction), **sparse labeling** (limited annotations on emission origins), **environmental noise** (temperature and humidity confound analysis), and **high dimensionality** (numerous simultaneous variables). Despite these challenges, few studies have applied EDA as a central methodology in VOC analysis. This paper aims to fill that gap by demonstrating how EDA — through methods such as PCA, hierarchical clustering, and correlation analysis — can generate interpretable insights into VOC behavior, especially in compost-derived air samples. Our contribution lies in showing that EDA not only reveals complex compound interdependencies, but also aids in identifying source-specific emissions and conditions that affect their distribution.

3 Methodology

This study adopted an Exploratory Data Analysis (EDA) approach with the primary objective of examining and understanding the structure and content of multiple datasets containing information on VOC emissions.

EDA techniques help to provide an understanding of data, without requiring the application of formal statistical procedures or the prior definition of assumptions about the data at hand. This methodology involves the use of graphical and non-graphical methods, such as descriptive statistics, to facilitate a comprehensive understanding of the data structure, quality, and relationship between its variables, among others. Therefore, it serves as a crucial initial step in the data analysis process [5].

At the initial stage of this research, three datasets were explored: the LCSC VOC Compost Dataset 2022 [4], Long-term variations of ambient VOCs [9], and Experiments on VOC uptake by the active layer soils of Greenlandic permafrost areas [2]. However, only the first was selected for analysis. The excluded datasets were not directly relevant to the study's objectives, one focused on urban air trends, and the other on VOC absorption in Arctic soils, neither directly related to the study's focus on VOC emissions

from hazardous waste sources. In contrast, the chosen dataset offers detailed, time-resolved VOC measurements from composting activities in a hazardous waste context, making it the most suitable for identifying emission patterns, clustering behavior, and contamination sources.

4 Dataset Analysis

The dataset analyzed in this study, the LCSC VOC 2022 Compost Dataset, was developed by the Lewis-Clark State College (LCSC) Air Research Group, led by Dr. Nancy A. C. Johnston. It is part of a broader NIH-funded project supported by the Institutional Development Award (IDeA) from the National Institute of General Medical Sciences in partnership with LCSC. The data was collected at the Washington State University Compost Facility to investigate the VOCs emitted from compost under different conditions.

Data collection occurred from July to September 2022, using high-resolution sampling intervals. The dataset comprises 90 samples: 84 combined air and water samples, 4 air-only samples, and 2 water-only samples. In total, 141 variables were measured, including chemical, environmental, and sampling-related parameters.

VOC concentrations are reported in parts per billion by volume (ppbv), with a measurement uncertainty of ±10%. The VOC concentrations in air samples were measured using thermal desorption tubes with a Markes-Agilent TD-GC-MS system, while water-phase VOCs were captured using impinger sampling and analyzed with an Agilent HS-FID-GC system.

This study primarily focuses on the analysis of air samples, as they are most relevant to our ongoing and future research. A total of 88 air samples were collected.

4.1 Sample Conditions Analysis

Table 1 presents the descriptive statistics of the air sample variables, while Fig. 1 illustrates the origin of the compost samples and their specific locations within the pile.

Table 1. Descriptive statistics for environmental variables in air samples

	Pile Temp. [ºC]	**Outside Temp. [ºC]**	**Humidity [%]**	**Pressure [atm]**	**Wind Speed [m/s]**
mean	50.35	28.22	26.32	0.91	3.98
std	17.60	4.35	10.75	0.002	1.65
min	22.22	21.70	10	0.90	0.45
25%	34.72	22.68	20	0.90	3.13
50%	55	28.75	20	0.91	4.02
75%	65.56	31.63	36.25	0.91	4.92
max	88.33	38	49	0.91	7.15

The average pile temperature (around 50°C) aligns with expected thermophilic composting conditions, while outside temperature, humidity, pressure and wind speed values fall within typical ranges for outdoor composting in warm environments.

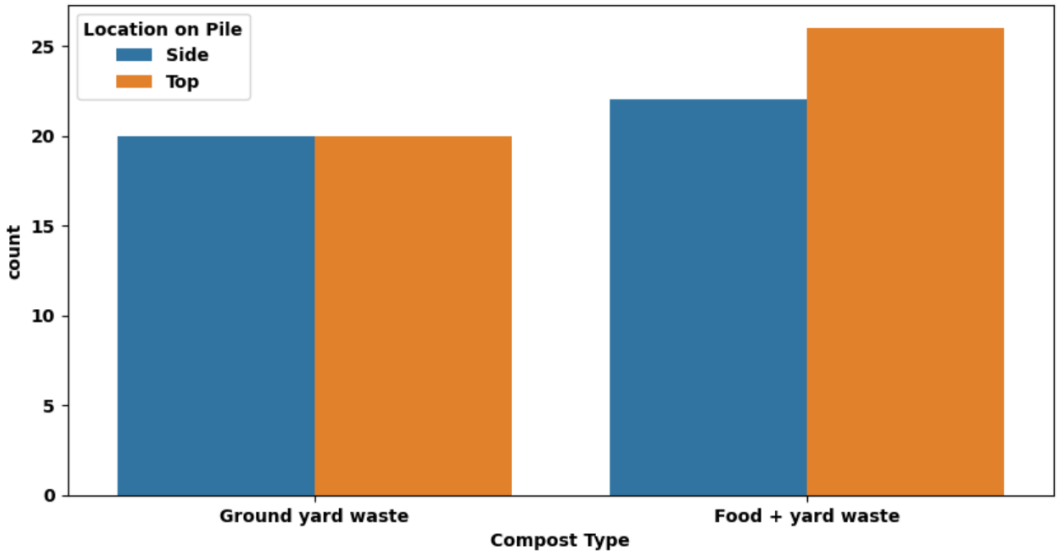

Fig. 1. Air sample frequency by compost type and pile location

Approximately half of the samples were taken from ground yard waste, while the other half corresponded to a mix of food and yard waste. Within each group, the samples are evenly distributed between the top and the side of the pile, ensuring a balanced representation of sampling locations across waste types.

4.2 VOCs concentration

Fig. 2 presents the top ten VOCs based on their average concentration (ppbv) in the samples. Fig. 3, on the other hand, illustrates the Pearson correlation coefficients observed between these VOCs.

Strong correlations are observed between the following VOC pairs: D-Limonene and α-Pinene; α-Pinene and β-Pinene; β-Pinene and γ-Terpinene; β-Pinene and Sabinene; γ-Terpinene and Sabinene; Camphor and L-Fenchone; Camphor and α-Humulene; and α-Humulene and L-Fenchone.

4.3 Specific Analysis for Ethanol

Fig. 4 plots ethanol concentration (the VOC with the highest average concentration) against the pile temperature, with data distinguished by waste type. A regression line is also shown to illustrate the overall trend for each compost type.

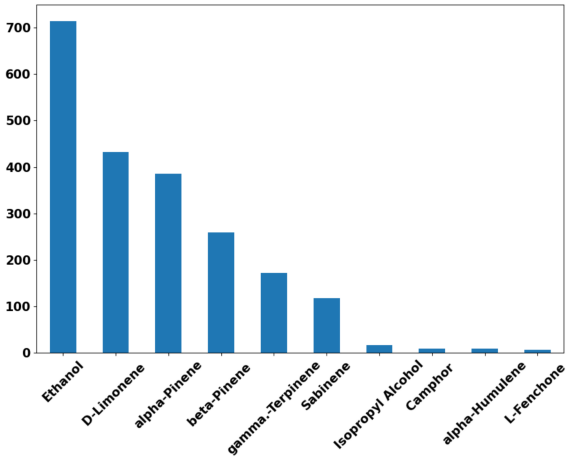

Fig. 2. Top ten VOCs by average concentration

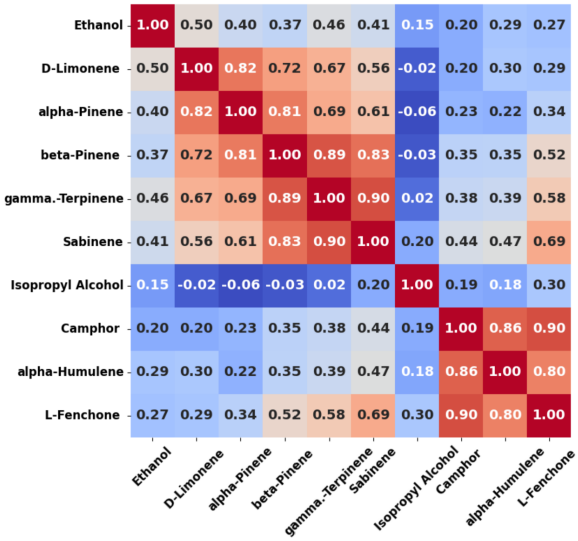

Fig. 3. Correlation between top 10 VOCs

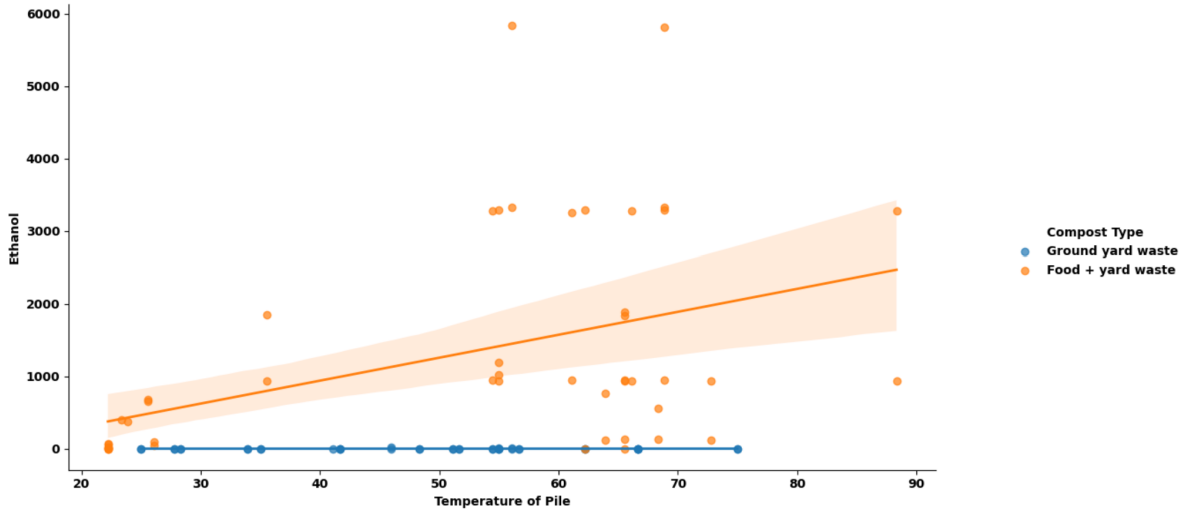

Fig. 4. Relationship between ethanol concentration and pile temperature by waste type, with fitted regression lines

Ethanol concentration shows a slight positive correlation with temperature, with a regression slope of 25.6292. Additionally, ethanol was not detected in ground waste samples, indicating no emissions from this type.

5 Results and Findings

The exploratory data analysis provided several noteworthy insights into the behavior of volatile organic compounds (VOCs) within compost environments. By focusing on a carefully selected dataset — comprising air samples collected from

compost piles of varying composition and sampling locations — key environmental patterns and compound relationships were uncovered.

Descriptive statistics highlighted distinct differences between internal and external environmental conditions. The pile temperature exhibited a broad range, with higher variability compared to the more stable external temperature. This internal heat may influence the microbial activity responsible for VOC production. Humidity and wind speed showed moderate fluctuations, while atmospheric pressure remained nearly constant, suggesting limited relevance to VOC dynamics in this context.

The dataset revealed a diverse VOC profile. Among the top 10 compounds identified by average concentration, ethanol, D-Limonene, and α-Pinene were particularly prominent. These compounds are commonly associated with microbial fermentation and the degradation of plant-based organic matter, indicating active biological decomposition within the compost.

A closer examination of VOC relationships through correlation analysis showed strong co-emission patterns between specific compounds, such as D-Limonene and α-Pinene, and β-Pinene with γ-Terpinene and Sabinene. These correlations suggest shared emission sources, likely tied to the decomposition of similar organic substrates, such as citrus residues or terpene-rich plant matter.

Ethanol emerged as the compound with the highest overall concentration. Its levels were found to increase with rising pile temperatures, hinting at a temperature-dependent fermentation process. Notably, ethanol was absent in samples from ground yard waste, while present in mixed food-and-yard waste piles. This suggests that food waste is the primary contributor to ethanol emissions, and that different waste compositions may lead to distinct VOC signatures.

The dataset's design ensured balanced spatial sampling across pile sides and tops. No significant concentration differences were observed based solely on sample location, reinforcing the idea that waste type and internal conditions are stronger drivers of VOC behavior than sampling orientation.

6 Conclusion and Future Work

This study applied exploratory data analysis (EDA) to a compost VOC dataset, yielding meaningful insights into emission patterns. Ethanol and several terpenes emerged as dominant compounds, particularly in mixed food-and-yard waste samples. Temperature and waste composition were found to significantly influence VOC levels.

Strong correlations among specific VOCs suggested shared sources, offering potential for simplified monitoring using key indicator compounds. The findings support the value of data-driven methods in complementing traditional chemical analysis for environmental monitoring.

Future work will expand the analysis using diverse datasets across seasons, compost types, and locations, while integrating machine learning and real-time sensors to improve detection and decision-making in hazardous waste environments.

Acknowledgment

This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope: 2024.07316.IACDC/2024.

References

1. Capitain, C., Weller, P.: Non-targeted screening approaches for profiling of volatile organic compounds based on gas chromatography-ion mobility spectroscopy (gc-ims). Molecules **26**(18), 5457 (2021)
2. Jiao, Y., Kramshøj, M., Davie-Martin, C., Elberling, B., Rinnan, R.: Dataset: experiments on volatile organic compounds uptake by the active layer soils of greenlandic permafrost areas (Nov 2024)
3. Jindamanee, K., Keawboonchu, J., Pinthong, N., Meeyai, A., Inchai, P., Thepanondh, S.: Environmental impacts and emission profiles of volatile organic compounds from petroleum refineries. Scientific Reports **15**(1) (2025)
4. Johnston, N.: LCSC VOC Compost Dataset 2022 (2023), mendeley Data, V1
5. Komorowski, M., Marshall, D.C., Salciccioli, J.D., Crutain, Y.: Exploratory data analysis, pp. 185—203. Springer International Publishing (1 2016)
6. Nalini, M., Poustchi, H., Bhandari, D., Blount, B.C., Kenwood, B.M., Chang, C.M., Gross, A., Ellison, C., Khoshnia, M., Pourshams, A., Gail, M.H., Graubard, B.I., Dawsey, S.M., Kamangar, F., Boffetta, P., Brennan, P., Abnet, C.C., Malekzadeh, R., Freedman, N.D., Etemadi, A.: Exposure to volatile organic compounds and chronic respiratory disease mortality, a case-cohort study. Respiratory Research **26**(1) (2025)
7. Shtepliuk, I., Domènech-Gil, G., Almqvist, V., Kautto, A.H., Vågsholm, I., Boqvist, S., Eriksson, J., Puglisi, D.: Electronic nose and machine learning for modern meat inspection. Journal of Big Data **12**(1) (2025)
8. Taiti, C., Vivaldo, G., Mancuso, S., Comparini, D., Pandolfi, C.: Volatile organic compounds (vocs) fingerprinting combined with complex network analysis as a forecasting tool for tracing the origin and genetic lineage of arabica specialty coffees. Scientific Reports **15**(1) (2025)
9. Yafei, L., Chenlu, L., Xingang, L.: Long-term variations of ambient volatile organic compounds (vocs) from 2016 to 2020 in beijing, china (Jun 2023)
10. Yang, Y., Sun, F., Hu, C., Gao, J., Wang, W., Chen, Q., Ye, J.: Emissions of biogenic volatile organic compounds from plants: Impacts of air pollutants and environmental variables. Current Pollution Reports **11**(1) (2025)