Prediction of Black Carbon Concentrations and Potential Sources

Master's Research Project Report

Supervisors:

Prof. Mathieu Salzman
Dr. Ekaterina Krymova
Dr. Imad El Haddad

Student:

Youssef Belghmi

June 6, 2025

Contents

1 Introduction

1.1 Abstract

This project aims to improve the estimation and harmonization of Black Carbon (BC) and Elemental Carbon (EC) concentrations across Europe through advanced outlier detection, statistical analysis, and machine learning techniques. Black Carbon is a major component of fine particulate matter with significant health and climate impacts. However, its measurement is complex and varies depending on both the instruments and the protocols used, introducing inconsistencies in large-scale datasets. We work with over 300,000 measurements from more than 200 monitoring stations across Europe, combining optically measured BC and thermally measured EC.

Our approach involves cleaning and standardizing the raw data, detecting and removing outliers using both statistical and machine learning-based methods, and harmonizing BC and EC measurements by modeling their relationship across instruments and protocols. We also examine the influence of secondary pollutants (such as aerosols and dust) on measurement bias through regression and correlation analyses. This harmonization process provides a foundation for developing reliable predictive models of daily EC concentrations and for understanding systematic biases in BC measurements.

1.2 Motivation

Black Carbon (BC), a key component of fine particulate matter, has major health and climate impacts, contributing to over four million deaths annually through respiratory and cardiovascular diseases [1]. As a short-lived climate pollutant, it also plays a significant role in global warming. Recognizing its importance, institutions like the World Health Organization and the EU have labeled BC a pollutant of emerging concern [2].

Yet a critical barrier persists: the lack of harmonized, high-resolution data necessary to accurately model BC-related exposure at the continental scale. Variations in measurement instruments and protocols, along with the complexity of emission sources, limit the reliability of current models. This project addresses these challenges by applying data-driven methods to detect inconsistencies, harmonize multi-source measurements, and lay the groundwork for robust, large-scale predictive modeling.

1.3 Context

This project was carried out as part of the Master's Research Project at EPFL, a core component of the Data Science program designed to provide students with hands-on experience in conducting original research. It was conducted in collaboration with the Swiss Data Science Center (SDSC), which plays a key role in applying data science techniques to real-world scientific and societal challenges across Switzerland. The project was also supervised by domain experts at the Paul Scherrer Institute (PSI), the largest natural and engineering sciences research center in Switzerland, which provided access to high-resolution atmospheric data and deep expertise in environmental monitoring.

This report constitutes the written component of the project, summarizing the methodology, data analysis, modeling, and results. The full implementation is available at the following GitHub repository: https://github.com/youssefbelghmi/predict-black-carbon-concentrations-and-sources.git.

2 Dataset Description and Preparation

2.1 Overview of Data Sources

The dataset used in this project was provided by the Paul Scherrer Institute (PSI) and compiled through collaborative efforts involving multiple European environmental monitoring programs. This comprehensive collection of black carbon (BC) and elemental carbon (EC) measurements in Europe covers a wide temporal and spatial scale. The dataset includes 316,951 individual observations, each corresponding to a single atmospheric measurement of either BC or EC at a specific station on a given day. In total, the dataset spans 260 monitoring stations geographically distributed across Europe.

Each row in the dataset contains a variety of metadata fields that provide essential context for interpreting the carbon measurements. The most important columns include:

- **Station and Datetime:** identify the location and timestamp of the measurement. These two variables define the spatial and temporal resolution of the dataset and are critical for aligning and merging observations.
- **BC and EC:** represent the concentrations of Black Carbon and Elemental Carbon. Although often used interchangeably, BC and EC are conceptually distinct [3]:
  - **Black Carbon (BC)** refers to the fraction of carbonaceous particles that strongly absorb light, typically measured using optical methods. BC is often considered a proxy for soot, originating mainly from incomplete combustion of fossil fuels and biomass.
  - **Elemental Carbon (EC)** refers to the mass of carbonaceous material that is thermally stable and graphitic, measured using thermal-optical analysis. EC quantification is protocol-dependent and sensitive to temperature ramping and split-point criteria.

  While related, BC and EC are not directly equivalent, and their measured values can vary depending on the instrument and measurement protocol used.
- **BC Instrument and EC Instrument:** indicate the type of instrument used to measure BC or EC, respectively.
  - **BC Instrument** can be either MAAP (Multi-Angle Absorption Photometer) or AE (Aethalometer), both of which estimate BC concentration based on light absorption. MAAP provides a more accurate quantification by accounting for angular scattering, while AE is widely used for its operational simplicity.
  - **Instrument_EC** can be either Low Volume Filter or High Volume Filter, which refer to the sampling rate of the air intake system used for thermal-optical EC measurement. Low volume filters typically collect smaller sample volumes and are more suitable for long-term monitoring, whereas high volume filters collect larger samples, useful for detailed chemical analysis.

  In some cases, the instrument used is labeled as unknown, either because the information was not recorded or the metadata was incomplete. This introduces uncertainty and complicates the harmonization process.

- **Protocol:** indicates the thermal-optical protocol used to determine EC concentration. The main protocols include:
  - **NIOSH** (National Institute for Occupational Safety and Health) is a protocol that uses a defined temperature ramp in an inert atmosphere followed by an oxidizing phase, commonly applied in workplace exposure studies.
  - **EUSAAR2** (European Supersites for Atmospheric Aerosol Research) is an optimized protocol for ambient atmospheric measurements, with modifications in the temperature sequence and split-point definition.

In addition to these primary variables, the dataset includes several other columns that support more advanced analysis, such as:

- `OC` (Organic Carbon), `MAC` (Mass Absorption Cross-section), `Size_cut`, `C_value`, and `H_value` which relate to the physical and chemical properties of the particles or the sampling conditions.
- `Flag_MCH` that indicates the validity or quality control status of the measurement.

Overall, the dataset combines a rich diversity of measurements, protocols, and instruments, which provides an opportunity for large-scale modeling of carbonaceous aerosol concentrations. However, this diversity also introduces substantial challenges in terms of data quality, consistency, and harmonization, which we address in the subsequent sections.

2.2 Preprocessing and Cleaning Strategy

Due to the dataset's heterogeneity, spanning different instruments, protocols, units, and formats, we developed a modular preprocessing pipeline to standardize and clean the data. Raw files with inconsistencies were harmonized using the following steps:

- Renamed all columns to follow a consistent and descriptive naming convention,
- Converted all concentration values to nanograms per cubic meter (ng/m^3),
- Parsed date fields into standardized datetime objects,
- Ensured that instrument and protocol fields were correctly assigned to BC or EC,
- Encoded categorical variables (e.g., instrument, protocol) using standardized labels,
- Assigned a default placeholder for all missing or undocumented fields,
- Verified consistency (e.g., no lowercase/uppercase mix, invalid identifiers, etc.),
- Created new columns for the logarithmic transformation of BC and EC values and the computed EC/BC ratio for harmonization analysis.
- Rows with missing values in essential fields such as Datetime, BC or EC, and with obviously erroneous values (e.g., negative concentrations) were removed.

This process resulted in a cleaned and uniform dataset ready for exploratory analysis. The modular pipeline developed will facilitate future studies by streamlining data preparation for the team.

Before modeling, we analyzed the empirical distributions of BC and EC values by visualizing the distributions on both the original and logarithmic scales. In their raw form, both BC and EC concentrations exhibit a pronounced right skew, characterized by long tails extending toward higher values and a substantial presence of extreme outliers.

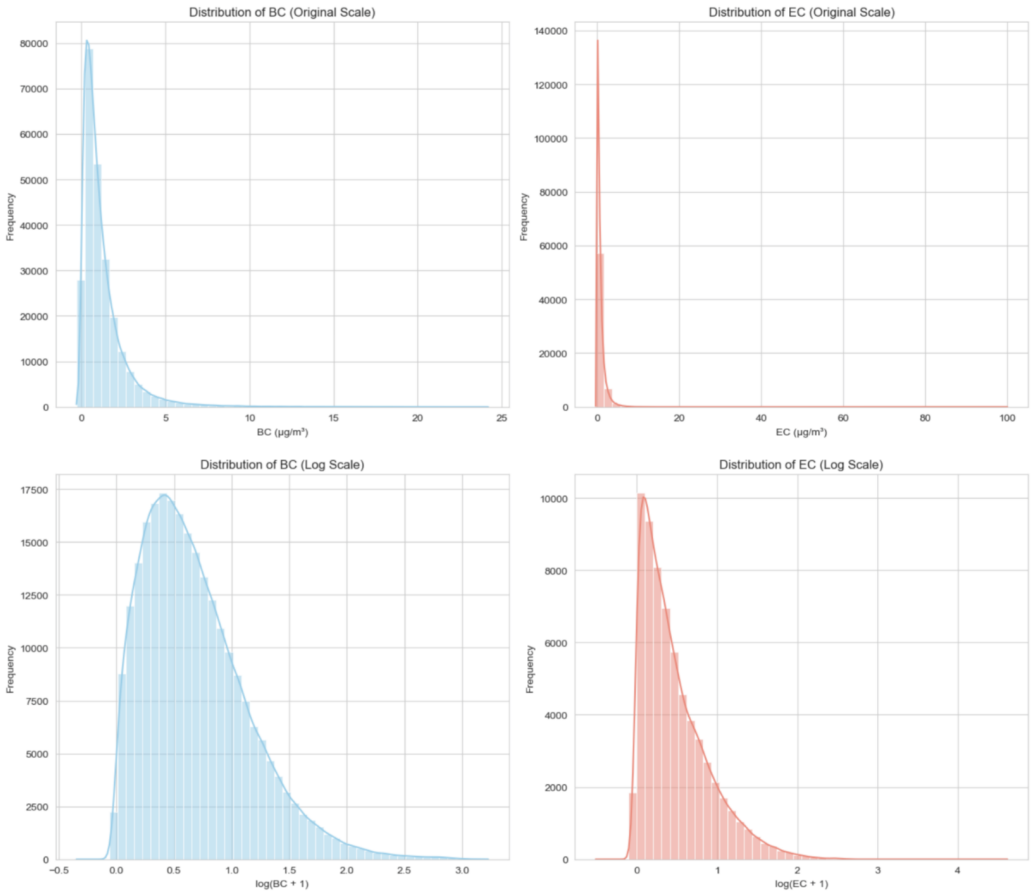

Figure 1: Distributions of BC and EC concentrations (original and log scale)

Such extreme values may result from instrumental errors, data reporting mistakes, or rare atmospheric conditions like pollution spikes. These anomalies can bias statistical models, especially those sensitive to distributional assumptions. To address this, we tested several statistical techniques for outlier detection, each offering different levels of sensitivity and robustness.

We first applied the Interquartile Range (IQR) method to the log-transformed BC and EC values [4]. The IQR is a robust univariate approach that defines outliers as observations falling outside the interval:

$$[Q1 - 1.5 \times IQR,\ Q3 + 1.5 \times IQR]$$

For BC, we obtained Q1 = 0.4017 and Q3 = 1.5814, resulting in an IQR of 1.1797 and bounds of approximately [-1.3680, 3.3510]. For EC, Q1 = 0.1621 and Q3 = 0.9900, giving an IQR of 0.8279 and bounds of [-1.0797, 2.2318]. Using these thresholds, we identified a total of 20,936 outlier rows across the dataset, corresponding to 5.07% of all BC values and 1.53% of EC values. These results are visualized in the figure below.

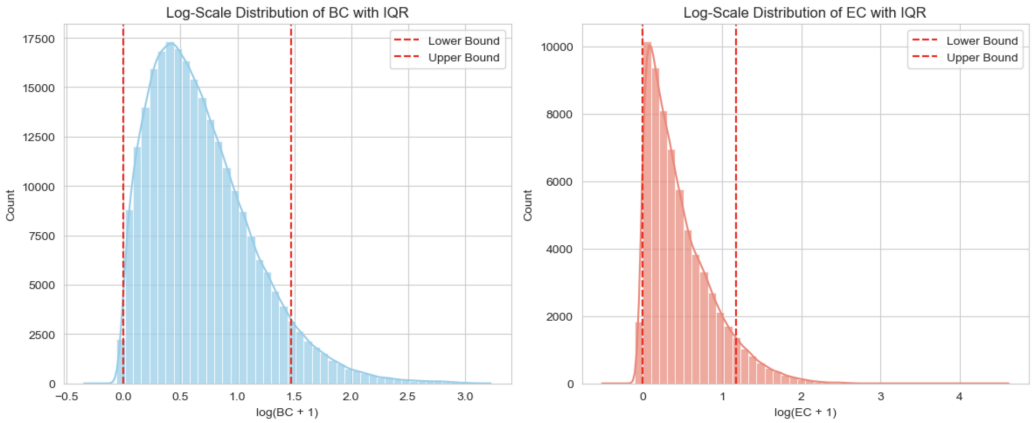

Figure 2: Log-scale distributions of BC and EC with IQR outlier bounds

While the method is intuitive and computationally simple, it can be overly aggressive for long-tailed environmental data. In particular, it may remove rare but valid high concentration events. Therefore, we tested the Isolation Forest algorithm [5], an unsupervised machine learning method that isolates anomalies via recursive data partitioning. The logic is that outliers require fewer splits to isolate due to their rarity and uniqueness.

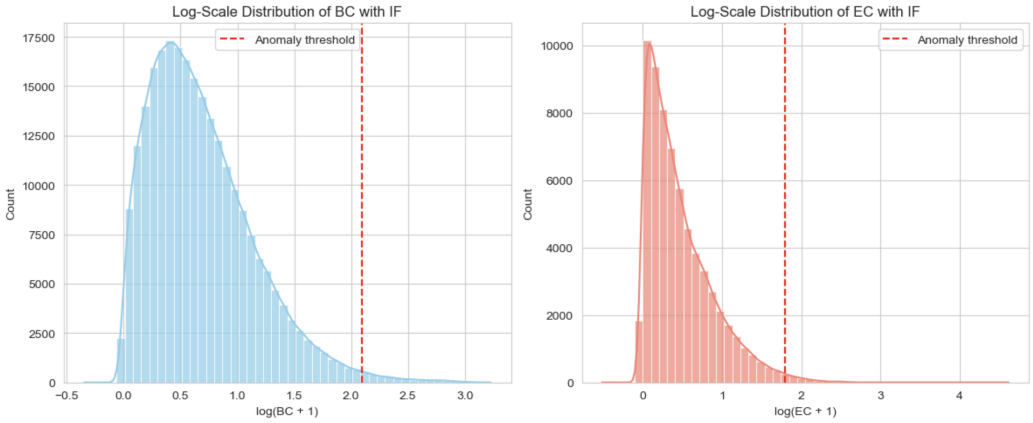

Figure 3: Log-scale distributions of BC and EC with IF anomaly threshold

Using the Isolation Forest algorithm on the log-transformed concentrations, we identified a total of 3,158 anomalous rows, which represents approximately 1.00% of the entire dataset. Compared to the IQR method, Isolation Forest is more conservative in its detection strategy and may overlook subtle anomalies embedded in complex patterns.

Thus, both IQR and IF have limitations: IQR may discard valid extremes in skewed distributions, while IF can miss subtle anomalies. Since preserving meaningful outliers is critical in environmental data, we adopted a more context-aware and interpretable method, quantile regression. This approach allowed us to finalize a clean and reliable dataset, ready for advanced modeling and harmonization tasks discussed below.

2.3 Construction of the Harmonized Dataset

After preprocessing and cleaning, we constructed a dataset for joint analysis of Black Carbon (BC) and Elemental Carbon (EC) by identifying co-located measurements, instances where both values were recorded at the same station on the same day. In total,

approximately 18,000 such records were extracted from the original dataset. This subset serves as the foundation for our modeling and outlier detection strategy, offering a coherent view of BC–EC relationships across Europe.

Given the drawbacks of the initial outlier detection methods, namely, the Interquartile Range (IQR), which proved too aggressive, and Isolation Forest (IF), which was too conservative, we explored a more flexible and interpretable alternative: quantile regression [6]. Unlike traditional regression, which estimates the mean response, quantile regression models conditional quantiles of BC given EC. This enables us to define a context-specific interval of expected values and identify measurements that deviate significantly from the typical BC–EC trend.

Optimized via the pinball loss function, quantile regression is well-suited for skewed or heteroscedastic environmental data. We fitted quantile curves at several levels to model the BC distribution across the EC spectrum, flagging as outliers any BC values lying outside the predicted 5–95% range, that is, outside the red-shaded area in the figure. This approach captures the central trend while accounting for variability, helping to avoid misclassifying values that may be extreme in magnitude but remain consistent with the underlying relationship between BC and EC.

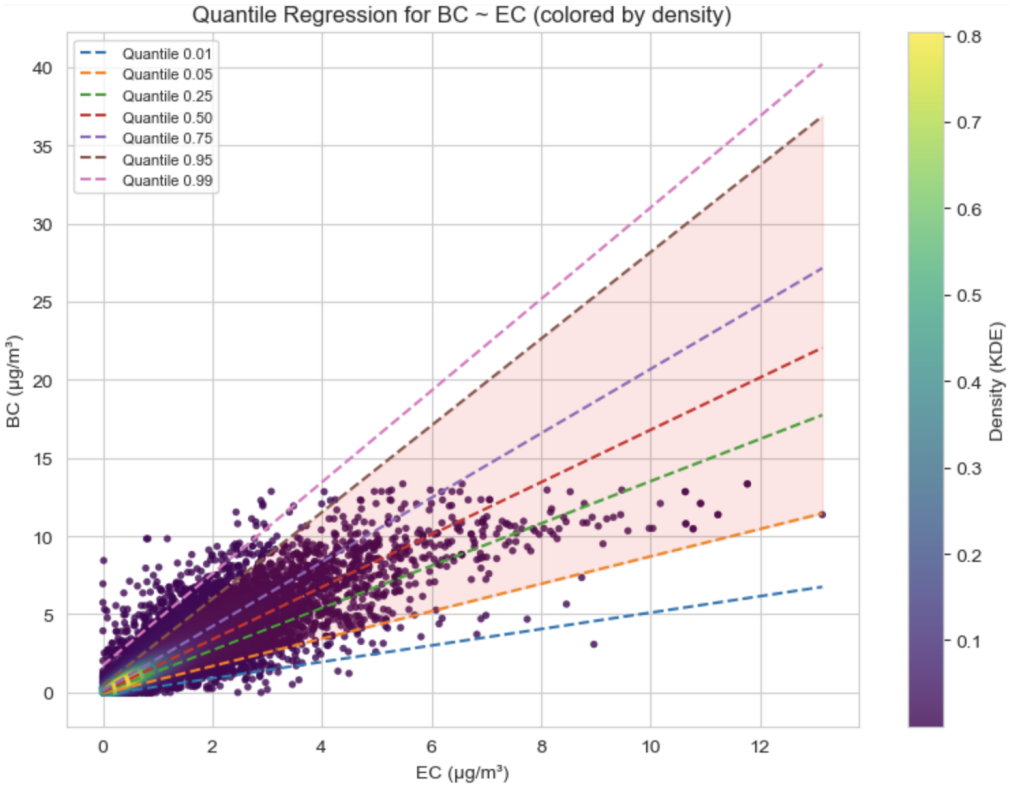

Figure 4: Quantile regression for BC ∼ EC with different quantile bands

This cleaned and harmonized dataset, built from co-located BC and EC measurements, provides a solid foundation for the next steps of our analysis. Thanks to the quantile regression approach, we can detect outliers more reliably by considering the expected relationship between BC and EC, rather than relying on fixed global thresholds. This helps preserve meaningful extreme values while removing inconsistent or biased ones. With this quality-controlled dataset, we are now ready to apply advanced harmonization methods and investigate systematic measurement biases across instruments and protocols.

3 Analysis of Protocol and Instrument Biases

3.1 Protocol Effects on Measurements

To assess whether measurement protocols introduce systematic biases in the EC–BC relationship of measured concentrations, we analyzed how the protocol affects the log-transformed ratio log(EC/BC), which provides a normalized metric to compare measurements across conditions.

We first visualized the log-ratio distribution globally and by protocol. Most values are centered below zero, indicating that EC is generally lower than BC, likely because thermal methods may misclassify some EC as OC or lose it during heating, while optical methods can overestimate BC by capturing additional particles. The EUSAAR and NIOSH distributions are similar in shape, but NIOSH is slightly shifted to the left, suggesting that EC is even lower under this protocol, possibly due to differences in protocol sensitivity.

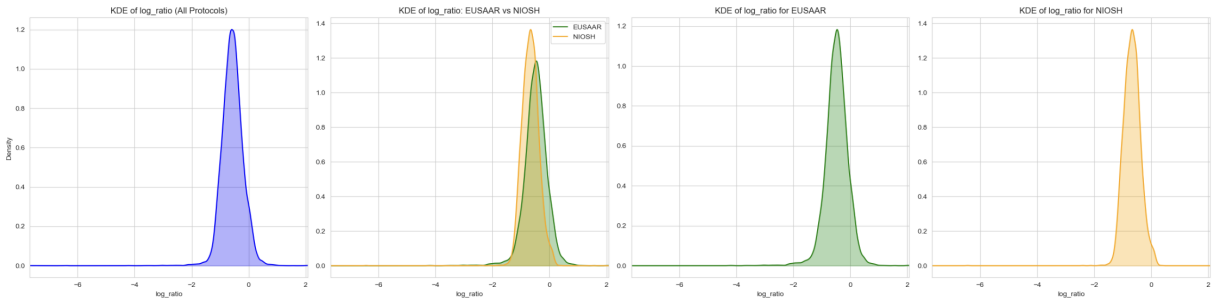

Figure 5: Distribution of log(EC/BC) for different protocols

To evaluate whether this visual difference is statistically significant, we performed a one-way ANOVA [7] with protocol as the grouping factor, to compare the mean log-ratios across the different protocols. The results show a very high F-statistic and a p-value close to zero, confirming that the differences in log-ratio across protocols are statistically significant. Thus, the measurement protocol has a systematic effect on the EC/BC ratio.

To assess how measurement protocols affect the EC/BC ratio, we fitted a linear regression model with dummy variables [8], using `UNKNOWN` as the reference category:

$$\log\left(\frac{\text{EC}}{\text{BC}}\right) = a + b \cdot \text{EUSAAR} + c \cdot \text{NIOSH} + \varepsilon$$

The regression yielded the following coefficients ($a = -0.6034$):

- **EUSAAR protocol effect**: $b = +0.1238 \Rightarrow \text{EC} \approx 0.62 \times \text{BC}$
- **NIOSH protocol effect**: $c = -0.0613 \Rightarrow \text{EC} \approx 0.51 \times \text{BC}$

All effects were statistically significant ($p < 0.05$), confirming protocol-dependent variations in EC/BC ratios. EUSAAR consistently yielded higher EC values than NIOSH, necessitating protocol-specific adjustments in harmonization models.

To further validate these results, we modeled the EC–BC relationship using quantile regression at the median ($q = 0.5$). This approach captures the central trend while remaining robust to outliers, and allows us to evaluate whether the relationship slope varies by protocol, a key indicator of structural bias in measurements.

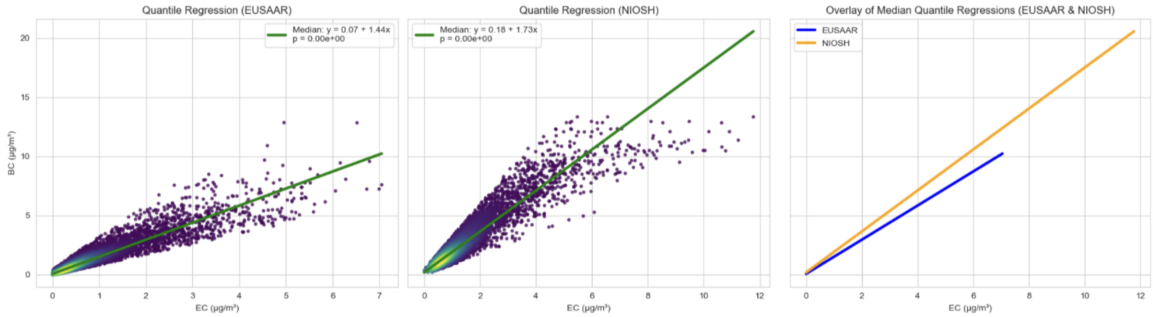

Figure 6: Median Quantile Regressions of BC as a function of EC by protocol

Median quantile regression confirms that the EC–BC relationship varies by protocol, supporting the inclusion of protocol as an essential explanatory variable in harmonization. As shown in the plots, NIOSH displays a steeper slope than EUSAAR, indicating lower EC concentrations for the same BC value. These results are consistent with the log-ratio analysis and highlight systematic differences across protocols.

3.2 Instrument Effects on Measurements

To determine whether the instruments used to measure Black Carbon (BC) and Elemental Carbon (EC) introduce systematic differences, we analyzed the log(EC/BC) ratio across all instrument types.

We began by visualizing the distribution of log-ratio values for each instrument. The figure below shows histograms grouped by BC instrument (MAAP vs. AE) and EC instrument (LVol_Filters vs. HVol_Filters). While the distributions appear similar overall, slight shifts in central tendency suggest potential systematic differences.

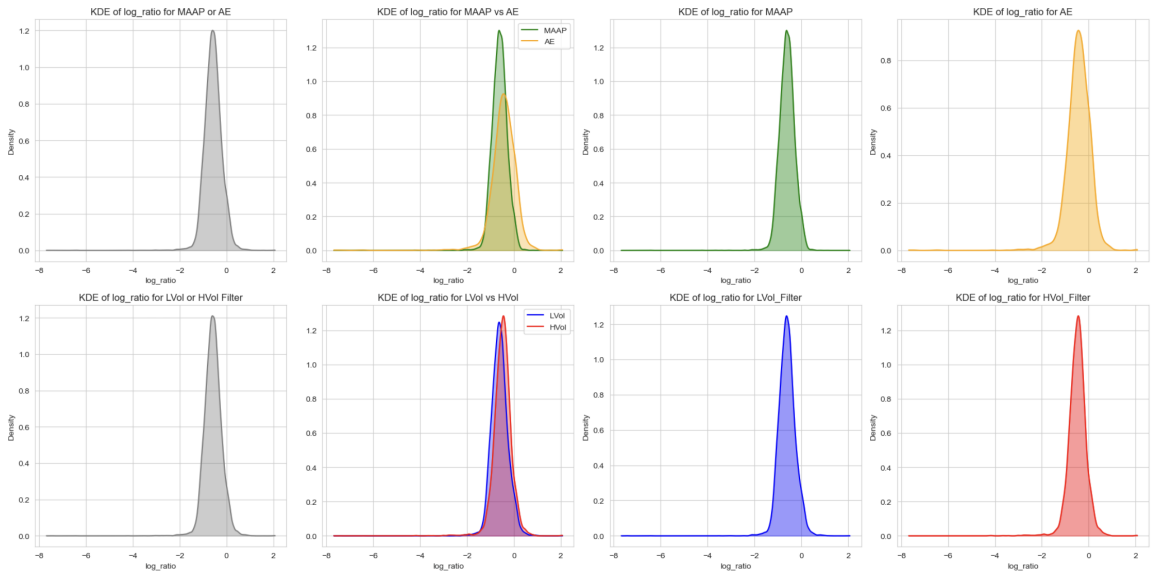

Figure 7: Distribution of log(EC/BC) by BC and EC instrument types

To statistically assess these differences, we performed two one-way ANOVA tests, one for BC instruments (MAAP vs. AE), and one for EC instruments (LVol_Filters vs. HVol_Filters). In both cases, the p-values were smaller than 0.05, confirming that the

observed differences in log-ratio are significant. The corresponding F-statistics were also high, indicating that the choice of instrument has a strong effect on the EC/BC ratio. To quantify the effect of each instrument, we fit two linear regression models of the form:

$$\log\left(\frac{\text{EC}}{\text{BC}}\right) = \beta_0 + \beta_1 \cdot \text{Instrument}$$

Measurement	Instrument Type	Estimated log(EC/BC)	EC ≈ Coeff x BC
BC	MAAP	-0.432700	0.65
BC	AE	-0.604700	0.55
EC	LVol_Filter	-0.483600	0.62
EC	HVol_Filter	-0.603100	0.55

Figure 8: OLS regression results of log(EC/BC) by instrument type

The table shows that AE yields lower EC values than MAAP for the same BC, and that LVol_Filters report higher EC than HVol_Filters. These results confirm that instrument type systematically affects the EC/BC ratio.

To further explore the structure of the EC–BC relationship, we used median quantile regression ($q = 0.5$) to estimate the central trend for each instrument.

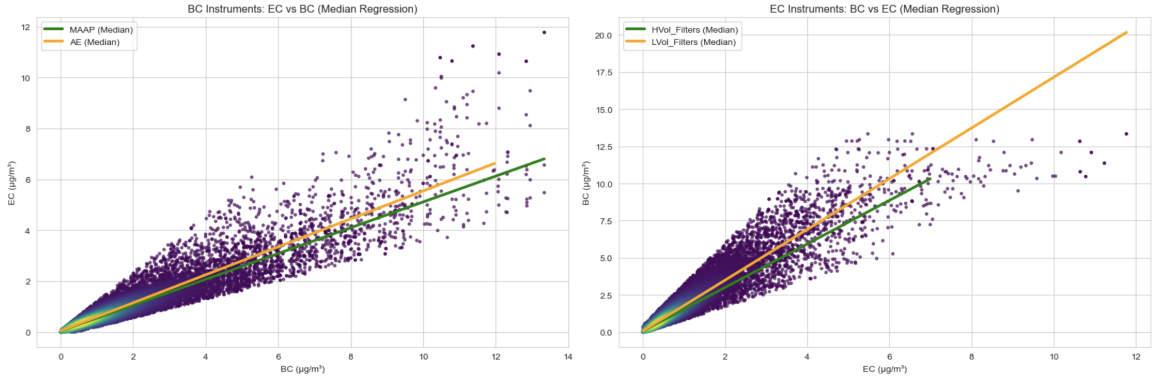

Figure 9: Median quantile regressions of EC and BC by instrument type

Overall, the median quantile regressions show that instrument choice affects both the level and the structure of EC–BC measurements. While MAAP and AE (BC instruments) display nearly identical slopes, suggesting consistent EC–BC relationships despite minor scale differences, Low Volume and High Volume filters (EC instruments) show a somewhat more pronounced divergence in both slope and magnitude. These results confirm that both BC and EC instruments introduce systematic measurement variations.

3.3 Combined Influence of Protocol and Instruments

After analyzing protocol and instrument effects individually, we evaluated their combined influence on log(EC/BC) by creating a categorical variable `combination` representing each unique protocol-BC-EC instrument triplet. Only the most frequent configurations were included to ensure statistical reliability.

A one-way ANOVA confirmed significant differences across combinations ($F = 250.4$, $p < 10^{-300}$), indicating that the interaction between protocol and instruments strongly influences the EC/BC relationship. To pinpoint which specific combinations differ, we applied Tukey's Honest Significant Difference (HSD) test [9], a post-hoc analysis that compares all possible pairs of means while adjusting the significance level to account for multiple comparisons. This method controls the overall false-positive rate, ensuring that observed differences are unlikely to arise by chance. The test revealed statistically significant differences in most pairwise comparisons, further confirming systematic variation between configurations.

To quantify these effects, again, we fitted a linear regression model using dummy variables for each combination:

$$\log(\text{EC/BC}) = \beta_0 + \sum_{i=1}^{k} \beta_i \cdot \text{Combination}_i + \varepsilon$$

The regression results confirmed that most combinations have statistically significant effects on the log-ratio. Here are the estimated EC/BC conversion factors:

- `MAAP_LVol_NIOSH`: EC $\approx 0.509 \times$ BC `MAAP_LVol_EUSAAR`: EC $\approx 0.584 \times$ BC
- `MAAP_HVol_EUSAAR`: EC $\approx 0.611 \times$ BC `AE_HVol_EUSAAR`: EC $\approx 0.645 \times$ BC
- `AE_LVol_EUSAAR`: EC $\approx 0.706 \times$ BC `AE_LVol_NIOSH`: EC $\approx 0.713 \times$ BC

To test if these differences affect the EC–BC relationship, we used median quantile regression ($q = 0.5$) for each configuration, providing robust trend estimates.

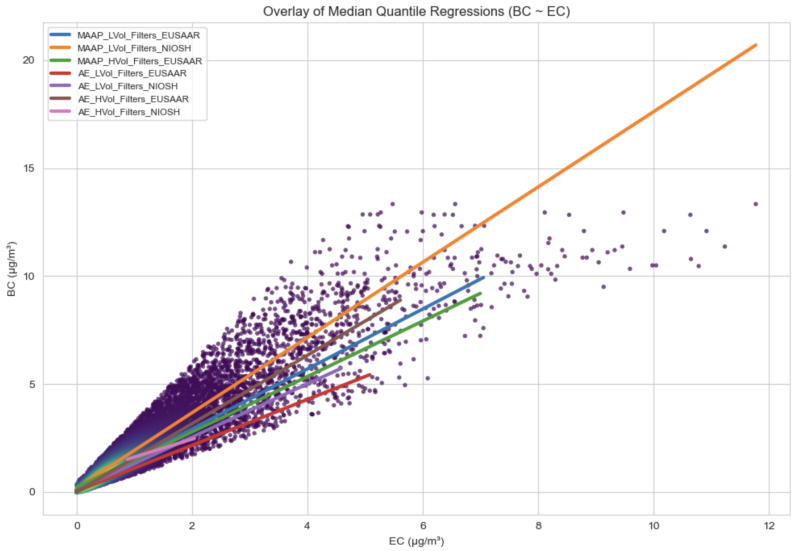

Figure 10: Overlay of median quantile regressions of BC on EC by combinations

As shown, each combination exhibits a distinct regression line. While all maintain a positive relationship between EC and BC, the slopes and intercepts vary considerably, highlighting that protocol–instrument interactions influence both the magnitude and shape of the EC–BC relationship.

These results confirm that the combined influence of measurement protocol and instrument choice introduces systematic and heterogeneous effects.

4 Environmental Influence on Measurements

4.1 Atmospheric Factors and Their Role

Beyond instrumentation and measurement protocols, environmental conditions can also influence the estimation of carbonaceous pollutant concentrations. In particular, two categories of atmospheric components, secondary aerosols and mineral dust, are known to affect the accuracy of Black Carbon measurements, especially when using optical methods such as MAAP or Aethalometer (AE).

In this study, we refer to **Secondary Aerosols** as atmospheric particles composed of both organic and inorganic components. Organic aerosols (OA) originate from natural sources such as vegetation or from human activities like transportation and industry, and can also form through chemical reactions in the atmosphere [10]. Inorganic aerosols, or Secondary Inorganic Aerosols (SIAs), include sulfate (SO_4^{2-}), nitrate (NO_3^-), and ammonium (NH_4^+), typically produced through atmospheric reactions involving sulfur and nitrogen compounds emitted by combustion processes. These aerosols frequently coexist with Black Carbon (BC) and can interfere with optical measurements used to estimate its concentration [11]. Since such methods rely on light attenuation by particles on filters, the presence of other light-absorbing but non-carbonaceous species can lead to a systematic overestimation of BC. Their abundance, which varies by location, season, and emission source, represents a confounding factor in interpreting BC measurements.

Another important contributor to measurement uncertainty is **Mineral Dust**. Originating from natural and anthropogenic sources such as desert transport, soil resuspension, or road abrasion, dust particles, although non-carbonaceous, can absorb and scatter light depending on their composition and size. When present in large quantities, this optical interference can be misinterpreted as Black Carbon (BC) absorption, especially in optical measurement methods [12]. This can lead to an overestimation of BC concentrations, a phenomenon known as dust-induced BC overestimation, which may distort the EC/BC relationship and introduce bias in harmonization.

Understanding how these atmospheric components affect BC and EC measurements is essential for accurate data interpretation and harmonization. In the following subsections, we assess in detail the influence of aerosols and dust on the EC/BC relationship and the resulting biases in concentration estimates.

4.2 Aerosol Impact on BC/EC Relationship

To investigate the potential influence of secondary aerosols on the relationship between Black Carbon (BC) and Elemental Carbon (EC), we analyzed how aerosol abundance relates to the log(BC/EC) ratio. Optical instruments used for BC detection may be sensitive to non-carbonaceous particles such as organic and inorganic aerosols, which can interfere with light absorption and lead to biased measurements. In this section, we explore whether these secondary components are associated with systematic deviations between BC and EC estimates.

We used aerosol concentrations from CAMx (Comprehensive Air Quality Model with Extensions), a chemistry-transport model that provides simulated data for organic aerosol (OA), sulfate (SO_4^{2-}), nitrate (NO_3^-), and ammonium (NH_4^+). Our hypothesis is that higher concentrations of these species may cause BC to be overestimated, increasing the log(BC/EC) ratio.

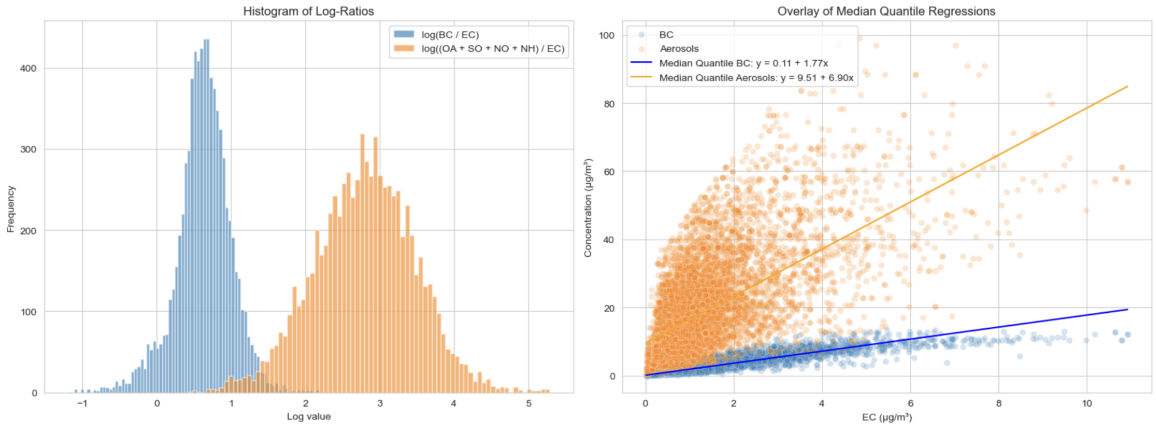

Figure 11: Distributions and quantile regressions of BC and aerosols with respect to EC

As shown in the figure above, the histogram of log(Aerosol/EC) is shifted to the right and exhibits greater dispersion than that of log(BC/EC), indicating that secondary aerosols are not only more abundant but also more variable than BC. To further examine this, we performed median quantile regressions of BC and aerosol concentrations against EC. The estimated regression lines are:

$$\text{BC}: \quad y = 0.11 + 1.77x$$

$$\text{Aerosols}: \quad y = 9.51 + 6.90x$$

These results show that BC increases moderately with EC, whereas aerosols display a much steeper and higher baseline trend. This pronounced divergence highlights the stronger dependency of aerosol levels on EC, and suggests that in aerosol-rich environments, optical BC measurements may be biased upward due to the presence of non-carbonaceous absorbing particles.

We then modeled the direct relationship between the aerosol-to-EC ratio and the log(BC/EC) value using two approaches: a degree-2 polynomial regression and a smoothing spline [13]. As shown below, both modeling techniques reveal a positive association. In raw scale, the relationship is slightly curved. However, in log-log space, the relationship is more linear and consistent, reinforcing the idea that BC is increasingly overestimated relative to EC as aerosol concentration grows.

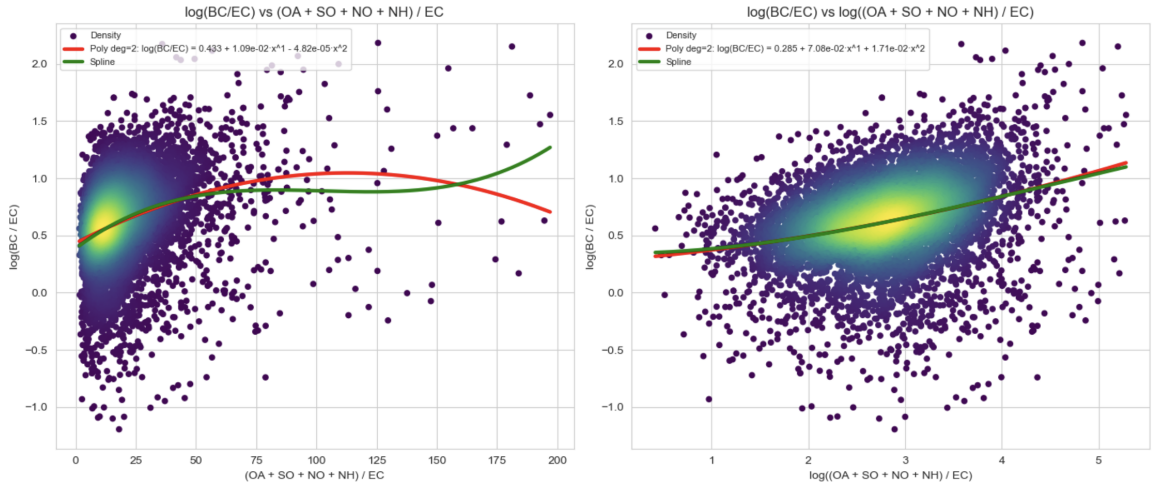

Figure 12: Modeling log(BC/EC) as a function of aerosol load

To statistically assess the association between aerosol presence and BC overestimation, we applied a Spearman correlation test between log(BC/EC) and log(Aerosol/EC). The result yielded a correlation coefficient of $\rho = 0.324$ with a p-value below 10^{-180}, indicating a strong and highly significant positive monotonic relationship. This confirms that higher aerosol load is statistically associated with increased BC values relative to EC, supporting the hypothesis of an optical measurement bias induced by secondary aerosols.

To deepen our analysis, we focused on nitrate (NO_3^-) and sulfate (SO_4^{2-}), two common and major secondary inorganic aerosols frequently present in polluted air that may interfere with optical BC measurements. We modeled their combined effect using a second-degree polynomial regression, both in absolute terms and normalized by EC.

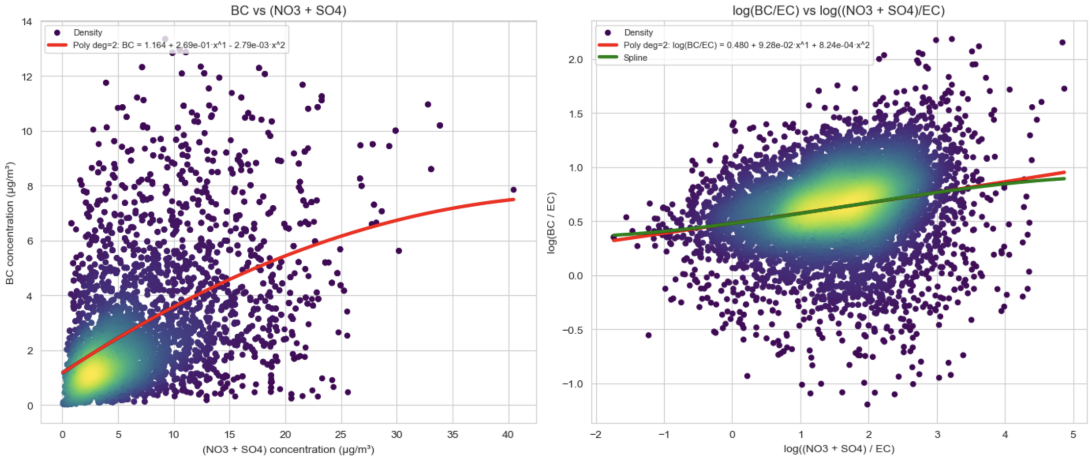

Figure 13: BC concentration and BC/EC ratio as functions of nitrate and sulfate concentration

As shown in the figure above, the left panel displays the fitted second-degree polynomial regression that showed a positive association between their combined concentration and BC levels.

$$\mathrm{BC} = 1.164 + 2.69 \times 10^{-1} \cdot (\mathrm{NO_3} + \mathrm{SO_4}) - 2.79 \times 10^{-3} \cdot (\mathrm{NO_3} + \mathrm{SO_4})^2$$

The regression suggests that BC increases with the concentration of these aerosols, but the negative quadratic term indicates a saturation effect at higher values. The right panel shows the relationship between the normalized variables:

$$\log\left(\frac{\mathrm{BC}}{\mathrm{EC}}\right) = 0.480 + 9.28 \times 10^{-2} \cdot \log\left(\frac{\mathrm{NO_3} + \mathrm{SO_4}}{\mathrm{EC}}\right) + 8.24 \times 10^{-4} \cdot \log^2\left(\frac{\mathrm{NO_3} + \mathrm{SO_4}}{\mathrm{EC}}\right)$$

This model reveals a steadily increasing trend, suggesting that the log(BC/EC) ratio increases with $\log((\mathrm{NO_3} + \mathrm{SO_4})/\mathrm{EC})$, supporting the hypothesis that these aerosols contribute to systematic overestimation of BC. So, indeed, the presence of secondary inorganic aerosols contributes to a systematic upward bias in BC relative to EC, even after adjusting for EC levels.

In summary, both organic and inorganic secondary aerosols are key confounders in optical BC measurements, often leading to overestimated values relative to EC. These biases affect the EC–BC relationship and underscore the need to consider environmental composition in harmonization models for carbonaceous pollutants.

4.3 Dust Contribution to Concentration Bias

Beyond secondary aerosols, mineral dust is another atmospheric component that may bias optical measurements of BC. Although generally considered non-carbonaceous, dust particles can scatter and absorb light depending on their mineral composition and size distribution. When such particles are present in significant quantities, their optical properties may lead to a misattribution of light absorption to BC, thereby inflating concentration estimates, the so-called dust-induced BC overestimation phenomenon.

To investigate this effect, we analyzed the relationship between dust and the log(BC/EC) ratio. As with aerosols, our hypothesis is that increasing dust concentrations could contribute to a systematic overestimation of BC when using optical methods. We first modeled the direct relationship between dust and log(BC/EC) using a second-degree polynomial regression. The results are shown in the left panel of the figure below.

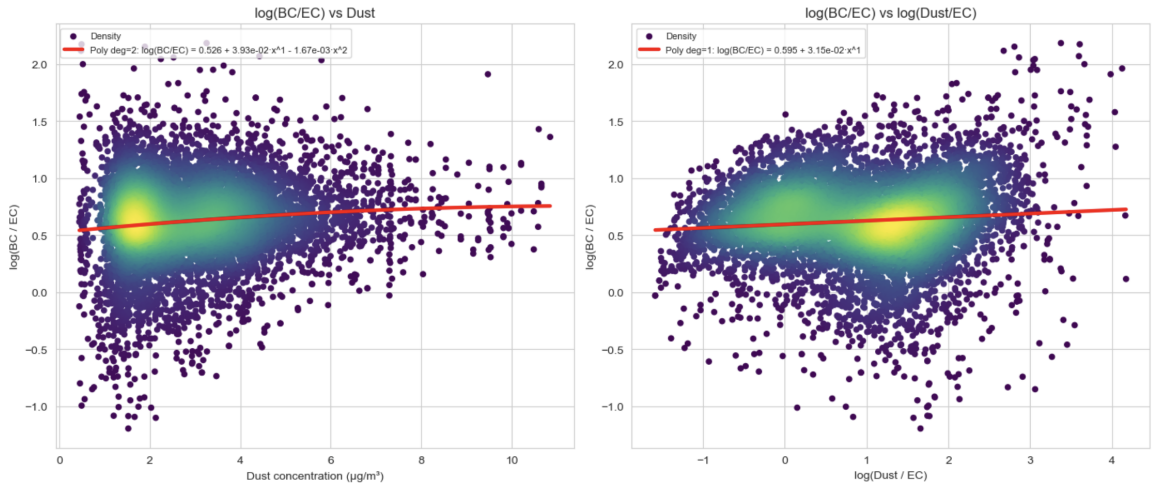

Figure 14: Modeling log(BC/EC) as a function of dust load

The polynomial regression applied to the raw scale dust concentration (left panel) reveals a mild but positive association with log(BC/EC), as captured by the fitted curve:

$$\log\left(\frac{\text{BC}}{\text{EC}}\right) = 0.526 + 3.93 \times 10^{-2} \cdot \text{Dust} - 1.67 \times 10^{-3} \cdot \text{Dust}^2$$

This trend is clearly confirmed and the relationship becomes more apparent when the dust concentration is normalized by EC and both variables are log-transformed. In the right panel, the model simplifies to a linear form:

$$\log\left(\frac{\text{BC}}{\text{EC}}\right) = 0.595 + 3.15 \times 10^{-2} \cdot \log\left(\frac{\text{Dust}}{\text{EC}}\right)$$

This suggests that the influence of dust on BC estimation is more effectively captured when expressed relative to EC. It supports the notion that measurement biases due to dust are proportional to its abundance in relation to carbonaceous sources.

To confirm this observation, we applied a Spearman correlation test between the two log-transformed variables. The result showed a statistically significant positive correlation, supporting the hypothesis that dust contributes to upward biases in BC measurements. Therefore, its presence in the atmosphere should be carefully accounted for in harmonization efforts, as it can introduce systematic misestimation of carbonaceous pollution.

5 Conclusion

5.1 Synthesis of Findings

This study investigated the factors influencing Black Carbon (BC) and Elemental Carbon (EC) concentration measurements, with a focus on identifying and correcting systematic discrepancies. By compiling a harmonized dataset of co-located measurements, we applied multiple outlier detection techniques, including quantile regression, to clean and prepare the data.

Our analysis revealed that both the measurement protocol and instrument type significantly impact the EC/BC ratio. We quantified these effects using regression models, highlighting the importance of protocol-specific and instrument-specific adjustments. Moreover, we examined how environmental components, namely secondary aerosols and mineral dust, affect optical BC measurements. These components were shown to systematically bias BC estimates upward.

Together, these findings underscore the need for harmonization strategies that account for both methodological and environmental variability. By systematically accounting for protocol, instrumentation, and environmental context, future studies can ensure more reliable and comparable carbon concentration assessments.

5.2 Limitations of the Study

While our analysis highlights systematic sources of bias in carbonaceous pollutant measurements, several limitations remain. First, the limited availability of co-located BC and EC data, measured simultaneously at the same station, restricted our ability to build a more continuous model of the EC–BC relationship. This data sparsity may have also affected the robustness of some analyses.

Second, certain combinations of instruments and protocols were underrepresented in the dataset. As a result, conclusions drawn for these configurations may not generalize beyond the specific contexts in which they were observed.

Finally, this study did not explicitly include temporal variables such as seasonality or meteorological conditions, which are known to influence aerosol formation, pollutant dispersion, and measurement accuracy. Future work should aim to integrate these factors for a more comprehensive understanding of measurement variability.

5.3 Perspectives and Next Steps

Due to time constraints, we were not able to develop predictive harmonization models that adjust BC and EC values based on the contextual factors identified in this study. However, a promising direction for future work involves building such models using protocol, instrument type, and ambient aerosol concentrations as input variables.

The objective would be to estimate the expected value of BC given EC, or vice versa, under specific measurement conditions. This would be particularly useful for expanding the set of co-located data, by enabling the imputation of missing BC or EC values when only one is available at a given station and time. By incorporating these contextual corrections, we can enhance dataset completeness and consistency, ultimately improving regional and temporal comparability in carbonaceous pollution monitoring.

6 Acknowledgments

I would like to express my sincere gratitude to Dr. Ekaterina Krymova, Lead Data Scientist at the Swiss Data Science Center (SDSC), for her supervision throughout this research project. Her guidance in machine learning and modeling, as well as her thoughtful suggestions for deeper and more relevant analytical approaches, played a crucial role in shaping the methodology and techniques applied in this work.

I am also deeply thankful to the team of domain experts at the Paul Scherrer Institute (PSI), and in particular to Dr. Imad El Haddad, Group Head of Molecular Cluster and Particle Processes. He proposed the research topic and gave us the opportunity to work on a project rooted in real-world environmental data and challenges. His insights were invaluable for understanding the data and making sense of our findings. I would also like to warmly thank the co-supervisors Dr. Marta Via Gonzalez and Dr. Abhishek Kumar Upadhyay for their continuous support throughout the project, especially with data access, interpretation, and their domain expertise.

Finally, I gratefully acknowledge Prof. Mathieu Salzmann for endorsing and officially supervising this project within the EPFL academic framework.

7 References

[1] Climate and Clean Air Coalition, "Black Carbon, An air pollutant with damaging effects," 2024, URL: https://www.ccacoalition.org/short-lived-climate-pollutants/black-carbon.

[2] European Commission, "Regulation on Ambient Air Quality and Cleaner Air for Europe, Directive 2024/2881," 2024, URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ%3AL_202402881.

[3] C. Long, M. Nascarella, and P. Valberg, "Carbon black vs. black carbon and other airborne materials containing elemental carbon: Physical and chemical distinctions," 2013, URL: https://www.sciencedirect.com/science/article/pii/S0269749113003266.

[4] Scribbr, "Interquartile Range (IQR): Definition, Calculation, and Interpretation," 2020, URL: https://www.scribbr.com/statistics/interquartile-range/.

[5] C. Maklin, "An Introduction to Isolation Forest," 2022, URL: https://medium.com/@corymaklin/isolation-forest-799fceacdda4.

[6] Scikit-Learn developers, "Quantile Regression," 2024, URL: https://scikit-learn.org/stable/auto_examples/linear_model/plot_quantile_regression.html.

[7] Scribbr, "One-way ANOVA: Definition, Formula and Examples," 2020, URL: https://www.scribbr.com/statistics/one-way-anova/.

[8] University of Southampton, "Multivariate analysis: Linear regression," 2014, URL: https://www.southampton.ac.uk/passs/confidence_in_the_police/multivariate_analysis/linear_regression.page.

[9] M. Terpilowski, "scikit-posthocs: Pairwise multiple comparison tests in Python," 2019, URL: https://github.com/maximtrp/scikit-posthocs.

[10] ScienceDirect, "Atmospheric Organic Aerosol," 2023, URL: https://www.sciencedirect.com/topics/earth-and-planetary-sciences/atmospheric-organic-aerosol.

[11] F. Granella, S. Renna, and L. Aleluia Reis, "The formation of secondary inorganic aerosols: A data-driven investigation of Lombardy's secondary inorganic aerosol problem," 2024, URL: https://www.sciencedirect.com/science/article/abs/pii/S1352231024001559.

[12] L. Zeng *et al.*, "Overestimation of black carbon light absorption due to mixing state heterogeneity," 2024, URL: https://www.nature.com/articles/s41612-023-00535-8.

[13] SciPy developers, "Smoothing Splines Manual," 2023, URL: https://docs.scipy.org/doc/scipy/tutorial/interpolate/smoothing_splines.html.