Modeling Unreported Pollution Events from Air Monitoring Data

Cuining Liu

STAT 202A Final Project

6 December 2024

Background

Although petroleum products are essential to energy, transportation, and myriad manufactured goods, their utility must be weighed against the adverse impacts of air pollutants on environmental and human health[1]. Unfortunately, their impacts are likely systemically underestimated due to outdated techniques or irregular monitoring[2,3]; for example, the predicted emission levels of the carcinogen benzene reported by refineries to regulatory agencies based on industry-standard simulations were up to 28-fold lower than the true values empirically measured by continuous air monitoring systems[4].

Accurate and transparent reporting is especially ethically imperative for petroleum operations in population-dense areas, such as the Suncor petroleum refinery located three miles from downtown Denver, contributing to what has been deemed "the most polluted zip code in America."[5] Suncor has extremely high rates of "**reportable events**"—refinery malfunctions and pollution violations that require reporting to federal and local agencies by law—relative to refineries of matched characteristics like yield[6]. Despite the high number of violations, regulatory action has been limited or extremely slow; for example, the refinery has operated on an expired Environmental Protection Agency (EPA) permit for 12 years[7]. Only in the past three years has the refinery been mandated to enact empirical, real-time air monitoring systems[8]. As described above, these monitors are necessary for accurate risk assessment and their long-standing absence suggests that Suncor's already high number of reportable events have likely been underestimated.

Consequently, a surprising degree of pressure on Suncor has come not from regulatory government agencies, but instead grassroots efforts from the most affected communities. including citizen scientists that have created their own air monitoring systems[9,10]. With this data and other grassroots analyses, local community members recently filed a lawsuit documenting 9,205 Clean Air Act (federal EPA) violations in last five years[11].

In the continued spirit of this community-based regulation and monitoring effort, this report interrogates whether the newly released public air pollutant monitoring data from Suncor could feasibly be used to predict unplanned pollutant events that may not be reported or even detected. Specifically, we (1) perform exploratory analyses describing the publicly available air monitoring data then (2) assess whether measured air pollutant concentrations on a given day are significantly associated with a reported emission violation on that day.

Approach and Data Sources

The outcome, aerial concentrations C_{pmd} of pollutant p from monitor m over varying dates and times were downloaded for $p \in$ {benzene, H_2S, SO_2, CO}[12,13]. The primary predictor of interest, X_{pd}—an indicator variable whether a reportable event occurred on day d related to each respective p were downloaded from the Commerce City website[14] for 2020 to 2023 events or scraped from .pdf reports from Suncor[15] for 2024 events. Lastly, weather variables W_d including maximum temperatures (°F), precipitation levels (mm), and snowfall levels (mm) were averaged across the three nearest weather stations[16] nearest to Suncor as a putative time-varying confounder or proxy for confounders (e.g., emergence of other air pollution sources like forest fires due to seasonality).

The data was *a priori* expected to feature complex spatialtemporal correlations and require non-Gaussian modeling as pollutant concentrations must be non-negative. We thus made two pre-analysis, pre-specified simplifications. Unless otherwise specified, the spatial effects and monitor effects were ignored by aggregating concentrations to a single daily value C_{pd} by taking the 90th quantile of all C_{pmd} for a given ($p = p^*$, $d = d^*$) pair. The outcome was modeled using a Generalized Additive Model (GAM) with a $k = 100$ B-spline basis and Gamma distribution assumed log-linear link function; this attempts to capture the mean patterns with time, resulting in an approximately stationary process and, by extension, valid inference on other covariates after adjustment for these splines.

Results: Benzene Fenceline Data

First, we examined benzene concentrations, measured once every fourteen days by sixteen monitors at the “fenceline” perimeter of the Suncor facility. We note the presence of spatial correlation between nearby monitors, with the expected trend of higher correlations when the monitors are adjacent (**Fig 1A, Fig 1B**). Most notably, the monitors south of the refinery exhibit higher benzene concentrations as shown by distance-weighted interpolations (**Fig 1B**) and kernel density estimates by monitor ID (**Fig 1C**).

After aggregating C_{pmd} values (**Fig 1D**) into 90th quantile C_{pd} values (**Fig 1E**) for modeling, the regression model for benzene (**Fig 2A**) suggests that dates within $\tau = 2$ days of a reportable event were associated with estimated 1.2-fold mean increase in benzene concentrations recorded at the fenceline monitor albeit not statistically significantly so (p-value = 0.07, 95% CI 0.99 - 1.40). Interestingly, precipitation was associated with decreases in benzene levels, which is consistent with past literature given that rain and snow may may sequester benzene out of the air[3,17]. The overall percent variance explained was $R^2 = 0.40$.

Some sensitivity analyses included alternative model specifications, such as different encodings of the weather terms (e.g., yes/no precipitation versus continuous measure) and checking model fit scores (e.g., AIC; data not shown). Further, more complex models were fit without quantile-based aggregation and instead adding random intercepts and

slopes for monitor ID; however, this resulted poor model diagnostics, such as high autocorrelation between residuals (data not shown) that were not present in the standard quantile aggregation (**Figure 2B**). Lastly, due to the biweekly sampling, the reportable event indicator variable was "padded" to instead indicate whether C_{pmd} was measured within $\pm \tau$ days from a reported event. $\tau = 2$ was arbitrarily used to deal with limited overlap between dates in the monitor dates and benzene reported emissions (i.e., to deal with the rareness of predictor events). Consequently, the number of days padding was titrated as a sensitivity analysis (**Fig 2C**). There remained a positive and non-significant association between report date for all choices of τ, with the highest association found for $\tau = 2$ to 4.

Results: H_2S, CO, and SO_2 Monitors

We attempted to repeat these analyses for H_2S, CO, and SO_2, with the modification that $\tau = 0$ was used due to much higher sampling rate (hourly data measured at the same 10 monitoring sites more distal to the refinery; in contrast to biweekly benzene sampling at the fenceline) and much higher frequency of reportable events, which was respectively 48 days with reportable events out of 908 days (5.2%), 87 out of 908 (9.6%), and 56 out of 641 (8.7%; fewer total days due to missing data in late 2023), as compared to the 6 out of 108 (5.5%) for benzene with $\tau = 2$.

We noted strong pollutant-specific patterns in spatial correlation patterns (**Fig 3A-B**) and temporal trends (**Fig 4A-B**), potentially suggesting to distinct emission sources and/or dispersal mechanisms for each pollutant, and by extension, different environmental and health impacts. Further, the pollutant-specific patterns have analysis implications, as each pollutant ideally must be carefully interpreted and modeled separately. Unlike the benzene monitors, the estimated kernel density distributions of pollutant concentration did not clearly differ by monitor (data not shown).

Similar diagnostics and interpretation to those presented for the benzene model were used to assess model fit, so are not discussed in detail here for succinctness. Like for benzene, the regression results for CO and H_2S (**Fig 4C-D**) suggest a positive association between pollutant concentration and reported events, with a statistically significant FC (95% CI) = 1.08 (1.04-1.13) for CO and FC = 1.48 (1.06-2.06) for H_2S. The respective R^2 values were 0.53 and 0.56. Interestingly, H_2S had an strong association with temperature that no other pollutant had (**Fig 4B**); again, this may reflect pollutant-specific etiologies. The SO_2 model was ultimately not presented due to poor model fit; this may be due to sparsity in the outcome, as only 515 of 6,474 C_{pmd} entries were non-zero.

Discussion

We found that reportable events to be associated with increases in pollutant concentrations, but with arguably modest effect sizes: the estimated mean fold change values ranged for 1.08 to 1.48, with a statistically significant association for H_2S and CO but not benzene. The overall explanatory power was low ($R^2 = 0.40$ to 0.54) for all models

despite allowing for arbitrary spline fits. Overall, the small associations and high residual variability in each model suggests there may be some merit in using the air monitoring data to predict when reportable events occur, but that the predictions may not be reliable.

This comes with the caveat that we pre-specified the GAM framework used for modeling and simplified the outcome to 90^{th} quantile concentration values at the cost of discarding data. Alternative modeling approaches may merit investigation; in particular, exploiting the spatial correlations in the data by for example assuming errors between monitors inversely correlated to their physical distance may be more powerful. However, these may require bespoke models not in existing R packages, particularly if a non-parametric form (GAM) is assumed for the mean time pattern. The non-Gaussian nature of the data also was challenging: SO_2 models in particular had poor fit (e.g., large, time-dependent residuals) for days where the concentration was presumably below the limited of detection for the air monitor, with $C_{pd} = 0$ ppb. Zero-inflated models may merit investigation for this and other low-concentration pollutants. With all of these limitations said, models for all pollutants except for SO_2 did reproduce effects like the known negative association between air pollutant concentration and precipitation (due to sequestering of chemicals), perhaps softly support the validity of the modeling approaches.

However, a fundamental issue regardless of modeling approach is the likely miscategorization bias in the predictor variable, emission reported (yes/no). Because the refinery is disincentivized to document violations and may not observe them depending on the intensity of their process monitoring, there are likely dates that should have values of $X_d = 1$ instead of $X_d = 0$. Ideally, the magnitude of the violation would also not taken into account for a more sensitive model (a severe and minor offence are both assigned $X_d = 1$), but this may require an intimate operational knowledge of refinery processes to generate. Finally, the sparsity of $X_d = 1$ events (<10% for all pollutants) makes these models unlikely to be robust: predicting the mean effect of a rare feature may be sensitive to noise and outliers at the few observed events, and from a predictive modeling standpoint would be less powerful than a “balanced class” problem with more comparable cases of $X_d = 1$ versus $X_d = 0$. Ultimately, it is challenging to construct a high quality ground truth predictor for this problem.

Altogether, we speculate that the limited success developing a model for air pollutant concentrations over time suggests that external, community-based monitoring of Suncor refinery based on these datasets would likely be challenging and unreliable. Although there is a positive association detected between air pollutant concentrations and reportable events, The seemingly low signal-to-noise in the data and lack of high quality ground truth data on reportable emissions. This raises questions about corporate responsibility: the analysis ultimately suggests the necessity of the refinery to release more data, be more transparent, and better monitor their systems, as the community may not be able to do alone. However, refineries are disincentivized from reporting their own malfunctions and excess pollution, requiring more support from regulatory agencies to back up community organizations at risk of environmental and health impacts.

Code Availability

The final code base for this project consisted of 2,834 lines of code (>40 pages) in R and C (kernel density estimation and inverse distance weighing interpolation functions), so was posted publicly on Github versus copied into this document.

Link: https://github.com/chooliu/STAT202A_RefineryEmissions

Figures and Tables

Fig 1. (A) A Spearman correlation matrix across benzene monitors shows higher correlation among adjacent monitors. The monitors' locations are shown in **(B)**, a plot of pollutant concentration on top of longitude and latitude. The "X" points represent the location of the Suncor refinery (consists of three plants). The measurements at each monitor are for one single (x, y) coordinate, but interpolated using inverse distance weighting using a custom C function for visualization. Each date presented was associated with a reported event within "padding" ± τ = 2 days. **(C)** Estimated kernel densities are different by monitor, further underscoring the spatial correlation patterns. **(D)** "Raw" C_{pmd} values and **(E)** aggregated 90th quantile measurements C_{pd} with overlaid GAM mean temporal trend prediction. Dates of reportable events are marked by vertical lines on the x-axis and are colored in maroon if within ± τ = 2 days of an event.

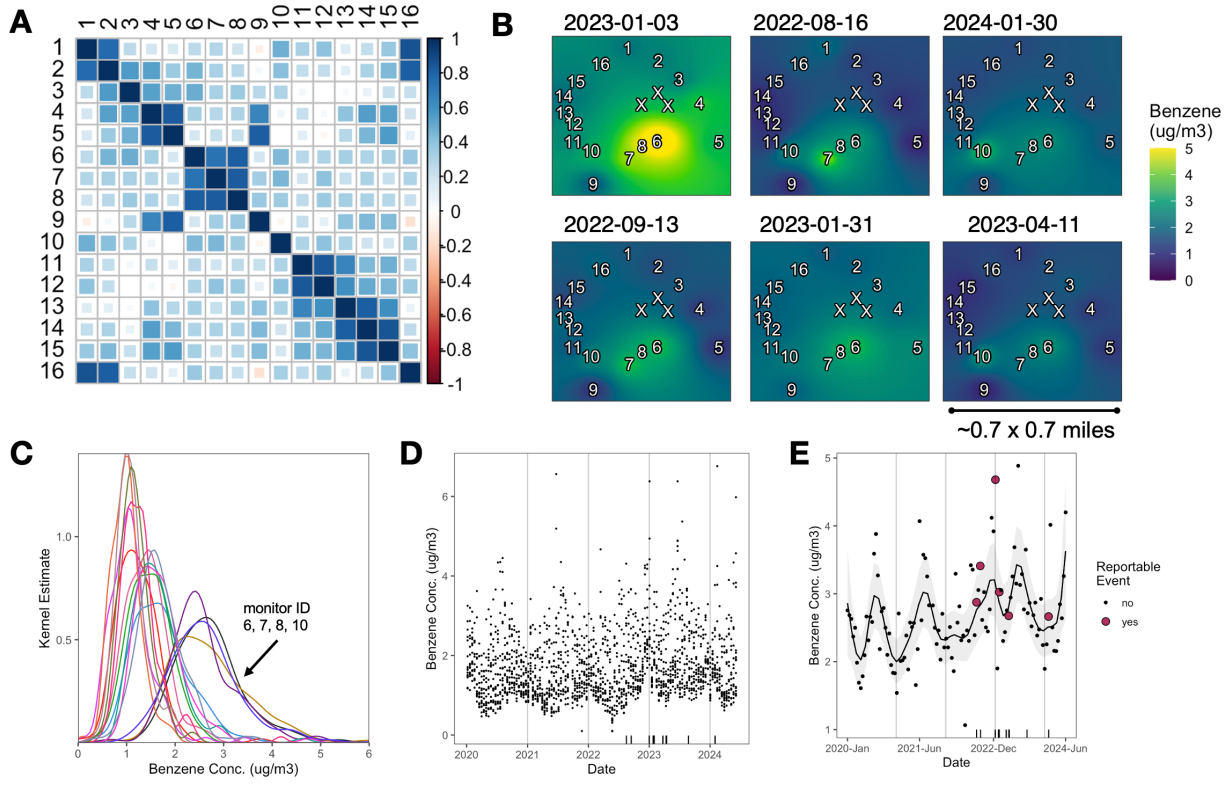
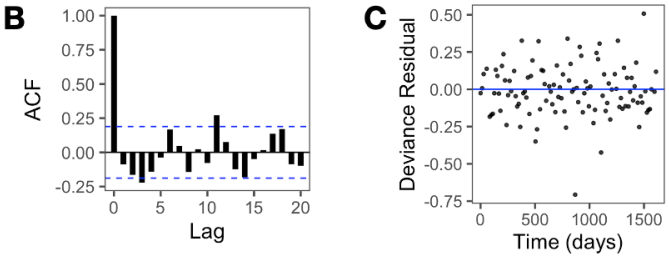


Fig 2. **(A)** Regression table for the GAM of benzene concentrations. The exponentiated regression coefficients exp(β) have been exponentiated for conversion from the log-link function into fold-change (FC) values and 95% CIs. **(B)** The limited autocorrelation function (ACF) lags and **(C)** association between deviance residuals and time (in days since start of monitoring suggest reasonable model fit for the benzene GAM. **(D)** A sensitivity analysis on the choice of τ. The error bars show the resulting FC (95%) values and number of days for which the reportable event predictor would equal to X_d = 1 when τ is changed. For example, only 1 day would intersect a reportable event date without this "padding", highlighting the sparsity of the predictor.

A

benzene	Fold Change (95% CI)	p-value
(Intercept)	2.48 (2.01-3.1)	7.60×10^{-13}
Emission Reported	**1.16 (0.97-1.4)**	**0.11**
Max Temp (°F)	1.00 (1.00-1.0)	0.54
Precipitation (mm)	0.71 (0.46-1.1)	0.13
Snow (mm)	1.01 (0.93-1.1)	0.81



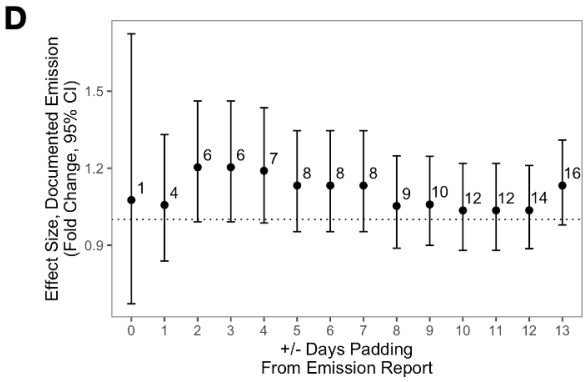


Fig 3. Interpolated **(A)** H_2S and **(B)** CO concentrations show different spatial patterns. Each panel represents one randomly sampled day on which there was a reported event for the respective pollutant. For H_2S, note that the limits of the color axis vary to emphasize relative patterns between monitors and locations, rather than the absolute pollutant values.

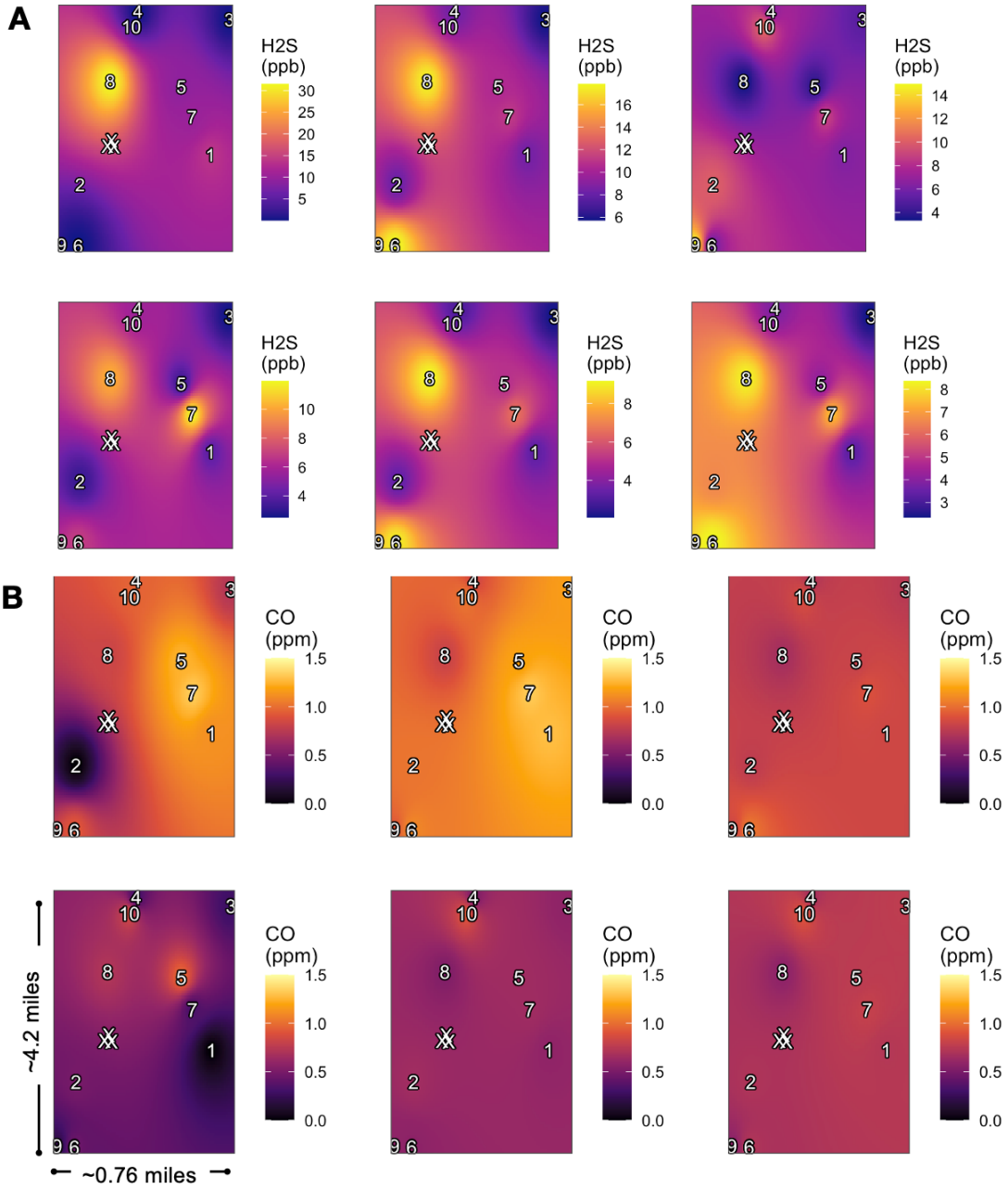
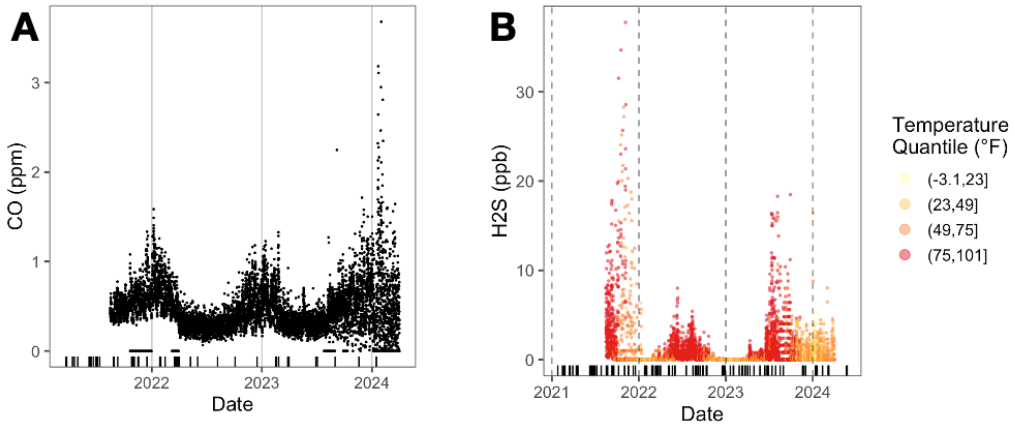


Fig 4. C_{pmd} values are plotted by day *d* for **(A)** CO and **(B)** H_2S are shown, with the latter colored by maximum temperature quantiles observed throughout the three year period modeled, showing a strong association between H_2S and temperature. Reportable events are shown again as vertical bars on the x-axis. The respective (C) CO and (D) H_2S regression tables show a statistically significant association between reportable events and pollutant concentrations.



C

CO	Fold Change (95% CI)	p-value
(Intercept)	0.42 (0.40-0.45)	6.9×10^{-152}
Emission Reported	1.08 (1.04-1.13)	0.00051
Max Temp (°F)	1.00 (1.00-1.00)	0.74
Precipitation (mm)	0.87 (0.83-0.92)	1.4×10^{-7}
Snow (mm)	0.91 (0.90-0.93)	3.6×10^{-34}

D

H_2S	Fold Change (95% CI)	p-value
(Intercept)	0.0069 (0.0039-0.012)	6.1×10^{-55}
Emission Reported	1.48 (1.06-2.06)	0.023
Max Temp (°F)	1.06 (1.05-1.07)	3.2×10^{-37}
Precipitation (mm)	0.75 (0.44-1.28)	0.29
Snow (mm)	1.22 (1.05-1.43)	0.0094

Works Cited

1. Adebiyi, F. M. Air quality and management in petroleum refining industry: A review. *Environ. Chem. Ecotoxicol.* **4**, 89–96 (2022).
2. He, M. *et al.* Total organic carbon measurements reveal major gaps in petrochemical emissions reporting. *Science (80-.).* **383**, 426–432 (2024).
3. Raynes, D., Klooster, A., Josh Eisenfeld, Levy, G. & Kirwin, M. *Certified Gaslighting*. (2024).

4. Brodzinsky, S. New study: Refineries under-reported benzene emissions by as much as 28-fold. *Oil and Gas Watch* (2023).
5. Horvath, A. How a Denver neighborhood became one of the most polluted zip codes in America. *Rocky Mountain PBS* (2023).
6. Dan Roper, Marissa Maier, E. R. G. I. *OCE-4 Technical Direction 20: Suncor Refinery Consent Decree Reportable Incident Analysis*. (2023).
7. Coghill, I. & Schluntz, A. Notice of Intent to Sue Over Failure to Act on Suncor Energy, Inc.'s Plant 2 (East Plant) Clean Air Act Title V Permit, 95OPAD108. (2023).
8. Curry, R. *Colorado's Suncor Refinery is Fighting a Plan to Monitor its Toxic Pollution*. (2022).
9. Detlev Helmig & Boulder A.I.R. Commerce City Current Air Conditions. (2023). Available at: https://www.bouldair.com/commerce_city.htm. (Accessed: 4th December 2024)
10. Brasch, S. A Latino-led group monitored the air near Suncor for more than a year. They found elevated levels of pollution and radioactive particles. *CPR News* (2023).
11. Booth, M. Nearly 10,000 Suncor air pollution violations in 5 years could lead to citizen lawsuit, groups warn. *The Colorado Sun* (2024).
12. US Environmental Protection Agency. The Benzene Fenceline Monitoring Dashboard. *AWSEDAP* (2024). Available at: https://awsedap.epa.gov/public/extensions/Fenceline_Monitoring/Fenceline_Monitoring.html?sheet=MonitoringDashboard&REGISTRY_ID=110032913024. (Accessed: 26th November 2024)
13. Montrose Air Quality Services. Commerce City-North Denver Air Monitoring. *CCND Air Monitoring program* (2024). Available at: https://www.ccnd-air.com/. (Accessed: 26th November 2024)
14. Commerce City. Living in Commerce City: Suncor Refinery. (2024). Available at: https://www.c3gov.com/living-in/energy-equity-and-the-environment/suncor-refinery. (Accessed: 26th November 2024)
15. Suncor. Reportable event Summaries. *Safety and Environment Information* (2024). Available at: https://www.suncor.com/en-ca/what-we-do/refining/commerce-city-refinery/safety-and-environment-information. (Accessed: 26th November 2024)
16. The Colorado Climate Center. Colorado Climate Center - Data Access. (2024). Available at: https://climate.colostate.edu/data_access_new.html. (Accessed: 27th November 2024)
17. Agency for Toxic Substances and Disease Registry. Public Health Statement for Benzene. *CDC Div. Toxicol. Environmrntal mediciine* 1–7 (2007).