Detecting plumes in mobile air quality monitoring time series with Density-based Spatial Clustering of Applications with Noise

Blake Actkinson[1], Robert J. Griffin[1,2*]

[1]Department of Civil and Environmental Engineering, Rice University, Houston, TX 77005, USA
[2] School of Engineering, Computing, and Construction Management, Roger Williams University, Bristol, RI 02809, USA

Correspondence to: Robert Griffin (rgriffin@rwu.edu)

Supplemental Information

Section S1. Temporal rescaling procedure for census tract comparisons.

Section S2. Anomaly type detection probability error estimation procedure.

Figures S1-S8

Tables S1-S8

S1 Temporal Rescaling Procedure for Census Tract Comparisons

To remove temporal effects from census tract comparisons of anomaly type detection probability, we perform a rescaling procedure. We transform each census tract's sampling distribution into a uniform distribution, then multiply each hour of the newly transformed uniform distribution by the fraction of detected anomalies in that hour.

Out of 35 census tracts sampled in the Houston area, we restrict our analysis to 19 to ensure that each hour between 8 AM and 4 PM CST had at least 1,000 samples for each individual census tract. The lowest number of samples in any given hour for a census tract was 1,061, which equates to ≈ 17 minutes of sampling. For each census tract, we calculate the average number of samples per hour, determined by calculating the total number of samples and dividing by 8, the number of analyzed hours. In addition to calculating the average number of samples, we calculate for each hour in each census tract the fraction of that hour's measurement that are of a given anomaly type ("CO_2 – Rich", "Transition", "BC/UFP – Rich"). In the final step, we multiply the hourly fraction of each anomaly type by the average number of measurements for the census tract and then sum the results. To determine the % probability of detection for a given anomaly type, we divide these weighted totals by the number of measurements made within the census tract.

Figures S2 and S3 display the effects of implementing the rescaling procedure on the calculated probabilities of anomaly detection for the 19 census tracts. In general, we note that implementing the rescaling procedure results in mostly modest increases in these probabilities across the board. A notable exception is the North Rice polygon for CO_2 anomaly detections. Figure S4 displays the (a) total sampling distribution and (b) anomaly sampling distribution for the North Rice polygon. We note that the 8 AM hour was oversampled relative to other hours sampled and argue

that implementing the rescaling procedure decreases the effects of this hour relative to other sampling times in the census tract.

S2. Anomaly Detection Type Probability Error Estimation Procedure

We provide error estimates of our calculated anomaly type detection probabilities and present them in Tabs. S3, S4, and S5. To do this, we implement the bootstrap for each anomaly detection type probability for each census tract to generate sampling distributions (Efron and Tibshirani, 1994).

We create 1000 synthetic distributions for each census tract by sampling with replacement measurements within each census tract. For each synthetic distribution, we calculate the probability of each anomaly detection type, repeating the same temporal rescaling procedure described in Sect. S1 1000 times for each census tract to generate 1000 probabilities of each type. From the resultant sampling distributions, we report the lower and upper bounds of the 90% confidence interval (5th to 95th percentiles), the mean, and bias. We define bias as the difference between the originally calculated probability and its mean probability estimate from its corresponding sampling distribution (in effect, taking the difference between columns in Tab. 2 and mean columns in Tabs. S3, S4, and S5).

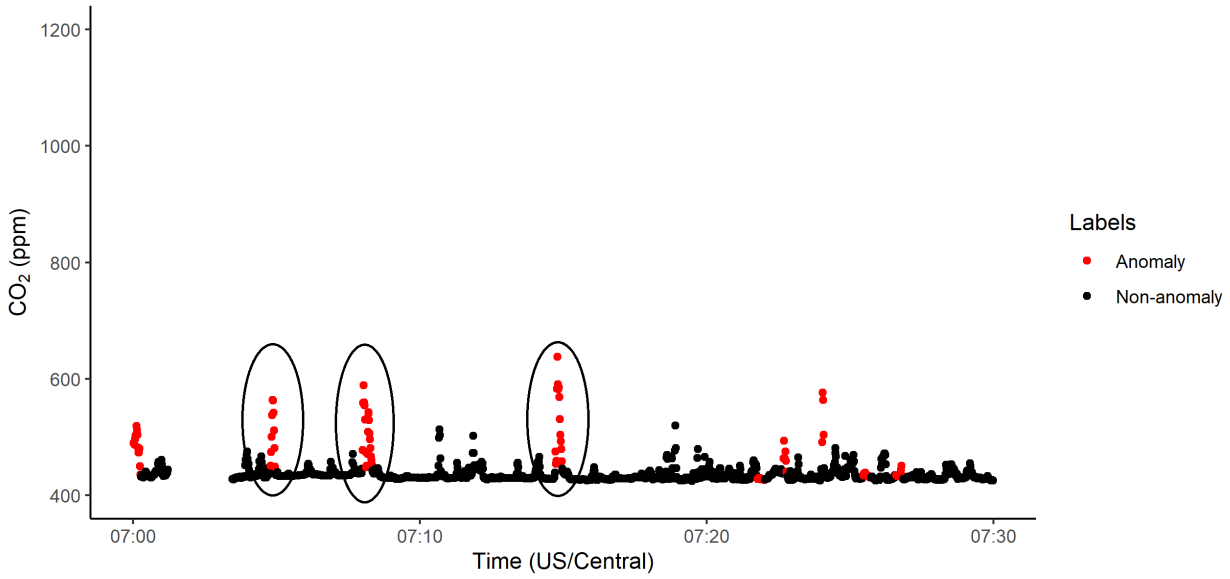

Figure S1. Illustration of manually flagged plumes for CO_2. Points in red are labeled as plume (anomaly), while points in black are labeled as normal (non-anomaly). Ovals represent manually flagged plumes for this portion of the CO_2 time series. Note – not all red colored points correspond to CO_2 plumes, but they can represent plume detections in other pollutants not shown here.

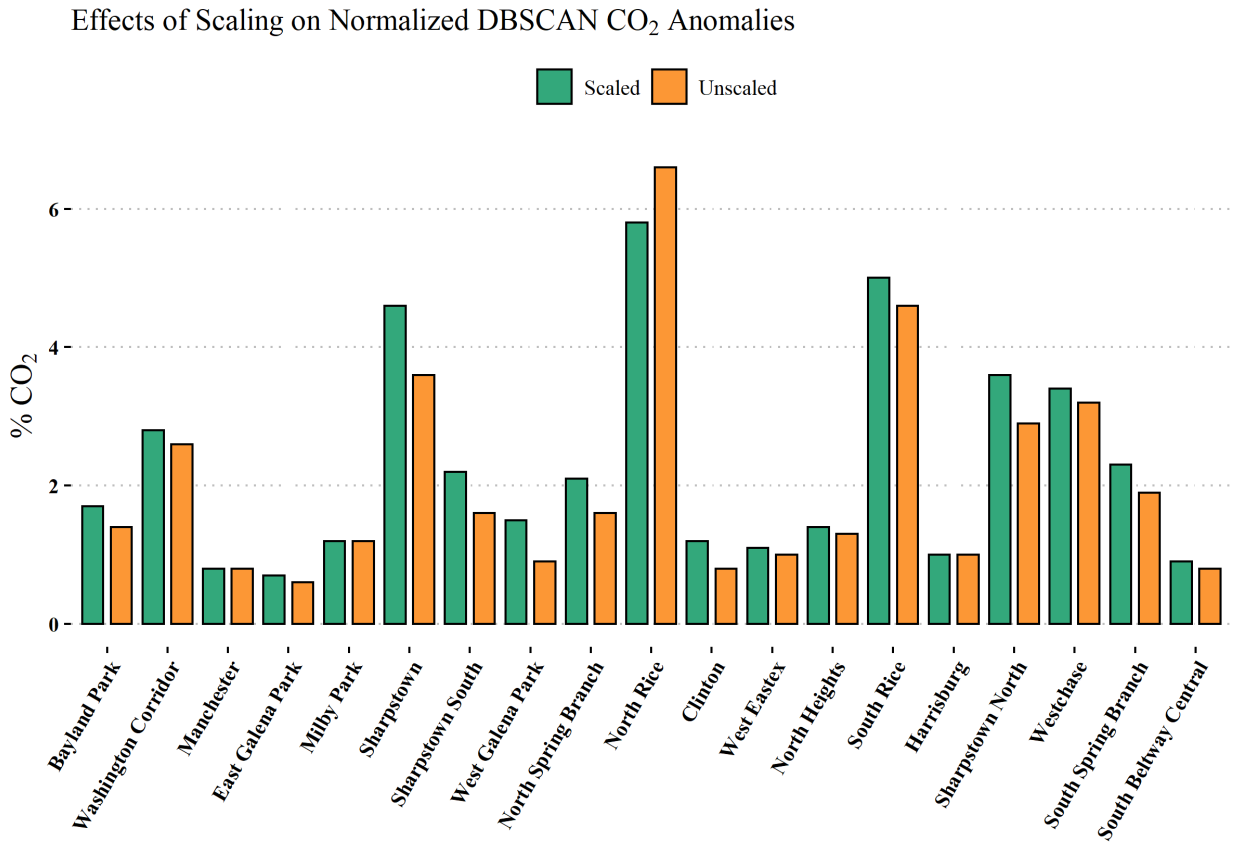

Figure S2. Effects of scaling on the probability of CO_2 anomaly type detection for each census tract (green/left bar for each tract is scaled).

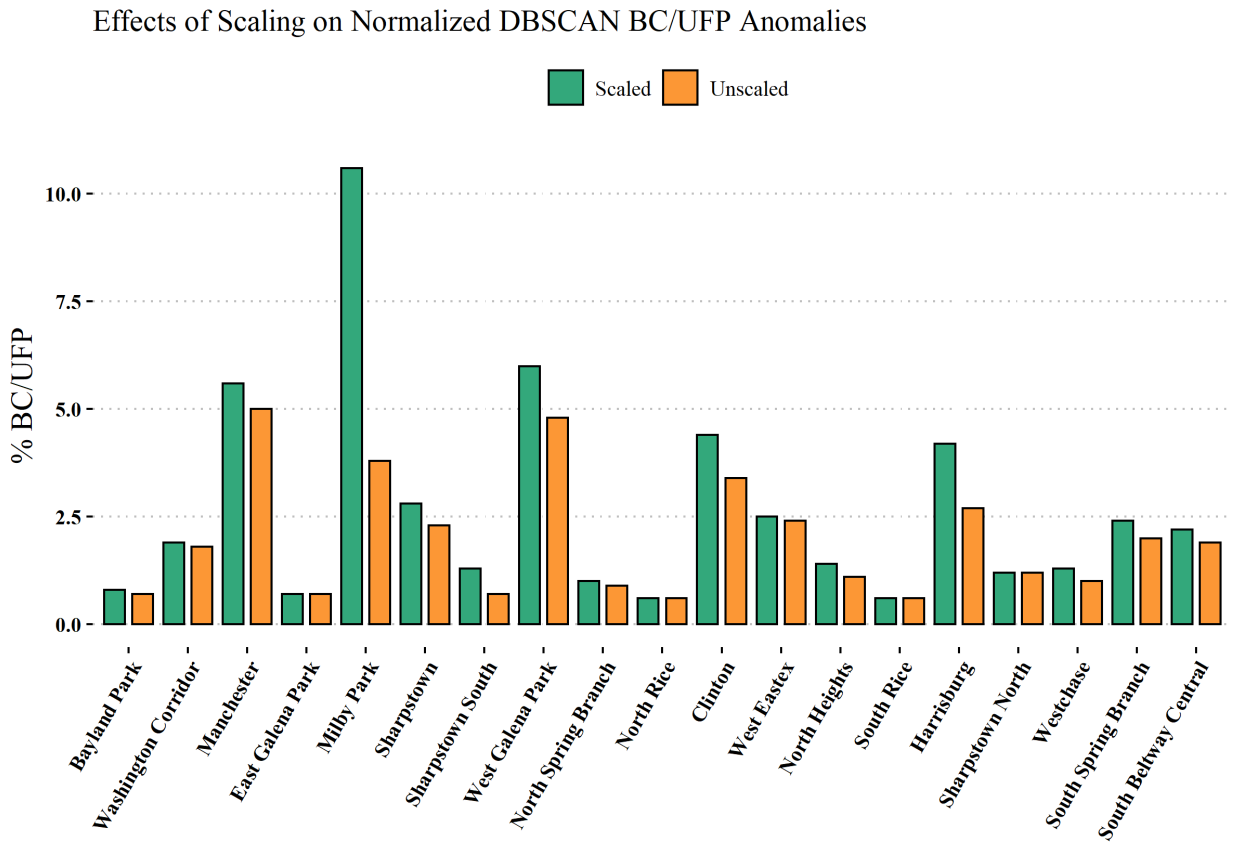

Figure S3. Effects of rescaling on probability of BC/UFP anomaly type detection for each census tract (green/left bar for each census tract is scaled).

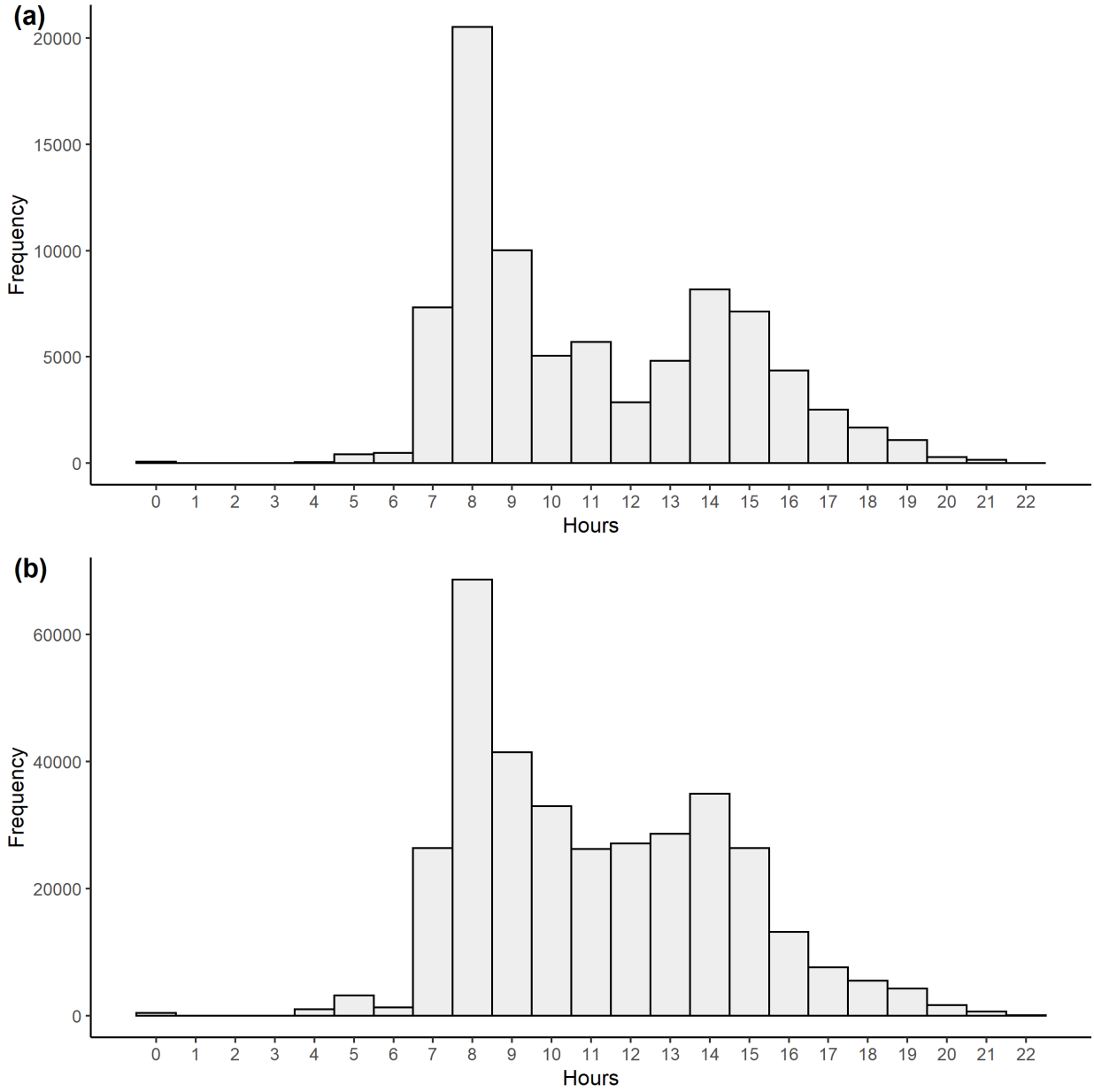

Figure S4. Sampling distributions for (a, top) all measurements and (b, bottom) anomalies in the North Rice census tract.

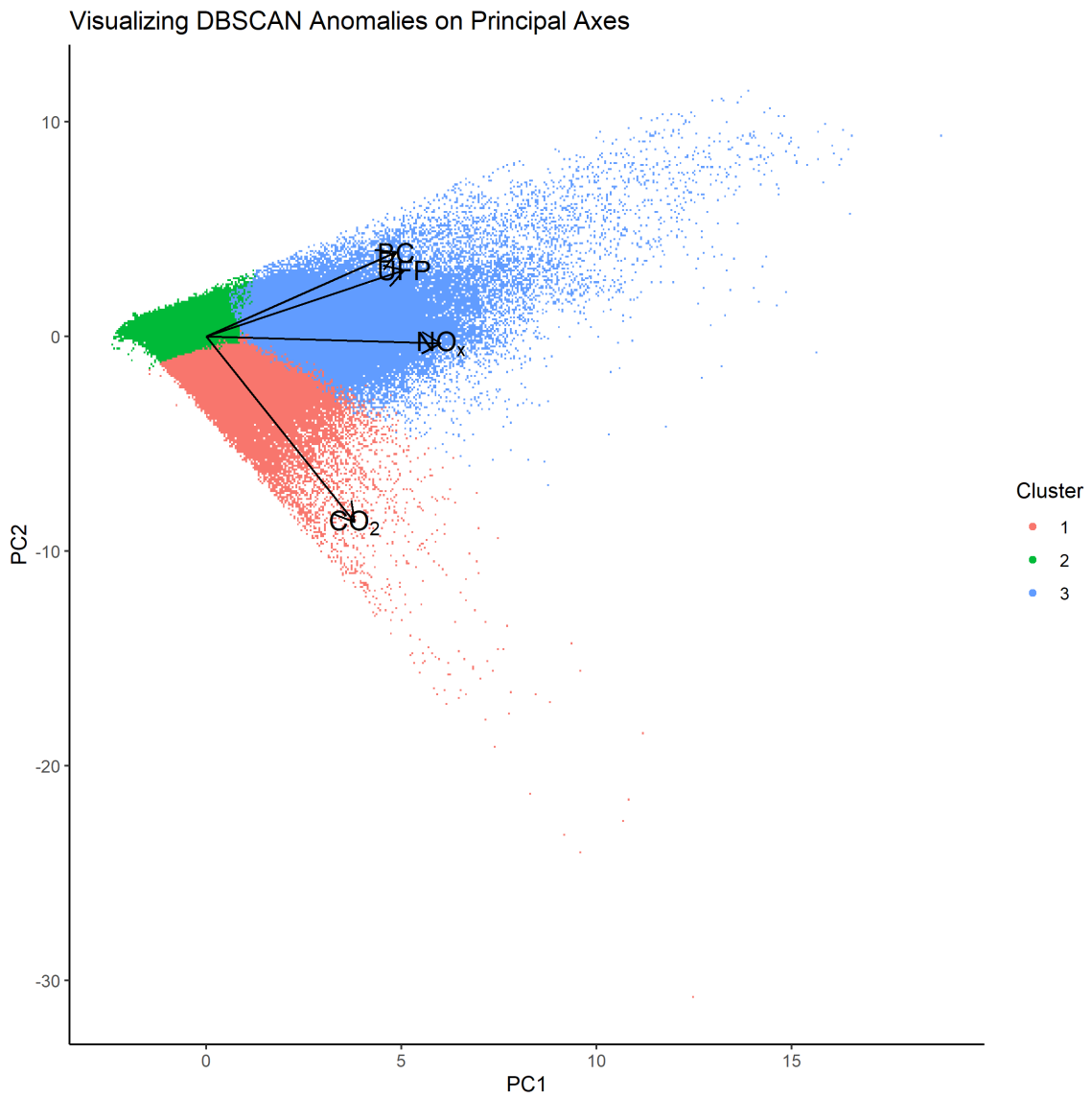

Figure S5. Visualizing cluster assignment on the first two principal component axes for DBSCAN-derived anomalies. Cluster 1 extends down and to the right from the origin, cluster 2 is around the origin, and cluster 3 extends up and to the right from the origin.

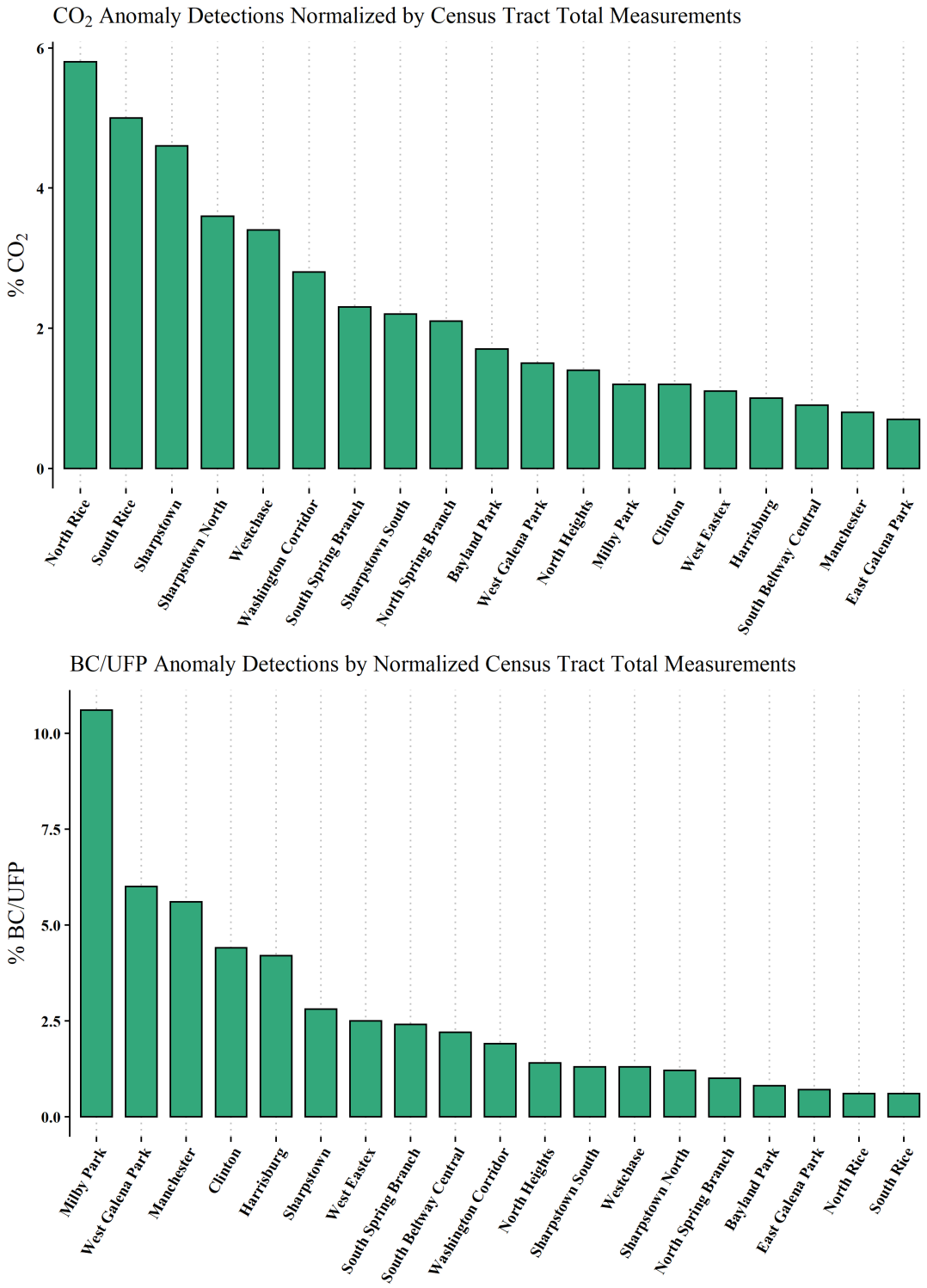

Figure S6. Total anomaly type counts per census tract normalized by the total number of measurements within each census tract. a) CO_2 (top) b) BC/UFP (bottom).

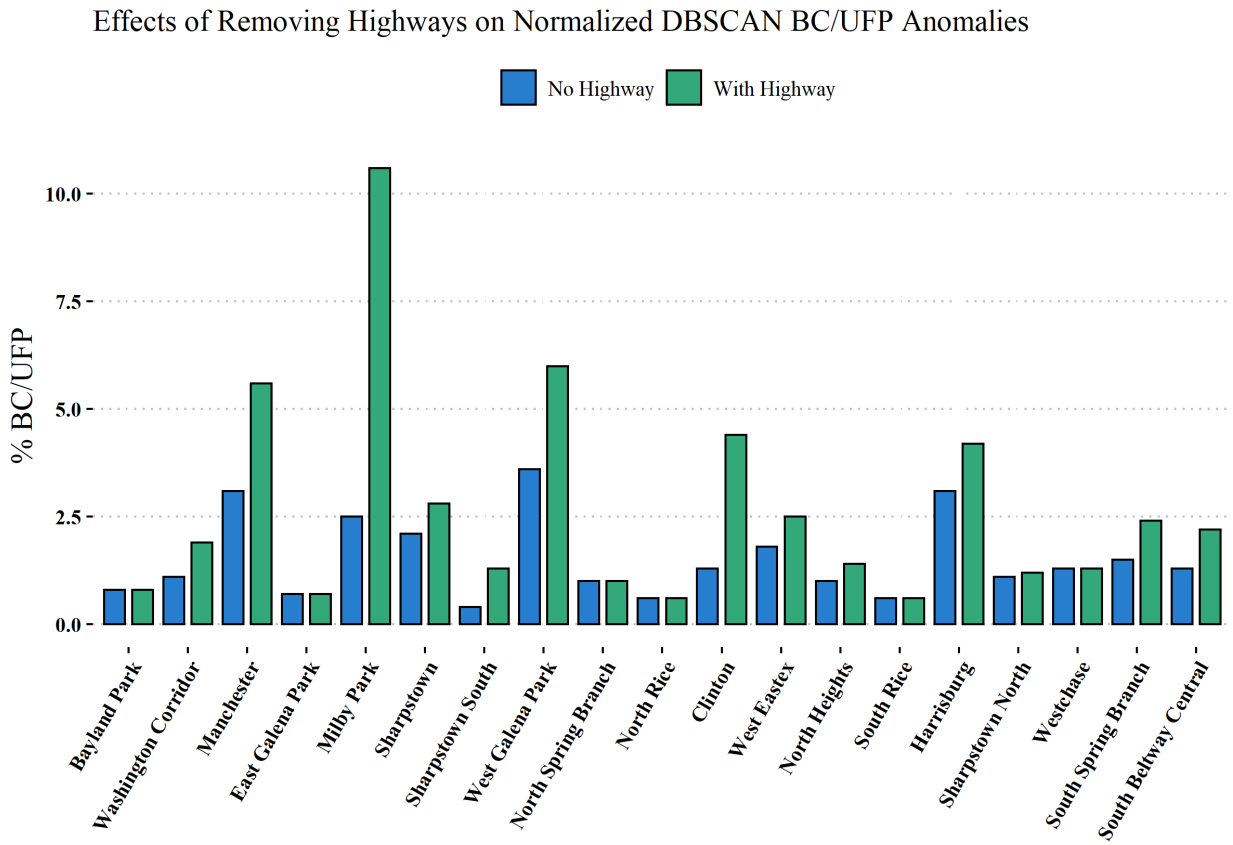

Figure S7. Probability of detecting BC/UFP anomaly type with highways in the analysis (green, right bar for each census tract) and without highways in the analysis (blue, left bar for each census tract).

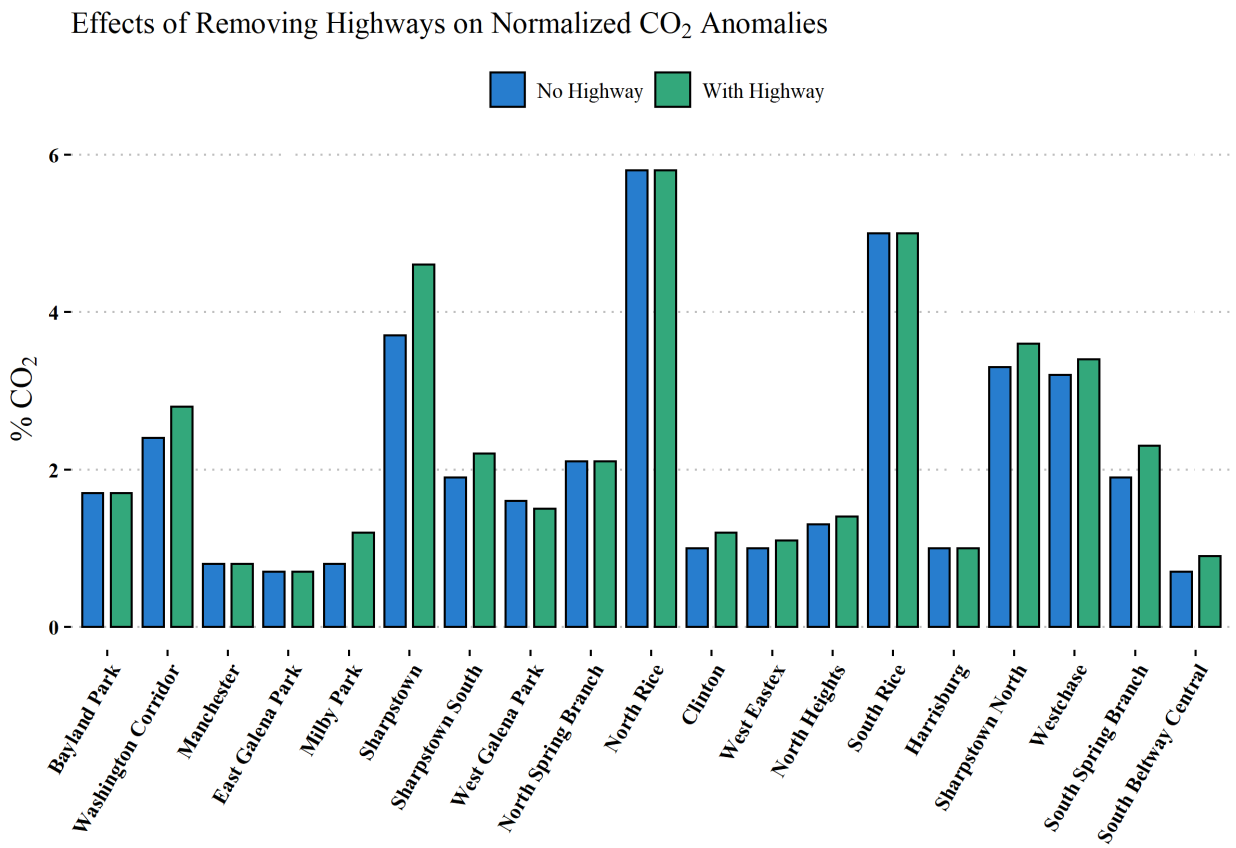

Figure S8. Probability of detecting CO_2 anomaly type with highways in the analysis (green, right bar for each census tract) and without highways in the analysis (blue, left bar for each census tract).

Table S1. Instruments used in the Houston mobile monitoring campaign.

Measured Pollutant	Instrument
Black Carbon (BC) (ng m^{-3})	Magee AE33 (Aethalometer)
Carbon Dioxide (CO_2) (ppm)	Li-COR LI-7000 CO_2/H_2O Analyzer (Spectroscopy)
Nitric Oxide (NO) (ppb)	Teledyne T200 (Chemiluminescence)
Nitrogen Dioxide (NO_2) (ppb)	Teledyne T500U (CAPS)
Ultrafine Particle Counts (UFP) (p cm^{-3})	Aerosol Dynamics MAGIC 200p (CPC)

Table S2. Cross validation results for 5 folds.

Fold	Trained f_{val}	Testing Performance (%)
1	0.01	85.05
2	0.03	85.93
3	0.03	87.39
4	0.03	84.09
5	0.03	88.57

Table S3. Error estimates for CO_2 anomaly detection type probabilities (in %) by census tract determined from a sampling distribution composed of 1000 bootstrap replicates. "Mean" is the mean of the sampling distribution, "Lower" is the 5th percentile of the sampling distribution, "Upper" is the 95th percentile of the sampling distribution, "Bias" is the originally calculated value – "Mean".

Census Tract	CO_2 Mean	CO_2 Lower	CO_2 Upper	Bias
Bayland Park	1.7	1.6	1.8	0
Washington Corridor	2.8	2.7	2.9	0
Manchester	0.8	0.8	0.9	0
East Galena Park	0.7	0.7	0.8	0
Milby Park	1.2	1.2	1.3	0
Sharpstown	4.6	4.5	4.8	0
Sharpstown South	2.2	2.1	2.3	0
West Galena Park	1.5	1.4	1.6	0
North Spring Branch	2.1	1.9	2.2	0
North Rice	5.8	5.7	5.8	0
Clinton	1.1	1.1	1.2	0.1
West Eastex	1.1	1.0	1.1	0
North Heights	1.4	1.4	1.5	0
South Rice	5.0	4.9	5.1	0
Harrisburg	1.0	1.0	1.1	0
Sharpstown North	3.6	3.4	3.7	0
Westchase	3.4	3.2	3.5	0
South Spring Branch	2.3	2.2	2.4	0
South Beltway Central	0.9	0.9	0.9	0

Table S4. Error estimates for BC/UFP anomaly detection type probabilities (in %) by census tract determined from a sampling distribution composed of 1000 bootstrap replicates. "Mean" is the mean of the sampling distribution, "Lower" is the 5th percentile of the sampling distribution, "Upper" is the 95th percentile of the sampling distribution, "Bias" is the originally calculated value – "Mean".

Census Tract	BC/UFP Mean	BC/UFP Lower	BC/UFP Upper	Bias
Bayland Park	0.8	0.8	0.9	0
Washington Corridor	1.9	1.8	2.0	0
Manchester	5.6	5.5	5.8	0
East Galena Park	0.7	0.6	0.7	0
Milby Park	10.6	10.3	11.0	0
Sharpstown	2.8	2.6	2.9	0
Sharpstown South	1.3	1.2	1.4	0
West Galena Park	6.0	5.8	6.1	0
North Spring Branch	1.0	0.9	1.0	0
North Rice	0.6	0.5	0.6	0
Clinton	4.4	4.3	4.5	0
West Eastex	2.6	2.5	2.6	-0.1
North Heights	1.4	1.4	1.5	0
South Rice	0.6	0.6	0.7	0
Harrisburg	4.2	4.0	4.3	0
Sharpstown North	1.2	1.1	1.2	0
Westchase	1.3	1.2	1.4	0
South Spring Branch	2.4	2.3	2.5	0
South Beltway Central	2.2	2.1	2.2	0

Table S5. Error estimates for Transition anomaly detection type probabilities (in %) by census tract determined from a sampling distribution composed of 1000 bootstrap replicates. "Mean" is the mean of the sampling distribution, "Lower" is the 5th percentile of the sampling distribution, "Upper" is the 95th percentile of the sampling distribution, "Bias" is the originally calculated value – "Mean".

Census Tract	Transition Mean	Transition Lower	Transition Upper	Bias
Bayland Park	8.6	8.4	8.7	0
Washington Corridor	13.3	13.2	13.4	0
Manchester	19.6	19.4	19.8	0
East Galena Park	8.6	8.5	8.8	0
Milby Park	16.7	16.4	17.1	0.1
Sharpstown	17.8	17.6	18.1	0
Sharpstown South	9.5	9.3	9.7	0
West Galena Park	16.5	16.3	16.7	0
North Spring Branch	12.0	11.7	12.2	0
North Rice	14.4	14.3	14.5	0
Clinton	20.1	19.9	20.3	0
West Eastex	12.7	12.6	12.9	0.1
North Heights	10.4	10.3	10.5	0
South Rice	13.4	13.2	13.5	0
Harrisburg	16.9	16.7	17.1	0
Sharpstown North	18.7	18.4	19.0	0
Westchase	12.7	12.5	13.0	0
South Spring Branch	13.3	13.1	13.6	0
South Beltway Central	16.3	16.2	16.4	0

Table S6. Counts of when QOR or DBSCAN outperform the other under different circumstances.

QOR Label	DBSCAN Label	Correct Label	Counts
"Anomaly"	"Normal"	"Normal"	19456
"Normal"	"Anomaly"	"Normal"	6739
"Normal"	"Anomaly"	"Anomaly"	8183
"Anomaly"	"Normal"	"Anomaly"	12174

Table S7. Loadings post varimax rotation from Fig. S5. Varimax rotated loadings from Larson et al. (2017) are also presented for reference.

	CO_2-rich (This work)	CO-rich (Larson)	BC-rich (This work)	BC-rich (Larson)
BC	-0.02	0.09	0.76	0.88
CO_2	0.97	0.76	0.07	0.19
NO_x	0.42	0.69	0.70	0.62
UFP	0.08	0.26	0.75	0.87

Table S8. Census tract characteristics reprinted from Actkinson et al. (2021). Data taken from U.S. Census (2010) and Environmental Defense Fund (2020).

Census Tract	Population Total	# Metal Recyclers	# Concrete Batch Plants	# Petrochemical Facilities	Area (sq. miles)	# Facilities (sq. mi)$^{-1}$
North Spring Branch	5126	0	0	0	0.57	0
South Spring Branch	3604	0	0	0	0.73	0
Washington Corridor	5432	2	0	0	1.39	1.44
West Eastex	2753	5	2	0	1.42	4.93
North Heights	6472	1	0	0	1.18	0.85
Westchase	5548	0	0	0	0.70	0
Sharpstown	5616	0	0	0	0.50	0
Sharpstown North	3484	0	1	0	0.56	1.79
Sharpstown South	5196	0	0	0	0.94	0
Bayland Park	5083	0	0	0	0.71	0
South Beltway Central	2530	3	8	0	12.28	0.90
North Rice	2892	0	0	0	0.58	0
South Rice	5355	0	0	0	0.93	0
Clinton	2127	2	1	1	1.50	2.67
West Galena Park	5245	0	0	0	2.90	0
East Galena Park	3000	0	0	0	0.97	0
Manchester	1647	0	0	1	2.80	0.36
Harrisburg	1496	2	0	2	1.01	3.96
Milby Park	6662	0	0	0	1.61	0

References

Actkinson, B., Ensor, K., and Griffin, R. J. SIBaR: A new method for background quantification and removal from mobile air pollution measurements, Atmos. Meas. Tech., 14 (8), 5809–5821, doi.org/10.5194/amt-14-5809-2021, 2021.

Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*; Chapman and Hall/CRC: New York, doi.org/10.1201/9780429246593, 1994.

Environmental Defense Fund, Finding pollution—and who it impacts most—in Houston, https://www.edf.org/maps/airqualitymaps/houston/pollution-map/ (accessed 2020 -11 -23).

Larson, T., Gould, T., Riley, E. A., Austin, E., Fintzi, J., Sheppard, L., Yost, M., and Simpson, C., Ambient air quality measurements from a continuously moving mobile platform: Estimation of area-wide, fuel-based, mobile source emission factors using absolute principal component scores, Atmos. Environ., 152, 201–211, doi.org/10.1016/j.atmosenv.2016.12.037, 2017.

U.S. Census, Census 2010 Tracts, https://cohgis-mycity.opendata.arcgis.com/datasets/census-2010-tracts (accessed 2020 -11 -23).