Detecting Plumes in Mobile Monitoring Time Series with Density-based Spatial Clustering of Applications with Noise (DBSCAN)

Blake Actkinson1, Katherine Ensor2, Robert J. Griffin3\*

1Department of Civil and Environmental Engineering, Rice University, Houston, TX 77005, USA

2Department of Statistics, Rice University, Houston, TX 77005, USA

3School of Engineering, Computing, and Construction Management, Roger Williams University, Bristol, RI 02809, USA

\*Corresponding Author: Robert Griffin, *E-mail:* [*rgriffin@rwu.edu*](mailto:rgriffin@rwu.edu)

**ABSTRACT**

Mobile monitoring is becoming an increasingly popular technique to analyze air pollution measurements on fine spatial scales, but development of appropriate methods to analyze source contributions is sorely needed. One approach is to isolate and analyze plumes from the rest of the mobile monitoring time series, but there is also a research gap in the availability of methods suitable for large mobile monitoring time series. Here we discuss a novel method used to detect and isolate plumes from an extensive mobile monitoring data set. The method relies on Density-based Spatial Clustering of Applications with Noise (DBSCAN), an unsupervised machine learning technique. The method systematically runs DBSCAN on mobile monitoring time series by day and identifies points as anomalies for further analysis. Analyzed anomalies reveal patterns previously associated with different types of vehicle emission profiles. We tabulate our anomalies and reveal striking disparities in their type by census tract. The method serves to further improve upon previously published plume detection algorithms and results can be used to inform stakeholders of spatial variations in emission profiles not obvious through the use of stationary monitors alone.

**TOC ART**

**Map

Description automatically generated**

1. **INTRODUCTION**

A central question of air pollution studies is to identify the sources which contribute to measured pollutant concentrations. This question becomes more complicated in a mobile monitoring context because measurements and concentrations vary as a function of both space and time, making conventional source apportionment techniques such as PMF and PCA harder to apply effectively.1

Past work has taken several approaches to analyze source effects on measured pollutants. One approach involves using PCA on source contributions to mobile measurements, such as in Larson et al.1 The approach described in their manuscript is problematic for extensive mobile monitoring campaigns, in that it involves defining a rolling minimum across a static time window which may not be feasible for extensive mobile monitoring campaigns with 20x the temporal coverage. Other approaches have focused on using Land Use Regression (LUR) models to identify relationships between pollutants and land use variables, such as in Messier et al.2 The problems with LUR are threefold: (i) they often involve aggregating across time and thus lose temporal resolution (though this is being addressed through the development of spatiotemporal models – see Hankey et al.3) ; (ii) derived relationships do not necessarily imply causality; and (iii) LUR use is limited by the availability of sophisticated land use variable databases. There is a need for the development of methods that can identify source influences in large mobile monitoring data sets at high time resolution and not subject to the availability of pre-existing, sophisticated databases.

Another factor that aggravates source identification in mobile monitoring contexts can be the nature of mobile monitoring data itself. If a mobile monitoring campaign were conducted focusing largely on residential areas with brief excursions onto busier thoroughfares, performing a PCA or other dimension reduction techniques to describe patterns in the data would likely return results that are weighted towards data in which source influences are largely non-existent. This type of analysis could frequently generate solutions in which there is a demarcation between the majority of points with little source influences and a smaller subset of source-influenced points, which may not be compelling if one’s objective is to determine the specific sources affecting the measurements.

This raises the question of how to identify source influences within mobile monitoring time series that cover locations that range from ‘background’ to ‘highly influenced by sources.’ If one could identify source spikes, or plumes within mobile monitoring time series, one could restrict their analysis to these plumes to categorize the different types of sources that affected their mobile monitoring measurements. Plume identification within mobile monitoring time series has been addressed previously. Hagler et al. addressed the plume identification question through use of a rolling coefficient of variation across a 5-s time interval, then flagging points with a coefficient greater than 2.4,5 Drewnick et al. used a different moving window algorithm which involves calculating the standard deviation of points below a defined background threshold and flagging points which are greater than 3 times the background standard deviation of the previous point. The algorithm would then flag subsequent points, increasing the threshold necessary for flagging for every subsequent point beyond the first flagged.6 Others have addressed the plume identification question indirectly through background estimation and removal methods, such as Larson et al. and Choi et al. 1,7

The above plume identification methods suffer from their own drawbacks. In this work, the Hagler et al. method flags little to no points at all, suggesting that the method is sensitive to the time series utilized. The Drewnick et al. algorithm suffers in situations where many plumes appear consecutively to one another, frequently leading to poor performance in those circumstances.6 The methods of Larson et al. and Choi et al. are time window dependent, which presents problems for complex, multi-day mobile monitoring time series previously discussed in Chapter 4.

The following manuscript discusses an algorithm created to address the plume identification problem in a different manner. The algorithm relies on Density-based Spatial Clustering of Applications with Noise (DBSCAN), a clustering algorithm originally conceived by Ester et al.8 DBSCAN clusters points based on whether they fall into predetermined neighborhoods with other points and is thus a nearest-neighbor clustering algorithm. The technique clusters points with complex shapes and is not susceptible to problems with starting values that are encountered with other clustering techniques such as k-means. Additionally, the algorithm does not require every single point to be clustered, allowing for those points that do not neatly fall into a given cluster to be defined as noise.

This work describes a novel algorithm utilizing DBSCAN by envisioning daily multivariate mobile monitoring time series consisting of BC, CO2, NOx and UFP as large numbers of points clustered around a 4-dimensional origin with plumes being labeled as noise by the DBSCAN algorithm. After describing DBSCAN, we describe an algorithm adapting it for mobile monitoring time series. To evaluate performance, we construct a validation set by manually flagging plumes via visual inspection from a randomly chosen subset of days from the Houston mobile monitoring campaign. We use the validation set to tune the DBSCAN and other time series-based models and compare the performances of all models. We apply the algorithm to our mobile monitoring dataset to identify anomalies, which are then clustered into anomaly types evidence suggests are related to heavy-duty and light-duty vehicle influences. We tabulate the number of these different anomaly types by census tract and discuss the implications of the method, the results, and future directions for this research.

1. **METHODS**
   1. **Data**

Data were collected during the Houston mobile monitoring campaign and are described in detail elsewhere. In brief, 2 Google Street View cars drove different census tract in the Greater Houston area systematically to evaluate spatial differences in the concentrations of 7 pollutants. In this analysis, we constrict the set of analyzed pollutants to be Black Carbon (BC), Carbon Dioxide (CO2), Ultrafine Particle Counts (UFP), Nitric Oxide (NO), and Nitrogen Dioixde (NO2). NOx is defined to be the sum of NO and NO2. We create a multivariate dataset consisting of these 4 air pollution variables, along with corresponding latitude/longitude coordinates and timestamps that spans 277 separate days of sampling, for a total of 5,301,507 observations. BC data were smoothed with a 10-s time window to limit the effects of noise on subsequent analysis. Missing data were imputed up to a 6-s time gap with monotone Hermitian splines. Any multivariate realization with at least one missing observation in a variable not imputed was excluded otherwise. Days had to possess a minimum of 600 realizations to be included in the analysis. We constrict our analysis to road segments with road classes S1100, S1200, S1400, S1630, S1640, and S1730, to points with logged GPS coordinates <30 m away from a TigerLINE shapefile point, and to points > 30 m from a dead end for potential self-sampling contamination for consistency with Miller and Actkinson et al.9,10 A table providing instruments used to measure each respective pollutant is given in Table S1.

The following R packages were used in the analysis and visualization of results: tidyverse,11 ggpubr,12 caret,13 dbscan,14 leaflet,15 leafem,16 sf,17 mapview,18 and data.table.19

* 1. **DBSCAN – A Brief Description**

DBSCAN is a clustering routine originally conceived by Ester et al.8 Using two predefined parameters, epsilon () and MinPts, DBSCAN seeks to label points that have MinPts points within a neighborhood defined with radius as core points, points that do not meet the MinPts criteria but have a core point within their - neighborhood as border points, and points that do not fit either above criteria as noise.

More formally, the -neighborhood is defined using the notation of Hahsler14 as

where is the neighborhood, is the set of points, and is a distance measure such as the Euclidean. A point is defined as a core point if

where MinPts is the minimum points parameter and || denotes cardinality. The algorithm systematically labels points as core points, border points, or noise points depending on these criteria.

* 1. **Algorithm Description**

An algorithm was created to systematically label anomalies within our mobile monitoring time series incorporating DBSCAN. A flowchart depicting this algorithm’s steps is given in Figure 1. The algorithm estimates eps () and MinPts parameters for daily time series in the campaign via statistical heuristics and performs DBSCAN using these estimated parameters. The MinPts parameter was reparameterized to be the product of the total number of points in the daily time series, , and a defined fractional value . More formally,

Where

An appropriate value of = 0.03 was determined using the external validation set and is described in Section 2.5. 0.5 were not considered due to rapidly increasing computational cost and apparent poor performance at higher values.

In addition to MinPts, DBSCAN also requires . was determined by a k-nearest-neighbor distance ordering procedure in which the value of k was set to be equal to MinPts. An ordered k-nearest-neighbor distance graph was constructed and the mean and standard deviation of the first 30 ordered distances determined. Then, was determined to be the first distance in the ordered set that met the criteria:

where is the ordered k-nearest-neighbor distance, is the mean of the ordered distances between indices 1 and , and is the standard deviation of ordered distances

between 1 and . A graphical example illustrating the estimated epsilon value for a knn ordered distance graph is given in Figure 2. Given both and MinPts, DBSCAN was run on the daily time series observations in which core points were labeled as normal and both border and noise points labeled as anomalies. An example of labeled DBSCAN output for a scatterplot of daily BC/CO2 time series is given in Figure 3.

Diagram

Description automatically generated

**Figure 1.** Flowchart depicting steps of the DBSCAN anomaly detection algorithm for mobile monitoring datasets.

Diagram

Description automatically generated with medium confidence

**Figure 2.** Graphical depiction of the selection criteria.

Chart, scatter chart

Description automatically generated **Figure 3.** Scatterplot example of DBSCAN labeled anomalies for CO2 against BC.

* 1. **Validation Set Construction**

To tune parameters and evaluate algorithm performance, a validation set was constructed from the mobile monitoring data by manually flagging visible plumes within 30 randomly selected mobile monitoring time series out of a total of 277. The total number of points in the validation set was 564,107, which amounts to 10% of the entire set (5,712,090 total points). A graphical user interface in IgorPro was used to flag plumes by visually inspecting the time series for spikes in pollutant concentrations for each pollutant (BC, CO2, NOx, UFP).

**2.4 Description of Other Algorithms**

To put the performance of the DBSCAN anomaly detection algorithm in context, we compare its labeled anomalies with output from other plume detection techniques previously published in the literature6 or base-case 90th-quantile algorithms.

We compare our plume detection algorithm’s performance to an algorithm described in Drewnick et al. (“Drewnick”).6 The Drewnick algorithm flagged points as anomalous based on outputs from moving window functions across CO2 and UFP measurements. The algorithm first calculates the standard deviation of all points in the time series below the median, which is assumed to be representative of the standard deviation of background measurements (). The algorithm then flags points as anomalous if the current point is greater than the previous point + 3\*, and continues to flag subsequent points as part of the same plume if they are also greater than that same previous point + 3\* + , where is the number of points between the last unflagged point and the point currently being considered for flagging. For more details on the algorithm, consult Drewnick et al.6

Two base-case algorithms were also created and evaluated. The two algorithms, the Quantile-OR (QOR) and the Quantile-AND (QAND) algorithms, flag points as anomalous based on the 90th quantile of pollutant distributions. In the QOR case, points are flagged as anomalous if any one pollutant measurement of BC, , , or UFP is above the 90th quantile for the given daily time series (if > 90th BC OR > 90th OR > 90th OR > 90th UFP, flag as anomaly). In the QAND case, points are flagged as anomalous if all pollutant measurements are greater than their respective 90th quantiles (if > 90th BC AND > 90th AND > 90th AND > 90th UFP, flag as anomaly). These algorithms, along with the Drewnick algorithm, were run on all daily time series to assess performance.

The quantile parameter was explored via tuning with the external validation set. Quantiles near the 90th were found to return only modest improvements, and thus the 90th quantile was analyzed for simplicity.

**2.5 Using the External Validation Set to Tune Parameters**

To determine an appropriate value of , we perform grid search on values in [0.01,0.10] in increments of 0.01 and [0.15,0.50] in increments of 0.05. Values above 0.5 were not considered due to computational cost and apparent poor performance at higher values of . Performance was validated using percentage agreement, defined as

Where is the indicator function which evaluates to 1 if the condition is true and 0 otherwise, is the prediction label at point , is the validation set label at point , and is the total number of points in the validation set. Tuning results indicated that a value of 0.03 was most appropriate for , which was used in subsequent analyses. To evaluate whether we overfit to this validation set, we perform k-fold cross validation with the number of folds k = 5. We train our models on 4 out of 5 folds, tuning the parameter such that the model performance agreement is maximized on the testing set. We find that the value of which results in superior performance to be 0.03, suggesting that our work above generalizes appropriately. K-fold cross validation results are given in Table S2.

**2.6 Interpretation: k-means Clustering and Principal Component Analysis**

To interpret our results, we perform k-means clustering on the extracted anomalies in R using the kmeans function available in base.20 The number of centers (clusters) was set to 3 and 200 iterations with different random starts were chosen to ensure the derived result was robust. Cluster labels were assigned based on the cluster means to ensure consistency in label assignment.

Boxplots of assigned roadway trucking variables were created to probe potential meanings of clustered anomalies. Roadway trucking variables were extracted from TxDOT’s roadway inventory with processing performed using R package sf.21 Records along the same road segment were averaged with weights equivalent to the distance between fields FROM\_DFO and TO\_DFO. Extracted roadway variables from the shapefile included Truck Annual Average Daily Traffic Percentage (TRUCK\_AADT\_PCT) and the number of all trucks in AADT (AADT\_TRUCKS).

**2.7 Census Tract Assignment**

To determine differences in anomaly frequency between census tracts, points were assigned to census tracts using tract boundaries stored in a shapefile used in previously published analyses of the same campaign data.9,22 Anomalies of a given cluster assignment were counted and divided by the total recorded measurements in each polygon. Since each census tract was imperfectly sampled, we implement a rescaling procedure described in detail in the SI. As part of that procedure, we constrict the comparisons to 19 of the 35 census tracts, to measurements taken between 8 AM and 4 PM, and to measurements taken on weekdays. To account for different polygons containing differing amount of measurements, we divide the total amount of rescaled anomaly types by the total number of measurements made in the census tract, in essence deriving a probability of encountering the specified anomaly type during the campaign in the constricted time interval above.

1. **RESULTS**

**3.1 External Validation**

All four algorithms: Drewnick, QOR, QAND, and DBSCAN were run on the Houston mobile monitoring campaign data. To compare performance, each algorithm’s labeled anomalies were compared with the anomalies of the validation set on the same subset of days, which were taken to be the ground truth.

Out of the 4 algorithms, DBSCAN had the best performance, with its labels exhibiting 86.9% with the validation set’s labels. The QOR, QAND, and Drewnick algorithms exhibit 85.5%, 77.0%, and 81.8% agreement, respectively. To put these results into context, an algorithm that simply labeled all points as normal would generate 74.7% agreement with the validation set. Because this baseline agreement is so high, confusion matrices were created to probe more fine-grained sources of agreement and disagreement between each algorithm’s predicted anomalies and the validation set labeled anomalies. Confusion matrices for the four algorithms are presented in Figure 4.

Figure 4 illustrates that even though the DBSCAN algorithm exhibits greater overall agreement with the validation set, it contains less matching predicted anomalies compared with the QOR algorithm. The DBSCAN algorithm outperforms the QOR algorithm in its ability to not predict normal points as anomalous. The results suggest that the QOR algorithm captures the most anomalies but is a coarse approach to doing so; the DBSCAN algorithm captures fewer anomalies but is less likely to predict something as anomalous when it is likely not. The QAND and Drewnick algorithms do not offer superior performance over the DBSCAN and QOR algorithms in either of Chart, bar chart

Description automatically generatedthese regards, and so will not be considered for further analysis.

**Figure 4.** Confusion matrices corresponding to the performance of (a) DBSCAN, (b) Drewnick, (c) QOR, and (d) QAND. Overall agreement between each algorithm and the validation set was (a) 86.9%, (b) 85.5%, (c) 81.8%, and (d) 77.0%

**3.2 K-means Clustering and Principal Components**

Detected anomalies were clustered using kmeans and consistently yielded one cluster rich in concentrations (“CO2 Cluster”), another cluster rich in (“BC/UFP Cluster”) concentrations, and a third cluster that contains lower emissions of all 4 for both QOR and DBSCAN derived anomalies (“Transition Cluster”). Table 1 contains cluster means of the three derived clusters from QOR and DBSCAN anomalies. Figure 5 shows cluster assignments of DBSCAN labeled anomalies projected onto the first 2 principal component axes. Figure 6 contains boxplots of the clustered DBSCAN pollutant anomalies. In Figure 5, it is evident that the CO2 cluster points in the direction of CO2 loadings, while the BC/UFP cluster points in the direction of the BC/UFP loadings. These loadings when varimax rotated are consistent in split with those reported in Larson et al. but differ in value and are given in Table 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | (ppm) | BC (ng/m3) | (ppb) | UFP (p/cc) |
| DBSCAN | | | | |
| 1st cluster | 555.65 | 1892.85 | 73.10 | 16297.67 |
| 2nd cluster | 443.69 | 1539.54 | 42.62 | 15411.41 |
| 3rd cluster | 493.37 | 6325.59 | 178.55 | 50244.41 |
| QOR | | | | |
| 1st cluster | 547.32 | 2142.34 | 83.18 | 17462.55 |
| 2nd cluster | 443.60 | 1597.30 | 41.98 | 16615.62 |
| 3rd cluster | 494.57 | 6638.60 | 184.43 | 51112.113 |

**Table 1.** DBSCAN and QOR k-means cluster means.

Chart, scatter chart

Description automatically generated**Figure 5.** Visualzing cluster assignment on the first two principal component axes for DBSCAN-derived anomalies.

**Chart

Description automatically generated with medium confidenceFigure 6.** Boxplots of clustered DBSCAN anomalies by cluster label. Red rectangles correspond to insets of CO2 and BC that are displayed on the right side of the plot.

**Table 2.** Loadings post varimax rotation from Figure 5. Varimax rotated loadings from Larson et al. are also presented for reference.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CO2-rich (This work) | CO-rich1 | BC-rich (This work) | BC-rich1 |
| BC | -0.02 | 0.09 | 0.76 | 0.88 |
| CO2 | 0.97 | 0.76 | 0.07 | 0.19 |
| NOx | 0.42 | 0.69 | 0.70 | 0.62 |
| UFP | 0.08 | 0.26 | 0.75 | 0.87 |

To probe for further interpretation of these clusters, we extract traffic variables from the TxDOT roadway inventory and assign these values to our clustered anomalies based on nearest neighbor assignment between the logged GPS coordinates of each clustered point and the latitude/longitude coordinates of the shapefile’s features.21 We plot these assignments in Figure 7. Panel (a) in Figure 7 contains the overall annual average daily traffic counts. Panel (b) in Figure 7 shows percentages of trucks in the estimated annual average daily traffic counts. The boxplots illustrate how a possible interpretation of the 3rd cluster corresponds to trucking activity, while an interpretation of the 1st cluster corresponds to gasoline-powered vehicle activity. The high percentage of trucks in AADT in the 3rd cluster clearly point to trucking activity, while the lower trucking percentage in combination with elevated AADT compared to the 2nd cluster suggest that the 1st cluster could be capturing some light duty gasoline vehicle activity. These assignments, in conjunction with their loading assignments, are consistent with previously published studies of vehicle source apportionment.1,23,24

**Chart, box and whisker chart

Description automatically generatedFigure 7.** Boxplot of traffic attributes corresponding to anomalies in labelled clusters (1- “CO2 Cluster”, 2 – “Transition Cluster”, 3 – “BC/UFP Cluster”). (a) Annual average daily traffic (AADT) by cluster label. (b) Percentages of trucks in the annual average daily traffic counts (AADT % Truck).

**3.3 Detected Anomaly Type by Census Tract**

We evaluate spatial differences in these clustered anomaly types around the city of Houston by tabulating anomaly types for visited census tracts. We report rescaled total number of detected anomalies of a given cluster type (“CO2 Cluster” for CO2 rich, “Transition Cluster”, “BC/UFP Cluster”) divided by the total number of measurements made in that census tract. Normalizing by the total number of measurements in this manner results in the probability of encountering that anomaly in the census tract during the study period, which is from 8 AM to 3 PM during weekdays. Figure 8 displays bar plots showing detected DBSCAN anomaly type by each census tract, while Figures 9 and 10 map the census tracts colored by their CO2 and BC/UFP anomaly detection probabilities.

**Chart, bar chart, histogram

Description automatically generatedFigure 8.** Bar plots displaying total anomaly type counts per census tract normalized by the total census tract measurements.

**Map

Description automatically generatedFigure 9.** Map depicting analyzed census tracts colored by their calculated CO2 anomaly detection probabilities (%).

**Map

Description automatically generatedFigure 10.** Map depicting analyzed census tracts normalized by their calculated BC/UFP anomaly detection probabilities (%).

The bar plots and maps illustrate stark spatial divides in distributions of anomaly type. With regards to cluster anomalies, neighborhoods in the Western parts of Houston (North Rice, South Rice, South River Oaks) consistently rank higher than neighborhoods in the Eastern part of Houston, with neighborhoods surrounding Rice University ranking the highest. In particular, the neighborhoods near Rice campus consists of busy thoroughfares that are often congested with traffic from light duty gasoline powered vehicles, especially around rush hour (8 AM US/Central). With regards to the BC/UFP clusters, heavily industrialized neighborhoods in the Eastern part of Houston (Milby Park, West Galena Park, Manchester, Clinton) are ranked the highest, with the Milby Park census tract exhibiting the highest probability of encountering one of these anomaly types (10.6%) during the study period. Many of these anomaly detection types were found to occur on highway; Figure S4 illustrates the changes in BC/UFP anomaly detection probabilities between when highways are included and excluded from the analysis. Nonetheless, even with highways removed from the analysis, neighborhoods in the Eastern part of Houston still rank consistently higher compared to neighborhoods in the western part of Houston. The mapped census tracts serve to paint a picture of two cities: one that is CO2 dominated, and one that is BC/UFP dominated, and suggest that residents living in different census tracts have a higher probability of exposure to different emissions profiles. To underscore differences in exposure, Table 3 contains probabilities of detecting each anomaly type by census tract. From the highlighted entries in Table, it is evident that there was a 10x greater chance of encountering a BC/UFP anomaly type in the Manchester census tract compared to encountering the same anomaly type in the North Rice census tract. Given the evidence suggesting that the BC/UFP anomalies are closely related to heavy duty vehicles and that previous modeling studies have shown large contributions of air pollution in Houston’s Ship Channel neighborhoods can be attributed to heavy-duty vehicles,25 these results are consistent with what has been previously published.

**Table 3.** Tabulated anomaly detection probability type (“CO2 – rich” = “CO2”, “Transition” = “Transition”, “BC/UFP – rich” = “BC/UFP) by census tract.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Census Tract | CO2 % | Transition % | BC/UFP % | Total |
| Bayland Park | 1.7 | 8.6 | 0.8 | 138367 |
| Washington Corridor | 2.8 | 13.3 | 1.9 | 206611 |
| Manchester | 0.8 | 19.6 | 5.6 | 97374 |
| East Galena Park | 0.7 | 8.6 | 0.7 | 77046 |
| Milby Park | 1.2 | 16.8 | 10.6 | 110019 |
| Sharpstown | 4.6 | 17.8 | 2.8 | 80560 |
| Sharpstown South | 2.2 | 9.5 | 1.3 | 114595 |
| West Galena Park | 1.5 | 16.5 | 6.0 | 134501 |
| North Spring Branch | 2.1 | 12.0 | 1.0 | 100391 |
| North Rice | 5.8 | 14.4 | 0.6 | 263585 |
| Clinton | 1.2 | 20.1 | 4.4 | 185196 |
| West Eastex | 1.1 | 12.8 | 2.5 | 144963 |
| North Heights | 1.4 | 10.4 | 1.4 | 246103 |
| South Rice | 5.0 | 13.4 | 0.6 | 139313 |
| Harrisburg | 1.0 | 16.9 | 4.2 | 127736 |
| Sharpstown North | 3.6 | 18.7 | 1.2 | 98743 |
| Westchase | 3.4 | 12.7 | 1.3 | 68620 |
| South Spring Branch | 2.3 | 13.3 | 2.4 | 78195 |
| South Beltway Central | 0.9 | 16.3 | 2.2 | 311589 |

1. **DISCUSSION AND FUTURE DIRECTIONS**

We discuss the successful development of an anomaly detection algorithm based on DBSCAN that offers improvements over other plume detection alternatives. While previous work has implemented DBSCAN in conjunction with deep learning models to analyze satellite PM2.5 measurements26 or used it to define microenvironments in air pollution exposure contexts,27 this is the first study (to our knowledge) to incorporate DBSCAN in plume detection problems. The algorithm offers comparable, if not superior, performance to previously published plume detection techniques for mobile monitoring time series and is justified in analyses warranting a conservative approach. In this work, we use the algorithm to illustrate different emission profiles in census tracts around the city of Houston and showcase striking disparities. In particular, we show how BC/UFP anomaly types were 10x more likely to be detected in census tracts in the Eastern part of Houston compared to neighborhoods in the western part of Houston. While it is not definitive that this cluster type represents impacts from heavy duty vehicles, for there is no observational evidence to connect those observations to those vehicle types directly, the data displayed in Figures 5 and 6 are consistent with previously published studies analyzing emissions from light and heavy-duty vehicles (e.g. Larson et al. and references therein1), as well as with modeling studies elucidating the large impacts of trucking on PM formation in the Houston Ship channel area (e.g. Zhang et al. 25). It is evident that more investigation is needed into the trucking activity in ship channel neighborhoods.

There are a variety of steps that could be taken to improve this algorithm. Future studies could evaluate this algorithm through using different external validation sets; presence of an observer writing down witnessed emission events could result in easier and more thorough construction of a validation set. Other nearest neighbor clustering techniques aside from DBSCAN could be explored; local outlier factors could be used to address situations where DBSCAN does not exhibit great performance.28 Finally, an ensemble approach utilizing both DBSCAN and other clustering techniques could be investigated for improved performance.6,10

ASSOCIATED CONTENT

**Supporting Information**.

Section S1 describes rescaling procedure for census tract comparisons.

Tables S1-S3.

Figures S1-S5

AUTHOR INFORMATION

Corresponding Author

\*Rob Griffin (rgriffin@rwu.edu)

Author Contributions

BA1 conceived, wrote, and analyzed the plume detection algorithm with helpful insight from RG3 and KE2. BA1 wrote the manuscript. RG3 and KE2 provided helpful edits and suggestions.

Funding Sources

Authors Blake Actkinson, Katherine Ensor, and Robert Griffin all received funding from NIEHS grant #R01ES028819-01.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of NIEHS (grant #R01ES028819-01). Additionally, we appreciate the support of Environmental Defense Fund for the collection and provision of the mobile data used to develop this algorithm.

REFERENCES

(1) Larson, T.; Gould, T.; Riley, E. A.; Austin, E.; Fintzi, J.; Sheppard, L.; Yost, M.; 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.* **2017**, *152*, 201–211. https://doi.org/10.1016/j.atmosenv.2016.12.037.

(2) Messier, K. P.; Chambliss, S. E.; Gani, S.; Alvarez, R.; Brauer, M.; Choi, J. J.; Hamburg, S. P.; Kerckhoffs, J.; LaFranchi, B.; Lunden, M. M.; Marshall, J. D.; Portier, C. J.; Roy, A.; Szpiro, A. A.; Vermeulen, R. C. H.; Apte, J. S. Mapping Air Pollution with Google Street View Cars: Efficient Approaches with Mobile Monitoring and Land Use Regression. *Environ. Sci. Technol.* **2018**, *52* (21), 12563–12572. https://doi.org/10.1021/acs.est.8b03395.

(3) Hankey, S.; Sforza, P.; Pierson, M. Using Mobile Monitoring to Develop Hourly Empirical Models of Particulate Air Pollution in a Rural Appalachian Community. *Environ. Sci. Technol.* **2019**, *53* (8), 4305–4315. https://doi.org/10.1021/acs.est.8b05249.

(4) Brantley, H. L.; Hagler, G. S. W.; Kimbrough, E. S.; Williams, R. W.; Mukerjee, S.; Neas, L. M. Mobile Air Monitoring Data-Processing Strategies and Effects on Spatial Air Pollution Trends. *Atmospheric Meas. Tech.* **2014**, *7* (7), 2169–2183. https://doi.org/10.5194/amt-7-2169-2014.

(5) Hagler, G. S. W.; Lin, M.-Y.; Khlystov, A.; Baldauf, R. W.; Isakov, V.; Faircloth, J.; Jackson, L. E. Field Investigation of Roadside Vegetative and Structural Barrier Impact on Near-Road Ultrafine Particle Concentrations under a Variety of Wind Conditions. *Sci. Total Environ.* **2012**, *419*, 7–15. https://doi.org/10.1016/j.scitotenv.2011.12.002.

(6) Drewnick, F.; Böttger, T.; von der Weiden-Reinmüller, S.-L.; Zorn, S. R.; Klimach, T.; Schneider, J.; Borrmann, S. Design of a Mobile Aerosol Research Laboratory and Data Processing Tools for Effective Stationary and Mobile Field Measurements. *Atmospheric Meas. Tech.* **2012**, *5* (6), 1443–1457. https://doi.org/10.5194/amt-5-1443-2012.

(7) Choi, W.; He, M.; Barbesant, V.; Kozawa, K. H.; Mara, S.; Winer, A. M.; Paulson, S. E. Prevalence of Wide Area Impacts Downwind of Freeways under Pre-Sunrise Stable Atmospheric Conditions. *Atmos. Environ.* **2012**, *62*, 318–327. https://doi.org/10.1016/j.atmosenv.2012.07.084.

(8) Ester, M.; Kriegel, H.-P.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. 6.

(9) Miller, D. J.; Actkinson, B.; Padilla, L.; Griffin, R. J.; Moore, K.; Lewis, P. G. T.; Gardner-Frolick, R.; Craft, E.; Portier, C. J.; Hamburg, S. P.; Alvarez, R. A. Characterizing Elevated Urban Air Pollutant Spatial Patterns with Mobile Monitoring in Houston, Texas. *Environ. Sci. Technol.* **2020**, *54* (4), 2133–2142. https://doi.org/10.1021/acs.est.9b05523.

(10) Actkinson, B.; Ensor, K.; Griffin, R. J. SIBaR: A New Method for Background Quantification and Removal from Mobile Air Pollution Measurements. *Atmospheric Meas. Tech.* **2021**, *14* (8), 5809–5821. https://doi.org/10.5194/amt-14-5809-2021.

(11) Tidyverse https://www.tidyverse.org/ (accessed 2022 -03 -31).

(12) Kassambara, A. *Ggpubr: “ggplot2” Based Publication Ready Plots*; 2020.

(13) Kuhn, M. *The Caret Package*.

(14) Hahsler, M.; Piekenbrock, M.; Doran, D. Dbscan: Fast Density-Based Clustering with R. *J. Stat. Softw.* **2019**, *91*, 1–30. https://doi.org/10.18637/jss.v091.i01.

(15) Leaflet for R - Introduction https://rstudio.github.io/leaflet/ (accessed 2022 -03 -31).

(16) Appelhans, T.; Reudenbach, C.; Russell, K.; Darley, J.; plugin), D. M. (Leaflet E.; Busetto, L.; Ranghetti, L.; McBain, M.; Gatscha, S.; plugin), B. H. (FlatGeobuf; Dufour (georaster-layer-for-leaflet), D.; Neuwirth, Y.; Cazelles, K. *Leafem: “leaflet” Extensions for “Mapview”;* 2021.

(17) Pebesma, E.; Bivand, R.; Racine, E.; Sumner, M.; Cook, I.; Keitt, T.; Lovelace, R.; Wickham, H.; Ooms, J.; Müller, K.; Pedersen, T. L.; Baston, D.; Dunnington, D. *Sf: Simple Features for R*; 2022.

(18) Interactive Viewing of Spatial Data in R https://r-spatial.github.io/mapview/ (accessed 2022 -03 -31).

(19) Dowle, M.; Srinivasan, A.; Gorecki, J.; Chirico, M.; Stetsenko, P.; Short, T.; Lianoglou, S.; Antonyan, E.; Bonsch, M.; Parsonage, H.; Ritchie, S.; Ren, K.; Tan, X.; Saporta, R.; Seiskari, O.; Dong, X.; Lang, M.; Iwasaki, W.; Wenchel, S.; Broman, K.; Schmidt, T.; Arenburg, D.; Smith, E.; Cocquemas, F.; Gomez, M.; Chataignon, P.; Blaser, N.; Selivanov, D.; Riabushenko, A.; Lee, C.; Groves, D.; Possenriede, D.; Parages, F.; Toth, D.; Yaramaz-David, M.; Perumal, A.; Sams, J.; Morgan, M.; Quinn, M.; @javrucebo; @marc-outins; Storey, R.; Saraswat, M.; Jacob, M.; Schubmehl, M.; Vaughan, D.; Hocking, T.; Silvestri, L.; Barrett, T.; Hester, J.; Damico, A.; Freundt, S.; Simons, D.; Andrade, E. S. de; Miller, C.; Meldgaard, J. P.; Tlapak, V.; Ushey, K.; Eddelbuettel, D.; Schwen, B. *Data.Table: Extension of “Data.Frame”;* 2021.

(20) R: The R Project for Statistical Computing https://www.r-project.org/ (accessed 2021 -06 -22).

(21) Roadway Inventory https://www.txdot.gov/inside-txdot/division/transportation-planning/roadway-inventory.html (accessed 2022 -03 -04).

(22) Simple Features for R https://r-spatial.github.io/sf/index.html (accessed 2022 -03 -11).

(23) Pachon, J. E.; Balachandran, S.; Hu, Y.; Mulholland, J. A.; Darrow, L. A.; Sarnat, J. A.; Tolbert, P. E.; Russell, A. G. Development of Outcome-Based, Multipollutant Mobile Source Indicators. *J. Air Waste Manag. Assoc.* **2012**, *62* (4), 431–442. https://doi.org/10.1080/10473289.2012.656218.

(24) Dallmann, T. R.; Kirchstetter, T. W.; DeMartini, S. J.; Harley, R. A. Quantifying On-Road Emissions from Gasoline-Powered Motor Vehicles: Accounting for the Presence of Medium- and Heavy-Duty Diesel Trucks. *Environ. Sci. Technol.* **2013**, *47* (23), 13873–13881. https://doi.org/10.1021/es402875u.

(25) Zhang, X.; Craft, E.; Zhang, K. Characterizing Spatial Variability of Air Pollution from Vehicle Traffic around the Houston Ship Channel Area. *Atmos. Environ.* **2017**, *161*, 167–175. https://doi.org/10.1016/j.atmosenv.2017.04.032.

(26) Lu, X.; Wang, J.; Yan, Y.; Zhou, L.; Ma, W. Estimating Hourly PM2.5 Concentrations Using Himawari-8 AOD and a DBSCAN-Modified Deep Learning Model over the YRDUA, China. *Atmospheric Pollut. Res.* **2021**, *12* (2), 183–192. https://doi.org/10.1016/j.apr.2020.10.020.

(27) Do, K.; Yu, H.; Velasquez, J.; Grell-Brisk, M.; Smith, H.; Ivey, C. E. A Data-Driven Approach for Characterizing Community Scale Air Pollution Exposure Disparities in Inland Southern California. *J. Aerosol Sci.* **2021**, *152*, 105704. https://doi.org/10.1016/j.jaerosci.2020.105704.

(28) Tan, P.-N.; Steinbach, M.; Karpatne, A.; Kumar, V. *Introduction to Data Mining*; Pearson, 2019.