Interim Report

Cluster Analysis of Metals and Organics in Dust Samples: Porter Ranch Area

Author: Bernard Beckerman, PhD, and Michael Jerrett, PhD

Date: Thursday May 12, 2016

This following description, analysis and reported results are limited to mapping and analysis of the analytic lab results obtained for metals and organics found in dust wipes.

Data Description and Processing

Geocoding residential addresses

We were provided with a list of 114 residential addresses and two school addresses that were geocoded using the Google Maps Geocoding Service API through QGIS software version 2.14.0. All of the geocodes were successfully located. Of the 114 addresses 104 were geocoded to a rooftop, eight of the geocodes were street number range interpolated, and two were assigned to the middle of the residential street route. Figure 1 illustrates approximate locations of the geocoded addresses.

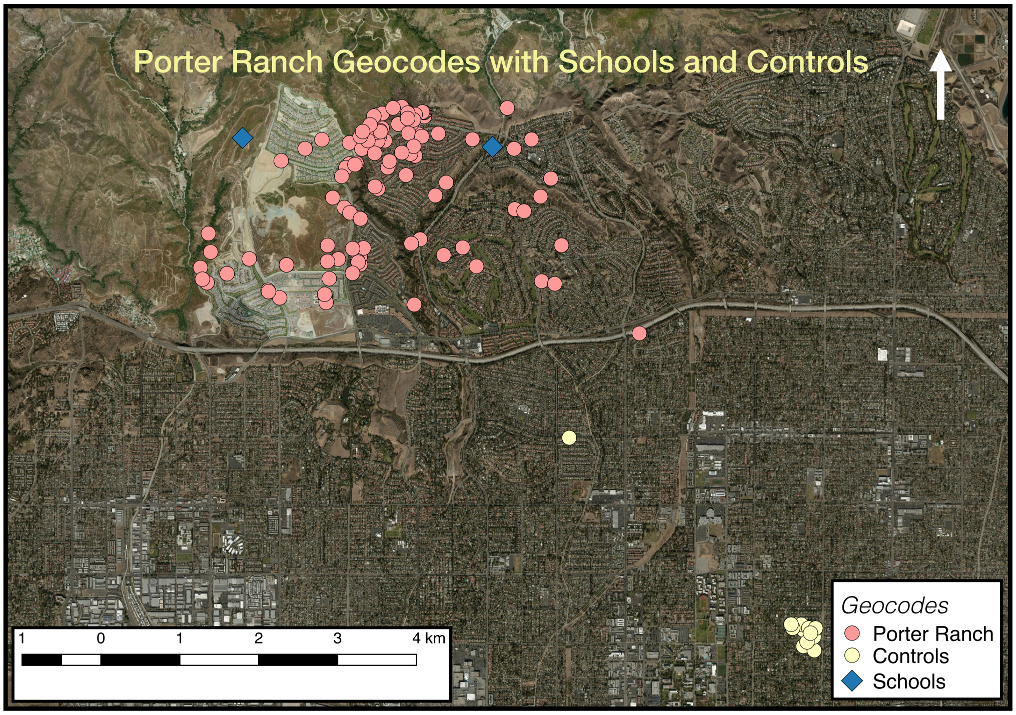

Figure 1: Map of Porter Ranch sampling locations geocoded by address, including two schools and the group of controls. Note: Coordinates elliptically shifted to protect privacy of the residents living in the homes.

Analytic datasets

The main dataset used for this analysis is a dataset reporting concentration of organics and metals in dust wipe samples taken at the sampling locations illustrated in Figure 1 in the Porter Ranch Area.

We used 11797 analyte samples in the following analyses; of those based on exclusion criteria, completeness, and numbers with estimated values for entries below detection limits, approximately 8568 analyte samples were used in the analysis. More specific descriptions of the exclusion criteria are given below. As part of the data processing, to ensure the data were balanced (i.e. data with no missing entries for a given analyte variable in all records), all data below the reporting limit was replaced with a value of half of the reporting limit (Burstyn & Teschke, 1999). Analyses, including mapping with analytic results, is restricted to only "Porter Ranch Area" homes; for this analysis, more distant control homes were excluded. This reduced the sample size for this analysis to 103 residential locations and 2 schools.

Cluster Analysis

The analyte data collected at the sample homes share a similar property of being high dimensional, i.e. having many variables, compared to most exposure-/impact assessments, which only attempt to assess a few compounds. This poses the challenge of having to potentially identify a complex mixture of compounds, rather than just a few analytes that are of interest. For this reason, two approaches were followed: the first was to identify houses that shared similar exposure profiles and post-hoc identify what made the groups different; and the second was to determine whether there were groups of underlying mixtures of components that were consistently represented in the sample.

The first of these approaches was conducted with k-means clustering (Hastie et al., 2001; Forgy, 1965). In using the k-means clustering method, the initial expectation was to be able to distinguish between high and low exposure profiles based on mixtures of the reported analytes. The second approach was to use principal components analysis (PCA) (Tabachnick and Fidell 2013; Hotelling, 1933) and factor analysis (Tabachnick and Fidell 2013; Catrell, 1952) to identity mixture trends in the datasets. These approaches would allow us to identify groups of variables that showed up together through a range of concentrations.

Initial screening on the variables was conducted to remove variables with no sample variability or detected quantities. After screening the variables in the dust samples, metals were represented by 13 elements, and organics by five compounds.

Dust wipe samples

The k-means cluster analyses of dust sample data were separately conducted on the subset of metals and organic compounds. We initially attempted to conduct the k-means cluster analysis on the continuous data of the laboratory results. This initial analysis was unable to find structural variability suggesting group membership configuration. In an attempt to identify structural mixtures in the data, a similar analysis was conducted on a categorical dataset that represented whether specific analytes were detected or not, i.e. detection = 1, non-detect=0. After these analyte detection mixture clusters were identified, they were then mapped to their geocoded locations. The goal of the post-hoc mapping of the mixture clusters was to determine whether certain analyte mixture profiles were more frequently represented in specific regions of the Porter Ranch area. Figures 2 and 3 illustrate the k-means clusters—based on the detection dataset—mapped at their sampling locations, for both organic compounds and metal, respectively. These mixture clusters were used later to visually evaluate whether patterns in analyte detection mixtures followed mixture structures identified with PCA and factor analysis..

We also conducted sensitivity analyses by removing iron (Fe) and aluminum (Al) from the analytic dataset and re-estimating all of the models to ascertain possible confounding by Fe and Al. As some of the samples were taken from window sills with Al casings which may have had steel fasteners we might expect to see levels of these metals elevated in a number of the samples due to dust settling on surfaces due to normal wear; if this was only present in certain residential developments due to age of the housing stock and materials used, it could have unduly biased the PCA and factor analysis.

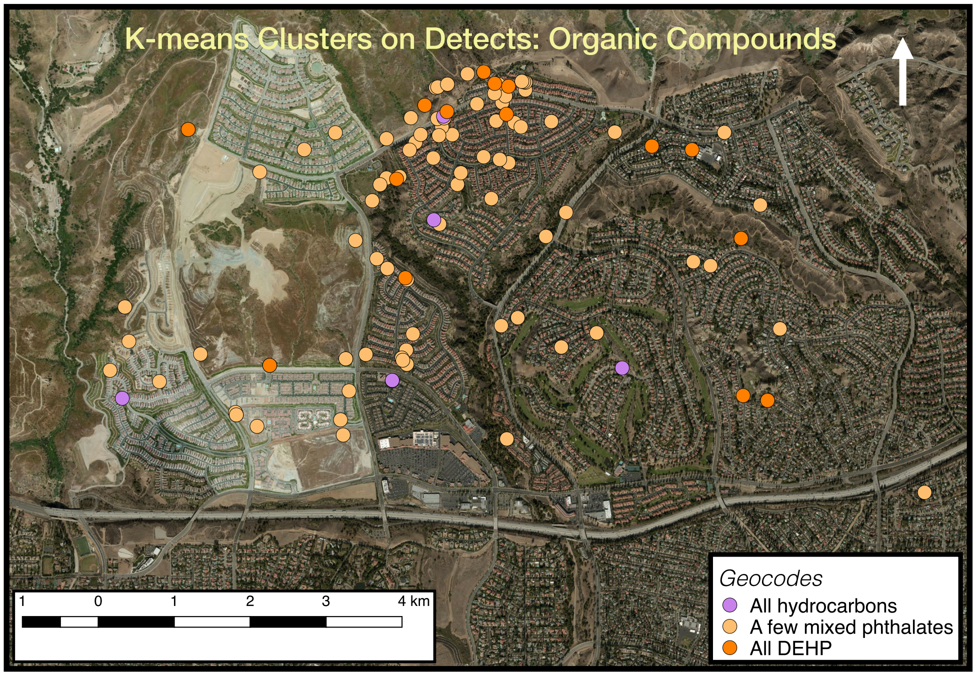

Figure 2: Map of results of k-means identified clusters for organic compounds

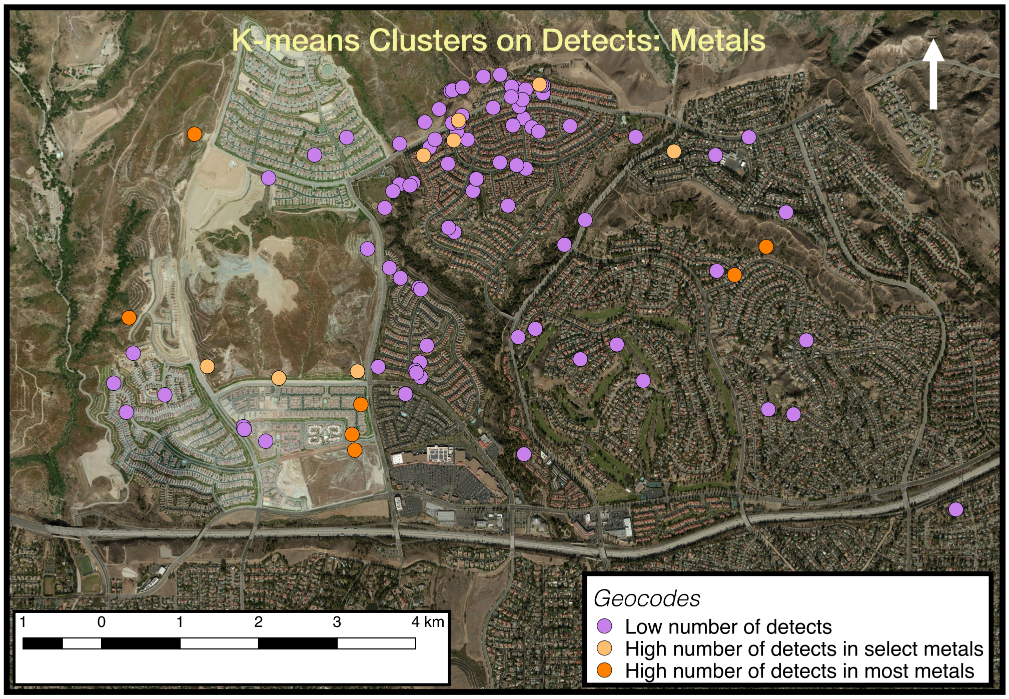
K-means Clusters on Detects: Metals
1
0
1
2
3
4 km
Geocodes
Low number of detects
High number of detects in select metals
High number of detects in most metals

Figure 3: Map of results of k-means identified clusters for metals

From the maps (Figures 2 and 3), there does not appear to be visual evidence that the detection-mixtures are spatially clustered beyond the clustering that already exists in the sampling locations . We also formally tested for spatial clustering in the mixture groups identified with the k-means clustering using the Ripley-K statistics for a marked non-homogenous Poisson point process (Cressie, 1993). The method would assess the level of spatial clustering inherent in the sampled data and determine if those homes identified as having larger numbers of detected compounds are more clustered than the already inherent clustering. The results of this analysis not provide evidence of clustering in the locations with higher levels of either metals or organics. It should be noted that the results of the cluster detection analyses suggest that even if there was clustering, the small available sample would have made it difficult to identify if clustering were present within a standard 95% confidence interval.

Under ideal conditions, to identify clustering or gradients in the observed levels associated with a distance decay process from a point source, each of the sampled houses would need uniform characteristics related to particle infiltration, particle settling, and interior environmental conditions like temperature, humidity and incident sun light on flat surfaces. Additionally, outdoor ambient conditions would need to be stable, especially with regard to wind speed and direction. Given the large number of uncharacterized variables related to the houses that may affect the rate of infiltration of outdoor air and dust into the homes, such as: air exchange rate;

insulation properties of the windows; seals on the window casing; seals around exterior doors, etc., it is not entirely unexpected that a clustered spatial pattern was not statistically observed. The meteorological conditions in the Porter Ranch area may have additionally contributed to our inability to detect clustering; the area is subject to high and variable wind speeds that could distribute contaminants at random, even if they were from a similar source.

Figure 4 is the screeplot of the PCA results for metals in dust. Results for the sensitivity analysis with Fe and Al removed are shown as an inset figure. In both cases, it illustrates that two components are able to explain a significant amount of the observed variability; and there is little difference between the two modeling specifications.

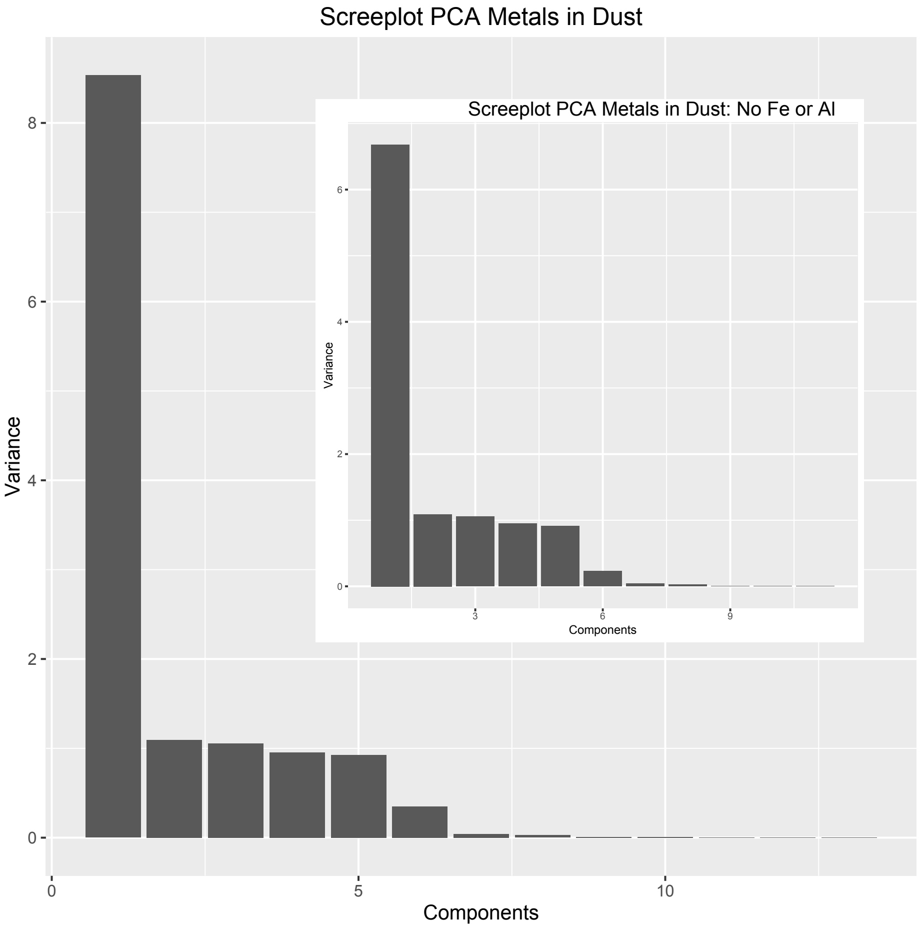

Figure 4: Screeplot of PCA analysis of metal in dust wipe samples with screeplot from sensitivity analysis (Fe and Al removed).

The first component was able to describe 66% of the variability and the second another 8%. Figure 5, the biplot of the PCA results for metals in dust, shows that there are a number of metals that are highly correlated (groups of arrows pointing in similar directions). Not shown is the related figure from the sensitivity analysis that removes Fe and Al. Both plots show nearly identical patterns in correlation structure and clustering of compounds. By following the arrows in Figure 5 to color coded points outside of the circle, we can identify that there are a few outlying points (green) that could possibly be classified by a smaller number of metals due to their leverage potential.

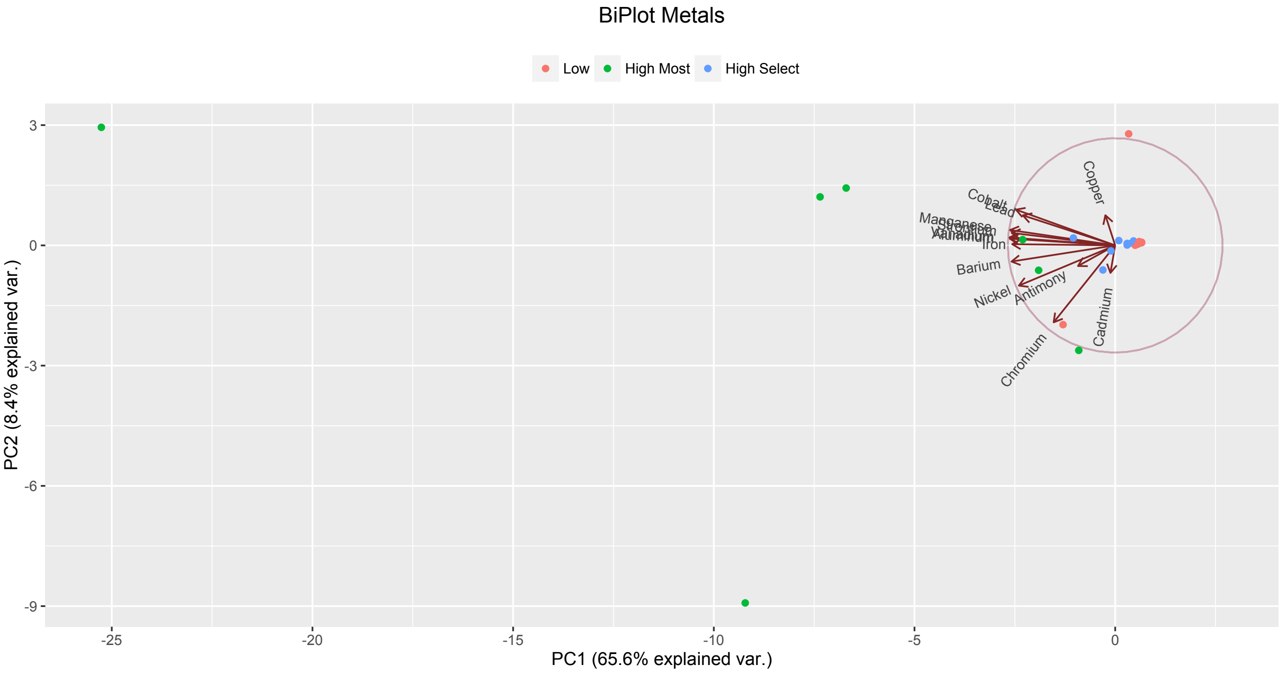

Figure 5: Biplot of PCA analysis of metal in dust wipe samples showing observations color coded by clusters from K-means analysis

Figure 6 illustrates the estimated analyte loading results of the factor analysis for metals in dust with an inset figure illustrating the sensitivity analysis without Fe or Al. It further illuminates the results from the PCA by showing that there is a mixture “fingerprint” of metals in dust. In both the main analysis and the sensitivity analysis, barium (Ba), manganese (Mn) and vanadium (V) are a stable cluster irrespective of the analysis. Striking differences are seen in the clustering memberships of lead (Pb) and chromium (Cr), both of which are markedly different between the main analysis and sensitivity analysis. While strontium (Sr), nickel (Ni) and cobalt (Co), have less stability in their clustering membership with Ba, Mn and V they generally remain in the same region suggesting that they might also be part of the metal dust characteristic “fingerprint”. It should be noted that both Fe and Al appear to be part of the Ba, Mn, and V cluster of metals.

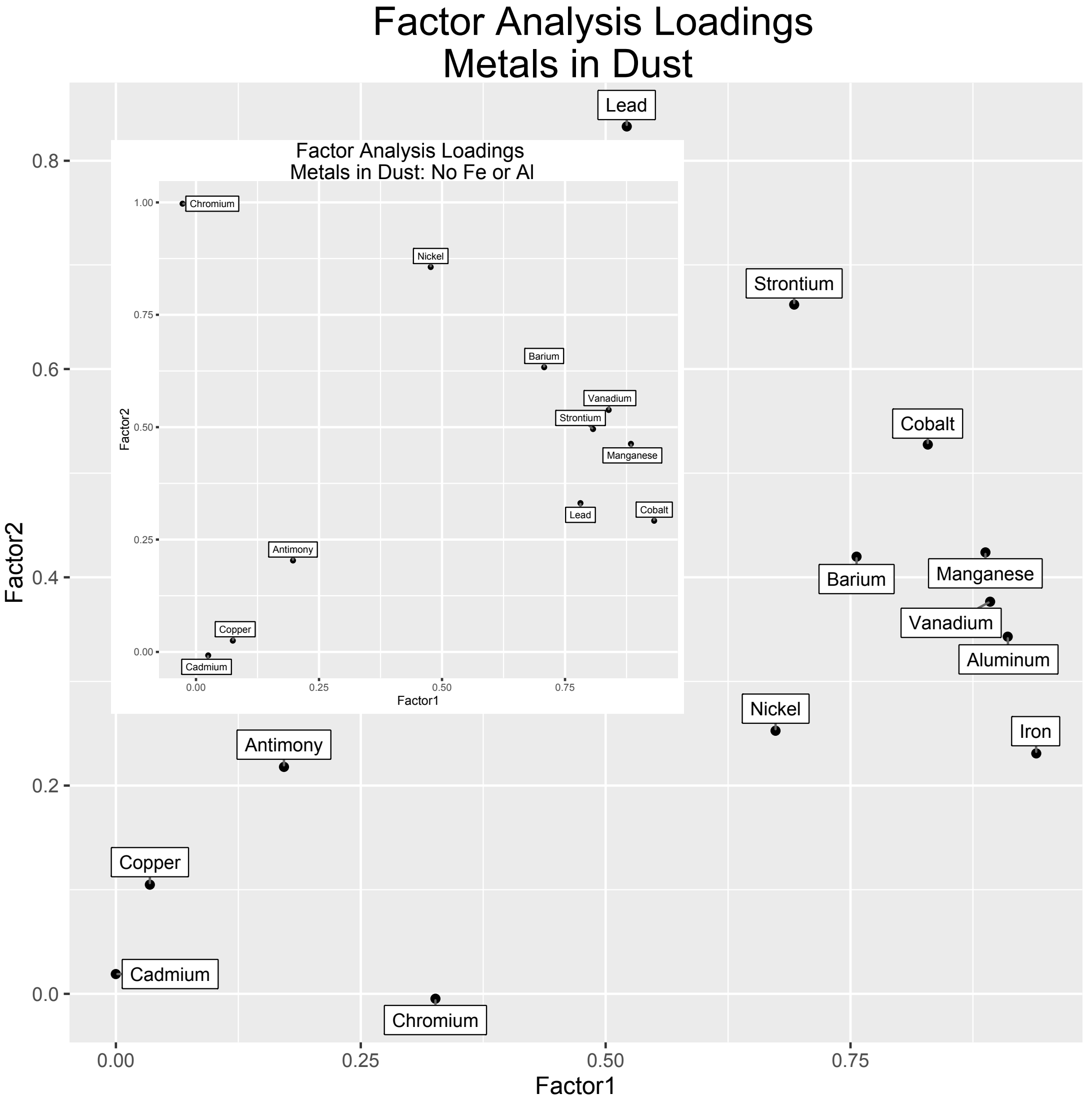

Figure 6: Plot of estimated loading factors for metals in dust

Results of the PCA for organic compounds did not show convincing evidence of a mixture fingerprint. The screeplot (Figure 7) did not show a large reduction in variance with the addition of more components. The biplot (Figure 8) shows that there is only one cluster of compounds; a cluster of two phthalates that only contains eight of the total detections of organics in dust. The only other unique signature of compounds in the dataset is a set of 15 observations, which have detections for bis(2-ethylhexyl) phthalate (DEHP), a common plasticizer in poly-vinyl chloride. These were previously identified in the k-means clustering show in Figure 2, labeled "All DEHP". Furthermore, the factor analysis of the organic compounds failed to find a combination of factors that could describe the co-variability in the analytes (p-value=0.617). In total, the results suggest that there is not a mixture fingerprint in the organic compounds found in the dust samples.

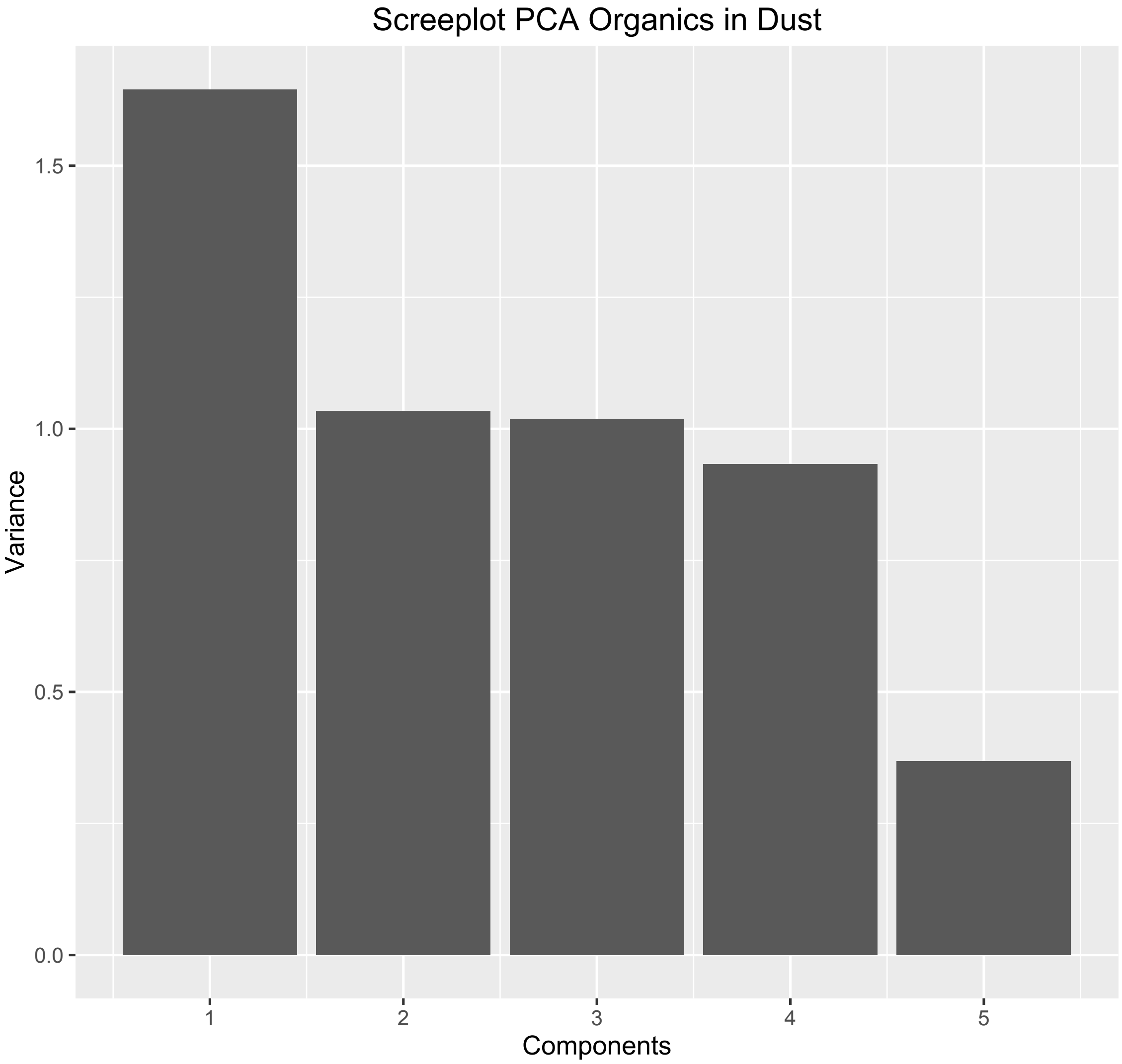

Figure 7: Screeplot of PCA analysis of organics in dust wipe samples

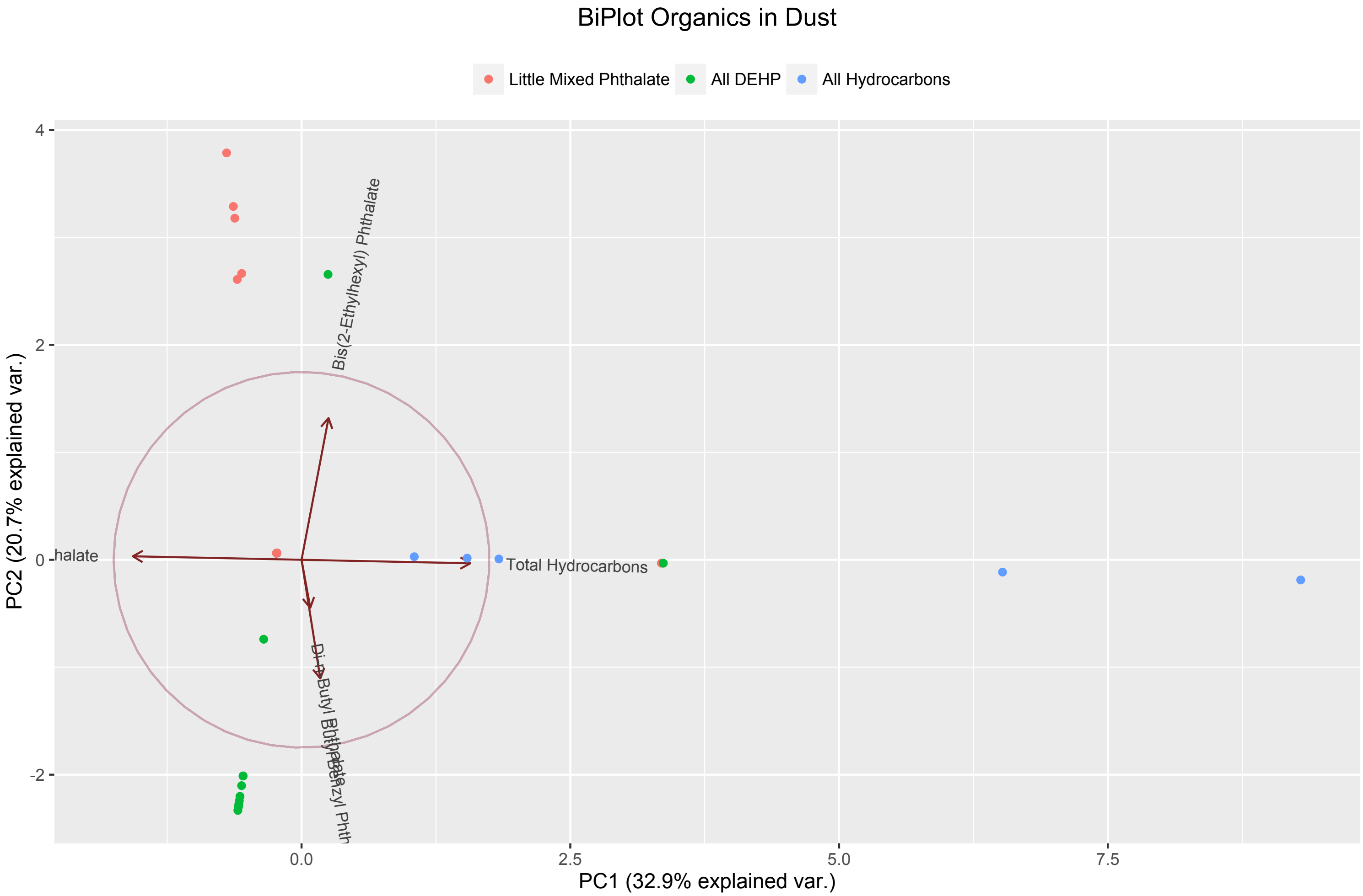

Figure 8: Biplot of PCA analysis of organics in dust wipe samples showing observations color coded by clusters from K-means analysis

Discussion

This statistical analysis cannot specifically identify a source of the mixture fingerprint. The results suggest that it is possible at some point in time prior to the sampling that a unique process liberated, and subsequently deposited in homes, a mixture of metal particulates that likely contained: Ba, Mn, V, and Al; to a lesser extent Fe, Co, and Ni; and perhaps Sr. The samples taken in the homes are cross-sectional, representing a static picture of the mixture of compounds in dust. Due to this, we cannot ascertain whether the metals were necessarily deposited at the same time. We can, however, conclude that there appears to be evidence that the relative amounts of certain metals in the samples are consistent through a range of concentrations. This consistency, previously referred to as the mixture fingerprint, is a logical prerequisite that should be met before being able to implicate a single source as being responsible for the contaminant impacts. The identification of the mixture fingerprint in this analysis, especially of metal not likely to be expected in household dust, meets this prerequisite and suggests that further investigation is necessary to determine the source of this contaminant impact and to whether there are other lasting measureable impacts that could be attributable to a single or multiple source event or activity.

References

Burstyn, I., & Teschke, K. (1999). Studying the determinants of exposure: a review of methods. American Industrial Hygiene Association journal, 60(1), 57-72.

Cattell, R. B. (1952). Factor analysis. New York: Harper.

Cressie, N. (1993). Statistics for spatial data: Wiley series in probability and statistics. Wiley-Interscience New York.

E.W. Forgy (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". Biometrics 21: 768–769. JSTOR 2528559.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning. 2001. NY Springer.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, 417–441, and 498–520.

Tabachnick, B. G., and Fidell, L. S. (2013). *Using Multivariate Statistics , 6th ed.* Boston : Pearson.

Disclaimer

The materials and opinions reported here are based on the following information and data:

Data List:

1. Database file of dust wipe samples: Wipe_Sample_Results_042816_v3.xlsx
2. Address list of residential locations: Household ID List.xlsx

We reserve the right to modify our results and opinions; or add additional opinions based on new materials reviewed.