## Title
Associations between socio-demographic characteristics and chemical concentrations contributing to cumulative exposures in the United States



**Journal**
Journal of Exposure Science and Environmental Epidemiology, 27(6)

**ISSN**
1559-0631 1559-064X

**Authors**
Huang, Hongtai
Tornero-Velez, Rogelio
Barzyk, Timothy M

**Publication Date**
2017-09-13

**DOI**
10.1038/jes.2017.15

# Associations between Socio-Demographic Characteristics and Chemical Concentrations Contributing to Cumulative Exposures in the United States

**Hongtai Huang[1,2], Ph.D., Rogelio Tornero-Velez[2], Ph.D., Timothy M. Barzyk[2], Ph.D.**

[1]Oak Ridge Institute for Science and Education (ORISE) and [2]National Exposure Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, NC 27709

Address correspondence to H. Huang, U.S. Environmental Protection Agency, National Exposure Research Laboratory, 109 T.W. Alexander Drive, Mail Code E205-2, Room D-482, Research Triangle Park, NC 27711 USA. Telephone: (919) 541- 5407. Fax: 919-541-9444. Email: Huang.Hongtai@epa.gov.

**Running Title**: Quantifying Combined Effects of Multiple Stressors

This research was supported in part by an appointment to the Post-doctoral Research Program at the U.S. Environmental Protection Agency's National Exposure Research Laboratory (Research Triangle Park, NC) administered by the Oak Ridge Institute for Science and Education through an Interagency Agreement between the U.S. Department of Energy and the U.S. Environmental Protection Agency. The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

All authors declare no actual or potential competing financial interests.

## Abstract

Background: Association rule mining (ARM) has been widely used to identify associations between various entities in many fields. Although some studies have utilized it to analyze the relationship between chemicals and human health effects, fewer have used this technique to identify and quantify associations between environmental and social stressors.

Methods: Socio-demographic variables were generated based on U. S. Census tract-level income, race/ethnicity population percentage, education level, and age information from 2010-2014, 5-year summary files in the American Community Survey (ACS) database, and chemical variables were generated by utilizing the 2011 National-Scale Air Toxics Assessment (NATA) census tract-level air pollutant exposure concentration data. ARM was then applied to quantify and visualize the associations between the chemical and socio-demographic variables.

Results: Census tracts with a high percentage of racial/ethnic minorities, and populations with low income, tended to have higher estimated chemical exposure concentrations (4$^{th}$ quartile), especially for diesel PM, 1, 3-butadiene, and toluene. In contrast, census tracts with an average population age of 40 to 50 years old, a low percentage of racial/ethnic minorities, and moderate-income levels, were more likely to have lower estimated chemical exposure concentrations (1$^{st}$ quartile).

Conclusion: Unsupervised data mining methods can be used to evaluate potential associations between environmental inequalities and social disparities, while providing support in public health decision-making contexts.

**Key words**: Multiple Stressors, Rule Mining, Cumulative Risks, Combined Effects, Environmental Justice

## INTRODUCTION

Quantitatively evaluating the combined effects of multiple chemical/non-chemical stressors has been simultaneously a crucial focus of and a challenge for cumulative risk assessment (CRA)[1]. CRA defines cumulative risk as 'the combined risks from aggregate exposures to multiple agents or stressors' [2]. Environmental Justice (EJ) communities are often host to multiple chemical and non-chemical stressors, such as poverty or pre-existing health conditions, which could decrease individual or population resilience, and increase the potential impacts from chemical exposures[3]. The role of CRA in public health decision making related to EJ is vital[4], and there have been a significant number of methodological approaches developed which intend to capture the combined effects of multiple stressors in addressing EJ issues[5].

In general, most of the approaches used in CRA chemical/non-chemical studies can be divided into three categories: effect-based (top-down), stressor-based (bottom-up) and the hybrid of these two, vulnerability-based[5, 6], which considers impacts from a number of chemical and non-chemical stressors. In practice, vulnerability-based studies utilize existing data and information, and can also effectively address the prioritized stressors without exhaustively considering all the non-chemical or chemical variables. Several quantitative CRA studies belong to this category[7-17]. Specifically, chemical or socio-demographic stressors of interest were quantified and used as the basis to either compare exposure levels or health effects among different groups in the population[8-16], or serve as a screening tool to address cumulative impacts in areas featured by social disadvantage[7, 17]. Other quantitative measures or indices such as Margin of Exposure (MOE), no observed adverse effect level (NOAEL), benchmark Dose (BMD) and reference dose (RfD) were also used to assess the combined health risk of chemical mixtures for regulatory purposes[18]. Regression models have proved useful in characterizing associations between exposure or health effects and different stressors[19-21], but this technique does require pre-defining the response variable and explanatory variables. Interpretation of the interaction term in the model can also be challenging, especially when there are a large number of variables involved[22].

Very few CRA studies adopt alternative data mining methods, such as unsupervised association rule mining techniques, to quantify associations between chemical/non-chemical stressors and health effects, especially those related to exposure and dose-response assessments.

Association rule mining (ARM)[23, 24] has been widely applied in many different scientific areas[25-29]. Recently, researchers used ARM to analyze the relationship between environmental stressors and adverse human health impacts[30, 31]. There are three main advantages of using ARM. First, it can provide better characterization of the interactions between multiple stressors without having to pre-define them as response or explanatory variables.

Second, outputs from this method are in general easily interpretable by those without an advanced mathematical background [31]. Finally, as a non-parametric method, ARM makes no assumptions about the probability distributions of the variables being assessed.

In this study, ARM was applied to analyze the inter-relationships between different chemical/non-chemical stressors, in order to demonstrate the use of advanced data mining techniques to understand social disparities and disproportionate environmental burdens. The null hypothesis is that increased chemical exposures are not associated with combinations of EJ-related variables.

## DATA AND METHODS

### Data

Socio-demographic data and chemical exposure estimates were collected for each census tract across the United States. In total, more than 73 000 census tracts were evaluated, representing more than 317 million people living in the U.S.

Socio-demographic variables were selected based on their relevance to EJ communities. These variables are individual income, race/ethnicity population percentage, educational attainment, and age by sex information at the census tract level from the 2010-2014, 5-Year Summary file in the American Community Survey (ACS) database. Note that the Summary file is not an average of the 5-year period but aggregated data collected continuously on a daily basis for 5 years[32].

Chemical variables were generated by utilizing the Environmental Protection Agency (EPA) 2011 National-Scale Air Toxics Assessment (NATA), census tract-level, modeled pollutant exposure estimates (http://www.epa.gov/national-air-toxics-assessment/2011-nata-assessment-results). Six pollutants were chosen for analysis, including acetaldehyde, benzene, cyanide, particulate matter components of diesel engine emissions (namely diesel PM), toluene, and 1,3-butadiene. These chemicals were selected based on their potential for health impacts as well as their relevance to mobile source (i.e., vehicular traffic) and industrial emissions, both of which are highly concentrated in EJ areas[33, 34].

Socio-demographic variables were binned such that every census tract had a score for each variable, and chemical exposure estimates were divided into quartiles for each census tract. Although variables were selected based on their relevance to EJ communities, given the national scale and lack of pre-defined associations, there was no assumption that EJ relationships would necessarily manifest themselves in the results.

## Method

Data analysis was performed using statistical software, R (version 3.2.1; R Core Team, Vienna, Austria). Execution of ARM and visualization of the resultant association rules were based on the R packages ‘arules’[35] and ‘arulesViz’[36] respectively.

### *Association Rule mining*

ARM, a form of frequent item set mining[37], is a tool used to search for associations between different variables within a database without explicitly specifying the cause (the left-hand-side, LHS) or corresponding effect (the right-hand-side, RHS). As is the case for many situations, if the values of all variables of concern are binary, i.e., either 0 or 1, the association rule is categorically referred to as market basket analysis[23]. Therefore, each observation or record constitutes a ‘transaction’ which, in our case, refers to a census tract. Each element within a record is an ‘item’ that corresponds to a stressor in this study. Essentially, ARM is mining co-occurrence relationships between two separate sets of items.

The proportion of transactions that contain the item set is defined as the *support* (i.e., the proportion of tracts that contain the stressor) and *confidence* is the estimated conditional probability of the co-occurrence of both LHS and RHS, or support of the rule given the support of the LHS[35]. *Lift* is defined as the confidence normalized by the support of the RHS, meaning the conditional probability of rule support given supports of the LHS and RHS[23]. High values of support, confidence, and lift are indicative of a strong association rule, in that it involves a large number of observations (i.e., tracts with those characteristics) and therefore can be generalized to a wider scope. When the rule size is only 2, which means that only one item showed up in both the LHS and RHS (such as an income score mapped to a chemical exposure score), the rule can be interpreted in the context of an odds ratio[38] and relative risks[39]. Mathematical relations/derivation between these measures can be found in Supplementary Material, Equations (1)-(9).

*Stressors*

Census tract-level individual income, race/ethnicity population percentage, and personal education attainment levels were obtained from the ACS 2010-2014, 5-Year Summary file to define, quantify, and assign scores for the demographic variables poverty, race, and education. Variable ‘poverty’ was defined as the percentage of people in each census tract whose ratio of income to the poverty level (over the past 12 months)[40] is below 1.5. Variable ‘race’ represents the non-white population percentage at each census tract. The definition of variable ‘education’ is the percentage of population who received a degree (Associate degree and above) at each census tract. Note that variables were initially calculated as a percentage value for each census tract. A score was then assigned to each census tract given the percentages ranging from score 1 (lowest percentage range – [0,10%]) to 10 (highest percentage range – [90%, 100%)). Note that the percentages are evenly divided into ten sub-ranges and therefore, 10 score categories. The education score 8-10 was merged into one score category, and poverty score 7-10 into another, due to the small sample size of these score categories. The number of census tracts associated with each score can be found in Supplementary Material, Table S-1.

The tract-level ‘age by sex’ variable in the ACS database was used, and the average weighted age calculated for each census tract by summing the products of the percentage of each age group and the median (or predefined value if there was no upper bound of the interval) of the corresponding age interval. This variable was then sub-divided into 7 variables, namely ‘0-20 years, ‘20-30 years, ‘30-35 years, ’35-38 years, ‘38-40 years, ‘40-50 years and ‘50-100 years. These age intervals were chosen based on biological stages and sample size (see Supplementary Material, Table S-1). We calculated the average of weighted age by sex assuming that the ratio of male to female was 1:1.

Each of the six chemical variables was converted into four quartile variables based on the chemical concentrations for each tract. Taking benzene as an example, the original benzene exposure concentration value for each census tract was converted into a label depending on which quartile that particular concentration value resides. For instance, if the value was within the first quartile of benzene exposure concentrations across all census tracts, the numeric value was converted to a category label ‘Q1’. As six chemical variables were considered, these became 24 distinct quartile variables.

In total, there were 56 variables: 10 race/ethnicity groups, 8 education groups, 7 poverty groups, 7 age groups, and 24 chemical quartile groups.

*Data Analysis*

Two separate experiments were conducted by applying the ARM method with different minimum support thresholds. In the first experiment, the LHS of the association rule was set to be only non-chemical stressors and the RHS to be only chemical variables for interpretation purposes. In order to understand the internal connections among non-chemical stressors, the second experiment was performed requiring both the LHS and RHS to be socio-demographic variables. The rules were only analyzed when the lift was greater than 1. In addition, the focus was on those rules with size equal to 2 (a 1-to-1 map of LHS and RHS) in order to better utilize the statistical measures Odds Ratio (OR) and Relative Risk (RR).

The 95% confidence intervals (CI) were estimated for OR using bootstrapping[41] random sampling for 10 000 times, for particular rules of interest. Specifically, a new data set was created each time using random sample records with replacement, and ARM was applied on these newly created data. The rule of interest was then obtained and the corresponding OR calculated. For 10 000 bootstrapping runs, we eventually had 10 000 new data sets and corresponding OR values. The 2.5 and 97.5 percentiles were identified among these 10 000 OR values, which was the estimated 95% CI.

The chemical exposure was also compared to the concentration levels associated with each of the three demographic variables (poverty, race/ethnicity & education attainment) using Student’s t tests, in order to examine the statistical significance of the differences between score categories of these variables.

## RESULTS

### *Association Rules*

Because there were 56 total variables, the possible number of item set combinations was $2^{56}-1$ ($\approx 7.2 \times 10^{16}$, or 72 quadrillion) as the basis for generating association rules. With confidence set to be 0.1 and support 0.1, 212 rules were obtained. Without setting a lower bound on the confidence value, there were 30 932 rules given a minimum support threshold of 0.1 (details in Supplementary Material, Table S-2). Imposed criteria regarding the content of the LHS or RHS further restricted the number of rules.

*-Rules with Larger Minimum Support Values*

Table 1 lists the rules for support >0.1 and lift >1.0 and shows that only two demographic variables, “Race Minority Score 1” (0-10% non-white) and “Age= 40-50” resulted as the LHS of these rules while most of the chemical variables represented first or second quartile concentrations, except cyanide. Odds ratios for these rules ranged from 1.433 to 2.947.

The graph-based visualization of all the association rules with support >0.1 and lift >1 is shown in Figure 1. All associations are connected through blank circles. The size of a circle represents the co-occurrence support value, and color indicates the lift value of the rule. Larger circles mean higher support values, while deeper colors suggest greater lift. It can be observed that both variables ‘Age = 40–50’ (average population age of 40 to 50 years old) and Race score 1 (low non-white percentage) were associated with 1st quartile chemicals.

Table 2 shows all the association rules with criteria that both the LHS and RHS were socio-demographic variables, and with minimum support value greater than 0.1 and lift greater than 1. Only three variables appeared in these 6 rules, including “Race Minority Score 1”, “Age=40-50” and “Poverty Score 2”. Interestingly, all three of these variables were interacting with each other, forming three loops.

*-Rules with Smaller Minimum Support Values*

If a similar criterion was applied, but with the minimum support value set to 0.01, more rules were found with size greater than 2 (see Supplementary Material, Table S-3). Not only did 1st and 2nd quartiles chemical variables show up in the RHS, but also those in the fourth quartiles. Corresponding LHS of the fourth quantile rules were high race minority scores (high non-white percentage), high poverty scores (high low-income percentage), and low education scores (low percentage of degree attainment).

Table 3 summarizes the total number of rules with particular LHS and RHS given a minimum support value of 0.01 and lift greater than 1. For the LHS, the focused was on low and high demographic scores. All the rules with race minority score 1 and race minority score 2 on the LHS were pooled together, since they both represent low percentages of non-white population, and so were race minority scores 7, 8, 9 and 10. Similarly, all the rules with poverty score 1, 2, and 3 were evaluated at the same time, and those with education score 1, 2, and 3 examined together. For the RHS, the total number of rules was counted that contained particular quartiles of chemical exposure concentrations given the specific LHS.

In general, rules containing low race score (low non-white percentage), low poverty score (less poor census tract), and average population age of 38 to 50 years old were more likely to contain the first quartile (i.e., Q1 or lower values) of chemical exposure concentrations, while rules encompassing high race score (high non-white percentage), high poverty score (poorer tracts), and high education score (high percentage of residents with education) tended to include the fourth quartile of chemical exposure concentration (or Q4, indicating high chemical exposure concentration). Specifically, 20 out of 29 rules (69%) that contained race score 7, 8, 9 or 10 had Q4 as their RHS, while only 16 out of 342 rules (5%) that contained race score 1 or 2 included Q4. The number of rules with high race score increased monotonically, as the chemical exposure concentration increased in the RHS (from 0 for Q1 to 20 for Q4). In contrast, the number of rules with low race scores gradually decreased as the chemical concentration became higher (from 144 for Q1 to 22 for Q4).

There were 9 out of 14 rules (64%) with poverty score 7-10 containing Q4, but there were only 27 out of 354 rules (8%) with poverty score 1, 2 or 3 containing Q4. A high poverty score was positively associated with chemical exposure concentrations in terms of rule number (from 1 rule for Q1, to 9 for Q4), while low poverty score had a negative association with chemical exposure concentration (144 for Q1, and only 28 for Q4).

Rules with average population age of 38-40 and 40-50 years old tended to have Q1 as their RHS (50% and 37% respectively). As the RHS of these rules changed from Q1 to Q4, the rule numbers decreased consistently (from 31 to 8, and 106 to 4 respectively).

Interestingly, rules with high education score (8-10) were associated with Q4 (46%), but those with low education score (1, 2, or 3) were more inclined to contain either Q1 (49%) or Q4 (22%). The number of rules with high education score increased gradually when RHS changed from Q1 to Q4. For rules with low education score, there was no monotonic change in rule numbers when RHS shifted from Q1 to Q4.

Supplementary Material, Table S-4 includes the top 100 rules with both LHS and RHS being demographic variables, minimum support value 0.01, and lift greater than 1. Highest poverty score was associated with average population age of 20-30 years old and the lowest education score.

On the other hand, lowest poverty score was related to high education scores and low race minority scores.

To explore further the one-to-one relationship between the LHS and RHS, the rule size was set to be 2 on top of other predefined criteria such as LHS being socio-demographic variables, RHS chemical variables, minimum support value 0.01 and lift greater than 1 (see sample rules in Supplementary Material, Table S-5). Table 4 lists complementary pairs of rules with high and low race scores for given high/low chemical quartiles. The rule with highest odds ratio (5.534, estimated 95% CI 5.102-6.008) had an LHS race score of 10 and RHS fourth quartile diesel. The rule with the same LHS and RHS but low race and exposure values was 'Race Minority Score = 1→ Diesel = Q1' for which the odds ratio was 2.893 (estimated 95% CI 2.818-2.969). The general form of these rules is that 'Race Minority Score = 10 → Chemical = Q4' and 'Race Minority Score = 1 → Chemical = Q1'. In addition, average population age of 20-30 and 30-35 years old were associated with 'Diesel = Q4' but average population age of 40-50 and 50-100 with Q1 chemical concentrations. All estimated 95% CI for the OR of all rules in Table 4 were well above 1 suggesting positive associations.

*Student's t-tests*

Regarding educational attainment, in general, chemical exposure concentration levels for different education scores were statistically different (Bonferroni's corrected ɑ level = $1.79\times10^{-3}$) except for cyanide compounds (see Supplementary Material, Table S-6). Also, differences between chemical concentration levels for each poverty score were statistically significant for all chemicals (details in Supplementary Material, Table S-7). Except for several pairs of race score categories associated with cyanide and acetaldehyde concentrations, statistically significant differences between different race scores in terms of chemical exposure concentration levels were observed (Supplementary Material, Table S-8).

## DISCUSSION

### Overview

*Major Association Rules*

Among the 212 rules with minimum support value greater than 0.1, 13 major rules were found with the strength measure ‘lift’ greater than 1 that contained socio-demographic variables as their LHS and chemical variables as their RHS. Results presented in Table 1 convey the main message that census tracts with low non-white population percentages (0-10%) or average population age of 40 and 50 years old (which happens to be associated with low poverty and low non-white populations, details in Table 2) are associated with low chemical exposure concentrations (mostly at the first quartiles).

Six major rules were also found when setting both the RHS and LHS to be socio-demographic variables with similar criteria (in Table 2). As with the results in Table 1, in addition to low percentage of non-white population and average population age of 40-50, poverty score 2 (or, 10% - 20% of the residents within a census tract having income below one-and-a-half times the poverty level) appeared and demonstrated key interactions with the other two socio-demographic variables. This suggests that income level is probably associated with chemical exposure concentration level. Another perspective is that predominantly white census tracts of middle aged people are directly related to lower exposure levels, and they happen to have low poverty levels, which are thus indirectly related to exposures.

*Association Rules and EJ Interpretation*

When the minimum support value was lowered to 0.01 and held other criteria the same, several interesting trends were found regarding the association between demographic variables and exposure concentration levels. Greater proportions of non-white populations and poorer census tracts tended to be exposed to higher chemical concentrations, while tracts with low non-white percentages, wealthy tracts, and those with average population age of 38 to 50 were more likely to have low chemical exposure concentrations (Table 3). Particularly, the number of stronger (lift > 1) and applicable (support > 0.01) association rules with high race score, high poverty score, and higher education scores (contrary to expectations) increased as the chemical exposure concentrations increased from the first to the fourth quartiles; while rules with low race score, low poverty score, and average population age of 38 to 50 decreased as chemical concentrations became higher.

Educational attainment did not show a clear inverse relationship with chemical concentrations when considered by itself on the LHS (Table 3). These may represent a limited sample of highly educated census tracts that were exposed to increased concentrations. However, in general, according to results when comparing socio-demographic variables as both LHS and RHS, (Table S-4), high education was associated with low poverty and low non-white population percentages, which experienced lower concentration levels and appeared to be more influential to exposures. Also, when considering multiple socio-demographic variables on the LHS and chemical concentrations on the RHS, educational scores were no greater than 4, suggesting that the majority of tracts that were associated with chemical concentrations (high or low) had populations where less than 40% of the residents have an associate’s degree, and were likely driven by the other EJ factors, especially race, income, and age. Wealthier, middle aged, white population experienced lower exposures, and low-income, younger, minority population experienced higher exposures. Education may not be as influential, as long as race and poverty had low scores (i.e., more non-white with higher incomes). Education could vary and still represent lower exposures but itself cannot sufficiently address environmental disparities.

*Graph-based Visualization*

Graph-based visualization of the identified association rules offers better illustrations of the combined effects of multiple chemical and socio-demographic variables. It can be rather useful in displaying associations between variables, especially when the number of involved variables increased and the size of a rule was more than 2 (see Supplementary Material, Figures S-1 & S-2). In conjunction with using other statistical methods such as regression analysis, the combined effects of multiple stressors upon one response variable can be identified and quantified, provided that the number of explanatory variables was small (<4) and the association of interest was statistically significant.

The graph-based visualization of the association rules can also serve as the basis for developing more complex mathematical models for environmental studies such as a system dynamic model[42, 43] or multi-objective model[44, 45], and provide hints for better ways of clustering and classifications (Supplementary Material, Figures S-1 & S-2). It may also shed lights on potential contributors to disproportionate environmental burdens for certain vulnerable populations such as pregnant women or children who suffer from obesity[46].

Along with the method developed to explore and identify a group of important variables[47], this approach can be applied to evaluate the internal relationships among a large number of multiple stressors, and potentially provides a systemic perspective into the environmental issues at hand.

*Limitations*

There are three limitations of this study. First, NATA exposure concentration are simulated data rather than actual observations. The results presented here may not perfectly reflect the actual chemical exposure levels. Second, ARM cannot provide exact quantitative relationships between variables. Therefore, the results cannot be directly compared with those from other studies. Third, interpretation of other measures such as OR and RR can be an issue when the rule size is greater than 2.

## Conclusion

Unsupervised data mining methods such as ARM can be applied to EJ-related evaluations of the combined effects of multiple stressors. It highlights some of the main variables associated with chemical exposures, in this case race, income, and population age, and suggests that other variables, such as education, may be less associated with exposures and more a secondary component of the other socio-demographic variables.

Other variables that could be included in future studies include pre-existing health conditions, access to health care, epigenetic predisposition, chemical mixtures, and chemical/non-chemical synergistic interactions (e.g., radon and smoking, or toluene and noise). ARM has proven to be an effective methodology for finding associations between specific categories/values (i.e., binned ranges) of EJ variables, which provides more insight into the specifically affected populations. In general, middle aged, white, non-poor tracts were associated with lower exposures, and younger, higher poverty, non-white tracts with higher exposures. ARM allows us to investigate each of these variables with respect to their associations to not only chemical exposures but to each other as well. This method could thus be used to target solutions to the most applicable variables.

Supplementary information is available at Journal of Exposure Science and Environmental Epidemiology’s website.

## Disclaimer

This article has been subject to review by the EPA and approved for publication. Although this work was performed as research for the U.S. Environmental Protection Agency, it does not necessarily represent endorsement of official Agency policies.

All authors declare no actual or potential competing financial interests.

## References

1. Callahan MA, Sexton K. If cumulative risk assessment is the answer, what is the question? Environmental Health Perspectives. 2007;115(5):799-806.

2. U.S. EPA (Environmental Protection Agency). Concepts, methods and data sources for cumulative health risk assessment of multiple chemicals, exposures and effects: A resource document. U.S. EPA, National Center for Environmental Assessment, Cincinnati, OH. EPA/600/R-06/013F2007.

3. Taylor WC, Poston WSC, Jones L, Kraft MK. Environmental justice: obesity, physical activity, and healthy eating. Journal of Physical Activity & Health. 2006;3:30-54.

4. Sexton K, Linder SH. The role of cumulative risk assessment in decisions about environmental justice. International Journal of Environmental Research and Public Health. 2010;7(11):4037-49.

5. Sexton K. Cumulative risk assessment: an overview of methodological approaches for evaluating combined health effects from exposure to multiple environmental stressors. International Journal of Environmental Research and Public Health. 2012;9(2):370-90.

6. Sexton K. Cumulative health risk assessment: finding new ideas and escaping from the old ones. Human and Ecological Risk Assessment: An International Journal. 2014;21(4):934-51.

7. Alexeeff GV, Faust JB, August LM, Milanes C, Randles K, Zeise L, et al. A screening method for assessing cumulative impacts. International Journal of Environmental Research and Public Health. 2012;9(2):648-59.

8. Apelberg BJ, Buckley TJ, White RH. Socioeconomic and racial disparities in cancer risk from air toxics in Maryland. Environmental Health Perspectives. 2005;113(6):693-9.

9. Barzyk TM, White BM, Millard M, Martin M, Perlmutt LD, Harris F, et al. Linking socio-economic status, adverse health outcome, and environmental pollution information to develop a set of environmental justice indicators with three case study applications. Environmental Justice. 2011;4(3):171-7.

10. Bell ML, Ebisu K. Environmental inequality in exposures to airborne particulate matter components in the United States. Environmental Health Perspectives. 2012;120(12):1699-704.

11. Clougherty JE, Levy JI, Kubzansky LD, Ryan PB, Suglia SF, Canner MJ, et al. Synergistic effects of traffic-related air pollution and exposure to violence on urban asthma etiology. Environmental Health Perspectives. 2007;115(8):1140-6.

12. Cutter SL, Boruff BJ, Shirley WL. Social vulnerability to environmental hazards. Social Science Quarterly. 2003;84(2):242-61.

13. Harner J, Warner K, Pierce J, Huber T. Urban environmental justice indices. The Professional Geographer. 2002;54(3):318-31.

14. Linder SH, Marko D, Sexton K. Cumulative cancer risk from air pollution in Houston: Disparities in risk burden and social disadvantage. Environmental Science & Technology. 2008;42(12):4312-22.

15. Morello-Frosch R, Pastor Jr M, Porras C, Sadd J. Environmental justice and regional inequality in southern California: implications for future research. Environmental Health Perspectives. 2002;110:149-54.

16. Perlin SA, Sexton K, Wong DW. An examination of race and poverty for populations living near industrial sources of air pollution. Journal of Exposure Analysis and Environmental Epidemiology. 1998;9(1):29-48.

17. Sadd JL, Pastor M, Morello-Frosch R, Scoggins J, Jesdale B. Playing it safe: assessing cumulative impact and social vulnerability through an environmental justice screening method in the South Coast air basin, California. International Journal of Environmental Research and Public Health. 2011;8(5):1441-59.

18. Sexton K, Linder SH. Cumulative risk assessment for combined health effects from chemical and nonchemical stressors. American Journal of Public Health. 2011;101(S1):81-8.

19. Chahine T, Schultz BD, Zartarian VG, Xue J, Subramanian SV, Levy JI. Modeling joint exposures and health outcomes for cumulative risk assessment: the case of radon and smoking. International Journal of Environmental Research and Public Health. 2011;8(9):3688-711.

20. Fox MA, J.D. G, Burke TA. Evaluating cumulative risk assessment for environmental justice: a community case study. Environmental Health Perspectives. 2002;110:203-9.

21. Morello-Frosch R, Pastor M, Sadd J. Environmental justice and southern california's "riskscape": the distribution of air toxics exposures and health risks among diverse communities. Urban Affairs Review. 2001;36(4):551-78.

22. Dawson JF, Richter AW. Probing three-way interactions in moderated multiple regression: development and application of a slope difference test. Journal of Applied Psychology. 2006;91(4):917.

23. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. Springer; 2005.

24. Agrawal R, Imieliński T, Swami A, editors. Mining association rules between sets of items in large databases. ACM SIGMOD 1993.

25. Becquet C, Blachon S, Jeudy B, Boulicaut J, Gandrillon O. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. Genome Biology. 2002;3(12):0067. 1-. 16.

26. Chen TJ, Chou LF, Hwang SJ. Application of a data-mining technique to analyze coprescription patterns for antacids in Taiwan. Clinical Therapeutics. 2003;25(9):2453-63.

27. Jiao J, Zhang Y. Product portfolio identification based on association rule mining. Computer-Aided Design. 2005;37(2):149-72.

28. Rajak A, Gupta MK, editors. Association rule mining-applications in various areas. International conference on data management. 2008. Ghaziabad, India.

29. Treinen JJ, Thurimella R. A framework for the application of association rule mining in large intrusion detection infrastructures. In International Workshop on Recent Advances in Intrusion Detection. Springer Berlin Heidelberg. 2006:1-18.

30. Bell SM, Edwards SW. Identification and prioritization of relationships between environmental stressors and adverse human health impacts. Environmental Health Perspectives. 2015;123(11):1193-9.

31. Bell SM, Edwards SW, editors. Building associations between markers of environmental stressors and adverse human health impacts using frequent itemset mining. Society for Industrial and Applied Mathematics (SIAM) international conference on data mining; 2014.

32. U.S. Census Bureau. A compass for understanding and using American Community Survey data: What general data users need to know. Washington, DC: U.S. government printing office; 2008.

33. Habermann M, Souza M, Prado R, Gouveia N. Socioeconomic inequalities and exposure to traffic-related air pollution in the city of São Paulo, Brazil. Cadernos de Saúde Pública. 2014;30(1):119-25.

34. Thompson U, Caquard S. Compiling a geographic database to study environmental injustice in Montréal: process, results, and lessons. In Mapping Environmental Issues in the City. Springer Berlin Heidelberg. 2011:10-29.

35. Hahsler M, Grün B, Hornik K, Buchta C. Introduction to arules-A computational environment for mining association rules and frequent item sets. 2009.

36. Hahsler M, Chelluboina S. Visualizing association rules: Introduction to the R-extension package arulesViz. 2011.

37. Borgelt C. Frequent item set mining. Wiley interdisciplinary reviews: data mining and knowledge discovery. 2012;2(6):437-56.

38. Ramsey F, Schafer D. The statistical sleuth: a course in methods of data analysis. Third ed. Boston, MA: Cengage Learning; 2012.

39. Zhang J, Kai FY. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA. 1998;280(19):1690-1.

40. U.S. Census Bureau. American community survey and puerto rico community survey 2014 subject definitions. Available from: https://www2.census.gov/programs-surveys/acs/tech_docs/subject_definitions/2014_ACSSubjectDefinitions.pdf.

41. Hillis DM, Bull JJ. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic biology. 1993;42(2):182-92.

42. Martínez-Fernández J, Esteve-Selma MA, Calvo-Sendín JF. Environmental and socioeconomic interactions in the evolution of traditional irrigated lands: a dynamic system model. Human Ecology. 2000;28(2):279-99.

43. Patterson T, Gulden T, Cousins K, Kraev E. Integrating environmental, social and economic systems: a dynamic model of tourism in Dominica. Ecological Modelling. 2004;175(2):121-36.

44. Kenney MA, Hobbs BF, Mohrig D, Huang H, Nittrouer JA, Kim W, et al. Cost analysis of water and sediment diversions to optimize land building in the Mississippi River delta. Water Resources Research. 2013;49(6):3388-405.

45. Trujillo-Ventura A, Ellis JH. Multiobjective air pollution monitoring network design. Atmospheric Environment. Part A. General Topics. 1991;25(2):469-79.

46. Nau C, Ellis H, Huang H, Schwartz BS, Hirsch A, Bailey-Davis L, et al. Exploring the forest instead of the trees: An innovative method for defining obesogenic and obesoprotective environments. Health & Place. 2015;35:136-46.

47. Huang H, Fava A, Guhr T, Cimbro R, Rosen A, Boin F, et al. A methodology for exploring biomarker--phenotype associations:

application to flow cytometry data and systemic sclerosis clinical manifestations. BMC bioinformatics. 2015;16:293.

**Table 1.** Association Rules (LHS socio-demographic variables and RHS chemical variables, minimum support value of 0.1, lift > 1)

**Table 2.** Association Rules (both LHS and RHS are socio-demographic variables, minimum support value of 0.1, lift > 1)

**Table 3.** Summary of Association Rules (LHS socio-demographic variables and RHS chemical variables, minimum support value of 0.01, lift > 1)

**Table 4.** Complementary Pairs of Rules with One-to-One Relationship (LHS socio-demographic variables and RHS chemical variables, minimum support value of 0.01, lift > 1, Size = 2)

Table 1. Association Rules (LHS socio-demographic variables and RHS chemical variables, minimum support value of 0.1, lift > 1)

| LHS | | RHS | Support | Confidence | Lift | Relative Risk | Odds Ratio |
|---|---|---|---|---|---|---|---|
| Race Minority Score 1 | => | BUTADIENE=Q1 | 0.146 | 0.448 | 1.793 | 2.074 | 2.947 |
| Race Minority Score 1 | => | DIESEL=Q1 | 0.145 | 0.445 | 1.780 | 2.051 | 2.893 |
| Race Minority Score 1 | => | TOLUENE=Q1 | 0.141 | 0.435 | 1.740 | 1.981 | 2.737 |
| Race Minority Score 1 | => | BENZENE=Q1 | 0.134 | 0.412 | 1.647 | 1.830 | 2.411 |
| Race Minority Score 1 | => | ACETALDEHYDE=Q1 | 0.129 | 0.396 | 1.585 | 1.734 | 2.216 |
| Age=40-50 | => | DIESEL=Q1 | 0.125 | 0.375 | 1.499 | 1.615 | 1.984 |
| Age=40-50 | => | BUTADIENE=Q1 | 0.119 | 0.356 | 1.425 | 1.512 | 1.795 |
| Age=40-50 | => | TOLUENE=Q1 | 0.117 | 0.349 | 1.396 | 1.473 | 1.726 |
| Age=40-50 | => | BENZENE=Q1 | 0.115 | 0.344 | 1.375 | 1.445 | 1.679 |
| Race Minority Score 1 | => | CYANIDE=Q3 | 0.108 | 0.332 | 1.328 | 1.383 | 1.573 |
| Age=40-50 | => | ACETALDEHYDE=Q1 | 0.109 | 0.324 | 1.297 | 1.346 | 1.512 |
| Race Minority Score 1 | => | DIESEL=Q2 | 0.102 | 0.315 | 1.259 | 1.297 | 1.433 |
| Race Minority Score 1 | => | TOLUENE=Q2 | 0.102 | 0.315 | 1.258 | 1.297 | 1.433 |

Table 2. Association Rules (both LHS and RHS are socio-demographic variables, minimum support value of 0.1, lift > 1)

| LHS | | RHS | Support | Confidence | Lift | Relative Risk | Odds Ratio |
|---|---|---|---|---|---|---|---|
| Race Minority Score 1 | => | Age=40-50 | 0.172 | 0.530 | 1.583 | 1.801 | 2.704 |
| Age=40-50 | => | Race Minority Score 1 | 0.172 | 0.514 | 1.583 | 1.801 | 2.650 |
| Poverty Score 2 | => | Race Minority Score 1 | 0.110 | 0.435 | 1.338 | 1.397 | 1.702 |
| Poverty Score 2 | => | Age=40-50 | 0.110 | 0.433 | 1.295 | 1.344 | 1.607 |
| Race Minority Score 1 | => | Poverty Score 2 | 0.110 | 0.340 | 1.338 | 1.397 | 1.601 |
| Age=40-50 | => | Poverty Score 2 | 0.110 | 0.329 | 1.295 | 1.344 | 1.512 |

Table 3. Summary of Association Rules (LHS socio-demographic variables and RHS chemical variables, minimum support value of 0.01, lift > 1)

| | **Number of Rules** | **Low Exposure (Q1)** | **Q2** | **Q3** | **High Exposure (Q4)** |
|---|---|---|---|---|---|
| Race Minority Score 7 or 8 or 9 or 10 | 29 | 0 (0%) | 1 (3.45%) | 8 (27.59%) | 20 (68.97%) |
| Race Minority Score 1 or 2 | 342 | 139 (40.64%) | 129 (37.72%) | 58 (16.96%) | 16 (4.68%) |
| Poverty Score 7-10 | 14 | 1 (7.14%) | 1 (7.14%) | 3 (21.43%) | 9 (64.29%) |
| Poverty Score 1 or 2 or 3 | 354 | 140 (39.55%) | 118 (33.33%) | 69 (19.49%) | 27 (7.63%) |
| Education Score 8-10 | 24 | 2 (8.33%) | 3 (12.5%) | 8 (33.33%) | 11 (45.83%) |
| Education Score 1 or 2 or 3 | 237 | 116 (48.95%) | 31 (13.08%) | 39 (16.46%) | 51 (21.52%) |
| Age 40-50 | 213 | 106 (49.77%) | 69 (32.39%) | 34 (15.96%) | 4 (1.88%) |
| Age 38-40 | 83 | 31 (37.35%) | 28 (33.73%) | 16 (19.28%) | 8 (9.64%) |

Table 4. Complementary Pairs of Rules with One-to-One Relationship (LHS socio-demographic variables and RHS chemical variables, minimum support value of 0.01, lift > 1, Size = 2)

| LHS | | RHS | Support | Confidence | Lift | Odds Ratio | Est. 95% CI | |
|---|---|---|---|---|---|---|---|---|
| Race Minority Score 10 | => | DIESEL=Q4 | 0.023 | 0.637 | 2.549 | 5.534 | 5.102 | 6.008 |
| Race Minority Score 1 | => | DIESEL=Q1 | 0.145 | 0.445 | 1.780 | 2.893 | 2.818 | 2.969 |
| Race Minority Score 10 | => | TOLUENE=Q4 | 0.018 | 0.501 | 2.002 | 3.081 | 2.851 | 3.335 |
| Race Minority Score 1 | => | TOLUENE=Q1 | 0.141 | 0.435 | 1.740 | 2.737 | 2.666 | 2.809 |
| Race Minority Score 10 | => | BUTADIENE=Q4 | 0.017 | 0.489 | 1.958 | 2.942 | 2.722 | 3.177 |
| Race Minority Score 1 | => | BUTADIENE=Q1 | 0.146 | 0.448 | 1.793 | 2.947 | 2.869 | 3.025 |
| Race Minority Score 10 | => | BENZENE=Q4 | 0.017 | 0.468 | 1.870 | 2.687 | 2.484 | 2.902 |
| Race Minority Score 1 | => | BENZENE=Q1 | 0.134 | 0.412 | 1.647 | 2.411 | 2.351 | 2.472 |
| Race Minority Score 10 | => | ACETALDEHYDE=Q4 | 0.013 | 0.369 | 1.475 | 1.768 | 1.636 | 1.914 |
| Race Minority Score 1 | => | ACETALDEHYDE=Q1 | 0.129 | 0.396 | 1.585 | 2.216 | 2.161 | 2.272 |

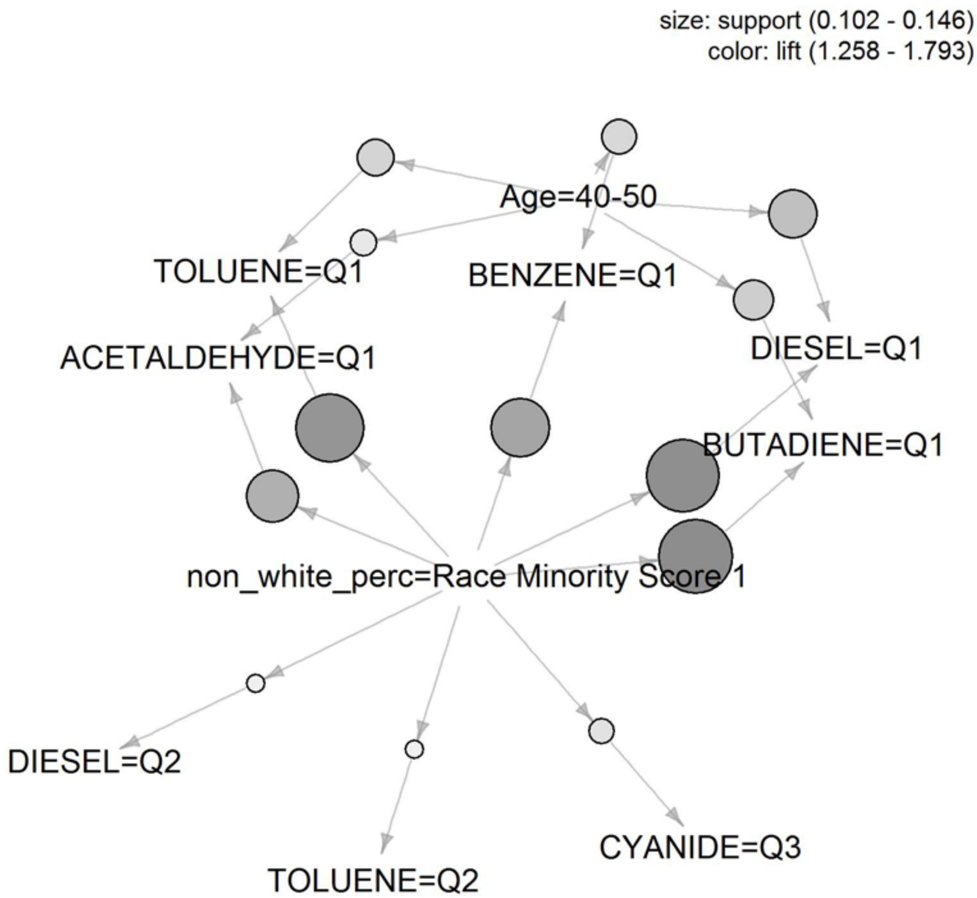


Figure 1. Graph-based Visualization of Association Rules (LHS is socio-demographic variables and RHS is chemical variables, minimum support value of 0.1, lift > 1)