Published in final edited form as:

Environ Int. 2014 February ; 63: 236–245. doi:10.1016/j.envint.2013.11.004.

Modeling and analysis of personal exposures to VOC mixtures using copulas

Feng-Chiao Su^a, Bhramar Mukherjee^b, and Stuart Batterman^{a,1}

^aEnvironmental Health Sciences, School of Public Health, University of Michigan, USA

^bBiostatistics, School of Public Health, University of Michigan, USA

Abstract

Environmental exposures typically involve mixtures of pollutants, which must be understood to evaluate cumulative risks, that is, the likelihood of adverse health effects arising from two or more chemicals. This study uses several powerful techniques to characterize dependency structures of mixture components in personal exposure measurements of volatile organic compounds (VOCs) with aims of advancing the understanding of environmental mixtures, improving the ability to model mixture components in a statistically valid manner, and demonstrating broadly applicable techniques. We first describe characteristics of mixtures and introduce several terms, including the mixture fraction which represents a mixture component's share of the total concentration of the mixture. Next, using VOC exposure data collected in the Relationship of Indoor Outdoor and Personal Air (RIOPA) study, mixtures are identified using positive matrix factorization (PMF) and by toxicological mode of action. Dependency structures of mixture components are examined using mixture fractions and modeled using copulas, which address dependencies of multiple variables across the entire distribution. Five candidate copulas (Gaussian, t, Gumbel, Clayton, and Frank) are evaluated, and the performance of fitted models was evaluated using simulation and mixture fractions. Cumulative cancer risks are calculated for mixtures, and results from copulas and multivariate lognormal models are compared to risks calculated using the observed data. Results obtained using the RIOPA dataset showed four VOC mixtures, representing gasoline vapor, vehicle exhaust, chlorinated solvents and disinfection by-products, and cleaning products and odorants. Often, a single compound dominated the mixture, however, mixture fractions were generally heterogeneous in that the VOC composition of the mixture changed with concentration. Three mixtures were identified by mode of action, representing VOCs associated with hematopoietic, liver and renal tumors. Estimated lifetime cumulative cancer risks exceeded 10^{-3} for about 10% of RIOPA participants. Factors affecting the likelihood of high concentration mixtures included city, participant ethnicity, and house air exchange rates. The dependency structures of the VOC mixtures fitted Gumbel (two mixtures) and t (four mixtures) copulas, types that emphasize tail dependencies. Significantly, the copulas reproduced both risk predictions and exposure fractions with a high degree of accuracy, and performed better than multivariate

© 2013 Elsevier Ltd. All rights reserved

¹Room 6507 SPH2, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA. Tel.: +1 734 763 2417; fax: +1 734 936 7283. stuartb@umich.edu (S. Batterman).

Appendix A. Supplementary data Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.envint.2013.11.004>.

lognormal distributions. Copulas may be the method of choice for VOC mixtures, particularly for the highest exposures or extreme events, cases that poorly fit lognormal distributions and that represent the greatest risks.

Keywords

Copula; Cumulative effects; Exposure determinants; Mixture; RIOPA; VOC

1. Introduction

Environmental mixtures have been defined as the combination of two or more chemical components, regardless of the sources or the spatial or temporal proximity where exposures occur (US EPA, 1986). Environmental exposures typically involve mixtures of pollutants that occur either simultaneously or sequentially, and over both short and long periods. Exposure to mixtures of environmental pollutants is very common and, if mixture components can interact or jointly contribute to adverse health or environmental impacts, then estimates of adverse effects and risks based on single compounds will be underestimated. As a result, there is growing interest and concern regarding the cumulative effects of mixtures. While a few regulations address mixtures, e.g., ambient air standards limit exposures to airborne particulate matter and diesel exhaust (US EPA, 2012a, 2012d), an occupational exposure limit exists for volatile organic compounds (VOCs) in gasoline vapor (ACGIH, 2012), and a drinking water regulation limits the sum of the four trihalomethanes (US EPA, 2013), most regulations and guidelines focus on single pollutants.

Effects of exposures to mixtures can be evaluated using empirical data from the same mixture or estimated using data from similar mixtures (ATSDR, 2004). However, the most common approach is to calculate risks with the assumption of response additivity among mixture components, as described later. To account for possible toxicological interactions between mixture components, especially for complex mixtures, a combination of mechanistic and statistical models has been suggested (Feron and Groten, 2002). As examples, physiologically-based pharmacokinetics (PBPK) models that incorporate interactions (Conolly, 2001; Feron and Groten, 2002; Haddad et al., 2001) have been used to evaluate the pharmacokinetics of exposures to gasoline (Dennison et al., 2004). The underlying data for these models (and for other analyses) must account for joint likelihood of exposure to the mixture components, especially high concentration mixtures. However, the literature addressing exposures to mixtures is not well developed, a meaningful limitation given that the composition of mixtures can vary considerably.

The objective of this study is to advance our understanding of exposures to environmental mixtures by improving the ability to model correlated mixture components. Using a large dataset of personal exposure measurements of VOCs, we first identify mixtures of interest using factor analysis, and empirically determine groups of VOCs based on correlation. We also select VOC mixtures using the toxicological mode of action. Second, dependency structures within the mixtures are modeled using copulas, a powerful and broadly applicable technique that can represent multiple types of dependencies that has not been used

previously to examine environmental exposures. Third, the importance of dependencies in mixture exposures is demonstrated using mixture fractions and by comparing risk estimates for VOC mixtures simulated using copulas and conventional multivariate distributions.

1.1. Definitions for mixtures assessment

Several definitions aid the understanding and analysis of mixtures. Three classes of mixtures have been defined (ATSDR, 2004): (1) “generated mixtures” composed of compounds which are generated concurrently from the same process, e.g., by-products of fuel combustion or cigarette smoke; (2) “intentional mixtures” composed of related compounds typically used to manufacture commercial products, e.g., gasoline; and (3) “coincidental mixtures” of unrelated compounds that are disposed or stored and reach the same target population, e.g., metals, solvents and semivolatile wastes at Superfund sites. Generated and intentional mixtures may be common in some settings, for example, in workplaces and homes. However, exposure to multiple air pollutants emitted from different outdoor sources, e.g., CO, PM_{2.5} and benzene from vehicles, and SO₂ from power plants is also common and can be considered coincidental mixtures.

Next, we define the “mixture fraction” as a component's fractional contribution to the total concentration of the mixture. These fractions help to show the influence of mixture components. Also, changes in the mixture fraction associated with the total mixture concentration (or other variables) can show trends and help reveal the mixture's source, e.g., fractions for generated or intentional mixtures should be constant. Mixtures with consistent mixture fractions across a population or over time are considered “homogeneous.” In contrast, highly variable or “heterogeneous” mixture fractions are likely to reflect coincidental mixtures.

“Dependencies” among components of mixtures refer to the statistical relationships among concentrations of each component in the mixture, and potentially to the composition of the mixture itself. The most common indicator of dependencies uses correlation measures, e.g., Pearson correlation coefficients (r), parametric measures that assume variables are normally distributed, and Spearman's rho and Kendall's tau, non-parametric measures that are robust with respect to outliers. Often, exposures are not normally distributed, but contain extreme values and remain right-skewed even after log-transformation (Jia et al., 2008), thus parametric correlation measures can have significant limitations. In addition, both types of correlation measures reveal only pair-wise dependencies (and not those involving three or more variables), and may not be reliable indicators in the presence of non-linear associations (Schmidt, 2006; Staudt, 2010).

Risk evaluations sometimes define “simple” and “complex” mixtures (Feron et al., 1998). Simple mixtures contain a relatively small number (<10) of components. Often, such mixtures have been identified and their components well quantified, e.g., medicines and pesticides. In contrast, “complex mixtures” include many more components, are usually incompletely quantified, and can be highly variable, e.g., gasoline vapor and tobacco smoke. Mixtures can also be classified by the nature of the effects or interactions between mixture components. Following the methods recommended to analyze “cumulative risks” of mixtures (ATSDR, 2004; US EPA, 2000a, 2003), mixture components can be considered to

have “independent toxicities,” meaning that each chemical has a different mode of action and that the overall response is obtained by adding responses of each component, which is called “response addition” (Bliss, 1939). For example, cumulative risks of cancer have been estimated using response addition across 13 VOCs (e.g., benzene, 1,3-butadiene, chloroform, formaldehyde, styrene, acetaldehyde, etc.), and 6 metals (chromium VI, nickel, arsenic, lead, cadmium, and beryllium) (Sax et al., 2006). If mixture components have similar toxicity effects or mechanisms, then doses can be added, called “dose addition.” An example is the use of toxic equivalency factors (TEFs) for polycyclic aromatic hydrocarbons (PAHs), which relate the relative potency of compounds in the mixture to a reference compound (benzo(a)pyrene), which are used as weights in summing doses or concentrations to estimate the mixture's toxicity (US EPA, 1993). US EPA (1986) suggests that if interaction information is unavailable, then the additive assumption should be adopted.

1.2. Application of copulas

Copulas represent a powerful technique for representing dependencies that can overcome shortcomings of conventional correlation measures. Introduced in 1959 by Sklar, a copula represents the dependency structure of two or more variables across the entire distribution (Frees and Valdez, 1998; Sklar, 1959). Copulas separate the dependency structure(s) from the variables' marginal distributions, a major advantage, and thus are unconstrained by the marginal distribution (Genest and Favre, 2007). While unrestricted, the choice of the marginal distributions affects the location and scale structure of copulas (Frees and Valdez, 1998).

Copulas transform the marginal distribution of each variable into a uniform distribution over the interval [0,1], after which the dependency structure is described following reference distributions. Once the dependency structure and marginal distributions are known (or estimated), the joint distribution function is:

$$C(u_1, u_2, \dots, u_p) = \text{prob}(U_1 \leq u_1, U_2 \leq u_2, \dots, U_p \leq u_p)$$

where C is a copula function, U_i for $i = 1, \dots, p$ are uniformly transformed random variables corresponding to the marginal distribution functions $F_i(x_i)$, and p is the number of variables. The joint distribution function can also be expressed as:

$$C[F_1(x_1), F_2(x_2), \dots, F_p(x_p)] = F(x_1, x_2, \dots, x_p).$$

According to Sklar's (1959) theorem, if F_i is continuous and x_i is over $[-\infty, \infty]$, then C is unique.

Copulas allow dependency structures to be weighted in different manners, and thus can be symmetric or asymmetric (Staudt, 2010). There are several families and many types of copulas, which have different origins and properties. The family of elliptical copulas is derived from distributions, e.g., the Gaussian copula is from the multivariate normal distribution, and the t copula from the multivariate Student t distribution. Given the same correlation coefficient, t copulas provide a better fit to distributions that include extreme

values than Gaussian copulas, i.e., the t copula more accurately models tail dependencies (Schmidt, 2006). Among Archimedean copulas, which are stated directly and not derived from distributions, Gumbel copulas emphasize upper tail dependency, Clayton copulas emphasize lower tail dependency, while Frank copulas have no emphasis on tail dependency, i.e., symmetrical dependencies on both tails (Schmidt, 2006). The product copula, the simplest copula, indicates independence between random variables (Trivedi and Zimmer, 2007).

Copulas have been widely applied in finance, especially for derivative pricing and financial risk management to address market, credit and operational risks where classical approaches to describe market and other fluctuations, i.e., using multivariate normal distributions, have been shown to be lacking (Cherubini et al., 2004; Jean-Frédéric et al., 2004). Given that environmental exposures also involve non-normal distributions and extreme values (Jia et al., 2008; Su et al., 2012), copulas could be a good tool to explore dependency structures of multivariate exposures. In earlier work using VOCs measured in the National Health and Nutrition Examination Survey (NHANES), we showed that marginal distributions fitted lognormal, Pareto and Weibull distributions (depending on the VOC), and that product, Gumbel, Clayton, Frank and Gaussian copulas fitted bivariate dependency structures (Jia et al., 2010). While there have been a few examples (Jia et al., 2010; Wang et al., 2012), environmental applications of copulas remain very limited.

2. Materials and methods

2.1. Data sources

The exposure dataset was drawn from the Relationship of Indoor Outdoor and Personal Air (RIOPA) study, which was designed to estimate the contribution of outdoor and indoor sources to personal exposures of air pollutants (Weisel et al., 2005a, 2005b). RIOPA was conducted in three U.S. cities (Elizabeth, NJ, Houston, TX, and Los Angeles, CA) from summer 1999 to spring 2001. Approximately 100 non-smoking households and adults and children in each city were recruited. Participants completed three questionnaires (based on the National Human Exposure Assessment Survey) that addressed demographics, health status (e.g., respiratory illness, heart condition), lifestyle factors and time activity patterns. Additional information was obtained by a walkthrough inspection conducted in each home by a technician. Outdoor, indoor, and personal VOC sampling of VOCs was conducted in visits to each household; this sampling was repeated about three months later. A total of 544 personal adult VOC samples were collected during these visits, including 299 samples collected at the first visit and 245 at the second (239 complete pairs). To avoid possible biases involved in the repeated measurements (i.e., cluster effects) and to maintain the largest sample size, the present study uses only the measurements collected at the first visit.

A total of 18 VOCs (benzene, toluene, ethylbenzene, m,p-xylene, o-xylene, methyl tertiary-butyl ether (MTBE), styrene, 1,4-dichlorobenzene (1,4-DCB) methylene chloride, trichloroethylene (TCE), tetrachloroethylene (PERC), chloroform, carbon tetrachloride (CTC), d-limonene, α -pinene, β -pinene, 1,3-butadiene and chloroprene) were collected using passive samplers (OVM3500, 3M Company, St. Paul, MN, USA) and 48-h sampling periods. Samples were analyzed using gas chromatography-mass spectrometry with method

detection limits (MDLs) from 0.21 (α -pinene and PERC) to 7.1 (toluene) $\mu\text{g m}^{-3}$ (Weisel et al., 2005b). Measurements below the MDL were substituted with one-half the MDL. 1,3-Butadiene, chloroprene and methylene chloride were excluded due to invalid measurements. Styrene has higher uncertainty due to biased inter-laboratory consistency. A new variable, TVOC (total volatile organic compound), was defined as the sum of the remaining 15 VOCs. Detection frequencies ranged from 31% (TCE) to 96% (MTBE), and only TCE and styrene had detection frequencies below 50%. Details of the RIOPA study design and several analyses of indoor, outdoor and personal VOC results are provided elsewhere (Kwon et al., 2006; Weisel et al., 2005a, 2005b).

2.2. Identification of mixtures

Exposure mixtures in the RIOPA dataset were selected using two approaches. The first identified common VOC mixtures using positive matrix factorization (PMF), a multivariate analysis similar to factor analysis, but with the ability to incorporate uncertainties on each measurement in order to represent sampling errors and MDLs (Anderson et al., 2001; Paatero and Tapper, 1994). PMF is often used for source apportionments, especially for ambient particulate matter, with the goal of identifying factors and quantifying factor contributions. Based on the uncertainty, PMF models variables as weak or strong, i.e., variables with high (or low) uncertainties are assigned weak (or strong) influence. Each VOC was given an uncertainty equal to the measurement precision, estimated as the pooled coefficient of variation for duplicate samples (Weisel et al., 2005b). Styrene was designated as “weak” due to potentially biased measurements, and TVOC was automatically designated as “weak” by the model. Measurements below MDLs were retained, but assigned large uncertainties in order to reduce their influence (USEPA, 2008).

PMF decomposes two matrices from the sample data: a matrix of factor profiles, which represents the mass and percentage of each species apportioned to the factor, and a matrix of factor relative contributions, which gives the contribution of each factor to the total concentration of each observation (USEPA, 2008). Because there is no optimal or *a priori* manner to select the number of factors, PMF analyses were conducted using 3, 4 and 5 factors, and each was tested using goodness-of-fit indicators, e.g., scaled residuals and Q values. The latter is the sum of squares of the residuals divided by the uncertainties for the concentrations of individual compounds (Anderson et al., 2001; USEPA, 2008). These analyses used PMF 3.0, a peer-reviewed receptor modeling tool (USEPA, 2008).

An analysis was undertaken to identify personal, behavioral and environmental variables associated with high exposure mixtures. VOC mixtures identified using PMF were divided into high and low groups using a 75th percentile cutoff of the mixture's total concentration, which was modeled as the dependent variable in bivariate logistic regression models. Candidate explanatory variables were based on our earlier work that identified determinants of VOC exposure, and included city, ethnicity, employment status, the presence of attached garage, self-service pumping gas, open doors or windows, other family members taking showers, the use of fresheners, and household air exchange rates. This analysis used PROC LOGISTIC in SAS 9.2 (SAS Institute, Cary, North Carolina, USA).

The second approach for selecting mixtures used the toxicological mode of action, which considers the biochemical pathways and outcomes potentially affected by pollutant exposure (Borgert et al., 2004). Two mixtures were considered that had cancer endpoints: (1) VOCs associated with hematopoietic cancers (lymphomas and leukemia), including benzene, MTBE, 1,4-DCB, TCE and PERC; and (2) VOCs associated with liver and renal tumors, including ethylbenzene, MTBE, 1,4-DCB, TCE, PERC, chloroform and CTC (Borgert et al., 2004; IARC, 2012). The two modes of action mixtures contained 5 and 7 components, respectively. It should be noted that mixtures based on mode of action represent a completely different approach from selecting variables using PMF or other correlation type measures, which are driven exclusively by the pattern of occurrence.

To reduce the number and complexity of analyses in mixtures containing a larger number of components, highly correlated VOCs were grouped together based on their likely emission sources or chemical characteristics. For example, the seven VOCs in the liver and renal tumor mixture were trimmed to a group of gasoline-related compounds (ethylbenzene and MTBE), and chlorinated hydrocarbons (1,4-DCB, TCE, PERC, chloroform and CTC). The analysis then proceeded with these groups.

2.3. Copula selection

The dependency structures of each mixture were fitted to copulas using maximum likelihood estimates (MLEs), five candidate copulas (Gaussian, t, Gumbel, Clayton, and Frank), and the observed marginal distributions. Goodness-of-fit (GOF) tests were conducted using Akaike and Bayesian information criteria (AIC and BIC), and the copula with the lowest criteria was chosen as the best-fit dependency structure.

After choosing the best-fit copulas, two sets of objects necessary for simulating joint distributions (discussed in the next section) were generated, namely, uniform [0,1] random variables for each component of the mixture that followed the copula-identifying correlations, and copula parameters estimated using MLE. Copula parameters were the covariance matrix of the mixture components for the Gaussian copula; the same matrix along with the number of degrees of freedom for the t copula; and (unique) correlation parameters estimated for the Gumbel, Clayton and Frank copulas.

2.4. Simulated distributions

GOF of the fitted copulas was tested using simulations, uniform random variables, parameters fitted for each copula (described above), and parametric marginal distributions fitted using MLE for each VOC. A large number ($n = 1000$) of pseudo observed data were generated for each mixture. Using the pseudo observed data, the probabilities that all components in the mixture exceeded the 50th, 75th, 90th and 95th percentile cutoffs were calculated and compared to observed data. We also calculated probabilities assuming independence among mixture components, e.g., the probability of a three component mixture in which each component exceeded the 90th percentile concentration is 0.001 ($p = 0.1^3$). Because styrene and TCE had low detection frequencies (49 and 31%, respectively), probabilities that all mixture components exceeded the 50th percentile cannot be calculated.

To examine the influence of each mixture component and identify any trends associated with concentration, mixture fractions were calculated for both observed and simulated data, and results were summarized using the median fraction in several bins (50–75th, 75–90th, 90–95th, 95–100th percentiles) for each mixture. This analysis indicates, for example, whether the composition of the mixture remained constant or shifted with concentration, the latter indicating more complex dependencies.

For VOC mixtures based on mode of action, cumulative cancer risks were estimated assuming response addition following EPA guidance (US EPA, 2000b):

$$\text{Excess cumulative lifetime cancer risk} = \sum R = \sum_i (C_i \text{URF}_i)$$

where $\sum R$ = excess individual lifetime cancer risk, C_i = concentration of i th VOC ($\mu\text{g m}^{-3}$), and URF_i = unit risk factor (cancer cases per $\mu\text{g m}^{-3}$). We also computed the fraction of individuals with cumulative risks exceeding 10^{-6} , 10^{-5} , 10^{-4} , 10^{-3} and 10^{-2} , and compared results obtained using observed data, copula simulations, and multivariate lognormal distributions based on the observed means and variance/covariance matrix. Cumulative probability plots were used to visualize differences between observed data and simulations.

Copula fitting and simulations were performed using ModelRisk 5 Industrial edition (Vose Software BVBA, Gent, Belgium). Multivariate lognormal distributions were simulated using `rlnorm.rplus` in R version 2.13.1 (R Development Core Team, Vienna, Austria) and Excel (Microsoft, Redmond, WA).

3. Results and discussion

3.1. VOC mixtures identified using PMF

The PMF analysis using 3-factors grouped many VOCs together in a single factor, e.g., benzene, MTBE, 1,4-DCB, TCE, PERC, chloroform and CTC. In comparison, the 4-factor analysis had two to five VOCs in each factor, while the 5-factor analysis placed a single VOC (d-limonene) in its own factor (other factors remained similar to those in the 4-factor analysis). The 4- and 5-factor results were considered the most informative for identifying mixtures and for fitting copulas. (Full results of the 3-, 4- and 5-factor PMF analyses are shown in Supplemental Table S1.) Based on the PMF analysis, we identified four VOC mixtures, which were designated as A1-A4 and which explained 20.5, 20.9, 16.3 and 42.3%, respectively, of the variation in ΣVOC levels in the RIOPA dataset (Table 1 and Fig. 1).

- Mixture A1 contained benzene (average contribution = $1.4 \mu\text{g m}^{-3}$) and MTBE ($11.2 \mu\text{g m}^{-3}$), and is identified as “gasoline vapor”. Both VOCs are highly volatile components of gasoline and reflect the gasoline composition during the RIOPA sampling era (1999 to 2001). Currently, benzene levels are considerably lower, e.g., benzene content is now limited to 0.62% of the fuel (US EPA, 2007), and MTBE use has been phased out (starting in 2000, fully in 2006) (US EPA, 2012c) after extensive use in California, New Jersey, and Texas (US EPA, 2008).

- Mixture A2 is designated as “vehicle exhaust” due to contributions from toluene ($4.9 \mu\text{g m}^{-3}$), ethylbenzene ($1.9 \mu\text{g m}^{-3}$), m,p-xylene ($5.5 \mu\text{g m}^{-3}$), o-xylene ($1.7 \mu\text{g m}^{-3}$) and styrene ($0.2 \mu\text{g m}^{-3}$). These VOCs are also highly volatile components of gasoline and diesel fuels as well as exhaust emissions from gasoline- and diesel-powered vehicles (ATSDR, 2007, 2010a, 2010b).
- Mixture A3 included several common indoor contaminants, including a moth repellent (1,4-DCB at $0.9 \mu\text{g m}^{-3}$), chlorinated solvents (TCE at $0.2 \mu\text{g m}^{-3}$, PERC at $1.7 \mu\text{g m}^{-3}$, CTC at $0.5 \mu\text{g m}^{-3}$), and a water disinfection by-product (chloroform at $0.8 \mu\text{g m}^{-3}$). These VOCs are fairly specific to these sources, e.g., 1,4-DCB is the major ingredient of mothballs (ATSDR, 2006) (although similar repellents often use naphthalene). PERC is a widely used dry cleaning solvent (ATSDR, 1997b). Chloroform is a by-product of water disinfection (ATSDR, 1997a). TCE and CTC have been widely used in industry as degreasers, chemical intermediates, and pesticides (ATSDR, 1997c, 2005).
- Mixture A4 contained d-limonene ($20.5 \mu\text{g m}^{-3}$), α -pinene ($1.5 \mu\text{g m}^{-3}$) and β -pinene ($2.7 \mu\text{g m}^{-3}$), which are fragrances and solvents indicative of “cleaning products and odorants.” Both d-limonene and pinene are widely used flavors and fragrance additives in cleaning products, fresheners, other consumer products, and even in foods and beverages (IARC, 1993; US EPA, 2012b). This factor explained the largest portion (42.3%) of the total VOC exposure, a result of the VOCs included in RIOPA, the large fraction of time most people spent indoors (see below), the wide use of the VOCs in this mixture, and the high concentrations of these VOCs relative to others measured in RIOPA.

Similar source profiles (gasoline vapor, vehicle exhaust, deodorizer and shower, and dry cleaning) have been observed in a study using PMF and the NHANES dataset, although NHANES did not measure d-limonene, α -pinene or β -pinene, and the dominant mixtures were gasoline vapor and the vehicle exhaust (Jia et al., 2010).

Because many RIOPA participants were older (average age = 45 years old; 24% were 60 years old) and predominantly female (75%), we suspected that the indoor residential fraction would be especially significant. Using the 48-h time-activity information collected in RIOPA, we calculated time fractions for each participant in seven compartments, i.e., time spent outdoors in the participant's neighborhood, outdoors away from neighborhood, indoors at home, indoors at school/work, other indoor, transportation, and unknown. RIOPA participants spent an average of 91% of time indoors – higher than the national average of 87% (Klepeis et al., 2001). Time fractions varied slightly, but significantly, by city (89, 92 and 92% for Los Angeles, Elizabeth and Houston, respectively, $p < 0.0001$).

Identifying emission sources is an essential step for implementing any exposure reduction strategy. PMF provides a concentration-based approach that can identify generated mixtures, defined earlier as mixtures that arise from a common source or multiple correlated emission sources. However, VOC mixtures identified using correlations can also reflect common factors affecting contaminant transport and fate, e.g., building air exchange rates, as well as common behavioral patterns, e.g., the use of certain types of cleaning products. Thus,

mixtures identified using PMF or other correlation-based methods may not be uniquely generated mixtures, but rather a combination of generated, intentional and possibly coincidental mixtures. It should also be noted that unlike the mixtures based on the mode of action (discussed below), mixtures identified by PMF should be orthogonal, that is, uncorrelated.

3.2. High exposure mixtures

The analysis of high exposure mixtures identified several variables associated with high exposures (Supplemental Table S2). Based on comparisons of the top quartile to the remainder of the data, the following variables were significant predictors (95% confidence interval excluding 1), except as noted:

- City effect: Compared to the Houston participants, individuals in Los Angeles and Elizabeth were less likely to have high exposure (75th percentile) of all mixtures (ORs from 0.18 to 0.63), except mixture A3 for the Elizabeth participants.
- Race/ethnicity: Mexicans were more likely to experience high exposures to mixtures A1 (benzene and MTBE), A3 (1,4-DCB, TCE, PERC, chloroform and CTC), and A4 (d-limonene, α -pinene and β -pinene) compared to Whites (ORs from 2.03 to 3.97). Compared to Whites, Hispanics had increased likelihood of high exposure to mixture A3 (OR = 1.78, 95% CI = 1.09–2.92), while Asians, Blacks and Indians were less likely to have high exposure to mixture A2 (toluene, ethylbenzene, xylene, and styrene, OR = 0.47, 95% CI = 0.24–0.92).
- Employment: Being employed decreased the likelihood of high exposure to mixture A4 (OR = 0.40, 95% CI = 0.27–0.61).
- Air exchange rates: Living in a well ventilated dwelling (higher log transformed air exchange rate) lowered risks of high exposure to all mixtures (ORs from 0.38 to 0.69), especially mixtures associated with strong indoor sources, e.g., mixture A4 containing d-limonene and pinene.
- Open doors or windows: Participants reporting opening doors or windows during the sampling period were less likely to have high exposure of all mixtures except A1 (ORs from 0.32 to 0.40). As seen for air exchange rates, effect of opening doors or windows was most pronounced for mixture A4.
- Attached garages: Participants living in houses with attached garages had increased odds of high exposure to mixtures A1 (gasoline vapor, OR = 2.27, 95% CI = 1.45–3.56) and A2 (vehicle exhaust, OR = 1.95, 95% CI = 1.25–3.05).
- Participant activities: Participants who self-pumped gas during the sampling period had greater chance of high exposure to the gasoline mixture A1 (OR = 2.10, 95% CI = 1.35–3.52). Participants using air fresheners had greater chance of high exposure to the d-limonene, α -pinene and β -pinene mixture A4 (OR = 2.20, 95% CI = 1.17–4.14).
- Activities of family members: Family members showering during the sampling period were more likely to have high exposure to mixtures A3 (moth repellents,

chlorinated solvents and water disinfection by-products, OR = 2.06, 95% CI = 1.20–3.56) and A4 (cleaning and odorant mixtures, OR = 2.45, 95% CI = 1.42–4.23).

Notably, city, ethnicity, and air exchange rates were significantly associated with all VOC mixtures. In addition, several factors identified for gasoline and vehicle exhaust mixtures for the RIOPA participants have been shown for the personal exposure measurements in NHANES, e.g., the presence of attached garages and self-pumped gas was associated with benzene, toluene and MTBE exposures (Jia et al., 2010). However, statistically significant factors were not identified for 1,4-DCB and chloroform in the NHANES dataset. The factors associated with this mixture may result from demographic differences between datasets. Specifically, in comparison to NHANES, participants in RIOPA tended to be older, female, unemployed, and at home more often (Su et al., 2012), all of which may have increased the importance of indoor sources of 1,4-DCB and chloroform.

3.3. Copulas

Copula types selected to match the PMF- and mode-of-action mixtures are listed in Table 2. (Parameters of marginal distributions, GOF statistics, and copulas are in Supplemental Tables S3 to S5.) The AICs and BICs for the different copulas were fairly similar for mixtures A1 (benzene, MTBE), A3/B3 (1,4-DCB, TCE, PERC, chloroform, CTC), A4 (d-limonene, α -pinene, β -pinene) and B1 (ethylbenzene, MTBE). However, AICs and BICs for mixtures A2 (toluene, ethylbenzene, xylene, styrene) and B2 (benzene, MTBE, 1,4-DCB, TCE, PERC) were much lower for Gaussian and t copulas, indicating that these copulas had enhanced ability for describing dependency structures. Gumbel copulas best fitted mixtures A1 and B1, which included two VOCs, while t copulas best fitted mixtures A2, A3, A4 and B2, which contained at least four VOCs. These results reflect the nature of the VOC exposure data in RIOPA, which tends to follow extreme value distributions (Su et al., 2012), since both Gumbel and t copulas are better able to represent extreme values than the types of copulas (Schmidt, 2006).

The fitting results might have also been affected by detection frequency. Since data below the MDLs were assigned a single value (0.5 MDL), non-detects formed “ties” in the distribution. Scatter plots for two variables in a mixture that contain many ties will display a star-like shape that fits the t copula. In contrast, mixtures A1 (benzene and MTBE) and B1 (ethylbenzene and MTBE) contained at least one component with a very high detection frequency (e.g., 96% for MTBE), and joint distributions did not show this star shape. To help test this explanation, we evaluated a mixture of two VOCs with low detection frequencies (styrene at 49% and α -pinene at 66%) and found that the t copula provided the best fit. This suggests that copula fits are not influenced by the number of mixture components; rather, mixtures containing components with low detection frequencies are better fitted by the t copula.

Table 2 contrasts the probability of exceeding specified percentile cut-offs between the observed data and predictions using the copula simulations. Differences between observations and predictions were usually small. For binary mixtures A1 and B1, differences ranged from 0.001 (A1 at the 90th percentile and B1 at 50th, 75th, and 95th percentiles) to

0.020 (B1 at the 75th percentile). For mixtures with three or more components, differences ranged from 0.001 (B2 at the 95th percentile) to 0.12 (A4 at the 50th percentile). These results suggest that copulas have better predictive ability for bivariate distributions than for distributions of three or more variables. Crossing probabilities calculated assuming that mixture components are uncorrelated (or independent), also shown in Table 2, fell far below observed data, especially at higher percentiles, as expected. For example, the observed 90th percentile probability for the odorant mixture A4 (d-limonene, α -pinene and β -pinene) was 0.023, but only 0.001 if components are assumed uncorrelated. Such large differences demonstrate the need to account for dependencies in mixtures.

As noted earlier, Gumbel and Gaussian copulas have been shown to best fit VOCs in NHANES that were highly correlated (Jia et al., 2010). However, this earlier study examined only bivariate mixtures, and did not consider the t copula that best fitted much of the RIOPA data. The present study also found a Gumbel dependency structure for the benzene and MTBE mixture.

3.4. Mixture fractions

Median mixture fractions are shown in Table 3. Copula simulations matched mixture fractions for the dominant components in all mixtures and at all levels, with the sole exception of mixture B2 at the 75 to 90th percentile level. Often, a single compound dominated the mixture, e.g., MTBE accounted for 78 to 94% of the exposure in mixtures A1 and B1. VOCs with strong indoor sources, e.g., 1,4-DCB and d-limonene, dominated mixtures A3 and A4, respectively, and their mixture fraction increased with percentile, e.g., the median observed and predicted mixture fractions for 1,4-DCB in mixture A3 (1,4-DCB, TCE, PERC, chloroform, CTC) at the 50–75th percentile were 0.33 and 0.45, respectively; these both increased to 0.99 at the 95–100th percentile, reflecting the extreme values previously identified for 1,4-DCB and d-limonene (Su et al., 2012). In contrast, mixture fractions varied little for mixtures A1, A2 and B1, e.g., toluene was the dominant component in mixture A2 (toluene, ethylbenzene, xylenes and styrene) with mixture fractions of 0.58 and 0.56 for observed data and simulations, respectively, at the 50–75th percentile level, and 0.57 and 0.53, respectively, at the 90–95th percentile. As noted earlier, consistent mixture fractions may suggest generated mixtures. In contrast, mixture B2 shifted composition at upper percentiles, e.g., the MTBE mixture fractions were 0.61 and 0.55 at the 50–75th percentile levels for observed data and simulations, respectively, but 1,4-DCB was dominant at the 95–100th percentile with mixture fractions of 0.98 and 0.94, respectively. Thus, mixture B2 is heterogeneous in that its composition differs by exposure level. In addition, it may be considered an “incidental” mixture as it likely combined VOCs from different sources. (Note that mixture B2 was selected based on the similar mode-of-action for the constituent VOCs, and not on the basis of common sources or high correlations.)

Mixtures A3/B3 and B2 were used to investigate whether mixture fractions estimated by the copulas were driven by the copula type or the marginal distribution of the mixture's components. These mixtures were simulated for five types of copulas, all using the same set of marginal distributions. (The marginal distributions for these simulations are shown in Supplemental Table 3; mixture fractions are shown in Supplemental Table S6.) For mixture

A3/B3, median mixture fractions showed only small changes, e.g., 1,4-DCB remained the dominant component at high exposure levels and its mixture fraction increased with percentile. Mixture B2 showed larger differences between median fractions for the (best-fit) t and other copulas, and the dominant VOC at the 90 to 95th percentile level differed among copulas, e.g., the dominant VOC was 1,4-DCB for both t and Clayton copulas, but MTBE for the Gaussian, Gumbel and Frank copulas. Even though t and Clayton copulas identified 1,4-DCB, its mixture fraction varied from 0.47 to 0.70 in the two copulas. This analysis highlights the importance of the copula type as well the marginal distributions of the mixture's components.

3.5. Estimated cancer risks

Estimated cancer risks for the mode-of-action mixtures B1 to B3 are shown in Table 4. The observed data reveal that these mixtures can present high cancer risks, e.g., about 10% of RIOPA participants had exposures of mixtures B2 and B3 associated with a 10^{-3} or higher lifetime risk of hematopoietic, liver and renal cancers. Mixture B1 (ethylbenzene and MTBE) posed lower risks of liver and renal cancers, but individuals still had a 25% chance of exceeding a 10^{-5} risk and a 1% chance of exceeding a 10^{-4} risk. For mixture B2 (benzene, MTBE, 1,4-DCB, TCE and PERC), 3% of participants exceeded a high risk level of 10^{-2} for hematopoietic cancers. Results for mixture B3 (1,4-DCB, TCE, PERC, chloroform and CTC) were similar.

Copula simulations gave risk predictions that resembled the observed data, although there was some divergence at the highest exposures, particularly for mixture B3 (Table 4, Fig. 2). For mixture B1 (liver and renal cancers), the highest risks ($>10^{-3}$) were underestimated by both copulas and the lognormal simulations, although copulas had smaller errors. Lognormal simulations slightly overestimated the chance of exceeding a 10^{-5} risk, but underestimated higher risks. For example, moving vertically on Fig. 2 at the 10^{-5} risk level, the observed data, copula simulations and lognormal simulations respectively predict that 25, 27 and 32% of individuals will exceed this risk level. For a 10^{-4} risk, the likelihoods are 1.0, 0.6, and 0.0%, respectively. For mixture B2 (hematopoietic cancers), lognormal simulations again overestimated low to moderate risks (10^{-6} to 10^{-4}), and both copula and lognormal simulations underestimated the highest risks (10^{-3} to 10^{-2}). For mixture B3 (liver and renal cancers), the lognormal simulations significantly underestimated the highest risks (10^{-2}). The cumulative probability plot (Fig. 2) shows that the copulas sometimes overpredicted the highest risks (not revealed by Table 4), e.g., the highest observed risk for mixture B3 was 3.0×10^{-2} ; the highest copula simulation was 8.1×10^{-2} . However, such cases were rare ($<1\%$ of the cases).

The preceding comparison of risk predictions confirms that lognormal distributions are a poor choice for representing extreme values, as noted earlier (Su et al., 2012). It also highlights important differences between predictions using lognormal distributions and copulas. First, copulas can use any marginal distribution for mixture components, e.g., our simulations used the best-fit marginal distribution (both type and parameters) for each VOC. While this increases flexibility, copula simulations will propagate any mismatches in the marginal distribution, which may explain the underprediction at the higher risk levels.

Second, copulas permit asymmetric dependency structures that can emphasize extreme values or other portions of the distribution that display “local” dependencies, e.g., mixture B1 fit the Gumbel copula which emphasizes upper tail dependencies. Lastly, copulas performed better than multivariate lognormal models in all cases. Despite their power and flexibility, copula predictions sometimes diverged from the very highest observed data, e.g., above the 95th percentile.

3.6. Strengths and limitations

This is the first study to estimate dependency structures of personal exposures to multivariate VOC mixtures using copulas, a powerful technique that is unrestricted with respect to the marginal distributions of the underlying mixture components. Since VOC exposures are right-skewed even after log-transformation, traditional methods do not properly capture the tail behavior of the distributions. Copulas can improve the precision of exposure estimates, and decrease the bias of risk estimates. Like the cumulative cancer risks predicted in this study, exposures to VOC mixtures should be modeled appropriately to obtain accurate estimates of cumulative risk. Another potential application of copulas is to predict the population attributable fraction (PAF), which quantifies the contribution of various risk factors to a disease, i.e., the number of cases that would not occur if the risk factor did not exist (WHO, 2013). In this case, the proportion of population exceeding certain exposure or risk levels could be estimated to obtain the PAF.

Using RIOPA data, we identified and modeled two sets of VOC mixtures: those based on correlative measures using PMF analyses, and those using toxicological mode-of-action. In the former set, the RIOPA data revealed four mixtures with sources that were easily identified. Generally, these mixtures are considered to be “generated” or “intentional” mixtures. The second set of mixtures was associated with high lifetime cancer risks, at least for the more exposed individuals. We also identified variables associated with high exposures using logistic regression models that do not require normality of the response variables, thus handling the right-skewed distributions typically encountered.

The study has several limitations. First, to avoid issues associated with repeated measurements, only data from the first visit in RIOPA were used, which decreased the sample size and did not permit the analysis of possible seasonal effects. Second, because PMF does not indicate the optimal number of factors, there is some arbitrariness in this analysis. However, the VOC components in each factor were quite consistent, and the factors were similar to those seen elsewhere. Third, only two families of copulas (elliptical and Archimedean) were tested due to the limitations of the software for copula simulations. While the elliptical and Archimedean copulas are best known and most commonly used, other copula families, e.g., extreme-value, meta-elliptical and vine copulas, recently have been applied in fields such as finance and hydrology (Acar et al., 2012; Genest et al., 2007; Ghorbel and Trabelsi, 2009). Fourth, the RIOPA dataset has limitations in that only 18 VOCs were measured and MDLs for several compounds were higher than desirable. Low detection frequencies can affect results of PMF, copula and risk evaluations. Fifth, data quality should be considered in analysis and modeling to help ensure that results are reasonable and robust (Kirchner, 2006; Wang et al., 2009). While PMF analyses

incorporated uncertainty, distribution and copula selection and fitting assumed that the measurements were error-free. We note that exposure measurements can involve many types of errors, and both the lowest and highest measurements may be especially prone to errors. Sixth, the RIOPA sample was not population-based, and thus results may not be generalizable to the entire population as a whole. Finally, the RIOPA dataset is now 13 years old, and changes in product formulation and other factors may have altered both the concentration and composition of some VOC exposures.

4. Conclusion

Personal VOC exposures in the RIOPA study were grouped into six mixtures based on the potential emission sources and toxicological effects. Identified VOC emission sources included gasoline vapor (mixture A1), vehicle exhaust (mixture A2), moth repellents, chlorinated solvents and water disinfection by-products (mixture A3), and cleaning products and odorants (mixture A4). These four mixtures were affected by city, ethnicity and air exchange rates. The influence of environmental factors and personal activities was also shown for certain mixtures, e.g., mixture A1 was associated with attached garages and self-service pumping gas. Three additional VOC mixtures were identified by mode of action, including liver and renal tumors (mixtures B1 and A3/B3), and hematopoietic cancers (mixture B2).

Copulas were demonstrated to describe dependencies between mixture components with a high degree of accuracy and flexibility. Several types of copulas are needed. Dependency structures of four mixtures in RIOPA were best described by the t copula, while two other mixtures best fitted Gumbel copulas, which better capture dependency structures of distributions containing extreme values. In all cases, the copulas clearly provided better fits than multivariate lognormal distributions. Copulas can provide accurate estimates and simulations of joint distributions of pollutants across the full range of concentrations, and they faithfully represent the correlation in the tails of the distributions. Thus, copulas may be the method of choice for estimating cumulative risks of exposure to mixtures, particularly for the highest exposures or extreme events, cases that poorly fit lognormal distributions and that represent the greatest risks.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research described in this article was conducted under contract to the Health Effects Institute (HEI), an organization jointly funded by the United States Environmental Protection Agency (EPA) (Assistance Award No. R-82811201) and certain motor vehicle and engine manufacturers. The contents of this article do not necessarily reflect the views of HEI, or its sponsors, nor do they necessarily reflect the views and policies of the EPA or motor vehicle and engine manufacturers. We appreciate the input of the HEI reviewers as well as the Environment International reviewers.

References

- Acar EF, Genest C, Nešlehová J. Beyond simplified pair-copula constructions. *J Multivar Anal.* 2012; 110:74–90.
- ACGIH. TLVs and BEIs. American Conference of Governmental Industrial Hygienists; Cincinnati: 2012.
- Anderson MJ, Miller SL, Milford JB. Source apportionment of exposure to toxic volatile organic compounds using positive matrix factorization. *J Expo Anal Environ Epidemiol.* 2001; 11:295–307. [PubMed: 11571609]
- ATSDR. Toxicological profile for chloroform. Agency for Toxic Substances and Disease Registry; Atlanta, GA: 1997a. Available from <http://www.atsdr.cdc.gov/ToxProfiles/tp.asp?id=53&tid=16>
- ATSDR. Toxicological profile for tetrachloroethylene. Agency for Toxic Substances and Disease Registry; Atlanta, GA: 1997b. Available from <http://www.atsdr.cdc.gov/ToxProfiles/tp.asp?id=265&tid=48>
- ATSDR. Toxicological profile for trichloroethylene. Agency for Toxic Substances and Disease Registry; Atlanta, GA: 1997c. Available from <http://www.atsdr.cdc.gov/toxprofiles/tp.asp?id=173&tid=30>
- ATSDR. Guidance manual for the assessment of joint toxic action of chemical mixtures. Agency for Toxic Substances and Disease Registry; Atlanta, GA: 2004.
- ATSDR. Toxicological profile for carbon tetrachloride. Agency for Toxic Substances and Disease Registry; Atlanta, GA: 2005. Available from <http://www.atsdr.cdc.gov/toxprofiles/tp30.pdf>
- ATSDR. Toxicological profile for dichlorobenzenes. Agency for Toxic Substances and Disease Registry; Atlanta, GA: 2006. Available from <http://www.atsdr.cdc.gov/toxprofiles/tp.asp?id=704&tid=126>
- ATSDR. Toxicological profile for xylene. Agency for Toxic Substances and Disease Registry; Atlanta, GA: 2007. Available from <http://www.atsdr.cdc.gov/toxprofiles/tp71.pdf>
- ATSDR. Toxicological profile for ethylbenzene. Agency for Toxic Substances and Disease Registry; Atlanta, GA: 2010a. Available from <http://www.atsdr.cdc.gov/toxprofiles/tp110.pdf>
- ATSDR. Toxicological profile for styrene. Agency for Toxic Substances and Disease Registry; Atlanta, GA: 2010b. Available from <http://www.atsdr.cdc.gov/PHS/PHS.asp?id=419&tid=74>
- Bliss C. The toxicity of poisons applied jointly. *Ann Appl Biol.* 1939; 26:585–615.
- Borgert CJ, Quill TF, McCarty LS, Mason AM. Can mode of action predict mixture toxicity for risk assessment? *Toxicol Appl Pharmacol.* 2004; 201:85–96. [PubMed: 15541748]
- Cherubini, U.; Luciano, E.; Vecchiato, W. Copula methods in finance. John Wiley and Sons; New York, NY: 2004.
- Conolly RB. Biologically motivated quantitative models and the mixture toxicity problem. *Toxicol Sci.* 2001; 63:1–2. [PubMed: 11509736]
- Dennison JE, Andersen ME, Dobrev ID, Mumtaz MM, Yang RSH. PBPK modeling of complex hydrocarbon mixtures: gasoline. *Environ Toxicol Pharmacol.* 2004; 16:107–19. [PubMed: 21782697]
- Feron V, Groten J. Toxicological evaluation of chemical mixtures. *Food Chem Toxicol.* 2002; 40:825–39. [PubMed: 11983277]
- Feron V, Groten J, van Bladeren P. Exposure of humans to complex chemical mixtures: hazard identification and risk assessment. *Arch Toxicol Suppl.* 1998; 20:363–73. [PubMed: 9442308]
- Frees E, Valdez E. Understanding relationships using copulas. *North Am Actuar J.* 1998; 2:1–25.
- Genest C, Favre A. Everything you always wanted to know about copula modelling but were afraid to ask. *J Hydrol Eng.* 2007; 12:347–68.
- Genest C, Favre AC, Béliveau J, Jacques C. Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data. *Water Resour Res.* 2007; 43:W09401.
- Ghorbel A, Trabelsi A. Measure of financial risk using conditional extreme value copulas with EVT margins. *J Risk.* 2009; 11:51–85.

- Haddad S, Béliveau M, Tardif R, Krishnan K. A PBPK modeling-based approach to account for interactions in the health risk assessment of chemical mixtures. *Toxicol Sci.* 2001; 63:125–31. [PubMed: 11509752]
- IARC. Some naturally occurring substances: food items and constituents, heterocyclic aromatic amines and mycotoxins. World Health Organization, International Agency for Research on Cancer; Lyon, France: 1993.
- IARC. Agents classified by the IARC monographs. World Health Organization, International Agency for Research on Cancer; Lyon, France: 2012. Available from <http://monographs.iarc.fr/ENG/Classification/index.php>
- Jean-Frédéric, J.; Gaël, R.; Thierry, R. Financial applications of copula functions. John Wiley and Sons; Par Giorgio Szego: 2004.
- Jia C, D'Souza J, Batterman S. Distributions of personal VOC exposures: a population-based analysis. *Environ Int.* 2008; 34:922–31. [PubMed: 18378311]
- Jia C, Batterman S, D'Souza J. Copulas and other multivariate models of personal exposures to VOC mixtures. *Hum Ecol Risk Assess.* 2010; 16:873–900.
- Kirchner JW. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. *Water Resour Res.* 2006; 42:W03S04.
- Klepeis NE, Nelson WC, Ott WR, Robinson JP, Tsang AM, Switzer P, et al. The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. *J Expo Anal Environ Epidemiol.* 2001; 11:231–52. [PubMed: 11477521]
- Kwon J, Weisel CP, Turpin BJ, Zhang J, Korn LR, Morandi MT, et al. Source proximity and outdoor-residential VOC concentrations: results from the RIOPA study. *Environ Sci Technol.* 2006; 40:4074–82. [PubMed: 16856719]
- Paatero P, Tapper U. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics.* 1994; 5:111–26.
- Sax S, Bennett D, Chillrud S, Ross J, Kinney P, Spengler J. A cancer risk assessment of inner-city teenagers living in New York City and Los Angeles. *Environ Health Perspect.* 2006; 114:1558–66. [PubMed: 17035143]
- Schmidt, T. Coping with copulas. In: Rank, J., editor. *Copulas from theory to applications in finance.* Risk Books; London, UK: 2006.
- Sklar A. Fonctions de répartition à n dimensions et leurs marges. *Publ Inst Stat Univ Paris.* 1959; 8:229–31.
- Staudt A. Tail risk, systemic risk and copulas. *Casualty Actuar Soc E-Forum.* 2010; 2
- Su F-C, Jia C, Batterman S. Extreme value analyses of VOC exposures and risks: a comparison of RIOPA and NHANES datasets. *Atmos Environ.* 2012; 62:97–106.
- Trivedi, P.; Zimmer, D. *Copula modeling: an introduction for practitioners.* World Scientific Publishing; Hackensack, NJ: 2007.
- US EPA. Guidelines for the health risk assessment of chemical mixtures. US Environmental Protection Agency; Washington, DC: 1986. Available from <http://cfpub.epa.gov/ncea/cfm/recorddisplay.cfm?deid=22567#Download>
- US EPA. Provisional guidance for quantitative risk assessment of polycyclic aromatic hydrocarbons. US Environmental Protection Agency; Washington, DC: 1993. Available from <http://cfpub.epa.gov/ncea/cfm/recorddisplay.cfm?deid=49732>
- US EPA. Supplementary guidance for conducting health risk assessment of chemical mixtures. US Environmental Protection Agency; Washington, DC: 2000a. Available from <http://cfpub.epa.gov/ncea/cfm/recorddisplay.cfm?deid=20533>
- US EPA. Supplementary guidance for conducting health risk assessment of chemical mixtures. US Environmental Protection Agency; Washington, DC: 2000b. Available from <http://cfpub.epa.gov/ncea/cfm/recorddisplay.cfm?deid=20533>
- US EPA. Framework for cumulative risk assessment. US Environmental Protection Agency; Washington, DC: 2003. Available from http://www.epa.gov/raf/publications/pdfs/frmwrk_cum_risk_assmnt.pdf
- US EPA. Mobile source air toxics — regulations. US Environmental Protection Agency; Washington, DC: 2007. Available from <http://www.epa.gov/oms/toxics-regs.htm>

- US EPA. Regulatory determinations support document for selected contaminants from the second drinking water contaminant candidate list (CCL 2). US Environmental Protection Agency; Washington, DC: 2008. Available from http://www.epa.gov/ogwdw/ccl/pdfs/reg_determine2/report_ccl2-reg2_supportdocument_full.pdf
- US EPA. Integrated Risk Information System (IRIS). US Environmental Protection Agency; Washington, DC: 2012a. Available from <http://www.epa.gov/IRIS/index.html>
- US EPA. An introduction to indoor air quality: Volatile organic compounds (VOCs). US Environmental Protection Agency; Washington, DC: 2012b. Available from <http://www.epa.gov/iaq/voc.html>
- US EPA. MTBE — recommendations and actions. US Environmental Protection Agency; Washington, DC: 2012c. Available from <http://www.epa.gov/mtbe/action.htm>
- US EPA. National Ambient Air Quality Standards (NAAQS). 2012. Available from <http://www.epa.gov/air/criteriahtml>
- US EPA. Basic information about disinfection byproducts in drinking water: total trihalomethanes, haloacetic acids, bromate, and chlorite. 2013. Available from <http://water.epa.gov/drink/contaminants/basicinformation/disinfectionbyproducts.cfm>
- USEPA. EPA Positive Matrix Factorization (PMF) 3.0 fundamentals & user guide. 2008. Available from http://www.epa.gov/heads/products/pmf/EPA%20PMF%2030%20User%20Guide%20v16_092208_final.pdf
- Wang D, Singh VP, Zhu Y-s, Wu J-c. Stochastic observation error and uncertainty in water quality evaluation. *Adv Water Resour.* 2009; 32:1526–34.
- Wang Y, Ma H, Sheng D, Wang D. Assessing the interactions between chlorophyll a and environmental variables using copula method. *J Hydrol Eng.* 2012; 17:495–506.
- Weisel CP, Zhang J, Turpin BJ, Morandi MT, Colome S, Stock TH, et al. Relationship of Indoor, Outdoor and Personal Air (RIOPA) study: study design, methods and quality assurance/control results. *J Expo Anal Environ Epidemiol.* 2005a; 15:123–37. [PubMed: 15213705]
- Weisel CP, Zhang J, Turpin BJ, Morandi MT, Colome S, Stock TH, et al. Relationships of Indoor, Outdoor, and Personal Air (RIOPA). Part I. Collection methods and descriptive analyses. *Health Eff Inst Res Rep.* 2005b:1–127.
- WHO. Metrics: Population Attributable Fraction (PAF). UN World Health Organization; Geneva, Switzerland: 2013. Available from http://www.who.int/healthinfo/global_burden_disease/metrics_paf/en/index.html

Fig. 1.

Factor profiles from PMF analyses for personal exposure measurements of VOCs. Red (inset) boxes indicate percentage of mass of each species apportioned to the factor; (larger) blue bars indicate concentrations of species.

Fig. 2.

Cumulative probability plots of cancer risks for VOC mixtures using observed data, copula and multivariate lognormal simulations in the RIOPA study. The y-axis scale emphasizes differences at upper percentiles.

Table 1

Sources and apportionments of mixtures of VOCs derived using PMF.

Mixture ID	Suggested source categories	VOC components	Fraction of TVOC	
			%	$\mu\text{g m}^{-3}$
A1	Gasoline	Benzene and MTBE	20.5	19.9
A2	Vehicle exhaust	Toluene, ethylbenzene, xylenes, and styrene	20.9	20.3
A3	Moth repellents, chlorinated solvents and disinfection by-products	1,4-DCB, TCE, PERC, chloroform, and CTC	16.3	15.9
A4	Cleaning products and odorants	d-Limonene, α -pinene, and β -pinene	42.3	41.1

Mixture ID: A indicates mixtures identified by PMF.

Table 2

Observed and estimated probability of high concentration mixtures.

Mixture ID	VOCs	Copula	Percentile	Probability ^a		
				Observed data (n = 299)	Uncorrelated	Copula simulations (n = 1000)
A1	Benzene and MTBE	Gumbel	50th	0.3545	0.2500	0.3470
			75th	0.1371	0.0625	0.1550
			90th	0.0502	0.0100	0.0510
			95th	0.0201	0.0025	0.0250
A2	Toluene, ethylbenzene, xylenes, and styrene	t	50th	NC	0.0625	0.1950
			75th	0.0635	0.0039	0.0500
			90th	0.0134	0.0001	0.0110
			95th	0.0033	0	0.0040
A3, B3	1,4-DCB, TCE, PERC, chloroform, and CTC	t	50th	NC	0.0313	0.0820
			75th	0.0067	0.0010	0.0040
			90th	0.0033	0	0
			95th	0	0	0
A4	d-Limonene, α -pinene, and β -pinene	t	50th	0.3244	0.1250	0.2070
			75th	0.1171	0.0156	0.0480
			90th	0.0234	0.0010	0.0060
			95th	0.0100	0.0001	0.0030
B1	Ethylbenzene and MTBE	Gumbel	50th	0.3478	0.0625	0.3490
			75th	0.1438	0.0039	0.1430
			90th	0.0435	0.0001	0.0510
			95th	0.0234	0.0000	0.0240
B2	Benzene, MTBE, 1,4-DCB, TCE, and PERC	t	50th	NC	0.0313	0.0630
			75th	0.0067	0.0010	0.0060
			90th	0.0033	0	0
			95th	0	0	0

Mixture ID: A indicates mixtures identified by PMF; B indicates mixtures identified by toxicological mode of action.

NC, not calculated as styrene and TCE had detection frequencies <50%.

^aProbabilities of all components in the mixture exceeding 50th, 75th, 90th and 95th percentiles.

Table 3

Median mixture fractions based on observed data (n = 299) and copula simulations (n = 1000).

Mixture ID	VOCs	Mixture fractions ^a for indicated percentile							
		Observed				Best-fit copula			
		50–75th	75–90th	90–95th	95–100th	50–75th	75–90th	90–95th	95–100th
A1	Benzene	0.222	0.150	0.169	0.099	0.179	0.177	0.137	0.173
	MTBE	0.778	0.850	0.831	0.901	0.821	0.823	0.863	0.827
A2	Toluene	0.578	0.555	0.571	0.484	0.557	0.572	0.533	0.547
	Ethylbenzene	0.072	0.071	0.085	0.083	0.073	0.072	0.080	0.074
	Xylenes	0.300	0.316	0.328	0.368	0.303	0.280	0.298	0.291
	Styrene	0.024	0.020	0.019	0.012	0.038	0.040	0.037	0.038
	1,4-DCB	0.333	0.842	0.972	0.993	0.447	0.786	0.968	0.994
A3, B3	TCE	0.026	0.009	0.001	0.000	0.031	0.010	0.002	0.000
	PERC	0.165	0.032	0.005	0.001	0.128	0.031	0.009	0.001
	Chloroform	0.180	0.053	0.015	0.003	0.134	0.052	0.013	0.001
	CTC	0.065	0.023	0.005	0.001	0.069	0.024	0.006	0.001
	d-Limonene	0.667	0.661	0.754	0.765	0.720	0.751	0.825	0.850
A4	α-Pinene	0.204	0.149	0.100	0.080	0.176	0.127	0.102	0.041
	β-Pinene	0.078	0.099	0.143	0.120	0.061	0.055	0.026	0.029
	Ethylbenzene	0.156	0.125	0.106	0.062	0.154	0.117	0.106	0.083
B1	MTBE	0.844	0.875	0.894	0.938	0.846	0.883	0.894	0.917
B2	Benzene	0.118	0.062	0.019	0.004	0.093	0.068	0.022	0.004
	MTBE	0.606	0.347	0.054	0.009	0.552	0.515	0.159	0.023
	1,4-DCB	0.134	0.411	0.857	0.982	0.127	0.170	0.484	0.943
	TCE	0.010	0.005	0.001	0.000	0.009	0.005	0.003	0.001
	PERC	0.054	0.019	0.004	0.001	0.031	0.016	0.012	0.001

Mixture ID: A indicates mixtures identified by PMF; B indicates mixtures identified by toxicological mode of action.

Dominant mixture fraction shown in bold.

Copula simulations use fitted marginal distributions shown in Supplemental Table S3, and best-fit copula type in Supplemental Table S4.

^aMedian fractions. They may not sum to 1.

Table 4

Percentage of individuals exceeding individual lifetime cancer risk thresholds for VOC mixtures: comparison of observed data, simulations using copulas, and simulations using multivariate lognormal distributions.

Mixture ID	VOC	Type	Percentage exceeding indicated cancer risks				
			1×10^{-6}	1×10^{-5}	1×10^{-4}	1×10^{-3}	1×10^{-2}
B1	Ethylbenzene and MTBE	Observed data	100.0	25.4	1.0	0.0	0.0
		Copula simulations	97.5	27.1	0.6	0.0	0.0
		Lognormal simulations	96.9	32.0	0.0	0.0	0.0
B2	Benzene, MTBE, 1,4-DCB, TCE and PERC	Observed data	100.0	100.0	34.8	9.7	3.0
		Copula simulations	100.0	99.5	35.9	6.6	1.6
		Lognormal simulations	100.0	99.2	40.1	5.6	0.7
B3	1,4-DCB, TCE, PERC, chloroform and CTC	Observed data	100.0	100.0	44.5	11.0	3.3
		Copula simulations	100.0	99.8	44.8	9.5	1.9
		Lognormal simulations	100.0	99.7	53.6	6.7	0.2

Mixture ID: B indicates mixtures identified by toxicological mode of action.