Atmospheric Environment 77 (2013) 464–473

Contents lists available at SciVerse ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Addressing extrema and censoring in pollutant and exposure data using mixture of normal distributions☆

Shi Li [a], Stuart Batterman [b,*], Feng-Chiao Su [b], Bhramar Mukherjee [a]

[a] Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, SPH II, Ann Arbor, MI 48109, USA
[b] Department of Environmental Health Sciences, School of Public Health, University of Michigan, 1420 Washington Heights, SPH II, Ann Arbor, MI 48109, USA

HIGHLIGHTS

- Fitting distributions of volatile organic compound concentrations.
- Finite mixture of normals and Dirichlet process mixture of normals.
- Superior performance compared to the traditional single normal distribution.
- Robustness and ability to characterize uncertainty for model parameters.
- Implemented via Relationship between Indoor, Outdoor and Personal Air study.

ARTICLE INFO

Article history:
Received 20 November 2012
Received in revised form
30 April 2013
Accepted 3 May 2013

Keywords:
Air pollution
Density estimation
Dirichlet process mixture
Limits of detection
Mixture of normal
Volatile organic compounds

ABSTRACT

Background: Volatile organic compounds (VOC), which include many hazardous chemicals, have been used extensively in personal, commercial and industrial products. Due to the variation in source emissions, differences in the settings and environmental conditions where exposures occur, and measurement issues, distributions of VOC concentrations can have multiple modes, heavy tails, and significant portions of data below the method detection limit (MDL). These issues challenge standard parametric distribution models needed to estimate the exposures, even after log transformation of the data.
Methods: This paper considers mixture of distributions that can be directly applied to concentration and exposure data. Two types of mixture distributions are considered: the traditional finite mixture of normal distributions, and a semi-parametric Dirichlet process mixture (DPM) of normal distributions. Both methods are implemented for a sample data set obtained from the Relationship between Indoor, Outdoor and Personal Air (RIOPA) study. Performance is assessed based on goodness-of-fit criteria that compare the closeness of the density estimates with the empirical density based on data. The goodness-of-fit for the proposed density estimation methods are evaluated by a comprehensive simulation study.
Results: The finite mixture of normals and DPM of normals have superior performance when compared to the single normal distribution fitted to log-transformed exposure data. The advantages of using these mixture distributions are more pronounced when exposure data have heavy tails or a large fraction of data below the MDL. Distributions from the DPM provided slightly better fits than the finite mixture of normals. Additionally, the DPM method avoids certain convergence issues associated with the finite mixture of normals, and adaptively selects the number of components.
Conclusions: Compared to the finite mixture of normals, DPM of normals has advantages by characterizing uncertainty around the number of components, and by providing a formal assessment of

Abbreviations: VOC, volatile organic compounds; MDL, method detection limit; DPM, Dirichlet process mixture; RIOPA study, Relationship between Indoor, Outdoor and Personal Air study; GEV, generalized extreme value; EM, expectation maximization; MLE, maximum likelihood estimation; AIC, Akaike information criterion; BIC, Bayesian information criterion; CDF, cumulative distribution function; MSE, mean squared error; MAE, mean absolute error; NHANES, National Health and Nutrition Examination Survey.

* Corresponding author. Department of Environmental Health Sciences, School of Public Health, University of Michigan, Room 6075, SPH2, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA. Tel.: +1 734 763 2417; fax: +1 734 763 8095.
E-mail address: stuartb@umich.edu (S. Batterman).

1352-2310/$ – see front matter
http://dx.doi.org/10.1016/j.atmosenv.2013.05.004

uncertainty for all model parameters through the posterior distribution. The method adapts to a spectrum of departures from standard model assumptions and provides robust estimates of the exposure density even under censoring due to MDL.

1. Introduction

Volatile organic compounds (VOCs) have been used extensively in personal, commercial and industrial products (MDE, 2010; Ling et al., 2011; Weschler, 2011; USEPA, 2012b), and these chemicals are widely found in air in indoor, outdoor and occupational settings. Many VOCs are hazardous, and exposure through inhalation has been associated with a variety of acute and chronic health effects, such as respiratory disease and cancer (Kim and Bernstein, 2009; USEPA, 2012a,b). While concentrations of VOCs in environmental settings are generally much lower than those in occupational settings (Rappaport and Kupper, 2004), moderate and sometimes high concentrations and exposures can be encountered among the general population during certain activities, such as filling vehicles with gasoline and home renovations, in hobbies such as furniture restoration, small engine repair and gun cleaning, and using cleaners, pesticides, pest repellants and air fresheners in poorly ventilated spaces (Batterman et al., 2006; Jia et al., 2008a; D'Souza et al., 2009; Jia and Batterman, 2010; USEPA, 2012b).

The high concentrations found for a portion of the population, along with the much lower concentrations for the bulk of the population, typically results in highly right skewed concentration distributions (Jia et al., 2008b). Extreme value theory and other techniques can model the upper percentiles of VOC concentration distributions, and generalized extreme value (GEV) distributions have been shown to fit VOC data much more closely than lognormal or other types of distributions (Jia et al., 2008b; Batterman et al., 2011; Su et al., 2012). Most data sets also contain many low observations, often including measurements that fall below the method detection limit (MDL). These "non-detects," which represent left-censored data, can be treated by substitution, single or multiple imputation, regression on order statistics (modeling using probability plots of known distributions to estimate summary statistics), and laboratory-generated data (using the original data without replacement) (Antweiler and Taylor, 2008). The extent of data below MDLs can significantly affect the quality of the results (Lubin et al., 2004; Antweiler and Taylor, 2008). The statistical issues associated with the analysis of data with MDL issues are well-known (Taylor et al., 2001; Krishnamoorthy et al., 2009).

Due to the variation in source emissions, differences in the settings and environmental factors where exposures occur, and the measurement issues just noted, distributions of VOC concentrations can have multiple modes, heavy tails, and significant portions of data falling below the MDL that are replaced by a single value. These issues, which can be encountered in exposure and as well as other types of data sets, challenge standard parametric distribution models. While GEV distributions can fit the upper portions of distributions, they do not represent the full distribution of the data. Information on the full distributions of exposure levels is needed to establish exposure/risk guidelines, to estimate health risks and uncertainty estimates across a population (Su et al., 2012), and to facilitate probabilistic analyses (Hammonds et al., 1994).

Mixture distributions, which extend parametric families of distributions to fit datasets that are not adequately fit by a single common distribution, provide a flexible and powerful approach of representing the distribution of a random variable (Titterington et al., 1985; McLachlan and Basford, 1988; McLachlan and Peel, 2000). As examples, a finite mixture of normals applies a set of 'mixing weights' to a specified and finite number of component distributions, while a nonparametric Dirichlet process mixture (DPM) of normals relaxes the need to pre-specify the number of component distributions and is potentially advantageous in terms of handling smoothing, modality and uncertainty (Escobar, 1994; Mueller and Quintana, 2004). Mixture of normals have been extensively used in a variety of important and practical situations, although environmental applications have been very limited (Burmaster and Wilson, 2000; Razzaghi and Kodell, 2000; Taylor et al., 2001; Chu et al., 2005).

This paper evaluates the applicability of mixture of normal distribution method to environmental data, specifically, air pollution concentration and exposure data. Both the traditional finite mixture of normal and the nonparametric DPM of normals are evaluated using a VOC exposure dataset that includes seasonal measurements for approximately 300 individuals, which was collected as part of the Relationship between Indoor, Outdoor and Personal Air (RIOPA) study. Goodness-of-fit for the density estimation methods are evaluated by a comprehensive simulation study.

2. Materials and methods

2.1. VOC measurements

The RIOPA study was designed to evaluate contributions of outdoor and indoor sources to personal exposures of air pollutants, including VOCs and $PM_{2.5}$, among residents of three cities (Elizabeth, NJ, Houston, TX and Los Angeles, CA) selected to reflect potential differences in emissions and other factors likely to influence exposures (Weisel et al., 2005a). Sampling was conducted in two seasons for approximately 100 adults (and a smaller number of children) in each city from summer 1999 through spring 2001. Indoor, outdoor and personal (worn by participants) measurements were obtained using passive samplers for 48 h periods, and 18 VOCs were measured using gas chromatography and mass spectrometry. Analytical work was performed by two laboratories. The RIOPA study represents one of the larger VOC studies in the USA that collected personal samples, which are generally considered to provide exposure estimates that are more accurate than indoor or outdoor samples.

Three VOCs (chloroform, 1,4-dichlorobenzene (1,4-DCB) and styrene) were selected to evaluate mixture distributions. These VOCs differ in terms of their distributions, detection frequencies and other properties. Personal samples for adults were selected, primarily because the sample size for the adult cohort ($n = 544$ for each VOC) was largest, and because the personal samples should best reflect exposure. The two laboratories used to analyze samples had different MDLs. Since the use of two laboratories is somewhat unusual, all data under MDLs were replaced with a single value using $0.5 \times$ the higher MDL. Because the VOC data in RIOPA had many extreme values (Su et al., 2012), the density estimation methods were implemented using logarithms of the concentration value, as described next.

2.2. Finite mixture of normal distributions

Finite mixture distributions are commonly used to identify and model sub-populations within an overall population. Rather than