

Published in final edited form as:

*Ann Epidemiol*. 2017 February ; 27(2): 145–153.e1. doi:10.1016/j.annepidem.2016.11.016.

# Current Approaches Used in Epidemiologic Studies to Examine Short-term Multipollutant Air Pollution Exposures

**Angel D Davalos**[1], **Thomas J Luben**[2], **Amy H Herring**[1], and **Jason D Sacks**[2,*]

[1]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, USA

[2]National Center for Environmental Assessment, Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, North Carolina, USA

## Abstract

**Purpose—**Air pollution epidemiology traditionally focuses on the relationship between individual air pollutants and health outcomes (e.g., mortality). To account for potential copollutant confounding, individual pollutant associations are often estimated by adjusting or controlling for other pollutants in the mixture. Recently, the need to characterize the relationship between health outcomes and the larger multipollutant mixture has been emphasized in an attempt to better protect public health and inform more sustainable air quality management decisions.

**Methods—**New and innovative statistical methods to examine multipollutant exposures were identified through a broad literature search, with a specific focus on those statistical approaches currently used in epidemiologic studies of short-term exposures to criteria air pollutants (i.e., particulate matter, carbon monoxide, sulfur dioxide, nitrogen dioxide, and ozone).

**Results—**Five broad classes of statistical approaches were identified for examining associations between short-term multipollutant exposures and health outcomes, specifically Additive Main Effects, Effect Measure Modification, Unsupervised Dimension Reduction, Supervised Dimension Reduction, and Nonparametric methods. These approaches are characterized including advantages and limitations in different epidemiologic scenarios.

**Discussion—**By highlighting the characteristics of various studies in which multipollutant statistical methods have been employed, this review provides epidemiologists and biostatisticians with a resource to aid in the selection of the most optimal statistical method to use when examining multipollutant exposures.

[*]**Corresponding author:** National Center for Environmental Assessment, Office of Research and Development, U.S., Environmental Protection Agency, Mailcode B-243-01, RTP, NC 27711, Ph: (919) 541-9729, Fax: (919) 541-2985, sacks.jason@epa.gov.

**Keywords**

Air pollution health effects; joint effects; multipollutant; dimension reduction; nonparametric methods; interactions; differential effects

## Introduction

The results of epidemiologic studies that examine the association between individual air pollutants and health effects have contributed enormously to understanding how air pollution impacts health and the dramatic improvement in air quality that has occurred since the inception of the Clean Air Act. Although understanding the independent effects of exposure to a single pollutant is essential, scientists also recognize that under normal ambient conditions, humans are not exposed to individual pollutants in isolation, but to a complex mixture of air pollutants. Recent publications convey this point by calling for research aimed at understanding the health effects of multipollutant exposures (i.e., the joint effect of two or more pollutants on a health outcome) with the aim of developing a catalogue of statistical methods to support multipollutant analyses [1] that can inform the development of more sustainable air quality regulations [2, 3].

Traditionally, epidemiologists examine whether there is evidence of an independent association between an individual pollutant on a health outcome (e.g., mortality) by including two or more air pollutants in a regression model and estimating the association attributable to each individual air pollutant after accounting for (or adjusting for) other measured pollutants co-occurring in the ambient air mixture. However, these types of models can become highly unstable when incorporating two or more pollutants that are highly correlated [2].

To examine the relationship between multipollutant exposures and health, new and innovative statistical methods are being developed and applied in epidemiologic studies. The purpose of this review is to highlight the variety of statistical methods currently available to examine the relationship between short-term exposures (i.e., single- or multi-day lags up to one week) to multipollutant mixtures and health effects. A number of these methods, specifically receptor modeling, have been used extensively to try and identify health risks associated with components and sources of fine particulate matter ($PM_{2.5}$), itself a multipollutant mixture. The multipollutant nature of $PM_{2.5}$ highlights a difficulty encountered when evaluating the current literature base of epidemiologic studies: the limited number of studies that focus specifically on examining the combined effect of multipollutant exposures to more than one criteria air pollutant (i.e., PM, $O_3$, $NO_2$, $SO_2$, and CO) on health. As such, for the purposes of this review, we focus on epidemiologic studies of multipollutant exposures that conduct a simultaneous evaluation of at least two criteria air pollutants, not studies focusing only on $PM_{2.5}$. Overall, this review is not intended to be a systematic evaluation of all available multipollutant statistical methods intended for use in short-term exposure epidemiologic studies, but instead is meant to highlight the broad classes of statistical approaches available to epidemiologists and statisticians as they continue to design, conduct and interpret multipollutant air pollution studies.

## Methods

We conducted a broad literature search for studies including at least two criteria air pollutants (i.e., PM, $O_3$, $NO_X$, $SO_X$, CO). The broad literature search was a multistep process in which search strings were composed and then run through the PubMed and Web of Science® databases. The search strings used for each pollutant are provided in Supplemental Table S1.

To the references retrieved by the broad literature search, a machine learning algorithm was applied to segregate references into domains of epidemiologic or other (e.g., experimental) studies (see [4] for details). The algorithm, developed from a seed of known relevant references that focused on studies of air pollution and health, had recall greater than 90% but lower precision, meaning the bins contained some references not relevant for this review. As a result, a title screen was then performed to exclude non-relevant references that were identified by the machine learning algorithm. Finally, an abstract review was conducted to exclude any non-relevant references that were not identified during the title screen. If we could not conclusively determine whether inclusion criteria were met from reviewing an abstract, we reviewed the reference's methods section. In addition, papers were identified for inclusion in several ways: specialized searches on specific topics, review of tables of contents for journals in which relevant papers may be published, identification of relevant literature by expert scientists, and review of citations in included studies. This is not intended to be a systematic review of the literature, but rather a broad overview of statistical methods and the feasibility and utility of their use to identify the combined effect of air pollutants in epidemiologic studies.

## Results

Within this review, statistical methods are categorized according to the mixture effect assumptions (pollutant mixture relationship (PMR) specification) in the regression analysis. Based on the literature evaluated, five broad classes of statistical approaches were identified: additive main effects (AME), which are those methods that assume each pollutant within the mixture has an additive effect; effect measure modification (EMM), which are regression-based methods to examine whether the level of one or more pollutants modify the health effect associated with another pollutant or group of pollutants; unsupervised dimension reduction (UDR) that transform multiple pollutants into a different set of variables independently of a health outcome of interest; supervised dimension reduction (SDR) where mixture transformation is dependent on the health outcome; and nonparametric methods, which are highly flexible methods that relax parametric assumptions of the interactive pollutant effects. Here, we use the language "effect" to refer to a general parameter of interest; we do not intend for the word "effect" to imply a necessarily causal association between exposure and outcome. The following sections provide a more detailed discussion of each broad class of multipollutant approaches along with the specific methods currently available.

### Additive Main Effects (AME)

AME approaches, which consist of multipollutant or joint effects models with no multiplicative pollutant interaction terms, may be used to estimate joint associations of multiple air pollutants. The statistical methods within this category have appeal due to the intuitive construction of regression models, allowing for the straightforward inclusion of terms to examine the potential immediate, delayed, or prolonged association between air pollution and health through either single or multi-day (e.g., distributed) lags.

Given the relative ease of construction and interpretability of AME models, surprisingly few air pollution studies utilize AME models to examine the combined association between multiple pollutants and health. Gold et al. [5] were one of the first to consider examining the combined effect of two pollutants ($PM_{2.5}$ and $O_3$) in a study of air pollution and lung function. They assessed pollutant specific differences in the temporal relationship with lung function by including differing lag structures for each pollutant. Unlike Gold et al. [5], Schildcrout et al. [6] included the same lag structure for each pollutant (linear 3-day moving average) to examine the combined effect of a simultaneous increase in air pollutant concentrations on asthma exacerbations. The authors also decomposed the effect of ambient concentrations of pairs of pollutants (e.g., CO + $NO_2$, CO + $PM_{10}$) into a within and between subject component. Decomposing effects is a useful tool for revealing intra- and inter-individual information, and may be used for any of the other methods described in this paper. However, the interpretability of effects and the additional number of coefficients to estimate will depend on the method chosen. Instead of focusing on two pollutant joint effects models, Winquist et al. [7] examined several pollutant mixtures (ranging from two to five pollutants) selected to represent pollutants that commonly occur together in ambient air, or that might have common mechanisms leading to pediatric asthma emergency department (ED) visits. Collinearity was acknowledged as an issue in the pairwise CO + $NO_2$ model [6] and multipollutant models explored by Winquist et al [7]. The AME specification does not in itself address multicollinearity and requires effect estimation procedures that can handle correlated variables in order to stabilize estimate precision, otherwise, estimates may not be obtainable or yield unreliable results.

Hierarchical models with AME specification have been used to study joint air pollution effects to overcome some difficulties with collinearity. Hierarchical models impose a distribution on effects (i.e., regression coefficients), where the effects can be assumed decomposed by a common property and pollutant-specific error resulting in pollutant effects being ‘shrunken’ toward the effect of the common property with improved precision. When an AME specification is used within a hierarchical model, joint effects are immediately obtained upon completion of the estimation procedure without need to aggregate pollutant specific effects to obtain a joint effect. Suh et al. [8] demonstrated the use of such models by examining the joint impacts of 65 pollutants by nine chemical properties on the odds of daily cause-specific hospital admission through a two-stage hierarchical model (i.e., model is fit in a two-step procedure). These types of estimates are called shrinkage estimates and can be obtained via numerous methods [9, 10].

Penalized regression methods produce another class of shrinkage estimates whose use has been proposed with AME specification. These methods impose mathematical constraints on

associations that introduce bias into estimates, but improve precision when pollutants are highly correlated. Roberts and Martin [11] compared least absolute selection operator (LASSO), ridge regression, and non-penalized regression models with five pollutants and linear AME specification in examining the relationship between daily changes in air pollution and mortality. The main difference between LASSO and ridge regression is that LASSO can assign air pollutant effect estimates of exactly zero because specific pollutant coefficients can be eliminated during the modeling process. While this is an appealing feature of LASSO, ridge regression was recommended over LASSO when the focus of a study is to assess the overall mixture effect on a health outcome rather than individual pollutant effects. As a result, further developments in penalized methods may be useful; see Chadeau-Hyam et al. [12] for a brief overview of recent advancements. One such example is elastic net [13], which offers improved effect estimates of highly correlated variables over LASSO while preserving its variable selection capability by combining it with ridge regression. The performance of elastic net has been studied via simulations within contexts similar to those in air pollution epidemiology with moderate correlation between chemical mixtures and the relationship with term birth weight [14], and high correlation between environmental factors mimicking exposure in mothers during pregnancy [15], both under AME specification linear regression.

By assuming a main effects structure, an AME model may not be flexible enough to capture key features of the true relationship between a health outcome and air pollutant mixture when the potential association with one pollutant may depend on the level of another, which may be addressed by EMM approaches.

### Effect Measure Modification (EMM)

Studies that assessed associations through effect measure modification by a pollutant (e.g., $PM_{2.5}$ concentrations) or multipollutant joint effects models that include multiplicative interaction terms are defined here as EMM approaches. The rationale behind grouping these methods together is they have similar properties in their PMR specification and explore a multidimensional response surface without assuming pollutant effects are solely additive as is done with AME models.

Katsouyanni et al. [16] examined whether there was evidence of EMM for the $PM_{10}$ and $O_3$ mortality association. The authors examined whether the mortality association changed between the $25^{th}$ and $75^{th}$ percentile of the coefficient of variation for $NO_2$ and $O_3$, mean $SO_2$, and the ratio of mean $NO_2$ to $PM_{10}$. Carbajal-Arroyo et al. [17] used a similar approach in the examination of potential EMM of the relationship between $PM_{10}$ exposure and infant mortality by $O_3$ quartile concentrations. Few studies have considered a joint effects model with interaction terms. One such study, Winquist et al. [7] discussed previously, considered all pairwise-interaction pollutant terms as part of sensitivity analysis on their primary models where they compared joint effect estimates between their EMM-like multiplicative interaction term and AME models.

The studies detailed above showcase the rich toolkit EMM models provide in exploring the multidimensional response surface beyond AME under different settings. EMM approaches are feasible when assessing the mixture relationship between pairs of pollutants, while the

Winquist et al. [7] sensitivity analysis approach is better suited for three or more pollutants. Even though the focus of the Winquist et al. [7] study was on effect estimation, they supplemented joint effect estimate comparison with model building to determine whether interaction terms were important via significance testing. If model uncertainty is of concern, some model building procedures may be helpful, such as: variable selection algorithms (forward, backward, stepwise selection, LASSO) or Bayesian methods (e.g. see 18, 19). In particular, a method that may be well suited for air pollution epidemiology is LASSO for hierarchical interactions, since it only allows interactions in models if at least one of the main effect variables is marginally important [20]. Some automated variable selection methods do not take this into account and may yield nonsensical models in terms of interpretation (i.e., models that imply absence of one pollutant implies absence of another).

Given the recent push toward multipollutant analyses and the relatively unknown interactive behavior between pollutants, model building strategies can be useful tools to account for model uncertainty in moving one step forward from AME approaches to including multiplicative interaction terms. Thus, there is a need to compare the performance of different model building strategies to gain knowledge for the development of optimal strategies under varying circumstances. It was surprising to find only one air pollution simulation study evaluated the performance of multiple strategies. Sun et al. [21] compared two model/variable selection (Bayesian Model Averaging (BMA), LASSO) and two dimension reduction (Projection to Latent Structures also known as Partial Least Squares (PLS), Supervised Principal Components Analysis (SPCA)) methods, in a time-series framework to identify and estimate a true model from a set of pollutants, ranging from four to ten, with pairwise–interaction pollutant terms. Briefly, BMA may be defined in multiple ways, but the general idea is to apply a prior distribution on a set of candidate models where effect estimates are then defined as weighted model-specific posterior estimates by corresponding posterior model probabilities for all models [22, 23]. LASSO outperformed all other procedures when sample size (N = 400) and number of pollutants (four) was smaller. However, BMA performed better when sample size (N = 800) and number of pollutants (ten) was increased.

EMM models offer greater flexibility in approximating the data generating mechanism over AMEs with similar ease of construction and are analyzed through familiar methods. However, model uncertainty may present challenges when the number of pollutants and interaction terms is large atop estimation difficulties in the presence of collinearity. In preliminary and hypothesis generating work, it may useful to use alternative approaches that transform the PMR into a different set of (ideally) lesser correlated variables, termed dimension reduction methods.

## Unsupervised Dimension Reduction (UDR)

Unsupervised Dimension Reduction (UDR) approaches consist of methods that transform pollutant mixture concentrations into a smaller set of variables that are then used to represent exposures to various pollutant combinations or sources. These transformations depend upon intrinsic natural structures within the data, without regard to the health outcome(s) being evaluated, to create clusters, groups, or indices of air pollutant exposures. UDR approaches

may be used to simplify multipollutant exposures or address multicollinearity, and may be appealing because they can be compared across different outcomes. However, pollutant groupings from UDR approaches are often specific to the geographic area being studied. For example, principal components computed based on air quality data from the U.S. can differ from those based on a European country. UDR approaches are subdivided below into two methods to differentiate those that reveal structures within the data according to (1) Statistical/Mathematical or (2) Scientific criteria. Note, we do not consider variable selection/selection operators (e.g. backward, forward, stepwise variable selection) to be dimension reduction methods since variable selection/selection operators are tools used to reduce the number of terms of an already specified PMR. Dimension reduction, on the other hand, refers to a characterization of the PMR.

**Statistically/Mathematically-Based UDR Methods—**Statistically/mathematically-based UDR methods create factor profiles beforehand and use these as the PMR specification. Some commonly used methods include chemical mass balance (CMB), principal component analysis (PCA), factor analysis (FA), latent class analysis, positive matrix factorization, multilinear engine and the EPA UNMIX model. Investigators have previously found relative agreement across methods in studies that use similar underlying data [24, 25]. Thurston et al [26] state that FA methods, such as PCA, have an advantage over mass balance methods in that they can incorporate nontraditional aerosols, such as secondary aerosols, and non-PM tracers, such as gaseous pollutants. This has recently been reflected in a study conducted by Sacks et al. [27] that used PCA to identify source-based factors using a combination of PM components and gaseous pollutants.

Another approach to PMR specification is to create indicator variable profiles by clustering exposures with similar properties, then using these clusters in regression models. Zanobetti et al. [28] implemented an approach proposed by Austin et al. [29] where daily pollutant concentrations were grouped using k-means and subsequently used as EMM variables in exploring the association between $PM_{2.5}$ exposures and mortality. Pearce et al. [30] also clustered days, however, they used a Self-Organizing Maps (SOM) algorithm to create day types that were then used to assess the effects of the air pollution mixture on ED visits for pediatric asthma. SOM differs from the clustering methods in that it is a learning process that produces a one- or two-dimensional array by grouping input data through the iterated estimation of distinct profiles, with the goal of minimizing information lost via grouping and retaining power and precision for statistical analysis [30]. These clustering methods have shown they can be used in different ways according to the intention of the study. These methods can reveal interesting cluster patterns to generate further investigation. However, the information clusters reveal may be limited by using clusters in the regression analysis. For instance, when clusters are directly used in regression models, joint effects are only obtainable as a comparison among cluster patterns. Also, it is not straightforward to tease out the interactive nature of the pollutants.

**Scientifically-Based UDR Methods—**Alternatively, scientifically-based UDR methods, in which the grouping or clustering of pollutant mixtures might be defined by scientific rationale, have also been used. For example, Hong et al. [31] developed a combined index of

pollutants as the sum of mean scaled pollutant concentrations, to examine the dose-response relationship between short-term exposures and mortality in South Korea, with indices selected to represent real ambient exposures. In a slightly different approach, Pachon et al. [32] proposed a summary indicator variable and demonstrated its use in estimating the association between air pollution exposures and cardiovascular disease ED visits. The indicator was defined as a sum of weighted normalized pollutant concentrations with weights computed as mobile-source-to-total emissions from the National Emission Inventory. The indicator was intuitively constructed as the weighted sums corresponding to differing sources of exposure. Scientifically-based UDR methods, such as those detailed here, are especially important when biological mechanisms support the construction of mixture transformation. However, when there is not strong biological evidence of such relations, these methods are less defensible.

### Supervised Dimension Reduction (SDR)

SDR approaches estimate a mixture transformation/dimension reduction concurrently with the regression analysis or with respect to a health outcome of interest. The methods that encompass SDR require a general specification of the relationship between the pollutants and exposure with specific components to be estimated with respect to a health outcome. SDR methods are similar in idea to UDR methods with the exception that pollutant groupings are developed specifically for a health outcome. As a result, as mentioned for UDR methods, the pollutant groupings identified using SDR methods may also be specific to the geographic region(s) examined within the study.

The following supervised methods specify a weighted sum relationship between the pollutants and outcome with the weights representing proportions that sum to one. Pachon et al. [32] proposed an outcome- or heath-based indicator where weights are a priori specified on a range of values and corresponding indicator candidates are defined as weighted sums between pairs of standardized pollutant concentrations. Significance testing is performed separately on each candidate as a predictor in univariate regression, where the candidate with the minimum p-value is defined as the health based indicator. This indicator was developed as a sensitivity analysis on the health effect of their UDR emissions-based indicator since atmospheric mixtures may differ from emissions-based fractions due to meteorological conditions. In a different approach, Roberts and Martin [33] developed a model where weights of normalized pollutant concentrations are computed concurrently with the regression analysis as they are treated as parameters and estimated via optimization. A benefit of this method is that the weighted model is able to address the important question of whether there is a biologically relevant pollutant mixture that is related to the health outcome. Another similar method has recently been proposed for dealing with highly correlated data called weighted quantile sum (WQS) regression [see 34]. In contrast to the Roberts and Martin [33] model, WQS regression sums a weighted linear index of quantile pollutant concentration categories and estimates the weights through a bootstrap resampling procedure on a training data set (random subset of the data). For each bootstrap sample of the training dataset, the weighted linear index is included as a predictor in a regression model, with the weights estimated via maximum likelihood constrained to sum to one, and the corresponding regression coefficient tested for significance. The estimate for the weight

of each pollutant category is then defined as the average across all bootstrap estimates with a significant regression coefficient.

Another outcome dependent method, originally developed to address problems where predictors greatly exceed observations, is SPCA. Roberts and Martin [35] noted that SPCA is an improvement to PCA that can identify which subset of predictors is most highly associated in terms of magnitude to estimate standard deviation with the outcome. An added benefit is that it constructs a cross-validated best model that reveals important outcome-specific profiles on a subset of pollutants. The performance of SPCA is useful in variable selection when there is "a moderately strong exposure response" in comparison to other methods for constructing multipollutant models [21].

The use of latent variables in defining a transformation of pollutant mixtures allows incorporating some uncertainty about the mixtures into the model. Latent or "unobserved" variables may be associated in numerous ways with pollutant mixtures and their relation with a health outcome. For example, PLS iteratively creates linear combinations of latent variables that best describe the response and predictor variables jointly [36, 12]. PLS regression was utilized to study the chemical composition of $PM_{2.5}$ with respect to lung toxicity [37], but has not been widely employed in epidemiologic studies. In assessing the creation of pollutant profiles, PLS has been compared to BMA, SPCA, and LASSO, where it was observed to estimate interaction effects with little bias [21]. The interpretation of the latent variables in PLS is not straightforward due to its mathematical construction; however, they may be used in alternative ways, such as can be done in structural equation modeling for example, where latent variables may be assigned a priori meaning. Nikolov et al. [38] proposed the use of a Bayesian structural equation model in analyzing the association between sources of PM and a cardiovascular outcome in dogs. Their modeling framework includes the specification of a receptor submodel, which is assumed dependent on unobserved pollution source profiles, and a health submodel where the receptor submodel source contributions (latent variables) are specified as the predictors. The Bayesian nature of this method allows the incorporation of prior knowledge on sources and their contribution through the specification of their respective prior distribution parameters. This method requires the number of sources to be fixed a priori; however, Park et al. [39] extended it to the case where the number of sources is unknown by incorporating model uncertainty through BMA.

Overall, SDR methods provide appealing ways to overcome specific issues with analyzing pollutant health effects, but require similar considerations as UDR methods. An appealing feature to consider is that SDR methods create mixture transformations that optimize associations with outcome by being outcome specific. However, one must carefully consider difficulties with interpretation and the possibility of data feature loss with SDR, as with UDR methods. It may not be straightforward to quantify pollutant specific or joint effects, or tease out the nature of the interactive effects, but these methods are designed to maximize the strength of association between a health outcome and mixture transformation.

## Nonparametric Methods

Methods that use nonparametric techniques to summarize the PMR are termed Nonparametric methods. The methods described below can be thought of as those with a PMR specification that is empirically or data driven. Semi-parametric models or techniques are included in this section if the multipollutant mixture relationship is explored via nonparametric techniques. Similar to the UDR and SDR methods discussed previously, some of the nonparametric methods may identify pollutant groupings that are specific to the geographic region being examined in the study. In particular, methods that rely on automated significance testing may be especially susceptible to the pollutant mixture characteristics of the study region. Investigators should implement these methods with caution and balance decision making based on the strengths and limitations of the statistical tool and data, despite the lure of robust results with nonparametric methods.

Nonparametric data partitioning methods have been used sparingly in the analysis of air pollution health effects, but can provide a wealth of tools for discovering the complex relationship between air pollution and health effects. A common technique used is recursive partitioning where pollutant concentrations (or the data) are recursively split into mixtures containing observations with similar health outcomes [42]. One example is Classification and Regression Trees (CART) [40, 41] which requires the specification of regression models (i.e., linear for continuous, logistic for binary outcomes, etc) within each partition and statistical significance tests determine splits in the data. Gass et al. [42] proposed a modified CART method, where an initial subset of data is withheld to represent an *a priori* selected referent mixture of pollutants and the remaining multipollutant exposure mixtures are partitioned all while controlling for confounding. The CART method was used to study the relationship between air pollution exposures (daily average ambient concentrations of $O_3$, $NO_2$, and $PM_{2.5}$) and pediatric asthma ED visits. Conclusions differed when compared to the joint effects in an AME model [43]. The authors suggest that some differences may be attributable to a non-synergistic effect between $PM_{2.5}$ and $NO_2$ because each can be correlated with $PM_{2.5}$ components.

With a slightly different use for CART, Sun et al. [21] proposed it as an initial screening tool for reducing the number of terms to be included in a model where interactive effects may be further examined. In simulation studies, CART was found to be a beneficial pre-screening tool in terms of reducing model dimensions when the number of candidate variables is large. This approach highlights that when the PMR is not well understood or the number of pollutants is high, pre-screening tools may be useful when combined with EMM and model building, UDR, or SDR methods to refine understanding of the PMR to suit the study objective. For instance, LASSO for hierarchical interactions may be useful to supplement CART to refine multipollutant joint effect estimates. Other methods that might be extended in a similar manner could include combining multiple tree models such as Bayesian Additive Regression Trees [44].

An alternative to data partitioning tools are regression methods that attempt to smooth a response surface. Kernel Machine Regression (KMR) is one such method where a response is regressed on a weighted sum of measures between subject exposure mixtures. The function used to define the measure is called a kernel and its specification in turn defines the

properties and form of the response surface. Kernels may introduce specific parameters that may be tuned or estimated. The weights are treated as parameters in the model and estimated along with the other regression parameters. Bayesian KMR was used by Bobb et al. [45] to estimate the health effects of multipollutant mixtures with a focus on exploring the exposure-response surface. It was showcased in both epidemiologic and toxicological studies that examined the effect of metals mixtures (including Pb) on neurodevelopment and exposure to air pollution mixtures on hemodynamics, respectively. The interplay between statistical techniques and machine learning with respect to model and variable selection within the KMR framework, is an area of active research that is in its infancy [46]. Thus, an exciting feature of this method, due to its Bayesian nature, is simultaneous health effect estimation and variable selection which allows it to account for model uncertainty. Because of this, it was shown to outperform frequentist methods [46, 47] in approximating exposure-response relationships via simulations. Another KMR variable selection method that may be useful was illustrated by Liu et al. [46], where kernel machine AIC and BIC values were proposed as model selection criteria. The authors illustrated its use while examining the complex joint effect of multiple genes within a pathway in the analysis of microarray data, by implementing an all-possible-subset procedure on a set of cell growth genes and selecting the combination producing the smallest AIC and BIC values.

Another way of exploring interactive joint effects is to model the joint distribution between multiple pollutants and the health outcome of interest where the multipollutant components are modeled non-parametrically. Bayesian Profile Regression is an example of a semi-parametric method that has been used to study air pollution health effects [48], where the joint distribution is characterized by multipollutant profile assignment and health effect submodels. The multipollutant profile assignment submodel is assumed to be a Gaussian-Dirichlet Process Mixture where the profiles are jointly assumed to follow a multivariate normal distribution and its parameters (i.e., mean and covariance) are assumed to follow a Dirichlet process prior. The Gaussian-Dirichlet Process Mixture specification inherently imposes an unknown random distribution on the multipollutant profiles which renders the estimation of their joint distribution a nonparametric procedure. By the parameter prior assumption, profiles are effectively assigned to clusters where then the health effect submodel is defined as a random effects model with a profile cluster assignment random effect. Molitor et al. [48] used this method in analyzing census block group multipollutant profile exposures and their association with term low birth weight. The exposure cluster random effect formulation allows the estimation of profile mixture cluster effects while controlling for relevant fixed effects such as confounders. Note, the Gaussian assumption does not limit the form of the pollutants as the authors suggest the use of latent continuous variables for categorical pollutant measures [49, 50]. When pollutant measures are solely categorical, other random distributions have been proposed that can accommodate multivariate discrete data [51, 52, and 53 for sparse data].

Nonparametric methods are promising alternatives to more traditional approaches for exploring complex non-linear interactions. The relaxing of PMR assumptions allows these methods to pick up interesting patterns of the PMR-response relationship especially when non-linear patterns exist. These can be invaluable tools during exploratory analyses or hypothesis-generating exercises. With recent extensive efforts in developing these methods

(e.g., machine learning, nonparametric Bayes), they are becoming feasible tools to be used in assessing health effects as computational and model assessment difficulties are addressed. However, currently, readily available software that incorporates these methods is limited and the interpretation of results may be difficult. Specifically, it may be difficult to estimate effects attributable to specific mixture components.

## Discussion

The growing body of air pollution health effects literature was broadly reviewed to summarize the numerous statistical methods available to examine health effect associations due to short-term multipollutant exposures. These methods can be grouped into five broad categories of statistical approaches according to their PMR specification: Additive Main Effects (AME), Effect Measure Modification (EMM), Unsupervised Dimension Reduction (UDR), Supervised Dimension Reduction (SDR), and Nonparametric methods.

The limited knowledge on the health effects associated with multipollutant exposures supports the need for having different analytical tools for assessing this complex outcome-mixture relationship. All of the methods reviewed across the five classes of statistical approaches provide different, possibly complementary, pieces of valuable information in assessing the health effects attributed to multipollutant air pollution exposures (see Table 1). For instance, if relatively little is known about how a mixture relates to a health outcome then nonparametric methods may be important tools since their purpose is to explore the potentially complex interactive PMR by relaxing parametric assumptions. Alternatively, UDR approaches are appropriate when the goal is to discover profiles or indices present in the data to then assess their effect. SDR approaches aim to discover outcome-specific profiles or indices that may be most appropriate for identifying pathways to disease. EMM and AME are the most interpretable and easiest to implement approaches for estimating joint effects when the PMR specification is more or less known. Table 2, provides specific details on the modeling methods used in the studies detailed throughout this review (i.e., regression assumptions, estimation method, software availability). The scope of this review was limited to focus on five criteria pollutants, however it is worth noting that the specific methods within the five classes of statistical approaches described could be applied more broadly and can accommodate more than just the five criteria pollutants, such as air toxics.

Despite the method of analysis implemented, exposure misclassification is a potential source of bias that can greatly impact air pollution health effects studies, but was beyond the scope of this review. All of the individual methods presented were entirely dependent on the assumption that exposure classification was appropriate for the subject unit on which a health outcome was assessed. As such, there was no intention to describe or analyze the effect of differing forms of exposure assessment or misclassification. It is possible that the statistical methods presented are sensitive to the different forms of exposure assessment or degree of misclassification, and obscure results. Some have argued that nonparametric or “data-driven” methods are especially constrained by data limitations [54]. Thus, it is recommended that when deciding on a statistical method, one must consider the complexity, difficulty of interpretation, study intent, and computational cost of the proposed model.

Furthermore, the overall analysis must be a balancing of data limitations, adequate exposure assessment, the current body of knowledge of the disease, and confounding.

Moving forward, understanding the independent effects of exposure to a single pollutant is essential, but research aimed at understanding the health effects of multipollutant exposures is necessary to potentially develop more sustainable air quality regulations. Epidemiologists have a clear role to play in this process. For years, copollutant models have been commonly employed to examine the role of individual pollutants in the complex air pollution mixture, all the while being recognized as a limited tool. Now is the time to move beyond copollutant models and take advantage of the multipollutant statistical methods currently available, to better evaluate the health effects of air pollution. As a first step in this process, it is beneficial for epidemiologists to familiarize themselves with new multipollutant statistical methods; specifically understanding when the different methods are best employed, and the strengths and limitations of each method. This can be accomplished by improving communication and collaboration between epidemiologists and biostatisticians, preferably early in the scientific process, beginning with the study design and analysis plan, and following through to the analysis and interpretation of results. In choosing a statistical method, it will be important for epidemiologists and biostatisticians to carefully consider the focus of the research and the limitations in the data. Each multipollutant statistical method explores the response surface in different ways, anticipates data limitations in the form of multicollinearity in order to adjust for them accordingly, and will have different implications for the types of conclusions that can be drawn.

In the end, one of the most important considerations for epidemiologists and biostatisticians, regardless of the statistical method employed, is the translation and application of results for use in a policy context. Although informative, not all of the multipollutant methods currently available can easily be used to inform policy decisions. However, by employing a wide range of multipollutant statistical methods across an array of epidemiologic study designs, we will begin to accumulate the scientific base necessary in order to potentially develop more sustainable, multipollutant air quality regulations.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We would like to acknowledge Danielle Moore for her help in developing the terms used in the broad literature search, and Ellen Kirrane and Ana Rappold for comments on early drafts of the manuscript.

Funding

This work was supported by the National Institute of Environmental Health Sciences [R01 ES020619, T32 ES007018].

## List of Abbreviations

| | |
|---|---|
| **PMR** | pollutant mixture relationship |

| | |
|---|---|
| **AME** | additive main effects |
| **ED** | emergency department |
| **LASSO** | least absolute selection operator |
| **EMM** | effect measure modification |
| **BMA** | Bayesian model averaging |
| **SPCA** | supervised principal components analysis |
| **UDR** | unsupervised dimension reduction |
| **CMB** | chemical mass balance |
| **PCA** | principal components analysis |
| **FA** | factor analysis |
| **SOM** | self-organizing maps |
| **SDR** | supervised dimension reduction |
| **WQS** | weighted quantile sum regression |
| **CART** | classification and regression trees |
| **KMR** | kernel machine regression |
| **PLS** | partial least squares/projection to latent structures |

## References


1. Billionnet C, Sherrill D, Annesi-Maesano I. Estimating the health effects of exposure to multi-pollutant mixture. Ann Epidemiol. 2012; 22(2):126–41. [PubMed: 22226033]
2. Dominici F, Peng RD, Barr CD, Bell ML. Protecting human health from air pollution: Shifting from a single-pollutant to a multipollutant approach. Epidemiology. 2010; 21:187–194. http://dx.doi.org/10.1097/ede.0b013e3181cc86e8. [PubMed: 20160561]
3. Greenbaum D, Shaikh R. First steps toward multipollutant science for air quality decisions [Comment]. Epidemiology. 2010; 21:195–197. http://dx.doi.org/10.1097/ede.0b013e3181ccc52a. [PubMed: 20160562]
4. Painter K, Dutton SJ, Owens EO, Burgoon LD. Automatic document classification for environmental risk assessment. Peer Journal. 2014; 2:e300v1. http://dx.doi.org/10.7287/peerj.preprints.300v1/supp-1.
5. Gold DR, Damokosh AI, Pope CA III, et al. Particulate and ozone pollutant effects on the respiratory function of children in southwest Mexico City. Epidemiology. 1999; 10:8–16. http://www.ncbi.nlm.nih.gov/pubmed/9888274. [PubMed: 9888274]
6. Schildcrout JS, Sheppard L, Lumley T, Slaughter JC, Koenig JQ, Shapiro GG. Ambient air pollution and asthma exacerbations in children: An eight-city analysis. Am J Epidem. 2006; 164:505–517. http://dx.doi.org/10.1093/aje/kwj225.
7. Winquist A, Kirrane E, Klein M, et al. Joint effects of ambient air pollutants on pediatric asthma emergency department visits in Atlanta, 1998–2004. Epidemiology. 2014; 25:666–673. http://dx.doi.org/10.1097/ede.0000000000000146. [PubMed: 25045931]

8. Suh HH, Zanobetti A, Schwartz J, Coull BA. Chemical properties of air pollutants and cause-specific hospital admissions among the elderly in Atlanta, GA. Environ Health Perspectives. 2011; 119:1421–1428. http://dx.doi.org/10.1289/ehp.1002646.
9. Friedman, J., Hastie, T., Tibshirani, R. The elements of statistical learning. Springer, Berlin: Springer series in statistics; 2001.
10. Carlin, BP., Louis, TA. Bayesian methods for data analysis. Boca Raton, FL, USA: Chapman & Hall/CRC; 2009.
11. Roberts S, Martin M. A critical assessment of shrinkage-based regression approaches for estimating the adverse health effects of multiple air pollutants. Atmospheric Environment. 2005; 39:6223–6230. http://dx.doi.org/10.1016/j.atmosenv.2005.07.004.
12. Chadeau-Hyam M, Campanella G, Jombart T, et al. Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. Environ Mol Mutagen. 2013; 54(7):542–557. http://dx.doi.org/10.1002/em.21797. [PubMed: 23918146]
13. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Stat Society. 2005; 67(2):301–320. https://web.stanford.edu/~hastie/Papers/elasticnet.pdf.
14. Lenters V, Portengen L, Rignell-Hydbom A, et al. Prenatal Phthalate, Perfluoroalkyl Acid, and Organochlorine Exposures and Term Birth Weight in Three Birth Cohorts: Multi-Pollutant Models Based on Elastic Net Regression. Environmental health perspectives. 2016; 124(3):365–72. http://dx.doi.org/10.1289/ehp.1408933. [PubMed: 26115335]
15. Agier, L., Portengen, L., Chadeau-Hyam, M., et al. A systematic comparison of linear regression-based statistical methods to assess exposome-health associations. In 27th Conference of the International Society for Environmental Epidemiology; São Paulo. 2015. http://dx.doi.org/10.1289/EHP172
16. Katsouyanni, K., Samet, JM., Anderson, HR., et al. Air pollution and health: A European and North American approach APHENA. Boston, MA: Health Effects Institute; 2009. Research Report No.: 142Available from: http://www.ncbi.nlm.nih.gov/pubmed/20073322
17. Carbajal-Arroyo L, Miranda-Soberanis V, Medina-Ramón M, et al. Effect of $PM_{10}$ and $O_3$ on infant mortality among residents in the Mexico City Metropolitan Area: A case-crossover analysis, 1997–2005. J of Epidem and Community Health. 2011; 65:715–721. http://dx.doi.org/10.1136/jech.2009.101212.
18. Thomas DC, Jerrett M, Kuenzli N, et al. Bayesian model averaging in time-series studies of air pollution and mortality. J Toxicol and Environ Health A. 2007; 70(3–4):311–315. http://dx.doi.org/10.1080/15287390600884941. [PubMed: 17365593]
19. Herring AH. Nonparametric Bayes shrinkage for assessing exposures to mixtures subject to limits of detection. Epidemiology. 2010; 21(Suppl 4):S71–S76. http://dx.doi.org/10.1097/ede.0b013e3181cf0058. [PubMed: 20526202]
20. Bien J, Taylor J, Tibshirani R. A LASSO for hierarchical interactions. Ann of Stat. 2013; 41(3): 1111–1141. http://dx.doi.org/10.1214/13-aos1096. [PubMed: 26257447]
21. Sun Z, Tao Y, Li S, et al. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. Environ Health. 2013; 12:85. http://dx.doi.org/10.1186/1476-069x-12-85. [PubMed: 24093917]
22. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial. Stat Science. 1999; 14(4):382–417.
23. O'Hara RB, Sillanpaa MJ. A Review of Bayesian Variable Selection Methods: What, How and Which. Bayesian Analysis. 2009; 4(1):85–118. http://dx.doi.org/10.1214/09-ba403.
24. Ito K, Christensen WF, Eatough DJ, et al. PM source apportionment and health effects: 2 An investigation of intermethod variability in associations between source-apportioned fine particle mass and daily mortality in Washington, DC. J of Exposure Science and Environ Epidem. 2006; 16:300–310. http://dx.doi.org/10.1038/sj.jea.7500464.
25. Mar TF, Ito K, Koenig JQ, et al. PM source apportionment and health effects: 3 Investigation of inter-method variations in associations between estimated source contributions of $PM_{2.5}$ and daily mortality in Phoenix, AZ. J of Exposure Science and Environ Epidem. 2006; 16:311–320. http://dx.doi.org/10.1038/sj.jea.7500465.

26. Thurston G, Ito K, Mar T, et al. Results and implications of the workshop on the source apportionment of PM health effects. Epidemiology. 2005; 16:S134–S135. http://dx.doi.org/10.1097/00001648-200509000-00339.
27. Sacks JD, Ito K, Wilson WE, Neas LM. Impact of covariate models on the assessment of the air pollution-mortality association in a single- and multipollutant context. Am J Epidem. 2012; 176:622–634. http://dx.doi.org/10.1093/aje/kws135.
28. Zanobetti A, Austin E, Coull BA, Schwartz J, Koutrakis P. Health effects of multi-pollutant profiles. Environment International. 2014; 71:13–19. http://dx.doi.org/10.1016/j.envint.2014.05.023. [PubMed: 24950160]
29. Austin E, Coull B, Thomas D, Koutrakis P. A framework for identifying distinct multipollutant profiles in air pollution data. Environment International. 2012; 45:112–121. http://dx.doi.org/10.1016/j.envint.2012.04.003. [PubMed: 22584082]
30. Pearce JL, Waller LA, Mulholland JA, et al. Exploring associations between multipollutant day types and asthma morbidity: epidemiologic applications of self-organizing map ambient air quality classifications. Environ Health. 2015; 14:55. http://dx.doi.org/10.1186/s12940-015-0041-8. [PubMed: 26099363]
31. Hong YC, Leem JH, Ha EH, Christiani DC. PM10 exposure, gaseous pollutants, and daily mortality in Inchon, South Korea. Environmental Health Perspectives. 1999; 107:873–878. http://www.ncbi.nlm.nih.gov/pubmed/10544154.
32. Pachon JE, Balachandran S, Hu Y, et al. Development of outcome-based, multipollutant mobile source indicators. J of the Air and Waste Manag Assoc. 2012; 62:431–442. http://dx.doi.org/10.1080/10473289.2012.656218.
33. Roberts S, Martin MA. Investigating the mixture of air pollutants associated with adverse health outcomes. Atmospheric Environment. 2006; 40:984–991. http://dx.doi.org/10.1016/j.atmosenv.2005.10.022.
34. Carrico C, Gennings C, Wheeler DC, Factor-Litvak P. Characterization of Weighted Quantile Sum Regression for Highly Correlated Data in a Risk Analysis Setting. J of Agri, Biol, and Environ Stat. 2015; 20(1):100–120. http://dx.doi.org/10.1007/s13253-014-0180-3.
35. Roberts S, Martin MA. Using supervised principal components analysis to assess multiple pollutant effects. Environ Health Perspectives. 2006; 114:1877–1882. http://dx.doi.org/10.1289/ehp.9226.
36. Wold S, Ruhe A, Wold H, Dunn WJ III. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM Journal on Scientific and Statistical Computing. 1984; 5(3):735–43.
37. Seagrave JC, Mcdonald JD, Bedrick E, et al. Lung toxicity of ambient particulate matter from southeastern US sites with different contributing sources: relationships between composition and effects. Environ Health Perspectives. 2006; 114:1387–1393. http://dx.doi.org/10.1289/ehp.9234.
38. Nikolov MC, Coull BA, Catalano PJ, Godleski JJ. An informative Bayesian structural equation model to assess source-specific health effects of air pollution. Biostatistics. 2007; 8(3):609–24. http://dx.doi.org/10.1093/biostatistics/kxl032. [PubMed: 17032699]
39. Park ES, Hopke PK, Oh MS, Symanski E, Han D, Spiegelman CH. Assessment of source-specific health effects associated with an unknown number of major sources of multiple air pollutants: a unified Bayesian approach. Biostatistics. 2014; 15(3):484–97. [PubMed: 24622036]
40. Breiman, L., Friedman, J., Stone, CJ., Olshen, RA. Classification and regression trees. CRC press; 1984.
41. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological methods. 2009 Dec.14(4):323. [PubMed: 19968396]
42. Gass K, Klein M, Chang HH, Flanders WD, Strickland MJ. Classification and regression trees for epidemiologic research: an air pollution example. Environmental Health. 2014; 13:17. http://dx.doi.org/10.1186/1476-069x-13-17. [PubMed: 24625053]
43. Gass K, Klein M, Sarnat SE, et al. Associations between ambient air pollutant mixtures and pediatric asthma emergency department visits in three cities: a classification and regression tree

approach. Environmental Health. 2015; 14(1):1. http://dx.doi.org/10.1186/s12940-015-0044-5. [PubMed: 25564290]

44. Chipman HA, George EI, Mcculloch RE. BART: Bayesian additive regression trees. Ann Appl Stat. 2010; 4(1):266–298. http://dx.doi.org/10.1214/09-aoas285.
45. Bobb JF, Valeri L, Claus Henn B, et al. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics. 2015; 16:493–508. http://dx.doi.org/10.1093/biostatistics/kxu058. [PubMed: 25532525]
46. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: Least-Squares Kernel Machines and Linear Mixed Models. Biometrics. 2007; 63(4):1079–1088. [PubMed: 18078480]
47. Maity A, Lin X. Powerful tests for detecting a gene effect in the presence of possible gene-gene interactions using Garrote Kernel Machines. Biometric. 2011; 67:1271–1284.
48. Molitor J, Coker E, Jerrett M, Ritz B, Li A, Health Review Committee. Part 3 Modeling of Multipollutant Profiles and Spatially Varying Health Effects with Applications to Indicators of Adverse Birth Outcomes. Research report (Health Effects Institute). 2016; (183 Pt 3):3. [PubMed: 27459845]
49. Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association. 1993 Jun 1; 88(422):669–79.
50. Canale A, Dunson DB. Bayesian multivariate mixed-scale density estimation. Stat and its Interface. 2015; 8(20):195–201.
51. Molitor J, Papathomas M, Jerrett M, Richardson S. Bayesian profile regression with an application to the National Survey of Children's Health. Biostatistics. 2010; 11(3):484–498. http://dx.doi.org/10.1093/biostatistics/kxq013. [PubMed: 20350957]
52. Dunson DB, Xing C. Nonparametric Bayes modeling of multivariate categorical data. J of Am Stat Assoc. 2009; 104(487):1042–1051. http://dx.doi.org/10.1198/jasa.2009.tm08439.
53. Zhou J, Bhattacharya A, Herring AH, Dunson DB. Bayesian factorization of big sparse tensors. J of Am Stat Assoc. 2015; 110(512):1562–1576.
54. Thomas DC, Witte JS, Greenland S. Who's afraid of informative priors? Epidemiology. 2007; 18(2):186–190. [PubMed: 17301703]

**Table 1**

Summary of current statistical approaches applied to examine association between short-term multipollutant air pollution exposures and various health effects.

| Statistical Approach | | Example References | Short Characterization | Ease | When to Use | Certain Strengths | Certain Limitations |
|---|---|---|---|---|---|---|---|
| Additive Main Effects (AME) | Joint effects | 4–7 | Multipollutant or joint effects models with no pollutant interaction terms | Relative ease of construction and interpretation | Want to evaluate additive effects | Improves effect precision; Improved estimates of highly correlated variables | Multicollinearity adjusted for by various methods; Can bias results in either direction |
| | Penalized Regression Methods | 10, 13, 14 | | | | | |
| Effect Measure Modification (EMM) | | 6, 15, 16, 20 | Assess differential effects by EMM or multiplicative interaction terms; explore a multidimensional response surface without assuming pollutant effects are solely additive | Relative ease of construction and interpretation | Want to evaluate effect measure modification or multiplicative interactions | Only capable of capturing a section of the response-PMR surface; Reliant on model selection | Model uncertainty may be of concern |
| Unsupervised Dimension Reduction (UDR) | Statistical/Mathematical-Based Dimension Reduction Methods | 23–29 | Transform pollutant mixture concentration into a smaller set of variables used to represent pollutant combinations or sources; create clusters, groups or indices (e.g., PCA, k-means clusters, SOM) | Moderate ease of construction and interpretation | Information of exposure measurements available and measurement error is low | Can be used to address multicollinearity, compared across outcomes | Often specific to geographic area of study; Unable to inform pollutant concentrations at which health effects occur |
| | Scientifically-Based Dimension Reduction Methods | 30, 31 | | | Information on biological mechanisms is available | | |
| Supervised Dimension Reduction (SDR) | | 20, 32 – 34, 36 – 38 | Estimate a mixture transformation/dimension reduction concurrently with the regression analysis; | Moderate ease of construction, can be difficult to interpret | General specification of the relationship between the pollutants and exposure available | Can identify biologically relevant pollutant mixture; Can accommodate highly correlated data | Possibility of data feature loss; Requires moderately strong exposure response |
| Nonparametric methods | | 20, 41, 42, 44 – 47 | Does not assume specific form for the PMR; often coupled with parametric approaches that address other parts of the regression analysis (e.g., confounding) to deem the model as semiparametric | Moderate/ difficult to construct and interpret | Exploratory analysis; Hypothesis generation | Helpful in reducing model dimensions when number of candidate variables is large (useful prescreening tool); relaxing of PMR assumptions allow these methods to pick up interesting patterns of the PMR-response relationship | Joint effect estimation may not be easy to obtain; computation may require method expertise |

**Table 2**

Summary of statistical methods applicable for short-term ambient air pollution exposure studies.

| Lead Author | Year | Reference number | Regression Type | Estimation Method | Exclusively Short Term?[a] | Use of Pollutant indicator/factors or latent variables?[b] | Software Availability[c] | Stand Alone Procedure |
|---|---|---|---|---|---|---|---|---|
| Gold et al. | 1999 | 5 | Linear Time Series | Polynomial Distributed Lag | No | No | Standard | Yes |
| Schildcrout et al. | 2006 | 6 | Logistic GLM | Estimating Equations | No | No | Standard | Yes |
| Winquist et al. | 2014 | 7 | Poisson GLM | Estimating Equations | No | No | Standard | Yes |
| Suh et al. | 2011 | 8 | Conditional Logistic | Two-Step Hierarchical | Possibly | No | Standard | Yes |
| Roberts et al. | 2005 | 11 | Linear Time Series | Ridge, LASSO | No | No | R (ordered LASSO only) | Yes |
| Lenters et al. | 2016 | 14 | Linear GLM | Elastic Net | Yes | No | SAS, R (elasticnet) | Yes |
| Agier et al. | 2015 | 15 | Linear GLM | Multiple | Yes | No | SAS, R | Yes |
| Katsouyanni et al. | 2009 | 16 | Poisson Time Series | Bayesian Hierarchical, Meta Analysis | No | No | SAS, R | No |
| Carbajal-Arroyo et al. | 2011 | 17 | Conditional Logistic | Maximum Likelihood | Possibly | No | Standard | Yes |
| Herring et al. | 2010 | 19 | Logistic GLM | Bayesian Hierarchical | Possibly | No | No (MCMC) | |
| Sun et al. | 2013 | 21 | Linear GLM, Poisson Time Series | Multiple | Some | Some | R | Yes, individually |
| Ito et al. | 2006 | 24 | Poisson Time Series | Multiple | No | Yes | SAS, R | No |
| Mar et al. | 2006 | 25 | Poisson GLM | Multiple | No | Yes | SAS, R | No |
| Thurston et al. | 2005 | 26 | Poisson GLM | Multiple | No | Yes | SAS, R | No |
| Sacks et al. | 2012 | 27 | Poisson Time Series | PCA + Maximum Likelihood | No | Yes | SAS, R | No |
| Zanobetti et al. | 2014 | 28 | Poisson Time Series | Clustering + Maximum Likelihood | Possibly | Yes | R | No |
| Pearce et al. | 2015 | 30 | Poisson GLM | Clustering + Maximum Likelihood | Possibly | Yes | R | No |
| Hong et al. | 1999 | 31 | Poisson GLM | Maximum Likelihood | Yes | Yes | SAS, R | No |
| Pachon et al. | 2012 | 32 | Poisson Time Series | Maximum Likelihood | Possibly | Yes | SAS, R | No |
| Roberts et al. | 2006 | 33 | Poisson Time Series | Constrained Maximum Likelihood | Yes | No | Standard | No |
| Carrico et al. | 2015 | 34 | Linear GLM | Constrained Maximum Likelihood + Bootstrap | Possibly | Yes | SAS, R (WQS) | No, Yes |
| Roberts et al. | 2006 | 35 | Poisson GLM | SPCA + Maximum Likelihood | Possibly | No | R (superpc) | Yes |
| Seagrave et al. | 2006 | 37 | Linear GLM | PLS | No | Yes | Standard | Yes |
| Nikolov et al. | 2007 | 38 | Linear Random Effects | Bayesian Hierarchical | Possibly | Yes | Possibly R | No |
| Park et al. | 2014 | 39 | Linear Random Effects | Bayesian Hierarchical | Possibly | Yes | Possibly R | No |

| Lead Author | Year | Reference number | Regression Type | Estimation Method | Exclusively Short Term?[a] | Use of Pollutant indicator/factors or latent variables?[b] | Software Availability[c] | Stand Alone Procedure |
|---|---|---|---|---|---|---|---|---|
| Gass et al. | 2015 | 43 | Conditional Logistic | CART + Maximum Likelihood | Possibly | Yes | Standard | No |
| Chipman et al. | 2010 | 44 | Linear, Logistic GLM | BART | Possibly | Yes | R (BayesTree) | Yes |
| Bobb et al. | 2015 | 45 | Linear Random Effects | Bayesian KMR | Yes | No | R (bkmr) | Yes |
| Liu et al. | 2007 | 46 | Linear Random Effects | KMR | Possibly | No | Standard | No |
| Maity et al. | 2011 | 47 | Linear Random Effects | KMR | Possibly | No | Standard | No |
| Molitor et al. | 2016 | 48 | Logistic Random Effects | Bayesian Hierarchical | No | Yes | R (PReMiuM) | Yes |

[a] Possibly implies the model idea may be applicable to analyze associations between long-term exposure and health effect so long as the computation allows.

[b] Some implies there were multiple estimation methods applied and some used latent variables.

[c] Standard implies computation is achievable in SAS, R, and Stata; items in parenthesis specify specific R packages necessary.