Curr Envir Health Rpt (2017) 4:481–490
DOI 10.1007/s40572-017-0162-z

Statistical Approaches to Address Multi-Pollutant Mixtures and Multiple Exposures: the State of the Science

Massimo Stafoggia[1,2] · **Susanne Breitner**[3] · **Regina Hampel**[3] · **Xavier Basagaña**[4,5,6]

Published online: 7 October 2017

Abstract
Purpose of Review The purpose of this review is to describe the most recent statistical approaches to estimate the effect of multi-pollutant mixtures or multiple correlated exposures on human health.
Recent Findings The health effects of environmental chemicals or air pollutants have been widely described. Often, there exists a complex mixture of different substances, potentially highly correlated with each other and with other (environmental) stressors. Single-exposure approaches do not allow disentangling effects of individual factors and fail to detect potential interactions between exposures. In the last years, sophisticated methods have been developed to investigate the joint or independent health effects of multi-pollutant mixtures or multiple environmental exposures.
Summary A classification of the most recent methods is proposed. A non-technical description of each method is provided, together with epidemiological applications and operational details for implementation with standard software.

Keywords Correlated variables · Environmental exposures · Epidemiology · Health · Multi-pollutant

This article is part of the Topical Collection on *Air Pollution and Health*

✉ Massimo Stafoggia
m.stafoggia@deplazio.it

[1] Department of Epidemiology, Lazio Region Health Service/ASL Roma 1, Via Cristoforo Colombo 112, 00147 Rome, Italy

[2] Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

[3] Institute of Epidemiology II, Helmholtz Zentrum München-German Research Center for Environmental Health (GmbH), Neurherberg, Germany

[4] ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain

[5] Pompeu Fabra University, Barcelona, Spain

[6] Ciber on Epidemiology and Public Health (CIBERESP), Madrid, Spain

Introduction

Associations between environmental chemicals or air pollution and adverse health outcomes have been reported worldwide [1–4]. However, the population is exposed simultaneously to a large number of air pollutants or chemical contaminants. The *exposome* paradigm is an attempt to focus the attention to the multiple environmental factors affecting health. The *exposome* was defined as the totality of environmental (non-genetic) exposures from conception onwards [5]. Thus, in its broad sense, it includes air pollutants and other contaminants, but also behavioral and socioeconomic characteristics. Environmental epidemiology studies are collecting data on an increasing number of exposures to try to capture parts of the *exposome*, which poses the challenge of analyzing the effects of mixtures of exposures. For example, in single-pollutant models, it is not clear if an observed association reflects the effect of the analyzed pollutant or if it acts as a surrogate for another pollutant possibly originating from the same source. Furthermore, single-pollutant models cannot capture the mixture and interplay of different exposures. Analyzing health effects of several pollutants by including them together in a regression model is in many cases not meaningful because of the usually high correlation between these air pollutants. This "naive" multi-pollutant model can result in unstable parameter estimates with large standard errors. Therefore, more sophisticated methods are needed to investigate the health effects of mixtures of exposures or

simultaneous effects of multiple exposures. The National Institute of Environmental Health Sciences (NIEHS), the Health Effects Institute (HEI), the U.S. Environmental Protection Agency (EPA), and other experts in the field identified statistical approaches to address multi-pollutant mixtures and multiple exposures as an important area of ongoing research, also for regulatory purposes [6•, 7–11].

Several statistical approaches have been proposed in the last few years to correctly estimate the independent and joint effects of multiple correlated exposures on human health [12]. Most of the methods were borrowed from other disciplines (such as epigenetics) and transferred to the multi-pollutant case; however, they can be easily adapted to the general setting of multiple correlated exposures. The purpose of this review is to present the latest advancements in this field and to describe the main statistical approaches to the broader audience of environmental epidemiologists in a non-technical way, focusing on the pros and cons of each method and providing software details for their implementation.

A Classification of Methods for the Analysis of Multiple Correlated Exposures

Broadly speaking, nearly all of the proposed methods can be classified into three groups: dimension reduction, variable selection, and grouping of observations.

The aim of the first group of methods is to transform a large number of correlated variables (pollutants, exposures, etc.) into a smaller set of independent factors to be related with the health outcome. Dimensionality reduction can be done in an *unsupervised* way, i.e., only taking into account the associations between exposures, or in a *supervised* way, which additionally takes into account the correlation between the exposures and the outcome. The disadvantage of unsupervised methods is that they may give a strong weight to pollutants that have no influence on the outcome or vice versa. The disadvantage of supervised methods is that they lead to outcome-specific independent variables, which makes comparability of results across outcomes difficult.

Methods belonging to the second group called variable selection methods aim to identify the "best" subset of exposure variables, either based on their mutual correlation (*unsupervised*) or on their relation with the study outcome (*supervised*). Compared with the methods in the first group, those in the second one do not produce transformations of variables but retain subsets of the original ones. The main advantage is that the estimated coefficients are directly interpretable as they are on the same scale as the original exposures.

The methods in the third group aim to group observations (rather than variables) with similar exposure profiles and use this grouping in the analysis on the health outcome. Grouping is generally achieved by simultaneously maximizing the intra-class and minimizing the inter-class similarity. Again, unsupervised and supervised methods to derive those groups have been proposed.

Figure 1 schematically displays the multi-pollutant/multiple exposure setting and the three groups of statistical approaches. Details of the methods, including references of the most relevant applications and R packages for their implementation, are displayed in Table 1.

Methods for Dimension Reduction

These methods aim at reducing the number of exposures to be used in a regression model for the health outcome by creating combinations of the original exposures. Unsupervised methods include principal component analysis (PCA) and positive matrix factorization (PMF). Supervised methods include supervised principal component analysis (SPCA), partial least squares (PLS) regression, sparse partial least squares (SPLS) regression, and weighted quantile sum regression (WQSR).

Principal Component Analysis

A common method for dimension reduction is principal component analysis (PCA). PCA searches for independent (i.e., not correlated with each other) linear combinations of the exposure variables which capture most of the variance of the initial exposure data [13]. Briefly, a decomposition of the covariance matrix is performed. Based on the resulting eigenvectors and eigenvalues, independent linear combinations of the exposure variables, so-called principal components (PCs), are identified. A component represents a mixture of pollutants, each contributing with different weights. By estimating the effect of such component on a health outcome, one obtains the effect of a pollution mixture on health. There are several approaches to choose the optimal number of PCs. A graphical possibility to choose the number of PCs is given by the scree plot which reports the fraction of the total variance in the data explained by each PC. In general, the first one or two PCs capture most of the variability in the data. The main advantages of PCA are the ease of application and the efficiency (most of the variability in the original exposures retained by few factors). The major limitations include the difficulty to interpret the results, as the components are not on the same units as the original exposure variables, and the potential lack of relationship of the derived components with the study outcome, as they were derived in an unsupervised way.

PCA analysis has been applied, among others, by Yang et al. to investigate the association between ambient air pollution and daily mortality in Beijing after the 2008 Olympics. The authors applied conventional single-pollutant models

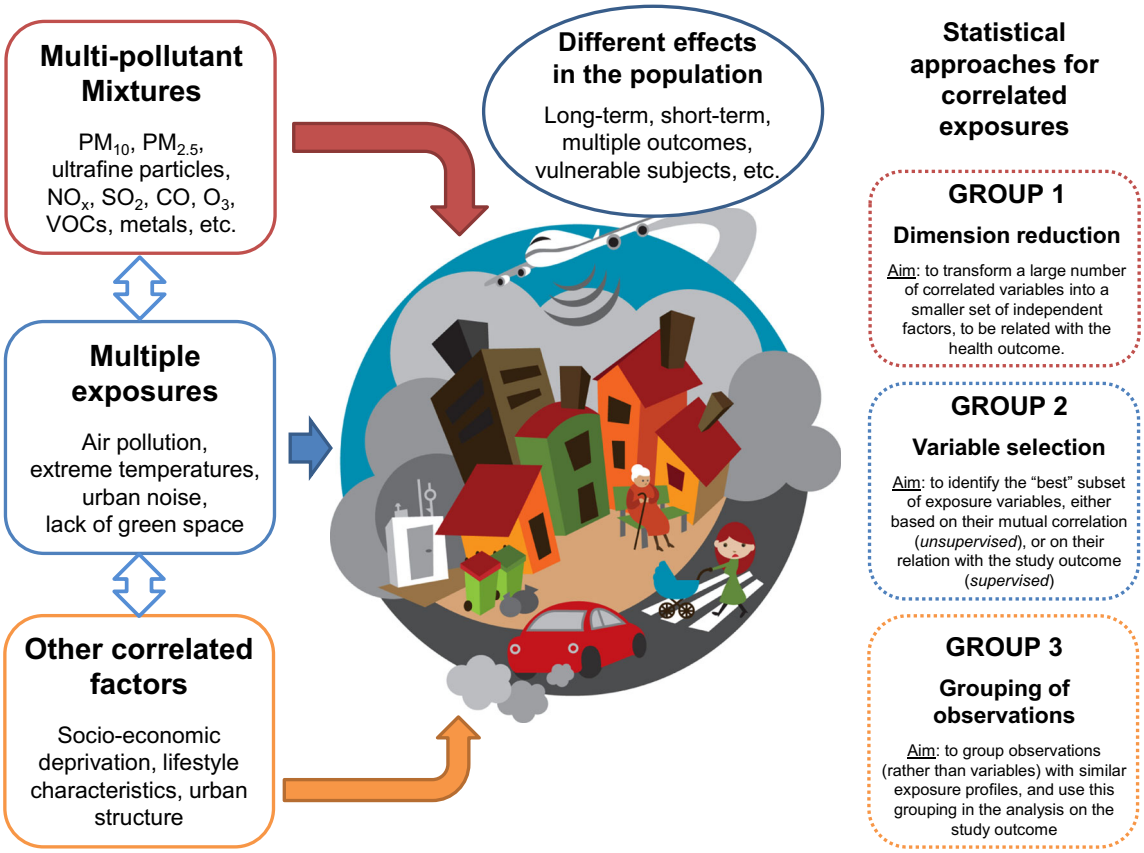

Fig. 1 A schematic representation of the statistical methods used for the analysis of the health effects of multiple correlated exposures

first, and then they replaced individual pollutants with latent variables identified through PCA, which resulted in significant effects on mortality and hospital admissions, though smaller than those found with single-pollutant analysis [14].

Positive Matrix Factorization

Positive matrix factorization (PMF) is a variant of PCA applied in the context of multi-pollutant profiles to derive air pollution sources from PM individual chemical components [15]. Specifically, PMF decomposes a matrix of PM-speciated data into two matrices—source contributions and source profiles: source contributions represent the amounts of PM mass contributed by each source to the measurement, while source profiles reflect the types of emissions that originate from a given source. Source contributions are constrained to be non-negative and the method can incorporate measures of uncertainty associated with the data to weight individual points. Algorithms such as multilinear engine (ME-2) allow incorporating any a priori knowledge, such as chemical ratios or profiles of involved sources, into the model [16]. Thus, this technique is particularly suitable for source apportionment studies. Several epidemiological studies have been conducted in the last years using source apportionment estimates based on PMF [17•, 18–23].

Supervised Principal Component Analysis

In unsupervised analyses like PCA, some pollutants might receive a (strong) weight even if they are not associated with the outcome of interest. Supervised principal component analysis (SPCA)—a modified version of PCA—overcomes this problem by excluding pollutants which do not provide information directly related to the outcome [24]. After the selection of influential exposure variables, a PCA is performed using the reduced number of exposure variables. Roberts and Martin [25] adapted this method for air pollution analysis, suggesting a recursive algorithm which identifies the best predictors of the study outcome (variable selection), and combined them into a few relevant PCs. The procedure has been implemented for Poisson regression but can easily be adapted for other models.

Partial Least Squares Regression

Partial least squares (PLS) regression takes the correlation between the outcome and exposure variables into account by combining PCA and multiple regression analysis [26]. Briefly, PLS regression searches for a linear decomposition of the exposure matrix which maximizes the covariance between the exposure and the outcome; the stronger the correlation

Table 1 Overview of statistical methods, with details, applications, and R packages for implementation

Method	Supervised	Theory	Application in environmental epidemiology	R package
Dimension reduction				
PCA	No	Anderson [13]	Yang et al. [14]	*stats*
PMF	No	Paatero and Tapper [15]	Krall and Strickland [17•]	ad-hoc software (example: EPA PMF 5.0)
SPCA	Yes	Bair et al. [24]	Roberts and Martin [25]	*superpc*
PLS	Yes	Wold [26]	Sun et al. [28]	*stats*, *pls*
SPLS	Yes	Chun and Keleş [29]	Agier et al. [30•]	*spls*
WQSR	Yes	Carrico et al. [31]	Czarnota et al. [32]	*wqs*
Variable selection				
Cluster prototypes	No	Reid and Tibshirani [33]	–	*protoclust and prototest*
D/S/A algorithm	Yes	Sinisi and van der Laan [36]	Beckerman et al. [37]	*modelUtils, DSA*
BMA	Yes	Amini and Parmeter [38]	Bobb et al. [40]	*bms*
LASSO	Yes	Tibshirani [41]	Sun et al. [28]	*glmnet*
ENET	Yes	Zou and Hastie [42]	Lenters et al. [44]	*glmnet*
GLINTERNET	Yes	Lim and Hastie [46]	Agier et al. [30•]	*glinternet*
R2GUESS	Yes	Liquet et al. [48]	Agier et al. [30•]	*R2GUESS*
BKMR	Yes	Bobb et al. [51•]	Bobb et al. [51•]	*bkmr*
Grouping of observations				
k-means	No	Steinley [53]	Ljungman et al. [55]	*stats*
Groups based on score	No	Lee et al. [58]	Lee et al. [58]	–
BPR	Yes	Molitor et al. [59]	Papathomas et al. [61]	*PReMiuM*
CART	Yes	Strobl et al. [64]	Gass et al. [65]	*party*

PCA principal component analysis, *PMF* positive matrix factorization, *SPCA* supervised principal component analysis, *PLS* partial least squares, *SPLS* sparse partial least squares, *WQSR* weighted quantile sum regression, *D/S/A* deletion/substitution/addition, BMA Bayesian model averaging, *LASSO* least absolute shrinkage and selection operator, *ENET* elastic net, *GLINTERNET* group-lasso INTERaction-NET, *BKMR* Bayesian kernel machine regression, *CART* classification and regression trees

between an exposure variable and the outcome, the larger the weight for this exposure variable in the linear combination. In PLS regression, it is also possible to include several outcome variables. The mean squared error of prediction using cross-validation is used in order to choose the optimal number of components [27].

In a recent study, Sun and colleagues compared the performance of five statistical methods, among which SPCA and PLS, under two different simulated settings: continuous outcome from a cross-sectional study and daily event counts from a time-series study. They simulated data on different number of pollutants, from four to 20, and different degrees of multicollinearity and designed "true" multivariate models with main effects of individual pollutants plus some pairwise interactions. All models displayed varying degrees of goodness-of-fit depending on the simulated scenario, with no clear superiority of one model over the others [28].

Sparse Partial Least Squares Regression

The disadvantage of PLS regression is the difficult interpretability of the linear combinations, especially in the case of a large number of original exposure variables. Chun and Keleş therefore introduced a method which incorporates variable selection and dimension reduction simultaneously [29]. This approach leads to linear combinations of a reduced number of exposure variables. Sparsity is introduced by a penalty term to the loadings of the exposure variables. The optimal number of components and the parameter for sparsity are chosen based on cross-validation.

The method has been recently applied in a simulation study on exposome-health association with 237 generated exposure covariates and realistic correlation structures [30•].

Weighted Quantile Sum Regression (WQSR)

In WQSR, all exposures are first categorized, e.g., using quartiles. Then, one finds a single index to summarize all exposures, obtained as a weighted average of the categorized exposures. The weights for each exposure are between zero and one, and they are chosen to maximize the likelihood of the regression model of the outcome variable against the exposure index, adjusted for covariates, using bootstrap samples [31]. This technique assumes that all exposures contribute in the

same direction to the outcome, and therefore, it should be used only with exposures that are hypothesized to act in the same direction on the investigated health outcome. The WQSR method was used in a study of the relationship between exposure to 27 chemicals and non-Hodgkin lymphoma [32].

Methods for Variable Selection

Methods for variable selection aim to identify the "best" subset of exposure variables either based on their mutual correlation (unsupervised) as is used in the cluster prototypes approach or on their relation with the study outcome (supervised approach; examples include the deletion/substitution/addition (D/S/A) algorithm, Bayesian model averaging (BMA), penalized methods, and Bayesian variable selection methods).

Cluster Prototypes

The use of cluster prototypes is a novel approach for data with correlated variables [33]. The procedure first clusters the variables; clustering can be thereby done using any particular clustering method, e.g., hierarchical clustering methods using the minimax linkage [34]. Having identified several clusters, the next step is to choose a single representative—so-called prototype—for each cluster. A key feature of the approach is its use of post-selection inference theory provided by Tibshirani et al. [35] to compute exact p values and confidence intervals that properly account for the selection of prototypes.

Deletion/Substitution/Addition Algorithm

The deletion/substitution/addition (D/S/A) algorithm is an iterative selection approach [36]. It builds a model space of candidate models based on the following three steps: (1) a deletion step which removes a term from the model, (2) a substitution step which replaces one term with another, and (3) an addition step which adds a term to the model. The choice of a move is based on a loss function-based estimation procedure (with the aim of minimizing a specific loss function). In the case of linear regression, the move which minimizes the sum of squared residuals is selected. The final model returned is identified via cross-validation. The algorithm also provides the possibility to include interaction terms. Compared to stepwise model selection procedures, D/S/A has the advantages of being less sensitive to outliers and of allowing the search to move between statistical models that are not nested.

Among other approaches, the D/S/A algorithm was applied in the context of cross-sectional analysis using data from the National Health and Nutrition Examination Survey (NHANES) [28]. In this application, several environmental contaminants (e.g., phthalates) were found to be associated with systemic markers of oxidative stress. Further, the D/S/A algorithm has been used to select predictor variables for land use regression (LUR) models [37].

Bayesian Model Averaging

Bayesian model averaging (BMA) is a method which takes model uncertainty into account. BMA judges the importance of single-exposure variables by estimating all possible models (i.e., all possible exposure combinations) and constructs a weight for each model [38]. For each possible model, a prior probability has to be selected. A simple choice is the use of uniform prior probabilities reflecting a lack of prior knowledge. The final weights for the exposure variables are derived from posterior model probabilities and reflect the impact of this variable. BMA automatically shrinks the number of exposure variables by giving weights of zero for some variables. The final exposure effect is calculated as a weighted average of the exposure effects from each of the models. The importance of a specific exposure variable can be judged by the posterior inclusion probability. A low value indicates that the exposure effect was zero for many models. As the number of possible models increases quickly in case of many exposure variables, it is not feasible to perform all models. Two main approaches have been proposed to overcome this problem—dimensionality reductions of the model space and stochastic searches through Markov chain Monte Carlo (MCMC) [39]. The R package *BMS* uses MCMC samplers, applying the Metropolis-Hastings algorithm, to identify the most important part of the posterior distribution.

BMA has been applied in several environmental studies, for example, in a study using time-series data from 105 US cities to estimate the relative risk of mortality associated with heat waves [40].

Penalized Methods

The LASSO (least absolute shrinkage and selection operator) is very similar to ordinary least squares, except that the coefficients are estimated by minimizing a slightly different quantity—it imposes a shrinkage penalty on the size of coefficients [41]. It penalizes the absolute size of the regression coefficients based on the value of a tuning parameter λ. In doing so, the LASSO can drive the coefficients of irrelevant variables to zero, thus performing automatic variable selection. When the tuning parameter λ is small, the result is essentially the least squares estimates.

Elastic net (ENET) combines the LASSO method and Ridge regression [42]. This model includes penalty terms of both first and second degree for the regression coefficients. Thus, not only the best subset of variables is selected by shrinking some effect estimates exactly to zero (LASSO) but

groups of (highly) correlated variables are kept in the model with similar effect estimates (ridge regression).

Using a later version of the LASSO, an example from the Veterans Affairs Normative Aging Study was able to select $PM_{2.5}$ components associated with blood pressure [43]. Based on elastic net penalty regression, a recent study found that, among others, two phthalate metabolites were most consistently associated with impaired fetal growth [44].

There are many developments around the LASSO with potential relevance for environmental epidemiology studies [45]. For example, the GLINTERNET method extends the LASSO to select two-way interaction terms [46], and a recent development allows using the LASSO while controlling for the false discovery rate [47].

Bayesian Variable Selection Methods

Several Bayesian variable selection methods have been proposed. These approaches incorporate a vector with the probability of each exposure having a zero (or non-zero) effect. Then, one assigns prior probabilities and lets the data inform on the value of such probabilities. One of such implementations is the GUESS method, available in the R package R2GUESS [48]. MacLehose et al. proposed a method that performed variable selection and also clustered regression coefficients into groups having similar effects based on prior knowledge and information in the data [49]. Hill et al. proposed one method to incorporate biological knowledge in the prior distributions, using available information from pathway maps [50].

Bayesian kernel machine regression (BKMR) is a recent approach based on a popular tool in the machine learning literature—kernel machine regression or Gaussian process regression [51•]. The idea behind BKMR is to flexibly model the relationship between a large number of exposure variables/mixture components and a particular health outcome. This is done by using a smooth function of the exposure variables/mixture components represented using a Gaussian kernel function. The kernel machine representation allows the incorporation of non-linear effects and/or interaction among mixture components. To systematically handle highly correlated exposures, a hierarchical variable selection approach within BKMR is used which can incorporate prior knowledge on the structure of how the exposure variables/mixture components are related. The hierarchical variable selection approach allows estimating the posterior inclusion probability for each exposure.

BKMR has been applied to a dataset on metal exposures and neurodevelopment in children in Bangladesh suggesting a non-additive and non-linear exposure-response function between the metals and a summary measure of psychomotor development [51•].

Methods for Grouping of Observations

The techniques described in this third group attempt to cluster observations (i.e., rows in a dataset) so that each of the resulting groups has a distinctive profile in terms of the exposures. The outcome obtained from this clustering or grouping is a categorical variable indicating cluster membership. The clusters or groups of observations can then be compared in terms of the health outcome. Note the differences between this procedure and the techniques described in the first group, which grouped exposures (i.e., columns in the dataset) and synthesized the information of each group by creating a new set of continuous variables.

As in the first group of methods, the grouping of observations can be done in an unsupervised or a supervised way, depending on whether the health outcome is used to form the clusters. The unsupervised option involves a two-step process. First, the groups of observations are formed using only exposure data. Second, the categorical variable that identifies the groups is used as predictor in a regression model for the health outcome. Examples of unsupervised analyses include cluster analysis or building groups based on an exposure score. In a supervised analysis, the grouping of observations is done taking the health outcome into account, favoring groupings that result in marked differences between the groups in terms of the outcome of interest. Examples of supervised techniques include Bayesian profile regression or recursive partitioning techniques.

Cluster Analysis

This approach first fits a cluster analysis technique to the exposure data in order to define the groups and then includes the indicators of group membership as predictors in a regression model for the health outcome. There are hundreds of clustering techniques and none of them can be considered to outperform the others in all situations [52]. Clustering techniques can be classified into different groups. The most widely used fall into the partitioning-based, of which k-means [53] or partitioning around medoids (PAM) are two of the most popular; the hierarchical-based, in which observations are sequentially grouped (agglomerative clustering) or separated (divisive clustering) according to a proximity measure; and the model-based, which assumes that the data were generated by a model and it estimates its parameters (this category includes finite mixture model, latent class model, or probabilistic self-organizing maps) [54].

Clustering has been used in several papers assessing the effects of multiple pollutants. For example, in the context of time-series analyses of air pollution, one study used k-means to classify days into five clusters, indicating days of low pollution levels, days with high concentrations of crustal particles, days with high levels of particles from traffic and oil

combustion, days affected by regional sources, and days with high levels of particles from wood burning or combustion of heating oil [55]. Some of the clusters were associated with pulse amplitude.

Groups Obtained Through a Cumulative Score

Another analysis option is to derive the groups of observations based on a score that accumulates the levels of all the exposures under study. This could be done through creating a principal component from PCA, in which case each exposure will contribute to the score with a different weight, or through other options. For example, some studies have summed the ranks of each exposure [56] or the number of exposures with values in the top decile [57]. By categorizing the resulting score, the different groups represent participants exposed to a large or small number of exposures simultaneously. Using one of these approaches, one study found that the group exposed simultaneously to high levels of six PCBs had a much higher prevalence of diabetes than those with lowest levels for all PCBs [58]. This type of analysis can be useful in situations in which populations are exposed to low levels of a large number of exposures, and even if exposure to low levels of one of the compounds can be considered safe, simultaneous exposure to many of them can result in some health effects.

Bayesian Profile Regression

Bayesian profile regression (BPR) is a model-based technique that aims at finding clusters of subjects sharing similar exposure profiles that at the same time show differences in the health outcome [59]. Unlike the clustering techniques described above, the clustering part and the regression of the outcome on the clusters are done in a single step. BPR is quite flexible, as it allows continuous and categorical exposures, it automatically handles missing data, and it calculates the optimal number of clusters, and the assignment to clusters is done probabilistically, i.e., it takes into account uncertainty in cluster assignment. In addition, one can incorporate variable selection into the algorithm, which can facilitate the interpretation of the clusters and can lead to better performance in studies with a large number of exposures relative to sample size [60]. BPR has been used in several studies. One of them identified a cluster of participants characterized by living close to a main road, having high exposure to PM_{10} and nitrogen dioxide, and carrying out manual work as being at high risk for lung cancer [61]. It has also been used in the context of time-series analysis in a study that identified days with high concentrations of nitrate and sulfate as having higher risk of respiratory mortality [62•].

A regression profile analysis can capture complex interactive effects of several exposures while producing an output that is more interpretable than the output of a regular regression model with several high-order interaction terms [63]. In addition, it has the advantage that it directly directs attention to patterns of values that are most characteristic of the data and that it does not assume a pre-specified form for the association.

Recursive Partitioning Techniques

The most basic techniques of recursive partitioning are classification and regression trees (CART). CART is a supervised technique in which a sequence of binary splits based on a subset of exposures creates groups of observations in a way that—within each group—observations have similar values of the outcome. The sequence of binary splits creates a tree structure. At each node of the tree, the observations are split into two groups based on one exposure, i, and a statement of the form $\text{exposure}_i < c$ vs. $\text{exposure}_i \geq c$, for a value of c chosen by the CART algorithm [64]. The tree structure, usually based on a small set of exposures, facilitates the interpretation of the clusters. CART can easily capture complex interactions and may facilitate their interpretation. In particular, any asymmetries in the tree encompass underlying interactions. In a time-series setting, a study used CART to characterize the joint effects of CO, NO_2, O_3, and $PM_{2.5}$ on asthma admissions [65].

There are other techniques closely related to CART which create trees based on "and" and "or" statements, such as logic regression or part DSA [66]. Cross-validated CARTs usually include a few terminal nodes, with limited explanatory power. Random forests and boosting are two techniques that have been used to combine the results of several CARTs in order to achieve better predictive performance [64, 67]. The drawback of those techniques is that the easy interpretability of single CARTs is lost and one has to rely on measures of variable importance that do not directly link the exposure with the outcome. Lampa et al. offer some tools on how to assess if interactions are present and how to visualize them within boosted regression trees [67].

Conclusion

Humans are exposed to multiple environmental stressors simultaneously. Especially in urban settings, where most of the population lives, air pollution, environmental chemicals, noise, temperature extremes, lack of green space, but also social inequality and detrimental lifestyle habits are likely to negatively affect health and quality of life. A better understanding of the independent and synergistic/antagonistic effects of these risk factors is mandatory for: (a) designing effective public health prevention strategies especially targeted on vulnerable population subgroups and (b) promoting urban policies aimed at reducing air pollution and environmental

chemical concentrations as well as noise and heat levels at the same time.

In this review, we proposed a classification of the statistical methods recently applied to jointly estimate the effects of multiple correlated exposures on human health or to disentangle the effects of the components to the multi-pollutant mixture. While the list is not presumed to be exhaustive, it categorizes all methods into one of three different groups (some of the methods overlapping more groups): dimension reduction, variable selection, and grouping of observations. Each class is characterized by different pros and cons and its applicability depends on the available data and the degree of multicollinearity among exposures. For each of the proposed methods, specific R packages have been developed and studies have been conducted in the last years with environmental epidemiology applications, especially in the multi-pollutant case.

The existing literature, despite being limited up to now, shows the high potential of these approaches for disentangling individual and joint effects, often non-linear, of multiple correlated exposures on human health. Further studies are needed, especially in the context of multiple exposures other than air pollution, to improve our understanding of the complex interrelationship among environmental, socio-economic, and lifestyle risk factors on human health.

Acknowledgments ISGlobal is a member of the CERCA Programme, Generalitat de Catalunya.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:
• Of importance

1. International Programme on Chemical Safety (IPCS)-World Health Organization (WHO). Public health impact of chemicals: knowns and unknowns. Geneva: World Health Organization; 2016.
2. International Agency for Research on Cancer (IARC). IARC monographs on the evaluation of carcinogenic risks to humans. Lyon: World Health Organization; 2015.
3. Lelieveld J, Evans JS, Fnais M, Giannadaki D, Pozzer A. The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature. 2015;525:367–71.
4. GBD 2013 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015;386:2287–323.
5. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomark Prev. 2005;14:1847–50.
6.• Taylor KW, Joubert BR, Braun JM, Dilworth C, Gennings C, Hauser R, et al. Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: lessons from an innovative workshop. Environ Health Perspect. 2016;124:A227–9. **This paper provides an important summary of a workshop organized by NIEHS on statistical methods for the analysis of environmental chemical mixtures.**
7. Dominici F, Peng RD, Barr CD, Bell ML. Protecting human health from air pollution: shifting from a single-pollutant to a multi-pollutant approach. Epidemiology. 2010;21:187–94.
8. Health Effects Institute (HEI). Strategic plan for understanding the health effects of air pollution 2015–2020. Boston: Health Effects Institute; 2014.
9. Johns DO, Stanek LW, Walker K, Benromdhane S, Hubbell B, Ross M, et al. Practical advancement of multipollutant scientific and risk assessment approaches for ambient air pollution. Environ Health Perspect. 2012;120:1238–42.
10. Mauderly JL, Burnett RT, Castillejos M, Ozkaynak H, Samet JM, Stieb DM, et al. Is the air pollution health research community prepared to support a multipollutant air quality management framework? Inhal Toxicol. 2010;22S1:1–19.
11. U.S. Environmental Protection Agency (EPA). The multi-pollutant report: technical concepts and examples. Washington, DC: US Environmental Protection Agency; 2008.
12. Billionnet C, Sherrill D, Annesi-Maesano I. Estimating the health effects of exposure to multi-pollutant mixture. Ann Epidemiol. 2012;22:126–41.
13. Anderson TW. An introduction to multivariate statistical analysis. 2nd ed. New York: John Wiley & Sons; 1984.
14. Yang Y, Li R, Li W, Wang M, Cao Y, Wu Z, et al. The association between ambient air pollution and daily mortality in Beijing after the 2008 Olympics: a time series study. PLoS One. 2013;e76759:8.
15. 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.
16. Paatero P. The multilinear engine—a table-driven least squares program for solving multilinear problems, including the n-way parallel factor analysis model. J Comput Graph Stat. 1999;8:1–35.
17.• Krall JR, Strickland MJ. Recent approaches to estimate associations between source-specific air pollution and health. Curr Environ Health Rep. 2017;4:68–78. **Krall et al. provide a thorough review of recent methodological developments in the study of the association between source-specific air pollution and health.**
18. Krall JR, Mulholland JA, Russell AG, Balachandran S, Winquist A, Tolbert PE, et al. Associations between source-specific fine particulate matter and emergency department visits for respiratory disease in four US cities. Environ Health Perspect. 2017;125:97–103.
19. Dai L, Bind M-A, Koutrakis P, Coull BA, Sparrow D, Vokonas PS, et al. Fine particles, genetic pathways, and markers of inflammation and endothelial dysfunction: analysis on particulate species and sources. J Expo Sci Environ Epidemiol. 2016;26:415–21.
20. Siponen T, Yli-Tuomi T, Aurela M, Dufva H, Hillamo R, Hirvonen M-R, et al. Source-specific fine particulate air pollution and systemic inflammation in ischaemic heart disease patients. Occup Environ Med. 2015;72:277–83.
21. Gass K, Balachandran S, Chang HH, Russell AG, Strickland MJ. Ensemble-based source apportionment of fine particulate matter

and emergency department visits for pediatric asthma. Am J Epidemiol. 2015;181:504–12.
22. Park ES, Symanski E, Han D, Spiegelman C. Part 2. Development of enhanced statistical methods for assessing health effects associated with an unknown number of major sources of multiple air pollutants. In: Development of statistical methods for multipollutant research. Res Rep Health Eff Inst. 2015; 183:51–113.
23. Basagaña X, Esnaola M, Rivas I, Amato F, Alvarez-Pedrerol M, Forns J, et al. Neurodevelopmental deceleration by urban fine particles from different emission sources: longitudinal observational study. Environ Health Perspect. 2016;124:1630–6.
24. Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. J Am Stat Assoc. 2006;101:119–37.
25. Roberts S, Martin MA. Using supervised principal components analysis to assess multiple pollutant effects. Environ Health Perspect. 2006;114:1877–82.
26. Wold H. Estimation of principal components and related models by iterative least squares. In: Krishnaiah PR, editor. Multivariate analysis. New York: Academic Press; 1966. p. 391–420.
27. Mevik BH, Wehrens R. The *pls* package: principal component and partial least squares regression in R. J Stat Softw. 2007;18:1–23.
28. Sun Z, Tao Y, Li S, Ferguson KK, Meeker JD, Park SK, et al. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. Environ Health. 2013;12:85.
29. Chun H, Keles S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J R Stat Soc B. 2010;72:3–25.
30.• Agier A, Portengen L, Chadeau-Hyam M, Basagaña X, Giorgis-Allemand L, Siroux V, et al. A systematic comparison of linear regression-based statistical methods to assess exposome-health associations. Environ Health Perspect. 2016;124:1848–56. **This study conducted a comparison of the performance of several variable selection methods in an exposome setting.**
31. Carrico C, Gennings C, Wheeler DC, Factor-Litvak P. Characterization of a weighted quantile sum regression for highly correlated data in a risk analysis setting. J Agric Biol Environ Stat. 2015;20:100. https://doi.org/10.1007/s13253-014-0180-3.
32. Czarnota J, Gennings C, Colt JS, De Roos AJ, Cerhan JR, Severson RK, et al. Analysis of environmental chemical mixtures and non-Hodgkin lymphoma risk in the NCI-SEER NHL study. Environ Health Perspect. 2015;123:965–70.
33. Reid S, Tibshirani R. Sparse regression and marginal testing using cluster prototypes. Biostatistics. 2016;17:364–76.
34. Bien J, Tibshirani R. Hierarchical clustering with prototypes via minimax linkage. J Am Stat Assoc. 2011;106:1075–84.
35. Tibshirani RJ, Taylor J, Lockhart R, Tibshirani R. Exact post-selection inference for sequential regression procedures. arXiv 2014:1401.3889v5 [stat.ME].
36. Sinisi S, van der Laan M. Deletion/substitution/addition algorithm in learning with applications in genomics. Stat Appl Genet Mol Biol. 2004;3:Article18.
37. Beckerman BS, Jerrett M, Martin RV, van Donkelaar A, Ross Z, Burnett RT. Application of the deletion/substitution/addition algorithm to selecting land use regression models for interpolating air pollution measurements in California. Atmos Environ. 2013;77:172–7.
38. Amini SM, Parmeter CF. Bayesian model averaging in R. J Econ Soc Meas. 2011;36:253–87.
39. Fragoso TM, Louzada Neto F. Bayesian model averaging: a systematic review and conceptual classification. arXiv 2015:1509.08864.
40. Bobb JF, Dominici F, Peng RDA. Bayesian model averaging approach for estimating the relative risk of mortality associated with heat waves in 105 US cities. Biometrics. 2011;67:1605–16.
41. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc B. 1996;58:267–88.
42. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc B. 2005;67:301–20.
43. Dai L, Koutrakis P, Coull BA, Sparrow D, Vokonas PS, Schwartz JD. Use of the adaptive LASSO method to identify PM2.5 components associated with blood pressure in elderly men: the Veterans Affairs Normative Aging Study. Environ Health Perspect. 2016;124:120–5.
44. Lenters V, Portengen L, Rignell-Hydbom A, Jönsson BAG, Lindh CH, Piersma AH, 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. Environ Health Perspect. 2016;124:365–72.
45. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. J R Stat Soc B. 2011;73 Part 3:273–82.
46. Lim M, Hastie T. Learning interactions via hierarchical group-lasso regularization. J Comput Graph Stat. 2015;24:627–54.
47. Huang H. Controlling the false discoveries in LASSO. Biometrics. 2017; https://doi.org/10.1111/biom.12665.
48. Liquet B, Bottolo L, Campanella G, Richardson S, Chadeau-Hyam M. R2GUESS: a graphics processing unit-based R Package for Bayesian variable selection regression of multivariate responses. J Stat Softw. 2016;69:2.
49. MacLehose RF, Dunson DB, Herring AH, Hoppin JA. Bayesian methods for highly correlated exposure data. Epidemiology. 2007;18:199–207.
50. Hill SM, Neve RM, Bayani N, Kuo WL, Ziyad S, Spellman PT, et al. Integrating biological knowledge into variable selection: an empirical Bayes approach with an application in cancer biology. BMC Bioinformatics. 2012;13:94.
51.• Bobb JF, Valeri L, Claus Henn B, Christiani DC, Wright RO, Mazumdar M. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics. 2015;16:493–508. **This study provides a thorough description of BKMR method.**
52. Wolpert DH. The lack of a priori distinctions between learning algorithms. Neural Comput. 1996;8:1341–90.
53. Steinley D. K-means clustering: a half-century synthesis. Br J Math Stat Psychol. 2006;59:1–34.
54. Fahad A, Alshatri N, Tari Z, Alamri A, Khalil I, Zomaya AY, et al. A survey of clustering algorithms for big data: taxonomy and empirical analysis. IEEE Trans Emerg Top Comput. 2014;2:267–79.
55. Ljungman PL, Wilker EH, Rice MB, Austin E, Schwartz J, Gold DR, et al. The impact of multi-pollutant clusters on the association between fine particulate air pollution and microvascular function. Epidemiology. 2016;27:194–201.
56. Lee DH, Steffes MW, Sjödin A, Jones RS, Needham LL, Jacobs DR Jr. Low dose of some persistent organic pollutants predicts type 2 diabetes: a nested case-control study. Environ Health Perspect. 2010;118:1235–42.
57. Pumarega J, Gasull M, Lee DH, López T, Porta M. Number of persistent organic pollutants detected at high concentrations in blood samples of the United States population. PLoS One. 2016;11:e0160432.
58. Lee DH, Lee IK, Song K, Steffes M, Toscano W, Baker BA, et al. A strong dose-response relation between serum concentrations of persistent organic pollutants and diabetes: results from the National Health and Examination Survey 1999-2002. Diabetes Care. 2006;29:1638–44.
59. Molitor J, Papathomas M, Jerrett M, Richardson S. Bayesian profile regression with an application to the National Survey of Children's Health. Biostatistics. 2010;11:484–98.
60. Wang Y, Miller DJ, Clarke R. Approaches to working in high-dimensional data spaces: gene expression microarrays. Br J Cancer. 2008;98:1023.

61. Papathomas M, Molitor J, Richardson S, Riboli E, Vineis P. Examining the joint effect of multiple risk factors using exposure risk profiles: lung cancer in nonsmokers. Environ Health Perspect. 2011;119:84–91.

62.• Pirani M, Best N, Blangiardo M, Liverani S, Atkinson RW, Fuller GW. Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. Environ Int. 2015;79:56–64. **Pirani and colleagues propose a Bayesian approach to analyze the impact of multiple particle metrics on daily mortality. The method enables a better understanding of hidden structures in multi-pollutant health effects and provides a tool to assess the changes in health effects from various policies to control the ambient particle matter mixtures.**

63. Bauer DJ, Shanahan MJ. Modeling complex interactions: person-centered and variable-centered approaches. In: Little TD, Bovaird JA, Card NA, editors. Modeling contextual effects in longitudinal studies. Mahwah: Lawrence Erlbaum Associates; 2007. p. 255–83.

64. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods. 2009;14:323–48.

65. Gass K, Klein M, Chang HH, Flanders WD, Strickland MJ. Classification and regression trees for epidemiologic research: an air pollution example. Environ Health. 2014;13:17.

66. Molinaro AM, Lostritto K, van der Laan M. partDSA: deletion/substitution/addition algorithm for partitioning the covariate space in prediction. Bioinformatics. 2010;26:1357–63.

67. Lampa E, Lind L, Lind PM, Bornefalk-Hermansson A. The identification of complex interactions in epidemiology and toxicology: a simulation study of boosted regression trees. Environ Health. 2014;13:57.