


Invited Perspective: Challenges and Opportunities for Missing Data in the Context of Environmental Mixture Methods

Stephanie M. Eick^{1,2}  and Anke Hüls^{1,2}

¹Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA

²Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA

<https://doi.org/10.1289/EHP12118>

Refers to <https://doi.org/10.1289/EHP10479>

Humans are exposed to myriad chemicals simultaneously. Based on their sources, many of these chemicals co-occur, leading to high correlations between certain chemical concentrations. For example, individuals who live in urban areas may have higher exposures to air pollutants relative to residents of rural areas,¹ and those who eat fish and red meat may have elevated levels of PFAS and mercury relative to those who eat these foods less frequently.^{2–4} Identifying behavioral patterns that reflect unique groups of multipollutant exposures may be useful as an intervention aimed at exposure reduction because regulations and policy changes to remove specific chemicals from commerce take years. Ascertaining upstream sources of chemical exposures is also useful for risk assessment because this may allow us to identify the most vulnerable populations.

To date, the vast majority of environmental epidemiology studies have assessed the effects of single chemicals one at a time, and regulatory guidelines set by the U.S. Environmental Protection Agency generally focus on a few specific chemicals, as opposed to considering chemical classes.^{5,6} A problem with these approaches is that they fail to consider coexposure to other pollutants, which may produce additive or synergistic health effects. To account for highly correlated coexposures, environmental epidemiologists have developed environmental mixture methods,^{7–9} which allow us to estimate the impact of exposure to a group of pollutants on adverse health outcomes.

Although exposure assessment and mixture methods have provided novel insights, challenges remain. Data obtained from laboratory-based exposure assessment often include observations below the limit of detection (LOD), leading to incomplete and missing data, which pose a challenge from a statistical standpoint because researchers are left with little to no information regarding the concentration. Given that the actual value lies somewhere between 0 and just below the LOD and that most current mixture methods cannot account for missing data, researchers commonly impute values below the LOD with $\text{LOD}/\sqrt{2}$ to retain the maximum possible sample size for downstream analyses. However, imputing with a fixed value falsely assumes that all values below the LOD are equal, making it difficult to identify patterns associated with low levels of exposure.

In their new study,¹⁰ Gibson et al. adapted principal component pursuit (PCP), a robust method for dimensionality reduction and pattern identification, to accommodate missing data

and values below the LOD. An important strength of this method is that it allows for the inclusion of missing values, such as those below the LOD, in a way beyond just imputing them as $\text{LOD}/\sqrt{2}$. To evaluate its performance, the authors compared their method, PCP-LOD, with traditional principal component analysis (PCA) with values $<\text{LOD}$ imputed as $\text{LOD}/\sqrt{2}$ in simulations and in a real-world application to data from the National Health and Nutrition Examination Survey (NHANES).¹¹

In simulations, PCP-LOD generally outperformed PCA (e.g., PCP-LOD recovered a higher percentage of the true number of patterns) when the percentage of observations below the LOD was $<50\%$ and in scenarios in which there was either low Gaussian noise or there were both low Gaussian noise and sparse events.¹⁰ Further, PCP-LOD largely outperformed PCA when 16 chemicals were included in the mixture, but performance decreased when the number of chemicals in the mixture increased to 48. In the application to NHANES data, PCP-LOD produced results similar to those of PCA when applied to a mixture of 21 chemicals (including dioxins, furans, and polychlorinated biphenyls) with $>50\%$ above the LOD.¹⁰

The simulation findings by Gibson et al. suggest there could be specific situations in which PCP-LOD would be preferred in practice. However, more studies are needed to demonstrate the advantage of PCP-LOD in real-world data applications because the authors' application to NHANES data indicates that PCP-LOD and PCA provide comparable estimates.

As the field moves toward incorporating information on increasingly large numbers of chemicals into mixture models, handling large amounts of missing data (e.g., concentrations $<\text{LOD}$) is becoming increasingly important. Current methodological approaches do not have high accuracy in situations with large amounts of missing data, which leads to difficulty in assessing the health effects of emerging chemicals that are readily detected only in small percentages of the population. In addition to PCP-LOD, there is a need for extensions accommodating missing data in other mixture methods, particularly those that aim to identify the most dangerous constituent chemicals or estimate the joint effects of the whole mixture. Using mixture methods is crucial for a holistic risk assessment and prevention of disease because these approaches provide an opportunity to answer research questions related to the health effects of simultaneous exposure to numerous chemicals.

References

- Landrigan PJ, Fuller R, Acosta NJR, Adeyi O, Arnold R, Basu NN, et al. 2018. The Lancet Commission on pollution and health. *Lancet* 391(10119):462–512, PMID: 29056410, [https://doi.org/10.1016/S0140-6736\(17\)32345-0](https://doi.org/10.1016/S0140-6736(17)32345-0).
- Christensen KY, Raymond M, Blackowicz M, Liu Y, Thompson BA, Anderson HA, et al. 2017. Perfluoroalkyl substances and fish consumption. *Environ Res* 154:145–151, PMID: 28073048, <https://doi.org/10.1016/j.envres.2016.12.032>.
- Tian Y, Zhou Y, Miao M, Wang Z, Yuan W, Liu X, et al. 2018. Determinants of plasma concentrations of perfluoroalkyl and polyfluoroalkyl substances in pregnant women from a birth cohort in Shanghai, China. *Environ Int* 119:165–173, PMID: 29958117, <https://doi.org/10.1016/j.envint.2018.06.015>.

Address correspondence to Stephanie M. Eick. Email: stephanie.marie.eick@emory.edu

The authors declare they have no conflicts of interest.

Received 8 September 2022; Revised 16 September 2022; Accepted 21 September 2022; Published 23 November 2022.

Note to readers with disabilities: *EHP* strives to ensure that all journal content is accessible to all readers. However, some figures and Supplemental Material published in *EHP* articles may not conform to 508 standards due to the complexity of the information being presented. If you need assistance accessing journal content, please contact ehpsubmissions@niehs.nih.gov. Our staff will work with you to assess and meet your accessibility needs within 3 working days.

4. Castaño A, Cutanda F, Esteban M, Pärt P, Navarro C, Gómez S, et al. 2015. Fish consumption patterns and hair mercury levels in children and their mothers in 17 EU countries. *Environ Res* 141:58–68, PMID: [25667172](https://pubmed.ncbi.nlm.nih.gov/25667172/), <https://doi.org/10.1016/j.envres.2014.10.029>.
5. U.S. EPA (U.S. Environmental Protection Agency). Drinking Water Health Advisories for PFOA and PFOS. <https://www.epa.gov/ground-water-and-drinking-water/drinking-water-health-advisories-pfoa-and-pfos> [accessed 27 January 2022].
6. U.S. EPA. Phthalates Action Plan. https://www.epa.gov/sites/default/files/2015-09/documents/phthalates_actionplan_revised_2012-03-14.pdf [accessed 7 September 2022].
7. Keil AP, Buckley JP, O'Brien KM, Ferguson KK, Zhao S, White AJ. 2020. A quantile-based g-computation approach to addressing the effects of exposure mixtures. *Environ Health Perspect* 128(4):47004, PMID: [32255670](https://pubmed.ncbi.nlm.nih.gov/32255670/), <https://doi.org/10.1289/EHP5838>.
8. Carrico C, Gennings C, Wheeler DC, Factor-Litvak P. 2015. Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *J Agric Biol Environ Stat* 20(1):100–120, PMID: [30505142](https://pubmed.ncbi.nlm.nih.gov/30505142/), <https://doi.org/10.1007/s13253-014-0180-3>.
9. Bobb JF, Valeri L, Claus Henn B, Christiani DC, Wright RO, Mazumdar M, et al. 2015. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* 16(3):493–508, PMID: [25532525](https://pubmed.ncbi.nlm.nih.gov/25532525/), <https://doi.org/10.1093/biostatistics/kxu058>.
10. Gibson EA, Zhang J, Yan J, Chillrud L, Benavides J, Nunez Y. 2022. Principal component pursuit for pattern identification in environmental mixtures. *Environ Health Perspect* 130(11):117008, <https://doi.org/10.1289/EHP10479>.
11. Zipf G, Chiappa M, Porter KS, Ostchega Y, Lewis BG, Dostal J. 2013. National Health and Nutrition Examination Survey: plan and operations, 1999-2010. *Vital Health Stat* 1 (56):1–37, PMID: [25078429](https://pubmed.ncbi.nlm.nih.gov/25078429/).