UC Davis
UC Davis Previously Published Works

## Title
Methodological challenges in spatial and contextual exposome-health studies

## Permalink
https://escholarship.org/uc/item/9258w5p7

## Journal
Critical Reviews in Environmental Science and Technology, 53(7)

## ISSN
1064-3389

## Authors
Hu, Hui
Liu, Xiaokang
Zheng, Yi
et al.

## Publication Date
2023-04-03

## DOI
10.1080/10643389.2022.2093595

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

*Crit Rev Environ Sci Technol*. Author manuscript; available in PMC 2023 May 02.

Published in final edited form as:

*Crit Rev Environ Sci Technol*. 2023 ; 53(7): 827–846. doi:10.1080/10643389.2022.2093595.

# Methodological Challenges in Spatial and Contextual Exposome-Health Studies

**Hui Hu**[1,*], **Xiaokang Liu**[2], **Yi Zheng**[1], **Xing He**[3], **Jaime Hart**[1,4], **Peter James**[4,5], **Francine Laden**[1,4,6], **Yong Chen**[2,*], **Jiang Bian**[3,*]

[1]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA

[2]Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[3]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA

[4]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

[5]Department of Population Medicine, Harvard Pilgrim Healthcare, Boston, Massachusetts, USA

[6]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

## Abstract

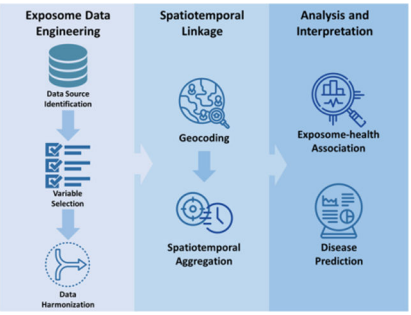
The concept of the exposome encompasses the totality of exposures from a variety of external and internal sources across an individual's life course. The wealth of existing spatial and contextual data makes it appealing to characterize individuals' external exposome to advance our understanding of environmental determinants of health. However, the spatial and contextual exposome is very different from other exposome factors measured at the individual-level as spatial and contextual exposome data are more heterogenous with unique correlation structures and various spatiotemporal scales. These distinctive characteristics lead to multiple unique methodological challenges across different stages of a study. This article provides a review of the existing resources, methods, and tools in the new and developing field for spatial and contextual exposome-health studies focusing on four areas: (1) data engineering, (2) spatiotemporal data linkage, (3) statistical methods for exposome-health association studies, and (4) machine- and deep-learning methods to use spatial and contextual exposome data for disease prediction. A critical analysis of the methodological challenges involved in each of these areas is performed to identify knowledge gaps and address future research needs.

## Graphical Abstract

[*]**Address correspondence to** Hui Hu, Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, 401 Park Drive, Room 301.48, Boston, Massachusetts 02215; Yong Chen, Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; and Jiang Bian, Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, 2004 Mowry Road, Gainesville, FL 32610.

**Conflicts of Interest:** The authors declare they have nothing to disclose.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript




**Keywords**

Exposome; External exposome; Spatial and contextual exposome; Statistical method; Machine learning; Deep learning

## 1. Introduction

The concept of the exposome was first proposed in 2005, and "*encompasses all life-course environmental exposures from the prenatal period onwards, complementing the genome*" (Wild, 2005, 2012). The exposome can be categorized into internal (e.g., metabolism) and external (i.e., including specific external factors such as pollutants and general external factors such as social capital) factors (Vrijheid, 2014). While a large number of exposome-health studies have been conducted over the past decade, the majority of them focused on the internal exposome. The external exposome is still a new and developing field, and very few external exposome studies have been conducted (Zheng et al., 2020). This is not surprising since exposome-health studies require well-characterized historical exposures before disease onset, as well as large sample sizes to ensure sufficient statistical power given the domain-agnostic approach often used. The internal exposome can usually be measured using biospecimens that allows researchers to retrospectively assess individuals' exposures from long time ago; on the other hand, measurement methods for the external exposome are more heterogeneous. As shown in Figure 1, a wide variety of tools and information can be leveraged to characterize external exposome factors, such as questionnaires and surveys (e.g., self-reported physical activity and diet), smartphone-based sensors (e.g., accelerometer-based physical activity measures), personal environmental monitors (e.g., passive silicone wristband-based polycyclic aromatic hydrocarbon measures), environmental specimens (e.g., house dust), and spatial and contextual information (e.g., ambient temperature) (Turner et al., 2017). However, most of these methods are not able to accurately characterize historical exposures decades ago (e.g., prior to 2000) because historical data on these exposures are scarce.

One exception is spatial and contextual information. There is a wealth of existing historical spatial and contextual data (Table 1 shows the examples of publicly available data sources) which can be linked to individuals based on their geolocations (e.g., residential address history, Global Positioning System [GPS]-derived time activity patterns) that are widely available in many studies and/or can also be obtained from public-record databases (Wheeler & Wang, 2015). Because of this, the few existing external exposome studies have

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

predominantly focused on exposures measured using spatial and contextual data (Hu, Zhao, Savitz, et al., 2020; Hu et al., 2021; Lynch et al., 2017; Mooney et al., 2017; Nieuwenhuijsen et al., 2019; Siroux et al., 2018; Vrijheid et al., 2020), instead of other methods (e.g., questionnaires and surveys, smartphone-based sensors, personal environmental monitors, and environmental specimens). However, the spatial and contextual exposome also has several unique methodological challenges due to its difference from the internal exposome and other external exposome factors (e.g., chemicals and toxicants based on personal environmental monitors). First, while routinely performed for single spatial and contextual factors in environmental health studies, data engineering and linkages are challenging when the exposome is of interest (i.e., a large number of exposome factors needs to be considered to account for the totality of the exposome) due to scalability issues. In addition, the spatial and contextual exposome are usually assessed using data from different sources, which are often heterogenous with (1) unique correlation structures (i.e., factors from the same data source are more likely to be highly correlated) (Hu, Zhao, Savitz, et al., 2020; Zheng et al., 2020), and (2) different spatiotemporal scales, leading to different aggregations and spatiotemporal linkages and potentially different exposure-health associations (e.g., the modifiable areal and temporal unit problems) (Cheng & Adepeju, 2014; Jelinski & Wu, 1996). While many statistical methods have been developed for exposome-health studies in the past decades (Agier et al., 2016), they have all focused on the exposome factors that can be directly measured at the individual-level, and it is largely unknown how these methods perform in spatial and contextual exposome-health studies. Furthermore, given the wealth of spatial and exposome data and relatively low cost to append them to individuals' geolocation data, it is appealing to leverage them in disease prediction (Hu, Zhao, Bian, et al., 2020). Machine and deep learning approaches are increasingly used to account for the high dimensionality of and high-order interactions among spatial and exposome factors and their nonlinear relationships with the outcomes of interest for disease prediction (Feng & Jiao, 2021; Hu, Zhao, Bian, et al., 2020; Mohnen et al., 2019). However, challenges exist to fully leverage and preserve the rich spatiotemporal structures in spatial and contextual exposome data using off-the-shelf machine and deep learning models.

In this review, we aimed to describe existing resources, methods, and tools available and major methodological challenges for spatial and contextual exposome-health studies, focusing on four major areas: (1) data engineering, (2) spatiotemporal data linkage, (3) statistical methods for exposome-health association studies, and (4) machine- and deep-learning methods to use spatial and contextual exposome data for disease prediction.

## 2. Engineering of the spatial and contextual exposome data

The first step in spatial and contextual exposome-health studies is data engineering, which involves three major steps (Figure 2): data source identification, variable selection, and data harmonization. The unique characteristics of the spatial and contextual exposome pose challenges to each of these three steps.

The spatial and contextual exposome can be characterized using data from a variety of data sources (Hu, Zhao, Savitz, et al., 2020; Hu et al., 2021; Lynch et al., 2017; Mooney et al., 2017; Nieuwenhuijsen et al., 2019; Siroux et al., 2018; Vrijheid et al., 2020). Table 1

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

shows the examples of publicly available spatial and contextual exposome data sources. It is not rare to have multiple data sources available to assess a specific spatial and contextual exposome construct. For example, air pollutants such as fine particulate matter ($PM_{2.5}$) and ozone can be assessed using stationary monitored data such as the U.S. Environmental Protection Agency (USEPA) Air Quality System (AQS, Lamsal et al., 2015), remote sensing data such as Sentinel-5P (Verhoelst et al., 2021), and modelled data from USEPA's Fused Air Quality Surface Using Downscaling (FAQSD, Berrocal et al., 2012), the Atmospheric Composition Analysis Group (ACAG, van Donkelaar et al., 2019), the Center for Air, Climate, and Energy Solutions (CACES, Kim et al., 2020), and many other sources (Di et al., 2017; Hart et al., 2009; Kloog et al., 2015; Lee et al., 2015; Logue et al., 2011; Sampson et al., 2013; Yanosky et al., 2014). Similarly, there are multiple indices developed to measure walkability such as the Walk Score (Carr et al., 2010), USEPA's National Walkability Index (Thomas & Zeller, 2021), and a few others (James et al., 2015; Rundle et al., 2019). For specific spatial and contextual exposome constructs of interest which can be assessed using multiple data sources, the decisions are usually based on their corresponding spatiotemporal coverage and scale. For example, studies focusing on acute impacts of air pollution usually prefer data sources with daily measures, while data sources with annual estimates work well for studies on long-term impacts. Nevertheless, for certain spatial and contextual exposome factors, there are multiple options with similar spatiotemporal coverage and scale available. The selection of data sources in this circumstance is usually subjective: while the same exposome factor from different data sources may have varying degree of bias by geographic areas and time (Ma et al., 2019; Mukhopadhyay & Sahu, 2018; van Donkelaar et al., 2019), there is not a framework or tool to enable objective selections of data sources. For traditional studies focusing on single or a small subset of environmental exposures, researchers' knowledge of the specific exposures is usually used to guide the selection of data sources and sensitivity analyses using different data sources are commonly conducted within and between research groups. However, this approach is increasingly challenging in spatial and contextual exposome studies. When adopting the exposome concept, individual researchers may lack the expertise to subjectively select the best data source for certain exposures, and the large number of exposome factors considered makes it infeasible to conduct sensitivity analyses due to the substantially increased computational burden and difficulties in interpretations. A potential way to address this challenge is to establish reference spatial and contextual exposome databases. Reference databases have long been used in other fields such as image classifications (Deng, 2012) to compare performance and guide selection of different methods for specific studies. There is an urgent need for the field to start the efforts to establish reference databases with gold-standard measures on different spatial and contextual exposome constructs across different geographic areas and time periods to provide guidance when multiple data sources exist to assess a specific exposure.

After the data sources are selected, the next step in data engineering is to determine the variables from each data source to include. A data source may include multiple variables measuring similar spatial and contextual exposome constructs. For example, data from the American Community Survey (ACS) contain thousands of variables characterizing contextual-level social environment. Some of the existing exposome-wide association

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

studies (ExWAS) included all of these individual variables in the analyses (Hu, Zhao, Savitz, et al., 2020; Mooney et al., 2017), while others performed dimension reduction and used indices such as the neighborhood deprivation index and index of concentration at the extremes (Andrews et al., 2020; Krieger et al., 2018, 2016; Messer et al., 2006). Similarly, many air pollution studies examined associations between health outcomes and individual air pollutants (Loxham et al., 2019; Mannucci et al., 2015), while others used indices such as the Air Quality Index to assess multiple air pollutants (Olalekan et al., 2018). There are different assumptions and hypotheses underlying these two approaches, which lead to substantially different numbers of variables to be included in subsequent analyses. The approach to consider all individual variables seeks to understand the impact of each variable separately with the assumption that these variables represent different constructs and consequently leads to many more variables to be included in the analyses compared with the other approach (i.e., $10^4$-$10^5$ vs. $10^3$), which assumes that these variables matter in aggregate, but quantifying their individual contributions is difficult or not of interest. This dramatic difference in the number of variables to be included has a large impact on the analyses: the p value cut points determined to account for multiple testing are very different and thus certain variables considered statistically significant in the second approach may no longer be significant when using the first approach. Interestingly, for many spatial and contextual exposome factors, it is possible to generate multiple variables with the exposures aggregated at different spatiotemporal windows. For example, social environment factors are available in ACS at different spatial (e.g., county, census tract, census block group, 5-digit zip code tabulation area) and temporal levels (e.g., 1-year, 3-year, and 5-year), and many of the variables from the Food Access Research Atlas (FARA) are the same exposures aggregated at spatial buffers with different sizes (e.g., percentages of low access population that are children at 0.5, 1, 10, or 20 miles). We will revisit this issue and its implications for statistical analyses and interpretations in section 4 (Cheng & Adepeju, 2014; Jelinski & Wu, 1996). The lack of consensus on variable selection in spatial and contextual exposome data engineering may dramatically impact the results of downstream studies, and therefore, it is critical for the field to make more efforts (e.g., develop ontology-based approaches (Heacock et al., 2022; Zhang et al., 2021)) to not only standardize the variables selected and data sources but also the approaches of making these choices.

The last step in spatial and contextual exposome data engineering is data harmonization. As shown in Table 1, data from different sources are very heterogeneous – with different formats, data structures, and spatiotemporal scales – making harmonization challenging. Data on some exposome factors may only be available for certain years. For example, air toxicants data from the USEPA's National Air toxics Assessment (NATA) are only available in 1996, 1999, 2002, 2005, 2011, and 2014 (Logue et al., 2011), and FARA data are only available in 2010, 2015, and 2019. As a result, interpolation is often needed to assign these exposures for other years. While many spatiotemporal interpolation methods are available (Li & Heap, 2014; Liu et al., 2021; Susanto et al., 2016), there is not a standardized, validated method and thus an urgent need to evaluate and establish the optimal method(s) for different spatial and contextual exposome factors.

In summary, it is critical for the field to start making efforts to tackle challenges in data engineering and to develop infrastructure to support streamlined and standardized

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

measures of the spatial and contextual exposome. As discussed previously (Zhang et al., 2021), the development of semantic standards of the spatial and contextual exposome is critically needed to provide an unambiguous and consistent understanding of the variables in heterogeneous data sources and to explicitly express the context of the variables as well as the relationships among them. Further, like in any other domain, reporting standards to capture the steps of the data engineering process from each data source and variable selection choices to the interpolation methods used for different variables are critical for transparency and reproducibility.

## 3. Spatiotemporal linkages of spatial and contextual exposome data to individuals

There are two major steps to link spatial and contextual exposome data to individuals (Figure 2): (1) when GPS locations are not available, geocoding is performed to derive individuals' geolocations from addresses, which are (2) then used to spatiotemporally link exposome data. Geocoding also has methodological challenges such as positional accuracy and privacy issues, which have been extensively discussed in the literature (Brokamp, 2018; Christen, 2006; Harris & Delcher, 2019; Kounadi et al., 2013) and are not our focus. In this section, we will focus on the second step – the spatiotemporal data linkage – which has unique scalability challenges in spatial and contextual exposome studies.

Two approaches are commonly used to assign spatial and contextual exposome data to individuals spatially. The first is the buffer-based approach (Kwan, 2012), which calculates area- or population-weighted averages to generate individuals' exposures based on preselected spatial buffers surrounding individuals' geolocations, and the other approach preserves the original spatial scale in the data source and directly assigns exposures from the geographic units corresponding to individuals' geolocations. Existing spatial and contextual exposome studies often use the buffer-based approach for all exposome factors (Hu, Zhao, Savitz, et al., 2020; Mooney et al., 2017; Nieuwenhuijsen et al., 2019) due to the lack of statistical methods available to assess multi-level exposome-health associations, a major methodological challenge that we will revisit in section 4. Lastly, temporal aggregation is performed to generate exposures in different time windows of interest. We further discussed the implications of different spatiotemporal aggregations for downstream analyses in Sections 4 and 5.

There are many packages and tools available to perform spatiotemporal data linkages. As a free software environment with an open source development model (R Core Team, 2013), R has an increasing number of contributed packages to handle spatial and contextual data, such as "sf" (Pebesma, 2018), "sp" (Pebesma & Bivand, 2005), "rgdal" (Bivand et al., 2010), "rgeos" (Bivand & Rundel, 2017), "raster" (Hijmans, 2021), and "exactextractr" (Baston et al., 2021). Python – as a general programing language becoming increasingly popular for data science projects – also supports a rich set of libraries for spatiotemporal data linkages. For example, GeoPandas (Jordahl et al., 2021) is widely used for manipulation of geospatial data allowing spatial operations on geometric objects (e.g., creating geometries representing all points within a given distance of each geometric object). For zonal statistics calculation

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and areal-weighted interpolation, several packages are often used along with Geopandas, such as Xarray-Spatial (xarray-spatial Development Team, 2022), Rasterstats (Perry, 2021), and Tobler (Knaap et al., 2021). QGIS (*QGIS*, 2022) and ArcGIS (ESRI, 2022) are also useful tools with good integrations with both R and Python. Other emerging tools have also been developed to facilitate spatiotemporal data linkages, such as the Decentralized Geomarker Assessment for Multi-Site Studies (DeGAUSS), a container-based application that performs both geocoding and exposure assessment (Brokamp, 2018).

While the buffer-based approach makes it possible to take advantage of the wealth of statistical methods available to assess individual-level exposome-health associations, it is more computationally challenging in the exposome setting given the large number of exposures and sample size. Parallelization, a type of computing paradigm designed to have multiple processes carrying out the computation simultaneously to speed up the processing time, is possible in many of the existing packages and tools to address some of the large dataset issues. For example, spatiotemporal data linkages in R and Python can be scaled with parallel computing techniques such as using the “parallel” package and Dask (Rocklin, 2015) in R and Python, respectively. However, as in-memory processing is implemented in many of these parallelization packages, the scalability is still a concern for extremely large datasets, which is common for spatiotemporal linkages in spatial and contextual exposome studies. There are several potential alternatives. For example, Google Earth Engine (GEE) is increasingly used (Gorelick et al., 2017). GEE leverages Google’s cloud infrastructure and distributed data processing algorithms such as MapReduce, which makes it extremely fast to handle large datasets such as remote sensing data. However, as a cloud infrastructure, GEE requires uploads of geolocation data for spatiotemporal data linkages, leading to potential privacy concerns as all sub-state geographic information is often considered as protected health information under the Health Insurance Portability and Accountability Act of 1996 (HIPAA) (Cohen & Mello, 2018). PostGIS is another alternative (The PostGIS Development Group, 2019), which provides parallel processing in spatial queries that can be further accelerated using PG-Strom (PG-Strom Development Team, 2021), a GPU acceleration extension. However, there is a steep learning curve with complex configurations required.

Despite the various packages and tools available, each of them has its limitations. Future efforts are warranted to develop tools specifically to facilitate spatiotemporal data linkages for spatial and contextual exposome studies.

## 4. Statistical methods for spatial and contextual exposome-health studies

Many statistical methods have been applied in exposome-health association studies to identify risk and protective factors from the exposome for a variety of health outcomes. Some commonly used variable selection methods perform well when model assumptions are satisfied and can be easily scaled to large-scale problems, but their performance can be affected by the special structure of exposome covariates and their complex effects on the health outcome. In particular, the unique challenges in exposome-health association studies are mainly manifested by the high correlation among exposures, with possible nonlinear relationships and nonadditive effects. When the model assumptions of some commonly used methods are violated, statistical approaches tailored to solve these issues

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

are needed to recover the underlying dependence relationship between exposures and health outcomes. In this section, we review representative approaches used in exposome-health association studies, followed by a discussion on major methodological challenges to apply these approaches to spatial and contextual exposome data. Table 2 shows a summary of the representative statistical methods. Details of model fitting and parameter specification can be found in reference literatures. Thus, our discussions here focus on the intuition underlying the various methods and their applicability under different use case scenarios. Continuous outcomes are mainly considered in this section while most approaches are also applicable to exponential family distributed variables.

Let us first consider an ideal situation where the assumption of a linear relationship between the outcome of our interest $Y \in R$ and environmental covariates $X_1, \ldots, X_p$ is satisfied. In this situation, the environment-wide association study (EWAS, Patel et al., 2010) or exposome-wide association study (ExWAS, Juarez & Matthews-Juarez, 2018) can be used to assess influential factors on the outcome. These methods measure the association between the outcome and covariates by regressing $Y$ on each $X_j$ separately, and then applying multiple adjustment techniques to control the inflated type I error caused by simultaneously conducting hypothesis testing for all $p$ covariates. Variable selection is conducted by examining the two-sided p values, and a covariate is claimed to be statistically significantly related to the outcome if its related p value is less than a threshold value after multiple adjustment. This method has a high sensitivity to detect predictive exposures, but it ignores the potential joint functioning mechanism of covariates on the outcome and has a high false discovery rate even after doing multiple adjustment (Agier et al., 2016).

To take into account the joint effect of covariates, the following multiple linear regression model is more appropriate

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon, \tag{1}$$

where $\beta_j$s are the regression coefficients and $\epsilon$ is the random error. In this formulation, an environmental exposure $X_j$'s effect on the health outcome is the part that is only attributable to $X_j$ after accounting for other covariates' effects. To do variable selection, penalized regression is commonly adopted for simultaneous variable selection and estimation. Specifically, variable selection is facilitated by adding penalties on the magnitude of coefficients to push noise signals which are often embodied by small coefficient values to be zeros, and therefore rules out irrelevant covariates. For example, lasso regression applies an $l_1$ penalty of the regression coefficients and can achieve variable selection consistency when the irrelevant covariates are "irrepresentable" by predictive covariates (Tibshirani, 1996; Zhao & Yu, 2006). To fulfill different needs, various variants of lasso regression can be applied in exposome-health association studies, e.g., group lasso to conduct group selection (Yuan & Lin, 2006) and adaptive lasso to adjust penalty levels for different exposures (Zou, 2006). Instead of penalized regression, Graphical Unit Evolutionary Stochastic Search (GUESS, Bottolo et al., 2013) can also be applied in large $p$ and small $N$ ($N$ is the sample size) scenarios by applying a Bayesian variable selection technique to select a combination of covariates that achieves the best prediction performance from a total of $2^p$ candidate models. Specifically, the candidate models with large posterior probabilities are retained,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and from the retained models, the involved exposures that have large marginal posterior probabilities of inclusion are selected. Rockova and George (2018) proposed a spike-and-slab lasso regression by introducing a mixture prior of two Laplace distributions on each $\beta_j$ with the spike part heavily concentrating around zero (Bai et al., 2021), and it achieves simultaneous variable selection and estimation using the posterior mode. Bai et al. (2020) further extended this method to handle grouped covariates and proposed the spike-and-slab group lasso regression.

The performance of the above linear regression-based statistical methods can be hampered by high-collinearity among exposures, which is a commonly seen issue in exposome-health association studies (Agier et al., 2016; Hu, Zhao, Savitz, et al., 2020). With the existence of high-collinearity, it is highly possible that the selected variable is the one that is correlated with the truly predictive covariate, or measured with the least amount of error (Agier et al., 2016). To relieve this issue, one choice is to apply the elastic-net regression which combines the $l_1$ and $l_2$ penalties as a weighted average so that strongly correlated covariates are included in or excluded from the model together (Zou & Hastie, 2005). The additionally incorporated $l_2$ penalty assigns almost equal coefficients to highly correlated covariates and avoids the situation that only one variable from the highly correlated covariates is selected while all the others are removed from the model (Zou & Hastie, 2005). The sparse partial least square (sPLS, Chun & Keleş, 2010) regression solves this problem from a different aspect. Instead of identifying a subset of variables that best predict the outcome, sPLS focuses on locating a set of latent variables by iteratively searching for projection directions that project the original set of covariates to directions that are most correlated with the outcome and best explain the variations in covariates as well. Therefore, this supervised dimension reduction procedure is robust to the high collinearity of the covariates. However, it has inferior interpretability than lasso although an $\ell_1$ penalty is imposed on each direction vector to select covariates when forming each latent variable.

All the methods we discussed above assume a linear association between the environmental covariates and the health outcome, which could be violated in exposome-health studies as dose-dependent effect and nonlinear relationships are observed for many environmental exposures (Claus Henn et al., 2010). Under these circumstances, additive models serve as a more flexible alternative to linear regression models. Consider the following additive model

$$Y = \widetilde{Z}\gamma + \sum_{j=1}^{p} f_j(X_j) + \epsilon, \tag{2}$$

where $\widetilde{Z}$ represents the confounding variables and $f_j(X_j)$ models the effect of $X_j$ on $Y$ with no assumption of a specific form. Model (1) can be regarded as a special case of (2) where each function $f_j(X_j)$ admits a linear form. In (2), the relationship between each exposure and the health outcome can be characterized more flexibly by using, e.g., smoothing splines. In general, to avoid overfitting and achieve some sort of smoothness, the penalty $\int f_j''(X_j)^2 dX_j$ is often exploited to remove excessive wiggliness during fitting (Marra & Wood, 2011). However, this penalty often cannot completely exclude an exposure from the model, e.g., for cubic splines there is no penalty on the linear part. Marra and Wood (2011) provided two choices of modifying the penalty $\int f_j''(X_j)^2 dX_j$ to make removing $X_j$ from the model

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

possible in the context of generalized additive models. Briefly speaking, both methods complete the penalty space to be of full rank but with different completing methods. If taking the cubic splines as an example, these methods added a penalty on the linear part to completely exclude the effects of an exposure on the health outcome from the model. They also provided a two-step nonnegative garrote component selection method to shrink the irrelevant model components to be zeros based on the originally obtained smooth function estimates.

The spike-and-slab group lasso method can also be applied in the context of additive models by treating the coefficients of the basis functions corresponding to $f_j(X_j)$ as a group (Bai et al., 2020). Another method for conducting variable selection from a Bayesian aspect is the Bayesian structured additive regression with spike-and-slab priors (BSTARSS, Scheipl et al., 2012). BSTARSS is an extension of the additive model (2) where a special prior specification is used to conduct simultaneous variable and function shape selection and the choice of functions $f_j(\widetilde{X}_i)$ where $\widetilde{X}_i = (X_{i1}, \ldots, X_{ip})^T$, $i = 1, \ldots, N$ is the exposure of the $i$-th subject can be more flexible. For example, $f_j(\widetilde{X}_i)$ can be smooth functions of one or more exposures, Markov random fields, random effects, and interactions between different terms (Scheipl et al., 2012). Specifically, representing each $f_j(\widetilde{X}_i)$ by a linear combination of $d_j$ basis functions as $f_j(\widetilde{X}_i) = \sum_{k=1}^{d_j} \beta_{jk} B_{jk}(\widetilde{X}_i) = \beta_j^T B_j(\widetilde{X}_i)$ with $B_j(\widetilde{X}_i) = \left(B_{j1}(\widetilde{X}_i), \ldots, B_{jd_j}(\widetilde{X}_i)\right)^T$ be the basis function value of the $i$-th subject. For simultaneously selecting variables and deciding their functional forms, let the basis function coefficient be $\beta_j = r_j \xi_j$ with mutually independent components $r_j$ and $\xi_j = \left(\xi_{j1}, \ldots, \xi_{jd_j}\right)^T$. The inclusion or exclusion of the component $f_j(\widetilde{X}_i)$ is decided by $r_j$ for which a spike-and-slab type prior is imposed, and the determination of the function shape is further controlled by all elements of $\xi_j$. Specifically, the spike-and-slab prior is imposed on the hyper-variance of $r_j$ with a narrow spike around zero and a slab of inverse Gamma distribution, and the posterior mixture weights for the spike component is used as the posterior probability of excluding $f_j(\widetilde{X}_i)$ from the model. BSTRASS can be implemented by R package spikeSlabGAM (Scheipl, 2011).

Although additive models relax the linear assumption imposed by model (1) and terms that capture more complex relationships between environmental exposures and the health outcome (e.g., the interaction terms) can be added into model (2) when necessary (Bai et al., 2020), there are more flexible methods that require no such specification. Bayesian kernel machine regression (BKMR) is a non-parametric method that commonly uses a Gaussian kernel machine to characterize the functioning mechanism of multiple covariates on the outcome (Bobb et al., 2015). Using $(Y_i, \widetilde{Z}_i, \widetilde{X}_i)$, $i = 1, \ldots, N$ to denote the observation for the $i$-th subject, the model is written as

$$Y_i = \widetilde{Z}_i^T \gamma + h(\widetilde{X}_i) + \epsilon_i$$

where $h(\widetilde{X}_i)$ is the target multivariate exposure-response function. Based on the Gaussian kernel machine, the functions $\left(h(\widetilde{X}_1), \ldots, h(\widetilde{X}_N)\right) \sim N(0, \tau \mathrm{K})$ where $\tau$ is a smoothing parameter and $K \in R^{N \times N}$ is the kernel matrix whose $(i_1, i_2)$-th element is in the form of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the augmented Gaussian kernel function $K\left(\widetilde{X}_{i_1}, \widetilde{X}_{i_2}; r\right) = \exp\left(-\sum_{j=1}^{p} r_j (X_{i_1 j} - X_{i_2 j})^2\right)$ with the smoothing parameters $r = (r_1, \ldots, r_p)$. As for the effects of the kernel matrix, we provide an intuitive explanation here. Firstly, the form of $K(\widetilde{X}_{i_1}, \widetilde{X}_{i_2}; r)$ directly transforms the similarity between exposures $\widetilde{X}_{i_1}$ and $\widetilde{X}_{i_2}$ of two subjects into the correlation strength between their corresponding exposure-response functions $h\left(\widetilde{X}_{i_1}\right)$ and $h\left(\widetilde{X}_{i_2}\right)$. Consider the following two extreme examples: when two subjects have exactly the same exposures, i.e., $\widetilde{X}_{i_1} = \widetilde{X}_{i_2}$, we have $K\left(\widetilde{X}_{i_1}, \widetilde{X}_{i_2}; r\right) = 1$ which means perfectly positively correlated $h\left(\widetilde{X}_{i_1}\right)$ and $h\left(\widetilde{X}_{i_2}\right)$. On the contrary, we have $K\left(\widetilde{X}_{i_1}, \widetilde{X}_{i_2}; r\right) = 0$ when $\sum_{j=1}^{p} r_j (X_{i_1 j} - X_{i_2 j})^2 = \infty$, i.e., when two subjects have totally different exposures, knowing one's health outcome cannot give any information of another's health outcome. Secondly, applying the kernel machine successfully allows for the nonlinear, nonadditive and multivariate functioning mechanism of exposures on the health outcome, and thus provides BKMR high flexibility in modeling. Finally, the parameter $r_j$ controls the weights when measuring the distance between $\widetilde{X}_{i_1}$ and $\widetilde{X}_{i_2}$, and when $r_j = 0$ the $j$-th exposure plays no role in the exposure-response function. Based on this intuition, BKMR imposes a spike-and-slab prior that comprises a spike part with a point mass at zero and a gamma density as the slab part on each $r_j$ to facilitate variable selection. Once $r_j$ has a large posterior probability to be nonzero, the $j$-th exposure is kept in the model. After specifying proper prior distributions on the unknown parameters in the model, BMKR is fitted by Markov Chain Monte Carlo (MCMC). The package bkmr in R can be used to fit the model (Bobb, 2022).

There are some other useful Bayesian non-parametric methods that provide a flexible modeling framework. For example, Bayesian additive regression trees (BART, Chipman et al., 2010) is a nonparametric ensemble method that uses the summation of a set of trees to model the outcome. Each tree consists of some nonterminal binary decision rules and terminal nodes, and each subject is assigned a leaf value at the terminal node. For the choice of the splitting variables and the related splitting values, uniform priors are used to allow all covariates to have the same probability to be selected thus different trees may represent the effects from different covariates. As for the tree size and leaf values at terminal nodes, to weaken the influence from any individual tree, the priors are set to prefer shallow trees with fewer splits with leaf values be centered around the mean value of the outcome. To understand this setting, suppose each tree has only one or two splitting variables and each different tree is built based on different covariates, then the summation of these trees plays a similar role as a model contains individual covariates and second-order interaction terms among them. Similarly, higher-order interaction terms can be built into the model by growing deeper tree structures. Therefore, BART allows both the nonlinear and nonadditive functioning effect of covariates to be modeled. Variable selection in BART is achieved by comparing the variable inclusion proportion to a pre-specified threshold (Bleich et al., 2014), and the R package bartMachine facilitates the application of BART in association studies (Kapelner & Bleich, 2016).

Most of the above methods can be used to examine the association between an individual exposure on a health outcome while keeping all other exposures fixed. If the interest is instead to see a joint effect of a mixture of exposures on the outcome, we can

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

resort to exposure-index methods. One application scenario is for designing public health interventions that act on multiple exposures simultaneously (e.g., interventions on reducing particulate matter usually also result in joint reductions of other pollutants from particulate matter sources such as sulfur dioxide), then it is more meaningful to measure the joint effect of a mixture of exposures on a health outcome. Weighted quantile sum regression (Carrico et al., 2015) and quantile G-computation approach (Keil et al., 2020) are popular exposure-index methods to form an index by taking a weighted average of quantiled exposures and then estimating an overall index effect and the weights of forming the index by fitting a linear model between the outcome and index. These methods enjoy a good interpretation gifted by decomposition of effects into the overall index effect and index weights, but they still suffer from the potential nonlinear and nonadditive relation between the exposures and the health outcome. To relieve these restrictions, the Bayesian multiple index model (BMIM, McGee et al., 2021) has been developed to leverage the strengths from both the index-based models and BKMR by incorporating more flexibility and interpretability. Suppose the exposures can be grouped into $M(< p)$ disjoint groups written as $X^m = \left(X_{m_1}, \ldots, X_{m_{l_m}}\right)^T$, $m = 1, \ldots, M$, then the model is

$$Y_i = \widetilde{Z}_i^T \gamma + h\left(\theta_1^T X_i^1, \ldots, \theta_M^T X_i^M\right) + \epsilon_i$$

with $\theta_m \in R^{l_m}$ is an index weights vector that satisfies some identifiability constraints. Then follow the notation of BKMR, the elements in the kernel matrix can be represented as $K\left(\widetilde{X}_{i_1}, \widetilde{X}_{i_2}; r\right) = \exp\left(-\sum_{m=1}^{M} r_m\left(\theta_m^T(X_{i1}^m - X_{i2}^m)\right)^2\right) = \exp\left(-\sum_{m=1}^{M}\left(\theta_m^{*T}(X_{i1}^m - X_{i2}^m)\right)^2\right)$ with a reparameterization $\theta_m^* = \sqrt{r_m}\theta_m$ to simplify computation. The spike-and-slab priors are imposed on each element of $\theta_m^*$ to facilitate variable selection, which naturally leads to index selection if all variables within an index are excluded from the model. Therefore, compared to index-based methods, this method allows for the characterization of nonlinear and nonadditive relations between the variables/index and the outcome. On the other hand, it also improves the interpretability of BKMR and achieves a dimension reduction from $p$ to $M$, which makes the inspection of the fitting results by visualization be less cumbersome than BKMR. Other highly flexible methods include the nonparametric Bayesian shrinkage method (Herring, 2010) and clustering-based Bayesian profile regression (Molitor et al., 2010, 2011), among others.

Despite the various methods available for exposome-health association studies, all of them were developed for exposure factors measured at the individual-level. Given the differences between individual- and contextual-level exposome factors, there are two major methodological challenges to apply these methods to study the spatial and contextual exposome.

The first challenge is the scalability of these methods. ExWAS and elastic-net have been applied to examine the totality of the external environment where thousands of covariates are involved with a large sample size (Hu, Zhao, Savitz, et al., 2020). GUESS is capable of handling a large number of covariates, but its computation time may be long when $N$ is large (Bottolo et al., 2013). On the contrary, the shrinkage additive models often require $N$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

> $kp$ with $k$ the average number of basis functions for constructing each function component (Marra & Wood, 2011). Other methods that we discussed above which impose very few or no assumptions on the underlying model have only been applied to problems with a subset of preselected exposures from a single source, e.g., mixtures of multiple pollutants (McGee et al., 2021), where $p$ is often not large. In addition, small scale scenarios are often considered in simulation studies investigating the performance of these methods. For example, Agier et al. (2016) considered a setting with 237 covariates and 1200 samples to test the variable selection ability of linear regression-based methods, while in investigations of the empirical performance of some more flexible methods include BKMR and BSTRASS (Hoskovec et al., 2021; Lazarevic et al., 2020), the scale of the simulation study is even much smaller with $p < 20$ and $N < 250$. Therefore, the scalability of most of the methods for spatial and contextual exposome-health studies that commonly have a large sample size and a large number of exposures remains unknown and warrants more evaluation.

The other major methodological challenge lies in the lack of consensus and guidance to optimally handle the heterogeneous spatiotemporal scales in spatial and contextual exposome data. When linking the spatial and contextual exposome data to individuals, different spatiotemporal aggregations may lead to subsequently different associations. This issue has long been recognized as the modifiable areal unit and uncertain geographic context problems spatially (Jelinski & Wu, 1996; Kwan, 2012), and their temporal analogue, the modifiable temporal unit problem (Cheng & Adepeju, 2014). To address data heterogeneity, two strategies have been used in existing spatial and contextual exposome studies: (1) calculate area- and time-weighted averages to generate individuals' exposures based on pre-selected spatiotemporal exposure windows (Hu, Zhao, Savitz, et al., 2020; Mooney et al., 2017; Nieuwenhuijsen et al., 2019) and then directly apply the existing statistical methods for exposome-health associations, and (2) preserve the original spatiotemporal scales in spatial and contextual exposome data and apply statistical methods that can account for multi-level data (Lynch et al., 2017). However, it is largely unknown about the performance of these strategies and their corresponding statistical methods as well as whether and how the modifiable areal/temporal unit problems may impact the findings. Future efforts are warranted to systematically evaluate these methods in spatial and contextual exposome-health studies.

## 5. Using spatial and contextual exposome data for disease prediction

It is attractive to leverage the spatial and contextual exposome for disease prediction given the wealth of information captured and the extremely low cost to obtain and append it to large numbers of individuals. A number of spatial and contextual factors have been used for disease prediction, such as air pollution (Jayaraj, 2021; Ku et al., 2022), climate (Ku et al., 2022), and neighborhood socioeconomic status (Bhavsar et al., 2018). Most existing studies only focused on single or very few spatial and contextual factors. It is increasingly acknowledged that most diseases are not caused by single exposure but rather the aggregate result of multiple exposures (Figueroa et al., 2020), and there are emerging efforts to leverage multiple spatial and contextual factors for disease prediction (Feng & Jiao, 2021; Hu, Zhao, Bian, et al., 2020; Mohnen et al., 2019).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Traditional machine learning models such as elastic net, support vector machine, random forest, and gradient boosting (Feng & Jiao, 2021; Hu, Zhao, Bian, et al., 2020; Mohnen et al., 2019) have been predominantly used in existing disease prediction models leveraging multiple spatial and contextual exposome factors. As shown in sections 2 and 3, spatial and contextual exposome data have rich and heterogeneous spatiotemporal structures, and they need to be manually aggregated spatiotemporally to be used by traditional machine learning models. However, the manually selected and engineered predictors from the aggregations unavoidably result in the loss of spatiotemporal structures and are likely to result in a loss of model performance.

Deep learning (also known as deep neural network) is a potential solution to preserve the rich spatiotemporal structures in spatial and contextual exposome data to improve performance for disease prediction. Since the breakthrough by AlexNet in 2012 (Krizhevsky et al., 2017), deep learning has dramatically improved state-of-the-art prediction performance in many domains including speech recognition, visual object recognition, object detection, drug discovery, and genomics (LeCun et al., 2015). Compared with traditional machine learning models, one of the biggest advantages of deep learning is its ability to perform automatic feature selection and engineering by using multiple processing layers to learn representations of data with multiple levels of abstraction. This is especially useful for data with spatiotemporal structures, such as images, time series, and sequential data, which traditionally rely on manually selected and engineered features. Given the successes of deep learning in these fields and the rich spatiotemporal structures in spatial and contextual exposome data, deep learning has great potential to boost the predictive power of the spatial and contextual exposome in disease prediction.

In the past decades, a number of deep learning model architectures have been developed, such as convolutional neural networks such as VGG (Simonyan & Zisserman, 2014), Inception (Szegedy et al., 2016, 2015), ResNet (He et al., 2016), and DenseNet (Huang et al., 2017), and recurrent neural networks such as Long-Short-Term Memory (Hochreiter & Schmidhuber, 1997) and Gated Recurrent Units (Cho et al., 2014). However, these well-designed architectures may not be directly applied to spatial and contextual exposome data, which have several major distinctions compared with images, time series and sequential data. First, spatial and contextual exposome data are much more heterogeneous. As shown in section 2, data obtained from different sources have different spatiotemporal scales, which need to be accounted for in the models. In addition, the number of variables characterizing the spatial and contextual exposome is much larger, which poses additional challenges to the design of the model architecture and computational efficiency. Lastly, for certain exposures such as air pollution and green space only a few “pixels” characterizing individuals’ immediate surroundings may matter. On the other hand, exposures such as neighborhood socioeconomic status and safety may be relevant at larger spatial scales (Kwan, 2012). This unique characteristic requires model architectures to have specific attention mechanisms for different exposures and individuals (as individuals’ activity patterns, when known, may vary). To fully leverage the information in the spatial and contextual exposome for disease prediction, new deep learning model architectures need to be developed to address these challenges.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

## 6. Summary

Although the field is still at an early stage, spatial and contextual exposome-health studies provide an effective pathway to advance our understanding of environmental determinants of health. Several critical steps are involved in spatial and contextual exposome-health studies, including extensive data engineering to provide standardized measures of the spatial and contextual exposome, spatiotemporal linkage to assign spatial and contextual exposome data to individuals, statistical analysis to identify risk and protective factors associated with the health outcomes of interest, and outcome prediction using machine and deep learning methods. The wealth of existing historical spatial and contextual data from publicly available databases provides rich resources and a solid basis for spatial and contextual exposome-health studies; however, on the other hand, the unique correlation structures, large scalability, and various spatiotemporal scales of spatial and contextual exposome data also bring challenges into each stage of studies and require more consideration and methodological investigation.

In this review, we surveyed the existing resources, methods, and tools available for conducting spatial and contextual exposome-health studies and elaborated on the current knowledge gaps and future research needs for each step. In general, methodological challenges existing across different steps involved in spatial and contextual exposome-health studies include (1) data source and variable selection as well as data harmonization for spatial and contextual exposome data engineering, (2) computational scalability issues in spatiotemporal data linkage, (3) scalability challenges of existing statistical methods and heterogeneous spatiotemporal scales of spatial and contextual exposome factors in association studies, and (4) inapplicability of existing deep learning model architectures to spatial and contextual exposome data for disease prediction. It is critical for the field to make efforts to tackle these challenges by (1) creating reference databases, developing semantic standards and ontology-based approaches, and establishing state-of-the-art data harmonization methods, (2) developing tools specifically to facilitate spatiotemporal data linkages for spatial and contextual exposome studies, (3) improving scalability and systematically evaluating the performance of existing statistical methods in spatial and contextual exposome-health association studies, and developing new multi-resolution data analysis methods to handle the heterogeneous spatiotemporal scales, and (4) developing deep learning model architectures specifically to leverage spatial and contextual exposome data for disease prediction.

## Funding:

Research reported in this publication was supported in part by the National Institute of Environmental Health Sciences under award numbers R21ES032762 and P30ES000002; and in part by the National Heart, Lung, and Blood Institute under award number K01HL153797. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## References

1. Agier L, Portengen L, Chadeau-Hyam M, Basagaña X, Giorgis-Allemand L, Siroux V, Robinson O, Vlaanderen J, González JR, Nieuwenhuijsen MJ, Vineis P, Vrijheid M, Slama R, & Vermeulen R

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(2016). A systematic comparison of linear regression-based statistical methods to assess exposome-health associations. Environmental Health Perspectives, 124(12), 1848–1856. [PubMed: 27219331]

2. Ahearn M, Baker G, Hastings A, Roof C, & Strocko E (2017). National transportation noise map. trid.trb.org. https://trid.trb.org/view/1439023
3. Akiyama Y, & Propheter SK (2005). Methods of Data Quality Control: For Uniform Crime Reporting Programs. https://it.ojp.gov/documents/CJIS_Methods_of_DQ_Control_for_UCR.pdf
4. Andrews MR, Tamura K, Claudel SE, Xu S, Ceasar JN, Collins BS, Langerman S, Mitchell VM, Baumer Y, & Powell-Wiley TM (2020). Geospatial Analysis of Neighborhood Deprivation Index (NDI) for the United States by County. Journal of Maps, 16(1), 101–112. [PubMed: 32855653]
5. Bai R, Moran GE, Antonelli JL, Chen Y, & Boland MR (2020). Spike-and-slab group lassos for grouped regression and sparse generalized additive models. Journal of the American Statistical Association, 1–14.
6. Bai R, Ročková V, & George EI (2021). Spike-and-slab meets LASSO: A review of the spike-and-slab LASSO. In Handbook of Bayesian Variable Selection (pp. 81–108). Chapman and Hall/CRC.
7. Baston D, ISciences LLC, & Baston MD (2021). Package 'exactextractr.' R Foundation for Statistical Computing. URL: https://cran.r-project.org….
8. Berrocal VJ, Gelfand AE, & Holland DM (2012). Space-time data fusion under error in computer model output: an application to modeling air quality. Biometrics, 68(3), 837–848. [PubMed: 22211949]
9. Bhavsar NA, Gao A, Phelan M, Pagidipati NJ, & Goldstein BA (2018). Value of Neighborhood Socioeconomic Status in Predicting Risk of Outcomes in Studies That Use Electronic Health Record Data. JAMA Network Open, 1(5), e182716. [PubMed: 30646172]
10. Bivand R, Keitt T, Rowlingson B, Pebesma E, Sumner M, Hijmans R, Baston D, Rouault E, Warmerdam F, Ooms J, & Rundel C (2010). rgdal : Bindings for the Geospatial Data Abstraction Library, R package version 0.6–28. Http://Cran.r-Project.Org/Package=rgdal. https://ci.nii.ac.jp/naid/10029343357/
11. Bivand R, & Rundel C (2017). RGeos: Interface to Geometry Engine-Open Source ('GEOS'). R package version 0.3–26.
12. Bleich J, Kapelner A, George EI, & Jensen ST (2014). Variable selection for BART: An application to gene regulation. The Annals of Applied Statistics, 8(3), 1750–1781.
13. Bobb JF (2022). Bayesian Kernel Machine Regression [R package bkmr version 0.2.1]. https://CRAN.R-project.org/package=bkmr
14. Bobb JF, Valeri L, Claus Henn B, Christiani DC, Wright RO, Mazumdar M, Godleski JJ, & Coull BA (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics (Oxford, England), 16(3), 493–508. [PubMed: 25532525]
15. Bottolo L, Chadeau-Hyam M, Hastie DI, Zeller T, Liquet B, Newcombe P, Yengo L, Wild PS, Schillert A, Ziegler A, Nielsen SF, Butterworth AS, Ho WK, Castagné R, Munzel T, Tregouet D, Falchi M, Cambien F, Nordestgaard BG, … Richardson S (2013). GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. PLoS Genetics, 9(8), e1003657. [PubMed: 23950726]
16. Breneman V (2013). Food environment atlas. https://agris.fao.org/agris-search/search.do?recordID=US2019X00062
17. Brokamp C (2018). DeGAUSS: Decentralized Geomarker Assessment for Multi-Site Studies. Journal of Open Source Software, 3(30), 812.
18. Carr LJ, Dunsiger SI, & Marcus BH (2010). Walk scoreTM as a global estimate of neighborhood walkability. American Journal of Preventive Medicine, 39(5), 460–463. [PubMed: 20965384]
19. 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. Journal of Agricultural, Biological, and Environmental Statistics, 20(1), 100–120. [PubMed: 30505142]
20. Cheng T, & Adepeju M (2014). Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection. PloS One, 9(6), e100465. [PubMed: 24971885]
21. Chipman HA, George EI, & McCulloch RE (2010). BART: Bayesian additive regression trees. The Annals of Applied Statistics, 4(1), 266–298.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript


22. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, & Bengio Y (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar. 10.3115/v1/d14-1179
23. Christen P (2006). Privacy-Preserving Data Linkage and Geocoding: Current Approaches and Research Directions. Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06), 497–501.
24. Chun H, & Keleş S (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 72(1), 3–25. [PubMed: 20107611]
25. Claus Henn B, Ettinger AS, Schwartz J, Téllez-Rojo MM, Lamadrid-Figueroa H, Hernández-Avila M, Schnaas L, Amarasiriwardena C, Bellinger DC, Hu H, & Wright RO (2010). Early postnatal blood manganese levels and children's neurodevelopment. Epidemiology (Cambridge, Mass), 21(4), 433–439. [PubMed: 20549838]
26. Cohen IG, & Mello MM (2018). HIPAA and Protecting Health Information in the 21st Century. JAMA: The Journal of the American Medical Association, 320(3), 231–232. [PubMed: 29800120]
27. Deng L (2012). The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. IEEE Signal Processing Magazine, 29(6), 141–142.
28. Di Q, Rowland S, Koutrakis P, & Schwartz J (2017). A hybrid model for spatially and temporally resolved ozone exposures in the continental United States. Journal of the Air & Waste Management Association, 67(1), 39–52. [PubMed: 27332675]
29. Eckert F, Fort TC, Schott PK, & Yang NJ (2020). Imputing Missing Values in the US Census Bureau's County Business Patterns (No. 26632). National Bureau of Economic Research. 10.3386/w26632
30. Elvidge CD, Baugh KE, Kihn EA, Kroehl HW, & Davis ER (1997). Mapping city lights with nighttime data from the DMSP Operational Linescan System. Photogrammetric Engineering and Remote Sensing, 63(6), 727–734.
31. ESRI. (2021). What is ArcGIS StreetMap Premium?—ArcGIS StreetMap Premium. https://doc.arcgis.com/en/streetmap-premium/get-started/overview.htm
32. ESRI. (2022). ArcGIS. https://www.esri.com/en-us/arcgis/about-arcgis/overview
33. Feng C, & Jiao J (2021). Predicting and mapping neighborhood-scale health outcomes: A machine learning approach. Computers, Environment and Urban Systems, 85(101562), 101562.
34. Figueroa JF, Frakt AB, & Jha AK (2020). Addressing social determinants of health: Time for a polysocial risk score. JAMA: The Journal of the American Medical Association, 323(16), 1553–1554. [PubMed: 32242887]
35. Folch DC, Arribas-Bel D, Koschinsky J, & Spielman SE (2016). Spatial Variation in the Quality of American Community Survey Estimates. Demography, 53(5), 1535–1554. [PubMed: 27541024]
36. Garvin E, Branas C, Keddem S, Sellman J, & Cannuscio C (2013). More than just an eyesore: local insights and solutions on vacant land and urban health. Journal of Urban Health: Bulletin of the New York Academy of Medicine, 90(3), 412–426. [PubMed: 23188553]
37. Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, & Moore R (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202, 18–27.
38. Harris DR, & Delcher C (2019). bench4gis: Benchmarking Privacy-aware Geocoding with Open Big Data. Proceedings : … IEEE International Conference on Big Data. IEEE International Conference on Big Data, 2019, 4067–4070.
39. Hart JE, Yanosky JD, Puett RC, Ryan L, Dockery DW, Smith TJ, Garshick E, & Laden F (2009). Spatial modeling of PM10 and NO2 in the continental United States, 1985–2000. Environmental Health Perspectives, 117(11), 1690–1696. [PubMed: 20049118]
40. He K, Zhang X, Ren S, & Sun J (2016, June). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. 10.1109/cvpr.2016.90

Author Manuscript Author Manuscript Author Manuscript Author Manuscript


41. Heacock ML, Lopez AR, Amolegbe SM, Carlin DJ, Henry HF, Trottier BA, Velasco ML, & Suk WA (2022). Enhancing data integration, interoperability, and reuse to address complex and emerging environmental health problems. Environmental Science & Technology. 10.1021/acs.est.1c08383
42. Herring AH (2010). Nonparametric bayes shrinkage for assessing exposures to mixtures subject to limits of detection. Epidemiology (Cambridge, Mass.), 21 Suppl 4(4), S71–6. [PubMed: 20526202]
43. Hhs, U. S. (2019). Area health resource file (AHRF). Rockville, MD: Health Resources and Services Administration.
44. Hijmans RJ (2021). The raster package. https://rspatial.org/raster/pkg/RasterPackage.pdf
45. Hochreiter S, & Schmidhuber J (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. [PubMed: 9377276]
46. Hoskovec L, Benka-Coker W, Severson R, Magzamen S, & Wilson A (2021). Model choice for estimating the association between exposure to chemical mixtures and health outcomes: A simulation study. PloS One, 16(3), e0249236. [PubMed: 33765068]
47. Hu H, Zhao J, Bian J, Zheng Y, & Pearson TA (2020). Abstract P428: A polyexposomic risk score for hypertensive disorders of pregnancy using external exposome data. Circulation, 141(Suppl_1). 10.1161/circ.141.suppl_1.p428
48. Hu H, Zhao J, Savitz DA, Prosperi M, Zheng Y, & Pearson TA (2020). An external exposome-wide association study of hypertensive disorders of pregnancy. Environment International, 141, 105797. [PubMed: 32413622]
49. Hu H, Zheng Y, Wen X, Smith SS, Nizomov J, Fishe J, Hogan WR, Shenkman EA, & Bian J (2021). An external exposome-wide association study of COVID-19 mortality in the United States. The Science of the Total Environment, 768, 144832. [PubMed: 33450687]
50. Huang G, Liu Z, Van Der Maaten L, & Weinberger KQ (2017, July). Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI. 10.1109/cvpr.2017.243
51. James P, Hart JE, & Laden F (2015). Neighborhood walkability and particulate air pollution in a nationwide cohort of women. Environmental Research, 142, 703–711. [PubMed: 26397775]
52. Jayaraj M (2021). Air quality monitoring and disease prediction using IoT and machine learning. In Advances in Intelligent Systems and Computing (pp. 18–32). Springer International Publishing.
53. Jelinski DE, & Wu J (1996). The modifiable areal unit problem and implications for landscape ecology. Landscape Ecology, 11(3), 129–140.
54. Jordahl K, Van den Bossche J, Fleischmann M, McBride J, Wasserman J, Badaracco AG, Gerard J, Snow AD, Tratner J, Perry M, Farmer C, Hjelle GA, Cochran M, Gillies S, Culbertson L, Bartos M, Ward B, Caria G, Taves M, … Wasser L (2021). geopandas/geopandas: v0.10.2. 10.5281/zenodo.5573592
55. Juarez PD, & Matthews-Juarez P (2018). Applying an exposome-wide (ExWAS) approach to cancer research. Frontiers in Oncology, 8, 313. [PubMed: 30211112]
56. Kapelner A, & Bleich J (2016). BartMachine: Machine learning with Bayesian additive regression trees. Journal of Statistical Software, 70(4). 10.18637/jss.v070.i04
57. 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. Environmental Health Perspectives, 128(4), 47004. [PubMed: 32255670]
58. Kim S-Y, Bechle M, Hankey S, Sheppard L, Szpiro AA, & Marshall JD (2020). Concentrations of criteria pollutants in the contiguous U.S., 1979 – 2015: Role of prediction model parsimony in integrated empirical geographic regression. PloS One, 15(2), e0228535. [PubMed: 32069301]
59. Kittel TGF, Royle JA, Daly C, Rosenbloom NA, Gibson WP, Fisher HH, Schimel DS, Berliner LM, & Participants V (1997). A gridded historical (1895–1993) bioclimate dataset for the conterminous United States. Proceedings of the 10th Conference on Applied Climatology, 20, 20–24.
60. Kloog I, Melly SJ, Coull BA, Nordio F, & Schwartz JD (2015). Using Satellite-Based Spatiotemporal Resolved Air Temperature Exposure to Study the Association between Ambient

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Air Temperature and Birth Outcomes in Massachusetts. In Environmental Health Perspectives (Vol. 123, Issue 10, pp. 1053–1058). 10.1289/ehp.1308075 [PubMed: 25850104]

61. Knaap E, Cortes RX, Rey S, Arribas-Bel D, Gaboardi J, Fleischmann M, & Frontiera P (2021). pysal/tobler: Release v0.8.2. 10.5281/zenodo.5047613
62. Kounadi O, Lampoltshammer TJ, Leitner M, & Heistracher T (2013). Accuracy and privacy aspects in free online reverse geocoding services. Cartography and Geographic Information Science, 40(2), 140–153.
63. Krieger N, Kim R, Feldman J, & Waterman PD (2018). Using the Index of Concentration at the Extremes at multiple geographical levels to monitor health inequities in an era of growing spatial social polarization: Massachusetts, USA (2010–14). International Journal of Epidemiology, 47(3), 788–819. [PubMed: 29522187]
64. Krieger N, Waterman PD, Spasojevic J, Li W, Maduro G, & Van Wye G (2016). Public Health Monitoring of Privilege and Deprivation With the Index of Concentration at the Extremes. American Journal of Public Health, 106(2), 256–263. [PubMed: 26691119]
65. Krizhevsky A, Sutskever I, & Hinton GE (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90.
66. Ku Y, Kwon SB, Yoon J-H, Mun S-K, & Chang M (2022). Machine learning models for predicting the occurrence of respiratory diseases using climatic and air-pollution factors. Clinical and Experimental Otorhinolaryngology. 10.21053/ceo.2021.01536
67. Kwan M-P (2012). The Uncertain Geographic Context Problem. Annals of the Association of American Geographers. Association of American Geographers, 102(5), 958–968.
68. Lamsal LN, Duncan BN, Yoshida Y, Krotkov NA, Pickering KE, Streets DG, & Lu Z (2015). U.S. NO2 trends (2005–2013): EPA Air Quality System (AQS) data versus improved observations from the Ozone Monitoring Instrument (OMI). Atmospheric Environment (Oxford, England: 1994), 110, 130–143.
69. Lazarevic N, Knibbs LD, Sly PD, & Barnett AG (2020). Performance of variable and function selection methods for estimating the nonlinear health effects of correlated chemical mixtures: A simulation study. Statistics in Medicine, 39(27), 3947–3967. [PubMed: 32940933]
70. LeCun Y, Bengio Y, & Hinton G (2015). Deep learning. Nature, 521(7553), 436–444. [PubMed: 26017442]
71. Lee M, Kloog I, Chudnovsky A, Lyapustin A, Wang Y, Melly S, Coull B, Koutrakis P, & Schwartz J (2015). Spatiotemporal prediction of fine particulate matter using high-resolution satellite images in the Southeastern US 2003–2011. Journal of Exposure Science & Environmental Epidemiology, 26(4), 377–384. [PubMed: 26082149]
72. Li J, & Heap AD (2014). Spatial interpolation methods applied in the environmental sciences: A review. Environmental Modelling & Software, 53, 173–189.
73. Liao L, Song J, Wang J, Xiao Z, & Wang J (2016). Bayesian Method for Building Frequent Landsat-Like NDVI Datasets by Integrating MODIS and Landsat NDVI. Remote Sensing, 8(6), 452.
74. Liu X, Wang F, & Zhang Z (2021). Research Progress of Spatio-Temporal Interpolation in the Field of Public Health. Journal of Physics. Conference Series, 1802(4), 042060.
75. Logue JM, Small MJ, & Robinson AL (2011). Evaluating the national air toxics assessment (NATA): Comparison of predicted and measured air toxics concentrations, risks, and sources in Pittsburgh, Pennsylvania. Atmospheric Environment (Oxford, England: 1994), 45(2), 476–484.
76. Loxham M, Davies DE, & Holgate ST (2019). The health effects of fine particulate air pollution [Review of The health effects of fine particulate air pollution]. BMJ, 367, l6609. [PubMed: 31776108]
77. Lynch SM, Mitra N, Ross M, Newcomb C, Dailey K, Jackson T, Zeigler-Johnson CM, Riethman H, Branas CC, & Rebbeck TR (2017). A Neighborhood-Wide Association Study (NWAS): Example of prostate cancer aggressiveness. PloS One, 12(3), e0174548. [PubMed: 28346484]
78. Ma X, Longley I, Gao J, Kachhara A, & Salmond J (2019). A site-optimised multi-scale GIS based land use regression model for simulating local scale patterns in air pollution. The Science of the Total Environment, 685, 134–149. [PubMed: 31174113]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript


79. Mannucci PM, Harari S, Martinelli I, & Franchini M (2015). Effects on health of air pollution: a narrative review. Internal and Emergency Medicine, 10(6), 657–662. [PubMed: 26134027]
80. Marra G, & Wood SN (2011). Practical variable selection for generalized additive models. Computational Statistics & Data Analysis, 55(7), 2372–2387.
81. McGee G, Wilson A, Webster TF, & Coull BA (2021). Bayesian multiple index models for environmental mixtures. Biometrics. 10.1111/biom.13569
82. Mennitt D, Sherrill K, & Fristrup K (2014). A geospatial model of ambient sound pressure levels in the contiguous United States. The Journal of the Acoustical Society of America, 135(5), 2746–2764. [PubMed: 24815258]
83. Mesinger F, DiMego G, Kalnay E, Mitchell K, Shafran PC, Ebisuzaki W, Jović D, Woollen J, Rogers E, Berbery EH, Ek MB, Fan Y, Grumbine R, Higgins W, Li H, Lin Y, Manikin G, Parrish D, & Shi W (2006). North American Regional Reanalysis. Bulletin of the American Meteorological Society, 87(3), 343–360.
84. Messer LC, Laraia BA, Kaufman JS, Eyster J, Holzman C, Culhane J, Elo I, Burke JG, & O'Campo P (2006). The development of a standardized neighborhood deprivation index. Journal of Urban Health: Bulletin of the New York Academy of Medicine, 83(6), 1041–1062. [PubMed: 17031568]
85. Mohnen SM, Schneider S, & Droomers M (2019). Neighborhood characteristics as determinants of healthcare utilization - a theoretical model. Health Economics Review, 9(1), 7. [PubMed: 30840211]
86. Molitor J, Papathomas M, Jerrett M, & Richardson S (2010). Bayesian profile regression with an application to the National Survey of Children's Health. Biostatistics (Oxford, England), 11(3), 484–498. [PubMed: 20350957]
87. Molitor J, Su JG, Molitor N-T, Rubio VG, Richardson S, Hastie D, Morello-Frosch R, & Jerrett M (2011). Identifying vulnerable populations through an examination of the association between multipollutant profiles and poverty. Environmental Science & Technology, 45(18), 7754–7760. [PubMed: 21797252]
88. Mooney SJ, Joshi S, Cerdá M, Kennedy GJ, Beard JR, & Rundle AG (2017). Contextual correlates of physical activity among older adults: A neighborhood environment-wide association study (NE-WAS). Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology, 26(4), 495–504.
89. Moore RB, McKay LD, Rea AH, Bondelid TR, Price CV, Dewald TG, & Johnston CM (2019). User's guide for the national hydrography dataset plus (NHDPlus) high resolution. In Open-File Report (No. 2019–1096). US Geological Survey. 10.3133/ofr20191096
90. Mukhopadhyay S, & Sahu SK (2018). A Bayesian spatiotemporal model to estimate long-term exposure to outdoor air pollution at coarser administrative geographies in England and Wales. Journal of the Royal Statistical Society. Series A, , 181(2), 465–486.
91. Nieuwenhuijsen MJ, Agier L, Basagaña X, Urquiza J, Tamayo-Uria I, Giorgis-Allemand L, Robinson O, Siroux V, Maitre L, de Castro M, Valentin A, Donaire D, Dadvand P, Aasvang GM, Krog NH, Schwarze PE, Chatzi L, Grazuleviciene R, Andrusaityte S, … Slama R (2019). Influence of the urban exposome on birth weight. Environmental Health Perspectives, 127(4), 47007. [PubMed: 31009264]
92. Olalekan RM, Timothy AA, Enabulele Chris E, & Olalekan AS (2018). Assessment of air quality indices and its health impacts in ilorin metropolis, Kwara State, Nigeria. Science Park Journals of Scientific Research and Impact, 4(4), 060–074.
93. Patel CJ, Bhattacharya J, & Butte AJ (2010). An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. PloS One, 5(5), e10746. [PubMed: 20505766]
94. Pebesma E (2018). Simple features for R: Standardized support for spatial vector data. The R Journal, 10(1), 439.
95. Pebesma E, & Bivand RS (2005). Classes and Methods for Spatial Data: the sp Package. https://cran.microsoft.com/snapshot/2017-08-01/web/packages/sp/vignettes/intro_sp.pdf
96. Perry M (2021, October 29). rasterstats. PyPI. https://pypi.org/project/rasterstats/
97. PG-Strom Development Team. (2021). PG-Strom. https://heterodb.github.io/pg-strom/

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

98. QGIS. (2022). https://www.qgis.org/en/site/
99. R Core Team. (2013). R: A language and environment for statistical computing.
100. Rocklin M (2015). Dask: Parallel computation with blocked algorithms and task scheduling. Proceedings of the 14th Python in Science Conference, 130, 136.
101. Ročková V, & George EI (2018). The spike-and-slab LASSO. Journal of the American Statistical Association, 113(521), 431–444.
102. Rundle AG, Chen Y, Quinn JW, Rahai N, Bartley K, Mooney SJ, Bader MD, Zeleniuch-Jacquotte A, Lovasi GS, & Neckerman KM (2019). Development of a Neighborhood Walkability Index for Studying Neighborhood Physical Activity Contexts in Communities across the U.S. over the Past Three Decades. Journal of Urban Health: Bulletin of the New York Academy of Medicine, 96(4), 583–590. [PubMed: 31214976]
103. Sampson PD, Richards M, Szpiro AA, Bergen S, Sheppard L, Larson TV, & Kaufman JD (2013). A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2.5 concentrations in epidemiology. Atmospheric Environment, 75, 383–392. [PubMed: 24015108]
104. Scheipl F (2011). SpikeSlabGAM: Bayesian variable selection, model choice and regularization for generalized additive mixed models inR. Journal of Statistical Software, 43(14). 10.18637/jss.v043.i14
105. Scheipl F, Fahrmeir L, & Kneib T (2012). Spike-and-slab priors for function selection in structured additive regression models. Journal of the American Statistical Association, 107(500), 1518–1532.
106. Simonyan K, & Zisserman A (2014). Very deep convolutional networks for large-scale image recognition. In arXiv [cs.CV]. arXiv. http://arxiv.org/abs/1409.1556
107. Siroux V, Agier L, Basagaña X, Urquiza J, Sunyer J, Casas M, Robinson O, Granum B, Oftedal B, Thomsen C, De Castro M, Nieuwenhuijsen M, Wright J, Mceachan R, Bird P, Uphoff N, Grazuleviciene R, Andrusaityte S, Petraviciene I, … Slama R (2018, September 15). Early life exposome and lung function in children from the HELIX cohort. Epidemiology. ERS International Congress 2018 abstracts. 10.1183/13993003.congress-2018.oa5184
108. Susanto F, de Souza P, & He J (2016). Spatiotemporal Interpolation for Environmental Modelling. Sensors, 16(8). 10.3390/s16081245
109. Szegedy C, Ioffe S, Vanhoucke V, & Alemi A (2016). Inception-v4, Inception-ResNet and the impact of residual connections on learning. In arXiv [cs.CV]. arXiv. http://arxiv.org/abs/1602.07261
110. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, & Rabinovich A (2015, June). Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA. 10.1109/cvpr.2015.7298594
111. The PostGIS Development Group. (2019). PostGIS 3.0.0alpha5dev Manual. https://postgis.net/docs/manual-dev/
112. Thomas J, & Zeller L (2021, May 17). National Walkability Index User Guide and Methodology. https://www.epa.gov/smartgrowth/national-walkability-index-user-guide-and-methodology
113. Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, 58(1), 267–288.
114. Turner MC, Nieuwenhuijsen M, Anderson K, Balshaw D, Cui Y, Dunton G, Hoppin JA, Koutrakis P, & Jerrett M (2017). Assessing the exposome with external measures: Commentary on the state of the science and research recommendations. Annual Review of Public Health, 38, 215–239.
115. van Donkelaar A, Martin RV, Li C, & Burnett RT (2019). Regional Estimates of Chemical Composition of Fine Particulate Matter Using a Combined Geoscience-Statistical Method with Information from Satellites, Models, and Monitors. Environmental Science & Technology, 53(5), 2595–2611. [PubMed: 30698001]
116. Verhoelst T, Compernolle S, Pinardi G, Lambert J-C, Eskes HJ, Eichmann K-U, Fjæraa AM, Granville J, Niemeijer S, Cede A, Tiefengraber M, Hendrick F, Pazmiño A, Bais A, Bazureau A, Boersma KF, Bognar K, Dehn A, Donner S, … Zehner C (2021). Ground-based validation

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of the Copernicus Sentinel-5P TROPOMI NO2 measurements with the NDACC ZSL-DOAS, MAX-DOAS and Pandonia global networks. Atmospheric Measurement Techniques, 14(1), 481–510.

117. VoPham T, Hart JE, Bertrand KA, Sun Z, Tamimi RM, & Laden F (2016). Spatiotemporal exposure modeling of ambient erythemal ultraviolet radiation. Environmental Health: A Global Access Science Source, 15(1), 111. [PubMed: 27881169]
118. Vrijheid M (2014). The exposome: a new paradigm to study the impact of environment on health. Thorax, 69(9), 876–878. [PubMed: 24906490]
119. Vrijheid M, Fossati S, Maitre L, Márquez S, Roumeliotaki T, Agier L, Andrusaityte S, Cadiou S, Casas M, de Castro M, Dedele A, Donaire-Gonzalez D, Grazuleviciene R, Haug LS, McEachan R, Meltzer HM, Papadopouplou E, Robinson O, Sakhi AK, … Chatzi L (2020). Early-life environmental exposures and childhood obesity: An exposome-wide approach. Environmental Health Perspectives, 128(6), 67009. [PubMed: 32579081]
120. Wheeler DC, & Wang A (2015). Assessment of residential history generation using a public-record database. International Journal of Environmental Research and Public Health, 12(9), 11670–11682. [PubMed: 26393626]
121. Wild CP (2005). Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology, 14(8), 1847–1850.
122. Wild CP (2012). The exposome: from concept to utility. International Journal of Epidemiology, 41(1), 24–32. [PubMed: 22296988]
123. xarray-spatial Development Team. (2022, February 4). xarray-spatial. PyPI. https://pypi.org/project/xarray-spatial/
124. Yanosky JD, Paciorek CJ, Laden F, Hart JE, Puett RC, Liao D, & Suh HH (2014). Spatio-temporal modeling of particulate air pollution in the conterminous United States using geographic and meteorological predictors. Environmental Health: A Global Access Science Source, 13, 63. [PubMed: 25097007]
125. Yuan M, & Lin Y (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 68(1), 49–67.
126. Zhang H, Hu H, Diller M, Hogan WR, Prosperi M, Guo Y, & Bian J (2021). Semantic standards of external exposome data. Environmental Research, 197, 111185. [PubMed: 33901445]
127. Zhao P, & Yu B (2006). On model selection consistency of Lasso. The Journal of Machine Learning Research, 7, 2541–2563.
128. Zheng Y, Chen Z, Pearson T, Zhao J, Hu H, & Prosperi M (2020). Design and methodology challenges of environment-wide association studies: A systematic review. Environmental Research, 183, 109275. [PubMed: 32105887]
129. Zou H (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101(476), 1418–1429.
130. Zou H, & Hastie T (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 67(2), 301–320.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

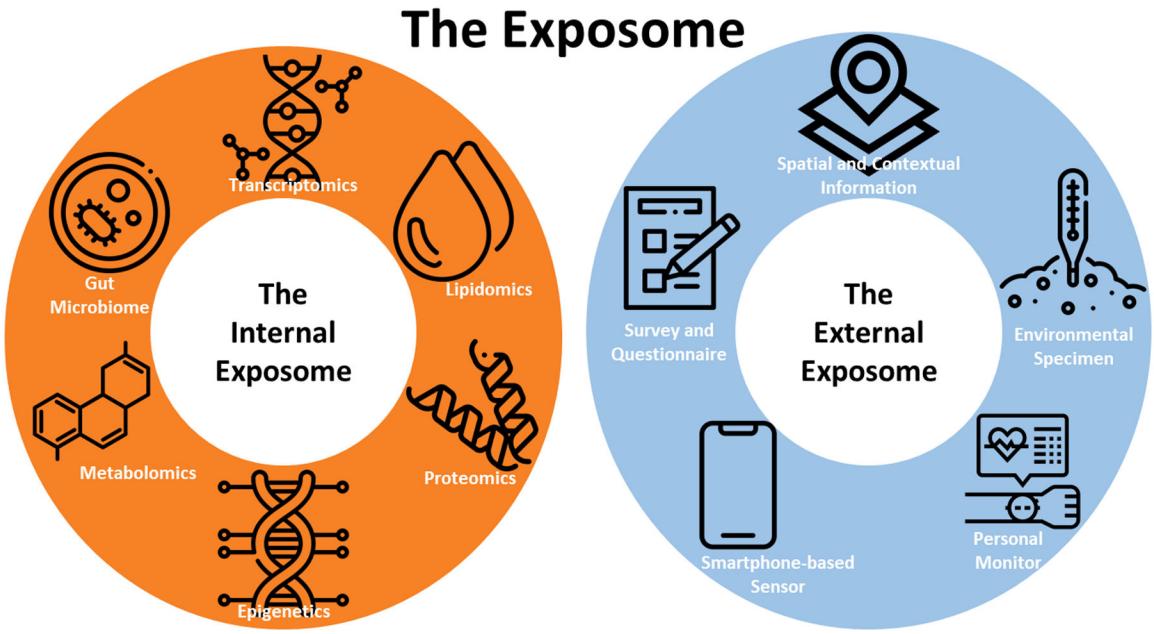



**Figure 1.**
Overview of the exposome categorizations and measurement methods

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

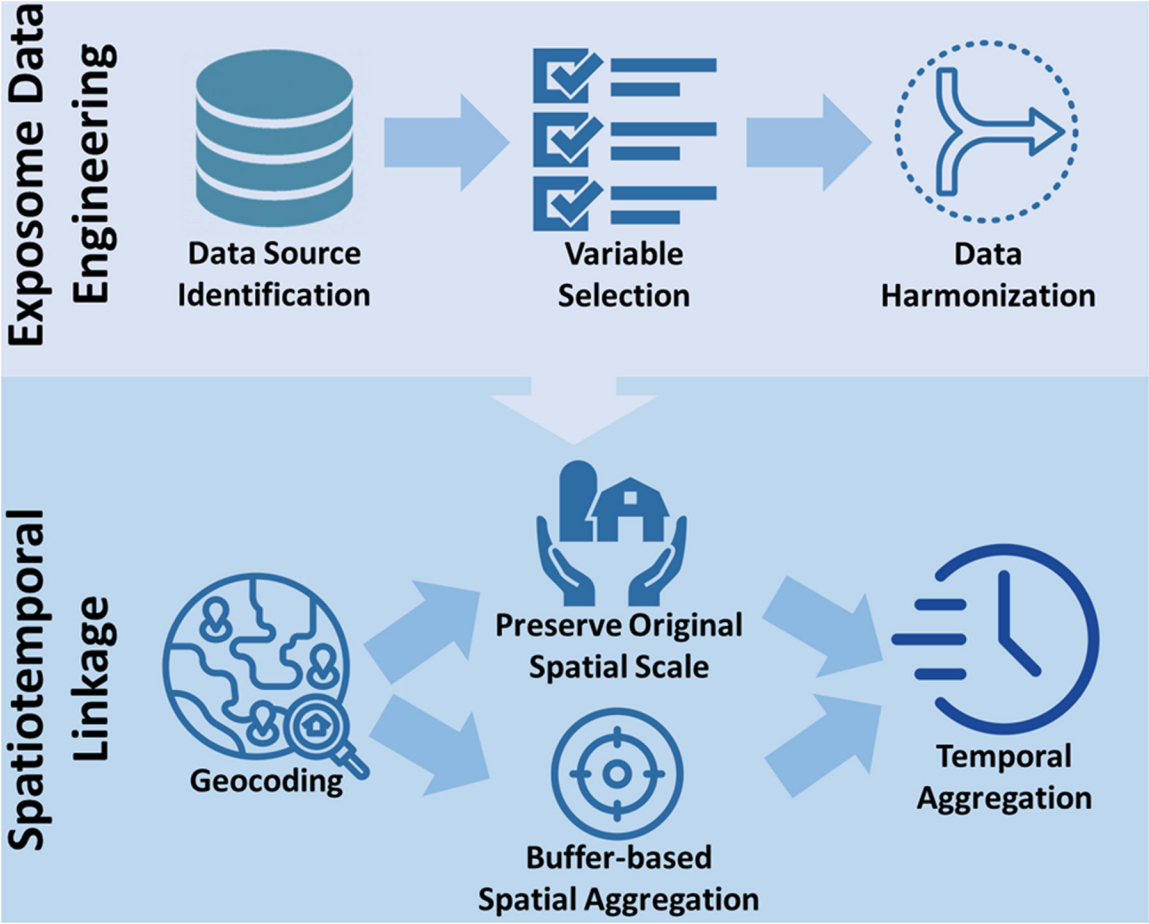



**Figure 2.**

Overview of data engineering and spatiotemporal data linkage in spatial and contextual exposome-health studies.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**Table 1.**

Examples of publicly available spatial and contextual exposome data sources.

| Exposure | Example Data Source[a] | Reference | Available Since | Geometry | Spatiotemporal Scale[b] | Example Measures |
|---|---|---|---|---|---|---|
| ***Natural Environment*** | | | | | | |
| *Air pollution* | ACAG | van Donkelaar et al. (2019) | 2000 | Raster | 0.01°/1-month | $PM_{2.5}$, sulfate, nitrate, ammonium |
| | CACES | Kim et al. (2020) | 1979 | Polygon | BG/1-year | $PM_{2.5}$, $PM_{10}$, $O_3$, $NO_2$, CO, $SO_2$ |
| | USEPA AQS | Lamsal et al. (2015) | 1980 | Point | PL/1-day | $PM_{2.5}$, $PM_{10}$, $O_3$, $NO_2$, CO, $SO_2$, Pb |
| | USEPA FAQSD | Berrocal, Gelfand, and Holland (2012) | 2002 | Polygon | CT/1-day | $PM_{2.5}$, $O_3$ |
| | Sentinel-5P | Verhoelst et al. (2021) | 2018 | Raster | 1113m/daily | $NO_2$, CO, $SO_2$ |
| | USEPA NATA | Logue Small and Robinson (2011) | 1996 | Polygon | CT/1-year | Acrolein, propylene oxide |
| *Ultraviolet radiation* | NASA TOMS, OMI, UVMRP | VoPham et al. (2016) | 1998 | Raster | 1000m/2-year | Erythemal UV |
| *Meteorology* | NCEP-NARR | Mesinger et al. (2006) | 1979 | Raster | 32km/1-day | Temperature, humidity |
| | PRISM | Kittel et al. (1997) | 1895 | Raster | 800m/1-month | |
| ***Built Environment*** | | | | | | |
| *Vacant land* | USHUD | Garvin et al. (2013) | 2005 | Polygon | CT/3-month | Percent addresses vacant |
| *Roads and traffic* | ESRI | ESRI (2021) | 1995 | Line | Line/1-year | Distance to nearest major roadway |
| *Noise* | NPS | Mennitt, Sherrill, and Fristrup (2014) | 2000 | Raster | 270m/1-year | A-weighted $L_{50}$ |
| | USDOT | Ahearn et al. (2017) | 2016 | Raster | 30m/1-year | Aviation noise, road noise |
| *Light at night* | DMSP | Elvidge et al. (1997) | 1992 | Raster | 1000m/1-year | Nighttime lights |
| *Walkability* | Walk Score | Carr, Dunsiger, and Marcus (2010) | 2009 | Raster | 0.0015°/CS | Walkability index |
| | InfoUSA, Census and ACS | James, Hart, and Laden (2015) | 2009 | Point | PL/1-year | |
| | BEH-NWI | Rundle et al. (2019) | 1990 | Point | PL/1-year | |
| *Food Access* | USDA FARA | Breneman (2013) | 2010 | Polygon | CT/1-year | % low-access population at 1 mile |
| *Green Space* | NASA Landsat | Liao et al. (2016) | 1981 | Raster | 30m/16-day | Normalized difference vegetation index |
| *Blue Space* | NHD | Moore et al. (2019) | 1990 | Polygon | FB/1-year | Presence of blue space within 250m |
| ***Social Environment*** | | | | | | |

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

| Exposure | Example Data Source[a] | Reference | Available Since | Geometry | Spatiotemporal Scale[b] | Example Measures |
|---|---|---|---|---|---|---|
| *Socio-demographic* | Census and ACS | Folch et al. (2016) | 1980 | Polygon | CT/10-year or 5-year | Neighborhood deprivation index |
| *Social Capital* | CBP | Eckert et al. (2020) | 1986 | Polygon | ZCTA5/1-year | Religious, civic, and social organizations |
| *Crime and Safety* | UCR | Akiyama and Propheter (2005) | 1974 | Polygon | County/1-year | Burglary rate, aggravated assault rate |
| *Healthcare Indicator* | AHRF | USHHS (2019) | 2010 | Polygon | County/1-year | Hospital utilizations and expenditures |

[a]ACAG, Atmospheric Composition Analysis Group; ACS, American Community Survey; AHRF: Area Health Resource File; CACES, Center for Air, Climate, and Energy Solutions; CBP, Census Business Patterns; DMSP, U.S. Defense Meteorological Satellite Program; FARA, Food Access Research Atlas; NARR, North American Regional Reanalysis; NASA, National Aeronautics and Space Administration; NATA, National Air Toxics Assessment; NCEP, National Centers for Environmental Prediction; NHD, National Hydrography Dataset; NPS, National Park Service; OMI, Ozone Monitoring Instrument; PRISM, Parameter-elevated Regressions on Independent Slopes Model; TOMS, Total Ozone Mapping Spectrometer; UCR, Uniform Crime Reporting; USDA, U.S. Department of Agriculture; USEPA, U.S. Environmental Protection Agency; USHUD, U.S. Department of Housing and Urban Development; UVMRP, UV-B Monitoring and Research Program.

[b]BG: Census Block Group; CS: Cross-sectional; CT: Census Tract; FB: Feature-based; PL: Point Location; ZCTA5: 5-digit ZIP Code Tabulation Area

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**Table 2.**

Statistical methods for exposome-health association studies.

| Method | Reference | Category[a] | Focus[b] |
|---|---|---|---|
| Environment-wide association study (EWAS) | Patel, Bhattacharya, and Butte (2010) | Linear | Individual |
| Exposome-wide association study (ExWAS) | Juarez and Matthews-Juarez (2018) | Linear | Individual |
| Lasso regression | Tibshirani (1996) | Linear | Individual |
| Group Lasso regression | Yuan and Lin (2006) | Linear | Individual/Joint |
| Adaptive Lasso regression | Zou (2006) | Linear | Individual |
| Elastic-net regression | Zou and Hastie (2005) | Linear | Individual |
| Graphical unit evolutionary stochastic search (GUESS) | Bottolo et al. (2013) | Linear | Individual |
| Spike-and-Slab Lasso regression | Rockova and George (2018) | Linear | Individual |
| Spike-and-Slab Group Lasso regression[c] | Bai et al. (2020) | Linear/Additive | Individual/Joint |
| Sparse partial least square regression (sPLS) | Chun and Keleş (2010) | Linear | Individual/Joint |
| Bayesian structured additive regression with spike-and-slab priors (BSTARSS) | Scheipl et al. (2012) | Additive | Individual/Joint |
| Bayesian kernel machine regression (BKMR) | Bobb et al. (2015) | NP | Individual/Joint |
| Bayesian additive regression trees (BART) | Chipman et al. (2010) | NP | Individual |
| Weighted quantile sum regression (WQS) | Carrico et al. (2015) | Index | Joint |
| Quantile G-computation approach | Keil et al. (2020) | Index | Joint |
| Bayesian multiple index model (BMIM) | McGee et al. (2021) | NP | Joint |

[a]Linear: linear regression-based; Additive: additive model-based; NP: Non-parametric model-based; Index: Index-based;

[b]Individual: method focuses on individual effect from an exposure on the outcome; Joint: method focuses on the joint effect of a mixture of exposures on the outcome

[c]With coefficients of basis functions as a group