scientific reports

Check for updates

OPEN

Investigating the association of environmental exposures and all-cause mortality in the UK Biobank using sparse principal component analysis

Mohammad Mamouei✉, Yajie Zhu, Milad Nazarzadeh, Abdelaali Hassaine, Gholamreza Salimi-Khorshidi, Yutong Cai & Kazem Rahimi

Multicollinearity refers to the presence of collinearity between multiple variables and renders the results of statistical inference erroneous (Type II error). This is particularly important in environmental health research where multicollinearity can hinder inference. To address this, correlated variables are often excluded from the analysis, limiting the discovery of new associations. An alternative approach to address this problem is the use of principal component analysis. This method, combines and projects a group of correlated variables onto a new orthogonal space. While this resolves the multicollinearity problem, it poses another challenge in relation to interpretability of results. Standard hypothesis testing methods can be used to evaluate the association of projected predictors, called principal components, with the outcomes of interest, however, there is no established way to trace the significance of principal components back to individual variables. To address this problem, we investigated the use of sparse principal component analysis which enforces a parsimonious projection. We hypothesise that this parsimony could facilitate the interpretability of findings. To this end, we investigated the association of 20 environmental predictors with all-cause mortality adjusting for demographic, socioeconomic, physiological, and behavioural factors. The study was conducted in a cohort of 379,690 individuals in the UK. During an average follow-up of 8.05 years (3,055,166 total person-years), 14,996 deaths were observed. We used Cox regression models to estimate the hazard ratio (HR) and 95% confidence intervals (CI). The Cox models were fitted to the standardised environmental predictors (a) without any transformation (b) transformed with PCA, and (c) transformed with SPCA. The comparison of findings underlined the potential of SPCA for conducting inference in scenarios where multicollinearity can increase the risk of Type II error. Our analysis unravelled a significant association between average noise pollution and increased risk of all-cause mortality. Specifically, those in the upper deciles of noise exposure have between 5 and 10% increased risk of all-cause mortality compared to the lowest decile.

Abbreviations

CI	Confidence Interval
CNOSSOS-EU	Common Noise Assessment Methods in Europe
CVD	Cardiovascular disease
ESCAPE	European Study for Cohorts of Air Pollution Effects
HR	Hazard ratio
ICD	International classification of diseases
LUR	Land Use Regression
N/EWAS	Neighbourhood-wide or environment-wide associations
NMF	Non-negative matrix factorisation
PC	Principal component

Deep Medicine, Nuffield Department of Women's & Reproductive Health, Oxford Martin School, University of Oxford, 1st Floor, Haye House, 75 George Street, Oxford OX1 2BQ, UK. ✉email: Mohammad.Mamouei@wrh.ox.ac.uk

PM	Particulate matter
PCA	Principal component analysis
SD	Standard deviation
SPCA	Sparse principal component analysis

Numerous studies have reported significant associations between individual environmental variables such as traffic noise, air pollution, green space and health outcomes[1–4]. Such findings are important but a key limitation of them is that they do not consider simultaneous exposure to key environmental stressors in the analysis. Since environmental variables are often highly correlated, this limitation can diminish the causal plausibility of the findings. For instance, multiple studies have reported significant associations between traffic noise, all-cause mortality and cardiovascular diseases[5–7]. However, exposure to higher levels of traffic noise, also increases the likelihood of exposure to particulate matter pollutants (PM), gaseous pollutants and other traffic-related stressors. Additionally, individuals exposed to these stressors are less likely to have access to domestic and urban green spaces which have been reported to have protective effects against adverse health outcomes.

The aforementioned gap in simultaneous analysis of multiple environmental stressors is partly due to "multicollinearity". Environmental variables such as green space, gaseous and particulate air pollution, noise pollution and traffic-related variables are usually temporally and/or spatially correlated; they are also often correlated with demographic and socioeconomic determinants. The inclusion of these correlated variables in regression models leads to erroneous estimation of the effect size, broad confidence intervals, and therefore, inaccurate interpretation. Methods for mitigating the effects of multiple correlated variables include dimensionality reduction (e.g. Principal Component Analysis (PCA)), partial least-squares, shrinkage regression models, mixture models, and Bayesian approach. However, several factors such as complexity of application, difficulty of interpretation, and high computational requirements have impeded their adoption in environmental research[8,9]. PCA and PLS are not interpretable[10,11]. Shrinkage regression models achieve sparsity by penalising nonzero model coefficients as well as regression error. While this mitigates multicollinearity, penalising regression coefficients has unfortunate implications for statistical inference where the aim is finding reliable estimates of model coefficients regardless of their contribution to predictive performance. Mixture models and Bayesian modelling are computationally demanding and become intractable as the number of variables and observations increase[12,13]. Therefore, a widely applicable, interpretable and computation-efficient statistical approach is needed to fill this gap.

The computational efficiency and desired statistical properties of PCA make it good candidate for big data studies where multicollinearity poses a problem. PCA transforms a group of correlated variables into a smaller group of independent variables, called principal components. Therefore, the use of principal components -instead of the set of variables- in regression analysis eliminates multicollinearity. Due to these advantages, the method has been widely used in epidemiological studies[14–17]. Yet, the main shortcoming of PCA is interpretability. Each principal component is a mix of all variables, making inference impossible. Since this difficulty is the result of a dense transformation, we hypothesise that a sparse transformation could facilitate the interpretation of findings. To this end, we investigated the usefulness of Sparse Principal Component Analysis (SPCA)[18]. As a case study, we focused on potentially modifiable but correlated environmental exposures. As such, the main contributions of the study are, firstly, we showcase the benefits of SPCA as an interpretable alternative to PCA, offering clear advantages for statistical inference in the presence of multicollinearity. Secondly, using SPCA, we showed a significant association between noise levels and all-cause mortality after adjusting for a comprehensive list of correlated environmental variables that could affect health outcomes independent of noise levels, namely residential traffic levels, vicinity to roads and major roads, green space, natural environment, domestic garden, proximity to water, and coastal proximity. While several studies in the past decade have largely addressed the questions around the confounding effects of air pollution and traffic noise[5–7,19], none of the available studies have adjusted for the comprehensive list of environmental confounders considered in our study.

Methods

Study population. This analysis was conducted using the UK Biobank cohort. The UK Biobank is a large prospective cohort study involving 502,527 participants aged 40–69 years who were recruited between 2006 and 2010 from 22 assessment centres across the UK[20]. The data is globally accessible to approved researchers. We excluded participants from analysis if they had any of the following: (a) withdrawal of consent for future data linkage from the UK Biobank after recruitment (158 individuals) (b) left the UK (1102 individuals) (d) deaths reported by relatives but not recorded in death registry data (38 individuals) (e) missing data on investigated environmental exposures (69,268 individuals) and (f) change of residential address after the baseline (52,271 individuals). The final sample size consisted of 379,690 individuals. All participants provided written consent, ethical approval was obtained from the North West Multi-Centre Research Ethical Committee and Patient Information Advisory Group and all methods were performed in accordance with the relevant guidelines and regulations.

Environmental exposures. Measures for exposure to air pollutants included: the annual average concentration of $PM_{2.5}$, PM_{10}, and PM_{coarse} (particulate matter (PM) with an aerodynamic diameter of less than 2.5 µm, 10 µm, and between 2.5 and 10 µm respectively), NO_2 (nitrogen dioxide), and NO_x (nitrogen oxides). These measures were calculated for year 2010 for each participant's residential address at recruitment using Land Use Regression model developed and validated by the ESCAPE project[21,22].

Exposures to traffic were also derived by the ESCAPE project for each participant's home: traffic intensity on the nearest road, traffic intensity on the nearest major road (*traffic intensity*, vehicles/day), and sum of major road length within 100 m buffer. Average daytime, evening time and night-time sound level of road traffic noise pollution were derived for year 2010 using the CNOSSOS model[23].

Other environmental indicators included the proportion of green space, natural environment, domestic garden, and water within 300 m and 1000 m of residential addresses, using the 2005 Generalised Land Use Database for England and Centre for Ecology and Hydrology 2007 Land Cover Map data for Great Britain[24]. The buffer sizes were decided based on relevant health evidence and public policy on both density and accessibility. Coastal proximity was estimated using Euclidean distance raster[25].

All the exposure indicators were only modelled or available to a single year, which may differ up to 4 years from recruitment. This may particularly affect air pollution and road traffic noise estimates, distributions of which tend to be spatially and temporally different. As with other studies[26,27] using these air pollution and noise data in UK Biobank, we made an assumption that whilst the absolute traffic volumes will have changed between earlier baseline periods and 2010, the relative difference in these exposures would likely have been spatially stable over this short period in the UK. This assumption is supported by findings for NO_2 air pollution in Great Britain, for which road traffic is a major source, where LUR-modelled NO_2 estimates for 2009 could be reliably back-extrapolated to earlier 1990s[28]. Between years 2010 and 2018, total annual emissions for PM_{10} and $PM_{2.5}$ have been stable across the UK while emissions for NO_2 have proportionally decrease according to the official statistics[29]. While we cannot exclude the possibility of exposure misclassification, the decision of using single-year annual average exposures at baseline to represent the annual average exposures during the entire follow-up period was deemed justifiable.

Additional covariates.

In the regression analysis, we adjusted for a number of sociodemographic, socioeconomic, physiological, behavioural and lifestyle determinants of health. Specifically, we adjusted for age, sex, ethnicity, Townsend Deprivation Index, household income, qualifications, employment status, standing height, body mass index, average systolic blood pressure (SBP), average diastolic blood pressure (DBP), average pulse rate (PR), alcohol consumption and smoking status. Table 1 provides a descriptive summary of the cohort.

Health outcome.

We used all-cause mortality as the outcome of interest. The date of death was extracted from the linked national death registries. An event was ascertained if death was recorded between the date of recruitment and the end of follow-up (censoring date: 1st May 2017). Fig. 1. shows the top 20 ICD10 codes that were registered as the primary causes of death.

Statistical analysis.

We used SPCA, which was originally proposed by Zou and colleagues[18]. Our hypothesis is that the sparsity of principal components in SPCA can help overcome the limitation of PCA for identifying important stressors. The term 'sparse' in SPCA means that most of the coefficients in the loading matrix will be zeros, thus each derived principal components in SPCA will only be related to a small subset of the variables. Additionally, in contrast to PCA, each variable can only contribute to a small numbers of principal components in SPCA. These two features are expected to facilitate the interpretability of results. This is schematically demonstrated in Fig. 2, where $\mathbf{x_i} \in \mathbb{R}^{\boldsymbol{n}}$ is the vector of variables for the observation i. The arrows represent the loading matrix $\mathbf{V} \in \mathbb{R}^{\boldsymbol{n} \times \boldsymbol{m}}$ and map the variables to principal components $\mathbf{z_i} \in \mathbb{R}^{\boldsymbol{m}}$ where often m $\ll n$. Following this projection, a regression analysis may map the principal components to the outcome of interest, **y.** Standard statistical hypothesis testing methods can determine the significance of associations between the principal components, **z,** and the outcome, **y,** however, the dense mapping between the variables, **x,** and the principal components, **z**, mean these associations cannot be traced back to the variables. We expect SPCA to resolve this by providing a sparse loading matrix.

In order to achieve sparsity, SPCA penalizes the absolute value of the loadings at the cost of loss of information. A hyper-parameter, λ, is used to balance the trade-off between information loss and the sparsity of the loading matrix. Several implementations of SPCA have been proposed, here, we used the implementation reported by Erichson et al.[30] which uses the following formulation:

$$\min_{B} \nu(B) + \psi(B),$$

where,

$$\nu(B) := \min_{A} \frac{1}{2} \left\| X - XBA^T \right\|_F^2 \text{subject to} A^T A = I,$$

$\nu(B)$ represents the reconstruction error, $\psi(B)$ is the penalty term which could be L1 norm (LASSO), L2 norm (RIDGE), or a combination of the two (elastic net). The hyperparameter λ controls the trade-off between the reconstruction error and sparsity; a larger value of λ produces a sparser model. Hereafter this parameter is denoted by λ_{SPCA} to distinguish it from the penalty coefficient in the penalised regression model (λ_{Cox}). The data matrix is denoted by X, B is the sparsely weighted matrix and A is an orthonormal matrix.

Data processing, modelling and visualisations were performed in Python v.3.8.3 and R v.4.1.0. Cox models and related plots were obtained using the Python library Lifelines v.0.25.10, the PCA was performed using the package scikit-learn v.0.25.10 and SPCA using the SPCA R library[30]. Sankey plots were obtained using Plotly v.5.3.1.

Results

We used Cox regression to evaluate the association of environmental variables with all-cause mortality after adjusting for the aforementioned covariates. We compared the results when,

Variables	**Women** **(*n* = 206,925)**	**Men** **(*n* = 172,765)**	**All** **(*n* = 379,690)**
Age, mean (SD)	56.66 (7.94)	57.15 (8.10)	56.88 (8.02)
Age at the time of event (SD)	66.39 (7.32)	67.28 (6.87)	66.93 (7.06)
Townsend deprivation index, mean (SD)	− 1.44 (2.96)	− 1.39 (3.06)	− 1.42 (3.01)
Ethnicity: British (%)	87.98	88.84	88.37
Ethnicity: Any other white background (%)	3.56	2.60	3.12
Ethnicity: Irish (%)	2.46	2.67	2.55
Ethnicity: Indian (%)	1.20	1.37	1.28
Ethnicity: other (%)	3.53	4.43	4.58
Annual average day-time noise level (dB(A))a, mean (SD)	55.35 (4.22)	55.39 (4.27)	55.37 (4.24)
Annual average evening noise level (dB(A))[a], mean (SD)	51.61 (4.22)	51.64 (4.27)	51.62 (4.24)
Annul average night-time noise level (dB(A))[a], mean (SD)	46.53 (4.22)	46.57 (4.27)	46.55 (4.24)
Domestic garden coverage (%) within 1000m[b], mean (SD)	24.46 (11.26)	24.24 (11.23)	24.36 (11.25)
Domestic garden coverage (%) within 300m[b], mean (SD)	31.49 (14.67)	31.27 (14.70)	31.39 (14.68)
Greenspace coverage (%) within 1000m[c], mean (SD)	45.28 (21.56)	45.37 (21.44)	45.32 (21.51)
Greenspace coverage (%) within 300m[c], mean (SD)	35.40 (23.20)	35.51 (23.07)	35.45 (23.14)
Natural environment coverage (%) within 1000m[d], mean (SD)	41.32 (25.67)	41.35 (25.59)	41.33 (25.63)
Natural environment coverage (%) within 300m[d], mean (SD)	26.68 (25.31)	26.78 (25.25)	26.72 (25.28)
Water body coverage (%) within 1000m[e], mean (SD)	1.24 (2.46)	1.25 (2.45)	1.25 (2.46)
Water body coverage (%) within 300m[e], mean (SD)	0.87 (2.88)	0.89 (2.90)	0.88 (2.89)
Costal distance (meter), mean (SD)	45.39 (26.82)	45.85 (26.77)	45.60 (26.80)
NO$_2$; (μg/m^3), mean (SD)	26.67 (7.50)	26.72 (7.58)	26.69 (7.54)
NO$_x$; (μg/m^3), mean (SD)	43.89 (15.21)	44.05 (15.55)	43.96 (15.36)
PM$_{10}$; (μg/m^3), mean (SD)	16.22 (1.87)	16.23 (1.88)	16.23 (1.87)
PM$_{coarse}$; (μg/m^3)[f], mean (SD)	6.42 (0.89)	6.42 (0.89)	6.42 (0.89)
PM$_{2.5}$; (μg/m^3), mean (SD)	9.98 (1.03)	9.99 (1.05)	9.98 (1.04)
Sum of major road length within 100 m (m) [g], mean (SD)	27.25 (75.41)	28.23 (77.80)	27.70 (76.51)
Traffic intensity on nearest major road (vehicles/day)[h], mean (SD)	23,472.94 (21,322.17)	23,477.37 (21,272.41)	23,474.95 (21,299.52)
Traffic intensity on nearest road (vehicles/day)[h], mean (SD)	1480.04 (4906.16)	1516.51 (5020.38)	1496.63 (4958.49)
Years of follow-up, mean (SD)	8.08 (1.03)	8.01 (1.19)	8.05 (1.10)
Number of events	5954 (2.88%)	9042 (5.23%)	14,996 (3.95%)
Incidence rate, per 1000 person-years	4	7	5

Table 1. Descriptive summary of the study sample, environmental exposures, and outcome. [a]Average sound level pressure LAeq between the hours of 07:00 to 19:00 for day-time; 19:00–23:00 for evening; 23:00–07:00 for night-time; [b]Derived from the land use types classed as 'domestic garden' from the Generalised Land Use Database (GLUD) 2005 for England at the Census Output Area level; [c]Derived from the land use types classed as 'greenspace' from the Generalised Land Use Database (GLUD) 2005 for England at the Census Output Area level; [d]Derived from the land cover classified as 'natural environment' from the Land Cover Map (LCM) 2007; [e]Derived from the land use types classed as 'water' from the Generalised Land Use Database (GLUD) 2005 for England at the Census Output Area level; [f]PM coarse (particulate matter between 2.5 and 10 μm); Land Use Regression (LUR) estimate for annual average 2010; [g]The definition of a major road for the local road network is a road with traffic intensity greater than 5000 motor vehicles per 24 h; [h]Traffic intensity is the average total number of motor vehicles per 24 h on the nearest major road based upon a local road network.

(a) the environmental variables were plugged into the model with no transformation (***Cox model*** hereafter).

(b) L1 penalty was included in the model (***penalised Cox model*** hereafter). We varied the coefficient of the L1 penalty term, λ_{Cox}, between 0 and 2e-3 at 5e-5 intervals producing different levels of sparsity (supplementary materials: Fig. S1).

(c) The environmental variables were transformed with PCA. The number of principal components was selected to explain 90% of the variance in the data, leading to seven principal components. The Cox regression model was fitted to the resulting principal components and other covariates (***PCA Cox model*** hereafter);

(d) We repeated step (c) using SPCA. The coefficient of the L1 penalty, λ_{SPCA}, was selected to increase model parsimony. Increasing the value of λ_{SPCA} results in principal components that consist of a smaller set of variables. To facilitate interpretability, we selected λ_{SPCA} such that no two principal components share the same variable, in other words each variable at most contributes to one principal component. More details about the selection of λ_{SPCA} is included in supplementary materials (Fig. S2). The number of principal components were similarly selected to explain 90% of the data variance, leading to seven principal components (S***PCA Cox model*** hereafter).

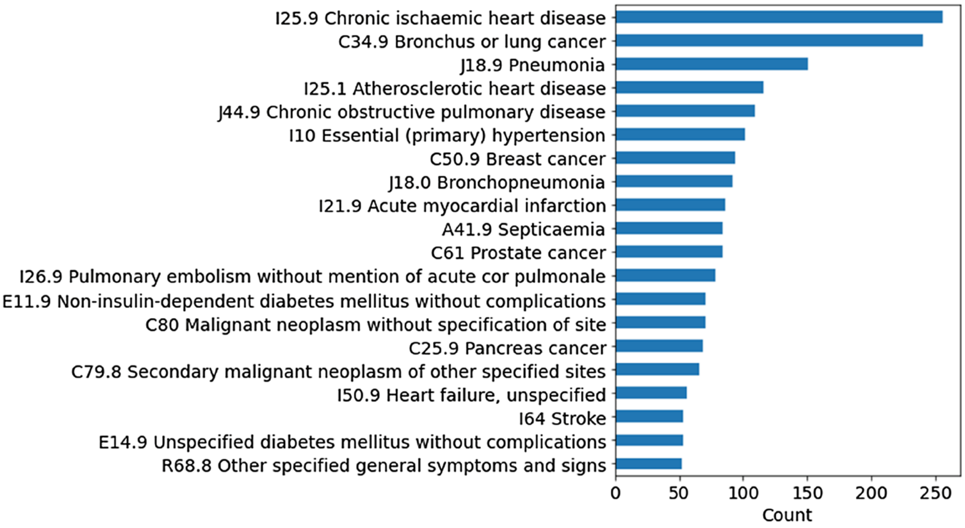

Figure 1. The top 20 primary causes of deaths within the cohort.

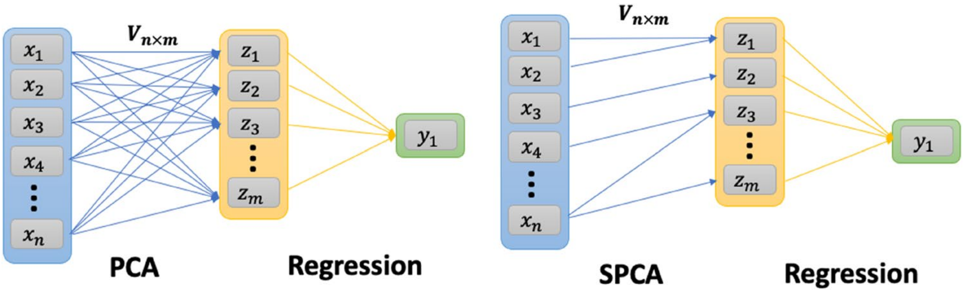

Figure 2. Schematic representation of PCA and Sparse PCA projection of the variables (**x**) to the latent space or principal components (**z**). The second layer shows a subsequent regression analysis for the outcome of interest (**y**).

The number of follow-up years was the underlying time variable for all Cox models. Prior to the analysis, all numeric variables were examined for normality and outliers. Subsequently, they were standardised and values above or below five, were set to five.

Figure 3a depicts the coefficient of the environmental variables in the Cox model. Multicollinearity in the Cox model results in high standard errors in the estimation of the coefficients, inhibiting reliable statistical inference. None of the environmental variables are found to be statistically significant. The detailed results are included in supplementary materials (Table S1). Figure 3b shows pairwise Pearson correlation between the variables. The block with high correlation coefficients pertains to the 20 environmental variables, underlining high collinearity within this class of variables. A moderate correlation is also observed between Townsend deprivation index and a number of environmental variables. A larger figure with detailed labels is included in supplementary materials (Fig. S3).

Adding the L1 penalty (penalised Cox model) attenuates the log(HR) associated with the environmental variables. Large values of λ_{Cox} result in log(HR) = 0 for all environmental variables. None of the intermediate values of λ_{Cox} produced any log(HR) values significantly different from zero at $\alpha = 5\%$. Lastly, similar to the Cox model without L1 penalty, multicollinearity led to convergence errors for several smaller values of λ_{Cox}. Figure 4 demonstrated the shrinkage of the log(HR) estimates and the 95% CI for different values of λ_{Cox}.

Figure 5, schematically compares PCA and SPCA results. The environmental variables are shown in the far left. The width of the links between the variables and the principal components are proportional with the loading coefficients. The links between the principal components and the outcome (i.e. all-cause mortality) are similarly proportional with the absolute value of the Cox coefficients (log(HR)). The associations that were found significant at $\alpha = 5\%$ are highlighted in red. Detailed results are included in supplementary materials (Table S2).

In the PCA Cox model, the seventh component has a negative association with the outcome, however, given the complex interrelationship between the variables and principal components, it is not possible to disentangle this association. On the contrary, in the SPCA Cox model, the second component has a positive association with mortality and this can be easily traced back to the three constituting variables of this component. Specifically, this component is the average of the three variables representing average level of sound pollution in daytime,

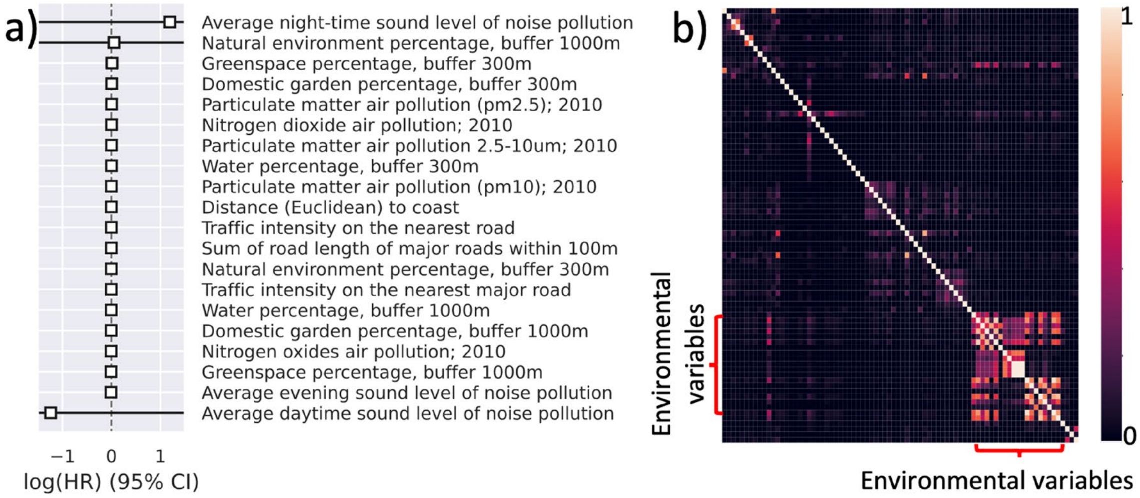

Figure 3. (**a**) The plot shows log(HR) per 1 standard deviation increase of the variables (**b**) Pairwise Pearson correlation between socioeconomic, demographic, physiological and environmental factors in a large cohort of 379,690 in the UK.

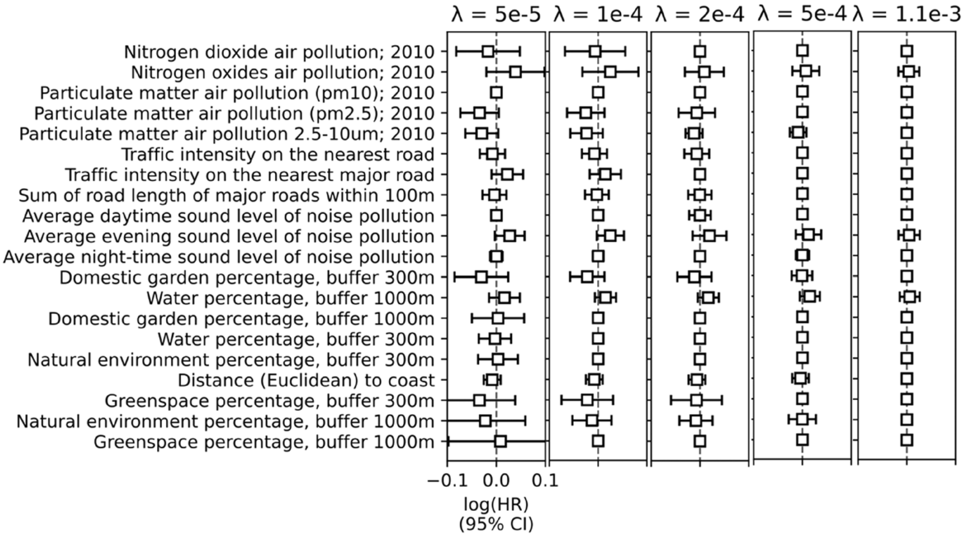

Figure 4. Log(HR) for different values of the L1 penalty coefficient (λ_{Cox}) in the penalised Cox model.

evening, and night-time. One unit change in this principal component, corresponds to 2.47 dB increase in the average daily noise pollution (details in supplementary materials) and this is associated with HR:1.017 (95% CI: 1.004–1.030). Although the list of covariates that we adjusted for is much more comprehensive than previous studies and included some traffic-related stressors that were correlated with noise pollution, our result (HR: 1.07, 95%CI: 1.02–1.13, per 10 dB increase in the average daily noise) is in agreement with the previous studies[5,7].

To further verify this association, we investigated whether it persists across different exposure levels. To this end, as suggested by the previous analysis, the three aforementioned noise pollution variables were averaged; forming a new variable that represents average daily noise pollution. This was then categorised into deciles and the hazard ratios were calculated for the nine top deciles relative to lowest decile. The lowest decile represents noise pollution levels between 46.72 and 47.23 dB. To address the multicollinearity of the environmental covariates, the remaining 17 environmental variables were transformed to principal components explaining 0.92 percent of the variance. The model was adjusted for all other covariates. The results are depicted in Fig. 6, showing an upward trend which underlines the plausibility of a causal link. Descriptive summary of the subpopulations in each category and more details about the model is included in the supplementary materials (Table S3 and 4).

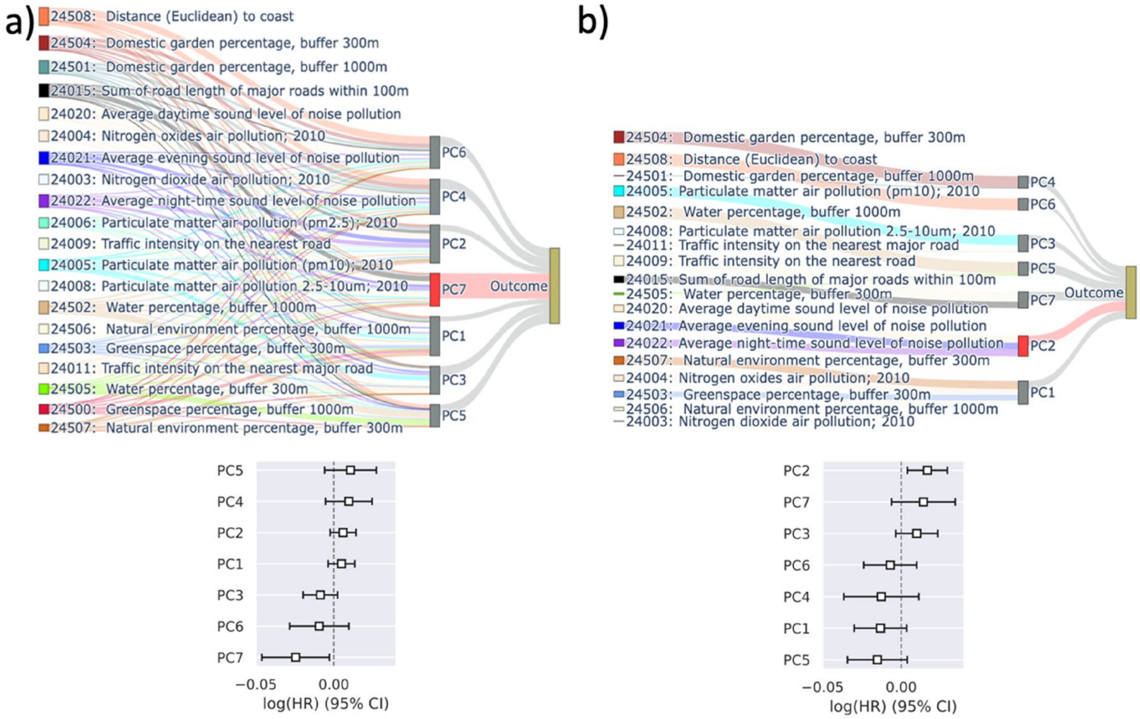

Figure 5. Schematic representation of the association of environmental variables with all-cause mortality using a two-stage regression analysis (**a**) with PCA and (**b**) with SPCA. In the first stage (i.e. dimensionality reduction), the variables are transformed to principal components. In the second stage, a Cox model was used to investigate the association of the transformed variables and all-cause mortality.

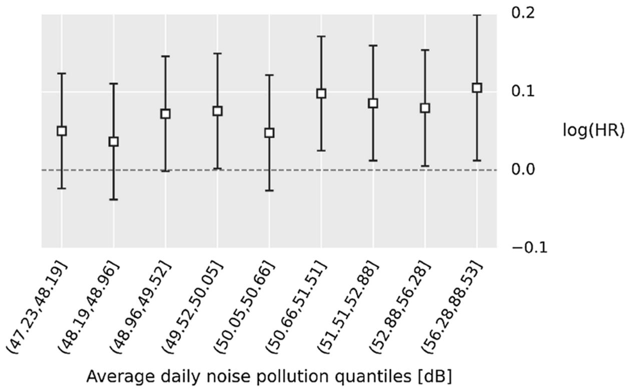

Figure 6. Log(Hazard Ratio) of all-cause mortality for different noise pollution exposure deciles compared to the lowest decile, i.e. (46.72, 47.23], after adjusting for socioeconomic, demographic, environmental, and physiological, and behavioural covariates.

Discussion

Main findings. The key finding of this study is the integration of SPCA in regression analysis provides a promising approach for conducting inference in the presence of multicollinearity. Multicollinearity leads to erroneous estimation of the coefficients and broad confidence intervals in regression models. This increases the likelihood of Type II errors. As a result, some important associations may remain concealed. As a case study, we investigated the association of various correlated environmental factors and all-cause mortality, adjusting for other established risk factors. SPCA resulted in a sparse and interpretable grouping of the environmental variables. For instance, all three noise related predictors were combined into one principal component. Additionally, the sparsity of the transformation enabled us to trace the statistical significance of the principal components back to the variables of interest.

Interpretation of findings in the context of previous studies. PCA, as a descriptive analysis tool is one of the oldest and most commonly used techniques for reducing the dimensionality of data[31]. But the lack of interpretability of the derived representations, i.e. principal components, has been recognised as one of its major drawbacks. As shown in our results, in PCA, the entangled relationship between the principal components and the variables hinders the interpretation of findings. While some seek to mitigate this issue by deselecting non-important variables[32] or selecting variables more relevant to outcomes using supervised methods[33,34], such interventions are not appropriate for statistical hypothesis testing where all relevant covariates should be adjusted for regardless of their contribution to predictive performance.

Over the years, other dimensionality reduction methods have been widely applied in different disciplines. Random Projection[35], Dictionary Learning[36], Factor Analysis, Independent Component Analysis[37], Non-negative Matrix Factorization (NMF)[38] are examples of these methods. Recently, Autoencoders including Denoising Autoencoder[39] and Variational Autoencoder[40] are increasingly used to learn a low dimensional representation of the input variables. But similar to PCA, the common limitation of these dimensionality reduction methods is the entangled relationship between the variables and the low dimensional representations. Enforcing sparsity in the transformation is recognised as an effective way to address this problem[41]. Inspired by this we investigated the use of SPCA for statistical hypothesis testing in the context of environmental health research and showed promising results. In light of the findings, we conclude that the integration of SPCA in statistical inference is a simple, computationally-efficient strategy for big data investigations when multicollinearity could lead to erroneous results.

Previously, environmental epidemiology studies have adopted dimensionality reduction methods, as well as one-stop methods such as Bayesian profile regression[42] to perform both dimensionality reduction and regression analysis for multiple pollutants. Some studies using PCA had previously identified a subset of air pollutants that were associated with mortality[34,43]. However, no studies have applied these statistical techniques to adjust for the wide range of environmental and non-environmental covariates that we considered in our analysis[44]. Neighbourhood-wide[45] and environment-wide[24] association studies (N/EWAS) have also been applied to high dimensional data in environmental epidemiology. These methods are inspired by genome-wide association studies[46] and their resources-intensiveness -in terms of data and computational power- hinders their wider adoption.

Our analysis led to a clear pattern of association between noise pollution and all-cause mortality. Noteworthy, the three indicators of noise pollution, day-time, evening, and night time noise levels, were combined into one principal component, all with the same weights. The resulting principal component (or a 2.47 dB increase in the average daily noise pollution) was associated with a HR:1.017 (95%CI:1.004–1.030) for all-cause mortality. This is translated to a HR: 1.07, 95%CI: 1.02–1.13 as per 10 dB increase in average noise level, in line with the only other study that showed a positive significant association between daily road traffic noise exposure and all-cause mortality (HR: 1.08, 95%CI: 1.04–1.12)[5]. A previous study in London reported the association between daytime noise and all-cause mortality in areas with noise pollution level greater than 60 dB compared to areas with noise pollution level less than 55 dB RR: 1.04 (95%CI: 1.00–1.07)[7]. While the hazard ratio calculated in our study is not directly comparable to the aforementioned studies, due to differences in the populations, study designs, data processing and covariates, the consistency of the findings are reassuring. Although the inclusion of correlated covariates can attenuate the significance of association, our results are largely in agreement with these studies, suggesting independent of gaseous pollutant, traffic-related stressors and other determinants, noise level is an important risk factor. Nonetheless, number of studies investigating the epidemiological link between road traffic noise exposure and all-cause mortality outcomes remains few, with a recent meta-analysis showing a weak association by pooling only five studies (HR: 1.01, 95%CI: 0.98–1.05)[47].

A subsequent exposure–response analysis showed that the four highest exposure deciles are associated with significant risk of all-cause mortality compared to the lowest exposure decile. Specifically, the top four average daily noise exposure deciles (50.66, 51.51], (51.51, 52.88], (52.88, 56.28], (56.28,88.53] dB were consistently associated with significant increase in all-cause mortality of HR:1.10 (95%CI: 1.03–1.19), 1.09 (95%CI: 1.01–1.17), 1.08 (95%CI: 1.01–1.17), 1.11 (95%CI: 1.01–1.22) compared to the lowest decile of (46.23, 47.23] dB. It should be noted that, 8 of the top 20 causes of death included in this study were cardiometabolic diseases. The finding of our trend analysis is similar to what was previously reported for the cardiovascular disease mortality, suggesting a possible effect threshold may start at around 50–53 dB[47–49].

Limitations and future works.

The key strengths of our analysis are, a large cohort, adjustment for a comprehensive list of environmental exposures, including correlated traffic-related exposures, which was facilitated by our methodological approach. This study has limitations. Firstly, we did not account for any potential non-linear exposure–response relationship. The inclusion of non-linear and interaction terms could reduce the risk of residual confounding. However, our primary objective was to study the usability of SPCA as a simple, computationally efficient and interpretable method to address collinearity. Second, as we already noted, exposure misclassification is inevitable for this type of study. Typically, if there was a true association with the health outcome, the effect estimates would be biased toward null for a classic random error. Third, SPCA approach is essentially a data-driven method without a priori variables hypotheses, without considering causal structures among the variables and/or variable-outcome links. In our study, 20 environmental exposures from UK Biobank were reduced in dimensionality using SPCA and were all used in the Cox regression under the assumptions of a causal structure linking each exposure and the outcome and the assumption that the exposures are independent of one another. However, in reality, some exposures may be on a specific causal pathway (e.g. traffic intensity–air pollution–mortality). It is beyond the scope of current study to investigate this complex causal structure which indeed requires a careful consideration of the causal inference analysis framework. Taking together all these limitations, the findings generated from our SPCA analysis are mainly exploratory and neither infers any potential causal relationship nor biological plausibility.

Conclusion

This study demonstrated that SPCA is a viable analytical approach to address, and enable interpretability of multiple environmental stressors-health associations. Using this method, our study further verified existing evidence on the association between noise as an important risk factor for adverse health outcomes in the UK Biobank. The strength of our analysis was observing this association even after adjusting for comprehensive list correlated stressors.

Data availability

The data that support the findings of this study are available from the UK Biobank but restrictions apply to the availability of these data, which were used under license for the current study. The raw data are only available to approved researchers via the UK Biobank.

Received: 17 January 2022; Accepted: 13 May 2022
Published online: 02 June 2022

References

1. Beelen R, Stafoggia M, Raaschou-Nielsen O, *et al.* Long-term exposure to air pollution and cardiovascular mortality: an analysis of 22 European cohorts. *Epidemiology* **25**(3) (2014).
2. Hansell, A. *et al.* Historic air pollution exposure and long-term mortality risks in England and Wales: Prospective longitudinal cohort study. *Thorax* **71**(4), 330–338. https://doi.org/10.1136/thoraxjnl-2015-207111 (2016).
3. Rajagopalan, S., Al-Kindi, S. G. & Brook, R. D. Air pollution and cardiovascular disease. *J. Am. Coll. Cardiol.* https://doi.org/10.1016/j.jacc.2018.07.099 (2018).
4. Liu, C. *et al.* Ambient particulate air pollution and daily mortality in 652 cities. *N. Engl. J. Med.* **381**(8), 705–715. https://doi.org/10.1056/NEJMoa1817364 (2019).
5. Thacher, J. D. *et al.* Long-term residential road traffic noise and mortality in a Danish cohort. *Environ. Res.* https://doi.org/10.1016/j.envres.2020.109633 (2020).
6. Kupcikova, Z., Fecht, D., Ramakrishnan, R., Clark, C. & Cai, Y. S. Road traffic noise and cardiovascular disease risk factors in UK Biobank. *Eur. Heart J.* https://doi.org/10.1093/eurheartj/ehab121 (2021).
7. Halonen, J. I. *et al.* Road traffic noise is associated with increased cardiovascular morbidity and mortality and all-cause mortality in London. *Eur. Heart J.* https://doi.org/10.1093/eurheartj/ehv216 (2015).
8. Billionnet, C., Sherrill, D. & Annesi-Maesano, I. Estimating the health effects of exposure to multi-pollutant mixture. *Ann. Epidemiol.* **22**(2), 126–141. https://doi.org/10.1016/j.annepidem.2011.11.004 (2012).
9. Sun, Z. *et al.* Statistical strategies for constructing health risk models with multiple pollutants and their interactions: Possible choices and comparisons. *Environ. Health* **12**(1), 85. https://doi.org/10.1186/1476-069X-12-85 (2013).
10. Westerhuis, J. A., Kourti, T. & MacGregor, J. F. Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemom.* **12**(5), 301–321. https://doi.org/10.1002/(SICI)1099-128X(199809/10)12:5%3c301::AID-CEM515%3e3.0.CO;2-S (1998).
11. Worley, B. & Powers, R. Multivariate analysis in metabolomics. *Curr. Metab.* **1**(1), 92–107. https://doi.org/10.2174/2213235X11301010092 (2012).
12. Marin, J.-M., Mengersen, K. & Robert, C. P. Bayesian modelling and inference on mixtures of distributions. *Handb. Stat.* **25**, 459–507. https://doi.org/10.1016/S0169-7161(05)25016-2 (2005).
13. Nasserinejad, K., van Rosmalen, J., de Kort, W. & Lesaffre, E. Comparison of criteria for choosing the number of classes in bayesian finite mixture models. *PLoS ONE* **12**(1), e0168838. https://doi.org/10.1371/journal.pone.0168838 (2017).
14. Zhang, Z. & Castelló, A. Principal components analysis in clinical studies. *Ann. Transl. Med.* https://doi.org/10.21037/atm.2017.07.12 (2017).
15. Homenauth, E., Kajeguka, D. & Kulkarni, M. A. Principal component analysis of socioeconomic factors and their association with malaria and arbovirus risk in Tanzania: A sensitivity analysis. *J. Epidemiol. Community Health* https://doi.org/10.1136/jech-2017-209119 (2017).
16. Greenfield, B. K., Rajan, J. & McKone, T. E. A multivariate analysis of CalEnviroScreen: Comparing environmental and socioeconomic stressors versus chronic disease. *Environ. Health A Glob. Access Sci. Source* https://doi.org/10.1186/s12940-017-0344-z (2017).
17. Welker-Hood, L. K., Hynes, H. P., Heeren, T., Snell, J. & Helmes, D. Principal component analysis as a new methodology for developing sensitive exposure measures for building dampness. *Epidemiology* https://doi.org/10.1097/00001648-200407000-00415 (2004).
18. Zou, H., Hastie, T. & Tibshirani, R. Sparse principal component analysis. *J. Comput. Graph. Stat.* **15**(2), 265–286 (2006).
19. Floud, S. *et al.* Exposure to aircraft and road traffic noise and associations with heart disease and stroke in six European countries: A cross-sectional study. *Environ. Health* **12**(1), 89. https://doi.org/10.1186/1476-069X-12-89 (2013).
20. Sudlow, C. *et al.* UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* https://doi.org/10.1371/journal.pmed.1001779 (2015).
21. Eeftens, M. *et al.* Spatial variation of PM2.5, PM10, PM2.5 absorbance and PMcoarse concentrations between and within 20 European study areas and the relationship with NO2 - Results of the ESCAPE project. *Atmos. Environ.* https://doi.org/10.1016/j.atmosenv.2012.08.038 (2012).
22. Beelen, R. *et al.* Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – The ESCAPE project. *Atmos. Environ.* **72**, 10–23. https://doi.org/10.1016/j.atmosenv.2013.02.037 (2013).
23. Kephalopoulos, S. *et al.* Advances in the development of common noise assessment methods in Europe: The CNOSSOS-EU framework for strategic environmental noise mapping. *Sci. Total Environ.* **482–483**, 400–410. https://doi.org/10.1016/j.scitotenv.2014.02.031 (2014).
24. Sheehan, A., Freni Sterrantino, A., Fecht, D., Elliott, P. & Hodgson, S. Childhood type 1 diabetes: An environment-wide association study across England. *Diabetologia* https://doi.org/10.1007/s00125-020-05087-7 (2020).
25. Wheeler, B. W., White, M., Stahl-Timmins, W. & Depledge, M. H. Does living by the coast improve health and wellbeing?. *Health Place* **18**(5), 1198–1201. https://doi.org/10.1016/j.healthplace.2012.06.015 (2012).
26. Cai, Y. *et al.* Road traffic noise, air pollution and incident cardiovascular disease: A joint analysis of the HUNT, EPIC-Oxford and UK Biobank cohorts. *Environ. Int.* **114**, 191–201. https://doi.org/10.1016/j.envint.2018.02.048 (2018).
27. Doiron, D. *et al.* Air pollution, lung function and COPD: Results from the population-based UK Biobank study. *Eur. Respir. J.* **54**(1), 1802140. https://doi.org/10.1183/13993003.02140-2018 (2019).
28. Gulliver, J., de Hoogh, K., Hansell, A. & Vienneau, D. Development and back-extrapolation of NO2 land use regression models for historic exposure assessment in Great Britain. *Environ. Sci. Technol.* **47**(14), 7804–7811. https://doi.org/10.1021/es4008849 (2013).

29. Mensah, G. A., Roth, G. A. & Fuster, V. The global burden of cardiovascular diseases and risk factors: 2020 and beyond. *J. Am. Coll. Cardiol.* https://doi.org/10.1016/j.jacc.2019.10.009 (2019).
30. Erichson, N B., Zheng, P., Manohar, K., Brunton, S. L., Kutz, J. N. & Aravkin, A. Y. Sparse principal component analysis via variable projection. *arXiv*. Published online 2018.
31. Jolliffe, I. T. & Cadima, J. Principal component analysis: A review and recent developments. *Philos. Trans. A Math. Phys. Eng. Sci.* **374**(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202 (2016).
32. Johnstone, I. M. & Lu, A. Y. On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Stat. Assoc.* **104**(486), 682–693. https://doi.org/10.1198/jasa.2009.0121 (2009).
33. Barshan, E., Ghodsi, A., Azimifar, Z. & Zolghadri, J. M. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recogn.* **44**(7), 1357–1371. https://doi.org/10.1016/j.patcog.2010.12.015 (2011).
34. Roberts, S. & Martin, M. A. Using supervised principal components analysis to assess multiple pollutant effects. *Environ. Health Perspect.* **114**(12), 1877–1882. https://doi.org/10.1289/ehp.9226 (2006).
35. Bingham, E. & Mannila, H. Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 245–250 (KDD '01. Association for Computing Machinery, 2001). https://doi.org/10.1145/502512.502546
36. Mairal, J., Bach, F., Ponce, J. Sapiro G. Online dictionary learning for sparse coding. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. 689–696 (Association for Computing Machinery, 2009). https://doi.org/10.1145/1553374.1553463
37. Hyvärinen, A. & Oja, E. Independent component analysis: Algorithms and applications. *Neural Netw.* **13**(4), 411–430. https://doi.org/10.1016/S0893-6080(00)00026-5 (2000).
38. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**(6755), 788–791. https://doi.org/10.1038/44565 (1999).
39. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. & Manzagol, P. A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **11**, 3371–3408 (2010).
40. Pu, Y. *et al.* Variational autoencoder for deep learning of images, labels and captions. In *Advances in Neural Information Processing Systems* (eds Lee, D. D. *et al.*) 2352–2360 (Curran Associates Inc, New York, 2016).
41. Hoyer, P. O. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.* **5**, 1457–1469 (2004).
42. Pirani, M. *et al.* Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. *Environ. Int.* **79**, 56–64. https://doi.org/10.1016/j.envint.2015.02.010 (2015).
43. Baxter, L. K., Duvall, R. M. & Sacks, J. Examining the effects of air pollution composition on within region differences in PM25 mortality risk estimates. *J. Expo. Sci. Environ. Epidemiol.* **23**(5), 457–465. https://doi.org/10.1038/jes.2012.114 (2013).
44. Chang, T. S. *et al.* Sparse modeling of spatial environmental variables associated with asthma. *J. Biomed. Inform.* **53**, 320–329. https://doi.org/10.1016/j.jbi.2014.12.005 (2015).
45. Lynch, S. M. *et al.* A neighborhood-wide association study (NWAS): Example of prostate cancer aggressiveness. *PLoS ONE* **12**(3), 1–13. https://doi.org/10.1371/journal.pone.0174548 (2017).
46. Zheng, Y. *et al.* Design and methodology challenges of environment-wide association studies: A systematic review. *Environ. Res.* **183**, 109275. https://doi.org/10.1016/j.envres.2020.109275 (2020).
47. Cai, Y., Ramakrishnan, R. & Rahimi, K. Long-term exposure to traffic noise and mortality: A systematic review and meta-analysis of epidemiological evidence between 2000 and 2020. *Environ. Pollut.* https://doi.org/10.1016/j.envpol.2020.116222 (2021).
48. Vienneau, D. *et al.* Transportation noise exposure and cardiovascular mortality: 15-years of follow-up in a nationwide prospective cohort in Switzerland. *Environ. Int.* **158**, 106974. https://doi.org/10.1016/j.envint.2021.106974 (2022).
49. Vienneau, D., Schindler, C., Perez, L., Probst-Hensch, N. & Röösli, M. The relationship between transportation noise exposure and ischemic heart disease: A meta-analysis. *Environ. Res.* **138**, 372–380. https://doi.org/10.1016/j.envres.2015.02.023 (2015).

Author contributions

M.M., Y.Z. and Y.C. conceived the idea. M.N. and A.H. contributed to processing the data and the selection of variables. M.M. carried out the analysis and wrote the manuscript. All authors reviewed the manuscript.

Funding

This work is supported by the PEAK Urban programme, funded by UKRI's Global Challenge Research Fund (Grant Number: ES/P011055/1).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-13362-3.

Correspondence and requests for materials should be addressed to M.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.