Environment International 158 (2022) 107015

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Machine learning approaches to characterize the obesogenic urban exposome

Haykanush Ohanyan [a,b,c,*], Lützen Portengen [b], Anke Huss [b], Eugenio Traini [b], Joline W. J. Beulens [a,c,d], Gerard Hoek [b], Jeroen Lakerveld [a,c], Roel Vermeulen [b]

[a] Department of Epidemiology and Data Science, Amsterdam Public Health Research Institute, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, Noord-Holland, the Netherlands
[b] Institute for Risk Assessment Sciences, Utrecht University, Utrecht, Utrecht, the Netherlands
[c] Upstream Team, www.upstreamteam.nl. Amsterdam UMC, VU University Amsterdam, Amsterdam, Noord-Holland, the Netherlands
[d] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherland

ARTICLE INFO

Handling Editor: Shoji F. Nakayama

Keywords:
Exposome
Random forest
Extreme gradient boosting (XGBoost)
Shapley values
Socioeconomic position
Air pollution

ABSTRACT

Background: Characteristics of the urban environment may contain upstream drivers of obesity. However, research is lacking that considers the combination of environmental factors simultaneously.
Objectives: We aimed to explore what environmental factors of the urban exposome are related to body mass index (BMI), and evaluated the consistency of findings across multiple statistical approaches.
Methods: A cross-sectional analysis was conducted using baseline data from 14,829 participants of the Occupational and Environmental Health Cohort study. BMI was obtained from self-reported height and weight. Geocoded exposures linked to individual home addresses (using 6-digit postcode) of 86 environmental factors were estimated, including air pollution, traffic noise, green-space, built environmental and neighborhood sociodemographic characteristics. Exposure-obesity associations were identified using the following approaches: sparse group Partial Least Squares, Bayesian Model Averaging, penalized regression using the Minimax Concave Penalty, Generalized Additive Model-based boosting Random Forest, Extreme Gradient Boosting, and Multiple Linear Regression, as the most conventional approach. The models were adjusted for individual sociodemographic variables. Environmental factors were ranked according to variable importance scores attributed by each approach and median ranks were calculated across these scores to identify the most consistent associations.
Results: The most consistent environmental factors associated with BMI were the average neighborhood value of the homes, oxidative potential of particulate matter air pollution (OP), healthy food outlets in the neighborhood (5 km buffer), low-income neighborhoods, and one-person households in the neighborhood. Higher BMI levels were observed in low-income neighborhoods, with lower average house values, lower share of one-person households and smaller amount of healthy food retailers. Higher BMI levels were observed in low-income neighborhoods, with lower average house values, lower share of one-person households, smaller amounts of healthy food retailers and higher OP levels. Across the approaches, we observed consistent patterns of results based on model's capacity to incorporate linear or nonlinear associations.
Discussion: The pluralistic analysis on environmental obesogens strengthens the existing evidence on the role of neighborhood socioeconomic position, urbanicity and air pollution.

1. Introduction

Obesity is a chronic, complex, multi-causal condition that increases the risk of a range of non-communicable diseases (Chooi et al., 2019; WHO. Obesity and Overweight. Factsheet, 2020). Whereas the essential cause of obesity is an imbalance between intake and expenditure of energy, the multiple underlying determinants of this imbalance are complex, extending beyond individual-level factors to contextual factors. The life in modern urban environments, with a large availability of energy-dense products, motorized transport, sedentary work and lack of

* Corresponding author at: de Boelelaan 1089A, 1081 BT Amsterdam, the Netherlands.
E-mail address: h.ohanyan@amsterdamumc.nl (H. Ohanyan).

https://doi.org/10.1016/j.envint.2021.107015
Received 18 July 2021; Received in revised form 26 November 2021; Accepted 29 November 2021
Available online 3 December 2021

physical activity has a huge impact on the growing rate of obesity worldwide (Lakerveld & Mackenbach, 2017).

The urban exposome can be defined as the continuous spatiotemporal variation of quantitative and qualitative indicators associated with the urban external and internal domains that shape the quality of life and the health of urban populations, using small city areas, such as neighborhoods (Andrianou & Makris, 2018). Some of the risk factors of urban exposome are well known and extensively studied, while other factors were not readily available until recent years. Among these less commonly studied factors are for instance, radiofrequency electromagnetic field (RF-EMF), light at night and meteorology. Mobile phone base station RF-EMF exposure has been suggested to be associated with sleep (Martens et al., 2017a). We hypothesized that RF-EMF exposure or outdoor exposure to high intensity electric lighting at night could have potential disruptive effect on sleep and therefore might be related to an increased BMI, as sleep deprivation is a known risk factor for obesity (Patel & Hu, 2008). High temperatures, in turn, may alter lifestyle behaviors. For example, high temperatures often lead individuals to be less physically active or discourage home-cooking, hence increasing the frequency of eating out, which has been linked with increased body weight (Bezerra et al., 2012; Gildner & Levy, 2021).

Previous studies indicate a consistent relationship between socioeconomic position (SEP) (Lam et al., 2021; McLaren, 2007). Studies from the Unite States or Australia consistently found a link with urban sprawl, - often based on population density and measures of land use mix -, and obesity. This trend is, however, less consistent in Europe (Mackenbach et al., 2014). Inconsistent findings are also reported with regards to other urban exposome attributes, such as green space, air pollution or noise exposure (An et al., 2018; Cai et al., 2020; De la Fuente et al., 2020). One of the explanations for these inconsistent associations may be the lack of adequate control for the interactions between exposures or combination of exposures. More specifically, two issues are especially important: 1) most research to date used a reductionistic approach, by focusing on single or a limited number of environmental exposures, 2) linear regression analysis was applied conventionally to link environmental factors to adult obesity. Although such studies are valuable to investigate a specific hypothesis regarding an underlying pathway from environmental exposure to obesity, it is unlikely that such methodological approaches do justice to the complexity of real life. Such a "naïve" linear regression model is limited in the context of multiple, highly correlated variables, because it can result in unstable parameter estimates with large standard errors (Stafoggia et al., 2017). Moreover, it fails to address nonlinear or complex non-additive associations, potentially present in multi-exposure models (Casson & Farmer, 2014). For these reasons, more advanced statistical methods are advocated to capture the complex association of multiple environmental exposures and obesity. However, it should be noted that such methods have a major application for prediction tasks, but not for hazard characterization. Therefore, currently in epidemiology the transition from conventional statistical inference based on hypothesis testing to more advanced methods lacks both methodological and applied framework.

The twin-revolution of data availability and advances in data science methods make it possible to pioneer and apply those methods in this field of research. Originally the term "machine learning" was defined as a program that learns automatically from data. However, this definition is very generic and could cover nearly any form of data-driven approach. For a given statistical method, the more we reduce the assumptions, the more the algorithm moves towards machine learning, but there is never a specific threshold where a model suddenly becomes "machine learning" (Beam & Kohane, 2018).

Currently, an increasing number of studies in environmental epidemiology are trying to capture numerous exposures as part of the exposome (Wild, 2012). The exposome approach has been applied previously to study obesity in children, but not in adults (Vrijheid et al., 2020). Some recent simulation studies have compared the performance of different machine learning models to identify exposome-health associations (Agier et al., 2016; Barrera-Gómez et al., 2017; Lenters et al., 2018; Sun et al., 2013). Results suggest that there is no one-size-fits-all model, because data-driven approaches have an ability to adapt to the information contained in the data, and each of them makes a different use of this information with any given dataset. Thus, it is still not known which environmental exposures of the urban exposome are consistently associated with adult BMI and which statistical approach (es) can deal better with the complex data and provide meaningful/interpretable output. According to the classification provided by Stafoggia et al. the multi-exposure models which are able to deal with a wide variety of correlated factors, may be grouped by their capacities to reduce data dimensionality, select more important variables within a group of highly correlated variables or cluster the observations (Stafoggia et al., 2017).

We therefore aimed to explore what environmental factors of the urban exposome are related to BMI, and evaluated the consistency of findings across multiple statistical approaches. We selected at least one statistical method from each of the three groups of approaches (dimension reduction, variable selection, clustering) to effectively combine the strengths and compensate the limitations of each method. We also compared the results of these approaches with a 'traditional' multiple linear regression model. The environmental factors representing the urban exposome included several groups of exposures: air pollution, road traffic noise, green space, light at night, radiofrequency electromagnetic field, meteorology, diverse socioeconomic and demographic factors of the neighborhood, food environment, urbanicity, road safety, access to neighborhood facilities and quality of drinking water.

2. Material and methods

2.1. Study design and population

We used baseline data of the prospective, population based Occupational and Environmental Health Cohort (AMIGO) study for a cross-sectional analysis. Participants of AMIGO were recruited from the general population in the Netherlands, through the Dutch national general practitioners network: the NIVEL Primary Care Database (NIVEL Primary Care Registry, 2021). A maximum of one adult per household was randomly selected from the NIVEL database. Overall, 14,829 adult cohort members (16% of those invited) consented and filled in the online baseline questionnaire. Participants were aged between 31 and 65 years during the data collection period (2011–2012). A detailed description of the recruitment process, a flowchart of participant data and the ethical approval is given elsewhere (Slottje et al., 2015). The data of 14,781 participants having complete cases on the outcome measure, were analyzed in the current study.

2.2. Outcome definition and covariates

Self-reported height and weight from the baseline questionnaire were used to calculate body mass index kg/m^2. Overweight and obesity were defined as BMI $\geq$ 25 kg/m^2 and BMI $\geq$ 30 kg/m^2, respectively (WHO. Obesity and Overweight. Factsheet, 2020). The following individual socio-demographic characteristics were considered as covariates in this study: age, sex (male/female), country of birth (Netherlands/other), country of birth of mother (Netherlands/other), country of birth of father (Netherlands/other), civil state (with/without a partner), current education (high, defined as college/university degree or higher/low or medium, including vocational education/community college or vocational/high school), employment status (employed/unemployed) and smoking (yes/no).

2.3. Characterization of the urban exposome

A large set of environmental factors were estimated by geospatial

models, monitoring stations, satellite data, and land use databases and linked to all respondents, using the geocoded residential addresses. The addresses were geocoded by the 6-digit postal codes in Netherlands, which cover a few houses at one side of a street section. Exposures were measured in a buffer of 100 m-20 km depending on the variable. Annual average exposure estimates were calculated for the data collection period of the baseline questionnaire (2011–2012) or the nearest date around in a range of five years.

We excluded factors that were judged uninformative based on the following reasons: i) variables with very low variability, e.g. when most observations (>99%) had the same value, assessed by histograms and descriptive statistics (see the list in supplementary material, Table A.1) ii) or if two variables were correlated at a level of $r_{spearman} \geq 0.95$. In the latter case only one of the correlated variables was included and considered as a proxy for the other variable (Table A.2). Overall, we retained 86 out of 99 exposures across a total of 11 exposure constructs: air pollution (19 factors), road traffic noise (1 factor), mobile phone base station radiofrequency electromagnetic field (1 factor), green space density (2 factors), outdoor light at night (1 factor), meteorology (2 factors), quality of the drinking water (29 factors), socio-demographic characteristics of the neighborhood (18 factors), food environment (3 factors), built environment (9 factors) and road safety (1 factor). The measurement of these constructs and variables are detailed in next paragraphs.

2.3.1. Air pollution

Land use regression (LUR) models were used to assess the air pollution exposure. The models provided data on particulate matter with aerodynamic diameter<10 μm (PM_{10}), <2.5 μm ($PM_{2.5}$), between 10 and 2.5 μm ($PM_{10-2.5}$, the coarse fraction of PM), and smaller than 100 nm (Ultrafine particles (UFP)), as well as the absorbance of $PM_{2.5}$ (a measure of black carbon particles) and NO_2, NO_X (sum of NO and NO_2). The median model explained variance (R^2) of LUR models was 71% for $PM_{2.5}$, 89% for $PM_{2.5\ absorbance}$, 68% for PM_{coarse}, 82% for NO_2 and 78% for NO_X (Beelen et al., 2013; Eeftens et al., 2012). Eight elemental components of particulate matter (copper, iron, potassium, nickel, sulfur, silicon, vanadium, zinc), were estimated in both PM_{10} and $PM_{2.5}$. Good models were developed for copper, iron, and zinc in both fractions (PM_{10} and $PM_{2.5}$) explaining on average between 67 and 79% of the concentration variance (R^2) with a large variability between areas. Models for vanadium and sulfur in the PM_{10} and $PM_{2.5}$ fractions and silicon, nickel, and potassium in the PM_{10} fraction performed moderately with R^2 ranging from 50 to 61%. silicon, nickel, and potassium models for $PM_{2.5}$ performed poorest with R^2 under 50% (De Hoogh et al., 2013). The oxidative potential (OP) was estimated in $PM_{2.5}$, by two metrics – electron spin resonance (OPESR) and dithiothreitol (OPDTT). The explained variance for these models were 67% and 60% accordingly (Yang et al., 2015). References on detailed description of each LUR model are provided in Table A.3.

2.3.2. Road traffic noise

Noise model maps were used from the Standard Model Instrumentation for Noise Assessments for the year 2016 (Baliatsas et al., 2016; Martens et al., 2018). For each participant the exposure to the road traffic noise levels (dB) was estimated over a whole day period (L_{den}), overweighing sound levels during evening and night, as the nuisance perception is higher during more quiet hours of the day. Following the recommendations of The Environmental Noise Directive, we used the threshold of 55 dB to define high and low noise levels (European Environment Agency, 2018).

2.3.3. Radiofrequency electromagnetic field

With regard to the exposure of radiofrequency electromagnetic field (RF-EMF), for each geocoded address the model estimates the total sum of the exposures to downlink field strength of GSM900 (Global System for Mobile Communication), GSM1800, and UMTS (Universal Mobile Telecommunications System) (mW/m2) (Martens et al., 2017b).

2.3.4. Green space

The neighborhood greenness was estimated for 100 m and 1000 m buffers around each respondent's home. Satellite images from Landsat 8, captured in September 2016 were used to generate the Normalized Difference Vegetation Index (NDVI), which quantifies vegetation density (Rhew et al., 2011).

2.3.5. Outdoor light at night

To measure the exposure to outdoor artificial light at night, the global low-light imaging data from Earth's surface was collected by the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) (Elvidge et al., 2017). The VIIRS-DNB map from 2015 was used to assign an exposure value to the home address.

2.3.6. Meteorology

An urban or metropolitan area usually tends to be warmer than its surrounding rural areas. This phenomenon is called "urban heat island effect". Satellite pictures from Landsat 8 were used to estimate the surface temperatures as measured from space on a hot day (20 July 2016). During heatwaves, the urban heat island effect is best assessed. In addition, we used a map to assign urban heat (difference in temperature to surrounding less urbanized areas), which was modeled using data from wind speed, population density and green space, as described in the heat island map from the Dutch National Institute for Public Health and the Environment (RIVM)(Remme, 2017).

2.3.7. Drinking water quality

To estimate the quality of drinking water the data was collected from drinking water quality map of RIVM, for annual average values of 2012 (Quality of Drinking Water in Netherlands, 2018). Twenty-eight water compounds were measured in the closest tested pump to the geocoded addresses. The measurements were done for two main types of contaminants: chemical and bacterial. Chemical contaminants, in turn, were from three sources: agricultural runoff (pesticides: ammonia, nitrate/nitrite, phosphate), industrial runoff (heavy metals: arsenic (can also be natural), mercury, cadmium, chromium), as well as fecal matter and urine. Biological compounds included coliform bacteria and *Escherichia coli*. The chlorine (disinfectant), which is used to reduce the bacterial contamination, and the acidity and turbidity of water were also measured.

2.3.8. Characteristics of neighborhoods

All geocoded residential addresses were linked to the neighborhood map of the Land Use Database of Statistics Netherlands for 2011, in order to collect a large set of neighborhood socio-demographic and built environmental data (Statistics Netherlands, 2012). The urbanicity level was calculated by Statistics Netherlands based on the density of addresses. The categories of urbanicity were defined as follows: 1 = very highly urban ≥ 2,500 addresses per km^2; 2 = highly urban 1 500–2 500 addresses per km^2; 3 = moderately urban 1,000–1,500 addresses per km^2; 4 = few urban 500–1,000 addresses per km^2; 5 = non-urban < 500 addresses per km^2 (Statistics Netherlands, 2012). Several socio-demographic characteristics of the neighborhoods were also collected for each participant, e.g., the percentage of residents in the neighborhood of different age groups, the relative number of residents belonging to different marital status, one-person households, and immigrants based on their country of birth, classified as Western and non-Western.

The economic status of the neighborhoods was estimated by the average value of the houses/apartments, the percentage of inhabitants with the lowest registered personal income (the lowest 40% after all persons have been ranked according to their personal income), the highest personal income (the highest 20% in the ranking), and the total number of cars registered in the neighborhood. Road safety was estimated by the number of registered cumulative road accidents from 2009

to 2011 in a buffer of 200 m around the home.

Accessibility of destinations including institutions of health, cultural institutions, schools, train stations, supermarkets, restaurants etc. was also mapped. For each destination two measures were used: the average distance (km) calculated by road for all residents of the neighborhood, and the count of specific destinations within buffers, extending from 1 to 20 km around the home (CBS, 2013). Some of these destinations were grouped together to reduce the number of highly correlated variables. For instance, the average distance to the general practitioner's office or pharmacies and hospitals were grouped as participants "access to medical facilities". Average distance to kindergartens, elementary, middle, and high school facilities were grouped as "access to educational facilities". Museums, cinemas, attraction parks, concert halls, swimming pools, ice skating halls, saunas and tanning clinics were grouped as "access to recreational activities". The exposure to the food environment in each neighborhood was assessed by the number of food retailers in 1 km and 5 km buffers. Food retailers were categorized into "healthy" and "less healthy" food options, based on the healthfulness of the food retailers. Although there is not a clear definition on the healthiness of food retailers, often the supermarkets and local food shops are considered as healthy food options (Pinho et al., 2019). Home-cooked meals are generally healthier in terms of salt or fat consumption as compared to restaurants. Thus, the supermarkets and local food shops (e.g., greengrocers, bakeries, butchers etc.) were considered as "healthy food" exposure, and the restaurants (including fast food restaurants), fast food take-away places, cafés, pancake houses, bars and pubs were classified as exposure to "non-healthy food" choices.

2.4. Statistical analysis

2.4.1. Data pre-processing

The highest percentage of missing values among the retained variables was 10.7%, for the variable 'share of immigrants in neighborhood with other non-western origins. A multiple imputation by chained equations was applied to handle missing data in the remaining dataset. Variables were transformed by logarithmic, root square or inverse functions as appropriate to best approach a Gaussian distribution (Osborne & Ph, 2005) (Table A.4). Note that the nature of association was inversed for the multiplicative inverse transformed variables (multiplicative inverse = 1/variable) (Table A.4). For some variables, notably those with a substantial proportion of zero values, transformation was impossible, and we therefore used the predictive mean matching algorithm for imputation. The outcome (BMI) was included in the imputation model as a predictor, but missing BMI values were not imputed (White et al., 2011). Only a single imputation set was retained, because there is not a widely accepted way to combine the results in different imputation sets from machine learning approaches applied in this study. Besides, repeating the analysis several times on large datasets would significantly increase the already high computational time. The distributions of imputed and non-imputed data were compared for all the imputed exposures. A sensitivity analysis was conducted to compare the estimates of linear models obtained from imputed and complete-case analysis. All continuous exposures were standardized to the same scale by their standard deviations, so that all the coefficients would weigh equally, independently of the original scales of the variables. As part of descriptive statistics, we calculated the absolute average values of Spearman's correlation coefficients between the groups of exposures.

2.4.2. Assessing exposome-health associations

Seven supervised algorithms for regression were applied to identify robustness of findings across methods. In a recent review, Stafoggia et al. suggested to classify the statistical approaches addressing multiple correlated exposures in three groups: dimension reduction, variable selection and grouping of observations (Stafoggia et al., 2017). Based on this classification, we applied one dimension reduction method: sparse group Partial Least Squares (sgPLS), three variable selection approaches: Bayesian Model Averaging (BMA), Minimax Concave Penalty (MCP) and Generalized Additive Model (GAM) boost, and two methods from grouping of observations: Random Forest (RF) and Extreme Gradient Boosting (XGBoost). Several other criteria shaped our choice for the selected approaches. For instance, we considered factors such as the popularity, the feasibility (e.g. with regard to computational burden), the interpretability or the ability to incorporate nonlinear associations or interactions. Besides the selected approaches we also considered applying Bayesian Kernel Machine Regression. However, as initially suggested by the size of our data, this was infeasible. Additionally, the results of these models were compared to results from a classical multiple linear regression (MLR) model, as the most conventional model to estimate exposure-effect associations.

In context of these data-driven approaches, there is no single method which is universally the best. Hence, we applied multi-model inference to avoid increasing the sensitivity at the expense of a lower specificity. The second advantage of applying different methods is the increased chances of capturing the entire picture of exposure-obesity interrelations, including nonlinear associations and interactions and allowing for a pluralistic approach for data interpretation.

To compare the results across multiple methods, we used the variable importance scores (VI). Each model generated VI scores associated to all variables, depending on their contribution to the final model. For each approach all exposures were ranked (89 exposures including dummy variables), with lower values corresponding to the more important variables. Then, an overall ranking across all approaches was calculated based on the median values of ranks across the approaches. In case of ties, each of them was replaced by their mean. The confounders were ranked separately from the exposures (9 confounders), thus they were excluded from the overall ranking of exposures.

In contrast to linear models, there is not a straightforward interpretation of coefficients of nonlinear models. Besides, the VI scores do not provide information about the directions of associations. In order to improve the interpretability of nonlinear models and to learn how a given exposure was related to BMI (positively, negatively or nonlinear associations), we used Shapley values, which is a concept coming from coalitional game theory (Shapley, 1953). During recent years Shapley values gained popularity for studies based on explainable machine learning (Smith & Alvarez, 2021). Shapley values are the average contribution of a feature value to the prediction when different combinations of features are used, rather than the difference in prediction when we would remove the feature from the model (Molnar, 2020). For each approach, scatterplots were used to visualize the relationship between the response and the most influential predictors. Shapley plots were used also to calculate the exposure effects of the most important predictors in nonlinear and nonparametric models.

In order to provide robust results and assure model stability, the parameters were optimized for efficiency using a three-times repeated cross validation (Table A.5). Tuning parameters were calibrated and set for each model individually (Table 1). All models included all the factors and were adjusted for age, sex, country of birth of participants and their parents, civil state, education, employment status and smoking, because of their potentially confounding role. The education and employment status were used as estimators of individual socio-economic status. All analyses were performed with the R statistical software (version 4.0.2). The package "mice" was used for the multiple imputation. All the variables, including also the confounders were included in multiple exposure models simultaneously. The entire dataset was used in all models to assess the VI scores.

A short summary of applied methods is given in the next paragraphs.

a) PLS is a supervised dimension reduction technique that builds summary variables as linear combinations of the original set of variables (Agier et al., 2016). sgPLS is a version of PLS regression, which allows sparsity both of groups of exposures, as well as within each group; only relevant variables within a group are selected (Liquet

Table 1
Summary characteristics of algorithms, packages and parameters optimized throughout the calibration process.

Model	Package	Tuning parameters	Characteristics
sgPLS	sgPLS	ncomp = 1 keepX = 1 alpha = 0.01 upper lambda = 1e + 05	• Dimension reduction and variable selection simultaneously • Creates linear combinations of the original predictors • Allows sparse variable selection both within groups and in groups
MCP	ncvreg stabsel	lambda = 0.0144 gamma = 3 alpha = 1 penalty="MCP" family="gaussian" nlambda = 300 max.iter = 100000 convex.min = 94 penalty.factor = rep(1, ncol(X))	• Linear regression shrinkage method • Good interpretability
BMA	BAS	prior = JZS modelprior = beta.binomial()	• Bayesian variable selection model • Takes into account model uncertainty • Considers the linear associations and interactions • Good interpretability • Low computational cost
GAM boost	mboost caret	mstop = 316 prune = no bol and bbs	• Generalized additive model by likelihood-based boost • Useful method for detecting nonlinear associations
RF	ranger tuneRanger	sample.fraction = 0.77 min.node.size = 987 importance ="permutation" mtry = 41 splitrule = variance	• Captures complex interactions and nonlinear associations • No information on magnitude or direction of the association • Less intuitive for interpretation • Computationally complex
XG Boost	xgboost caret	nrounds = 1000 max_depth = 3 eta = 0.005 verbose = 1 gamma = 2 colsample_bytree = 0.5 min_child_weight = 0 subsample = 0.632 objective = "reg: linear" eval_metric = "rmse"	• Tree based, supervised technique • Captures the interaction and nonlinearity in the dependence structure • Iteratively combines the output of multiple decision trees in a stepwise manner to improve performance at each iteration to make a strong classifier • VI scores facilitate the interpretability

sgPLS = sparse group Partial Least Squares; MCP = Minimax Concave Penalty; GAM boost = Generalised Additive Model boost; BMA = Bayesian Model Averaging; RF = Random Forest; XGBoost = Extreme Gradient Boosting.

et al., 2016). For selected variables, the absolute values of weights in the linear combination were used to calculate the VI. No VI score was attributed to the non-selected variables. The R package "sgPLS" was used for this analysis (Liquet et al., 2016). In sgPLS function calibration parameter "ncomp" refers to the number of components to be included in model, "keepX" shows the number of variables kept in the model on each component and the parameter "alpha" is related to the within group sparsity.

b) Penalized regression approaches were developed to address the problems of multicollinearity in high-dimensional settings causing the MLR models to produce unstable parameter estimates. Penalized regression by Minimax Concave Penalty (MCP) is a high-dimensional, linear variable selection approach, similar to the well-known LASSO penalty, but with better performance quality (Zhang, 2010). The variable selection was based on a ten-fold cross-validation, which helped to determine the optimal degree of the penalization. The final set of selected variables was decided by the stability selection procedure. The VI score was calculated by ranking the largest lambda (penalty) for which a variable was still included in the model. The package "ncvreg" was used to run the MCP regression model (Breheny & Huang, 2011). One of the important parameters to calibrate for the MCP regression model is the "lambda", which represents the degree of penalization. The parameter "penalthy.factor" allows to apply differential penalization if some coefficients are thought to be more likely than others to be in the model, for example the confounders. In this study no differential penalization was applied.

c) GAM boost is a variable selection approach. It fits a generalized additive model by likelihood-based boost and is particularly useful in the case of nonlinearly associated predictors (Tutz & Binder, 2006). Variable selection was done by the stability selection procedure. In this study the VI score obtained from the cross-validation model was used to rank the exposures. The "mboost" package was used for running the GAM boost model (Bühlmann & Hothorn, 2007; Hofner et al., 2014). The parameter "mstop" represents the optimal number of boosting iterations which was calibrated by the cross-validation.

d) BMA is a linear, variable selection approach, that considers model uncertainty by evaluating all possible exposure combinations and assigns weights for each model (Sun et al., 2013). The final exposure effect is calculated as a weighted average of the exposure effects from each model (Stafoggia et al., 2017). The variable selection was assessed after the control for false discovery rate. The posterior inclusion probabilities were used as estimator of VI score. The R package called "BAS" was used to run the BMA model (Clyde M, 2020). For BMA model it is required to specify the prior distribution for regression coefficients (function "prior") and to assign a priori probabilities for each model (function "modelprior").

e) Random forest algorithm is a recursive partitioning method, based on conduction of many decision trees and the aggregation of the predictions from these trees. For each iteration randomly selected predictors and a random subset of data are chosen (Ishwaran & Lu, 2019). It can capture complex interactions and nonlinear associations among the exposures (Stafoggia et al., 2017). We applied a model-based optimization, evaluating the out-of-bag predictions, to tune random forests. The permutation importance approach was used to assess the relative VI score. For non-selection methods, such as RF, we used scatterplots of VI score to decide the selected variables. The variables with highest VI, which were standing out from the others were considered selected. Packages "tuneRanger" and "ranger" were used to calibrate and to run the RF model (Probst et al., 2019; Wright & Ziegler, 2017). There are several important parameters to calibrate for the RF model, e.g. "sample.fraction" which shows the number of observations to sample, "min.node.size" refers to the minimal node size and the "mtry", shows the number of variables to possibly split at in each node.

f) Extreme Gradient Boosting (XGBoost) is another tree based technique, that captures the interactions and nonlinearity in the dependence structure (Z. Y. Chen et al., 2019). In fact, the main difference between the XGBoost and random forest is that, XGBoost iteratively combines the output of multiple decision trees in a stepwise manner to improve performance at each iteration to make a strong classifier, while in RF the decision trees are independent from each other, and the results are combined at the end of the process (T. Chen & Guestrin, 2016). A repeated cross-validation was used to assess the optimal tuning parameters for the model. The VI score was obtained using the fractional contribution of each feature to the model based on the total gain of this feature's splits. Like RF, the variable selection was based on scatterplots of VI. The R package called "xgboost" was used for the XGBoost model (T. Chen & Guestrin, 2016). Following parameters were calibrated for the final XGBoost model: "eta", which

is the learning rate; "gamma" refers to the minimum loss reduction required to make a further partition on a leaf node of the tree; "max_dept" is the maximum depth of a tree; "min_child_weight" is the minimum sum of instance weight (hessian) needed in a child; "subsample" corresponds to the ratio of the training instances at each iteration; "colsample_bytree" is the subsample ratio of columns when constructing each tree ("XGBoost Parameters — xgboost 1.5.0-dev documentation").

g) Multiple linear regression was applied as a simple model, simultaneously including all explanatory variables and BMI being the explained variable. The VI score from MLR model was based on the p-values for each exposure.

Additionally, a sensitivity analysis was performed to compare the effect estimates from single exposure models with the multiple exposure models. Single exposure models included all the confounders and a single exposure at a time. Associations were deemed significant or not after correction for multiple testing by Bonferroni method.

Most of the statistical approaches used in this study do not provide p-values. Consequently, it was impossible to control for the familywise error rate and the false discovery rate for all the models at a similar level. We considered, however, to have reduced the false discovery rate by applying stability selection procedure to two variable selection approaches: Minimax Concave Penalty regression and Generalized Additive Model boost. Stability selection is a subsampling based method used in high-dimensional variable selection setting, which provides a finite sample control for false discovery rate (Meinshausen & Bühlmann, 2010). The false discovery rate was controlled at the level of 0.2 in the Bayesian Model Averaging approach. A correction for multiple testing by the Bonferroni method was implemented in the Multiple Linear Regression model.

3. Results

3.1. Study population

The analytical sample consisted of 14,781 participants, with an average age of 50.7 ± 9.4 years and the majority were women (55.8%) (Table 2). More than one third had a higher education (38.2%) and most were employed (71.8%). About two-third was overweight (65.1% including obesity). Less than a fourth of participants (15.7%) were active smokers (Table 2).

Table 2
Baseline characteristics of the participants.

Characteristics	Complete cases	Mean ± SD or n(%)
BMI (kg/m2)	14,781 (99.7%)	26.1 ± 4.4
Age	14,829 (100%)	50.7 ± 9.4
Sex	14,829 (100%)	
Female		8268 (55.8%)
Male		6561 (44.2%)
Country of origin	14,829 (100%)	
Netherlands		14,127 (95.3%)
Other		702 (4.7%)
Country of birth of mother	14,793 (99.8%)	
Netherlands		13,750 (92.9%)
Other		1043 (7.1%)
Country of birth of father	14,787 (99.7%)	
Netherlands		13,776 (93.1%)
Other		1011 (6.8%)
Civil state	14,805 (99.8%)	
Having a partner/being married		11,902 (80.4%)
Not having a partner		2903 (19.6%)
Education	14,820 (99.9%)	
Low/Medium		9164 (61.8%)
High		5656 (38.2%)
Employment status	14,829 (100%)	
Employed		10,641 (71.8%)
Unemployed		4167 (28.2%)
Smoking	14,806 (99.8%)	
Yes		2322 (15.7%)
No		12,484 (84.2%)

Table A.6 contains a detailed description of all the factors of the urban exposome included in this study. The range of absolute average inter- or intra- group correlations between the exposures extended from low to moderate ($|r_{sp}|$=0.1–0.6) (Fig. 1). The intragroup correlations were relatively higher for air pollutants ($|r_{sp}|$=0.4), green space ($|r_{sp}|$=0.5) and meteorology ($|r_{sp}|$=0.4) (Fig. 1).

The directions of associations are mentioned for the selected exposures by each approach. "Positive" stands for a positive association and "Negative" for a negative association. Empty cells indicate that the given variable was not selected by the given approach. All models were adjusted for the following covariates: age, sex, country of birth, country of birth of mother, country of birth of father, civil state, education, employment status and smoking.

3.2. Urban exposome and BMI

Table 3 shows an overview of the variables that were selected by at least one of the approaches. The BMA was the only method to select up to six variables. The sgPLS method, on the other hand, did not select any variables associated with BMI. It selected only the group of covariates/confounders and all the variables within that group (since the variables were grouped, covariates were introduced as one group) (Fig. A.1).

Exposures reflecting the socioeconomic position of the neighborhoods were observed to have the most consistent associations with BMI. The average price of the houses was selected by six out of seven methods (besides sgPLS), and the highest rank was attributed to it by five methods. All models detected a negative association between the average price of houses and BMI (median rank = 1[min–max = 1–3]) (Fig. 2, Fig. 4, Table 3, Table A.7).

The second most important exposure was the oxidative potential of $PM_{2.5}$ (2.5 [2–5]), although it was only selected by RF, BMA and MLR (Fig. 2, Table 3). The association between oxidative potential and BMI was a positive association for all models (Fig. 4, Table 3).

Some consistency was found for particular area level characteristics, such as the number of healthy food outlets in a 5 km buffer (6 [1–20]) and the percentage of one-person households in the neighborhood,

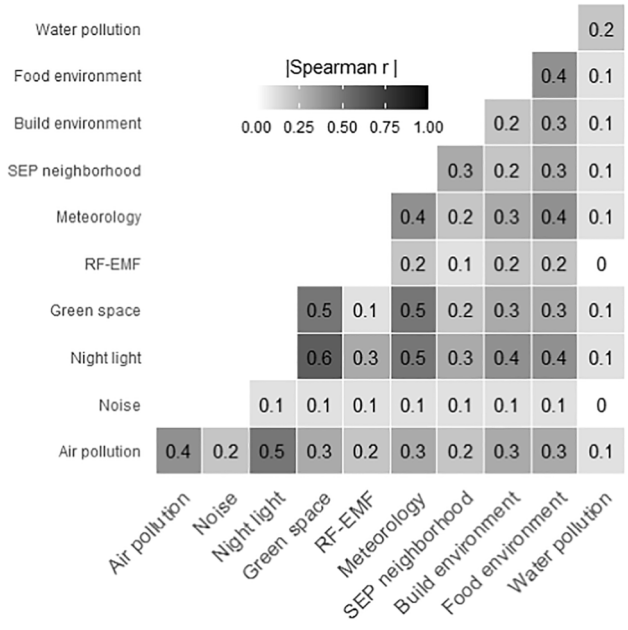

Fig. 1. Absolute average intra- and intergroup correlations between constructs of exposures. Higher values indicate higher absolute value of correlation coefficient. RF-EMF = Radiofrequency electromagnetic field.

Table 3
Overview of the variables that were selected by at least one of the approaches.

Exposures	MLR	sgPLS[1]	MCP	BMA	GAM Boost	XG Boost	RF	Median rank [Range] [2]
Average house price	Negative		Negative	Negative	Negative	Negative	Negative	1[1–3]
OPESR	Positive			Positive			Positive	2.5[2–5]
Healthy food (5 km)	Negative			Negative				6[1–20]
One-person households (%)				Negative				10[4–63]
Non-Western immigrants				Positive				22[7–62.5]
Amenity shops in 20 km				Positive				33[5–47]

OPESR = Oxidative potential of $PM_{2.5}$ measured by electron spin resonance; MLR = Multiple Linear Regression; sgPLS = sparse group Partial Least Squares; MCP = Minimax Concave Penalty; BMA = Bayesian Model Averaging; GAM boost = Generalized Additive Model boost; XGBoost = Extreme Gradient Boosting; RF = Random Forest.
[1]The cells are empty for sgPLS because no exposures were selected by this method.
[2]Reported median ranks of exposures are calculated based on VI scores of MCP, BMA, GAM Boost, XGBoost, RF and MLR. Lower ranks indicate higher VI score and the opposite for higher ranks.

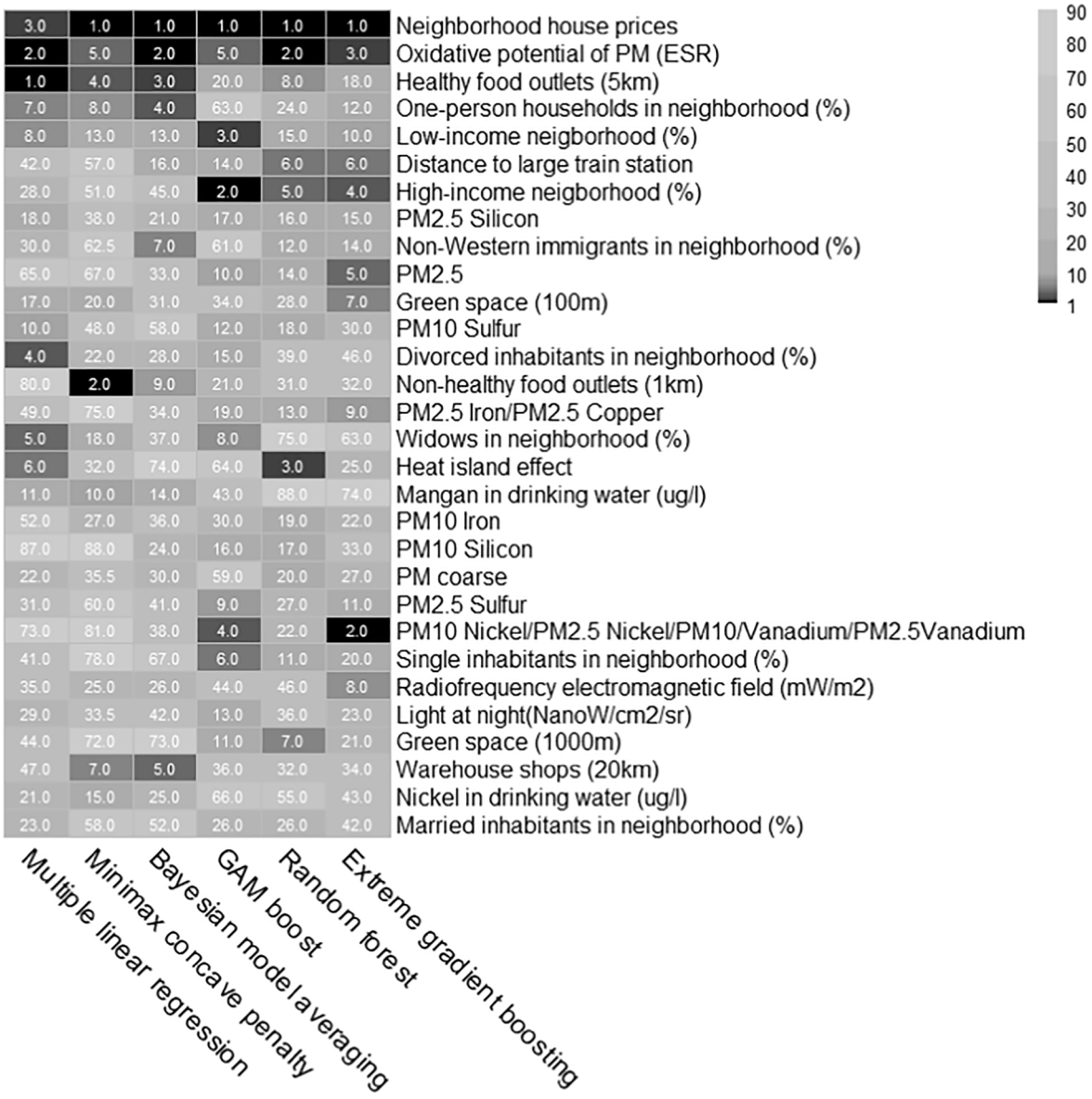

Fig. 2. The ranking of top 30 important exposures across different approaches. The rankings are sorted by the overall rank across approaches. The decimal values represent tie ranks, where the ranks have been replaced by their mean. The results of the sparse group Partial Least Squares are not presented, because no VI was attributed to the non-selected variables. GAM boost = Generalized Additive Model boost; PM = particulate matter; OP (ESR) = particulate matter measured by electron spin resonance.

which were associated with lower BMI (10 [4–63]) (Fig. 2). Participants living in neighborhoods where the percentage of inhabitants with the lowest registered personal income was higher, tended to have increased levels of BMI (11.5 [3–15]). The proximity to a large train station (15 [6–57]), high-income neighborhoods (16.5 [2–51]), the share of non-Western immigrants in a neighborhood (22 [7–62.5]), fine particles $PM_{2.5}$ (23.5 [5–67]), silicon in $PM_{2.5}$ (17.5 [15–38]) and amount of green in a 100 m buffer (24 [7–37]) were also related to BMI, but with less consistency across approaches as evidenced by Fig. 2.

For most exposures we observed a certain variability across approaches. The rankings between various approaches correlated with the overall median rank in a range of $r_{spearman} = 0.5–0.7$ (Fig. 3). The output of MLR model was assimilating to the overall rank $r_{spearman} = 0.6$, slightly better than other linear models $r_{spearman} = 0.5$. The sgPLS was the only method that correlated poorly with other approaches. Interestingly, a hierarchical clustering on the VI rankings from each approach showed that the results of linear and non-linear methods clustered separately (Fig. A.3). For instance, we observed that healthy food outlets in 5 km buffer and one-person households in the neighborhood received a much higher rank by linear models, as compared to nonlinear models (1–4 vs 8–20 and 4–8 vs 12–63 accordingly) (Fig. 2). In contrast, nonlinear models detected a strong, positive association with the share of inhabitants with highest registered personal income (ranked 2, 5 and 4 corresponding to GAMboost, RF and XGBoost), whereas the association received much less importance in linear models (ranked 45, 51, 28 corresponding to BMA, MCP regression and MLR).

In terms of effect sizes, generally, the estimated associations were in line with previously reported findings. The exposure association was strongest for the average value of houses and considerably less important for the rest of the exposures. Shapley plots (Fig. 4) showed that going from maximum to minimum values of the average house prices resulted in a change of BMI of 0.5 kg/m^2 for the RF model, whereas for oxidative potential it was 0.2. In XGboost model these values were 0.4 and 0.09 accordingly. Similar results were observed by linear models: the coefficients of BMA were 0.11/0.04, for MLR 0.07/0.05 and for the MCP 0.11/0.05 (Table A.8). Shapley plots of RF and XGBoost representing other influential variables can be found in Fig. 4 and Fig. A.2.

The sensitivity analysis showed that the effect estimates were different between single-exposure models and the multiple linear regression model. Overall, effect estimates from single-exposure models were lower compared to multiple linear regression. After the correction for multiple testing, ten significant associations were found by single exposure models: oxidative potential of $PM_{2.5}$ measured by ESR (p-value < 0.01) and DTT ($p < 0.01$), Nickel in PM_{10} ($p < 0.01$), Silicon in $PM_{2.5}$ ($p < 0.02$), share of divorced inhabitants (%) ($p < 0.01$), share of non-western immigrants ($p < 0.01$), average value of houses ($p < 0.01$), high- and low-income neighborhoods ($p < 0.01$ for both) and the amount of Mecoprop (herbicide) in drinking water ($p < 0.01$). In contrast, only three factors were found significant by multiple linear regression model average value of houses ($p < 0.01$), oxidative potential of $PM_{2.5}$ (ESR) ($p < 0.01$) and number of healthy food outlets in 5 km buffer ($p < 0.01$). The R^2 was on average 3.6% for single exposure models and 4.7% for multiple linear regression model. The results of this analysis are summarized in Table A.9. Furthermore, another sensitivity analysis on complete cases showed comparable results with the imputed dataset.

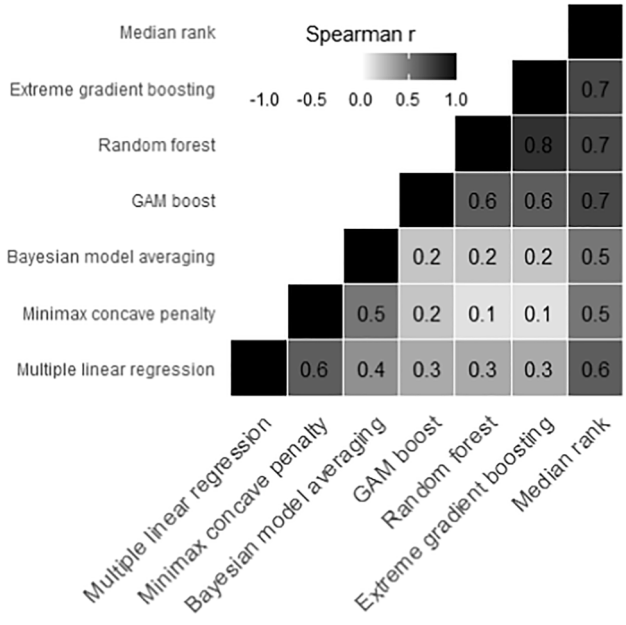

Fig. 3. Spearman correlation coefficients between variable importance scores attributed by each method and the median rank across these scores. GAM boost = Generalized Additive Model boost.

4. Discussion

In this study we assessed how the urban exposome relates to adult BMI. We addressed many environmental exposures and used a comprehensive pluralistic statistical strategy to account for linear as well as nonlinear associations and complex interactions among the exposures. The results from multi-model inference were consistent only for the strongest associations. Residents of neighborhoods with higher values of average house prices had a lower BMI and oxidative potential of $PM_{2.5}$ was related to increased BMI. Living in neighborhoods with higher share of one-person households was associated with lower levels of BMI. Furthermore, a clustering of results was observed based on model's capacity to deal with linear or non-linear associations. These results offer suggestions for further studies to explore the interactions between social structures, specific urban characteristics, and their effect on health-related behaviors.

Studying multiple environmental factors as one system offers advantages. Mainly, multiple exposure studies allow to disentangle the effects of individual factors, considering the complex interactions between them. Generalizing the results of this study, many of the selected environmental factors were reflecting urbanisation. As we did not find any associations between BMI and urbanicity level (measured by residential density), this suggests that urban obesogenic environments are not simply driven by population densities but rather by specific neighborhood characteristics associated with population density.

We confirmed that the neighborhood SEP is an important health-related component of the neighborhood environment. We found clear links between a variety of indicators of neighborhood SEP and BMI irrespective to individual-level SEP, which is in line with the previous literature (Kim et al., 2019). As mentioned in a recently published meta-analysis by Mohammed et al., low neighbourhood SEP might promote unhealthy dietary practice and sedentary lifestyle, as health-enhancing facilities are often limited, whereas energy-dense food items and alcohol are more readily available (Mohammed et al., 2019). Besides, in low SEP neighbourhoods residents are exposed to more psychosocial stressors and higher risk of depression, which is likely to influence one's unhealthy lifestyle choices, resulting in higher risk of obesity (Gary-Webb et al., 2011). Moreover, in low SEP neighbourhoods, streets walkability and safety might negatively influence the mobility and physical activity of residents (Popkin et al., 2005).

Some might argue to consider the area-level SEP variables as confounders rather than exposures. If area-level SEP would indeed have been considered as a confounder rather than an exposure in this study, reported results would not be distorted regarding to other exposures, because all the exposures and confounders were introduced simultaneously in all the models, including the area-level SEP proxies. This would mean that three variables representing proxies for neighborhood SEP (average house prices, low- and high-income neighborhoods) would

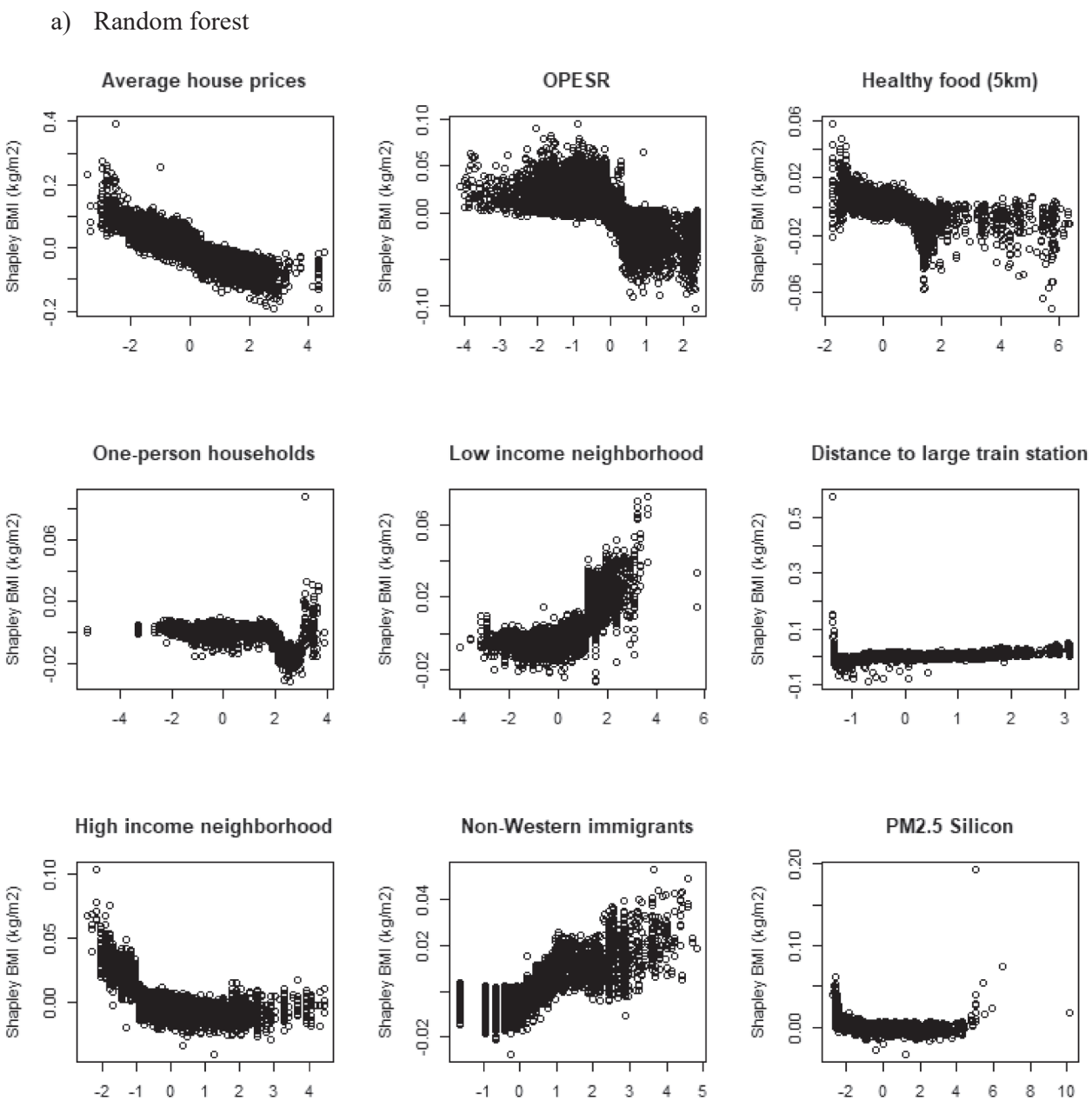

Fig. 4. Shapley plot illustrations of the most influential exposures in Random Forest (a) and Extreme Gradient Boosting (b) models. Shapley BMI (kg/m^2) represents the difference between a prediction and the average prediction of BMI. OPESR = Oxidative potential of particulate matter ($PM_{2.5}$) measured by electron spin resonance. For the interpretation of OPESR, note that the direction of association should be inversed, as it was multiplicative inverse transformed (multiplicative inverse = 1/variable) to approach normality.

be taken out of the overall ranking of exposures. This would cause a slight shift in the ranking of other variables, such that the number of educational facilities in 10 km, road traffic noise and the share of young children (0–14 years old) in neighborhood will then be included among the 30 most important predictors of BMI.

Social networks are important upstream determinants of obesity (Lakerveld & Mackenbach, 2017). In our study, linear models observed a link between the share of one-person households in neighborhood and BMI which highlights the importance of social networks, as a neighborhood of singles is probably more active, thriving neighborhood with many more social connections (Sarkisian & Gerstel, 2016). It is interesting to note that this outcome is contrary to that of individual level household composition found by Barnes et al. who observed that living with others is associated with lower body weight (Barnes et al., 2013).

The present study raises the possibility that oxidative potential of $PM_{2.5}$ is related to obesity. Oxidative potential measures the inherent capacity of PM to oxidise target molecules (Yang et al., 2016). Previously published studies observed higher concentrations of oxidative potential (assessed by ESR) in urban as compared to rural areas (Janssen et al., 2014) and near major roads compared to urban background. To the best of our knowledge, this is the first study to investigate associations between exposure to oxidative potential (DTT and ESR) of $PM_{2.5}$ and adult obesity. As documented for other air pollutants, the oxidative potential of $PM_{2.5}$, might also contribute to cardiometabolic disorders by inducing oxidative stress and inflammatory processes (Gangwar et al., 2020; Viehmann et al., 2015). In an animal study, Xu et al. showed that

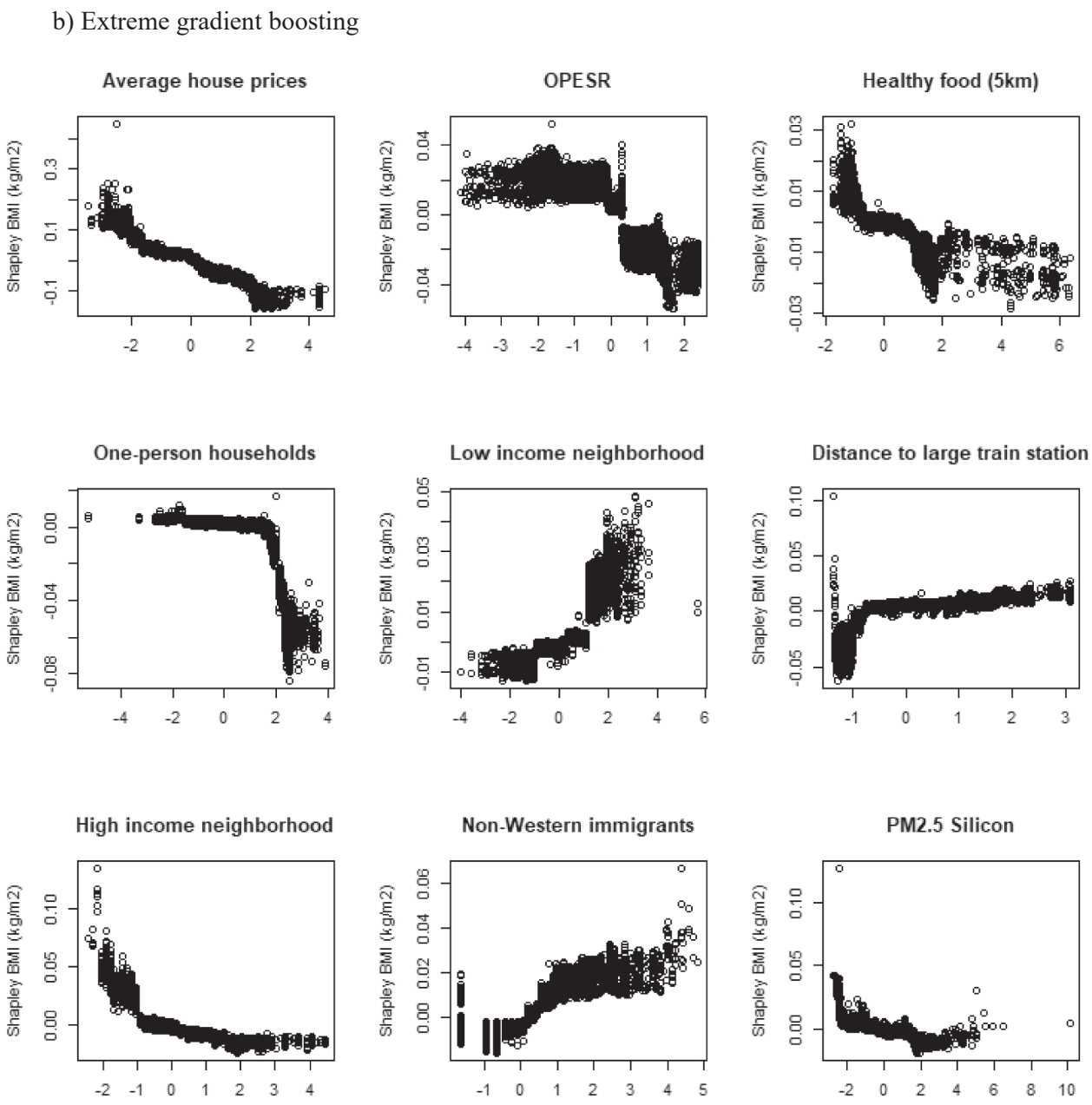

Fig. 4. (*continued*).

$PM_{2.5}$ exposure triggers oxidative stress in brown adipose tissue in mice, and results in key alterations in mitochondrial gene expression and mitochondrial alterations that are pronounced in brown adipose tissue. The authors assume that exposure to $PM_{2.5}$ may induce imbalance between white and brown adipose tissue functionality and thereby predispose to metabolic dysfunction (Xu et al., 2011). However, we must note that oxidative potential is reported to be highly sensitive to the components of PM sourced in atmosphere by road traffic (Yang et al., 2015). These data therefore must be interpreted with caution, as the observed association could be explained by unmeasured exposures related to urbanicity or road traffic. However, no associations were observed with other traffic related factors like UFP, NO or $PM_{absorbance}$ arguing perhaps for an actual role of the oxidative potential of particulate matter in obesity.

Previous studies have observed inconsistent results on whether the food environment is associated with weight status (Lam et al., 2021). Our findings add to the existing literature in that healthy food outlets were related to lower BMI. However, we also observed inconsistency of direction of association between non-healthy food outlets and BMI. A possible explanation to this finding is that in the Netherlands healthy and unhealthy food retailers often co-locate. Furthermore, in this study healthy food outlets were defined by supermarkets and convenience stores, whereas these types of retail offer both healthy as well as unhealthy food.

As expected, sensitivity analysis showed that the results of single exposure models differed from the results of multiple exposure models, mainly by an increased proportion of identified significant associations and underestimated effect estimates. One of the reasons for this is the insufficient control for confounding factors from the environment.

We combined seven different statistical approaches to find the most consistent exposure-BMI associations. For the strongest associations, the results were consistent across all approaches besides the sparse group PLS, which did not select any exposures. As the exposure effects were modest, this could be the reason why sgPLS did not pick up any

exposures. The performance of the multiple linear regression model was comparable with the other linear models. Our analysis was still in a rather low-dimensional setting (<100 variables) with relatively high number of observations and moderate correlations between variables. In this scenario the MLR model showed not to be compromised by multicollinearity problems. In terms of variable selection, MLR correlated with the overall median rank slightly better compared to other linear models, but less well than nonlinear methods (Fig. 3). Data-driven methods such as RF or XGboost proved useful for the incorporation of nonlinear and non-additive associations, especially combined with Shapley plot visualizations, which drastically improved the interpretability of these models by plotting nonlinear associations. Despite these remarkable advances in the interpretation of these powerful methods, we did not investigate interactions in the data, because with current variable importance measures the detection of interactions is not straightforward, especially for modest effect sizes, as in this study (Wright et al., 2016). RF approach provided a better balance between the capacity to address multiple exposure data, the computational burden and the robustness of model. However, as we do not know the true exposure-health associations in the data it is difficult to compare these statistical methods with each other and give strong recommendations on their future use. We therefore advocate a pluralistic approach in which multiple methods with different model specifications are used to derive inference across different models.

One of the challenges of a multi-inference approach is collective interpretation of the results. We handled this issue by using VI scores from various models to get a ranking of exposures, and evaluate the consistency of findings across approaches. A multi-inference approach also helps in achieving a balance between false positives and false negatives, i.e., multiple approaches reduce false negatives and the combination of stability selection/cross validation and collective evaluation of the results reduces false positives. It should be noted that despite the robustness of the pluralistic approach, it imposes a computational burden and does not offer a solution for the exact effect estimates. This question is however, a topic for further discussion indicating the need for more research in this field and advocates for the development of adapted statistical methods.

The strengths of this study include, first, the large number of participants (n > 14,000), sampled from the general population (age 30–65) in Netherlands. Second, we have analyzed a large number of individual level and contextual urban exposures and important confounding factors. Third, we have accounted for complex interactions and non-linear associations by applying six different statistical approaches, which are scarcely used in epidemiological studies.

We also acknowledge several limitations of this study. First, our study has a cross-sectional design, which makes it impossible to establish the temporal link between the events and limiting the causal interpretation of associations. This is particularly important for the vulnerability to residential self-selection bias, resulting from health-related attitudes, neighborhood preferences, or other unmeasured characteristics related to both neighborhood choice and health-related outcomes (James et al., 2015). For instance, despite that many of the identified factors reflect the degree of urbanization, it could also be the case that participants with lower SEP are generally living in these urban areas. Second, our outcome was self-reported BMI. Literature suggests that adults tend to under-report their own weight and that the gap between self-reported weight and actual weight increases with obesity (Maukonen et al., 2018). Third, the issue of measurement error could also apply to the environmental exposures, which consisted of a mix of modelled and measured values with heterogeneous error structures. Therefore, complex combinations of both types of measurement errors: classical and Berkson's, might be present in the given dataset. In general terms, this means that the sensitivity of models is lower for highly variable factors (if we repeat the exposure assessment several times, those with the lowest intra-class coefficient of correlation) compared to factors that are more stable over time (Agier et al., 2020).

Despite its limitations, a pluralistic multi-model inference seems yet to be the best approach when an established method of reference is lacking. Although such an approach is computationally heavy, with current computational resources it is implementable. This study is, to our knowledge, one of the first to assess the contributions of a large set of urban environmental factors in adult overweight and obesity. It indicates that urban infrastructure and socioeconomic and demographic characteristics at the neighborhood level could be drivers for obesity. As these neighborhood characteristics are modifiable, it strengthens the evidence base for targeted built environmental policies and intervention approaches. Further research should be undertaken to confirm and further investigate the role of oxidative potential of particulate matter in relation to obesity.

CRediT authorship contribution statement

Haykanush Ohanyan: Investigation, Formal analysis, Software, Visualization, Writing – original draft. **Lützen Portengen:** Validation, Software, Supervision, Writing – review & editing. **Anke Huss:** Data curation, Writing – review & editing. **Eugenio Traini:** Data curation, Writing – review & editing. **Joline W.J. Beulens:** Conceptualization, Supervision, Writing – review & editing. **Gerard Hoek:** Conceptualization, Supervision, Writing – review & editing. **Jeroen Lakerveld:** Conceptualization, Supervision, Writing – review & editing. **Roel Vermeulen:** Resources, Conceptualization, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2021.107015.

References

Agier, L., Portengen, L., Hyam, M.C., Basagaña, X., Allemand, L.G., Siroux, V., Robinson, O., Vlaanderen, J., González, J.R., Nieuwenhuijsen, M.J., Vineis, P., Vrijheid, M., Slama, R., Vermeulen, R., 2016. A systematic comparison of linear regression–Based statistical methods to assess exposome-health associations. Environ. Health Perspect. 124 (12), 1848–1856. https://doi.org/10.1289/EHP172.

Agier, L., Slama, R., Basagaña, X., 2020. Relying on repeated biospecimens to reduce the effects of classical-type exposure measurement error in studies linking the exposome to health. Environ. Res., 186(July 2019), 109492. https://doi.org/10.1016/j.envres.2020.109492.

An, R., Ji, M., Yan, H., Guan, C., 2018. Impact of ambient air pollution on obesity: A systematic review. Int. J. Obesity, 42(6), 1112–1126. https://doi.org/10.1038/s41366-018-0089-y.

Andrianou, X.D., Makris, K.C., 2018. The framework of urban exposome: Application of the exposome concept in urban health studies. Sci. Total Environ. 636, 963–967. https://doi.org/10.1016/j.scitotenv.2018.04.329.

Baliatsas, C., van Kamp, I., Swart, W., Hooiveld, M., Yzermans, J., 2016. Noise sensitivity: Symptoms, health status, illness behavior and co-occurring environmental sensitivities. Environ. Res. 150, 8–13. https://doi.org/10.1016/j.envres.2016.05.029.

Barnes, M.G., Smith, T.G., Yoder, J.K., 2013. Effects of household composition and income security on body weight in working-age men. Obesity 21 (9), E483–E489. https://doi.org/10.1002/oby.20302.

Barrera-Gómez, J., Agier, L., Portengen, L., Chadeau-Hyam, M., Giorgis-Allemand, L., Siroux, V., Robinson, O., Vlaanderen, J., González, J.R., Nieuwenhuijsen, M., Vineis, P., Vrijheid, M., Vermeulen, R., Slama, R., Basagaña, X., 2017. A systematic comparison of statistical methods to detect interactions in exposome-health associations. Environmental Health: A Global Access Science Source 16 (1). https://doi.org/10.1186/s12940-017-0277-6.

Beam, A.L., Kohane, I.S., 2018. Big data and machine learning in health care. JAMA – J. Am. Med. Assoc. 319 (13), 1317–1318. https://doi.org/10.1001/jama.2017.18391.

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M.-Y., Künzli, N., Schikowski, T., Marcon, A., Eriksen, K.T., Raaschou-Nielsen, O., Stephanou, E., Patelarou, E., Lanki, T., Yli-Tuomi, T., Declercq, C., Falq, G., Stempfelet, M., Birk, M., Cyrys, J., von Klot, S., Nádor, G., Varró, M.J., Dėdelė, A.,

Gražulevičienė, R., Mölter, A., Lindley, S., Madsen, C., Cesaroni, G., Ranzi, A., Badaloni, C., Hoffmann, B., Nonnemacher, M., Krämer, U., Kuhlbusch, T., Cirach, M., de Nazelle, A., Nieuwenhuijsen, M., Bellander, T., Korek, M., Olsson, D., Strömgren, M., Dons, E., Jerrett, M., Fischer, P., Wang, M., Brunekreef, B., de Hoogh, K., 2013. 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.

Bezerra, I.N., Curioni, C., Sichieri, R., 2012. Association between eating out of home and body weight. Nutr. Rev. 70 (2), 65–79. https://doi.org/10.1111/j.1753-4887.2011.00459.x.

Breheny, P., Huang, J., 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann. Appl. Stat. 5 (1), 232–253. https://doi.org/10.1214/10-AOAS388.

Bühlmann, P., Hothorn, T., 2007. Boosting algorithms: Regularization, prediction and model fitting. Stat. Sci. 22 (4), 477–505. https://doi.org/10.1214/07-STS242.

Cai, Y., Zijlema, W.L., Sørgjerd, E.P., Doiron, D., de Hoogh, K., Hodgson, S., Wolffenbuttel, B., Gulliver, J., Hansell, A.L., Nieuwenhuijsen, M., Rahimi, K., Kvaløy, K., 2020. Impact of road traffic noise on obesity measures: Observational study of three European cohorts. Environ. Res. 191, 110013. https://doi.org/10.1016/j.envres.2020.110013.

Casson, R.J., Farmer, L.DM., 2014. Understanding and checking the assumptions of linear regression: A primer for medical researchers. Clin. Exp. Ophthalmology 42 (6), 590–596. https://doi.org/10.1111/ceo.2014.42.issue-610.1111/ceo.12358.

CBS, 2013. *Wijk en buurtkaart 2012*. Buurt 2012. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2011.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-Augu, 785–794. https://doi.org/10.1145/2939672.2939785.

Chen, Z.Y., Zhang, T.H., Zhang, R., Zhu, Z.M., Yang, J., Chen, P.Y., Ou, C.Q., Guo, Y., 2019. Extreme gradient boosting model to estimate PM2.5 concentrations with missing-filled satellite data in China. Atmos. Environ., 202(December 2018), 180–189. https://doi.org/10.1016/j.atmosenv.2019.01.027.

Chooi, Y.C., Ding, C., Magkos, F., 2019. The epidemiology of obesity. Metab. Clin. Exp. 92, 6–10. https://doi.org/10.1016/j.metabol.2018.09.005.

de Hoogh, K., Wang, M., Adam, M., Badaloni, C., Beelen, R., Birk, M., Cesaroni, G., Cirach, M., Declercq, C., Dėdelė, A., Dons, E., de Nazelle, A., Eeftens, M., Eriksen, K., Eriksson, C., Fischer, P., Gražulevičienė, R., Gryparis, A., Hoffmann, B., Jerrett, M., Katsouyanni, K., Iakovides, M., Lanki, T., Lindley, S., Madsen, C., Mölter, A., Mosler, G., Nádor, G., Nieuwenhuijsen, M., Pershagen, G., Peters, A., Phuleria, H., Probst-Hensch, N., Raaschou-Nielsen, O., Quass, U., Ranzi, A., Stephanou, E., Sugiri, D., Schwarze, P., Tsai, M.-Y., Yli-Tuomi, T., Varró, M.J., Vienneau, D., Weinmayr, G., Brunekreef, B., Hoek, G., 2013. Development of land use regression models for particle composition in twenty study areas in Europe. Environ. Sci. Technol. 47 (11), 5778–5786. https://doi.org/10.1021/es400156t.

De la Fuente, F., Saldías, M.A., Cubillos, C., Mery, G., Carvajal, D., Bowen, M., Bertoglia, M.P., 2020. Green space exposure association with type 2 diabetes mellitus, physical activity, and obesity: a systematic review. Int. J. Environ. Res. Public Health 18 (1), 97. https://doi.org/10.3390/ijerph18010097.

Eeftens, M., Beelen, R., Hoogh, K. De, Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., De, A., Dons, E., Nazelle, A. De, Dimakopoulou, K., Eriksen, K., Fischer, P., Galassi, C., Graz, R., Heinrich, J., Ho, B., Jerrett, M., Keidel, D., … Hoek, G., 2012. Development of Land Use Regression Models for PM 2 . 5 , PM 2 . 5 Absorbance , PM 10 and PM coarse in 20 European Study Areas ; Results of the ESCAPE Project. https://doi.org/10.1021/es301948k.

Elvidge, C.D., Baugh, K., Zhizhin, M., Hsu, F.C., Ghosh, T., 2017. VIIRS night-time lights. Int. J. Remote Sens. 38 (21), 5860–5879. https://doi.org/10.1080/01431161.2017.1342050.

European Environment Agency, 2018. Environmental Noise. https://www.eea.europa.eu/airs/2018/environment-and-health/environmental-noise.

Gangwar, R.S., Bevan, G.H., Palanivel, R., Das, L., Rajagopalan, S., 2020. Oxidative stress pathways of air pollution mediated toxicity: Recent insights. In Redox Biology (Vol. 34, p. 101545). Elsevier B.V. https://doi.org/10.1016/j.redox.2020.101545.

Gary-Webb, T.L., Baptiste-Roberts, K., Pham, L., Wesche-Thobaben, J., Patricio, J., Pi-Sunyer, F.X., Brown, A.F., Jones-Corneille, LaShanda, Brancati, F.L., 2011. Neighborhood Socioeconomic Status, Depression, and Health Status in the Look AHEAD (Action for Health in Diabetes) Study. BMC Public Health 11 (1). https://doi.org/10.1186/1471-2458-11-349.

Gildner, T.E., Levy, S.B., 2021. Intersecting vulnerabilities in human biology: Synergistic interactions between climate change and increasing obesity rates. Am. J. Hum. Biol. 33 (2), 1–14. https://doi.org/10.1002/ajhb.23460.

Hofner, B., Mayr, A., Robinzonov, N., Schmid, M., 2014. Model-based boosting in R: A hands-on tutorial using the R package mboost. Comput. Statistics 29 (1-2), 3–35. https://doi.org/10.1007/s00180-012-0382-5.

Ishwaran, H., Lu, M., 2019. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. Stat. Med. 38 (4), 558–582. https://doi.org/10.1002/sim.7803.

James, P., Hart, J., Arcaya, M., Feskanich, D., Laden, F., Subramanian, S.V., 2015. Neighborhood self-selection: the role of pre-move health factors on the built and socioeconomic environment. Int. J. Environ. Res. Public Health 12 (10), 12489–12504. https://doi.org/10.3390/ijerph121012489.

Janssen, N.A.H., Yang, A., Strak, M., Steenhof, M., Hellack, B., Gerlofs-Nijland, M.E., Kuhlbusch, T., Kelly, F., Harrison, R., Brunekreef, B., Hoek, G., Cassee, F., 2014. Oxidative potential of particulate matter collected at sites with different source characteristics. Sci. Total Environ. 472, 572–581. https://doi.org/10.1016/j.scitotenv.2013.11.099.

Kim, Y., Cubbin, C., Oh, S., 2019. A systematic review of neighbourhood economic context on child obesity and obesity-related behaviours. Obes. Rev. 20 (3), 420–431. https://doi.org/10.1111/obr.v20.310.1111/obr.12792.

Lakerveld, J., Mackenbach, J., 2017. The upstream determinants of adult obesity. Obesity Facts 10 (3), 216–222. https://doi.org/10.1159/000471489.

Lam, T.M., Vaartjes, I., Grobbee, D.E., Karssenberg, D., Lakerveld, J., 2021. Associations between the built environment and obesity: an umbrella review. Int. J. Health Geographics 20 (1), 1–24. https://doi.org/10.1186/s12942-021-00260-6.

Lenters, V., Vermeulen, R., Portengen, L., 2018. Performance of variable selection methods for assessing the health effects of correlated exposures in case-control studies. Occup. Environ. Med. 75 (7), 522–529. https://doi.org/10.1136/oemed-2016-10423110.1136/oemed-2016-104231.supp1.

Liquet, B., De Micheaux, P.L., Hejblum, B.P., Thiébaut, R., 2016. Group and sparse group partial least square approaches applied in genomics context. Bioinformatics 32 (1), 35–42. https://doi.org/10.1093/bioinformatics/btv535.

M., C. (2020, March 1). *BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling*. https://cran.r-project.org/package=BAS.

Mackenbach, J.D., Rutter, H., Compernolle, S., Glonti, K., Oppert, J.M., Charreire, H., De Bourdeaudhuij, I., Brug, J., Nijpels, G., Lakerveld, J., 2014. Obesogenic environments: A systematic review of the association between the physical environment and adult weight status, the SPOTLIGHT project. BMC Public Health 14 (1). https://doi.org/10.1186/1471-2458-14-233.

Martens, A.L., Reedijk, M., Smid, T., Huss, A., Timmermans, D., Strak, M., Swart, W., Lenters, V., Kromhout, H., Verheij, R., Slottje, P., Vermeulen, R.C.H., 2018. Modeled and perceived RF-EMF, noise and air pollution and symptoms in a population cohort. Is perception key in predicting symptoms? Sci. Total Environ. 639, 75–83. https://doi.org/10.1016/j.scitotenv.2018.05.007.

Martens, A.L., Slottje, P., Timmermans, D.R.M., Kromhout, H., Reedijk, M., Vermeulen, R.C.H., Smid, T., 2017. Modeled and perceived exposure to radiofrequency electromagnetic fields from mobile-phone base stations and the development of symptoms over time in a general population cohort. Am. J. Epidemiol. 186 (2), 210–219. https://doi.org/10.1093/aje/kwx041.

Maukonen, M., Männistö, S., Tolonen, H., 2018. A comparison of measured versus self-reported anthropometrics for assessing obesity in adults: a literature review. Scand. J. Public Health 46 (5), 565–579. https://doi.org/10.1177/1403494818761971.

McLaren, L., 2007. Socioeconomic status and obesity. In Epidemiologic Reviews (Vol. 29, Issue 1, pp. 29–48). Oxford University Press. https://doi.org/10.1093/epirev/mxm001.

Meinshausen, N., Bühlmann, P., 2010. Stability selection. In J. R. Statist. Soc. B.

Mohammed, S.H., Habtewold, T.D., Birhanu, M.M., Sissay, T.A., Tegegne, B.S., Abuzerr, S., Esmaillzadeh, A., 2019. Neighbourhood socioeconomic status and overweight/obesity: A systematic review and meta-analysis of epidemiological studies. BMJ Open 9 (11), 1–12. https://doi.org/10.1136/bmjopen-2018-028238.

Molnar, C., 2020. Shapley Values | Interpretable Machine Learning. Interpretable Machine Learning, 5.9-5.10. https://christophm.github.io/interpretable-ml-book/shapley.html.

NIVEL Primary Care Registry, 2021. https://www.nivel.nl/en.

Osborne, J.W., Ph, D., 2005. Notes on the use of data transformations. Osborne, Jason. 1989, 1–8. papers3://publication/uuid/E75F318E-0E64-4077-B703-D26AA2B022DC.

Patel, S.R., Hu, F.B., 2008. Short sleep duration and weight gain: A systematic review. Obesity 16 (3), 643–653. https://doi.org/10.1038/oby.2007.118.

Pinho, M.G.M., Mackenbach, J.D., Oppert, J.M., Charreire, H., Bárdos, H., Rutter, H., Compernolle, S., Beulens, J.W.J., Brug, J., Lakerveld, J., 2019. Exploring absolute and relative measures of exposure to food environments in relation to dietary patterns among European adults. Public Health Nutr. 22 (6), 1037–1047. https://doi.org/10.1017/S1368980018003063.

Popkin, B.M., Duffey, K., Gordon-Larsen, P., 2005. Environmental influences on food choice, physical activity and energy balance. Physiol. Behav. 86 (5), 603–613. https://doi.org/10.1016/j.physbeh.2005.08.051.

Probst, P., Wright, M.N., Boulesteix, A.L., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (3), 1–15. https://doi.org/10.1002/widm.1301.

Quality of Drinking Water in Netherlands, 2018. https://www.rivm.nl/en/soil-and-water/drinking-water/quality-of-drinking-water.

Remme, R., 2017. *Netherlands Natural Capital Model-Technical Documentation*. www.atlasnatuurlijkkapitaal.nl.

Rhew, I.C., Vander Stoep, A., Kearney, A., Smith, N.L., Dunbar, M.D., 2011. Validation of the normalized difference vegetation index as a measure of neighborhood greenness. Ann. Epidemiol. 21 (12), 946–952. https://doi.org/10.1016/j.annepidem.2011.09.001.

Sarkisian, N., Gerstel, N., 2016. Does singlehood isolate or integrate? Examining the link between marital status and ties to kin, friends, and neighbors. J. Social Personal Relationships 33 (3), 361–384. https://doi.org/10.1177/0265407515597564.

Shapley, L.S., 1953.“A value for n-person games. In: Contributions to the Theory of Games (AM-28) Volume II.

Slottje, Pauline, Yzermans, C Joris, Korevaar, Joke C, Hooiveld, Mariëtte, Vermeulen, Roel C H, 2014. The population-based occupational and environmental health prospective cohort study (AMIGO) in the Netherlands. BMJ Open 4 (11), e005858. https://doi.org/10.1136/bmjopen-2014-005858.

Smith, M., Alvarez, F., 2021. Identifying mortality factors from Machine Learning using Shapley values – a case of COVID19. Expert Systems with Applications, 176(June 2020), 114832. https://doi.org/10.1016/j.eswa.2021.114832.

Stafoggia, M., Breitner, S., Hampel, R., Basagaña, X., 2017. Statistical Approaches to Address Multi-Pollutant Mixtures and Multiple Exposures: the State of the Science.

In: Current environmental health reports (Vol. 4, Issue 4, pp. 481–490). https://doi.org/10.1007/s40572-017-0162-z.
Statistics Netherlands, 2012. District and neighborhood map. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2011.
Sun, Z., Tao, Y., Li, S., Ferguson, K.K., Meeker, J.D., Park, S.K., Batterman, S.A., Mukherjee, B., 2013. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: Possible choices and comparisons. Environmental Health: A Global Access Science Source 12 (1). https://doi.org/10.1186/1476-069X-12-85.
Tutz, G., Binder, H., 2006. Generalized additive modeling with implicit variable selection by likelihood-based boosting. Biometrics 62 (4), 961–971. https://doi.org/10.1111/j.1541-0420.2006.00578.x.
Viehmann, A., Hertel, S., Fuks, K., Eisele, L., Moebus, S., Möhlenkamp, S., Nonnemacher, M., Jakobs, H., Erbel, R., Jöckel, K.H., Hoffmann, B., 2015. Long-term residential exposure to urban air pollution, and repeated measures of systemic blood markers of inflammation and coagulation. Occup. Environ. Med. 72 (9), 656–663. https://doi.org/10.1136/oemed-2014-102800.
Vrijheid, Martine, Fossati, Serena, Maitre, Léa, Márquez, Sandra, Roumeliotaki, Theano, Agier, Lydiane, Andrusaityte, Sandra, Cadiou, Solène, Casas, Maribel, de Castro, Montserrat, Dedele, Audrius, Donaire-Gonzalez, David, Grazuleviciene, Regina, Haug, Line S., McEachan, Rosemary, Meltzer, Helle Margrete, Papadopouplou, Eleni, Robinson, Oliver, Sakhi, Amrit K., Siroux, Valerie, Sunyer, Jordi, Schwarze, Per E., Tamayo-Uria, Ibon, Urquiza, Jose, Vafeiadi, Marina, Valentin, Antonia, Warembourg, Charline, Wright, John, Nieuwenhuijsen, Mark J., Thomsen, Cathrine, Basagaña, Xavier, Slama, Rémy, Chatzi, Leda, 2020. Early-life environmental exposures and childhood obesity: An exposome-wide approach. Environ. Health Perspect. 128 (6), 067009. https://doi.org/10.1289/EHP5975.
White, Ian R., Royston, Patrick, Wood, Angela M., 2011. Multiple imputation using chained equations: Issues and guidance for practice. Stat. Med. 30 (4), 377–399. https://doi.org/10.1002/sim.v30.410.1002/sim.4067.
WHO. Obesity and overweight. Factsheet, 2020. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.
Wild, C.P., 2012. The exposome: from concept to utility. Int. J. Epidemiol. 41 (1), 24–32. https://doi.org/10.1093/ije/dyr236.
Wright, M.N., Ziegler, A., 2017. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77 (1). https://doi.org/10.18637/jss.v077.i01.
Wright, M.N., Ziegler, A., König, I.R., 2016. Do little interactions get lost in dark random forests? BMC Bioinf. 17, 145. https://doi.org/10.1186/s12859-016-0995-8.
XGBoost Parameters — xgboost 1.5.0-dev documentation, n.d. Retrieved September 30, 2021, from https://xgboost.readthedocs.io/en/latest/parameter.html.
Xu, Z., Xu, X., Zhong, M., Hotchkiss, I.P., Lewandowski, R.P., Wagner, J.G., Bramble, L.A., Yang, Y., Wang, A., Harkema, J.R., Lippmann, M., Rajagopalan, S., Chen, L.C., Sun, Q., 2011. Ambient particulate air pollution induces oxidative stress and alterations of mitochondria and gene expression in brown and white adipose tissues. Part. Fibre Toxicol. 8 (1), 20. https://doi.org/10.1186/1743-8977-8-20.
Yang, A., Janssen, N.A.H., Brunekreef, B., Cassee, F.R., Hoek, G., Gehring, U., 2016. Children's respiratory health and oxidative potential of PM2.5: The PIAMA birth cohort study. Occup. Environ. Med. 73 (3), 154–160. https://doi.org/10.1136/oemed-2015-103175.
Yang, A., Wang, M., Eeftens, M., Beelen, R., Dons, E., Leseman, D.L.A.C., Brunekreef, B., Cassee, F.R., Janssen, N.A.H., Hoek, G., 2015. Spatial variation and land use regression modeling of the oxidative potential of fine particles. Environ. Health Perspect. 123 (11), 1187–1192. https://doi.org/10.1289/ehp.1408916.
Zhang, C.H., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. Vol. 38, Issue 2 https://doi.org/10.1214/09-AOS729.