Full length article

Associations between the urban exposome and type 2 diabetes: Results from penalised regression by least absolute shrinkage and selection operator and random forest models

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

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

ARTICLE INFO

ABSTRACT

Background: Type 2 diabetes (T2D) is thought to be influenced by environmental stressors such as air pollution and noise. Although environmental factors are interrelated, studies considering the exposome are lacking. We simultaneously assessed a variety of exposures in their association with prevalent T2D by applying penalised regression Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF), and Artificial Neural Networks (ANN) approaches. We contrasted the findings with single-exposure models including consistently associated risk factors reported by previous studies.

Methods: Baseline data (n = 14,829) of the Occupational and Environmental Health Cohort study (AMIGO) were enriched with 85 exposome factors (air pollution, noise, built environment, neighbourhood socio-economic factors etc.) using the home addresses of participants. Questionnaires were used to identify participants with T2D (n = 676(4.6 %)). Models in all applied statistical approaches were adjusted for individual-level socio-demographic variables.

Results: Lower average home values, higher share of non-Western immigrants and higher surface temperatures were related to higher risk of T2D in the multivariable models (LASSO, RF). Selected variables differed between the two multi-variable approaches, especially for weaker predictors. Some established risk factors (air pollutants) appeared in univariate analysis but were not among the most important factors in multivariable analysis. Other established factors (green space) did not appear in univariate, but appeared in multivariable analysis (RF). Average estimates of the prediction error (logLoss) from nested cross-validation showed that the LASSO outperformed both RF and ANN approaches.

Conclusions: Neighbourhood socio-economic and socio-demographic characteristics and surface temperature were consistently associated with the risk of T2D. For other physical-chemical factors associations differed per analytical approach.

1. Introduction

Type 2 diabetes (T2D) is a chronic disease with high individual and societal burden. Despite the genetic predisposition, environmental factors and lifestyle behaviours are important behavioural determinants in the etiology of T2D (Zheng et al., 2018). Environmental factors can affect the risk of T2D either directly (air pollutants, residential noise) or indirectly, by influencing lifestyle behaviours such as dietary habits and physical activity (walkability, green space)(Beulens et al., 2022). For instance, high neighbourhood walkability with more green areas and low levels of air pollution is associated with more physical activity (An et al., 2018; Barnett et al., 2017). With regard to direct environmental drivers, there is some evidence suggesting a potential link between T2D and exposure to arsenic in drinking water, persistent organic pollutants,

pesticides, antibiotics and several other drugs, as well as atmospheric pollutants, such as nitrogen dioxide and fine particulate matter (Misra and Misra, 2020).

A recent systematic review by Beulens et al. summarized the existing evidence on the environmental risk factors of T2D (Beulens et al., 2022). There was robust evidence for the associations of air pollution, residential noise, neighbourhood walkability, green space, area-level socio-economic deprivation and T2D and inconclusive evidence for association with outdoor temperature, neighbourhood social environment and food environment (Beulens et al., 2022; Pitt et al., 2017).

In real life environmental exposures and lifestyle behaviours are inseparable parts of the same complex structure that we call exposome. However, current evidence is mostly based on studies investigating single exposures (Wild, 2012). This might be problematic especially because such studies do not disentangle risks from associated environmental stressors. For instance, most studies focusing on air pollution did not consider green space or road traffic noise (Yang et al., 2020; Zare Sakhvidi et al., 2018). In the context of the exposome, only one study examined the association of 266 environmental factors measured in blood and urine samples with T2D. This was an environment wide association study (EXWAS) within the NHANES dataset based on biological measures of environmental, lifestyle and dietary exposures. Significant positive associations were found for the pesticide-derivative heptachlor epoxide, the vitamin c-tocopherol, and polychlorinated biphenyls and b-carotenes (Patel et al., 2010). Although the authors give a collective interpretation of results, using single or multivariable linear regression models may be ill-suited for exposome studies, because of a few reasons. First, these approaches do not consider complex interdependencies that exist between the exposures. Second, potential nonlinear exposure-outcome associations are mostly ignored. Third, when analysing a combination of highly correlated factors in a linear regression model simultaneously, generated effect estimates become unstable. Hence, more advanced statistical methods are required to analyse this type of high dimensional data.

In our previous work we showed how different methods could be applied to build interpretable and robust multi-exposure models (Ohanyan et al., 2022). Although there is not a gold standard approach for this type of analysis, the results showed that a combination of methods that complement each other by dealing with linear or nonlinear associations could be useful in capturing the overall picture of associations. For this reason, we used a linear model Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF) and Artificial Neural Networks (ANN) approaches. The last two methods can process nonlinear and non-additive associations without making any assumptions about the nature of the variables (Krogh, 2008; Stafoggia et al., 2017). LASSO is a more conventionally applied method for the purpose of variable selection among highly correlated variables (Petrovic et al., 2022; Stafoggia et al., 2017; Tibshirani, 1996). We also applied univariate analyses as to study the associations with established factors that in a multivariable model may not be selected due to low contribution to the overall fit.

The aim of this study was to examine the associations of a combination of 85 urban exposome factors and the prevalence of T2D, considering the nonlinear and non-additive associations and assess how our findings compare with the prior knowledge on established risk factors of T2D.

2. Methods

2.1. Study design and participants

We conducted a cross-sectional analysis using baseline data of the Occupational and Environmental Health Cohort (AMIGO) study. Participants across the Netherlands were randomly selected from the Dutch National General Practitioners Network database ("NIVEL Primary Care Registry," 2021). The only inclusion criterium was the age between 31

and 65 years old, as the target population were adults of working age from general population. Maximum one person per household were invited to complete the online questionnaire. Overall, 14,829 (16 % out of 93,550 invited) participants were included. A detailed description of the recruitment process, a flowchart of participant data and the ethical approval is provided elsewhere (Slottje et al., 2015).

2.2. Outcome variable

The outcome measure was prevalent T2D, assessed by self-reported questionnaires. Each participant responded to two questionnaire items: "Have you ever been diagnosed by a doctor with T2D ("non-insulin-dependent" or late-onset diabetes)?" and "Have you ever been diagnosed by a doctor with unknown type of diabetes?"(Slottje et al., 2015). Considering that T2D accounts for approximately 90 % of diabetes cases, we considered unknown type of diabetes as T2D. Among the participants who reported to have been diagnosed with T2D or unknown type of diabetes, a low number 43 (6 %) had reported age at the diagnostic less than 40 years, which might be type 1 diabetes. Therefore, a sensitivity analysis was performed where participants who reported unknown type of diabetes with an age at diagnoses less than 40 years, were not considered T2D cases.

2.3. Covariates

Self-reported questionnaire data on the duration of living at the current address, age, sex (male/female), country of birth (Netherlands/other), country of birth of mother and of father (Netherlands/other), civil state (with/without a partner), current education (high (college, university degree) / low or medium (vocational education, community college, high school)), employment status (employed/unemployed), smoking (yes/no) were considered as covariates in this study.

2.4. Exposome factors

Geospatial models, monitoring stations, satellite data, and land use databases were used to assess a large set of environmental factors. These data were then linked to each respondent's geocoded residential address, to assess exposure at the home addresses of participants as a proxy for actual exposure (Martens et al., 2018). Exposure estimates were calculated for the questionnaire data collection period (2011–2012). Overall, 85 exposures across a total of 12 exposure constructs were analysed: 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 neighbourhood (16 factors), food environment (3 factors), built environment (10 factors) and road safety (1 factor). The assessment of these constructs and variables are detailed in Table 1 and supplementary material Table S1.

2.5. Statistical analysis

2.5.1. Data pre-processing
We excluded exposures 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 S2) or ii) if two variables were correlated at a level of $r_{spearman} \geq 0.95$. In the latter case only one out of the correlated variables was included in the analysis and was considered as a proxy for the other variable(s) (Table S3). Overall, 85 exposure variables were included in the analytical models.

Before building the models, all continuous exposures were standardized to the same scale by their standard deviations (Z-score). This step helps to maximize the comparability of variable importance scores

Table 1
Description of data sources for each exposure. More detailed explanations can be found in the supplementary material.

Variables	Description	References
Temperature (C)		
Surface temperature	Surface temperature measured using Satellite pictures taken on a hot day (20 July 2016)	(Environmental Health Atlas Atlasleefomgeving, 2016; Remme, 2017)
Heat island effect	Temp difference to rural surrounding	
Combined traffic noise (Lden)	Road traffic noise levels (dB) estimated using noise model maps. It covered the whole day period and included an overweighting for noise levels during evening and night (Lden), as the nuisance perception is higher during more quiet hours of the day	(Baliatsas et al., 2016; Martens et al., 2018)
Green space (NDVI)	Normalized Difference Vegetation Index (NDVI) was used to quantify the vegetation density. Satellite images from Landsat 8, captured in September 2016 were used to generate NDVI for 100m and 1000m buffers around residential addresses	(Rhew et al., 2011)
100 m buffer		
1000 m buffer		
Electric lights at night (NanoW/cm2/sr)	Outdoor artificial light at night was assessed by the global low-light imaging data from Earth's surface. Maps (2015) from Visible Infrared Imaging Radiometer Suite Day/Night Band was used to assign an exposure value to the home address.	(Elvidge et al., 2017)
RF-EMF (mW/m^2)	Model estimates from 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/m^2)	(Beekhuizen et al., 2015; Bürgi et al., 2008; Martens et al., 2018)
Total = GSM900 + GSM1800 + UMTS		
Air pollution	Land use regression models were used to estimate annual average concentrations of the air pollutants listed below.	
NO2 (microg/m^3)		(Beelen et al., 2013; Eeftens et al., 2012)
NOx (microg/m^3)		
PM$_{2.5}$ absorbance		
PM$_{10}$ (microg/m^3)		
PM$_{2.5}$ (microg/m^3)		
PMcoarse		
UFP particle (count/cm^3)		(Kerckhoffs, 2021)
Oxidative Potential (dithiothreitol)		(Yang et al., 2015)
Oxidative Potential (electron spin resonance)		
Copper in PM$_{10}$ (ng/m^3)		(De Hoogh et al., 2013)
Iron in PM$_{10}$ (ng/m^3)		
Potassium in PM$_{10}$ (ng/m^3)		
Nickel in PM$_{10}$ (ng/m^3)		
Sulfur in PM$_{10}$ (ng/m^3)		
Silicon in PM$_{10}$ (ng/m^3)		
Vanadium in PM$_{10}$ (ng/m^3)		
Zinc in PM$_{10}$ (ng/m^3)		
Copper in PM$_{2.5}$ (ng/m^3)		
Iron in PM$_{2.5}$ (ng/m^3)		
Potassium in PM$_{2.5}$ (ng/m^3)		
Nickel in PM$_{2.5}$ (ng/m^3)		
Sulfur in PM$_{2.5}$ (ng/m^3)		
Silicon in PM$_{2.5}$ (ng/m^3)		
Vanadium in PM$_{2.5}$ (ng/m^3)		
Zinc in PM$_{2.5}$ (ng/m^3)		
Neighborhood socio-demographic characteristics	Land Use Database of Statistics Netherlands provided data on neighbourhood level socio-demographic and socio-economic factors for 2011	(Statistics Netherlands, 2012)
Inhabitants aged 0–14 years (%)		
Inhabitants aged 15–24 years (%)		
Inhabitants aged 25–44 years (%)		
Inhabitants aged 45–64 years (%)		
Inhabitants aged 65+ years (%)		
Single inhabitants (%)		
Married inhabitants (%)		
Divorced inhabitants (%)		
Widowed inhabitants (%)		
One-person households (%)		
Inhabitants with western origins (%) (Europe, North America, Oceania, Indonesia, Japan)		
Inhabitants with non-western origins (%)		
Average value of houses (x 1000 euros)		
Inhabitants with income below 40th percentile (%)		
Inhabitants with income above 20th percentile (%)		

Table 1 (*continued*)

Variables	Description	References
The number of total passenger cars		
Road safety by accident count in 200m buffer		
Built environment		
Degree of rurality (categories from 1 (high) to 5(low)	Based on address density: 1=very highly urban $\geq$ 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 = less urban 500–1,000 addresses per km^2; 5 = non-urban < 500 addresses per km^2	(Statistics Netherlands, 2012)
Distance to major road (km)		(Statistics Netherlands, 2012)
Distance to train station (km)		
Distance to larger train station (km)		
Distance to medical facilities (km)	Average distance to the general practitioner's office, pharmacies and hospitals were grouped as "access to medical facilities".	
Distance to educational facilities (km)	Average distance to kindergartens, elementary, middle, and high school facilities were grouped as "access to educational facilities".	
Distance to recreational facilities (km)	Museums, cinemas, attraction parks, concert halls, swimming pools, ice skating halls, saunas and tanning clinics were grouped as "access to recreational activities"	
Educational facilities in 10km buffer		
Distance to warehouse shops (km)		(Statistics Netherlands, 2012)
Warehouse shops in 20km buffer		
Food environment	Supermarkets and local food shops (e.g., greengrocers, bakeries, butchers etc.) were categorised as "healthy food" exposure, and restaurants, fast food restaurants and take-away places, cafés, pancake houses, bars and pubs were classified as "non-healthy food" exposure.	
Distance to healthier food retailers (km)		
Healthy food retailers (1km and 5km buffer)		
Non healthy food retailers (1km buffer)		
Quality of drinking water	Countrywide maps of drinking water quality are annually generated by the Dutch National Institute for Public Health and the Environment. Data from 2012 was used to assess the annual average values of 29 bacterial and chemical compounds, that were measured in the closest tested pump.	(Quality of Drinking Water in Netherlands, 2018)
Aluminium (µg/l)		
Natrium (ug/l)		
Nickel (ug/l)		
Nitrate (mg/l)		
Chloride (ug/l)		
Turbidity (FTE = Formazine Turbidity Units)		
Acidity, pH		
Fluoride (mg/l)		
Iron (µg/l)		
Copper (µg/l)		
Magnesium (mg/l)		
Total organic carbon (mg/l)		
Sulfate (mg/l)		
Color intensity (Pt/Co-schaal)		
Electrical conductivity (microS/cm)		
Aminomethyl phosphonic acid (Pesticide) (µg/l)		
Arsen (µg/l)		
Bentazon (herbicide)(µg/l)		
Bromat (µg/l)		
Chrome (µg/l)		
Diprogulic acid (µg/l)		
Lead (µg/l)		
Mangan (µg/l)		
pesticide: Mecoprop (µg/l)		
Nitrite (µg/l)		
Trihalomethanes (µg/l)		
Tritium (Becquerel)		
Bacteria of the coli group (kve/100 ml)		
Escherichia coli (kve/100 ml)		
Taste or smell		

RF-EMF = Radiofrequency electromagnetic field; GSM = Global System for Mobile Communication; UTMS = Universal Mobile Telecommunications System; PC4 = four digit postal code.

and minimize the impact of the measurement unit on the coefficients, independently from variable's original measurement units.

2.5.2. Missing values

The highest percentage of missing values was 11 % for the neighbourhood non-western immigrants. For the remainder of variables, the proportion of missing data was < 7 %, and the outcome measure had 2.8 % missing values. Five imputed datasets were generated using Multivariate Imputation via Chained Equations (MICE). Given the absence of a widely accepted way to combine the results from multiple imputation sets, as well as high computational cost of used statistical approaches, the imputed values were averaged across imputed datasets. Before introducing variables into the imputation model, some of them were transformed by logarithmic, root square, or inverse functions, to best

approach a Gaussian distribution, as the imputation model assumes normal distribution for predictors (Osborne and Ph, 2005)(Table S4). Note that the nature of association was inversed for the multiplicative inverse transformed variables (multiplicative inverse = 1/variable) throughout the rest of statistical analysis (Table S4). The outcome measure was imputed using all variables.

2.5.3. Univariate and multivariate analysis

Most previously published studies had a single-exposure approach. To see if established risk factors were also existent in our sample, we analysed exposome factors in univariate logistic regression models adjusted for individual-level confounders (EXWAS). We interpreted the result in light of prior evidence as summarised by Beulens et al. (2022). We followed up with an agnostic analysis by including all variables simultaneously in multivariate models using penalised regression LASSO, RF and ANN. All multivariable models were based on the same pre-processed dataset.

2.5.4. Nested cross-validation

Current state-of-the-art suggests to use nested cross-validation for the combined tuning of hyperparameters and model selection (Krstajic et al., 2014). Nested cross-validation implies that hyperparameters are selected using the inner folds of cross-validation and, an unbiased estimate of the expected accuracy of the algorithm is computed across the outer folds of cross-validation (Wainer and Cawley, 2021). Thus, we divided the dataset into training and test sets (80 % and 20 % accordingly). The training data was in turn divided into five inner folds, each including 20 % of the data. During the cross-validation, the model was iteratively trained on four inner folds. The fifth fold was used as a validation set for hyperparameter tuning. In the outer loop of cross-validation each of the outer folds was iteratively held out as a test set for the evaluation of model performance.

To maximize the comparability between statistical methods the same cross-validation folds were used for all models and stratified sampling on T2D case status (prevalence < 5 %) was used to create training and test sets. We compared predictive performances of multivariate models, using the logLoss metric (logistic loss or cross-entropy loss), which is based on probabilities and was suggested to be a better metric for model evaluation in imbalanced classification tasks (Harris and Samorani, 2021). A model that predicts perfectly would achieve a logLoss of zero, therefore the lower the logLoss, the better the prediction.

2.5.5. Penalised regression LASSO

LASSO is a penalised regression method that is commonly used in high dimensional data setting for variable selection (Tibshirani, 1996). LASSO forces the sum of the absolute value of the regression coefficients to be less than the tuning parameter lambda (λ)(Tibshirani, 1996). This causes the shrinkage of some coefficients to be zero, hence conducting a variable selection. The optimal value of lambda was selected using 5-fold cross-validation. We used subsampling based stability selection to provide finite sample control of the family-wise error rate (Meinshausen and Bühlmann, 2010). Packages "glmnet" and "stabsel" in R were used to fit the LASSO model and for stability selection respectively.

The advantage for using LASSO is that it has good properties for variable selection among highly correlated variables, hence a good interpretability. It is easy to tune and requires a low computational time. However, it is a linear model, therefore it cannot disentangle complex nonlinear or non-additive associations.

2.5.6. Random forest

RF is an ensemble learning method where at each iteration a random subset of predictors and observations is selected to build a decision tree (Ishwaran and Lu, 2019). The predictions from these trees are then aggregated to form the forest. Permutation importance was used to assess the variable importance score and Shapley values were used to assess the directions of associations (Molnar, 2020). We used a scree plot

to select variables with the highest variable importance. Packages "tuneRanger" and "ranger" were used to calibrate and to run the RF model. We calibrated the number of observations to sample for each decision tree ("sample.fraction"), the minimal size of terminal nodes to control for the depth of decision trees ("min.node.size"), and the number of variables to possibly split at each node ("mtry") using the package *tuneRanger* (Ohanyan et al., 2022).

RF can capture nonlinear and non-additive associations and recent developments in R software packages have drastically improved both the ease of the parameter tuning and the interpretability.

2.5.7. Artificial Neural Networks

The main structure of ANN consists of layers: one input layer, one or more hidden layers and one output layer. Each layer consists of neurons and weights attributed to neurons. The information passes along the network of layers until it reaches the output neurons. This is an artificial feed-forward neural network since the signals go towards one direction. The aim of a feed-forward ANN is pattern recognition; namely to find how input neurons (i.e., independent variables and covariates) predict output neurons (T2D: Yes/ No). The loss function then compares these predictions to the targets, producing a loss value: a measure of how well the network's predictions match what was expected (Chollet and Allaire, 2018). The optimizer uses this loss value to update the network's weights (Chollet and Allaire, 2018).

Parameters were calibrated through the nested cross-validation: number of hidden layers, epochs, learning rate of the Adam optimizer and penalization. The number of neurons on hidden layers (nodes) was set to 98 in each layer. The number of epochs is the number of iterations when the entire training data passes through the network, and after each epoch the weights are updated. A learning rate of $1e^{-5}$ was used for the Adam optimizer. The results of cross-validation suggested an optimal value of 0.001 L1 penalty on weights starting from the second layer. The model used sigmoid activation function by design and cross-validated batch normalisation. The "keras" package in R was used for running ANN.

Similar to the RF, ANN can incorporate nonlinear associations and interactions. In recent years ANN gained popularity for its high predictive performance in various fields. The main disadvantages of the ANN are the high computational cost and poor interpretability.

3. Results

3.1. Urban exposome and participants

Most participants (n = 14,829) were female(55.8 %) and were on average 50.7 ± 9.4 years old. Over 70 % were employed and more than one third had a higher education(38.2 %). Most respondents were originally from the Netherlands(95.3 %) and were living with a partner (80.4 %). A total of 676(4.6 %) respondents had T2D (Table 2).

The correlation plot (Fig. 1) shows that intragroup correlations (based on untransformed data) were the strongest between air pollutants and socio-demographic characteristics of neighbourhoods. Variables representing the quality of drinking water had the lowest inter- and also intragroup correlations. Moderate level correlations existed between air pollutants and neighbourhood built environmental and socio-demographic factors. Green space was negatively correlated with air pollutants. Descriptive statistics of the factors of urban exposome can be found in Table S5.

3.2. Single exposure analysis

Our univariate logistic regression results confirmed most of the established risk factors, such as air pollutants (oxidative potential of PM2.5 (DTT), potassium in PM_{10}), neighbourhood SEP (neighbourhood average home values, high-income and low-income neighbourhoods), and urbanicity level. Despite the evidence from previous studies for an

Table 2

Characteristics of the participants from baseline data of Occupational and Environmental Health cohort (AMIGO).

Characteristics	Complete cases	Mean ± SD or n(%)
Diagnosed T2D	14,410 (97.2 %)	
Yes		676 (4.6 %)
No		13,734 (92.6 %)
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 %)

T2D = Type 2 diabetes.

association between outdoor noise, green space and T2D, our results did not confirm these associations. Among suspected risk factors, we found that surface temperature, heat island effect, and the share of non-Western immigrants in neighbourhood were associated with higher odds of having T2D. We also identified a few risk factors which were never studied before in relation to diabetes (sulphate in drinking water and neighbourhood proportion of divorced inhabitants). It should be noted that the p-values generally were not very low. Eight factors had p-values lower than 0.01, among which the average home values ($p < 0.001$), high-income neighbourhoods ($p < 0.001$), low-income

neighbourhoods ($p < 0.01$), temperature ($p < 0.01$), share of non-Western immigrants ($p < 0.01$) and heat island ($p < 0.01$), share of divorced inhabitants ($p < 0.01$), urbanicity level ($p < 0.01$). An overview of results from the univariate analysis is given in Table 3 and Table 4.

3.3. Multivariable analysis LASSO

Similar to the results of the univariate analyses, LASSO showed that living in economically deprived neighbourhoods (low neighbourhood home values, low share of high-income residents) was associated with a higher risk of T2D. Living in areas with higher proportion of non-Western immigrants and higher surface temperatures was also related to a higher risk of T2D. Residents of highly urban areas had a higher risk of T2D as compared to residents from less urban areas (Table 4).

Some other factors were identified by the LASSO, but were not selected after the stability selection procedure (Table 4, Table 5). From Table 4 it can be noted that the coefficients were generally low for all the factors.

3.4. Multivariable analysis random forest

Based on estimated variable importance from the RF model, four factors were selected: neighbourhood average home values, surface temperature, share of non-Western immigrants, and green space in 1 km buffer. Lower neighbourhood SEP and higher proportion of non-Western immigrants were related with a higher risk of T2D. In Shapley plots, the associations with temperature and green space appeared non-linear (Fig. 2). Similar to the coefficients from LASSO, relative effect sizes were generally low for all predictors, as indicated by the Shapley plots (Fig. 2).

3.5. Multivariable analysis ANN

The higher prediction error rate from the nested cross-validation of ANN indicated poor performance of this model. For comparison, the average logLoss from an empty model (random classification into two groups) was 0.186(0.0006) and for the ANN: 0.177(0.006). The performance of the ANN was thus only slightly better than the random classification. In addition, comparison of logLoss estimated during hyper-parameter tuning on the inner folds (0.173) to that obtained on
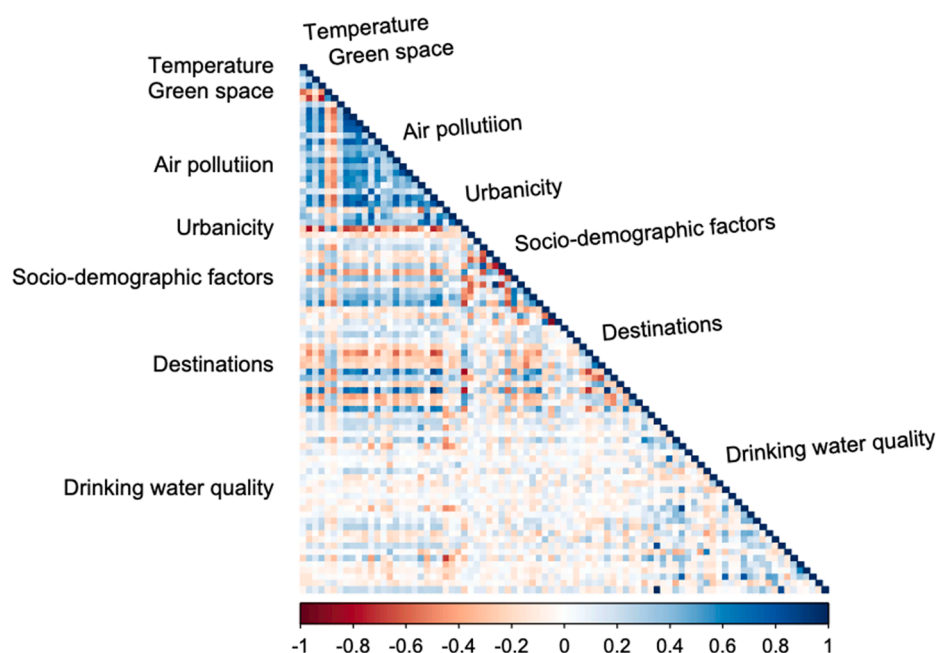

Fig. 1. Spearman correlations between constructs of exposures from the urban exposome.

Table 3

Results from the univariate analysis (EXWAS). All models were adjusted for individual confounding factors that were also used in all multivariate analysis. The exposures in all models were standardized(z-transformed).

Exposures	Estimate	Standard error	z-value	*p*-value
Established risk factors				
Average value of houses (×1000 euros)	−0,1938	0,0463	−4,19	<0.001***
Inhabitants with income below 40th percentile	0,1156	0,0429	2,69	0,0071**
Inhabitants with income above 20th percentile	−0,1716	0,0455	−3,77	<0.001***
Number of total passenger cars	−0,0315	0,0421	−0,74	0,4554
Road traffic noise (Lden > 55 dB)	0,0647	0,0874	0,74	0,4589
Green space (NDVI) 100m buffer	−0,0663	0,0419	−1,58	0,1138
Green space (NDVI) 1000m buffer	−0,0654	0,0429	−1,52	0,1277
Degree of urbanicity (categories from 1 (high) to 5 (low))	−0,0971	0,0327	−2,97	0,003**
NO2 (µg/m^3)	0,0475	0,0413	1,15	0,2501
NO$_x$ (µg /m^3)	0,0317	0,0398	0,80	0,4252
PM$_{2.5}$ absorbance (10^{-5}m^{-1})	0,0045	0,0416	0,11	0,9147
PM10 (µg /m^3)	0,0211	0,0409	0,52	0,6065
PM$_{2.5}$ (µg /m^3)	−0,0073	0,0414	−0,18	0,8602
PM coarse (µg /m^3)	0,0165	0,041	0,40	0,6882
Oxidative Potential (dithiothreitol)	0,0873	0,0419	2,08	0,0375*
Oxidative Potential (electron spin resonance)	0,0475	0,0408	1,17	0,244
UFP particle count (in cm^3)	0,012	0,0416	0,29	0,7726
Copper in PM$_{10}$ (ng/m^3)	0,0324	0,0399	0,81	0,4168
Iron in PM10 (ng/m^3)	0,0594	0,0402	1,48	0,1398
Potassium in PM10 (ng/m^3)	0,0872	0,0408	2,14	0,0325*
Nickel in PM$_{10}$ (ng/m^3)	−0,018	0,0409	−0,44	0,6601
Sulphur in PM$_{10}$ (ng/m^3)	0,0023	0,0398	0,06	0,954
Silicon in PM$_{10}$ (ng/m^3)	0,0513	0,0404	1,27	0,2045
Iron in PM$_{2.5}$ (ng/m^3)	0,0548	0,0404	1,36	0,1741
Potassium in PM$_{2.5}$ (ng/m^3)	0,0243	0,0398	0,61	0,5405
Sulphur in PM$_{2.5}$ (ng/m^3)	0,0384	0,04	0,96	0,3374
Silicon in PM$_{2.5}$ (ng/m^3)	−0,0016	0,039	−0,04	0,9676
Suspected risk factors				
Temperature	0,1194	0,0435	2,75	0,006**
Heat island effect	0,1143	0,0413	2,77	0,0057**
Electric light at night (NanoW/ cm^2/sr)	−0,0011	0,0418	−0,03	0,9785
Neighborhood inhabitants aged 0–14 years	−0,0675	0,0457	−1,48	0,1396
Neighborhood inhabitants aged 15–24 years	−0,042	0,0448	−0,94	0,3484
Neighborhood inhabitants aged 25–44 years	0,0367	0,0441	0,83	0,4045
Neighborhood inhabitants aged 45–64 years	−0,0154	0,0436	−0,35	0,7232
Neighborhood inhabitants aged 65+ years	0,0341	0,0406	0,84	0,4001
Single inhabitants in neighborhood	−0,0513	0,0454	−1,13	0,2582
Married inhabitants in neighborhood	−0,0341	0,0428	−0,80	0,4262
Divorced inhabitants in neighborhood	0,1294	0,04	3,24	0,0012**
Widowed inhabitants in neighborhood	0,0665	0,0378	1,76	0,0787
One-person households	0,0555	0,0412	1,35	0,1778
Inhabitants with western origins	0,0807	0,0408	1,98	0,0478
Inhabitants with non-western origins	0,1305	0,0402	3,24	0,0012**
Distance to healthy food outlets	−0,0532	0,0459	−1,16	0,2469
Healthy food outlets in 5km buffer	0,0424	0,0424	1,00	0,3173

Table 3 (*continued*)

Exposures	Estimate	Standard error	z-value	*p*-value
Non-healthy food outlets in 1km buffer	−0,0125	0,0424	−0,30	0,7679
Unknown risk factors				
Aluminium (µg/l)	−0,0751	0,0442	−1,70	0,0893
Mangan (µg/l)	−0,7592	0,4581	−1,66	0,0975
pesticide: Mecoprop (µg/l)	0,2566	0,2604	0,99	0,3243
Sulphate (mg/l)	−0,0888	0,0418	−2,12	0,0337*

**p*-value < 0.05, ***p*-value < 0.01, ****p*-value < 0.001.

Table 4

Cross-classification of the findings across statistical methods and previous evidence from literature as reported in the systematic review by Beulens et al. (2022).

Statistical method	Findings among the established risk factors	Findings among the inconsistent risk factors	New findings
Univariate analysis by Logistic regression	1. Neighbourhood SEP (average home values*, high- income neighbourhood*, low-income neighbourhood*) 2. Urbanicity level*4. Air pollution (oxidative potential of PM$_{2.5}$ (DTT), potassium in PM$_{10}$)	1. Temperature* 2. Heat island* 3. Non-Western immigrants*	1. Sulphate in drinking water (mg/l) 2. Proportion of divorced inhabitants*
Multivariate analysis by LASSO	1. Neighbourhood SEP (average home values*, high- income neighbourhood) 2. Urbanicity level 3. Absorbance of PM$_{2.5}$	1. Temperature* 2. Non-Western immigrants	1. 15–24 years old inhabitants of neighbourhood(%) 2. Single inhabitants in neighbourhood (%) 3. Sulphate in drinking water4. Mecoprop (herbicide) in drinking water 5. Aluminium in drinking water 6. Manganese in drinking water
Multivariate analysis by RF	1. Average home values*2. Green space (100 m and 1 km*) 3. Iron in PM$_{2.5}$	1. Temperature* 2. Heat Island 2. Non-Western immigrants*	1. Proportion of divorced inhabitants 2. Electric light at night

For random forest approach presented exposures are the top 10 important exposures.

SEP = Socio-economic position; RF = Random Forest; DTT = dithiothreitol.

[*] Indicates the factors for which the *p*-values were lower than 0.01 in single-exposure models, the factors that were selected after the stability selection in LASSO and the factors that had the highest variable importance scores in RF, as identified on scatterplot.

the outer folds (0.177) suggests possible overfitting of the model. For these reasons, we do not report results of the ANN in detail.

3.6. Comparison of the predictive performances and sensitivity analysis

Prediction error from the nested cross validation was lowest for the LASSO, when compared to RF and ANN. Average logLoss(sd) error across the outer folds was 0.168(0.003) for LASSO, 0.172(0.001) for RF and 0.177(0.006) for the ANN. Although the absolute value of the

Table 5
Results from the penalised regression: LASSO.

Variables	Beta coefficients	Probability of selection[1]
Average home values	−0,2911	0,93*
Temperature (C°)	0,0316	0,89*
Urbanicity level: 2 (high)	0,1719	0,84
Non-Western immigrants (%)	0,0412	0,7
High income neighbourhood (%)	−0,0073	0,53
Urbanicity level: 4 (low)	−0,0544	0,38
15–24 years old inhabitants (%)	−0,0594	0,35
Sulphate in drinking water (mg/l)	−0,0011	0,32
Aluminium in drinking water (µg/l)	−0,0116	0,28
Mangan in drinking water (µg/l)	−0,0685	0,16
Mecoprop (herbicide) in drinking water (µg/l)	0,0678	0,16
Single inhabitants in neighbourhood (%)	−0,1802	0,16
Absorbance of $PM_{2.5}$	−0,0636	0,05

[1] Maximum of selection probabilities after the stability selection procedure.
* Selected factors after the stability selection procedure.

average logLoss of RF was higher as compared to LASSO, the small standard error indicated on the possibility that the RF model was more stable.

The sensitivity analysis where eight participants with unknown type of diabetes who reported being diagnosed at <40 years old, were not considered as cases of T2D, showed very similar results for the univariate analysis, LASSO and RF (data not shown).

4. Discussion

We analysed a large set of environmental factors from the urban exposome as one complex system to identify the strongest predictors of T2D. Furthermore, we used univariate logistic regression to compare our findings with the knowledge from the literature, given that previously published studies mostly focused on single exposures without accounting for other related factors. Our univariate analyses based on known and suspected risk factors confirmed the associations of air pollution (oxidative potential of $PM_{2.5}$ (DTT), potassium in PM_{10}), urbanicity level, neighbourhood socio-economic position (SEP) (neighbourhood average home values, high- and low- income neighbourhoods), surface temperature and the share of non-Western immigrants in the neighbourhood with T2D risk, but not for road traffic noise and green space. The analyses in an agnostic framework (multivariable models of 85 exposures) identified associations for neighbourhood average home values, surface temperature, neighbourhood share of non-Western immigrants and green space in 1 km buffer. The factors with lower variable importance score (RF) or probability of selection (LASSO) were less consistent as they fluctuated between model runs.

Some known risk factors (air pollutants), which were identified in the univariate model, had more modest effects in more complex multivariable models. This may be because these risk factors have relatively small effects and do not add significantly to the predictive performance of the model. It could also be that they were falsely identified in previous research since other variables were not considered at the same time. The latter is possible, but as correlation patterns are likely to be different in
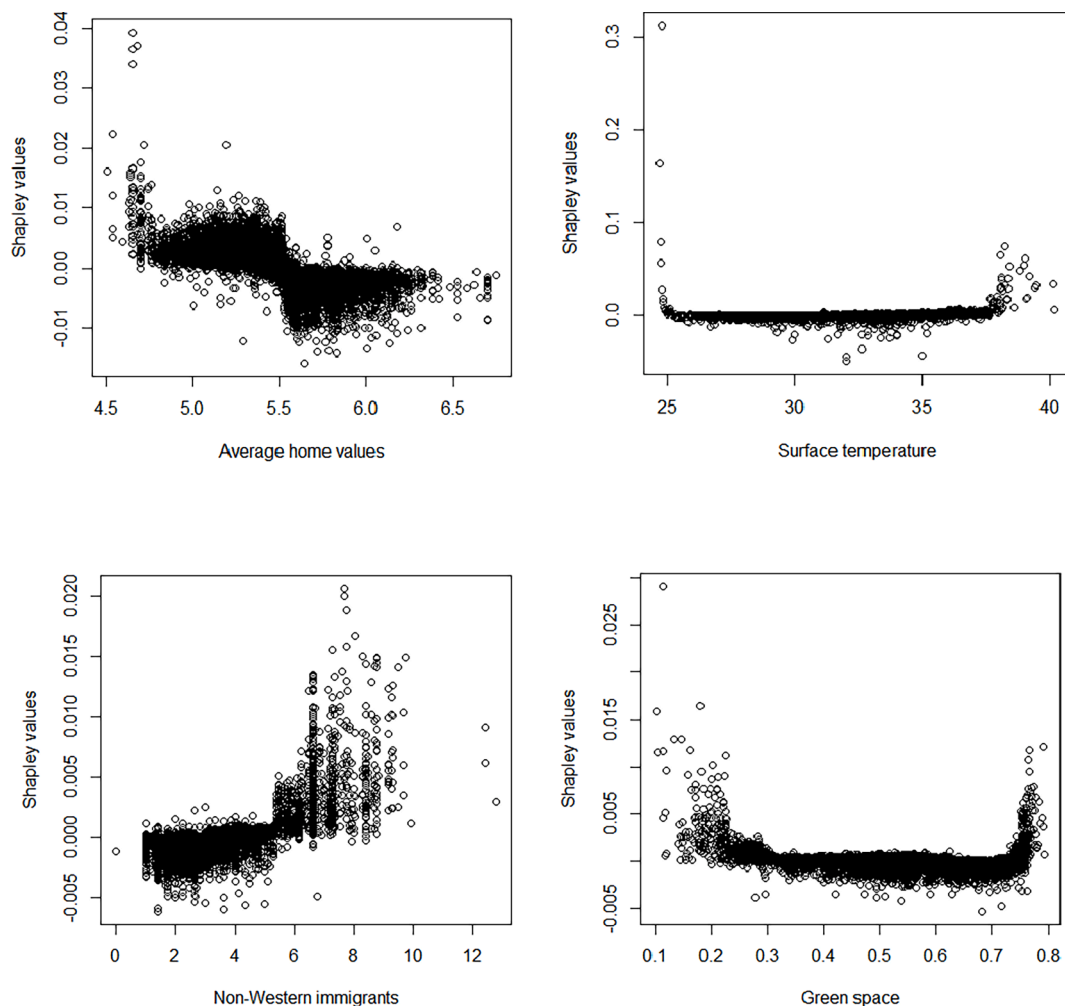

Fig. 2. Shapley plots of the top predictors of RF model. It shows the average effect of each predictor on the predicted outcome.

each study, the chances are low that they would always result in the same bias. It should be noted that only a few risk factors were selected (stability selection-LASSO and scatterplot-RF) in both multivariable models, perhaps arguing for the small effect sizes. However, these results suggest that confounding by other risk-factors could be important.

The results of this study show that neighbourhood SEP and socio-demographic characteristics are associated with T2D. The negative association with neighbourhood SEP is largely supported by earlier studies, but little is known on the question of neighbourhood socio-demographic factors like the share of immigrants (Beulens et al., 2022). The share of non-Western immigrants is a multi-component factor. It contains elements of SEP, social and cultural interactions, eating behaviours and other genetic or biological factors, which could influence the risk of T2D. For example, South-East Asians have a higher genetic or biological risk of developing T2D for the same level of body mass index or waist circumference (Chan et al., 2014; Meeks et al., 2016; Yoon et al., 2006). Some relate this risk to the propensity to store fat viscerally rather than subcutaneously, the higher degree of insulin resistance, and the lower beta-cell function (Yoon et al., 2006). In our study the models were adjusted for the participant's and their parents' countries of origin in a crude way (Dutch vs non-Dutch), therefore we cannot exclude that this finding is reflecting the higher genetic/biological risk for T2D in certain ethnic groups.

We observed a nonlinear association with surface temperature, which was also seen in the univariate analysis and LASSO (Fig. 2). The association was positive in linear models, but took a parabolic shape with highlighted extreme values on Shapley plots for RF. This discrepancy brings uncertainty for the interpretation of this association. To gain more insight on the nature of this association, we compared Shapley plots with multivariable generalised additive model splines, which showed a rather positive association (Fig.S1). More longitudinal studies are required to confirm the potential association and measure the role of temperature heat extremes for diabetes.

The plotted association for the density of green space (1 km) looked similar to the plot of temperature (Fig. 2). Recent systematic reviews and meta-analysis showed that green space has been associated with 10–20 % lower risk of T2D (Beulens et al., 2022; Bilal et al., 2018; Dendup et al., 2018). The mechanism of association might be related to walkability and more physical activity, but green space is also related with reduced air pollution, noise and heat, compensating possible harmful effects. Furthermore, the density of green space is related with urbanicity level and other characteristics of the built environment. These interrelations with different aspects of the urban exposome could perhaps explain this nonlinear relationship (Fig. 2, Fig.S1).

The oxidative potential of $PM_{2.5}$ (DTT), potassium in PM_{10}, the absorbance of $PM_{2.5}$ and the iron in $PM_{2.5}$ were identified by univariate and multivariable models, but none of these pollutants were selected. Although our univariate analysis did not find associations with $PM_{2.5}$, PM_{10} and NO_2, a large meta-analysis showed a positive association with these risk factors (Yang et al., 2020). Current evidence suggests that air pollution might change endothelial function, trigger inflammation and insulin resistance (Beulens et al., 2022; Yang et al., 2020). An explanation for the differences observed between our univariate and multivariate models could be that there is a lack of studies exploring combinations of pollutants and other related factors. Another possible explanation could be that the effect sizes of these pollutants and also the contrast in our data sample were too small, therefore do not contribute to the predictive performance of the models.

Nested cross validation allowed the comparison of nearly unbiased estimates for model performance, because the final model performance evaluation was done on the holdout test set, which was not used during the training process. Surprisingly, our analysis showed that the predictive accuracy of the ANN model for our data was lower as compared to other multivariable methods. It should be noted that we only trained a regular feed-forward ANN. It is possible that ANNs with different architectures could have had a better performance.

Despite the increasing popularity of ANNs in a wide range of fields, in this setting, with many weak predictors and very imbalanced data (due to low prevalence of diabetes (<5%)), more simple methods like penalized regression may perform better. ANNs have been developed to learn from the data and are very powerful in strong predictive tasks, such as image or text processing. Considering ANN's computational burden, the need for a lot of data and difficulties in the training, in an exposome context it is perhaps better to use alternative methods with similar properties for dealing with multiple interrelated factors with complex nonlinear or non-additive associations.

Our study was based on data from a large, nationwide cohort enriched with a wide variety of urban exposome risk factors. All exposome factors were analysed simultaneously, using RF next to penalized linear model LASSO, to identify potential nonlinear and non-additive associations. Our study has several limitations. First, the lack of availability of the timing of T2D diagnosis and the cross-sectional design is limiting causal interpretation of the findings, as the temporal link cannot be established between the exposures and the outcome. Second, the AMIGO cohort study has a potential limitation by selection bias, given the low participation rate (<16 %). Slottje et al. compared the baseline and health-related characteristics of study participants with the source population. They found no consistent indications of systematic health-related participation bias, but men below 50 years of age and those with an intermediate level of education were under-represented among cohort members, while those born in the Netherlands were over- represented, probably in part due to the fact that the questionnaire was in Dutch. However, the authors concluded that given the achieved contrast between sociodemographic, environmental factors and the results of the health-related bias analysis, limited differences with the source population are not a major concern for the internal validity of the study and if generalization to general adult population is desired, these results can be used for weighting purposes (Slottje et al., 2015). Third, the outcome measure as well as the confounding factors were assessed through self-reported questionnaire data, which is prone to errors and to the bias of desirable reporting. However, studies show that self-reported T2D is a valid measure in large-scale epidemiological studies (Li et al., 2020; Pastorino et al., 2015; Sluijs et al., 2010). In addition, we performed a sensitivity analysis excluding participants with unknown type of diabetes and age at diagnostic less than 40 years (potentially misclassified as T2D), which generated similar results. Fourth, the environmental factors were a mixture of modelled and measured factors, likely containing both types of measurement errors: classical and Berkson's error (Agier et al., 2020). 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; Ohanyan et al., 2022).

This study is one of the first to investigate the relations of various stressors from the urban exposome and the risk of T2D. Neighbourhood socio-economic and socio-demographic characteristics, surface temperature, urbanicity, and green space were related with the prevalence of T2D. Although effect sizes were small, on the population level the impact of these factors could be substantial. Therefore, targeted policy approaches that address socio-economic disparities on neighbourhood-level and measures for a better urban planning with more green areas could help to improve public health.

CRediT authorship contribution statement

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

Conceptualization, Supervision, Writing – review & editing. **Roel Vermeulen:** Resources, Conceptualization, Supervision, Writing – review & editing.

Funding

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.

Data availability

The data that has been used is confidential.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2022.107592.

References

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, 109492 https://doi.org/10.1016/j.envres.2020.109492.

An, R., Zhang, S., Ji, M., Guan, C., 2018. Impact of ambient air pollution on physical activity among adults: a systematic review and meta-analysis. Perspect. Public Health 138, 111–121. https://doi.org/10.1177/1757913917726567.

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.

Barnett, D.W., Barnett, A., Nathan, A., Van Cauwenberg, J., Cerin, E., 2017. Built environmental correlates of older adults' total physical activity and walking: A systematic review and meta-analysis. Int. J. Behav. Nutr. Phys. Act. 14, 1–24. https://doi.org/10.1186/s12966-017-0558-z.

Beekhuizen, J., Kromhout, H., Bürgi, A, Huss, A., Vermeulen, R., 2015. What input data are needed to accurately model electromagnetic fields from mobile phone base stations? J. Expo. Sci. Environ. Epidemiol. 25, 53–57. https://doi.org/10.1038/jes.2014.1.

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., Dèdelè, 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.

Beulens, J.W.J., Pinho, M.G.M., Abreu, T.C., den Braver, N.R., Lam, T.M., Huss, A., Vlaanderen, J., Sonnenschein, T., Siddiqui, N.Z., Yuan, Z., Kerckhoffs, J., Zhernakova, A., Brandao Gois, M.F., Vermeulen, R.C.H., 2022. Environmental risk factors of type 2 diabetes—an exposome approach. Diabetologia 65, 263–274. https://doi.org/10.1007/s00125-021-05618-w.

Bilal, U., Auchincloss, A.H., Diez-Roux, A.V., 2018. Neighborhood Environments and Diabetes Risk and Control. Curr. Diab. Rep. 18 https://doi.org/10.1007/s11892-018-1032-2.

Bürgi, A., Theis, G., Siegenthaler, A., Röösli, M., 2008. Exposure modeling of high-frequency electromagnetic fields. J. Expo. Sci. Environ. Epidemiol. 18, 183–191. https://doi.org/10.1038/sj.jes.7500575.

Chan, J.C.N., Yeung, R., Luk, A., 2014. The Asian diabetes phenotypes: Challenges and opportunities. Diabetes Res. Clin. Pract. 105, 135–139. https://doi.org/10.1016/j.diabres.2014.05.011.

Chollet, F., & Allaire, J.J., 2018. Deep Learning with R.

De Hoogh, K., Wang, M., Adam, M., Badaloni, C., Beelen, R., Birk, M., Cesaroni, G., Cirach, M., Declercq, C., Dèdelè, 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, 5778–5786. https://doi.org/10.1021/es400156t.

Dendup, T., Feng, X., Clingan, S., Astell-Burt, T., 2018. Environmental risk factors for developing type 2 diabetes mellitus: A systematic review. Int. J. Environ. Res. Public Health 15. https://doi.org/10.3390/ijerph15010078.

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., Korek, M., Lanki, T., Lindley, S., Madsen, C., Mo, A., Na, G., Nieuwenhuijsen, M., Nonnemacher, M., Pedeli, X., Raaschou-nielsen, O., Patelarou, E., Quass, U., Ranzi, Я.A., Schindler, Å. C., Stempfelet, M., Stephanou, E., Sugiri, D., Tsai, M., Yli-tuomi, T., Varro, J., Vienneau, D., Klot, S. Von, Wolf, K., Brunekreef, B., 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. 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, 5860–5879. https://doi.org/10.1080/01431161.2017.1342050.

Harris, S.L., Samorani, M., 2021. On selecting a probabilistic classifier for appointment no-show prediction. Decis. Support Syst. 142, 113472 https://doi.org/10.1016/j.dss.2020.113472.

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

Environmental Health Atlas (Atlasleefomgeving) [WWW Document], 2016, URL https://www.atlasleefomgeving.nl/en (accessed 7.22.22).

J.Kerckhoffs, 2021. Modelling Nationwide Spatial Variation of Ultrafine Particles based on Mobile Monitoring (in revision).

Krogh, A., 2008. What are artificial neural networks? Nat. Biotechnol. 26, 195–197. https://doi.org/10.1038/nbt1386.

Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. J. Cheminform. 6, 1–15. https://doi.org/10.1186/1758-2946-6-10.

Li, H.L., Fang, J., Zhao, L.G., Liu, D.K., Wang, J., Han, L.H., Xiang, Y.B., 2020. Personal characteristics effects on validation of self-reported type 2 diabetes from a cross-sectional survey among Chinese adults. J. Epidemiol. 30, 516–521. https://doi.org/10.2188/jea.JE20190178.

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.

Meeks, K.A.C., Freitas-Da-Silva, D., Adeyemo, A., Beune, E.J.A.J., Modesti, P.A., Stronks, K., Zafarmand, M.H., Agyemang, C., 2016. Disparities in type 2 diabetes prevalence among ethnic minority groups resident in Europe: a systematic review and meta-analysis. Intern. Emerg. Med. 11, 327–340. https://doi.org/10.1007/s11739-015-1302-9.

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

Misra, B.B., Misra, A., 2020. The chemical exposome of type 2 diabetes mellitus: Opportunities and challenges in the omics era. Diabetes Metab. Syndr. Clin. Res. Rev. 14, 23–38. https://doi.org/10.1016/j.dsx.2019.12.001.

Molnar, C., 2020. Shapley Values | Interpretable Machine Learning. Interpret. Mach. Learn. 5 (9-5), 10.

Ohanyan, H., Portengen, L., Huss, A., Traini, E., Beulens, J.W.J., Hoek, G., Lakerveld, J., Vermeulen, R., 2022. Machine learning approaches to characterize the obesogenic urban exposome. Environ. Int. 158, 107015 https://doi.org/10.1016/j.envint.2021.107015.

Osborne, J.W., Ph, D., 2005. Notes on the use of data transformations. Osborne, Jason 1–8.

Pastorino, S., Richards, M., Hardy, R., Abington, J., Wills, A., Kuh, D., Pierce, M., 2015. Validation of self-reported diagnosis of diabetes in the 1946 British birth cohort. Prim. Care Diabetes 9, 397–400. https://doi.org/10.1016/j.pcd.2014.05.003.

Patel, C.J., Bhattacharya, J., Butte, A.J., 2010. An environment-wide association study (EWAS) on type 2 diabetes mellitus. PLoS ONE 5. https://doi.org/10.1371/journal.pone.0010746.

Petrovic, D., Bodinier, B., Dagnino, S., Whitaker, M., Karimi, M., 2022. Epigenetic mechanisms of lung carcinogenesis involve differentially methylated CpG sites beyond those associated with smoking. Eur. J. Epidemiol. https://doi.org/10.1007/s10654-022-00877-2.

Pitt, E., Gallegos, D., Comans, T., Cameron, C., Thornton, L., 2017. Exploring the influence of local food environments on food behaviours: A systematic review of qualitative literature. Public Health Nutr. 20, 2393–2405. https://doi.org/10.1017/S1368980017001069.

NIVEL Primary Care Registry [WWW Document], 2021. URL https://www.nivel.nl/en (accessed 1.13.21).

Quality of Drinking Water in Netherlands [WWW Document], 2018. URL https://www.rivm.nl/en/soil-and-water/drinking-water/quality-of-drinking-water (accessed 12.17.20).

Remme, R., 2017. Netherlands Natural Capital Model-Technical Documentation.

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, 946–952. https://doi.org/10.1016/j.annepidem.2011.09.001.

Slottje, P., Yzermans, C.J., Korevaar, J.C., Hooiveld, M., Vermeulen, R.C.H., 2015. The population-based occupational and environmental health prospective cohort study (AMIGO) in the Netherlands. BMJ Open 4. https://doi.org/10.1136/bmjopen-2014-005858.

Sluijs, I., van der A, D.L., Beulens, J.W.J., Spijkerman, A.M.W., Ros, M.M., Grobbee, D.E., van der Schouw, Y.T., 2010. Ascertainment and verification of diabetes in the EPIC-NL study. Neth. J. Med. 68, 333–339.

Stafoggia, M., Breitner, S., Hampel, R., Basagaña, X., 2017. Statistical Approaches to Address Multi-Pollutant Mixtures and Multiple Exposures: the State of the Science. Curr. Environ. Heal. reports. 10.1007/s40572-017-0162-z.

Statistics Netherlands, 2012. District and neighborhood map [WWW Document]. URL https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2011 (accessed 12.18.20).

Tibshirani, R., 1996. Regression Shrinkage and Selection Via the Lasso. J. R. Stat. Soc. Ser. B 58, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Wainer, J., Cawley, G., 2021. Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Syst. Appl. 182, 115222 https://doi.org/10.1016/j.eswa.2021.115222.

Wild, C.P., 2012. The exposome: From concept to utility. Int. J. Epidemiol. https://doi.org/10.1093/ije/dyr236.

Yang, B.-Y., Fan, S., Thiering, E., Seissler, J., Nowak, D., Dong, G.-H., Heinrich, J., 2020. Ambient air pollution and diabetes: A systematic review and meta-analysis. Environ. Res. 180, 108817 https://doi.org/10.1016/J.ENVRES.2019.108817.

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, 1187–1192. https://doi.org/10.1289/ehp.1408916.

Yoon, K.H., Lee, J.H., Kim, J.W., Cho, J.H., Choi, Y.H., Ko, S.H., Zimmet, P., Son, H.Y., 2006. Epidemic obesity and type 2 diabetes in Asia. Lancet 368, 1681–1688. https://doi.org/10.1016/S0140-6736(06)69703-1.

Zare Sakhvidi, M.J., Zare Sakhvidi, F., Mehrparvar, A.H., Foraster, M., Dadvand, P., 2018. Association between noise exposure and diabetes: A systematic review and meta-analysis. Environ. Res. 166, 647–657. https://doi.org/10.1016/J.ENVRES.2018.05.011.

Zheng, Y., Ley, S.H., Hu, F.B., 2018. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat. Rev. Endocrinol. 14, 88–98. https://doi.org/10.1038/nrendo.2017.151.