Planning: Measuring the association between urban environment and risk behaviours

Polibienestar (UVEG)

2024-03-27

This strategic document provides an overview of the planned approach to analyze the association between urban environment characteristics and risk behaviors for Non-Communicable Diseases (NCDs), specifically Type 2 Diabetes Mellitus (T2DM) and Cardiovascular Diseases (CVD). The analysis adopts a cross-sectional design, leveraging self-reported data from surveys. It is important to note that the analysis is non-causal - as it does not provide an estimate of the effect of an event's occurrence on the disease occurrence - but aims to provide insights into association patterns and risk factors.

Characterisation of urban environments

Initially, the urban environment will be characterised through a clustering algorithm selected based on data characteristics, incorporating variables representative of the following categories (prepared by BAX): continuous or discrete variables - following a Poisson distribution -, at the neighborhood level. Classification will be split by cities - Valencia, Rotterdam and Rijeka -. By applying the clustering algorithm to each city's dataset independently, the analysis can take into account the possible heterogeneity of the data structures resulting from the correlative heterogeneity in the processes generating the data. The following table contains a non-exhaustive summary of the categories, which may include one or more variables.

Physical & Functional Attributes	
Dimension	**Category**
Density	Population density
	Business and retail density
Mobility	Street connectivity
	Location connectivity
	Cyclability
	Walkability
	Public transport
	Traffic
Mixticity	Proximity to social services
	Proximity to sport infrastructure
	Proximity to recreational & commercial amenitites
	Proximity to public spaces
	Food environment
Green infrastructure	Green coverage
	Green space diversity
	Green space continuity

Social attributes	
Dimension	**Category**
(not decided yet)	(not decided yet)

According to these categories, we assume as plausible the existence of four main clusters: High NCD-protective environments - High Socio-Economic Status (SES; HH), Low NCD-protective environments - High SES (LH), HL, and LL. However, due to heterogeneity within neighbourhoods, it is possible that the optimal number of clusters may vary between cities. Nevertheless, given our plausible classification, the structure of the data can be examined using Principal Component Analysis (PCA) and the comparison of the clusters obtained in a biplot. For illustration purposes, this is what the analysis would look like in the R environment.

We generate a synthetic dataset simulating 20 variables across four predefined clusters (HH, HL, LH, LL) with 20 observations each, using specific mean values and a structured covariance matrix to model variable correlations within two groups of variables (1:10, 11:20) but not between them. The `factoextra` and `MASS` libraries support the creation, manipulation, and analysis of this dataset, including the application of multivariate normal distribution and the assembly of a complex correlation structure. This process is intended to simulate obtaining the dataset of a city, with each row corresponding to a neighbourhood and each column to a variable characterising the urban environment.

```
library(factoextra)
library(MASS)
```

```
set.seed(123)

n <- 20 # Number of observations per cluster
k <- 20 # Number of variables
sd <- 0.35 # Standard deviation
mean_high <- 0.75 # Mean for High
mean_low <- -0.75 # Mean for Low

# Correlation within groups of variables
corr_within <- 0.8

# Create a correlation matrix for the first 10 variables
corr_matrix1 <- matrix(corr_within, nrow = 10, ncol = 10)
diag(corr_matrix1) <- 1

# Create a correlation matrix for the next 10 variables
corr_matrix2 <- matrix(corr_within, nrow = 10, ncol = 10)
diag(corr_matrix2) <- 1

# Create zero matrices for correlations between groups of variables.
zero_matrix <- matrix(0, nrow = 10, ncol = 10)

# Combine the matrices into a 20x20 correlation matrix
corr_matrix <- rbind(cbind(corr_matrix1, zero_matrix),
                     cbind(zero_matrix, corr_matrix2))

# Convert correlation matrix to covariance matrix
cov_matrix <- corr_matrix * (sd^2)

# Initialize the data frame and generate data
```

```r
data <- matrix(NA, nrow = 4 * n, ncol = k)

for (i in 1:4) {

  if (i == 1) {
    mean_vec <- rep(mean_high, k)  # HH
  } else if (i == 2) {
    mean_vec <- c(rep(mean_low, k/2), rep(mean_high, k/2))  # LH
  } else if (i == 3) {
    mean_vec <- c(rep(mean_high, k/2), rep(mean_low, k/2))  # HL
  } else {
    mean_vec <- rep(mean_low, k)  # LL
  }

  data[((i-1)*n + 1):(i*n), ] <- mvrnorm(n, mu = mean_vec, Sigma = cov_matrix)
}

# Adding cluster labels
clusters <- factor(rep(c("HH", "LH", "HL", "LL"), each = n))

# Create a data frame
data <- as.data.frame(data)
data$cluster <- clusters
```

We perform k-means clustering on the synthetic dataset. The final clustering method may vary according to the intrinsic characteristics of the data.

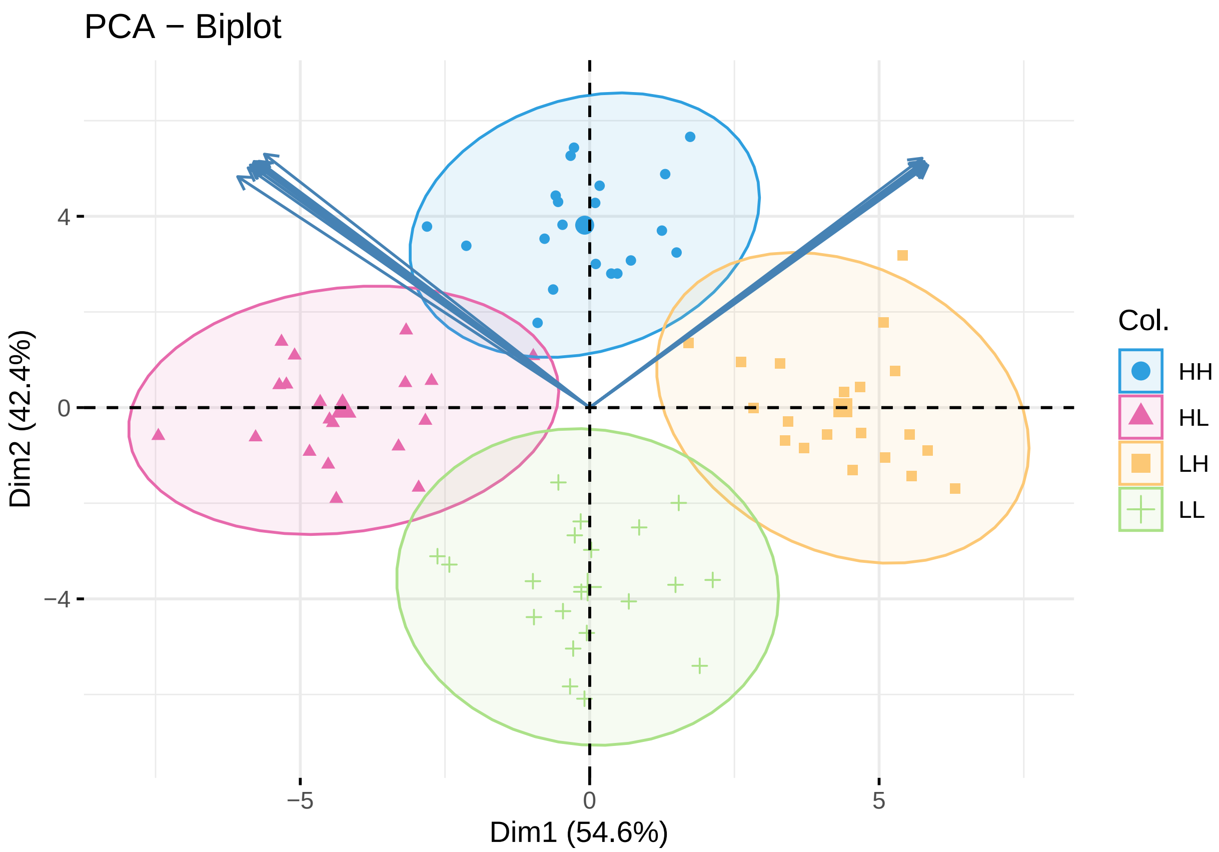
```r
# Exclude the true cluster labels from the clustering process
clusters_kmeans <- kmeans(scale(data[, 1:k]), centers = 4, nstart = 25)

# Add the k-means cluster assignments to the data frame
data$kmeans_cluster <- as.factor(clusters_kmeans$cluster)
```

We conduct PCA and visualize the clusters using a biplot.

```r
# Running PCA on the scaled dataset without the true cluster labels
pca_result <- prcomp(scale(data[, 1:k]), center = TRUE, scale. = TRUE)

# Visualizing the PCA biplot with factoextra
fviz_pca_biplot(pca_result, label = "none",
                col.ind = data$cluster, # color by true cluster label
                palette = c("#2E9FDF", "#E769AC", "#FCC875", "#ABE188"),
                addEllipses = TRUE, # add ellipses to identify clusters
                ellipse.level = 0.95) # confidence level for the ellipses
```



This code illustrates the generation of a fictional dataset and its analysis using k-means classification and PCA in the R environment, with the aim of exploring complex data structures in the context of urban environments and their association with NCDs. Based on the assumption of the existence of four main clusters - NCD-protective environments with high and low SES, as well as their counterparts with low levels of protection - the code generates a dataset with 20 normally distributed continuous variables to simulate indicators of interest. Subsequently, the k-means algorithm is applied to identify clusters within this dummy data.

PCA is performed to reduce the dimensionality of the data, allowing effective visualisation through a biplot. This biplot not only demonstrates how the clusters are distributed in principal component space, but also facilitates comparison between the actual classifications and those obtained by k-means. Such a visualisation exemplifies how planned analysis can identify patterns of association and underlying structures in the data.

Risk variables

Secondly, we have a set of variables of interest related to NCD risk factors from self-reported surveys and barometers. We want to study the effect of neighbourhood types (clusters previously obtained) on these behavioural variables and surrogate markers, adjusting for socio-demographic variables. We have a set of variables common to all three cities, with data at the individual level:

- At least n days per week spending x minutes or more on commuting {binary: yes/no}
- Doing sport in leisure time {binary: yes/no}
- At least n days per week spending x minutes or more doing sport {binary: yes/no}
- Smoking habit {binary: yes/no}

- Alcohol consumption {binary: yes/no}
- Frecuency of alcohol consumption {ordinal}
- Intensity of alcohol consumption {ordinal}
- Body Mass Index (BMI) {continuous}

In addition to these common variables, each city will analyse the effect of the typology of the respondent's neighbourhood of residence on their own variables of interest, not necessarily available in all cities. This approach will allow us to have a measure for the specific effect of the environment, characterised according to a set of multivariate criteria, on the outcome of interest.

The estimation approach will be based on the use of logistic regressions, selecting between binary, ordinal or multinomial variants according to the specific nature of the Dependent Variable (DV). For variables specified in binary, ordinal or multinomial form, the corresponding type of logistic regression will be applied. The variables of interest, whether risk behaviours or surrogate markers, will be treated as the DV within the model. The cluster will be treated as a categorical Independent Variable (IV) with four different levels, with level 'LL' designated as the reference category. This will allow the Odds Ratio (OR) to be calculated for each of the other levels in comparison to 'LL'. Additionally, the application of Poisson Regression will be considered for the analysis of count variables where relevant, as well as Ordinary Linear Regression (OLS) for the treatment of continuous variables, as appropriate.

A generic outline for the implementation of binary logistic regression in R is provided below:

```r
data <- within(data, cluster <- relevel(cluster, ref = 'LL'))
model <- glm(outcome ~ cluster + individual_covariates,
             family = binomial(link='logit'), data = data)
```

Another generic example for the implementation of Poisson regression:

```r
data <- within(data, cluster <- relevel(cluster, ref = 'LL'))
model <- glm(count_outcome ~ cluster + individual_covariates,
             family = poisson(link='log'), data = data)
```