



Discussion Paper

Mixed-Methods Clustering of Multi-source Data to Identify Population Subgroups Based on BMI-Associated Variables

Maik Beuken
Kees van Berkel
Melanie Kleynen
Iris Kanera
Jan van den Brakel

Oktober 2026

Abstract

Purpose: Overweight and obesity are growing public health concerns in the Netherlands. This study aims to assess a healthcare informatics-driven analytical approach for uncovering BMI-related population subgroups in a large multi-source health dataset using variable selection, dimensionality reduction, and clustering, to support tailored health behavior interventions.

Methods: A mixed methods approach combined DHS data with proximity-to-facilities and air pollution data, supplemented by expert evaluation. Missing data were imputed using MICE with Random Forests. Statistically Equivalent Signature (SES) variable selection identified 47 BMI-related variables. Clustering was performed using k-means on principal components derived via PCAMIX. Silhouette scores guided the optimal number of clusters, and cluster stability was assessed using bootstrap-based Jaccard similarity measures. An expert panel explored the practical relevance of the findings.

Results: Data from 15,202 respondents revealed 10 clusters in each of the total, overweight, and obese groups, and 9 in the morbid obesity group. Cluster profiles differed in socioeconomic position, and health conditions. Bootstrap-based stability analyses showed that most clusters were stable to highly stable. Experts recognized associations between BMI and 42 of the 47 selected variables and identified four overarching patterns: (1) drug use and STI testing, (2) chronic health conditions, (3) low socioeconomic status, and (4) migration background.

Conclusion: This study demonstrates the value of a healthcare informatics workflow for identifying stable BMI-related population subgroups in heterogeneous health data. The findings suggest that variable selection enhances cluster interpretability and stability and highlight the multifaceted nature of BMI. Although the identified clusters are not yet suitable for direct clinical application, they provide a foundation for future research on personalized prevention and targeted interventions.

Keywords: BMI, Clustering, Health Data Analytics, Obesity, Machine Learning

Contents

1	Introduction	4
2	Description of the Data	5
3	Methods	7
3.1	Data Preparation	7
3.2	Variable Selection	9
3.3	Identification of Clusters	11
3.4	Software	14
3.5	Method Expert Panel	14
4	Results	15
4.1	Results Data Analysis	16
4.2	Results Expert Panel	22
5	Discussion	25
6	Conclusion	28
	References	29
	Appendices	34
A	Data Cleaning Procedure	34
B	Transformation of the data	34
C	Removing near zero variance variables	35
D	List of the 47 variables used in the PCAMIX and clustering analysis	37
E	Scree plots	41
F	Table with SC scores to determine the amount of clusters	44
G	Explanation of the interpretation of the clustering results by investigating the different visualizations.	45
H	Tables with a description of the results for each BMI category separately.	69

1 Introduction

The increasing prevalence of obesity and lifestyle-related chronic diseases has become a pressing global health concern Blüher 2019; World Health Organization 2025; according to the Dutch Health Survey (DHS), obesity prevalence among adults aged 20 years or older increased threefold, from 5% in the early 1980s to 16% in 2023, with severe obesity rising concurrently from 1% to 4% over the same period Statistics Netherlands 2024. Research has shown that tailored health behavior interventions lead to significantly better health outcomes than non-tailored interventions Noar et al. 2007; Lustria et al. 2013. Now that healthcare is becoming increasingly complex and the needs of individual patients vary, healthcare professionals can benefit from data-driven insights to better understand which factors are most important within specific patient groups and adjust their decisions accordingly. Identifying distinct subgroups within the population, each characterized by specific combinations of (health-related) characteristics, is a key step toward developing such tailored interventions. Consequently, data-driven approaches, particularly those based on machine learning, have gained attention for their ability to uncover hidden patterns in large-scale health datasets Benke and Benke 2018; Zhang et al. 2019. Recent research has shown that unsupervised machine-learning methods, particularly k-means clustering, are effective in segmenting heterogeneous populations into meaningful health-related subgroups based on shared characteristics in health data Aljohani 2024. For example, Nagamine et al. 2022 deployed the k-means algorithm to detect symptom patterns reported by heart failure patients ($n = 25,861$), identifying clusters based on distinguishing variables to better understand disease characteristics and progression. Similarly, Nicolet et al. 2022 applied a combination of Random Forest (RF) and HDBSCAN clustering on healthcare claims data from Swiss residents aged 50 and above ($n = 18,732$) to explore multimorbidity and patient complexity in clinical settings. Furthermore, Elbattah et al. 2017 grouped elderly hip fracture patients in Ireland into subpopulations to optimize care strategies tailored to individual needs and Takeshita et al. 2024 identified seven distinct obesity clusters ($BMI \geq 35$) using healthcare claims data ($n = 9,494$), aiming to inform more targeted interventions. Collectively, these studies demonstrate the value of unsupervised machine learning for identifying meaningful subgroups and uncovering complex relationships within large-scale health data. Notably, most of these studies used the k-means clustering algorithm, highlighting its potential to enhance our understanding of patterns and relationships in lifestyle-related data.

While previous studies have demonstrated the value of clustering techniques for identifying subgroups in obesity and healthcare data, most have applied clustering directly to a predefined set of variables. In contrast, this study proposes an integrated analytical framework designed for complex, heterogeneous, multi-source health data. The framework combines statistically guided variable selection using the Statistically Equivalent Signatures (SES) algorithm, dimensionality reduction for mixed data using Principal Component Analysis for mixed type data (PCAMIX), k-means clustering, and expert interpretation within a single workflow. By first identifying variables that provide unique information with respect to BMI, the approach aims to reduce noise from uninformative variables that may hinder generalization and distort dimensionality reduction and clustering outcomes. This may improve the interpretability and stability of the resulting clusters. In addition, the framework integrates DHS data with environmental and proximity-to-facility data, enabling the exploration of BMI-related patterns. To the best of our knowledge, there are no studies that combine BMI-oriented variable selection, mixed-data dimensionality reduction, clustering, and expert interpretation within a nationally representative population health dataset. This study therefore contributes on three levels: methodologically, by evaluating this analytical workflow; epidemiologically, by identifying BMI-related population

subgroups; and translationally, by exploring the clinical relevance of these subgroups.

The aim of this study was twofold: (1) to evaluate a healthcare informatics-driven analytical workflow for identifying and describing BMI-related population subgroups in the overall population and within BMI categories and (2) to explore the practical relevance of the resulting subgroup profiles through expert interpretation. Importantly, the objective was not to maximize differences in BMI between groups, but to identify subgroups with similar multivariate profiles within a BMI-related variable space.

The remainder of this paper is organized as follows: section 2 describes the dataset used. Section 3 details the methodology, including data preprocessing, imputation of missing values, variable selection, dimensionality reduction, clustering of respondents, data visualization techniques, and the method of the expert panel study. Section 4 presents the results. Section 5 discusses the findings. Finally, Section 6 concludes the paper.

2 Description of the Data

We combined 3 datasets. The main part of the data used is the DHS data which is an annual, nationally representative survey conducted by Statistics Netherlands (CBS) to monitor the health status, healthcare utilization, lifestyle behaviors, and preventive practices of the Dutch population. The survey has been conducted continuously since 1981, with methodological refinements over time to enhance data quality and comparability. The survey targets individuals of all ages residing in private households in the Netherlands. The survey is based on stratified simple random sampling using equal selection probabilities within strata from the Dutch municipal population register. The survey covers approximately 9,500 respondents per year. Data are collected using a mixed-mode approach. Initially, respondents are invited to complete the questionnaire online (Computer-Assisted Web Interviewing, CAWI). According to an adaptive survey design, a selection of non-respondents is subsequently approached for face-to-face interviews (Computer-Assisted Personal Interviewing, CAPI) Berkel et al. 2020. The DHS covers a broad range of health-related topics, including self-perceived health and functional limitations, chronic conditions, mental health, healthcare utilization, medication use, lifestyle behaviours, and preventive health measures. Healthcare utilization variables include visits to general practitioners, specialists, and hospital admissions. Lifestyle variables include smoking, alcohol consumption, physical activity, and dietary habits, while preventive measures include influenza vaccination uptake Statistics Netherlands 2025a. In this study the responses observed in 2022 and 2023 are used. The DHS data are widely used in public health research, policy evaluation, and epidemiological studies.

The DHS data were enriched with The Proximity to Facilities register, published by Statistics Netherlands (CBS). These data provide detailed spatial statistics on the accessibility of various public and commercial facilities across the Netherlands. The dataset is designed to support spatial planning, urban development, and social policy by quantifying both the distance to and the density of essential services. It offers insights into the spatial distribution of services by reporting: the shortest travel distance from residential addresses to the nearest facility in meters over the road network and the number of facilities that can be reached, by road, from a residential address within a specified radius (e.g., 1 km, 3 km, 5 km, 10 km or 20 km). The dataset includes proximity and density measures for a wide range of facilities, such as: educational institutions (primary and secondary schools), healthcare services (general practitioners,

hospitals, pharmacies), retail (supermarkets, shopping centers), public transport (train station), and social and cultural amenities (libraries, childcare centers, restaurants, theaters) Statistics Netherlands 2025b. The Proximity to Facilities data were linked to respondents at the individual level using a unique key variable.

We also used data from the archive of the large-scale concentration (GCN) maps for the Netherlands, provided by the Dutch National Institute for Public Health and the Environment (RIVM) National Institute for Public Health and the Environment 2025a. The GCN data comprise annual large-scale air pollution concentration maps for the Netherlands, covering a range of atmospheric pollutants including nitrogen dioxide (NO_2), ammonia (NH_3), particulate matter (PM_{10} and $\text{PM}_{2.5}$), ozone (O_3), carbon monoxide (CO), and sulfur dioxide (SO_2). These maps are generated through a combination of atmospheric dispersion modeling and empirical measurements, ensuring spatially detailed and temporally consistent representations of air quality across the Netherlands National Institute for Public Health and the Environment 2025b. The air-pollution data were linked to individuals at the neighbourhood level.

3 Methods

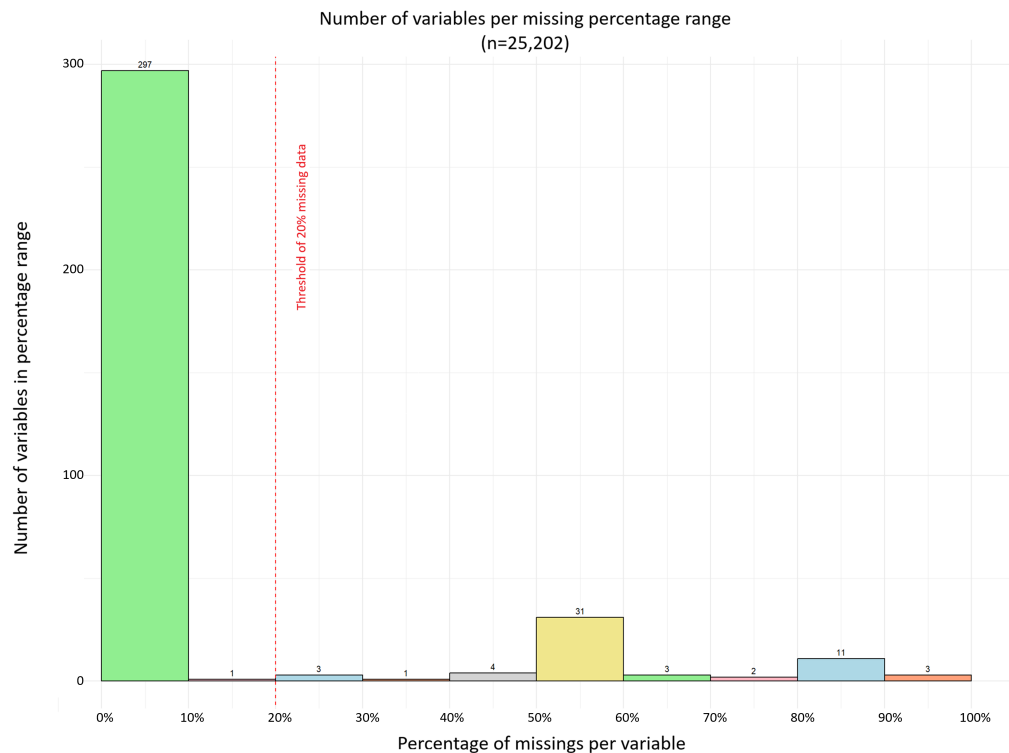
3.1 Data Preparation

All data were assessed for aberrant measurement data and missing data. First, variables measuring the same concept were removed. This is due to the dataset characteristics in which variables are derived from responses to a single question in the DHS questionnaire. In addition, respondents who did not report on weight or height were excluded. The same applied to respondents younger than 18 years, who were excluded to avoid systematic missing data. Respondents with other systematic missing data were also excluded from the analysis. A detailed overview of all respondent- and variable-level exclusion steps is provided in **Appendix A (Fig. A.1)**. In the next step, all remaining missing values that can be filled in, in a deductive manner, were imputed using rule-based procedures. As part of the data preparation process, variables with more than 20% missing data were removed prior to clustering as this may be due to the routing of the questionnaire. The choice of a 20% threshold for variable exclusion is motivated by evidence that imputation performance deteriorates as missingness increases. Simulation studies have shown that Multiple Imputation by Chained Equations (MICE) maintains high imputation performance up to approximately 20% missingness on evaluation metrics such as Root Mean Square Error (RMSE), Mean Absolute Deviation (MAD), bias, and proportion variance. From 30% missingness onward, some simulations show a drop in performance, particularly in terms of imputation bias Junaid et al. 2025. Other research indicates that bias and RMSE grow steadily with increasing missingness, with noticeable deviations in bias occurring when using a Random Forest-based imputation method on variables with more than 20% missing values Feng et al. 2021. While some methods are more robust than others, imputation generally becomes problematic under an MNAR mechanism when missing data exceeds 20% for many techniques Clavel et al. 2014. These findings support the use of a conservative cutoff, such as 20%, to balance data retention with imputation reliability.

To provide more detail on the amount of missing data, an overview of the missingness is given in **Fig. 3.1**.

To impute the missing data in the remaining variables, we used the MICE method from the mice package in R Buuren and Groothuis-Oudshoorn 2011. This method models each incomplete variable separately conditional on the other variables. The parametric use of MICE assumes linear relationships between variables. However, the data used exhibits non-linearities and interactions. Random Forest (RF) can accommodate such complexities without requiring explicit model specification, potentially leading to less biased and more efficient estimates of the missing data Shah et al. 2014; Doove et al. 2014; Biau and Scornet 2016. RF can also impute heterogeneous data, unlike most standard methods, with high imputation accuracy Stekhoven and Bühlmann 2012. RF combines predictions from multiple decision trees, thereby improving robustness and prediction accuracy. Traditional linear and distance-based algorithms face limitations in detecting complex structural relationships and latent patterns present in the data. Additionally, RF demonstrates robust performance when working with datasets containing many variables, while remaining resistance to overfitting phenomena Tang and Ishwaran 2017. This is in contrast to techniques such as K-Nearest-Neighbors (KNN), which experience degraded performance in high-dimensional environments Pestov 2013. These advantageous characteristics supported the decision to employ the MICE methodology paired with RF as an appropriate strategy for data imputation. We generated one imputed dataset ($m = 1$) using 10 iterations ($\text{maxit} = 10$) and 100 trees per random forest ($\text{ntree} = 100$). Additionally, a fixed random seed was used to guarantee the reproducibility of these outcomes. A single imputed dataset was

Figure 3.1 Histogram of the Amount of Missing Data in the DHS 2022-2023



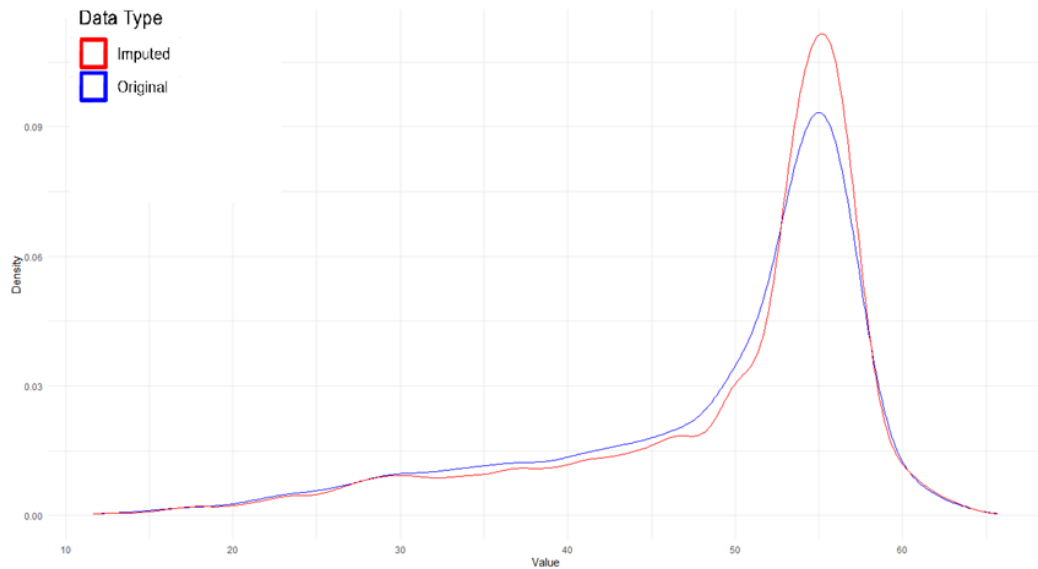
generated because imputation served as a preprocessing step for subsequent unsupervised learning analyses rather than for statistical inference. To verify that this choice did not distort the underlying data structure, the imputation procedure was subsequently evaluated through distributional comparisons under MCAR, MAR, and MNAR missingness scenarios.

MICE is generally considered suitable for handling data that are Missing Completely at Random (MCAR) and Missing at Random (MAR). Simulation studies suggest that MICE can also perform relatively well at low to moderate levels of missingness (up to 60%) under the Missing Not At Random (MNAR) mechanism Pereira et al. 2024. Given that the proportion of missing data in our study falls within this range, MICE was considered an appropriate and pragmatic choice. Despite this knowledge, we have evaluated the quality of the imputation by simulating these 3 different missingness patterns in a random sample of the data. To reduce computational burden, we randomly selected 5000 records and 60 variables. Thereafter, we used the `amputeData` function from the `miceRanger` package in R to randomly amputate the data to simulate an MCAR mechanism. To evaluate the results, we plotted density plots for continuous variables, drawing the imputed data over the data with missing values and plotting bar charts for the categorical variables. To simulate a MAR mechanism, we introduced 50% missing values in a specific group of respondents. After imputation, we evaluated the results using the same plotting methods as in the MCAR scenario.

Furthermore, we simulated the MNAR pattern by dividing the data into two groups based on the median value of the variable `PUBSF12Fysiek`, a summary measure for physical health being: (1) a group with low physical health and (2) a group with high physical health. We introduced 30% missing values in the low physical health group and 5% in the high physical health group. After imputing the simulated incomplete dataset generated under the MNAR mechanism, we compared the imputed data with the original complete data (before introducing missingness). By analyzing the plot in **Fig. 3.2**, we concluded that the imputation did not substantially alter the

distribution of the variable. This provided confidence to proceed with imputing the actual missing values in the real dataset. We then compared the imputed data with the observed data to check for consistency in the distribution of this variable. **Fig. 3.3** shows the distribution of the variable after imputation compared to the original observed distribution with actual non-response. Together, **Fig. 3.2** and **Fig. 3.3** indicate that the imputation preserved the original distribution.

Figure 3.2 Comparison of densities of original data and simulated MNAR nonresponse, which was subsequently imputed using MICE for the SF-12 health measure variable



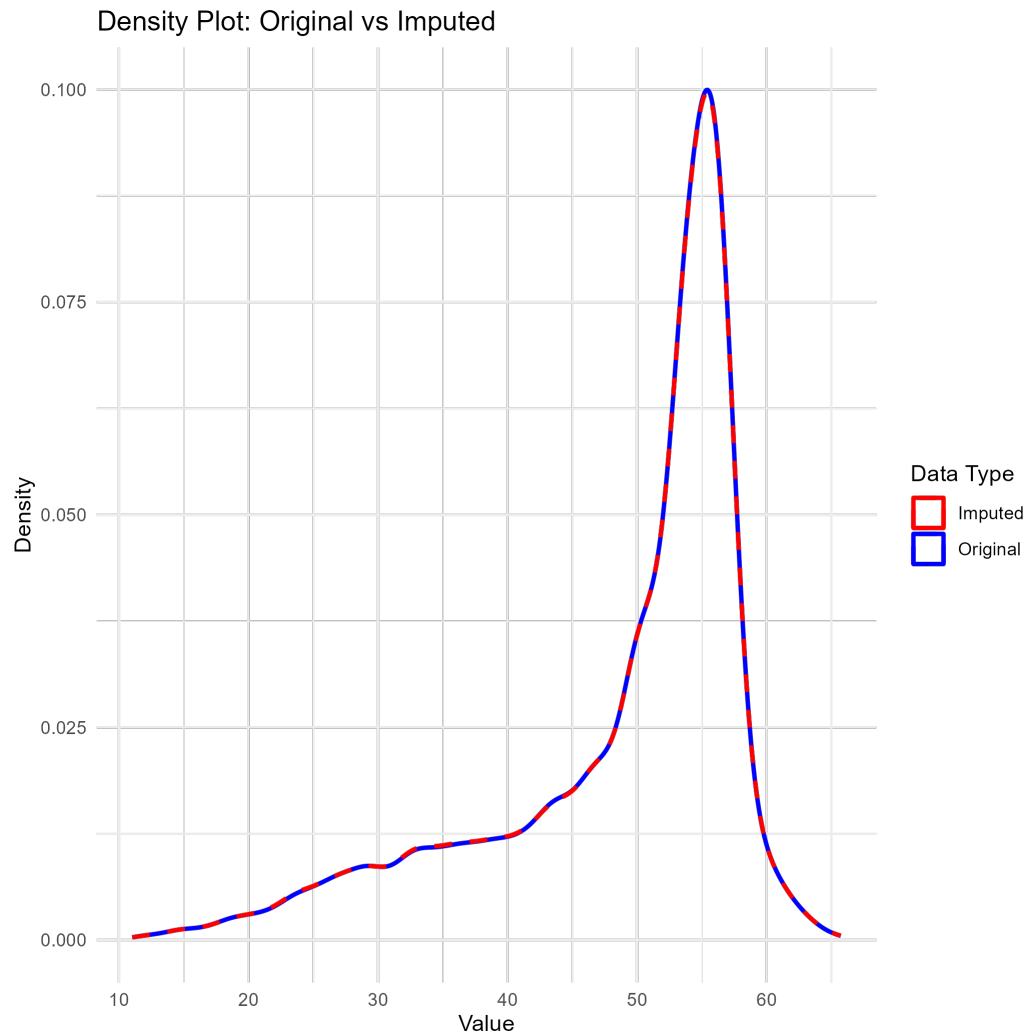
Before inclusion in the variable selection algorithm, the data was transformed to satisfy the key regression assumptions. This enhances the reliability and interpretability of regression models, which form the foundation of the variable selection algorithm. See **Appendix B** for a deeper explanation. To reduce the risk of including uninformative variables, we pre-filtered variables before applying the Statistically Equivalent Signatures (SES) algorithm by removing those with low or zero variance, i.e., variables with mostly identical or very few unique values. For a detailed explanation of this step, see **Appendix C**.

3.2 Variable Selection

We identified the most relevant subset of variables related to our target variable, BMI, from the available variables. This subset was subsequently used as input for a PCAMIX-based dimensionality-reduction and cluster analysis aimed at identifying patterns across BMI-related indicators. In the variable-selection step, continuous BMI was used as the target variable rather than BMI categories. While BMI categories are useful for interpretation, categorization inevitably results in a loss of information, as individuals in the middle of a category are treated as equivalent to those at the boundaries, obscuring meaningful variation. By retaining BMI as a continuous variable, the full spectrum of variability is preserved.

To reduce unnecessary complexity and enhance the interpretability of the results, while retaining the most relevant information for the target variable, we applied a parsimonious variable-selection approach prior to the cluster analysis. A parsimonious model avoids unnecessary complexity by retaining only the most informative predictors, facilitating

Figure 3.3 Comparison of densities of original data and actual nonresponse, which was subsequently imputed using MICE for the SF-12 health measure variable



interpretation of the results. To reduce redundancy among the retained variables, we applied the Statistically Equivalent Signatures (SES) algorithm Tsamardinos et al. 2012, a constraint-based feature-selection method that iteratively adds and removes variables based on conditional independence tests. The SES hyperparameters α (significance-level threshold) and max_k (maximum conditioning-set size) were optimized using the 10-fold cross-validation procedure (`cv.ses`) implemented in the MXM package Lagani et al. 2017. Candidate values of $\alpha \in \{0.10, 0.05, 0.01\}$ and $max_k \in \{2, 3, 5\}$ were evaluated. Larger values of max_k allow conditional independence to be assessed using larger conditioning sets, but at substantially increased computational cost. The optimal configuration was $\alpha = 0.10$ and $max_k = 3$, which was subsequently used in the SES algorithm. Additionally, a fixed random seed was used to guarantee the reproducibility of these outcomes.

In contrast to the least absolute shrinkage and selection operator (LASSO) or elastic net (EN), which may over-select variables and thereby reduce predictive performance by including predictors that contribute little additional information to the outcome Liu et al. 2023, SES focuses on identifying minimal, non-redundant variable subsets that retain predictive information. This is particularly advantageous when model simplicity, interpretability, and stability are prioritized. Furthermore, SES explicitly accounts for the possibility that multiple non-redundant variable

subsets may contain essentially the same information with respect to the target variable. Rather than identifying a single optimal predictor set, SES identifies statistically equivalent signatures that achieve comparable predictive performance. By identifying such signatures while reducing redundancy, SES offers a parsimonious and interpretable representation of the data that is well aligned with the objectives of the present study.

Constraint-based algorithms have been shown to retrieve highly predictive variable subsets Aliferis et al. 2010. These algorithms construct networks of variables connected by edges, which represent potential relationships. The connections are evaluated through repeated statistical tests of conditional independence. A variable is considered conditionally independent of the target if, given any subset of other variables, it does not provide additional predictive power. When such independence is established, the corresponding edge is removed from the network Lagani et al. 2017.

By assessing conditional independence, we identified how variables influence one another, providing insight into the underlying mechanisms that generated the data Lagani et al. 2017; Tsagris et al. 2018. The SES algorithm uses an iterative forward–backward filtering process. In the forward phase, it adds variables that are significantly associated with the target, conditional on the variables already selected. In the backward phase, it removes variables that have become redundant by testing whether they remain significantly associated with the target, given all other subsets of selected variables. Because our dataset contained mixed predictor variables and a continuous target variable, we used the `testIndReg` conditional independence test implemented in SES Lagani et al. 2017.

We consider this variable-selection approach prior to dimensionality reduction important, since in most cases including all variables in a statistical analysis is unnecessary and may even hinder correct interpretation of the data Pacheco et al. 2013. Although PCAMIX produces a lower-dimensional representation, it does not reduce dimensionality with respect to the original variables, as each principal component depends on all variables. As stated by McCabe McCabe 1984, “interpretation of the results and possible subsequent data collection and analysis still involve all of the variables”. In addition, this variable selection step is particularly important for subsequent dimensionality reduction using PCAMIX because, as noted in previous research Bai and Ng 2008, including redundant variables can deteriorate the estimation of principal components, as these variables still receive a non-zero weight in the construction of the components. By filtering out redundant variables prior to PCAMIX, we aim to obtain more reliable component estimates, improving both the stability and interpretability of the subsequent clustering results.

3.3 Identification of Clusters

Prior to clustering, dimensionality reduction was performed using PCAMIX Kiers 1991. Because the selected variables consisted of both continuous and categorical variables, PCAMIX was used to obtain a lower-dimensional representation that appropriately accommodates both data types. The method combines the matrix of standardized numerical values with the centered indicator matrix for categorical variables, thereby effectively weighting the different variable types. This allows for the extraction of principal components that simultaneously represent the variance in both continuous and categorical variables Chavent et al. 2022.

PCAMIX projects data onto a subspace spanned by eigenvalues that capture most of the variance, helping to identify the directions in which the data vary most, thus illuminating the data structure and making clusters more distinct Olsson et al. 2013. The contributions of variables to

the variance of principal components are called squared loadings, defined as squared correlations (R^2) for numerical variables and correlation ratios (η^2) for categorical variables. Squared loadings quantify the strength of the association between variables and principal components. To facilitate interpretation, the component structure was rotated Kiers 1991, which improves the clarity of results by redistributing variance along rotated axes. This made it easier to identify groups of variables with high loadings on the same component, simplifying the structure for better understanding. Importantly, this does not change the underlying relationships in the data; however, it does provide a better presentation of the results, which may lead to more meaningful conclusions Chavent et al. 2012. Based on these considerations, the data on rotated axes were processed before the k-means algorithm was applied.

We used scree plots, in which the proportion of explained variance for each Principal Component (PC) is depicted, to determine the number of PCs to retain. In particular, we used the elbow method to interpret the scree plots by identifying points at which the decrease in explained variance started to level off. These points indicate where additional components contribute progressively smaller amounts of explained variance. These scree plots were used to guide the selection of a limited number of components, rather than to interpret the substantive meaning of individual PCs.

The final dimensionality was not selected solely on the basis of a single elbow location. Instead, all plausible elbow locations identified from the scree plots were evaluated. Specifically, dimensions were considered when the decrease in explained variance appeared to stabilize, and additional candidate dimensions were examined when a subsequent larger decline in explained variance was followed by a new stabilization pattern. The resulting dimensionality solutions were then compared in terms of cluster structure before selecting the final solution.

We used these PCs to group this multidimensional data into clusters of respondents, with the aim of minimizing within-cluster variance and thus forming groups with similar patterns in the response outcomes. To accomplish this, a proven unsupervised machine learning technique, the k-means algorithm, was used. Since k-means begins with randomly chosen coordinates for the initial centroids, we set the number of starts to 50 (nstart=50) to ensure the algorithm is executed with different random initial cluster assignments. The maximum number of iterations allowed was set to 100 (iter.max=100). The solution with the lowest total within-cluster sum of squares, i.e., the most compact clusters, is automatically selected as the final result. Additionally, a fixed random seed was used to guarantee the reproducibility of clustering outcomes. Specifically, we used the MacQueen k-means algorithm. This method can be extended in a straightforward fashion to general metric spaces, including those where categorical variables introduce non-Euclidean distances, in which the notion of the arithmetic mean becomes problematic MacQueen 1967.

K-means struggles when variables are irrelevant to the underlying cluster structure and exhibit intercorrelations. This may negatively influence the clustering process and interfere with the identification of an optimal clustering subspace Yamamoto and Hwang 2014. Therefore, dimensionality reduction is an important step prior to clustering. These PCs are, by construction, non-correlated linear combinations of the variables and can be viewed as new numerical variables with maximum dispersion and maximum linkage to the original variables Chavent et al. 2022. Because k-means relies on Euclidean distances and seeks to minimize within-cluster variation while maximizing between-cluster variation, its performance may be adversely affected not only by correlated variables but also by differences in variable scales. The orthogonal component scores obtained through PCAMIX provide a more suitable clustering space by yielding centered and standardized representations of the major dimensions of variation in the data Jansson et al. 2022. In addition, the reduced-dimensional representation facilitates the

visualization and interpretation of the identified clusters, allowing both respondents and variables to be explored within a common low-dimensional space. Within this reduced-dimensional space, highly similar respondents form clusters.

That is why the k-means algorithm was performed on the retained PCs, with the number of clusters evaluated for values of $n \leq k \leq 30$, where n denotes the number of retained PCs. The lower limit ensures that variation represented by each retained PC can contribute to cluster formation. To assess clustering performance and determine the appropriate number of clusters, we used the average Silhouette Coefficient (SC), defined as the mean silhouette score across all respondents in the dataset. The average SC quantifies within-cluster cohesion and between-cluster separation, with values approaching 1 indicating compact and well-separated clusters Rousseeuw 1987. For each diagnostic group, the optimal number of clusters was identified as the configuration yielding the highest average SC value.

In addition to the average SC value, we assessed cluster-wise stability using a non-parametric bootstrap-based approach, in which cluster-wise stability was quantified by the mean Jaccard similarity between each original cluster and its most similar counterpart in the bootstrapped datasets Hennig 2007. This procedure was implemented using the `cClusterboot` function with the number of resampling runs for each scheme set to 100 ($B=100$). The same initialization settings were used as in the primary clustering analysis to ensure consistency.

This procedure was applied not only to grouped data segmented by adult BMI categories: (1) underweight ($BMI < 18.5$), (2) normal weight ($BMI 18.5 - 24.9$), (3) overweight ($BMI \geq 25.0$), (4) obese ($BMI \geq 30.0$), and (5) morbid obesity ($BMI \geq 40.0$), but also to the entire dataset. We employed this two-fold analysis to obtain a more nuanced and comprehensive perspective on the results of this study.

To provide a comprehensive overview of the data structure for the entire group, as well as for subgroups based on BMI categories, we explored relationships between mixed-type data using various visualizations. These plots help assess the PCs and interpret the underlying patterns. For continuous variables, we used a correlation plot to visualize their interrelationships. For categorical variables, we employed a level plot, where the categories of each variable are projected onto the PCs. To interpret clusters, we included individual plots that display all respondents on the PC dimensions, with each cluster represented by a distinct color and its corresponding centroid. To investigate the relationship between continuous and categorical variables, we examine the squared loadings plot. This allows us to relate clusters to both types of variables by observing their behavior on the PCs. Clusters with centroids located far from the origin along a PC can be interpreted by identifying similar patterns in both continuous and categorical variables. Categories that frequently co-occur tend to be positioned close together; the same principle applies to continuous variables. The distance between categories or variables reflects the strength of their association, with those farther from the origin being more distinctive. Respondents and cluster centroids, exhibit similar behavior in individual plots. The interpretation process can be viewed as an overlay procedure, in which the individual plot, correlation plot, and level plot were conceptually superimposed, enabling us to focus on the most distinctive categories and continuous variables and to describe the characteristics of each cluster. A more detailed explanation of our approach to interpreting the visualization of the cluster results is provided in **Appendix G: Fig. G.1 through Fig. G.27**, where we describe the method used to analyze the clusters for the total group.

3.4 Software

R version 4.4.0 (2024-04-24), R Foundation for Statistical Computing, was used. RStudio (2024.12.1.563) was used as an integrated development environment for the R programming language. R was used together with the following packages:

- **mice**: The R package mice imputes incomplete multivariate data by chained equations Buuren and Groothuis-Oudshoorn 2011.
- **MXM**: Provides many variable selection methods for a wide range of response variables, including minimal, statistically equivalent, and equally predictive variable subsets Lagani et al. 2017; Biza et al. 2025.
- **stats** (kmeans): K-means clustering MacQueen 1967.
- **PCAmixdata**: Implements principal component analysis, orthogonal rotation, and multiple factor analysis for a mixture of quantitative and qualitative variables Chavent et al. 2022.
- **corrplot**: Visualization of a correlation matrix Wei and Simko 2024.
- **fpc** (clusterboot): Visualization of a correlation matrix Hennig 2026; Hennig 2007.
- **BestNormalize**: Performs a suite of normalizing transformations and selects the best one based on the Pearson P -test statistic for normality Peterson 2021; Peterson and Cavanaugh 2020.

3.5 Method Expert Panel

An expert panel was asked to assess the applicability and the potential practical value of the quantitatively derived clusters. Expert consultation is a commonly used approach to contextualize data-driven findings and evaluate their relevance for clinical practice Fetters et al. 2013. The aims of the expert panel were to (1) explore whether an association between the selected variables and BMI was recognized by experts and to (2) discuss the clinical and practical relevance of cluster analysis. To achieve these aims, a mixed-methods design was employed. First, an online survey was conducted, followed by in-depth qualitative group discussions to capture group dynamics and collective views on the topic Tenny et al. 2022. The expert panel focused on the mutually exclusive BMI categories overweight, obesity and morbid obesity, as individuals falling within these BMI categories are often the target groups for interventions.

3.5.1 Study Population

Experts, with relevant expertise, were recruited through the professional networks of the research team, with additional participants identified through referrals from experts initially contacted. The study population comprised researchers from a variety of relevant fields, including obesity research, medical and psychosocial health, lifestyle behavior, and environmental health sciences. Participants were eligible if they had demonstrable expertise in one or more of the aforementioned domains and experience in data interpretation and/or public health. All participants were proficient in Dutch and willing to take part in either the online survey, the expert panel discussions, or both. Informed consent was obtained from all participating experts.

3.5.2 Procedure Expert Panel

For the online survey, the questionnaire was developed by the research team to investigate and describe expert opinions on the perceived relationships between the selected variables (see 3.2) and BMI. For each variable, the experts were invited to respond to the following multi-choice, question "Do you recognize a relationship between the identified variable and BMI?" The answering options were (1) "No, I see no relationship"; (2) "Yes, there may be a relationship"; and (3) "I do not know". Following completion of the online survey, the expert panel group interviews were conducted. Before the discussion sessions, participants received a standardized template containing the cluster results for the overweight, obesity, and morbid obesity BMI categories, ensuring consistency in interpretation across participants. In addition to the distinguishing variables per cluster, the following descriptive variables were presented for each cluster: BMI (mean; standard deviation (SD)), age (mean; SD), percentage of male participants, and the percentage of participants reporting (very) good self-rated health. The survey results were described in terms of the variables that were recognized or not recognized by the experts.

During the first part of the group sessions, to guide the discussion, experts were asked to first record their interpretations on several key questions, in writing, individually. This procedure was applied to minimize the potential influence of group dynamics and dominant participants while encouraging independent reflection prior to collective discussion Jünger et al. 2017. The key questions were: Which clusters appear more or less clinically or practically relevant? Which clusters could potentially be merged into one subgroup? Which cross-category clusters do you recognize? Following this individual reflection, participants were encouraged to identify and discuss similarities and differences between answers. In the second part, the discussion focused on identifying content-related patterns or constellations that could represent specific subgroups for future interventions.

The discussions were guided by a semi-structured discussion guide. A moderator facilitated the sessions, encouraged participation from all experts, and ensured that all predefined key questions were addressed. The discussions were audio taped, while a second researcher took field notes and documented key observations throughout the sessions. Following each discussion, the field notes were reviewed and expanded to capture contextual observations, reflections on the discussion process, and key themes raised by participants. Audio recordings were reviewed alongside the field notes to verify their accuracy and completeness, and the notes were refined where necessary to ensure a comprehensive representation of the discussion and the key issues raised by participants Halcomb and Davidson 2006.

The data were analyzed using thematic content analysis. The main outcomes of the group discussions were clustered and described for each BMI category (overweight, obesity, and morbid obesity). Initial themes were identified based on the field notes and audio recordings. To enhance the credibility of the analysis, a second researcher independently reviewed the materials and the preliminary themes. The themes were subsequently refined through iterative review of the recordings. Illustrative quotes from the experts were included to further substantiate the findings.

4 Results

4.1 Results Data Analysis

After data preparation, a total of 15,202 respondents were included in the analysis, out of an initial 15,515 participants aged 18 years and older. Respondents ranged in age from 18 to 99 years. Characteristics of the total group and the subgroups underweight, normal weight, overweight, obese, and morbid obesity are presented in **Table 4.1**.

Table 4.1 Basic characteristics of respondents (N = 15,202)

	Total	Underweight	Normal weight	Overweight (not obese)	Obese (not Morbid Obesity)	Morbid Obesity
	(A)	(U)	(N)	(O)	(B)	(M)
N	15202	248	7244	5395	2134	181
Female (%)	51.8	71.4	55.2	44.5	55.3	63.5
Age (mean, SD)	52.0 (18.7)	41.7 (21.7)	49.3 (19.7)	54.9 (17.3)	54.9 (16.2)	51.6 (16.3)
BMI (mean, SD)	25.8 (4.6)	17.5 (0.8)	22.5 (1.7)	27.1 (1.4)	32.9 (2.4)	44.6 (5.2)
Education (%)						
Primary	2.9	4.8	2.0	3.3	4.5	5.0
Lower secondary	17.6	16.9	13.5	20.6	23.2	25.4
Upper secondary	35.9	27.0	32.6	38.1	42.0	41.4
Bachelor	28.2	30.6	32.6	25.2	21.8	16.6
Master/PhD	15.4	20.6	19.3	12.8	8.4	11.6
Cluster membership (%)						
Cluster no. 1	7.8	10.5	6.2	8.9	7.4	11.6
Cluster no. 2	8.2	11.3	14.0	7.1	6.1	2.2
Cluster no. 3	3.8	10.1	8.8	19.1	2.0	12.7
Cluster no. 4	14.7	14.9	6.1	8.8	4.0	13.8
Cluster no. 5	5.7	16.5	7.2	10.9	7.6	17.7
Cluster no. 6	6.8	6.0	17.2	13.0	18.4	13.3
Cluster no. 7	10.6	7.7	4.8	5.3	16.2	11.6
Cluster no. 8	8.2	23.0	10.3	16.4	9.6	8.8
Cluster no. 9	15.2	NA	6.5	7.5	12.6	8.3
Cluster no. 10	18.9	NA	18.9	3.1	16.2	NA

Alphabetical prefixes (A, U, N, O, B, M) indicate population-specific clustering solutions and do not imply correspondence across groups (e.g., A1 $\neq$ O1).

4.1.1 Variable Selection

Applying the SES algorithm to our dataset resulted in the selection of 47 variables (see **Appendix D**) from an initial set of 285 variables. Each selected variable contributed unique information with respect to the target variable, BMI. Owing to the nature of the SES algorithm, highly dominant variables can overshadow the contribution of other potentially informative variables from a domain knowledge perspective. Preliminary analyses indicated that age acted as such a dominant variable, resulting in a substantially smaller and less diverse set of selected variables. Because the aim of this study was to identify a broad range of BMI-related characteristics for exploratory clustering rather than to identify the strongest predictors of BMI, age was excluded from the SES procedure. However, age was reintroduced during the interpretation of the clustering results to ensure that its relevance was incorporated into the broader analysis.

The SES variable selection process produced several signatures, representing alternative combinations of variables. Specifically, three pairs of interchangeable variables were identified:

1. Region, expressed either as a province of the Netherlands or as a Municipal Health Service region (GGD).
2. Smoking status, either as a general indicator or specifically for ex-smokers.
3. Diabetes status, either as a general indicator or specifically for type 2 diabetes.

We selected the following variables: Region expressed as a province, Smoking status as a general indicator, and Diabetes status defined as type 2 diabetes. These choices were motivated as follows:

1. Province provides a more intuitive and widely understood regional classification.
2. Overall smoking status offers a more comprehensive representation of respondents' smoking behavior than the ex-smoker category.
3. Type 2 diabetes was included, next to the already selected type 1 diabetes variable, as a distinct variable to ensure a clear and non-overlapping representation of diabetes-related characteristics within the selected variable set.

4.1.2 Identification of the Clusters

Based on scree plots (see **Fig. E.1** through **Fig. E.6** in **Appendix E**), we determined the appropriate number of PCs to retain by identifying the elbow points. To assess the robustness of the dimensionality selection, alternative solutions corresponding to other plausible elbow locations in the scree plots were also evaluated. For the overall population, solutions with 10, 11, and 12 dimensions were examined; for the overweight group, 8, 10, and 11 dimensions; for the obese group, 8, 10, and 12 dimensions; and for the morbid obesity group, 4, 6, and 8 dimensions. Lower-dimensional solutions generally resulted in a loss of finer-grained cluster differentiation, whereas higher-dimensional solutions provided limited evidence of additional cluster structure. These findings supported the selected dimensionalities presented in **Table 4.2**.

After selecting the number of PCs, SC scores were calculated for a range of cluster solutions within each group. **Appendix F** provides the SC scores for all evaluated cluster counts. **Table 4.2** presents an overview of the selected number of PCs and the corresponding number of clusters for each respondent group.

Table 4.2 Number of principal components and clusters for each group

Cluster prefix	Group	Principal components (PCs)	Clusters (<i>k</i>)
A	Total group (<i>N</i> = 15,202)	10	10
U	Underweight (<i>n</i> = 248)	8	8
N	Normal weight (<i>n</i> = 7,244)	10	10
O	Overweight (not obese) (<i>n</i> = 5,395)	10	10
B	Obese (not morbidly obese) (<i>n</i> = 2,134)	10	10
M	Morbid obesity (<i>n</i> = 181)	6	9

Alphabetical prefixes (A, U, N, O, B, M) indicate population-specific clustering solutions and do not imply correspondence across groups (e.g., A1 ≠ O1).

4.1.3 Assessment of the Clusterwise stability

Cluster stability was assessed using a non-parametric bootstrap-based resampling approach. Mean Jaccard similarity values were calculated between each original cluster and its most similar counterpart in the bootstrapped datasets (**Table 4.3**). These values were interpreted using established thresholds Hennig 2007: values ≤ 0.5 indicate unstable (dissolved) clusters; values between 0.6 and 0.75 indicate patterns in the data, although cluster membership remains uncertain; values ≥ 0.75 indicate valid, stable clusters; and values ≥ 0.85 indicate highly stable clusters.

Table 4.3 Cluster Stability Assessment Based on Mean Jaccard Similarity Values

Cluster	Overall	Overweight	Obese	Morbid obesity
1	0.989***	0.984***	0.839**	0.820**
2	0.990***	0.926***	0.966***	0.798**
3	0.986***	0.974***	0.915***	0.951***
4	0.994***	0.935***	0.674*	0.785**
5	0.978***	0.987***	0.913***	0.862***
6	0.965***	0.958***	0.981***	0.673*
7	0.992***	0.928***	0.985***	0.849**
8	0.991***	0.986***	0.762**	0.662*
9	0.996***	0.989***	0.946***	0.880***
10	0.995***	0.962***	0.973***	NA

*** Highly stable cluster (≥ 0.85)

** Stable cluster (≥ 0.75)

* Showing patterns (0.6–0.75)

Table 4.4 Comparison of Cluster Stability With and Without SES Feature Selection

Ranked Cluster	With Feature Selection		Without Feature Selection	
	Mean Jaccard	SD	Mean Jaccard	SD
1	0.996	0.003	0.986	0.021
2	0.995	0.002	0.980	0.009
3	0.994	0.003	0.976	0.010
4	0.992	0.003	0.960	0.021
5	0.991	0.004	0.959	0.032
6	0.990	0.007	0.941	0.042
7	0.989	0.007	0.935	0.037
8	0.986	0.007	0.815	0.197
9	0.978	0.012	0.803	0.202
10	0.965	0.012	0.801	0.187

Ranked clusters were ordered from highest to lowest mean Jaccard similarity within each analysis. Because cluster labels are arbitrary, ranks rather than original cluster numbers were used to facilitate comparison between solutions with and without feature selection.

All clusters in the overall group and the overweight subgroup showed high stability, indicating that the identified structure is robust and not driven by random variation. In the obese subgroup ($n = 2,134$), 9 out of 10 clusters showed (high) stability, while 1 cluster could be characterized as indicating patterns in the data. In the morbid obesity subgroup ($n = 181$), bootstrap-based stability values indicated that most clusters were (highly) stable, while 2 clusters showed patterns with uncertain membership. Given the limited sample size, this level of stability was considered acceptable.

To evaluate the impact of SES feature selection on cluster robustness, we compared bootstrap-based stability estimates from clustering solutions with and without prior feature selection (**Table 4.4**). Because cluster labels are arbitrary and do not necessarily correspond across clustering solutions, clusters were ranked according to their mean Jaccard similarity value before comparison. Across all ranked clusters, feature selection consistently resulted in higher means and lower standard deviations. The mean values ranged from 0.965 to 0.996 with feature selection, compared to 0.801 to 0.986 without feature selection. The largest improvements were observed among the lower-ranked clusters, where stability increased substantially and variability decreased markedly. These findings suggest that the feature-selection step contributes to a more stable and reproducible cluster structure by reducing the influence of redundant and uninformative variables.

4.1.4 Visualization and description of clusters

Based on the 47 selected variables, we described respondent characteristics of 10 clusters in the total group, 10 clusters in the overweight group, 10 clusters in the obese group and 9 clusters in the morbid obesity group (**Table 4.2**). To describe these clusters and the underlying relationships between heterogeneous data, we visualized the data in correlation plots for all consecutive dimension pairs (dimensions 1-2, 2-3, and so forth). The same approach was applied to the level plots and individual plots. See **Appendix G** to examine the plots based on the total group sample.

To examine the association between the identified clusters (A1 through A10) and BMI category, we computed Pearson residuals for each cell of the two-way contingency table. These residuals compare the observed cell counts with the expected counts Agresti 2002. Positive residuals indicate that the observed count exceeds the expected count (overrepresentation), whereas negative residuals indicate underrepresentation. In our visualization (**Fig. 4.1**), overrepresented cells are shown as darker blue and underrepresented cells as darker red, allowing the pattern of deviations to be interpreted at a glance.

We identified ten BMI-related clusters in the total group sample ($N = 15,202$), each with distinct sociodemographic and health profiles. Cluster A1 comprised individuals in the lowest group of prosperity, wealth, and education, with morbid obesity and underweight overrepresented and normal weight underrepresented. Cluster A2 consisted mainly of older adults with low prosperity, poorer dental status, higher diabetes prevalence, and a history of heart attack, with overweight/obesity overrepresented. Cluster A3 represented the youngest group and was characterized by higher smoking rates, greater drug use, and more frequent sexually transmitted infection (STI) testing. Normal-weight individuals were overrepresented, with a smaller overrepresentation of underweight individuals. Informal caregivers were relatively uncommon. Cluster A4 represented the highest-prosperity households, those with children aged 18 or older, married couples, or single-parent households, with more never-smokers and individuals of non-European origin; normal weight was overrepresented and obesity underrepresented. Cluster A5 included residents of highly urban areas, with relatively few informal caregivers and an overrepresentation of normal weight, but to a lesser extent than in cluster A4. Cluster A6 exhibited the poorest physical health and relatively poor mental health, multiple chronic conditions, and medication use, with obesity and morbid obesity clearly overrepresented; relatively few informal caregivers. Cluster A7 encompassed lower prosperity single or single-parent households, more individuals of non-Dutch origin, and a mixed weight profile slightly favoring obesity; education levels were also lower. Cluster A8 comprised older, high-prosperity former smokers, with a higher prevalence of prior heart attack and a slight over-representation of normal weight/overweight; relatively few informal caregivers. Cluster A9 reflected middle-class prosperity, with overweight overrepresented and normal weight underrepresented; relatively few informal caregivers. Cluster A10 captured upper-middle prosperity and wealth, with overweight overrepresented and morbid obesity underrepresented. Across clusters, mean BMI ranged from 23.7 (Cluster A3) to 28.5 (Cluster A6), and mean age from 29.2 years (Cluster A3) to 74.0 years (Cluster A2). A structured overview of these cluster profiles is provided in **Table 4.5**. These cluster descriptions are supported by the visualizations presented in **Appendix G**, which provide additional insight into the observed BMI-related patterns.

Appendix H contains **Tables H.1** through **H.3**, which describe the clustering results for the BMI categories overweight, obesity, and morbid obesity separately, as presented to the expert panel. Association plots were not generated for these subgroup analyses because the variable "BMI category" was no longer present within the analyzed subsets. Including such plots would not be meaningful in this context, given the absence of variation in the BMI category variable.

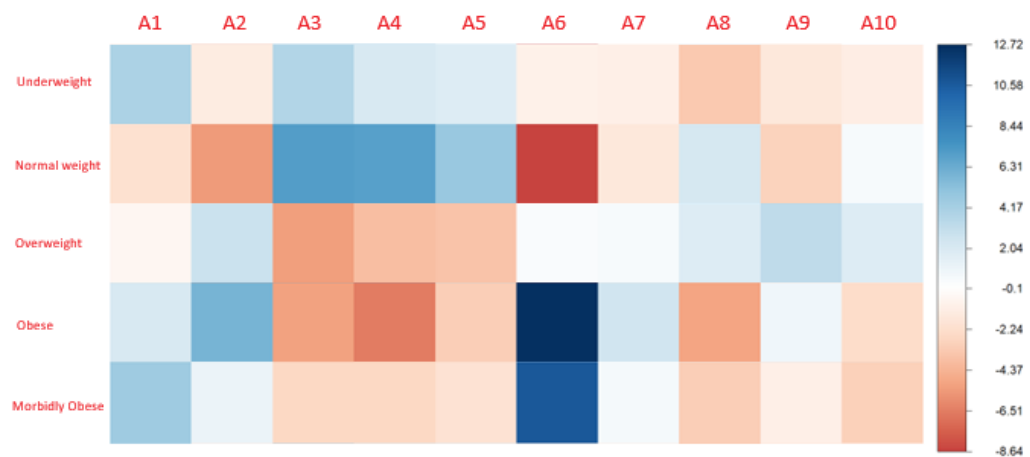
Table 4.5 Description of the Respondent Clusters for the Total Group ($N = 15,202$)

	Description clusters by variables related to BMI	Main characteristics
Cluster A1	Overrepresented in terms of lowest prosperity (lowest 20%) Individuals with only primary education were overrepresented Tend to be, mainly, in the 2nd decile in terms of wealth Morbid obesity and underweight are overrepresented Normal weight is underrepresented	N=1180; Mean BMI: 26.24 (SD 5.6); Mean Age: 46.5 (SD 18.7); Good or very good perceived health: 52.2%; Male: 42.5%
Cluster A2	Tend to have a fewer number of teeth Tend to have more frequently dentures Have more frequently no contact with a dentist Overrepresented in the household position: single and over 65 years old Overrepresented in household position: with partner, without children, over 65 years old Overrepresented with respect to education level: only primary education and vocational training level 1 (MBO1) Have more frequently reported heart attack Have more frequently had a Flu vaccination Overrepresented in the Prosperity category: 2nd decile (almost lowest category) Suffer more frequently from Type 1 diabetes And suffer more frequently from Type 2 diabetes Overweight and obesity are overrepresented	N=1244; Mean BMI: 26.91 (SD 4.9); Mean Age: 74.0 (SD 10.6); Good or very good perceived health: 56.4%; Male: 48.7%
Cluster A3	Slightly lower SF-12 mental health score Tested more frequently for sexually transmitted infection (STI) Tend to use more frequently drugs, especially ecstasy Overrepresented in household position: Couples without children under 40 years old and singles under 40 years old Are more frequently smokers Spread across all prosperity levels Normal weight is overrepresented Underweight is overrepresented to a lesser extent Very few people in this group provide informal care (care-giving)	N=574; Mean BMI: 23.70 (SD 3.6); Mean Age: 29.2 (SD 9.3); Good or very good perceived health: 82.6%; Male: 62.2%
Cluster A4	Overrepresented in the highest prosperity (> 80%) class Overrepresented are children aged 18 or older in a household with a married couple or in a single-parent household more frequently answer yes to the question have you never smoked More frequently had a migration background from Africa, Asia, America, Oceania Normal weight is overrepresented Obesity is clearly underrepresented	N=2238; Mean BMI: 24.70 (SD 4.0); Mean Age: 49.2 (SD 16.5); Good or very good perceived health: 86.9%; Male: 47.9%
Cluster A5	Live more frequently in a very strongly urban area Live in an area with more museums, more restaurants, more music venues, more cinemas nearby more frequently located in the provinces of North Holland and South Holland Less often located in the provinces of Limburg, Gelderland, North Brabant, Flevoland, Zeeland, Drenthe, Friesland Relatively few informal caregivers Normal weight is overrepresented, but to a lesser extent than in cluster A4	N=869; Mean BMI: 24.67 (SD 4.1); Mean Age: 47.5 (SD 17.8); Good or very good perceived health: 78.4%; Male: 49.3%

Table 4.5 Description of the Respondent Clusters for the Total Group (continued)

	Description clusters by variables related to BMI	Main characteristics
Cluster A6	Tend to have a low score on SF-12 physical health Also tend to have a low score on SF-12 mental health, though to a lesser extent Suffer more frequently from the following chronic conditions: joint inflammation, bronchitis, back disorders, urinary incontinence, perceived health, joint wear, asthma, type 1 and type 2 diabetes, diarrhea, and high blood pressure To a lesser extent overrepresented in terms of household position: single individuals over 65 years old, so not only elderly people Take more frequently prescribed medication Consists of relatively few informal caregivers Obesity and morbid obesity are clearly overrepresented	N=1037; Mean BMI: 28.5 (SD 6.0); Mean Age: 63.5 (SD 14.1); Good or very good perceived health: 7.0%; Male: 35.3%
Cluster A7	Are overrepresented in the lower prosperity classes: 1st, 2nd, 3rd decile, and to a lesser extent the 4th decile Individuals with only primary education were overrepresented Overrepresented in household positions: single individuals aged under 40, between 40–65 years, single-parent household with child aged 18 or older, and single-parent household more frequently have the origin: Africa, Asia, America, Oceania, Suriname, Dutch Caribbean (slightly more represented) In terms of weight categories there is a mixed profile, with a slight overrepresentation of obesity	N=1619; Mean BMI: 26.1 (SD 4.8); Mean Age: 49.1 (SD 18.7); Good or very good perceived health: 76.0%; Male: 45.1%
Cluster A8	Overrepresented in the highest prosperity (> 90%) class Are more frequently former smokers Have suffered more frequently from a heart attack Overrepresented in the household position: partner in a couple without children, aged 65 or older Relatively few informal caregivers Normal weight and overweight are slightly more represented Obesity, morbid obesity, and underweight are underrepresented	N=1251; Mean BMI: 25.3 (SD 3.5); Mean Age: 60 (SD 13.6); Good or very good perceived health: 83.2%; Male: 54.2%
Cluster A9	Are more frequently middle class in terms of prosperity: deciles 5 and 6 Relatively few informal caregivers Overweight is overrepresented Normal weight is underrepresented	N=2318; Mean BMI: 26.0 (SD 4.4); Mean Age: 48.7 (SD 17.5); Good or very good perceived health: 83.9%; Male: 49.4%
Cluster A10	Overrepresented in the prosperity and wealth class between middle class and upper class (7th and 8th decile). Morbid obesity is underrepresented Overweight is overrepresented	N=2872; Mean BMI: 25.6 (SD 4.1); Mean Age: 49.5 (SD 17.3); Good or very good perceived health: 85.7%; Male: 50.3%

Figure 4.1 Pearson Residual plot of the association between BMI category, on the vertical axis, and clusters on the horizontal axis



4.2 Results Expert Panel

Overall, ten experts completed the online survey, and seven of these experts participated in two expert panel group discussions. All experts were researchers specializing in obesity, lifestyle epidemiology, and metabolic health.

The experts reported in the online survey whether they recognized an association between the selected variables and BMI. Three of the 47 selected variables were recognized by all ten experts as variables associated with BMI (education, self-perceived health, and comply with physical activity guidelines). Fourteen additional variables were recognized as BMI-related variables by eight or nine experts: SF-12 health measure (physical health), SF-12 health measure (psychological health), chronic arthritis, arthrosis hips/knees, diabetes type 2, myocardial infarction, prosperity, wealth, former smoker, current smoker, sufficient fruit during 7 days per week, and three variables on adherence to the physical activity guidelines (weekly sporter, sufficient muscle-/bone-strength activities, and sufficient balance activities). Five of the 47 variables were not recognized as BMI-related by any of the experts: Diarrhea, tested for STI and three variables related to the proximity-to-facilities data (number of cinemas within a 10 km radius, number of performing arts venues within a 5 km radius, and number of museums within a 20 km radius). All other variables were recognized to varying degrees as BMI-related. These findings demonstrated areas of agreement alongside differences between data-driven variable selection and expert recognition of BMI-related factors.

The main themes discussed during the expert discussion are summarized in **Table 4.6** for BMI category overweight, **Table 4.7** for BMI category obese, and **Table 4.8** for BMI category morbid obesity. During the expert panel discussions, there was broad agreement among the experts on the clinical relevance of the described clusters. The experts appreciated the clustering approach and noted that its value could be further enhanced through the inclusion of additional contextual information, thereby improving the interpretability of the identified clusters. They emphasized that the inclusion of additional variables would enable even deeper and more detailed interpretations of the clusters. The experts also recognized that the variable selection method was effective, and they suggested that expanding it could reveal even more practically meaningful insights. Differences arose in the discussion regarding the possible merging of clusters within each BMI category. The experts made suggestions regarding subgroups that may

Table 4.6 Main themes in expert discussion regarding BMI category overweight

Overweight	Description of the clusters by the experts	Illustrative expert quotes on interpretation and context
Cluster O1	Includes older individuals living alone or with a partner and experiencing poor oral health.	<i>"Possibly difficult to reach." "This poses a risk of underweight in later life due to reduced energy intake caused by chewing difficulties or pain when eating. Loneliness may exacerbate this. In clinical terms, this is a group that is at risk of underweight"</i>
Cluster O2	Remarkably low self-rated physical and mental health and a high prevalence of chronic conditions.	<i>"A high need for health care."</i>
Clusters O3, O4, and potentially O5 and O8*	High levels of wealth and prosperity, high self-rated health, and comparable age ranges.	Cluster O3 and O4: <i>"Might be highly educated." "Might be willing to pay for treatments." "May be able to manage with digital guidance and take control themselves."</i>
Clusters O6 and O7	Similar to clusters O3, O4, O5, and O8, but include individuals with a migration background.	<i>"Culture-adapted intervention indicated."</i>
Cluster O9 (potentially merged with O1)	Very low financial assets.	<i>"Very low prosperity, so in terms of approach, first look at what else is causing stress rather than immediately addressing the issue of being overweight." "Health problems that may be related to low prosperity."</i>
Cluster O10	Younger individuals (< 40 years) with mental health concerns, drug use, and more frequently tested for STIs.	<i>"...apparently still in good health, but possibly a risk group." "But I do think it's a group that needs attention in the near future."</i>

* Different opinions arose regarding which clusters could be merged and in what configuration; italicized text represents expert quotations illustrating interpretation and context.

be at increased risk of (additional) health complications, as well as considerations for tailored intervention strategies.

Building on the findings for the overweight category, the analysis of the obesity category (**Table 4.7**) and morbid obesity category (**Table 4.8**) revealed several overlapping cluster profiles. (Quote: *"There are highly similar clusters as observed for overweight. Therefore, I would propose the same approach here. It could be considered to combine the categories of overweight and obesity. This is also commonly done in healthcare practice, and individuals with overweight and cardiometabolic risk are eligible for combined lifestyle interventions."*)

Experts agreed that the clusters (M1 through M9) identified within the morbid obesity category (**Table 4.8**) were difficult to interpret because of the limited sample size in this category ($n = 181$), which resulted in relatively small numbers of respondents within the individual clusters. In general, the experts found it remarkable that some individuals within the morbid obesity clusters met the physical activity guidelines (clusters M5 and M8).

Table 4.7 Main themes in expert discussion regarding BMI category obese

Obese	Description of the clusters by the experts	Illustrative expert quotes
Clusters B1, B4, and B8	Relatively older individuals with poor oral health and one or more chronic conditions. Cluster B1 shows very low self-rated health compared with clusters B4 and B8.	<i>"Oral health stands out."</i>
Cluster B3	Younger individuals (< 40 years) with risk behaviours comparable to those observed in overweight cluster O10, including drug use and tested for sexually transmitted infections.	<i>"Cluster B3 appears to be young again with risky behavior, compared to cluster O10 in the overweight category. Perhaps overarching category."</i>
Clusters B6, B7, and B10*	Relatively prosperous, middle-aged individuals (approximately 50–55 years) with a balanced gender distribution.	<i>"Merge clusters B6, B7, and B10: prosperous to very prosperous, middle-aged, similar male–female ratio."</i>

*Different opinions arose regarding which clusters could be merged and in what configuration; italicized text represents expert quotations illustrating interpretation and context.

Table 4.8 Main themes in expert discussion regarding BMI category morbid obesity

morbid obesity	Description of the clusters by the experts	Illustrative expert quotes
Clusters M4 and M9	Poor oral health.	<i>"Dental problems caused by chronic inflammation."</i>
Clusters M3, M5, and M8*	Physically active and relatively wealthy individuals. Cluster M8 consists of older individuals with very low self-rated health.	<i>"Characterized by compliance with the exercise guidelines; interventions could focus on components other than exercise." "I cannot interpret cluster M8: compliance with exercise guidelines, high socioeconomic status, but very low perceived health." "Wealth is a dominant recurring variable."</i>
Cluster M6	Very poor self-rated health, low physical functioning, and multiple chronic conditions.	<i>"Diabetes: what is cause and what is effect here?"</i>

*Different opinions arose regarding which clusters could be merged and in what configuration; italicized text represents expert quotations illustrating interpretation and context.

Based on the expert panel discussions, four overarching clusters were identified: (1) individuals who use drugs and are more frequently tested for STI, (2) individuals with chronic health conditions and inflammatory diseases, (3) individuals with lower prosperity and wealth, and (4) individuals with a migration background. These themes corresponded to characteristics observed in multiple clusters within the quantitative analysis. The experts noted that the identified clusters were not truly specific to BMI categories or instead maybe reflect broader population-level patterns (*Quote “You may want to know how unique this clustering is, or whether it is not a clustering of people, ... because we now also see such overlap between the clustering of overweight people and the clustering of obese people”*).

5 Discussion

The primary contribution of this study was the evaluation of an integrated healthcare informatics workflow that combines SES variable selection, PCAMIX dimensionality reduction, k-means clustering, and expert interpretation to identify BMI-related subgroup structures in heterogeneous population health data (N = 15,202), enriched with multiple complementary data sources.

A key feature of the strategy used is the Statistically Equivalent Signature (SES) step, that isolates variables that contribute unique information to the target variable (BMI), while excluding redundant variables. We consider this to be important because including variables that convey redundant information does not necessarily add information; instead, it may deteriorate the estimation of principal components. Our findings further suggest that this variable selection step enhances the interpretability and stability of the clusters. In addition, the bootstrap-based stability analysis demonstrated consistently higher mean Jaccard similarity values for the clustering solution obtained after SES-based variable selection than for the clustering solution without prior variable selection. While we acknowledge the potential concern that domain-relevant variables might be unintentionally omitted, the excluded variables did not provide additional information beyond that already contained in the retained set. This approach reduces the likelihood that variables disproportionately influence the principal component structure in PCAMIX, when many variables reflect a similar underlying concept. Exploratory analyses conducted without prior variable selection indicated that concepts represented by many variables tended to dominate the principal components. These concepts were often reflected across multiple dimensions, whereas concepts represented by fewer variables were less prominent. As a result, relevant patterns may become less visible, not due to lack of importance, but due to imbalance in variable representation. For example, from the large number of variables added from the Proximity to Facilities data, only a few were retained by the SES algorithm, suggesting that much of the information contained in these variables can be represented by a smaller subset of variables. The same applies to the GCN air-pollution data, for which none of the available variables were retained by the SES algorithm. This finding should not be interpreted as evidence that environmental factors are unimportant for understanding BMI-related patterns. Rather, it indicates that, within the present dataset, these variables did not provide substantial unique information with respect to BMI. It is also possible that environmental variables are associated with BMI through more dominant socioeconomic and behavioral factors. Therefore, it may be worthwhile to explore the effects of omitting such dominant variables in future applications of this analytical approach, particularly when the goal is to uncover more subtle or context-specific patterns. Additionally, to gain deeper insight into how these relationships evolve over time, future research could focus on the analysis of longitudinal data. This would facilitate the investigation of potential pathways and the directionality of these associations.

The analysis of heterogeneous complex datasets depends on how variables are selected and represented. Studies in medical image analysis have highlighted the impact of data complexity on model performance Kujur et al. 2022 and demonstrated that improved feature representations can substantially enhance downstream analytical performance Aung and Khan 2025. Although these studies were conducted in different application domains, they underscore the broader importance of selecting informative variables when analyzing complex datasets.

The coexistence of modest silhouette coefficients and high bootstrap-based cluster stability suggests that the identified subgroup structures are stable and show limited separation in the multidimensional BMI-related variable space. Rather than representing sharply separated classes across the entire multidimensional principal component space, the clusters appear to reflect stable and reproducible population profiles.

Using our variable selection technique, 47 of the 285 variables included in the highly diverse dataset were found to be related to BMI. Experts with scientific knowledge and experience in various healthcare fields recognized correlations between 42 of these 47 variables and BMI to varying degrees. Diarrhea, Sexual Transmitted Infection (STI) testing, and three environmental variables were not considered to be related to BMI by the experts, although these variables emerged as distinguishing characteristics in several clusters. Which is in line with previous literature that reported associations between obesity and diarrhea Han et al. 2025; Linghu 2022. Furthermore, Leal and Chaix 2011 found that residing in urbanized neighborhoods or areas with higher residential density in developed countries was associated with lower body weight. Other reviews also reported that less urbanized environments can be positively associated with obesity Feng et al. 2010; Mackenbach et al. 2014. In our study urbanization is reflected not only by the urbanization variable, particularly at the level “very highly urbanized”, but also by proximity-to-facility variables, such as the number of performing arts venues within 5 km, cinemas within 10 km, restaurants within 5 km, and museums within 20 km. These variables collectively characterize cluster A5 as an environment with a dense concentration of facilities. In addition, we observed that clusters with higher levels of prosperity and wealth show an overrepresentation of individuals with normal weight, while lower prosperity levels correspond to a slight overrepresentation of obese individuals, particularly in cluster A7. These findings are consistent with those in Wolfe et al. 2019. Interestingly, for the lowest prosperity deciles, we found not only a slight overrepresentation of individuals with morbid obesity but also of underweight individuals in cluster A1. Prosperity has already been considered a key indicator of socioeconomic position (SEP) and a known correlate of both health and obesity Hajat et al. 2010; Pool et al. 2018; Connolly et al. 2010. Wealth and prosperity, or more broadly, SEP, emerge as common denominators in our earlier research Beuken et al. 2022; Beuken et al. 2025, alongside educational level. These findings suggest that differences in socioeconomic position, prosperity, and education may play a role in shaping health outcomes and lifestyle-related behaviors.

Overall, experts advised that the findings should be interpreted with caution. The clusters were not yet considered ready for direct clinical use but were seen as potentially meaningful for future research and for identifying at-risk populations, in the sense of a higher risk of developing obesity. One example was a cluster of younger individuals characterized by mental health problems, drug use, and more frequent STI testing. Experts considered this a clearly identifiable high-risk group that may benefit from tailored preventive interventions. However, the observed cluster characteristics may partly reflect underlying factors that were not explicitly disentangled within the current analysis. Therefore, further research is needed to better understand the mechanisms underlying these cluster patterns before they can inform targeted interventions.

Despite these considerations, our approach helps identify how variables related to lifestyle, education, and health co-occur within BMI-related patterns. As a result, variables that receive

relatively little attention in obesity research, such as dental health, may emerge as distinguishing characteristics worthy of further investigation. In this methodologically oriented study, we chose to describe the clusters based solely on the selected variables. This can be considered a trade-off between maintaining a strict analytical approach and incorporating domain knowledge to enhance clinical relevance, as the clusters were not yet suitable for direct clinical use. Future research could therefore focus on exploring these clusters in relation to additional variables, potentially offering deeper insights into the underlying dynamics and enhancing both theoretical understanding and practical application.

Limitations

Data preparation was necessary to enable clustering analysis. For example, some variables contained indistinguishable answer options, resulting in significant overlap that limited their ability to differentiate between clusters. Consequently, these variables were excluded from the dataset.

As part of our methodology, we applied a variable selection technique to identify variables most informative with respect to BMI, with the aim of reducing noise from statistically uninformative variables that could impair generalization and distort dimensionality reduction and clustering outcomes. However, due to the trade-off between a strict analytical approach and ensuring clinical relevance, a potential limitation of this approach is the exclusion of variables that, while not informative from a statistical standpoint, may hold value from a domain-specific perspective. During variable selection, age was intentionally omitted. This decision was made to reduce the likelihood that age would dominate the variable-selection process and limit the selection of other potentially relevant variables. However, a number of selected variables may be related to age. Given the complex relationships among the assessed variables, some cluster characteristics may therefore partly reflect age-related patterns. Further research is needed to better disentangle the contribution of age from that of the other variables included in the analysis.

From a domain-specific perspective, a broader range of nutrition-related variables could offer valuable information on BMI and related health outcomes. However, such a broad range was not included in the current dataset, limiting our ability to assess their potential contribution. In 2023, a wider range of nutrition-related data was added to the DHS data, but we merged two years of data, 2022 and 2023, to obtain a sufficient number of respondents in our study. Including these variables would have introduced substantial missingness because they were only available in the 2023 survey wave. Therefore, these variables were excluded from the analysis. Future studies could address this gap by incorporating detailed nutritional information, which may enrich the analysis and provide a more comprehensive understanding of the factors influencing BMI.

The expert panel was relatively small and selected in a specific manner, which may limit its representativeness for the broader domain-specific knowledge.

As the analysis was partially based on self-reported survey data, some variables may be subject to reporting bias, recall bias, or social desirability bias.

Due to the cross-sectional nature of the data, no conclusions can be drawn regarding the directionality or causality of the observed associations.

Although the DHS is designed to be nationally representative of the Dutch population, the generalizability of these findings to other populations, countries, time periods, or datasets may be limited. In addition, the identified clusters are dependent on the variables available in the current dataset and the analytical choices made throughout the study.

The imputation process inherently introduces some uncertainty. However, imputation was restricted to variables with a maximum missing rate of 20%, reflecting a conservative strategy to limit its impact. In addition, comparison of variable distributions before and after imputation indicated that the imputation procedure did not materially alter the underlying data structure.

The average silhouette coefficients were relatively low across all BMI groups, indicating modest separation between clusters. Furthermore, some clusters, particularly within the morbid obesity subgroup, were based on relatively small sample sizes. Although bootstrap-based stability analyses indicated acceptable stability for most clusters, smaller cluster sizes may limit robustness, interpretability, and generalizability, and increase uncertainty in cluster membership. Consequently, the findings for the morbid obesity subgroup should be interpreted with caution and regarded as exploratory in nature.

6 Conclusion

This study demonstrates the value of a methodologically guided, healthcare informatics-driven analytical approach that combines SES variable selection, PCAMIX dimensionality reduction, k-means clustering, and expert interpretation to extract meaningful insights from heterogeneous population health data. The findings suggest that variable selection improves the interpretability and stability of the resulting clustering solution. Furthermore, the combination of modest silhouette coefficients and high bootstrap-based cluster stability indicates that the proposed workflow identifies stable subgroup structures, even when clusters are not sharply separated in the multidimensional variable space.

By applying a statistically guided variable-selection algorithm focused on BMI, we isolated 47 informative variables out of an initial set of 285 from a diverse dataset of over 15,000 respondents. The resulting clusters support the view that BMI is not merely an individual lifestyle outcome but may be embedded within multifaceted constellations of socioeconomic, environmental, mental, and health-related factors. In addition, some less obvious determinants, such as STI testing and dental health, also emerged as potential differentiators, highlighting the complex and systemic nature of patterns across BMI-related indicators. Taken together, these findings support the value of the proposed workflow as an exploratory population-health analytics strategy for identifying stable BMI-related population profiles. Although the identified clusters are not yet ready for clinical application, they highlight potentially meaningful subgroup patterns that warrant further investigation and may ultimately contribute to more personalized weight-management and preventive-care strategies, particularly for high-risk populations. Longitudinal studies are recommended to explore how these patterns evolve over time.

Acknowledgments

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands.

The authors gratefully acknowledge the expert panel members for their contributions to the interpretation of the study findings and for sharing their expertise and recommendations. Special thanks are extended to Jan van der Laan for careful reading of the manuscript and Jan-Willem Bruggink for his valuable insights into the DHS dataset, which were essential for developing a comprehensive understanding of the data.

References

- Agresti, A. (2002). *Categorical Data Analysis (Second Edition)*. Hoboken, New Jersey: John Wiley & Sons, p. 81. DOI: 10.1002/0471249688.
- Aliferis, C. F., A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos (2010). "Local causal and Markov blanket induction for causal discovery and feature selection." In: *Journal of Machine Learning Research* 11.1.
- Aljohani, A. (2024). "Optimizing patient stratification in healthcare: a comparative analysis of clustering algorithms for EHR data." In: *International Journal of Computational Intelligence Systems* 17.1, p. 173. DOI: 10.1007/s44196-024-00568-8.
- Aung, T. M. M. and A. A. Khan (2025). "Enhanced U-Net with Attention Mechanisms for Improved Feature Representation in Lung Nodule Segmentation." In: *Current Medical Imaging* 21. ISSN: 1573-4056. DOI: <https://doi.org/10.2174/0115734056386382250902064757>. URL: <https://www.sciencedirect.com/science/article/pii/S1573405625001092>.
- Bai, J. and S. Ng (2008). "Forecasting economic time series using targeted predictors." In: *Journal of Econometrics* 146.2. Honoring the research contributions of Charles R. Nelson, pp. 304–317. ISSN: 0304-4076. DOI: <https://doi.org/10.1016/j.jeconom.2008.08.010>.
- Benke, K. and G. Benke (2018). "Artificial Intelligence and Big Data in Public Health." In: *International Journal of Environmental Research and Public Health* 15.12. ISSN: 1660-4601. DOI: 10.3390/ijerph15122796. URL: <https://www.mdpi.com/1660-4601/15/12/2796>.
- Berkel, K. van, S. van der Doef, and B. Schouten (2020). "Implementing adaptive survey design with an application to the Dutch Health Survey." In: *Journal of Official Statistics* 36.3, pp. 609–629. DOI: 10.2478/jos-2020-0031.
- Beuken, M. J. M., I. M. Kanera, N. P. M. Ezendam, S. Braun, and M. Zoet (2022). "Identification and Potential Use of Clusters of Patients With Colorectal Cancer and Patients With Prostate Cancer in Clinical Practice: Explorative Mixed Methods Study." In: *JMIR Cancer* 8.4, e42908. ISSN: 2369-1999. DOI: 10.2196/42908. URL: <https://cancer.jmir.org/2022/4/e42908>.
- Beuken, M. J. M. et al. (2025). "Identification of Clusters in a Population With Obesity Using Machine Learning: Secondary Analysis of The Maastricht Study." In: *JMIR Med Inform* 13, e64479. ISSN: 2291-9694. DOI: 10.2196/64479. URL: <https://medinform.jmir.org/2025/1/e64479>.
- Biau, G. and E. Scornet (2016). "A random forest guided tour." In: *TEST* 25.2, pp. 197–227. DOI: 10.1007/s11749-016-0481-7.
- Biza, K., I. Tsamardinos, V. Lagani, G. Athineou, M. Tsagris, G. Borboudakis, et al. (2025). *MXM: feature selection (including multiple solutions) and Bayesian networks*. URL: <https://github.com/mensxmachina/MXM-R-Package>.
- Blüher, M. (2019). "Obesity: global epidemiology and pathogenesis." In: *Nature Reviews Endocrinology* 15.5, pp. 288–298. DOI: 10.1038/s41574-019-0176-8.
- Buuren, S. van and K. Groothuis-Oudshoorn (2011). "mice: Multivariate Imputation by Chained Equations in R." In: *Journal of Statistical Software* 45.3, pp. 1–67. DOI: 10.18637/jss.v045.i03. URL: <https://www.jstatsoft.org/index.php/jss/article/view/v045i03>.
- Chavent, M., V. Kuentz-Simonet, A. Labenne, and J. Saracco (2022). *Multivariate Analysis of Mixed Data: The R Package PCAmixdata*. arXiv: 1411.4911 [stat.CO]. URL: <https://arxiv.org/abs/1411.4911>.
- Chavent, M., V. Kuentz-Simonet, and J. Saracco (2012). "Orthogonal rotation in PCAMIX." In: *Advances in Data Analysis and Classification* 6.2, pp. 131–146. DOI: 10.1007/s11634-012-0105-3.

- Clavel, J., G. Merceron, and G. Escarguel (2014). "Missing Data Estimation in Morphometrics: How Much is Too Much?" In: *Systematic Biology* 63.2, pp. 203–218. ISSN: 1063-5157. DOI: 10.1093/sysbio/syt100.
- Connolly, S., D. O'Reilly, and M. Rosato (2010). "House value as an indicator of cumulative wealth is strongly related to morbidity and mortality risk in older people: a census-based cross-sectional and longitudinal study." In: *International Journal of Epidemiology* 39.2, pp. 383–391. ISSN: 0300-5771. DOI: 10.1093/ije/dyp356.
- Doove, L., S. Van Buuren, and E. Dusseldorp (2014). "Recursive partitioning for missing data imputation in the presence of interaction effects." In: *Computational Statistics & Data Analysis* 72, pp. 92–104. ISSN: 0167-9473. DOI: <https://doi.org/10.1016/j.csda.2013.10.025>.
- Elbattah, M. and O. Molloy (2017). "Data-driven patient segmentation using K-means clustering: the case of hip fracture care in Ireland." In: *Proceedings of the Australasian Computer Science Week Multiconference*. ACSW '17. Geelong, Australia: Association for Computing Machinery. ISBN: 9781450347686. DOI: 10.1145/3014812.3014874.
- Feng, J., T. A. Glass, F. C. Curriero, W. F. Stewart, and B. S. Schwartz (2010). "The built environment and obesity: A systematic review of the epidemiologic evidence." In: *Health & Place* 16.2, pp. 175–190. ISSN: 1353-8292. DOI: <https://doi.org/10.1016/j.healthplace.2009.09.008>.
- Feng, S., C. Hategeka, and K. A. Grépin (2021). "Addressing missing values in routine health information system data: an evaluation of imputation methods using data from the Democratic Republic of the Congo during the COVID-19 pandemic." In: *Population Health Metrics* 19.1, p. 44. DOI: 10.1186/s12963-021-00274-z.
- Fetters, M. D., L. A. Curry, and J. W. Creswell (2013). "Achieving Integration in Mixed Methods Designs—Principles and Practices." In: *Health Services Research* 48.6pt2, pp. 2134–2156. DOI: <https://doi.org/10.1111/1475-6773.12117>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-6773.12117>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-6773.12117>.
- Hajat, A., J. Kaufman, K. Rose, A. Siddiqi, and J. Thomas (2010). "Do the wealthy have a health advantage? Cardiovascular disease risk factors and wealth." In: *Social Science & Medicine* 71.11, pp. 1935–1942. ISSN: 0277-9536. DOI: <https://doi.org/10.1016/j.socscimed.2010.09.027>.
- Halcomb, E. J. and P. M. Davidson (2006). "Is verbatim transcription of interview data always necessary?" In: *Applied Nursing Research* 19.1, pp. 38–42. DOI: 10.1016/j.apnr.2005.06.001.
- Han, K., X. Wang, Y. Wang, X. Niu, J. Xiang, N. Ru, et al. (2025). "Prevalence of chronic diarrhea and its association with obesity in a Chinese community-based population." In: *Chinese Medical Journal* 138.13. DOI: 10.1097/CM9.0000000000003190.
- Hennig, C. (2007). "Cluster-wise assessment of cluster stability." In: *Computational Statistics & Data Analysis* 52.1, pp. 258–271. ISSN: 0167-9473. DOI: <https://doi.org/10.1016/j.csda.2006.11.025>. URL: <https://www.sciencedirect.com/science/article/pii/S0167947306004622>.
- Hennig, C. (2026). *fpc: Flexible Procedures for Clustering*. Available at <https://CRAN.R-project.org/package=fpc>. R package version 2.2-14. URL: <https://CRAN.R-project.org/package=fpc>.
- Jansson, N. F., R. L. Allen, G. Skogsmo, and S. Tavakoli (2022). "Principal component analysis and K-means clustering as tools during exploration for Zn skarn deposits and industrial carbonates, Sala area, Sweden." In: *Journal of Geochemical Exploration* 233, p. 106909. ISSN: 0375-6742. DOI: <https://doi.org/10.1016/j.gexplo.2021.106909>. URL: <https://www.sciencedirect.com/science/article/pii/S0375674221001886>.
- Junaaid, K. P., T. Kiran, M. Gupta, K. Kishore, and S. Siwatch (2025). "How much missing data is too much to impute for longitudinal health indicators? A preliminary guideline for the choice of

- the extent of missing proportion to impute with multiple imputation by chained equations." In: *Population Health Metrics* 23.1, p. 2. DOI: 10.1186/s12963-025-00364-2.
- Jünger, S., S. A. Payne, J. Brine, L. Radbruch, and S. G. Brearley (2017). "Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: Recommendations based on a methodological systematic review." In: *Palliative Medicine* 31.8, pp. 684–706. DOI: 10.1177/0269216317690685.
- Kiers, H. A. L. (1991). "Simple Structure in Component Analysis Techniques for Mixtures of Qualitative and Quantitative Variables." In: *Psychometrika* 56.2, pp. 197–212. DOI: 10.1007/BF02294458.
- Kuhn, M. and K. Johnson (2013). *Applied Predictive Modeling*. Springer Series in Statistics. New York: Springer. ISBN: 978-1461468493.
- Kujur, A., Z. Raza, A. A. Khan, and C. Wechtaison (2022). "Data Complexity Based Evaluation of the Model Dependence of Brain MRI Images for Classification of Brain Tumor and Alzheimer's Disease." In: *IEEE Access* 10, pp. 112117–112133. DOI: 10.1109/ACCESS.2022.3216393.
- Lagani, V., G. Athineou, A. Farcomeni, M. Tsagris, and I. Tsamardinos (2017). "Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets." In: *Journal of Statistical Software* 80.7, pp. 1–25. DOI: 10.18637/jss.v080.i07. URL: <https://www.jstatsoft.org/index.php/jss/article/view/v080i07>.
- Leal, C. and B. Chaix (2011). "The influence of geographic life environments on cardiometabolic risk factors: a systematic review, a methodological assessment and a research agenda." In: *Obesity Reviews* 12.3, pp. 217–230. DOI: <https://doi.org/10.1111/j.1467-789X.2010.00726.x>.
- Linghu, E. (2022). "Obesity and chronic diarrhea: a new syndrome?" In: *Chinese Medical Journal* 135.15, pp. 1806–1807. DOI: 10.1097/cm9.0000000000002147.
- Liu, L., J. Gao, G. Beasley, and S.-H. Jung (2023). "LASSO and Elastic Net Tend to Over-Select Features." In: *Mathematics* 11.17. ISSN: 2227-7390. DOI: 10.3390/math11173738. URL: <https://www.mdpi.com/2227-7390/11/17/3738>.
- Lustria, M. L. A., S. M. Noar, J. Cortese, S. K. Van Stee, R. L. Glueckauf, and J. Lee (2013). "A meta-analysis of web-delivered tailored health behavior change interventions." In: *Journal of Health Communication* 18.9, pp. 1039–1069. DOI: 10.1080/10810730.2013.768727.
- Mackenbach, J. D., H. Rutter, S. Compornolle, K. Glonti, J.-M. Oppert, H. Charreire, et al. (2014). "Obesogenic environments: a systematic review of the association between the physical environment and adult weight status, the SPOTLIGHT project." In: *BMC Public Health* 14.1, p. 233. DOI: 10.1186/1471-2458-14-233.
- MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations." In: 1, pp. 281–297. URL: https://digitalassets.lib.berkeley.edu/math/ucb/text/math_s5_v1_article-17.pdf.
- McCabe, G. P. (1984). "Principal Variables." In: *Technometrics* 26.2, pp. 137–144. DOI: 10.1080/00401706.1984.10487939.
- Nagamine, T., B. Gillette, J. Kahoun, R. Burghaus, J. Lippert, and M. Saxena (2022). "Data-driven identification of heart failure disease states and progression pathways using electronic health records." In: *Scientific Reports* 12.1, p. 17871. DOI: 10.1038/s41598-022-22398-4.
- National Institute for Public Health and the Environment (2025a). *Archive of previous GCN rounds*. URL: <https://www.rivm.nl/gcn-gdn-kaarten/concentratiekaarten/downloaden/archief>.
- National Institute for Public Health and the Environment (2025b). *GCN and GDN maps*. URL: <https://www.rivm.nl/gcn-gdn-kaarten>.
- Nicolet, A., D. Assouline, M.-A. Le Pogam, C. Perraudin, C. Bagnoud, J. Wagner, et al. (2022). "Exploring patient multimorbidity and complexity using health insurance claims data: a cluster analysis approach." In: *JMIR Medical Informatics* 10.4, e34274. DOI: 10.2196/34274.

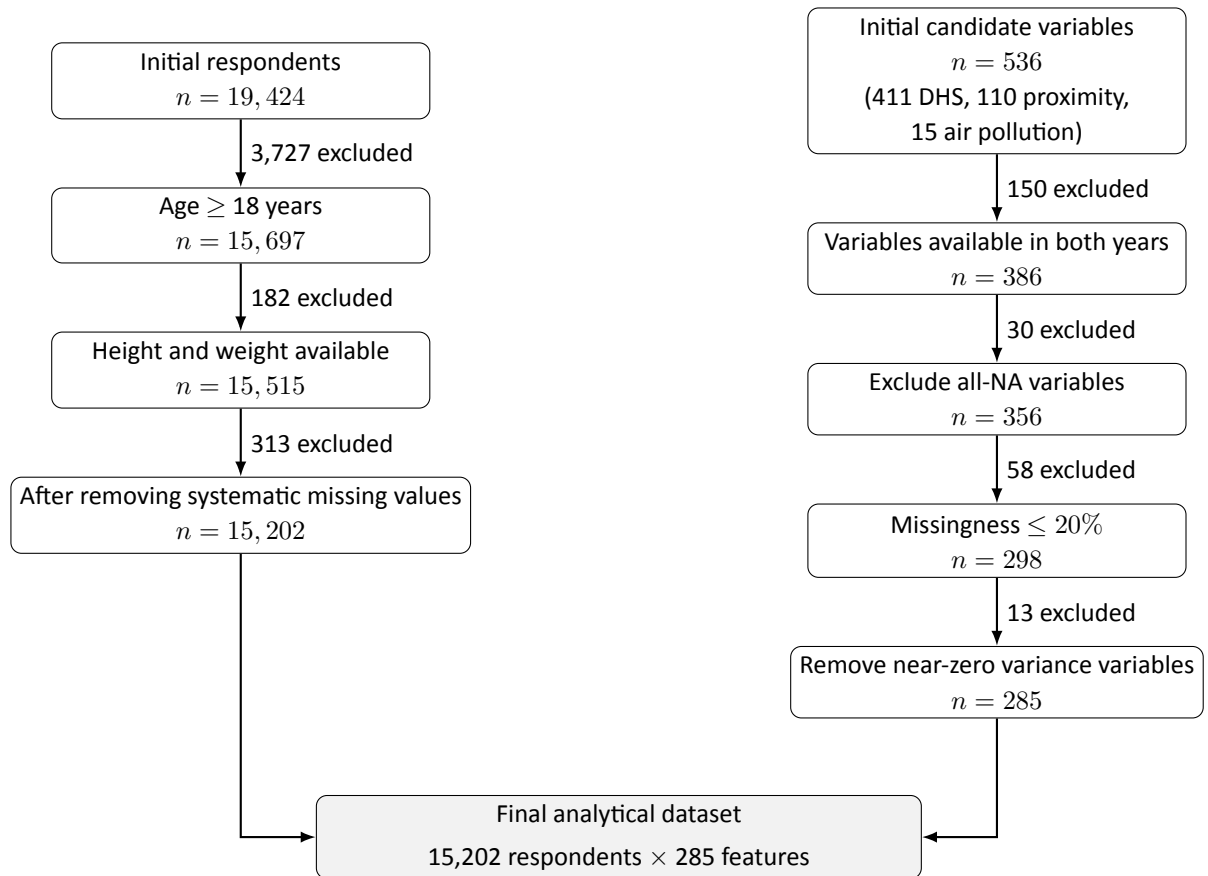
- Noar, S. M., C. N. Benac, and M. S. Harris (2007). "Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions." In: *Psychological Bulletin* 133.4, pp. 673–693. DOI: 10.1037/0033-2909.133.4.673.
- Olsson, D., P. Georgiev, and P. M. Pardalos (2013). "Kernel Principal Component Analysis: Applications, Implementation and Comparison." In: *Models, Algorithms, and Technologies for Network Analysis*. Ed. by B. I. Goldengorin, V. A. Kalyagin, and P. M. Pardalos. New York, NY: Springer New York, pp. 127–148. ISBN: 978-1-4614-8588-9.
- Pacheco, J., S. Casado, and S. Porras (2013). "Exact methods for variable selection in principal component analysis: Guide functions and pre-selection." In: *Computational Statistics & Data Analysis* 57.1, pp. 95–111. ISSN: 0167-9473. DOI: <https://doi.org/10.1016/j.csda.2012.06.014>.
- Pereira, R. C., P. H. Abreu, P. P. Rodrigues, and M. A. Figueiredo (2024). "Imputation of data Missing Not at Random: Artificial generation and benchmark analysis." In: *Expert Systems with Applications* 249, p. 123654. ISSN: 0957-4174. DOI: <https://doi.org/10.1016/j.eswa.2024.123654>.
- Pestov, V. (2013). "Is the k-NN classifier in high dimensions affected by the curse of dimensionality?" In: *Computers & Mathematics with Applications* 65.10. Grasping Complexity, pp. 1427–1437. ISSN: 0898-1221. DOI: <https://doi.org/10.1016/j.camwa.2012.09.011>.
- Peterson, R. A. (2021). "Finding Optimal Normalizing Transformations via bestNormalize." In: *The R Journal* 13 (1). <https://doi.org/10.32614/RJ-2021-041>, pp. 294–313. ISSN: 2073-4859. DOI: 10.32614/RJ-2021-041.
- Peterson, R. A. and J. E. Cavanaugh (2020). "Ordered quantile normalization: a semiparametric transformation built for the cross-validation era." In: *Journal of Applied Statistics* 47.13-15. PMID: 35707424, pp. 2312–2327. DOI: 10.1080/02664763.2019.1630372.
- Pool, L. R., S. A. Burgard, B. L. Needham, M. R. Elliott, K. M. Langa, and C. F. Mendes de Leon (2018). "Association of a Negative Wealth Shock With All-Cause Mortality in Middle-aged and Older Adults in the United States." In: *JAMA* 319.13, pp. 1341–1350. ISSN: 0098-7484. DOI: 10.1001/jama.2018.2055.
- Rousseeuw, P. J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65. ISSN: 0377-0427. DOI: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- Shah, A. D., J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway (2014). "Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study." In: *American Journal of Epidemiology* 179.6, pp. 764–774. ISSN: 0002-9262. DOI: 10.1093/aje/kwt312.
- Statistics Netherlands (2024). *Obesity rate has tripled in the last 40 years*. URL: <https://www.cbs.nl/en-gb/news/2024/10/obesity-rate-has-tripled-in-the-last-40-years>.
- Statistics Netherlands (2025a). *Gezondheidsenquête 2010–2024*. URL: <https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/gezondheidsenquête-2010-2024>.
- Statistics Netherlands (2025b). *Nabijheidsstatistieken per wijk/buurt*. URL: <https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/wijk-en-buurtstatistieken/nabijheidsstatistieken-per-wijk-buurt>.
- Stekhoven, D. J. and P. Bühlmann (2012). "MissForest—non-parametric missing value imputation for mixed-type data." In: *Bioinformatics* 28.1, pp. 112–118. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr597.
- Takeshita, S., Y. Nishioka, Y. Tamaki, F. Kamitani, T. Mohri, H. Nakajima, et al. (2024). "Novel subgroups of obesity and their association with outcomes: a data-driven cluster analysis." In: *BMC Public Health* 24.1, p. 124. DOI: 10.1186/s12889-024-17648-1.

- Tang, F. and H. Ishwaran (2017). "Random forest missing data algorithms." In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 10.6, pp. 363–377. DOI: <https://doi.org/10.1002/sam.11348>.
- Tenny, S., J. M. Brannan, and G. D. Brannan (2022). "Qualitative Study." In: *StatPearls*. Treasure Island, FL: StatPearls Publishing. URL: <https://www.ncbi.nlm.nih.gov/books/NBK470395/>.
- Tsagris, M., V. Lagani, and I. Tsamardinos (2018). "Feature selection for high-dimensional temporal data." In: *BMC Bioinformatics* 19.1, p. 17. DOI: 10.1186/s12859-018-2023-7.
- Tsamardinos, I., V. Lagani, and D. Pappas (2012). "Discovering multiple, equivalent biomarker signatures." In: *Proceedings of the 7th Conference of the Hellenic Society for Computational Biology and Bioinformatics*, pp. 54–56.
- Wei, T. and V. Simko (2024). *R package 'corrplot': Visualization of a Correlation Matrix*. (Version 0.95). URL: <https://github.com/taiyun/corrplot>.
- Wolfe, J. D., E. H. Baker, and I. C. Scarinci (2019). "Wealth and Obesity Among US Adults Entering Midlife." In: *Obesity* 27.12, pp. 2067–2075. DOI: <https://doi.org/10.1002/oby.22625>.
- World Health Organization (2025). *Obesity and overweight*. URL: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>.
- Yamamoto, M. and H. Hwang (2014). "A General Formulation of Cluster Analysis with Dimension Reduction and Subspace Separation." In: *Behaviormetrika* 41.1, pp. 115–129. DOI: 10.2333/bhmk.41.115.
- Zhang, S., S. M. H. Bamakan, Q. Qu, and S. Li (2019). "Learning for Personalized Medicine: A Comprehensive Review From a Deep Learning Perspective." In: *IEEE Reviews in Biomedical Engineering* 12, pp. 194–208. DOI: 10.1109/RBME.2018.2864254.

Appendices

A Data Cleaning Procedure

Figure A.1 Respondent- and feature-level data cleaning procedure.



B Transformation of the data

To meet the key assumptions of linear regression, namely:

1. a linear relationship between the independent variables and the dependent variable,
2. normality of the error distribution,
3. homoscedasticity, and
4. independence of the residuals,

variables were transformed using the `bestNormalize` package in R. This package applies a range of normalization techniques and selects the most effective transformation based on the Pearson P test for normality. Specifically, it chooses the transformation that yields the highest p -value, indicating the best approximation to a normal distribution Peterson 2021.

The target variable, body mass index (BMI), exhibited a right-skewed distribution and was therefore transformed using the natural logarithm of the value plus one. To assess whether the regression assumptions were satisfied after transformation, the following diagnostic plots were examined (Fig. B.1):

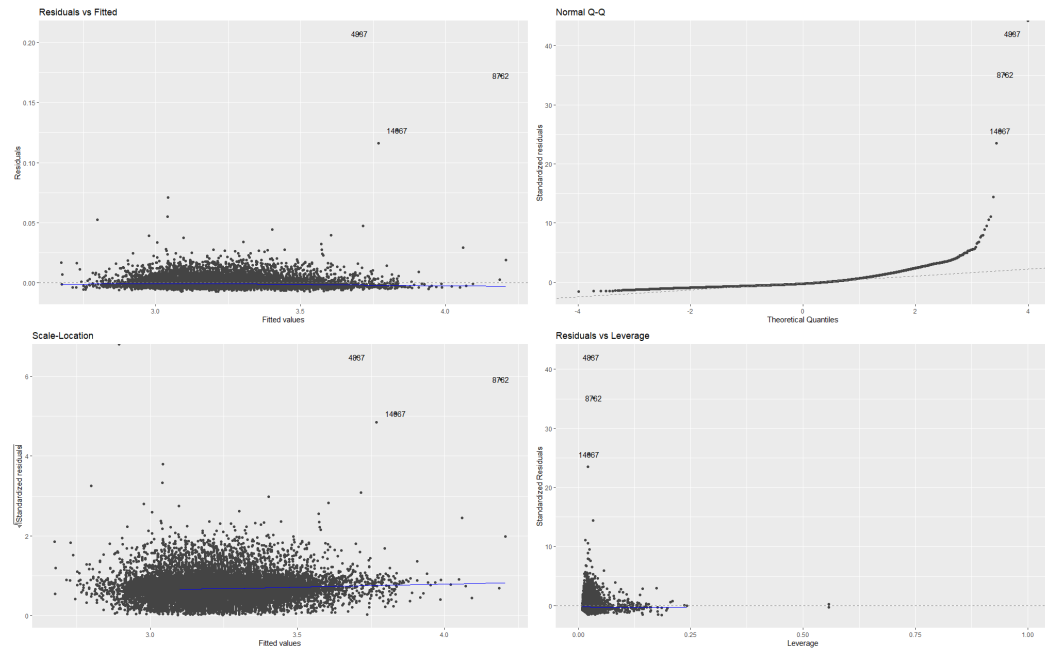


Figure B.1 Regression Diagnostic Plots

- **Residuals vs. Fitted (top left):** This plot assesses the linearity assumption. A random scatter of residuals around the horizontal axis indicates an appropriate linear relationship. In this case, the blue smoothing line is approximately horizontal and centered near zero, suggesting that the linearity assumption is satisfied.
- **Normal Q–Q Plot (top right):** This plot evaluates the normality of the residuals by comparing their empirical quantiles to those of a theoretical normal distribution. Most observations closely follow the diagonal reference line, with minor deviations in the right tail. Given the large sample size, the normality assumption is considered adequately met.
- **Scale–Location Plot (bottom left):** This plot was used to assess homoscedasticity. A roughly horizontal blue line and an even spread of residuals across fitted values indicate constant variance. The observed pattern supports the assumption of homoscedasticity.
- **Residuals vs. Leverage (bottom right):** This plot identifies potentially influential observations. Points with both high leverage and large residuals would indicate violations of model assumptions. Although some observations deviate from the main cluster, none display high Cook’s distance, suggesting that no influential outliers are present. Consequently, the independence assumption is also considered satisfied.

C Removing near zero variance variables

Variables with near-zero variance were removed, as they do not contribute meaningful information and may allow the minority of observed values to exert undue influence on the model Kuhn and Johnson 2013. Near-zero variance variables are defined as variables that either exhibit a single invariant value (resulting in zero variance) or contain a very limited number of unique values relative to the total number of observations.

A cutoff of 10% was applied for the proportion of unique values. In addition, near-zero variance variables are characterized by a highly imbalanced frequency distribution, in which the most frequent value substantially outweighs the second most frequent value. Following standard

practice, a frequency ratio threshold of 95:5 was used, corresponding to the default setting of the applied function.

D List of the 47 variables used in the PCAMIX and clustering analysis

Table D.1 Overview of variables included in the analysis

Variables	Translation	Description
AFLPositieInHuishoudenOP	Position in household	Levels; single < 40 year; single 40 to 64 years; single ≥ 65 years; child < 18 yrs one-parent; child ≥ 18 yrs, one-parent household; child < 18 yrs two-parents; child ≥ 18 yrs two-parents; single parent; Partner in couple, no child with child; partner in couple, no child < 40 yrs; partner in couple, no child, 40 to 64 yrs; partner in couple, no child, ≥ 65 yrs; other member of household
PUBHogeBloeddruk	Hypertension last 12 months	Hypertension, 12 plus; Persons aged 12 years and over who answered yes to the question: Have you suffered from hypertension in the last 12 months?
PUBSF12Fysiek	SF-12 health measure, a summary measure for physical health	The Short Format 12 or SF-12 questionnaire is a selection of 12 questions from the SF-36
AFLOpleidingsniveauGevolgd	Education Attained, 5 divisions	Levels: Pre-primary and primary education; lower secondary education; upper secondary education; bachelor degrees; master degrees and doctorate
PUBGewrichtsslijtage	Arthrosis hips/knees last 12 month	Persons who answered 'yes' to the question: 'Have you suffered from arthrosis of hips or knees in the last 12 months?'
PUBMedicijnVoorgeschreven	Prescribed medicines	Persons who in 14 days preceding the inquiry date used medicines prescribed by a physician. The contraceptive pill is not taken into consideration here.
RIVMBewRichtlBotSpier	Physical activity, Sufficient muscle-/bone-strength activities	People who perform sufficient muscle- and bone-strengthening activities. Persons aged 18 and over should perform muscle- and bone-strengthening activities at least twice a week
RIVMWekelijksSporter	Physical activity, Weekly sporter	People who practise sport at least once a week.
RIVMBeweegRichtlijn	Physical activity, Comply with physical activity guidelines	People who comply with the physical activity guidelines. Adults should be physically active at moderate intensity for at least two and a half hours every week, spread over several days, such as walking and cycling. In addition, they should perform muscle- and bone-strengthening activities at least twice a week.
RIVMBewRichtlBalans	Physical activity, Sufficient balance activities	People who perform sufficient balance activities.

Table D.1 Overview of variables included in the analysis (continued)

Variables	Translation	Description
PUBTandenAantal	Oral Health: real teeth and molars	Average number of real teeth in the upper and lower jaw of people aged 18 years or older
VZAANTPODIUMKUNST05KM	The number of performing arts that can be reached, by road, from a residential address within a 5 km radius.	Art forms performed by professional actors and actresses for a live audience. The professional performing arts take place at locations such as theatres and concert halls, but also in neighbourhood centres and on music venues. Locations with arts as a collateral activity are included in the figures. Festivals aren't included in this variable.
REGVRLLandAchtDeling	Country of origin condensed to 6 levels	Dutch background; With migration background: (1) Europe (excluding Dutch background), (2) Turkey/-Morocco, (3) Surinam/Caribbean Netherlands, (4) Indonesia, (5) America/Africa/Asia/Oceania; Don't know
PUBOraleGezondheid	Oral Health: Good or very good	levels: yes, no; Persons of 12 years or older who answered 'good' or 'very good' to the question: 'How do you assess your general state of teeth and gum?'
PUBHulpmiddelZien	Technical aids/Visual aids	Persons who answer 'yes' to the question: 'Do you sometimes wear glasses or contact lenses?' or 'yes' to the question: 'Do you have any other visual aid?'
PUBType1Diabetes	Chronic disorders: Diabetes	Persons who answered 'yes' to the question: 'Do you suffer from diabetes?' and 'yes' to the following two questions: 'Do you use insulin at the moment?' and 'Did you start using insulin within 6 months after you were diagnosed with diabetes?'
PUBOnvrijwilligUrineverlies	Non-voluntary urine loss	Persons who answered 'yes' to the question: 'Have you suffered from non-voluntary loss of urine (incontinence) in the last 12 months?'
PUBFruitNorm	Sufficient fruit 7 during days per week	People who eat at least 2 pieces of fruit every day. Eat at least 2 portions of fruit every day. One portion of fruit corresponds to 100 grams. These amounts are advised by the Dutch Nutrition Center. These guidelines are based on the Health Council's Guidelines for a Healthy Diet.
PUBTandloos	Oral health: no real teeth/-molars	People aged 18 or older without real teeth in the upper and lower jaw.
PUBRugaandoening	Chronic disorders: Spinal disorder last 12 month	Persons who answered 'yes' to the question: 'Have you suffered from a serious or stubborn spinal disorder in the last 12 months, including a slipped disc?'

Table D.1 Overview of variables included in the analysis (continued)

Variables	Translation	Description
REGVrlPercWelvaartTminEen	Percentile group of household prosperity for year t-1	Condensed into 10 uniform distributed deciles.
REGProvincie	Province	Administrative classification of the Dutch territory.
PUBAstma	Chronic disorders: Asthma last 12 months	Persons who answered 'yes' to the question: 'Have you suffered from asthma in the last 12 months?'
PUBEcstasy12Mnd	Drugs use: Use of amphetamine (last year)	Levels: yes; no
PUBType2Diabetes	Chronic disorders: Diabetes type 2	Persons who answered 'yes' to the question: 'Do you suffer from diabetes?' and who did not suffer from diabetes type 1.
PUBGewrichtontsteking	Chronic disorders: Chronic arthritis	Persons who answered 'yes' to the question: 'Have you suffered from chronic arthritis (chronic rheumatism, rheumatoid arthritis) in the last 12 months?'
PUBGriep prik	Influenza inoculation, 16 years or older	Persons aged 16 years or older who was vaccinated against influenza during the last 12 months.
PUBDrugs12Mnd	Drug use (last year)	Levels: yes; no. Based on the questions about the use of cannabis, amphetamine, LSD, mushrooms, cocaine, heroine, GHB, methadone, 4-Fluoroamphetamine, and other drugs.
VZAANTBIOSCOOP10KM	Number of cinemas within 10 km	The number of cinemas that can be reached, by road, from a residential address within a 10 km radius.
PUBMantelzorg	Giving informal care	Persons giving informal care for at least three months or at least eight hours per week.
PUBErvarenGezondheid	Self-perceived health: good or very good	Levels: yes; no. Persons who answered 'good' or 'very good' to the question: 'How do you assess your general state of health?'
PUBDiarree	Acute diseases: Diarrhoea	Persons who answered 'yes' to the question: 'Have you suffered from diarrhoea during the last 2 months (liquid faeces at least 3 times within 24 hours)?'
PUBSOA	STI testing last 12 months	Persons aged 16 years or older tested for a sexually transmittable infection other than HIV in the last 12 months.
PUBHartinfarct	Chronic disorders: myocard infarctation	Persons who answered 'yes' to the question: 'Have you ever had a myocardial infarction?'
PUBKunstgebit	Oral health: denture	People aged 18 or older with full or partial dentures in the upper and/or lower jaw.
PUBBronchitis	Chronic disorders: COPD, chronic bronch., pulmonary emphysema	Persons who answered 'yes' to the question: 'Have you suffered from COPD, chronic bronchitis or pulmonary emphysema' in the last 12 months?

Table D.1 Overview of variables included in the analysis (continued)

Variables	Translation	Description
VZAANTRESTAU05KM	Number of restaurants within 5 km	The number of restaurants that can be reached, by road, from a residential address within a 5 km radius.
PUBTandartsContact	Visits to a dentist in the Netherlands	Contacts with the dental hygienist, orthodontist and oral surgeon are not included.
AFLExSigarettenRoker	Smoking behaviour: Former smokers	Persons who answered No to the question: Do you ever smoke? and Yes to the question: Have you ever smoked?
AFLRookStatus	Smoking behaviour: Smoking status	Levels: Smokers; Former smokers; Never smokers.
PUBOorontsteking	Acute diseases: Inflammation of the ear	Persons who answered 'yes' to the question: 'Have you suffered from an inflammation of the ear during the last 2 months?'
AFLWelvaartsKwintiel	Prosperity quintile	Households subdivided by level of equivalised prosperity into five equal groups.
REGVrlPercVermogenTminEen	Percentile group of household wealth for year t-1	Condensed into 10 uniform distributed deciles.
REGStedgem	Urbanization	Levels: Very highly urbanized; highly urbanized; moderately urbanized; low urbanization; not urbanized
PUBSF12Psychisch	SF12 health measure, a summary measure for psychological health	The 'Short Format 12' or SF-12 questionnaire is a selection of 12 questions from the SF-36
VZAANTMUSEUM20KM	Number of museums within 20 km	The number of museums that can be reached, by road, from a residential address within a 20 km radius.
PUBUrenMantelzorg	Hours of informal care per week	The average number of hours of care given per week, by an informal care giver

E Scree plots

Figure E.1 Screeplot of the total population respondents in this study

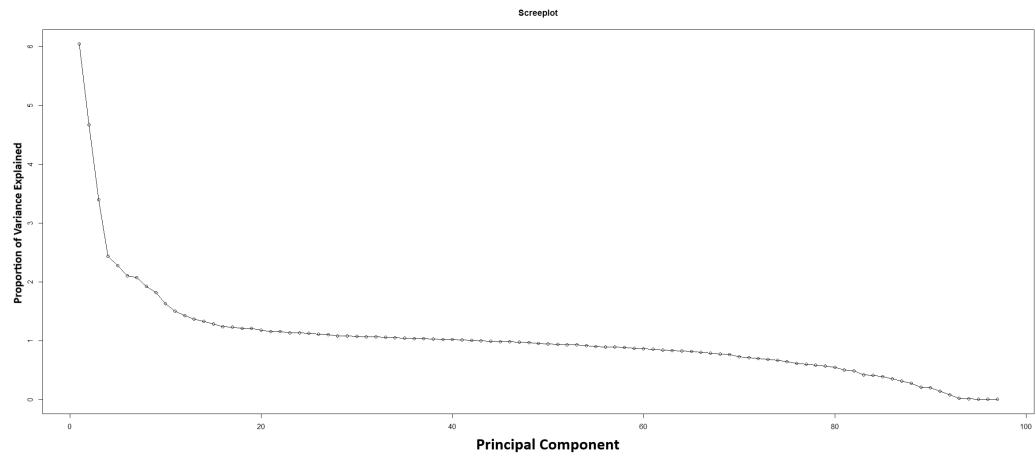


Figure E.2 Screeplot of the underweighted respondents in this study

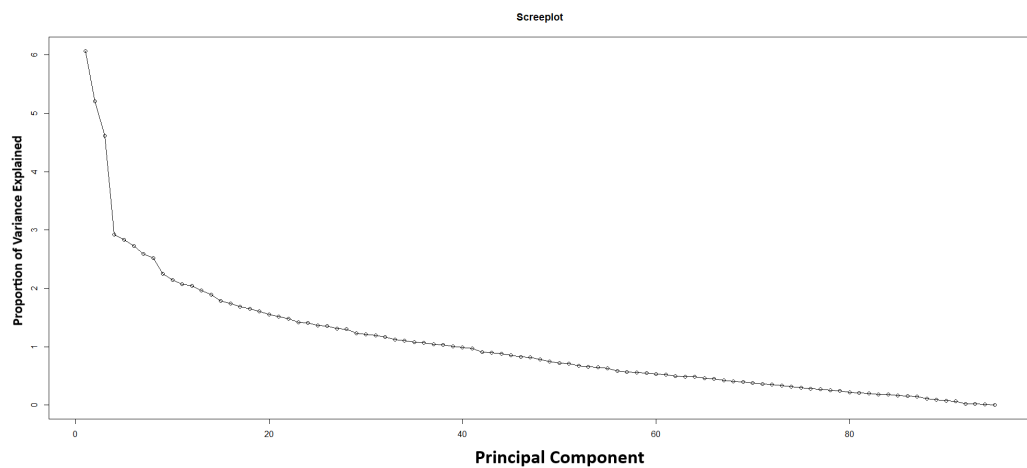


Figure E.3 Screeplot of the normal weighted respondents in this study

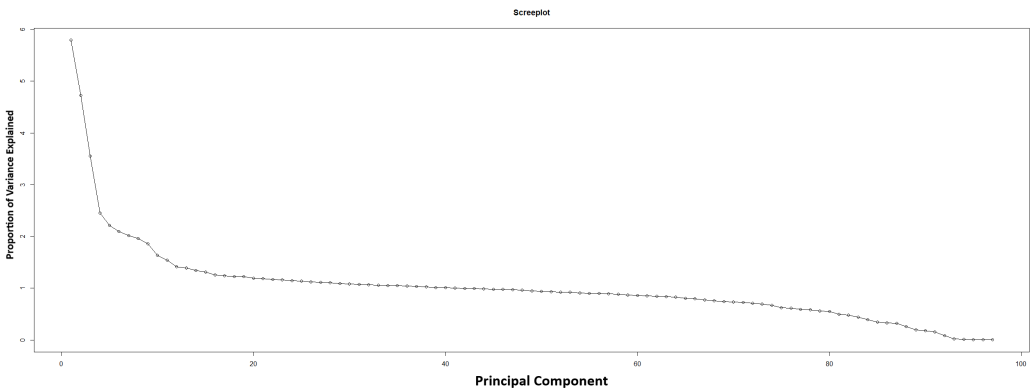


Figure E.4 Screeplot of the overweighted respondents in this study

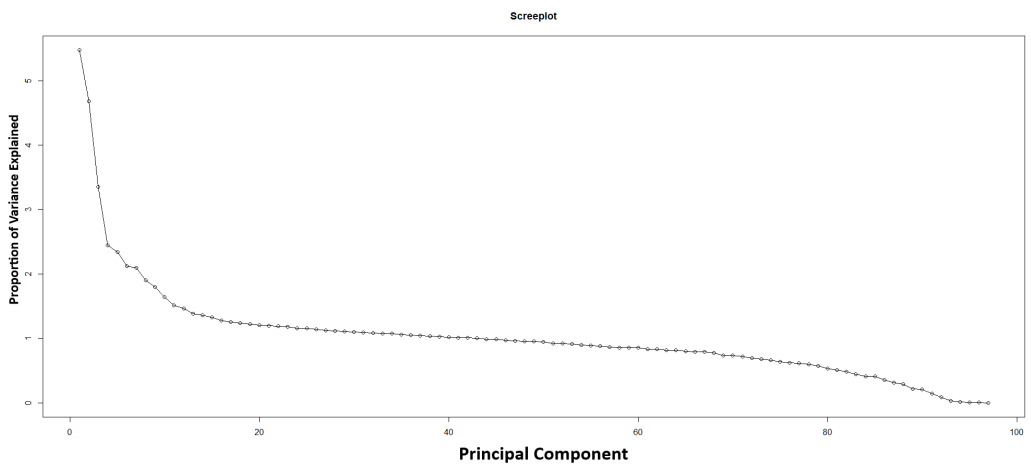


Figure E.5 Screeplot of the obese respondents in this study

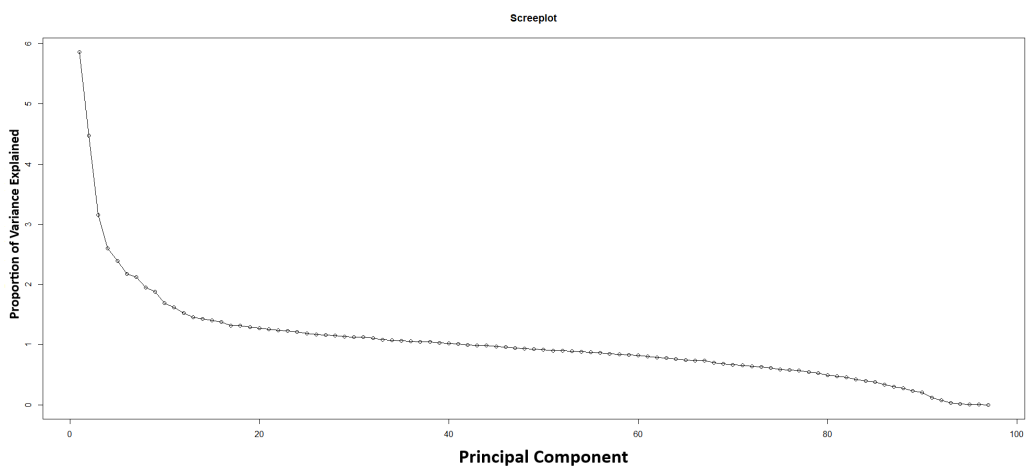
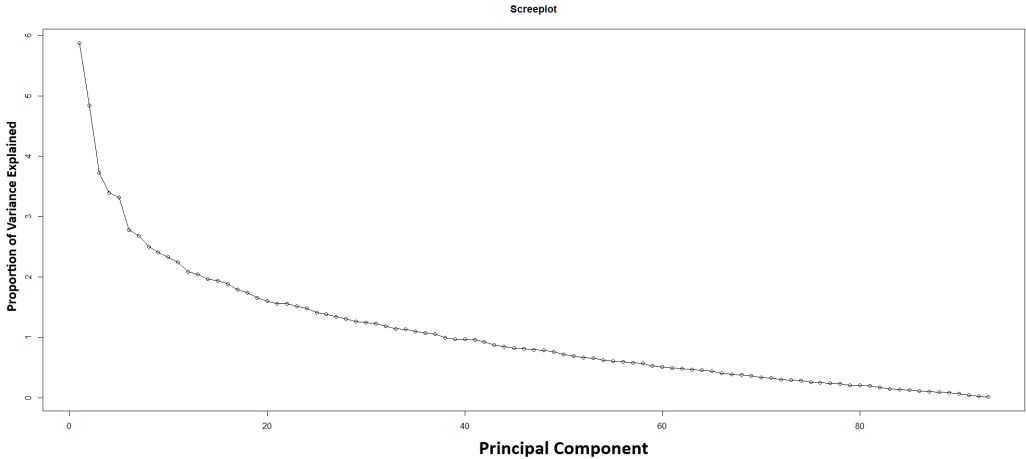


Figure E.6 Screeplot of the morbid obese respondents in this study



F Table with SC scores to determine the amount of clusters

Table F.1 Silhouette coefficients for different numbers of clusters across BMI groups

Clusters	Overall	Underweight	Normal weight	Overweight	Obese	Morbid obesity
6	NA	NA	NA	NA	NA	0.2012
7	NA	NA	NA	NA	NA	0.2177
8	NA	0.1947	NA	NA	NA	0.2213
9	NA	0.1934	NA	NA	NA	0.2303
10	0.2037	0.1930	0.2106	0.2000	0.2172	0.2202
11	0.1940	0.1929	0.2065	0.1863	0.2128	0.2220
12	0.1909	0.1897	0.2006	0.1898	0.1948	0.2301
13	0.1913	0.1881	0.1958	0.1886	0.1822	0.2210
14	0.1889	0.1828	0.2013	0.1833	0.1839	0.2179
15	0.1983	0.1818	0.1982	0.1833	0.1738	0.2289
16	0.1902	0.1809	0.1863	0.1879	0.1693	0.1843
17	0.1890	0.1801	0.1953	0.1902	0.1724	0.2031
18	0.1916	0.1796	0.1904	0.1896	0.1700	0.2027
19	0.1950	0.1792	0.2044	0.1787	0.1665	0.1833
20	0.1865	0.1778	0.1766	0.1875	0.1682	0.2099
21	0.1869	0.1754	0.1901	0.1730	0.1693	0.1757
22	0.1825	0.1723	0.1910	0.1901	0.1713	0.2022
23	0.1907	0.1717	0.2016	0.1869	0.1682	0.1808
24	0.1888	0.1702	0.1929	0.1802	0.1704	0.1738
25	0.1810	0.1701	0.1769	0.1812	0.1657	0.1907
26	0.1747	0.1693	0.1873	0.1825	0.1754	0.1892
27	0.1810	0.1691	0.1868	0.1808	0.1684	0.1825
28	0.1857	0.1679	0.1820	0.1798	0.1672	0.1775
29	0.1795	0.1581	0.1893	0.1768	0.1758	0.1684
30	0.1849	0.1459	0.1852	0.1752	0.1739	0.1850

G Explanation of the interpretation of the clustering results by investigating the different visualizations.

To explain how we analyzed and interpreted the clustering results, we used three complementary visualizations. In this particular example, we used the total study population. In **Fig. G.1** through **Fig. G.27**, three types of plots are shown:

1. the individual plots, in which individual respondents are displayed and colored according to their cluster assignment. Cluster centroids are shown to facilitate interpretation of the cluster positions within the principal component space;
2. the correlation plots, which display the coordinates of continuous variables on the dimensions; and
3. the level plots, which display the coordinates of categorical variable levels on the dimensions.

Cluster profiling was performed through a structured interpretation procedure based on the simultaneous examination of these three visualizations. For each pair of dimensions, cluster centroids that were clearly separated from the origin, either along a single dimension or along a direction spanning both dimensions, were first identified. Cluster characteristics were derived by relating the location of cluster centroids to the locations of continuous variables and categorical variable levels within the same principal component space. Variables and category levels pointing in the same direction as a cluster centroid were interpreted as positively associated with that cluster, whereas variables and category levels pointing in the opposite direction along the same dimension were interpreted as negatively associated. Variables and categories located close to the origin were considered less informative for cluster interpretation because they contributed less strongly to the corresponding dimensions. Across different dimensions, different clusters emerge. The dimensions therefore act as linking mechanisms between the three visualization types. By jointly interpreting centroid positions, continuous variables, and categorical variable levels, relationships between clusters and their distinguishing characteristics can be systematically identified. The procedure described above provides a systematic framework for cluster profiling. To demonstrate its application, we now present an example based on the total study population. In this example, the individual plot, correlation plot, and level plot are interpreted jointly to show how cluster centroids were linked to distinguishing continuous variables and categorical variable levels.

In the first individual plot, two clusters stand out most clearly. Cluster A2 is positioned on the positive side of Dimension 1, whereas Cluster A5 is positioned on the positive side of Dimension 2. Most remaining clusters are located closer to the origin and become more clearly separated in higher dimensions. Consequently, these clusters may appear similar in a two-dimensional projection while being distinguishable in the higher-dimensional principal component space.

Applying the interpretation procedure described above, Cluster A2 can be related to characteristics positioned along Dimension 1. In the corresponding correlation plot, the variable number of teeth (PUBTandenAantal) is strongly positioned on the negative side of this dimension, indicating that respondents in Cluster A2 tend to have fewer teeth. Physical health (PUBSF12Fysiek) is also positioned on the negative side of Dimension 1, although less strongly, suggesting somewhat poorer physical health. Cluster A5 is associated primarily with Dimension 2. The corresponding correlation plot shows strong positive coordinates for variables describing the number of nearby restaurants (VZAANTRESTAU05KM), museums (VZAANTMUSEUM20KM), cinemas (VZAANTBIOSCOOP10KM), and performing arts venues (VZAANTPODIUMKUNST05KM). This suggests that respondents in Cluster A5 tend to reside in environments with a relatively high concentration of amenities.

The level plot provides further support for the interpretation of Cluster A2. Respondents in this cluster are more frequently associated with levels such as:

- A “Yes” answer to the question: Are you toothless? (PUBTandeloos=Ja)
- A “Yes” answer to the question: Do you have dentures? (PUBKunstgebit=Ja)
- More frequently being single and ≥ 65 years old in terms of household position (AFLPositieInHuishoudenOP=Alleenstaand, 65 jaar of ouder)

- More frequently being a partner in a couple without children, ≥ 65 years old (AFLPositieInHuishoudenOP=Partner in paar zonder kind; ≥ 65 jaar)
- A “Yes” answer to the question: Have you ever had a myocardial infarction? (PUBHartinfarct=Ja)
- A “No” answer to the question: Do you have contact with a dental hygienist? (PUBTandartsContact=Nee)
- More frequently having followed pre-primary and primary education (AFLopleidingsniveauGevolgd=Basisonderwijs)

This approach facilitates a systematic and reproducible interpretation of the cluster profiles by explicitly linking cluster centroids to the continuous variables and categorical levels within the same principal component space. While the resulting profiles remain interpretative rather than algorithmically generated, the procedure provides a transparent framework for deriving cluster characteristics. The figure captions summarize the cluster centroids, variable names, and category levels included in the analysis and provide access to the complete underlying structure used for interpretation.

Figure G.1 Projection of individual observations onto dimensions 1 and 2, colored by cluster assignment, with cluster centroids indicated. For interpretability, emphasis is placed on the most strongly represented centroids along each dimension (dimension 1, positive side: cluster 2; dimension 2, positive side: cluster 5).

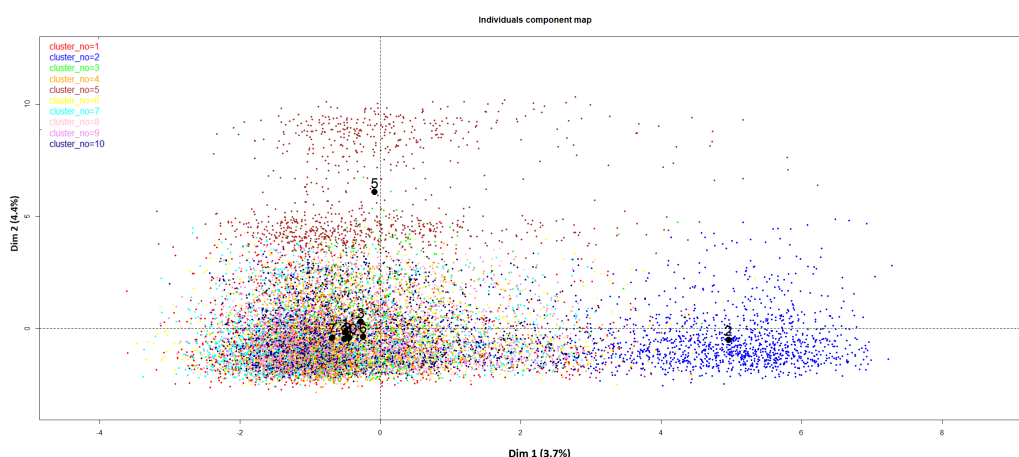


Figure G.2 Projection of individual observations onto dimensions 2 and 3, colored by cluster assignment, with cluster centroids indicated. For interpretability, emphasis is placed on the most strongly represented centroids along each dimension (dimension 2, positive side: cluster 5; dimension 3, positive side: cluster 1 and 7, negative side: 4 and 8).

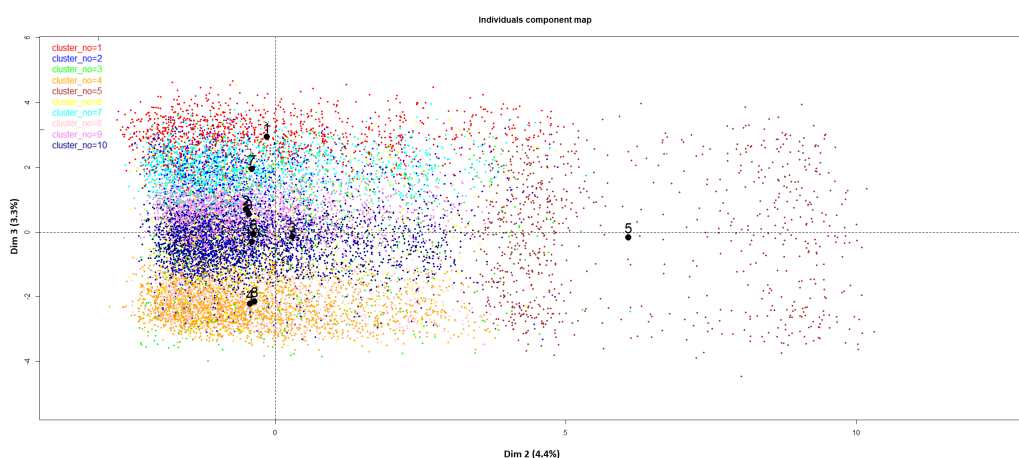


Figure G.3 Projection of individual observations onto dimensions 3 and 4, colored by cluster assignment, with cluster centroids indicated. For interpretability, emphasis is placed on the most strongly represented centroids along each dimension (dimension 3, positive side: cluster 1 and 7, negative side: 4 and 8; dimension 4, positive side: cluster 9 and 10, negative side: cluster 1, 4 and 8).

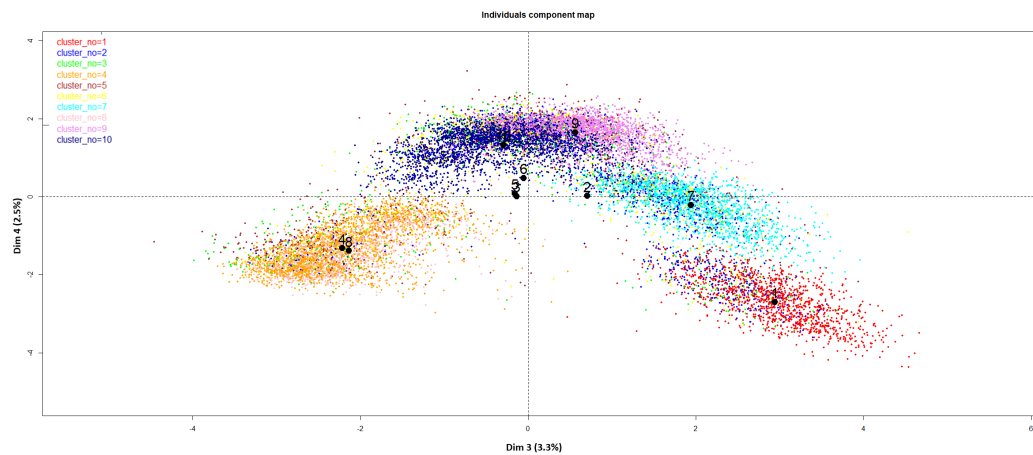


Figure G.4 Projection of individual observations onto dimensions 4 and 5, colored by cluster assignment, with cluster centroids indicated. For interpretability, emphasis is placed on the most strongly represented centroids along each dimension (dimension 4, positive side: cluster 9 and 10, negative side: cluster 1, 4 and 8; dimension 5, positive side: cluster 6).

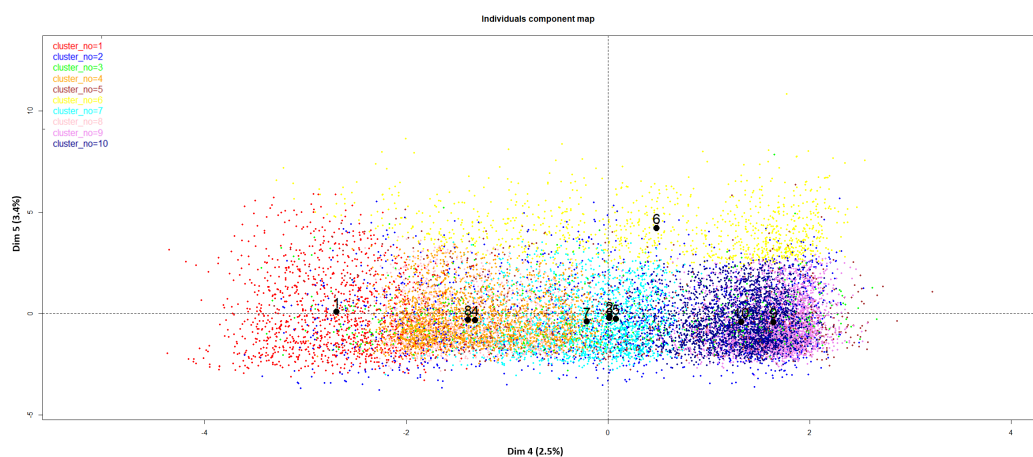


Figure G.5 Projection of individual observations onto dimensions 5 and 6, colored by cluster assignment, with cluster centroids indicated. For interpretability, emphasis is placed on the most strongly represented centroids along each dimension (dimension 5, positive side: cluster 6; dimension 6, positive side: cluster 7, negative side: cluster 1 and 10).

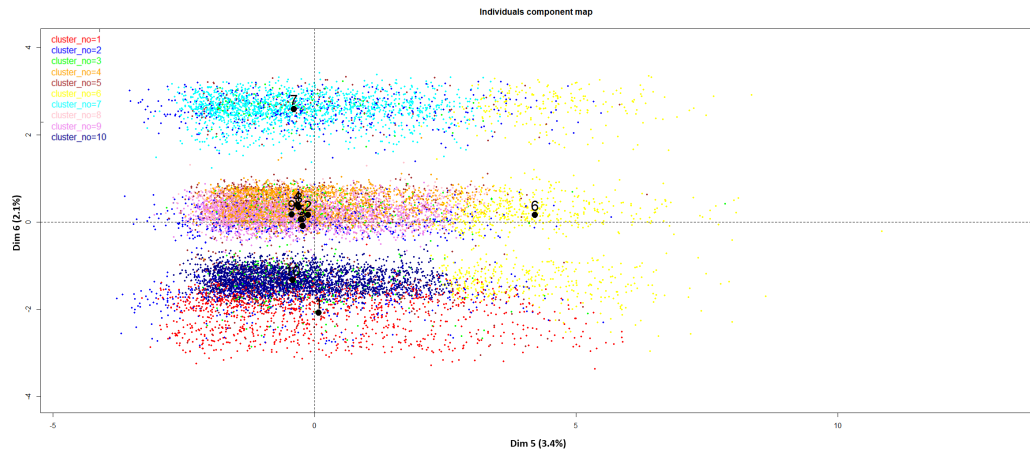


Figure G.6 Projection of individual observations onto dimensions 6 and 7, colored by cluster assignment, with cluster centroids indicated. For interpretability, emphasis is placed on the most strongly represented centroids along each dimension (dimension 6, positive side: cluster 7, negative side: cluster 1 and 10; dimension 7, positive side: cluster 1 and 9, negative side: cluster 7 and 10).

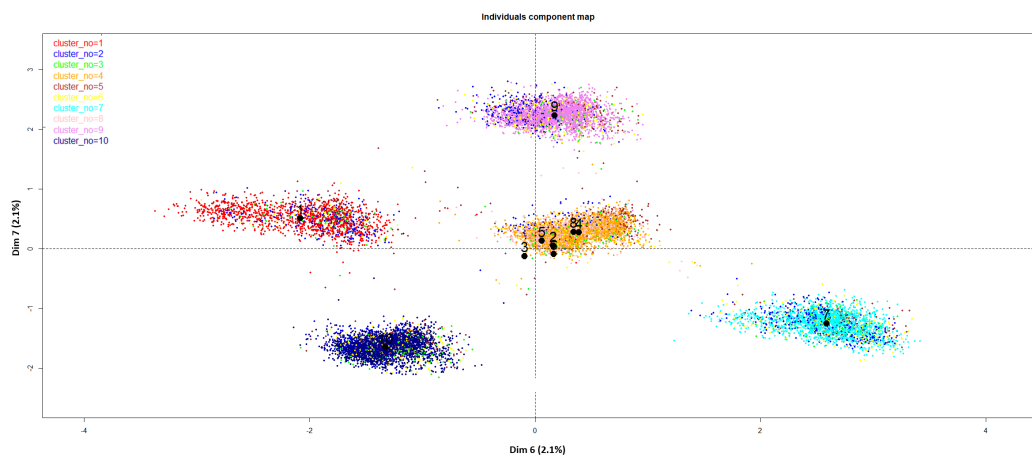


Figure G.7 Projection of individual observations onto dimensions 7 and 8, colored by cluster assignment, with cluster centroids indicated. For interpretability, emphasis is placed on the most strongly represented centroids along each dimension (dimension 7, positive side: cluster 1 and 9, negative side: cluster 7 and 10; dimension 8, positive side: cluster 3).

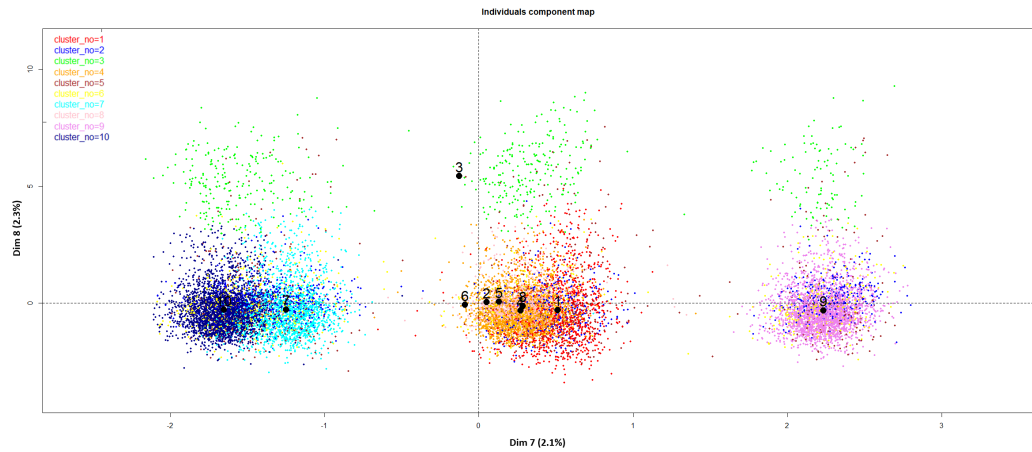


Figure G.8 Projection of individual observations onto dimensions 8 and 9, colored by cluster assignment, with cluster centroids indicated. For interpretability, emphasis is placed on the most strongly represented centroids along each dimension (dimension 8, positive side: cluster 3; dimension 9, positive side: cluster 4, negative side: cluster 8).

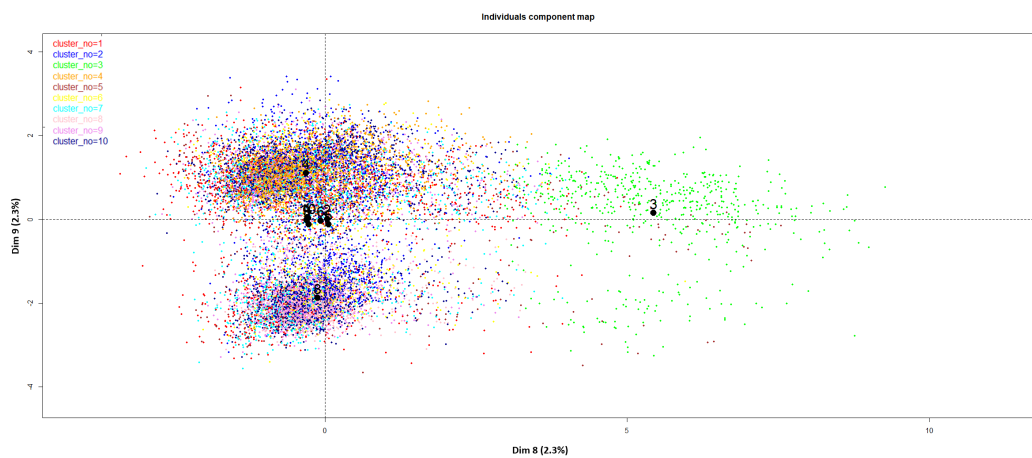


Figure G.9 Projection of individual observations onto dimensions 9 and 10, colored by cluster assignment, with cluster centroids indicated. For interpretability, emphasis is placed on the most strongly represented centroids along each dimension (dimension 9, positive side: cluster 4, negative side: cluster 8).

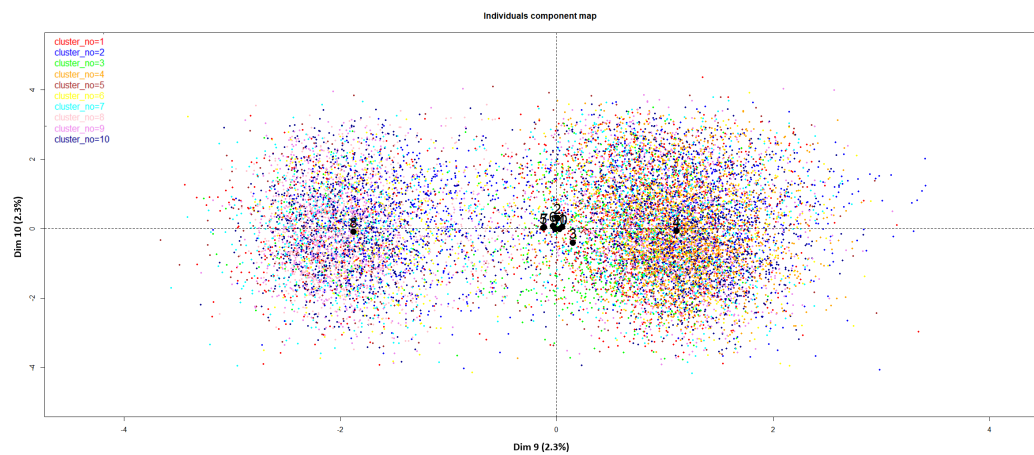


Figure G.10 Correlation plot of continuous variables projected onto dimensions 1 and 2. For interpretability, emphasis is placed on the variables with the strongest contributions to these dimensions (dimension 1, negative side: PUBTandenAantal; dimension 2, positive side: VZAANTBIOSCOOP10KM, VZAANTRESTAU05KM, VZAANTMUSEUM20KM, VZAANTPODIUMKUNST05KM). Definitions of the variable labels are provided in Table D.1.

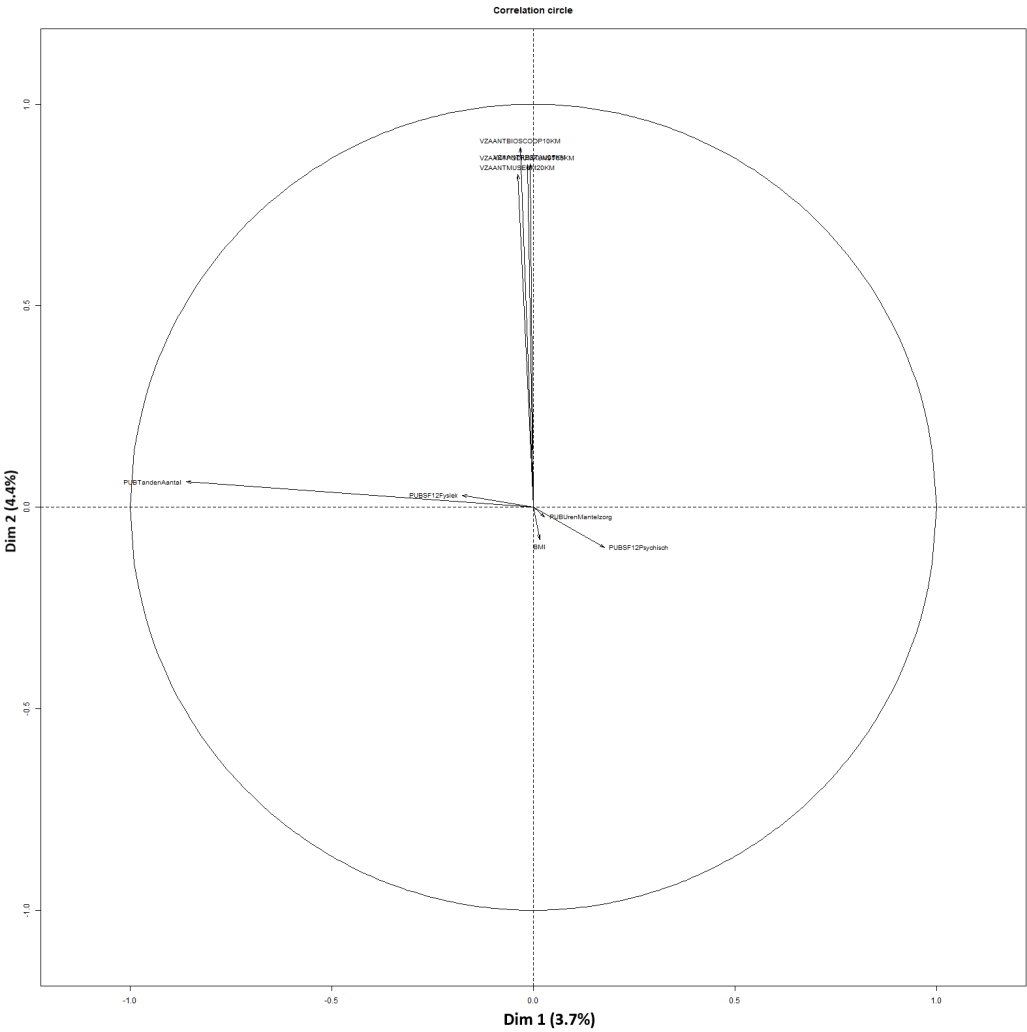


Figure G.11 Correlation plot of continuous variables projected onto dimensions 2 and 3. For interpretability, emphasis is placed on the variables with the strongest contributions to these dimensions (dimension 2, positive side: VZAANTBIOSCOOP10KM, VZAANTRESTAU05KM, VZAANTMUSEUM20KM, VZAANTPODIUMKUNST05KM). Definitions of the variable labels are provided in Table D.1.

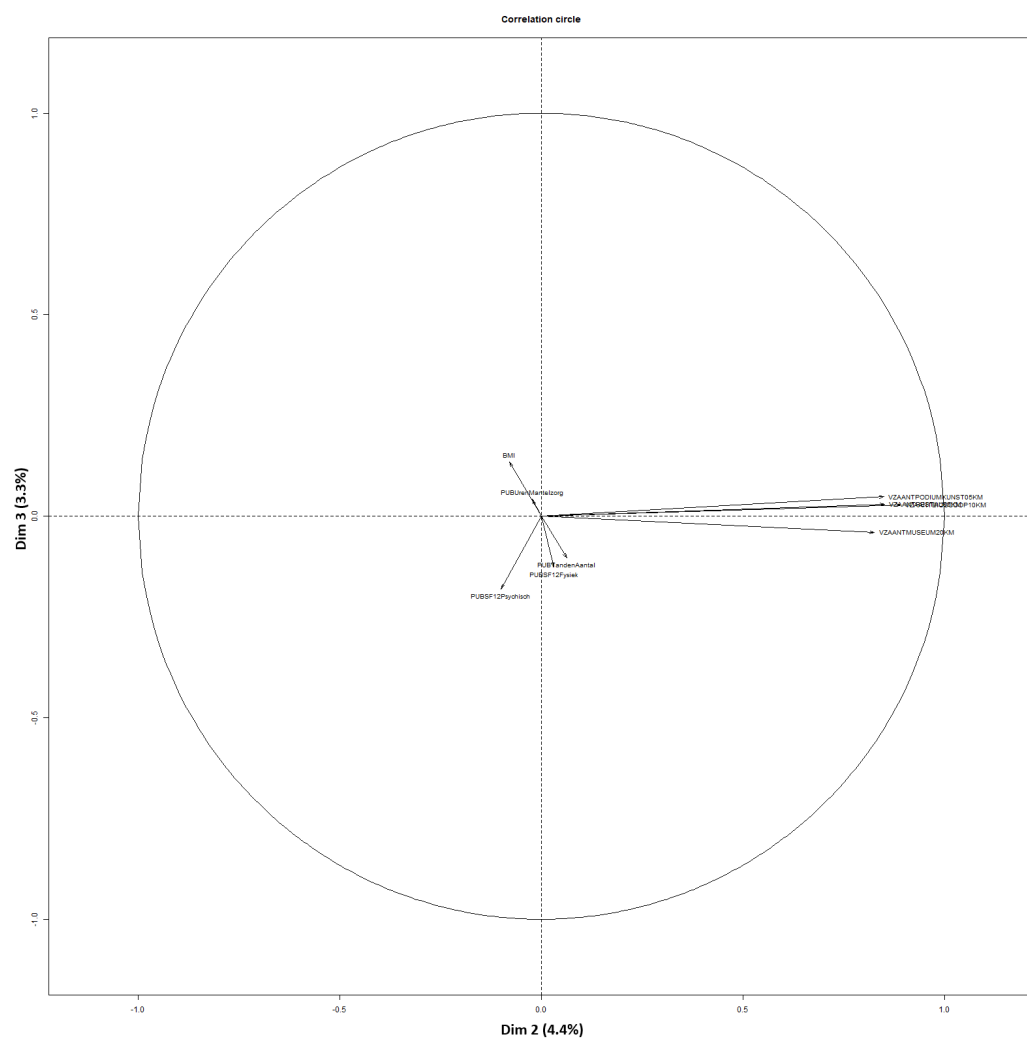


Figure G.12 Correlation plot of continuous variables projected onto dimensions 3 and 4.

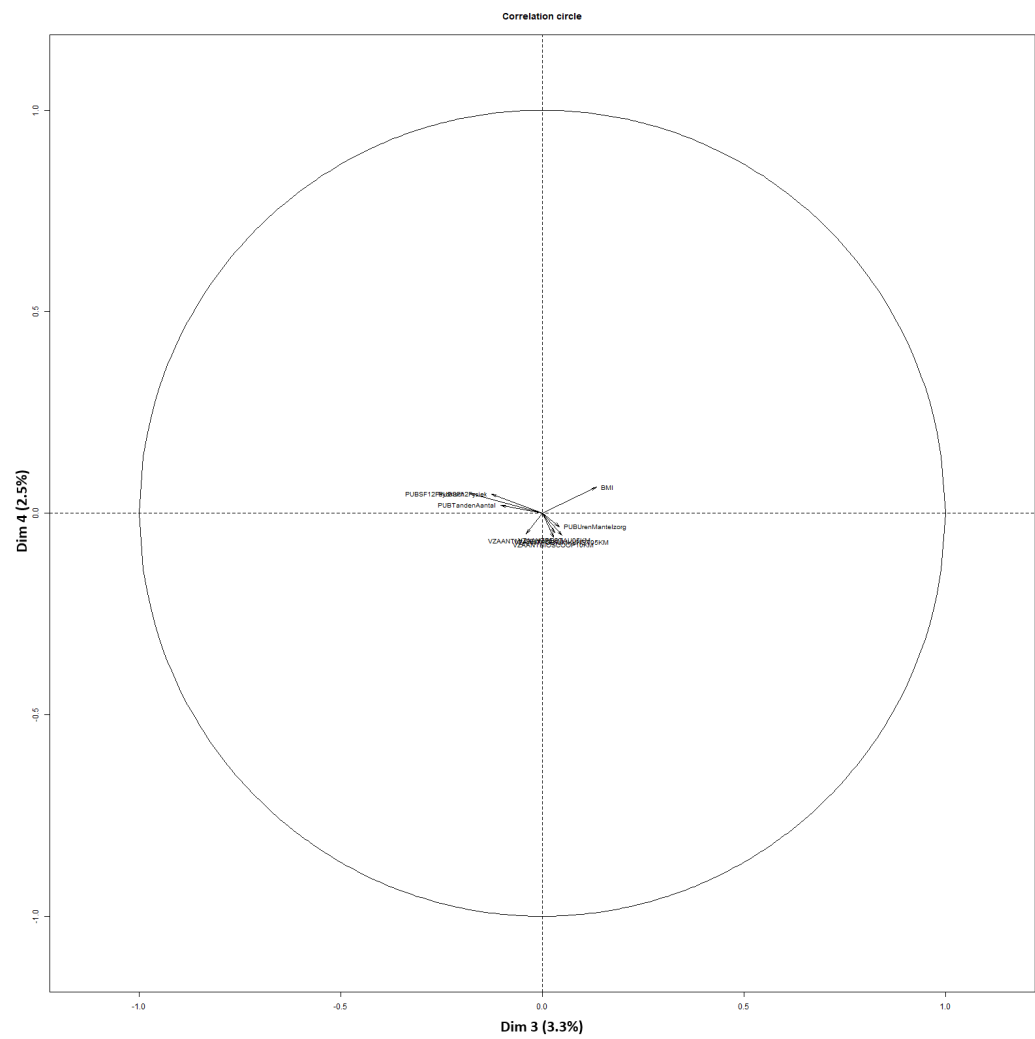


Figure G.13 Correlation plot of continuous variables projected onto dimensions 4 and 5. For interpretability, emphasis is placed on the variables with the strongest contributions to these dimensions (dimension 5, negative side: PUBSF12Fysiek and PUBSF12Psychisch, positive side: BMI). Definitions of the variable labels are provided in Table D.1.

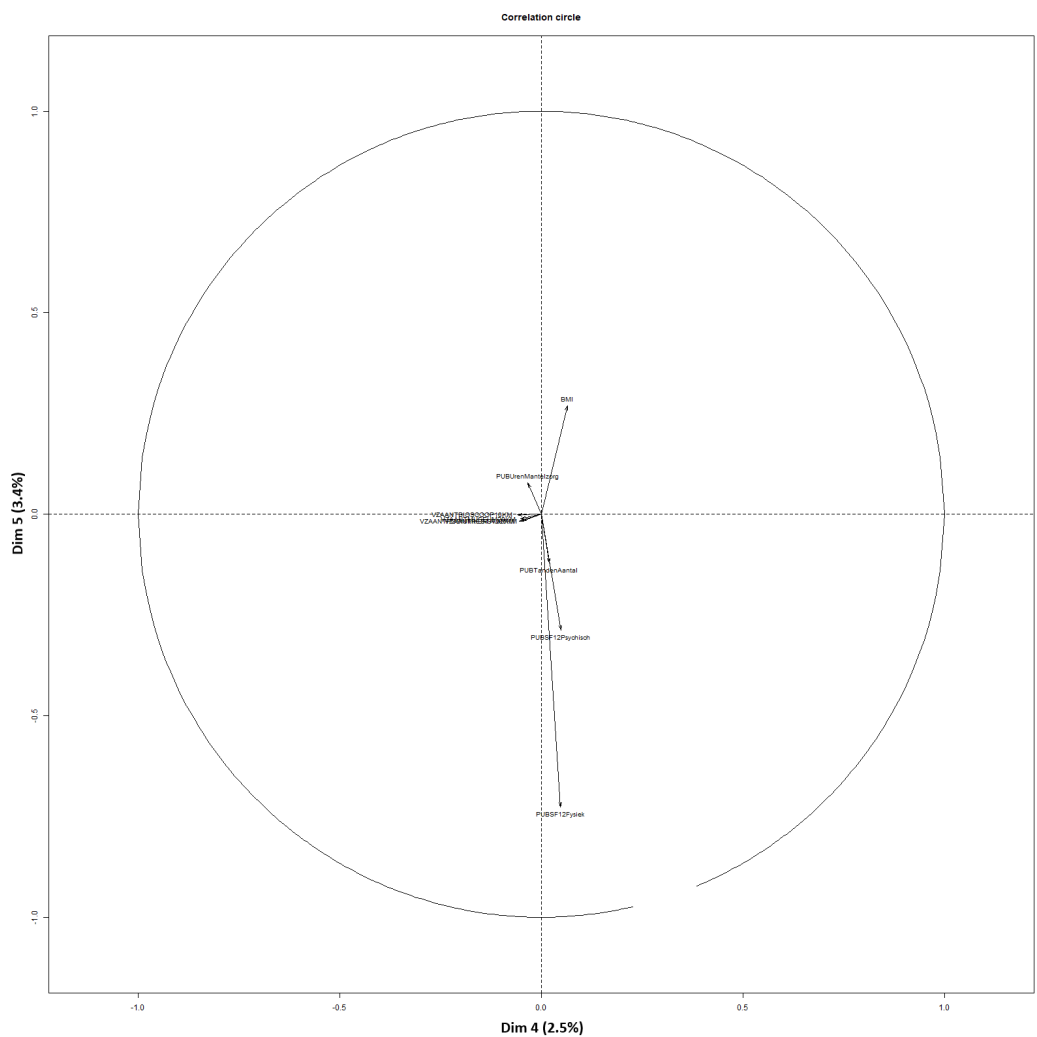


Figure G.14 Correlation plot of continuous variables projected onto dimensions 5 and 6. For interpretability, emphasis is placed on the variables with the strongest contributions to these dimensions (dimension 5, negative side: PUBSF12Fysiek and PUBSF12Psychisch, positive side: BMI). Definitions of the variable labels are provided in Table D.1.

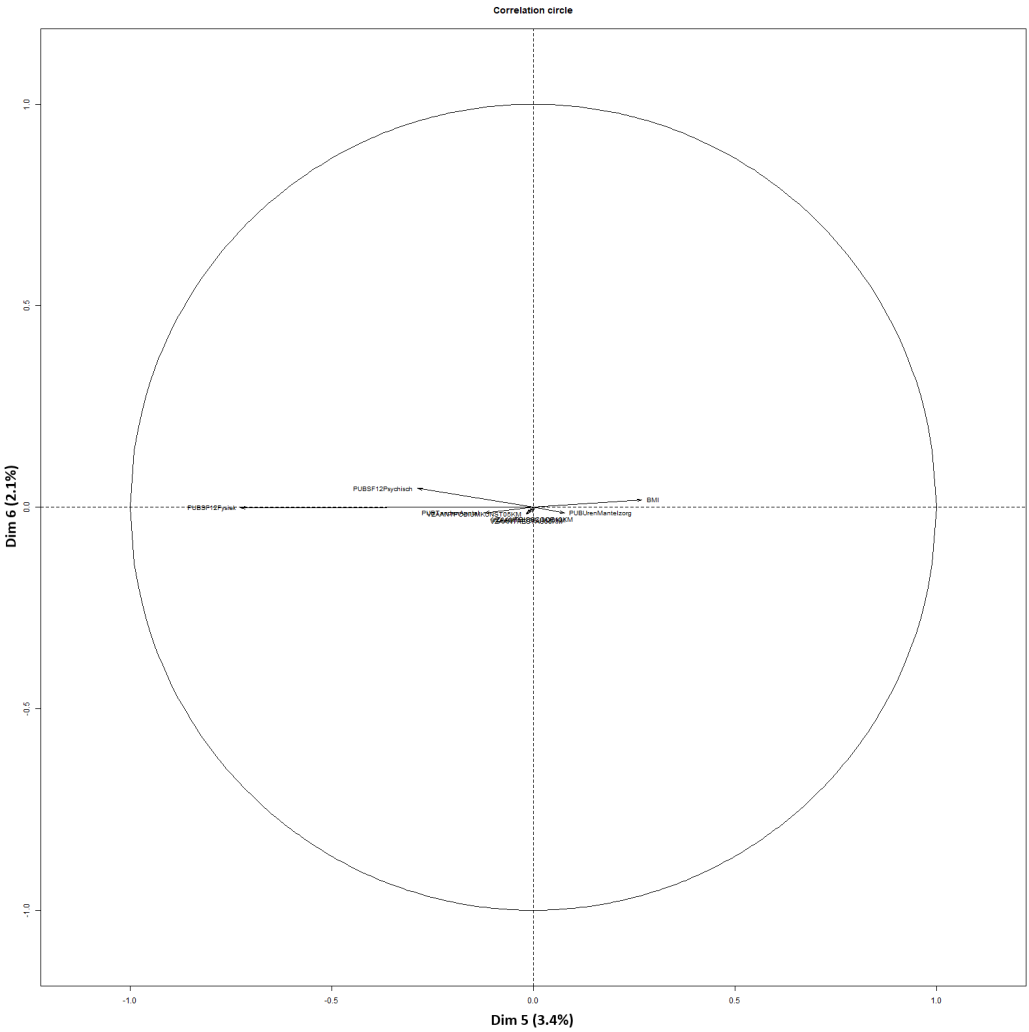
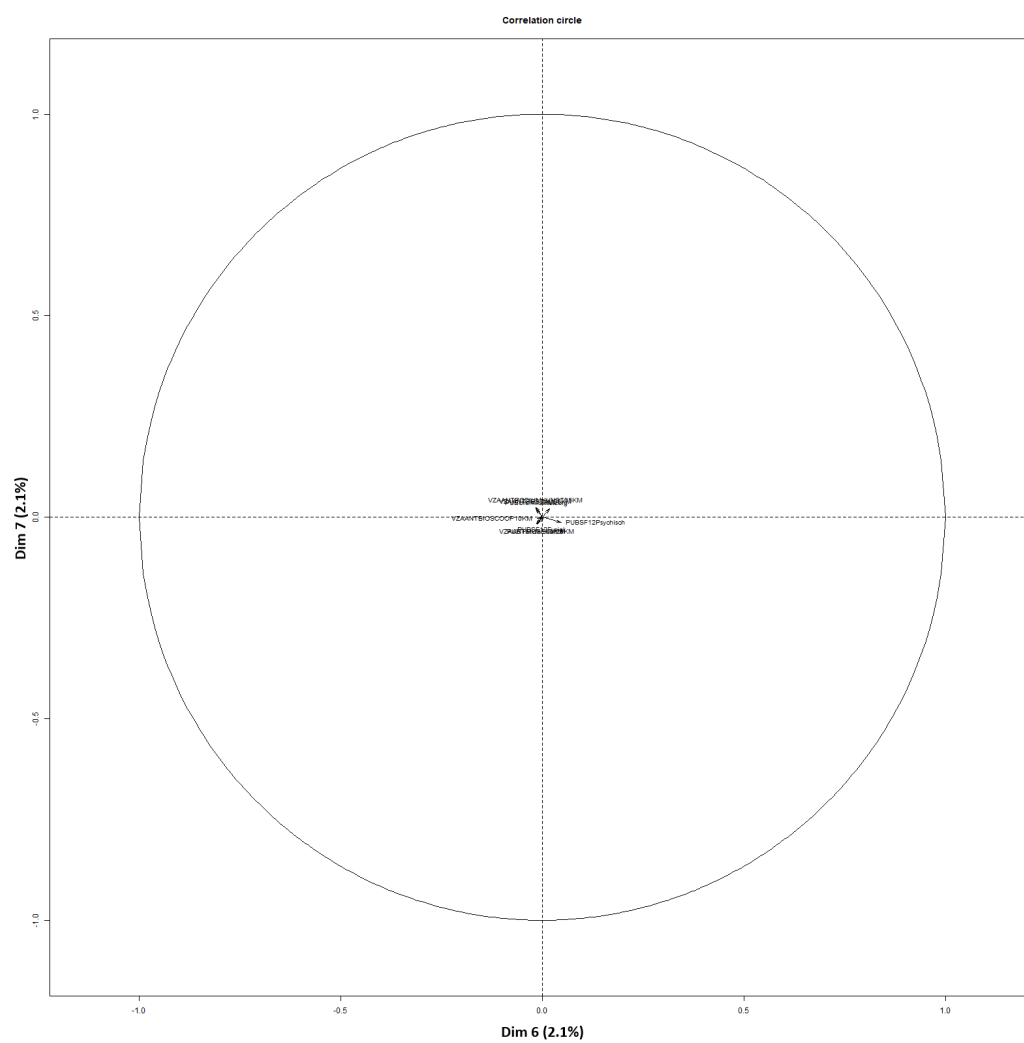


Figure G.15 Correlation plot of continuous variables projected onto dimensions 6 and 7.



Correlation circle

Figure G.17 Correlation plot of continuous variables projected onto dimensions 8 and 9. For interpretability, emphasis is placed on the variables with the strongest contributions to these dimensions (dimension 8, negative side: PUBSF12Psychisch). Definitions of the variable labels are provided in Table D.1.

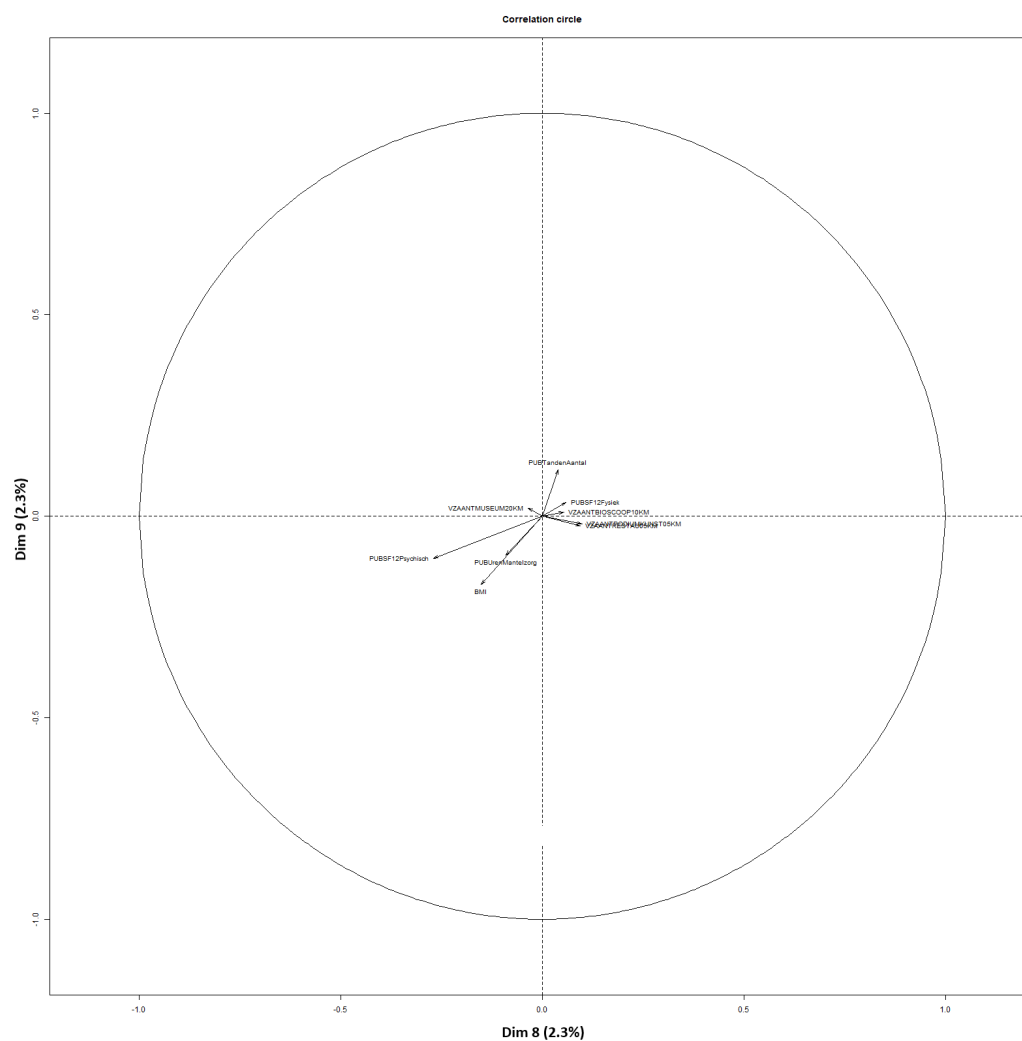


Figure G.18 Correlation plot of continuous variables projected onto dimensions 9 and 10.

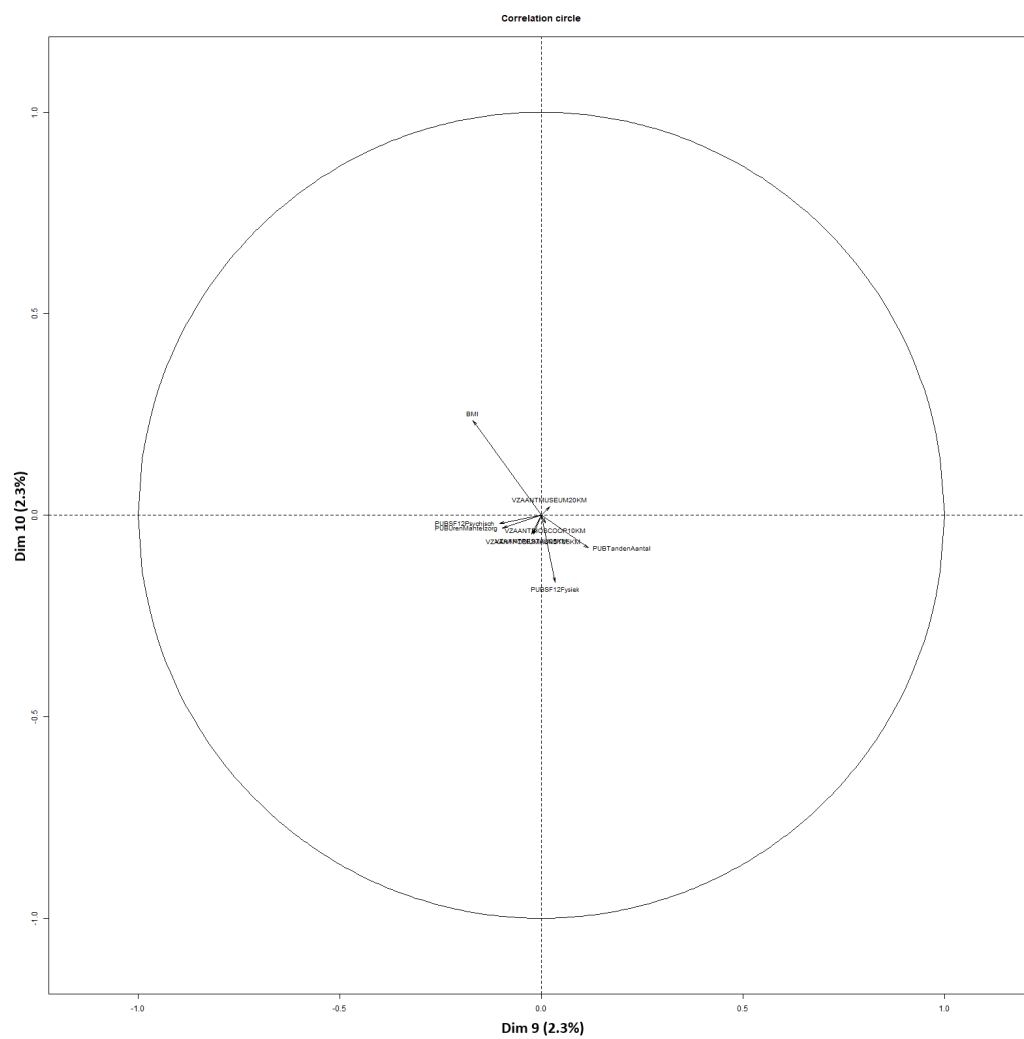


Figure G.19 Projection of categorical variable levels onto dimensions 1 and 2. For interpretability, emphasis is placed on the levels with the strongest contributions to these dimensions. (dimension 1, positive side: PUBTandloos=Yes, PUBKunstgebit=Yes, AFLPositieInHuishoudenOP=Single $\geq$ 65 years, AFLOpleidingsniveauGevolgd=Pre-primary and primary education, PUBTandartsContact=No, PUBHartinfarct=Yes, AFLPositieInHuishoudenOP=Partner in couple, no child $\geq$ 65 yrs, PUBType1Diabetes=Yes, AFLOpleidingsniveauGevolgd=Lower secondary education, REGVrlPercWelvaartTminEen=2nd decile, PUBType2Diabetes=Yes, PUBGriep prik=Yes; dimension 2, positive side: REGStedgem=Very highly urbanized, REGProvincie=Noord-Holland, REGProvincie=Zuid-Holland, negative side: REGProvincie=Zeeland, REGProvincie=Drenthe, REGStedgem=Not urbanized, REGProvincie=Friesland, REGProvincie=Flevoland, REGStedgem=Low urbanization, REGProvincie=Overijssel). Definitions of the variable labels are provided in Table D.1.

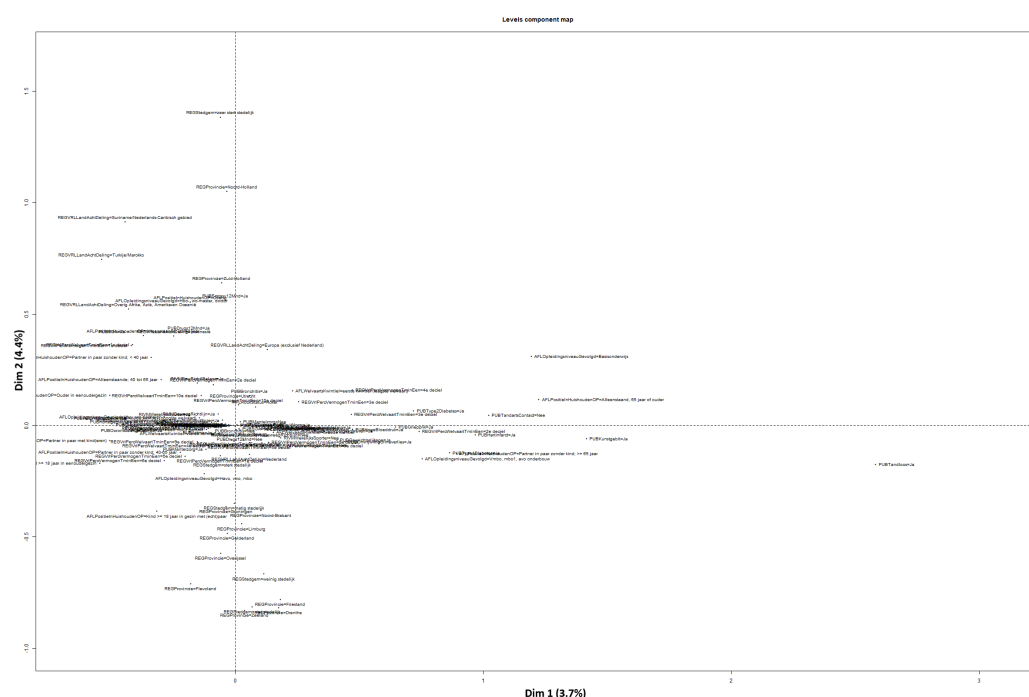


Figure G.20 Projection of categorical variable levels onto dimensions 2 and 3. For interpretability, emphasis is placed on the levels with the strongest contributions to these dimensions. (dimension 2, positive side: REGStedgem=Very highly urbanized, REGProvincie=Noord-Holland, REGProvincie=Zuid-Holland, negative side: REGProvincie=Zeeland, REGProvincie=Drenthe, REGStedgem=Not urbanized, REGProvincie=Friesland, REGProvincie=Flevoland, REGStedgem=Low urbanization, REGProvincie=Overijssel; dimension 3, positive side: REGVrlPercWelvaartTminEen=1st decile, REGVrlPercVermogenTminEen=2nd decile, AFLWelvaartsKwintiel=1st quintile (lowest prosperity), REGVrlPercWelvaartTminEen=2nd decile, REGVrlPercVermogenTminEen=3rd decile, AFLOpleidingsniveauGevolgd=Pre-primary and primary education, REGVrlPercWelvaartTminEen=3rd decile, AFLWelvaartsKwintiel=2nd quintile, REGVrlPercWelvaartTminEen=4th decile, REGVrlPercVermogenTminEen=1st decile, AFLPositieInHuishoudenOP=Child $\geq$ 18 years in one-parent household, negative side: REGVrlPercWelvaartTminEen=10th decile, REGVrlPercVermogenTminEen=10th decile, AFLWelvaartsKwintiel=5th quintile (highest prosperity), REGVrlPercWelvaartTminEen=9th decile, REGVrlPercVermogenTminEen=9th decile). Definitions of the variable labels are provided in Table D.1.

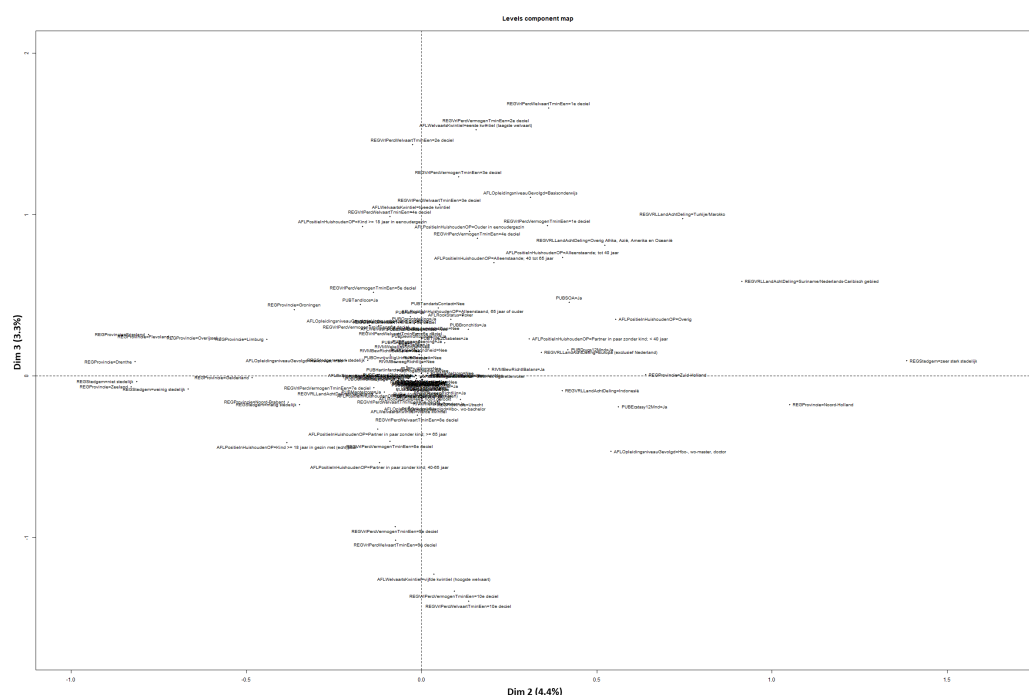


Figure G.21 Projection of categorical variable levels onto dimensions 3 and 4. For interpretability, emphasis is placed on the levels with the strongest contributions to these dimensions. (dimension 3, positive side: REGVrIPercWelvaartTminEen=1st decile, REGVrIPercVermogenTminEen=2nd decile, AFLWelvaartsKwintiel=1st quintile (lowest prosperity), REGVrIPercWelvaartTminEen=2nd decile, REGVrIPercVermogenTminEen=3rd decile, AFLOpleidingsniveauGevolgd=Pre-primary and primary education, REGVrIPercWelvaartTminEen=3rd decile, AFLWelvaartsKwintiel=2nd quintile, REGVrIPercWelvaartTminEen=4th decile, REGVrIPercVermogenTminEen=1st decile, AFLPositieInHuishoudenOP=Child $\geq$ 18 years in one-parent household, negative side: REGVrIPercWelvaartTminEen=10th decile, REGVrIPercVermogenTminEen=10th decile, AFLWelvaartsKwintiel=5th quintile (highest prosperity), REGVrIPercWelvaartTminEen=9th decile, REGVrIPercVermogenTminEen=9th decile; dimension 4, positive side: REGVrIPercWelvaartTminEen=6th decile, AFLWelvaartsKwintiel=3rd quintile, REGVrIPercWelvaartTminEen=5th decile, REGVrIPercWelvaartTminEen=7th decile, AFLWelvaartsKwintiel=4th quintile, REGVrIPercWelvaartTminEen=8th decile, REGVrIPercVermogenTminEen=6th decile, REGVrIPercVermogenTminEen=7th decile, REGVrIPercVermogenTminEen=5th decile, REGVrIPercVermogenTminEen=8th decile, negative side: REGVrIPercWelvaartTminEen=1st decile, AFLWelvaartsKwintiel=1st quintile (lowest prosperity), REGVrIPercWelvaartTminEen=2nd decile, REGVrIPercVermogenTminEen=2nd decile, REGVrIPercWelvaartTminEen=10th decile, REGVrIPercVermogenTminEen=10th decile). Definitions of the variable labels are provided in Table D.1.

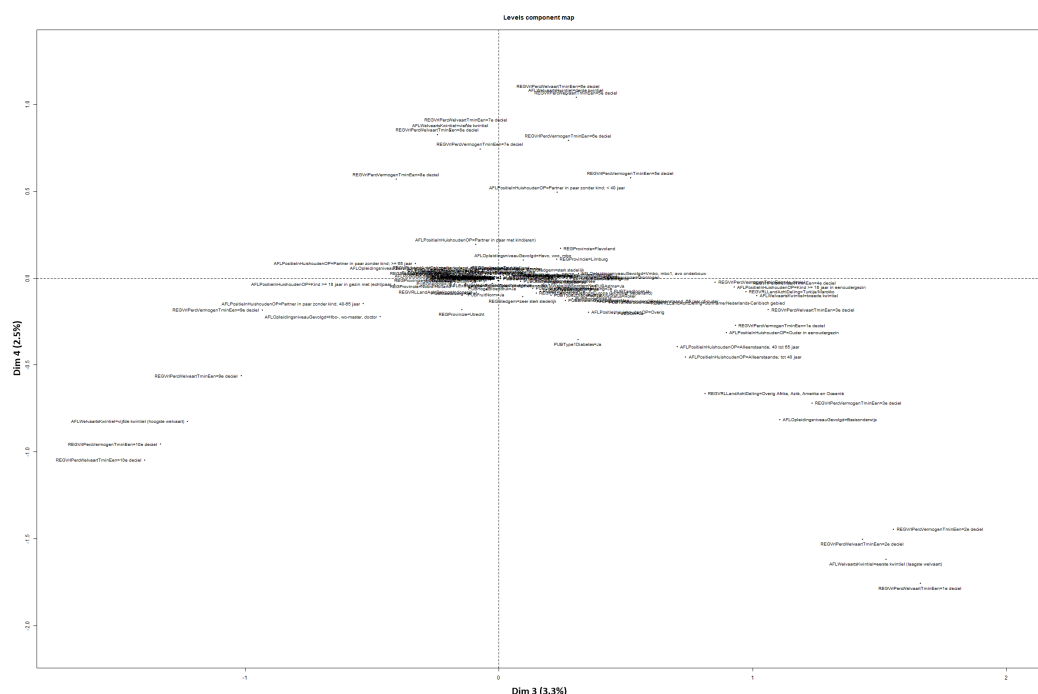


Figure G.22 Projection of categorical variable levels onto dimensions 4 and 5. For interpretability, emphasis is placed on the levels with the strongest contributions to these dimensions. (dimension 4, positive side: REGVrIPercWelvaartTminEen=6th decile, AFLWelvaartsKwintiel=3rd quintile, REGVrIPercWelvaartTminEen=5th decile, REGVrIPercWelvaartTminEen=7th decile, AFLWelvaartsKwintiel=4th quintile, REGVrIPercWelvaartTminEen=8th decile, REGVrIPercVermogenTminEen=6th decile, REGVrIPercVermogenTminEen=7th decile, REGVrIPercVermogenTminEen=5th decile, REGVrIPercVermogenTminEen=8th decile, negative side: REGVrIPercWelvaartTminEen=1st decile, AFLWelvaartsKwintiel=1st quintile (lowest prosperity), REGVrIPercWelvaartTminEen=2nd decile, REGVrIPercVermogenTminEen=2nd decile, REGVrIPercWelvaartTminEen=10th decile, REGVrIPercVermogenTminEen=10th decile; dimension 5, positive side: PUBGewrichtontsteking=Yes, PUBBronchitis=Yes, PUBRugaandoening=Yes, PUBOnvrijwilligUrineverlies=Yes, PUBErvarenGezondheid=No, PUBGewrichtsslijtage=Yes, PUBType1Diabetes=Yes, PUBAstma=Yes, PUBOorontsteking=Yes, PUBType2Diabetes=Yes, PUBHartinfarct=Yes, PUBDiarree=Yes, PUBHogeBloeddruk=Yes, AFLPositieInHuishoudenOP=Single ≥ 65 years, PUBMedicijnVoorgescreven=Yes). Definitions of the variable labels are provided in Table D.1.

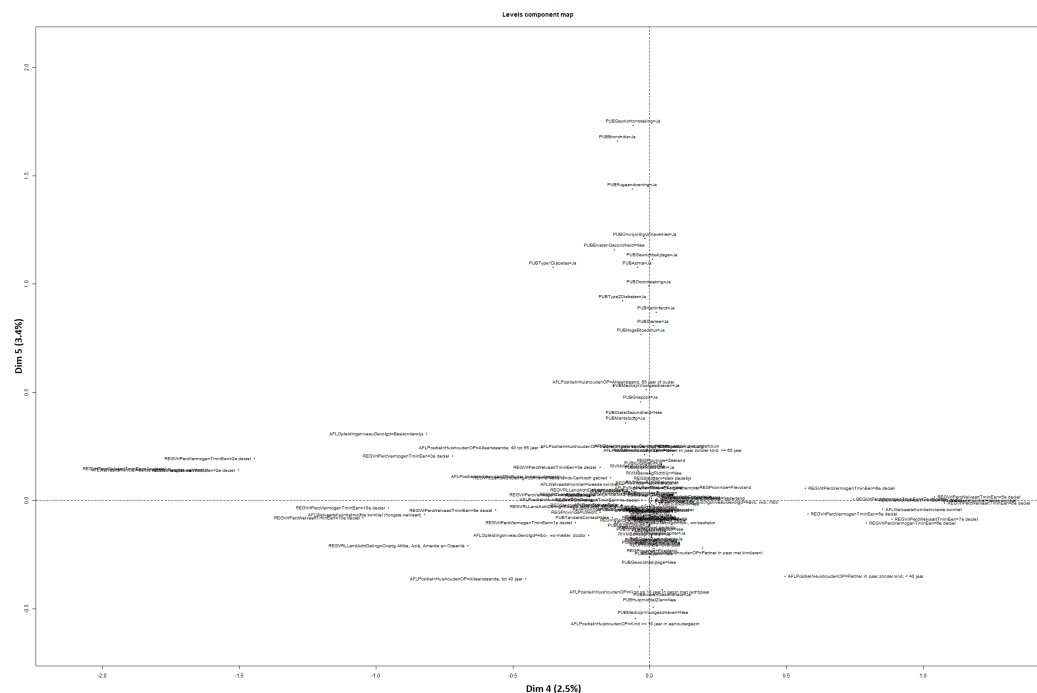


Figure G.23 Projection of categorical variable levels onto dimensions 5 and 6. For interpretability, emphasis is placed on the levels with the strongest contributions to these dimensions. (dimension 5, positive side: PUBGewrichtontsteking=Yes, PUBBronchitis=Yes, PUBRugaandoening=Yes, PUBOnvrijwilligUrineverlies=Yes, PUBErvarenGezondheid=No, PUBGewrichtsslijtage=Yes, PUBType1Diabetes=Yes, PUBAstma=Yes, PUBOorontsteking=Yes, PUBType2Diabetes=Yes, PUBHartinfarct=Yes, PUBDiarree=Yes, PUBHogeBloeddruk=Yes, AFLPositieInHuishoudenOP=Single $\geq$ 65 years, PUBMedicijnVoorgeschreven=Yes; dimension 6, positive side: REGVrlPercWelvaartTminEen=3rd decile, AFLWelvaartsKwintiel=2nd quintile, REGVrlPercWelvaartTminEen=4th decile, REGVrlPercVermogenTminEen=4th decile, negative side: REGVrlPercWelvaartTminEen=1st decile, AFLWelvaartsKwintiel=1st quintile (lowest prosperity), REGVrlPercWelvaartTminEen=2nd decile, REGVrlPercWelvaartTminEen=8th decile, AFLWelvaartsKwintiel=4th quintile, REGVrlPercWelvaartTminEen=7th decile, REGVrlPercVermogenTminEen=2nd decile). Definitions of the variable labels are provided in Table D.1.

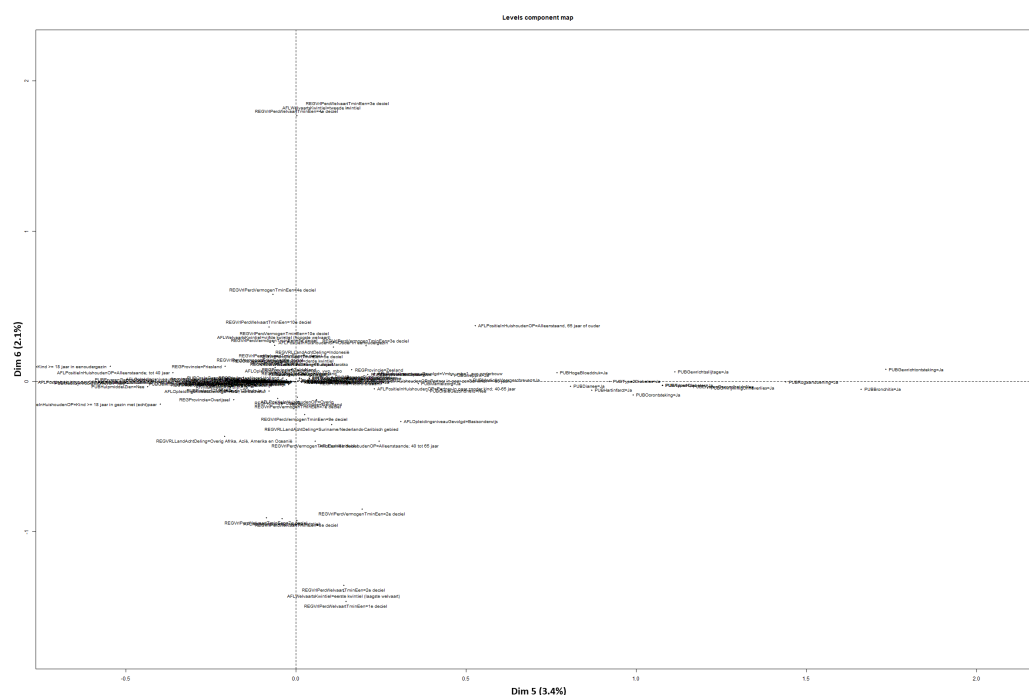


Figure G.24 Projection of categorical variable levels onto dimensions 6 and 7. For interpretability, emphasis is placed on the levels with the strongest contributions to these dimensions. (dimension 6, positive side: REGVrIPercWelvaartTminEen=3rd decile, AFLWelvaartsKwintiel=2nd quintile, REGVrIPercWelvaartTminEen=4th decile, REGVrIPercVermogenTminEen=4th decile, negative side: REGVrIPercWelvaartTminEen=1st decile, AFLWelvaartsKwintiel=1st quintile (lowest prosperity), REGVrIPercWelvaartTminEen=2nd decile, REGVrIPercWelvaartTminEen=8th decile, AFLWelvaartsKwintiel=4th quintile, REGVrIPercWelvaartTminEen=7th decile, REGVrIPercVermogenTminEen=2nd decile; dimension 7, positive side: REGVrIPercWelvaartTminEen=5th decile, AFLWelvaartsKwintiel=3rd quintile, REGVrIPercWelvaartTminEen=6th decile, REGVrIPercWelvaartTminEen=1st decile, AFLWelvaartsKwintiel=1st quintile (lowest prosperity), REGVrIPercWelvaartTminEen=2nd decile, negative side: REGVrIPercWelvaartTminEen=8th decile, AFLWelvaartsKwintiel=4th quintile, REGVrIPercWelvaartTminEen=7th decile, REGVrIPercWelvaartTminEen=3rd decile, AFLWelvaartsKwintiel=2nd quintile, REGVrIPercWelvaartTminEen=4th decile). Definitions of the variable labels are provided in Table D.1.

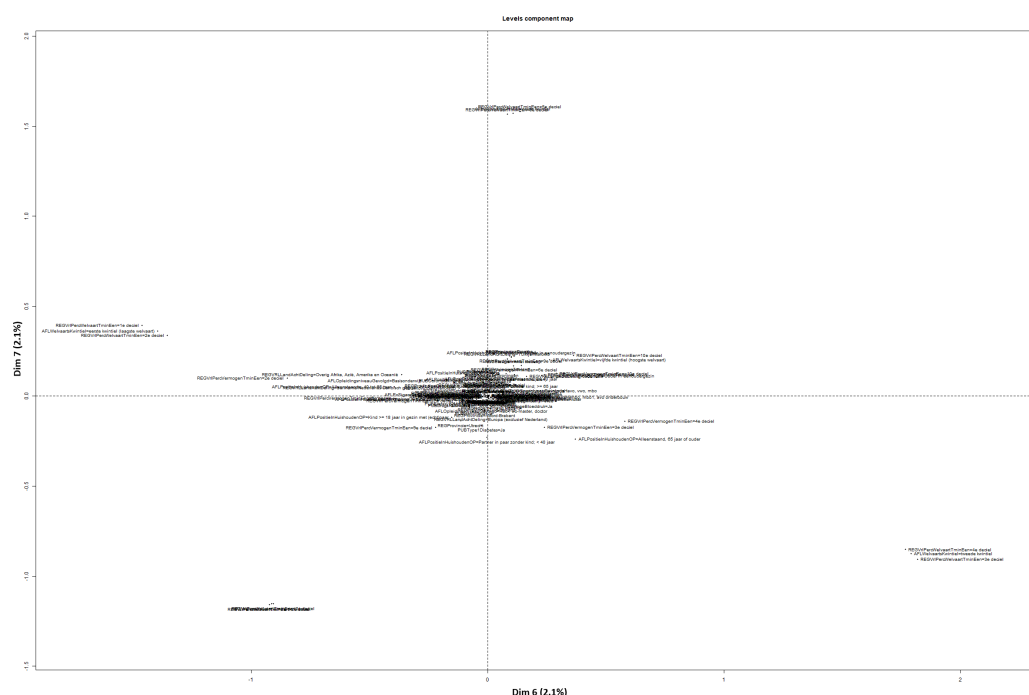
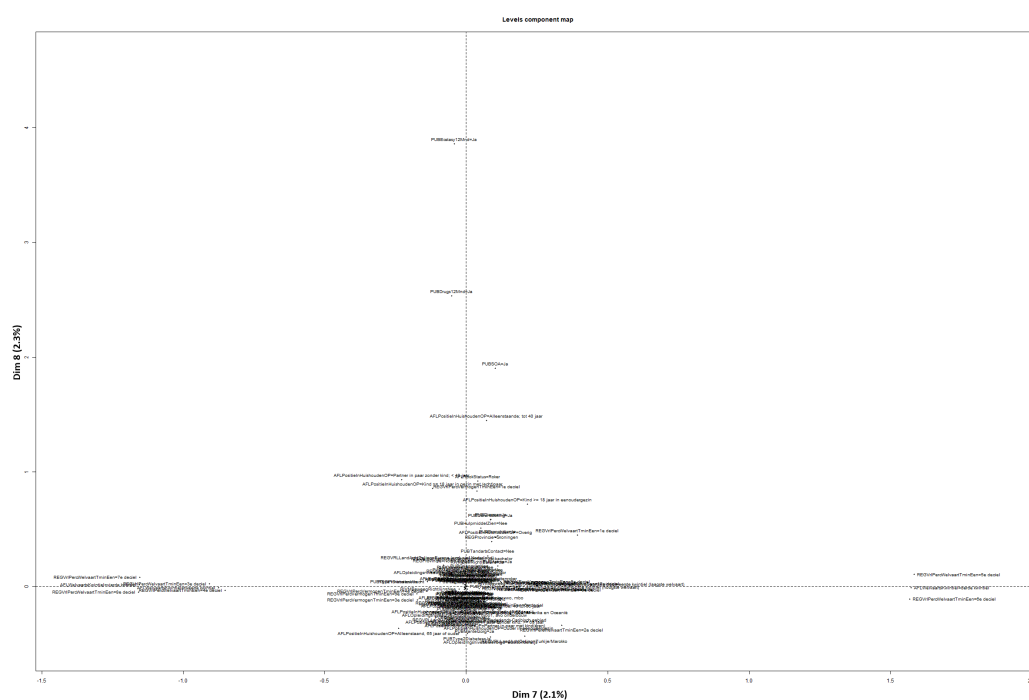
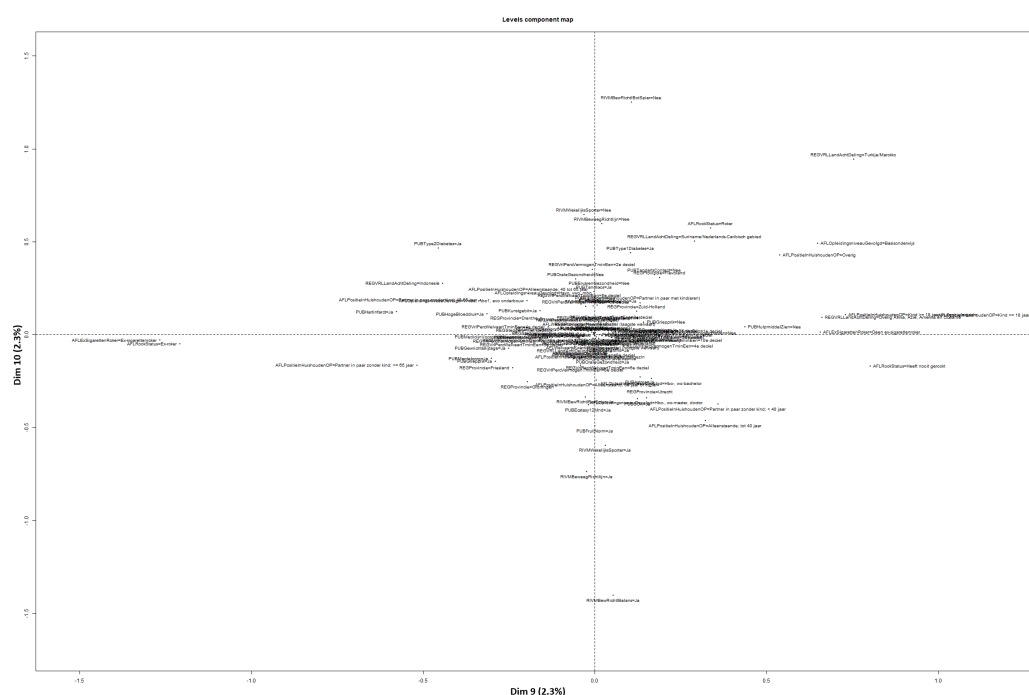


Figure G.25 Projection of categorical variable levels onto dimensions 7 and 8. For interpretability, emphasis is placed on the levels with the strongest contributions to these dimensions. (dimension 7, positive side: REGVrIPercWelvaartTminEen=5th decile, AFLWelvaartsKwintiel=3rd quintile, REGVrIPercWelvaartTminEen=6th decile, REGVrIPercWelvaartTminEen=1st decile, AFLWelvaartsKwintiel=1st quintile (lowest prosperity), REGVrIPercWelvaartTminEen=2nd decile, negative side: REGVrIPercWelvaartTminEen=8th decile, AFLWelvaartsKwintiel=4th quintile, REGVrIPercWelvaartTminEen=7th decile, REGVrIPercWelvaartTminEen=3rd decile, AFLWelvaartsKwintiel=2nd quintile, REGVrIPercWelvaartTminEen=4th decile; dimension 8, positive side: PUBEcstasy12Mnd=Yes, PUBDrugs12Mnd=Yes, PUBSOA=Yes, AFLPositieInHuishoudenOP=Single up to 40 years, AFLPositieInHuishoudenOP=Partner in couple, no child up to 40 years, AFLRookStatus=Smoker). Definitions of the variable labels are provided in Table D.1.



[illegible]

Figure G.27 Projection of categorical variable levels onto dimensions 9 and 10. For interpretability, emphasis is placed on the levels with the strongest contributions to these dimensions. (dimension 9, positive side: AFLPositieInHuishoudenOP=Child $\geq$ 18 years in household with (married) couple, AFLRookStatus=Never smoker, AFLPositieInHuishoudenOP=Child $\geq$ 18 years in one-parent household, REGVRLandAchtDeling=Other Africa, Asia, America and Oceania, AFLExSigarettenRoker=Never cigarette smoker, negative side: AFLExSigarettenRoker=Former cigarette smoker, AFLRookStatus=Former smoker PUBHartinfarct=Yes, AFLPositieInHuishoudenOP=Partner in couple, no child $\geq$ 65 years; dimension 10, positive side: RIVMBewrichtBotSpier=No, REGVRLandAchtDeling=Turkey/Morocco, RIVMWekelijksSporter=No, RIVMBeweegRichtlijn=No, AFLRookStatus=Smoker, negative side: RIVMBewrichtBalans=Yes, RIVMBeweegRichtlijn=Yes, RIVMWekelijksSporter=Yes). Definitions of the variable labels are provided in Table D.1.



H Tables with a description of the results for each BMI category separately.

Table H.1 Description of the respondent clusters for the overweight group
(*N* = 5,395)

	Description clusters by variables related to BMI	Main characteristics
Cluster O1	Tend to have fewer teeth; tend to have dentures more frequently; tend to be toothless more frequently; more frequently have no contact with a dentist. Overrepresented household positions include partners without children aged > 65 years and single individuals aged > 65 years.	N =480; Mean BMI: 27.3 (SD 1.4); Mean age: 73.8 (SD 10.4); Good or very good perceived health: 65.0%; Male: 54.2%.
Cluster O2	Low scores on SF-12 physical and mental health. More frequent chronic conditions include high blood pressure, ear infection, type 1 and type 2 diabetes, asthma, bronchitis, diarrhea, joint wear, joint inflammation, heart attack, back disorders, and involuntary urine loss.	N =383; Mean BMI: 27.5 (SD 1.4); Mean age: 64.2 (SD 14.3); Good or very good perceived health: 6.0%; Male: 44.4%.
Cluster O3	Overrepresented in prosperity and wealth between the middle and upper class (7th and 8th deciles).	N =1,030; Mean BMI: 27.0 (SD 1.4); Mean age: 53.3 (SD 15.6); Good or very good perceived health: 86.8%; Male: 57.3%.
Cluster O4	Overrepresented in the upper prosperity and wealth class (9th and 10th deciles). More frequently former smokers, particularly cigarette smokers.	N =476; Mean BMI: 27.0 (SD 1.4); Mean age: 60.6 (SD 12.2); Good or very good perceived health: 81.7%; Male: 67.4%.
Cluster O5	Overrepresented in prosperity and wealth categories (3rd and 4th deciles). Overrepresented household positions include child ≥ 18 y in a single-parent households, parents in single-parent households, and single individuals aged < 40 years.	N =588; Mean BMI: 27.2 (SD 1.4); Mean age: 51.2 (SD 17.9); Good or very good perceived health: 73.8%; Male: 50.0%.
Cluster O6	Overrepresented in the upper prosperity and wealth categories (9th and 10th deciles). More frequently never smokers. More frequently had a migration background from Turkey and Morocco. Overrepresented household positions include household position: child aged ≥ 18 years.	N =699; Mean BMI: 27.0 (SD 1.4); Mean age: 52.8 (SD 14.7); Good or very good perceived health: 87.4%; Male: 56.2%.
Cluster O7	More frequently live in very highly urbanized areas, particularly in the province of North Holland. More frequently have a Surinamese or Dutch Caribbean origin.	N =284; Mean BMI: 27.1 (SD 1.4); Mean age: 51.5 (SD 16.6); Good or very good perceived health: 71.8%; Male: 55.6%.
Cluster O8	Overrepresented in the middle prosperity and wealth class (5th and 6th deciles).	N =883; Mean BMI: 27.1 (SD 1.4); Mean age: 50.6 (SD 17.0); Good or very good perceived health: 85.1%; Male: 57.9%.
Cluster O9	Overrepresented in the lowest prosperity classes (1st and 2nd deciles; lowest 40%).	N =403; Mean BMI: 27.2 (SD 1.4); Mean age: 51.6 (SD 18.5); Good or very good perceived health: 55.3%; Male: 44.9%.
Cluster O10	More often tested for sexually transmitted infections. More frequent drug use, particularly ecstasy. Overrepresented household positions include couples without children and single individuals aged < 40 years. Tend to have lower SF-12 mental health scores.	N =169; Mean BMI: 26.9 (SD 1.4); Mean age: 32.1 (SD 9.5); Good or very good perceived health: 79.3%; Male: 69.2%.

Note: In this table, the plots used to describe the clusters are summarized to provide insight into the differences between the clusters. In this description, each cluster is contrasted with the rest of the population.

Table H.2 Description of the respondent clusters for the obese group ($N = 2,134$)

	Description clusters by variables related to BMI	Main characteristics
Cluster B1	Tend to have a low score on SF-12 physical health and mental health. More frequently suffer from chronic conditions such as joint disorders, urinary incontinence, asthma, joint wear, type 1 diabetes, diarrhea, ear infection, and high blood pressure. More frequently do not answer “good” or “very good” to the question “How do you assess your general state of health?” Overrepresented in the household position: single, 65 years and older.	N =157; Mean BMI: 33.7 (SD 2.6); Mean age: 64.1 (SD 12.6); Good or very good perceived health: 5.1%; Male: 23.6%.
Cluster B2	Live more frequently in very strongly urbanized areas. More frequently located in the provinces of North Holland and South Holland. More frequently have Turkish or Moroccan origin.	N =130; Mean BMI: 32.8 (SD 2.4); Mean age: 50.4 (SD 16.2); Good or very good perceived health: 60.0%; Male: 46.9%.
Cluster B3	Consume drugs more frequently, especially ecstasy. More frequently tested for sexually transmitted infections (STIs). Overrepresented in household positions: couples without children and single individuals under 40 years old.	N =42; Mean BMI: 33.1 (SD 2.2); Mean age: 33.0 (SD 9.6); Good or very good perceived health: 81.0%; Male: 64.3%.
Cluster B4	Tend to have fewer teeth, more frequently have dentures, and are more frequently toothless. More frequently have no contact with a dentist. Overrepresented household positions include single individuals aged over 65 years and partners in couples without children aged over 65 years. More frequently have reported a heart attack. Overrepresented with respect to education level: only primary education. Overrepresented in prosperity and wealth categories: 1st, 2nd, 3rd, and 4th decile.	N =86; Mean BMI: 32.8 (SD 2.2); Mean age: 70.7 (SD 10.6); Good or very good perceived health: 43.0%; Male: 36.1%.
Cluster B5	Overrepresented in the lowest prosperity and wealth categories (1st, 2nd, 3rd, and 4th decile; lowest 40%).	N =163; Mean BMI: 33.5 (SD 2.5); Mean age: 48.7 (SD 16.8); Good or very good perceived health: 41.1%; Male: 39.3%.
Cluster B6	Overrepresented in the higher prosperity categories (7th and 8th decile).	N =392; Mean BMI: 32.8 (SD 2.4); Mean age: 51.1 (SD 14.6); Good or very good perceived health: 73.0%; Male: 48.5%.
Cluster B7	Overrepresented in the highest prosperity and wealth categories (9th and 10th decile).	N =345; Mean BMI: 32.7 (SD 2.3); Mean age: 55.7 (SD 13.0); Good or very good perceived health: 68.1%; Male: 48.1%.
Cluster B8	Tend to have fewer teeth, more frequently have dentures, and are more frequently toothless. More frequently have no contact with a dentist. Overrepresented household positions include single individuals aged over 65 years and partners in couples without children aged over 65 years. More frequently reported a heart attack. Overrepresented with respect to education level: only primary education.	N =205; Mean BMI: 32.8 (SD 2.2); Mean age: 71.0 (SD 9.8); Good or very good perceived health: 44.9%; Male: 50.2%.
Cluster B9	Overrepresented in prosperity and wealth categories: 3rd and 4th decile. Overrepresented household positions include singles aged 40 years and older, single-parent households with children aged 18 years or older, and single-parent households. More frequently individuals with only primary education. More frequently suffer from chronic conditions such as type 1 diabetes and asthma. More frequently located in the provinces of Friesland and Flevoland. More frequently have Surinamese or Dutch Caribbean origin.	N =268; Mean BMI: 32.9 (SD 2.4); Mean age: 52.2 (SD 17.3); Good or very good perceived health: 67.5%; Male: 39.2%.
Cluster B10	More frequently middle class in terms of prosperity (5th and 6th decile).	N =346; Mean BMI: 32.9 (SD 2.4); Mean age: 50.2 (SD 15.3); Good or very good perceived health: 73.1%; Male: 48.8%.

Note: In this table, the plots used to describe the clusters are summarized to provide insight into the differences between the clusters. In this description, each cluster is contrasted with the rest of the population.

CBS | Discussion paper | Oktober 2026 71

Table H.3 Description of the respondent clusters for the morbid obesity group
(*N* = 181)

	Description clusters by variables related to BMI	Main characteristics
Cluster M1	Tend to have more teeth. Overrepresented in the household position: parent in a single-parent household. More frequently located in the province of Groningen. More frequently have origin from Turkey, Morocco, Surinamese, or Caribbean regions. Consume drugs more frequently. More frequently suffer from type 1 diabetes. Overrepresented in the lowest prosperity categories (1st and 2nd decile) and lowest wealth category (1st decile).	N =21; Mean BMI: 45.8 (SD 6.8); Mean age: 36.8 (SD 12.7); Good or very good perceived health: 33.3%; Male: 28.6%.
Cluster M2	Live more frequently in a very strongly urbanized area. Overrepresented in the 5th quintile of prosperity and the 10th decile of wealth.	N =4; Mean BMI: 44.0 (SD 2.7); Mean age: 48.0 (SD 5.2); Good or very good perceived health: 0.0%; Male: 75.0%.
Cluster M3	Tend to have a higher SF-12 physical health score. Overrepresented in the highest wealth categories (9th and 10th decile). More frequently have origin from Africa, the Americas, Asia, or Oceania.	N =23; Mean BMI: 43.4 (SD 3.9); Mean age: 51.5 (SD 14.1); Good or very good perceived health: 34.8%; Male: 34.8%.
Cluster M4	Tend to have fewer teeth, more frequently have dentures, and are more frequently toothless. Also overrepresented in household positions with a partner aged over 65 years.	N =25; Mean BMI: 45.8 (SD 6.6); Mean age: 67.0 (SD 13.1); Good or very good perceived health: 48.0%; Male: 28.0%.
Cluster M5	More frequently middle class in terms of prosperity (6th and 7th deciles). Overrepresented household positions include children aged 18 years or older in single-parent or two-parent households. Overrepresented among weekly sports participants and more frequently meet physical activity guidelines.	N =32; Mean BMI: 43.4 (SD 2.2); Mean age: 42.6 (SD 12.2); Good or very good perceived health: 65.6%; Male: 53.1%.
Cluster M6	Tend to have a low score on SF-12 physical health. More frequently suffer from back disorders, joint inflammation, type 2 diabetes, joint wear, and bronchitis. More frequently reported a heart attack. More frequently located in the province of Utrecht.	N =24; Mean BMI: 45.4 (SD 7.8); Mean age: 56.8 (SD 10.2); Good or very good perceived health: 0.0%; Male: 29.2%.
Cluster M7	Overrepresented in prosperity and wealth categories (3rd and 4th decile).	N =21; Mean BMI: 44.3 (SD 3.4); Mean age: 47.8 (SD 16.0); Good or very good perceived health: 57.1%; Male: 38.1%.
Cluster M8	More frequently located in the province of Drenthe. Overrepresented in prosperity and wealth categories between the middle and upper class (7th and 8th decile). More frequently meet the physical activity guidelines.	N =16; Mean BMI: 43.2 (SD 2.5); Mean age: 69.4 (SD 6.9); Good or very good perceived health: 18.8%; Male: 25.0%.
Cluster M9	Tend to have a poor oral health. Overrepresented in household positions as partnered with children. Tend not to use visual aids. More frequently live in urbanized areas.	N =15; Mean BMI: 45.5 (SD 5.6); Mean age: 44.4 (SD 15.4); Good or very good perceived health: 33.3%; Male: 40.0%.

Note: In this table, the plots used to describe the clusters are summarized to provide insight into the differences between the clusters. In this description, each cluster is contrasted with the rest of the population.

Colophon

Publisher

Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague, Netherlands
www.cbs.nl

Prepress

Statistics Netherlands Grafimedia

Design

Edenspiekermann

Information

Telephone +31 88 570 70 70
Via contact form: www.cbs.nl/information

© Statistics Netherlands, The Hague/Heerlen/Bonaire 2026.
Reproduction is permitted, provided Statistics Netherlands is quoted as the source