UNIVERSITÉ DE
NEUCHÂTEL

Institut de statistique

# Application on biochemical cluster detection

Master Thesis submitted in partial fulfillment of the requirements for the degree of

**Master of Science in Statistics**

**by**

**Denis Dériaz**

Thesis Director
**Dr Alina Matei**
University of Neuchâtel,
Faculty of Science
Institute of Statistics

**Neuchâtel, January 2019**

# Acknowledgments

It was a real pleasure to study Statistics at the University of Neuchâtel. All the teachers were kind and encouraging. These studies allowed me to have a deeper understanding of the biomedical research. Moreover, the R courses given by Dr Alina Matei developed my knowledge in R such that I currently couldn't imagine another software than R to do Statistics.

I would like also to thank Prof. Murielle Bochud at the Institute of Social and Preventive Medicine in Lausanne for giving her medical expertise and providing the medical data, and Claire Mudge for her English revision.

# Contents

# Chapter 1

# Introduction

*Cluster analysis is the art of finding groups in data.*

— Rousseeuw

Irregularities in blood pressure are related to the exposure of some low-molecular biochemical elements ($< 69$ kDa). For example, sodium, potassium, nitrates and some metal-mixtures are associated with high blood pressure in the general population [41, 25, 11, 33, 32, 7, 8, 3]. Excretion of those elements in human urine is a reflection of their total exposure. Recently, a study reported that environmental agents, such as ambient air pollution and metal-mixtures, are underappreciated as risk factors for cadiovascular diseases such as hypertension that is an abnormally high blood pressure [9]. Linear or logistic regressions are often used in order to determine the effect of biochemical elements on blood pressure [41, 25, 11, 33, 18]. However, multiple regression methods face three problems: 1) the tendency to make a true association statistically non-significant through multicollinearity; 2) the tendency to make a true association non-significant and smaller than it is through random measurement errors of covariates (regression dilution bias) as seen in the INTERSALT study where the effect of sodium on blood pressure was underestimated [16]; 3) difficulties in dissociating the effect of a predictor from that of a confounder. In order to account for this last point, supplementary variables and interactions are often added in the regression models, however the number of all possible interactions is often important and interpretation of interactions of the order of more than two is non-trivial. These three problems partially stem from the fact that the biochemical elements are intercorrelated. Firstly, correlations can be consequences of physiological processes. For example, potassium supplementation promotes sodium retention and potassium loss [4]. Secondly, dietary patterns can induce correlations between molecules. For example, both nitrate and potassium are known to decrease blood pressure in the general population and both come from a diet rich in fruits and vegetables [33].

Since clustering puts in evidence patterns of biochemical elements rather than elements taken alone, we think that clustering methods can mitigate these problems. For example, clustering is applied to metabolomics to learn how biochemical patterns from urine or blood samples can be associated with dietary patterns [27]. We think that clustering and dimentionality reduction have the potential to better predict hypertension in regression in addition to a better visual summary of the data.

To test these hypotheses, we will use the data from the Swiss Kidney Project on Genes in Hypertension (SKIPOGH) which is a Swiss study collecting medical data. Firstly, we will use principal component analysis (PCA) and cluster analysis to see if we can distinguish heterogenous groups of individuals in the data. The emerging principal components and biochemical clusters are expected to be meaningful in terms of physiological processes and dietary patterns. Then, we will see how these exploratory findings are associated (or not) with hypertension. We will perform regression both for hypertension status (binary variable) and blood pressure (continuous variable).

We will also construct regression models with observed covariates and compare their results with the methods mentioned above.

Note that all the analysis was conducted in `R version 3.5.0` (2018-04-23) using the packages `''FactoMineR''` version 1.41 for PCA and `''cluster''` version 2.0.7-1 for PCA and cluster analysis, respectively.

In the following chapter called *Context*, we present some biomedical notions that can be useful for the interpretation of the results. The subsequent chapters relate directly to the statistical analyses: *Data description and preprocessing*, *Methods* and *Results*. R-code and abbreviations glossary are shown in the *Appendix*.

# Chapter 2

# Context

## 2.1 An overview of hypertension

In 2005, Kearney et al. estimated the worldwide prevalence of hypertension [21]. They combined reported prevalences of hypertension in 31 articles published from 1980 to 2002. Using demographic data as auxiliary variables, the worldwide prevalence in 2000 was estimated at 26.4% (95% Confidence Interval, abbrev. 95% CI, was 26.0-26.8%) by age and sex using post-stratification. Using extrapolation from Taylor series approximation methods, the projected prevalence for 2025 was 29.2% (95% CI 28.8 – 29.7%) [21]. An article published in the medical journal *The Lancet* in 2002, concluded that hypertension is the leading risk factor for mortality and is ranked third as a cause of disability-adjusted life-years, which is a measure of morbidity [12]. Hypertension belongs to the Metabolic Syndrome which is defined by the authors of *Harrisons Principles of Internal medicine* as "a constellation of metabolic abnormalities that confer increased risk of cardiovascular disease and diabetes mellitus" (Table 2.1) [19]. Consequently, the exploration of risk factors for hypertension is a critical step in the prevension of hypertension, and more generally a public health issue.

## 2.2 Measuring and defining hypertension

The *sphygmomanometer* is an instrument which measures blood pressure expressed in *mmHg* (see Figure 2.1). The instrument gives two values: the *systolic pressure* (SBP) and the *diastolic pressure* (DBP) that are respectively the blood pressure during cardiac muscle contraction (systole) and the blood pressure during cardiac muscle relaxation (diastole). The SBP is always greater than the DBP. The bivariate result is conventionally presented under the form SBP/DBP. For example, an individual with SBP=130 and DBP=80 has a bivariate result of 130/80.

Hypertension is defined as a SBP greater than 140 mmHg, a DBP greater than 90 mmHg and/or use of antihypertensive medication [21, 22]. This accepted definition allows for comparisons between different studies.

So, hypertension can be expressed either as a binary variable using the definition above (presence of hypertension or absence of hypertension) or as a continuous variable (values of SBP/DBP). Thereby, a wide range of different statistical models can be used to model hypertension. The binary variable contains much less information than the continuous variable because of dichotomization. Consequently, the dichotomized variable can loose information useful in statistical analyses. On the other hand, the binary variable is often more meaningful from a clinical point of view (increased cardiovascular risk and decision boundary for treatment).

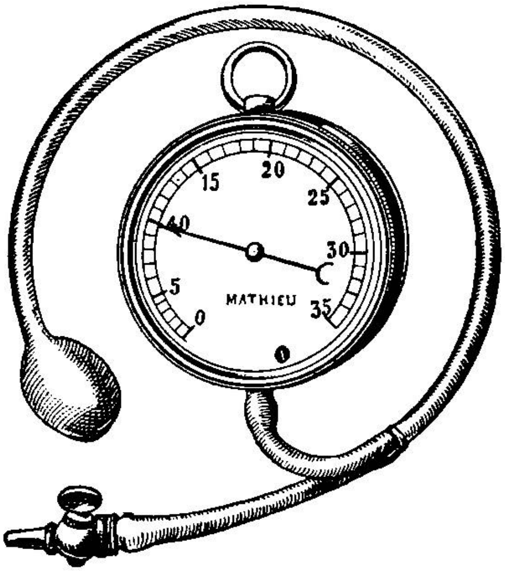


FIG. 1021.
Sphygmomanomètre du D$^{r}$ Potain.

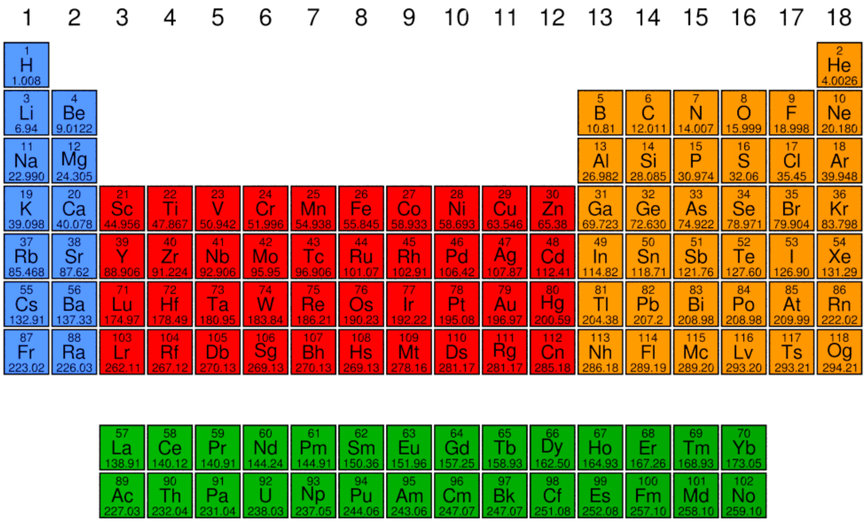
**Figure 2.1** Historical sphygmomanometer.

*Larousse médical illustré*, 1912. Ed. Librairie Larousse, Paris [14].

**Table 2.1:** Definition Criteria for the Metabolic Syndrome [19].

**Three or more of the following:**

- Central obesity: waist circumference $>$ 102 cm (M), $>$ 88 cm (F)
- Hypertriglyceridemia: triglyceride level $\geq$ 150 mg/dL or specific medication
- Low HDLc cholesterol: $<$ 40 mg/dL and $<$ 50 mg/dL for men and women, respectively, or specific medication
- Hypertension: blood pressure $\geq$ 130 mmHg systolic or $\geq$ 85 mmHg diastolic or specific medication
- Fasting plasma glucose level $\geq$ 100 mg/dL or specific medication or previously diagnosed type 2 diabetes



| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1<br>H<br>1.008 | | | | | | | | | | | | | | | | | 2<br>He<br>4.0026 |
| 3<br>Li<br>6.94 | 4<br>Be<br>9.0122 | | | | | | | | | | | 5<br>B<br>10.81 | 6<br>C<br>12.011 | 7<br>N<br>14.007 | 8<br>O<br>15.999 | 9<br>F<br>18.998 | 10<br>Ne<br>20.180 |
| 11<br>Na<br>22.990 | 12<br>Mg<br>24.305 | | | | | | | | | | | 13<br>Al<br>26.982 | 14<br>Si<br>28.085 | 15<br>P<br>30.974 | 16<br>S<br>32.06 | 17<br>Cl<br>35.45 | 18<br>Ar<br>39.948 |
| 19<br>K<br>39.098 | 20<br>Ca<br>40.078 | 21<br>Sc<br>44.956 | 22<br>Ti<br>47.867 | 23<br>V<br>50.942 | 24<br>Cr<br>51.996 | 25<br>Mn<br>54.938 | 26<br>Fe<br>55.845 | 27<br>Co<br>58.933 | 28<br>Ni<br>58.693 | 29<br>Cu<br>63.546 | 30<br>Zn<br>65.38 | 31<br>Ga<br>69.723 | 32<br>Ge<br>72.630 | 33<br>As<br>74.922 | 34<br>Se<br>78.971 | 35<br>Br<br>79.904 | 36<br>Kr<br>83.798 |
| 37<br>Rb<br>85.468 | 38<br>Sr<br>87.62 | 39<br>Y<br>88.906 | 40<br>Zr<br>91.224 | 41<br>Nb<br>92.906 | 42<br>Mo<br>95.95 | 43<br>Tc<br>96.906 | 44<br>Ru<br>101.07 | 45<br>Rh<br>102.91 | 46<br>Pd<br>106.42 | 47<br>Ag<br>107.87 | 48<br>Cd<br>112.41 | 49<br>In<br>114.82 | 50<br>Sn<br>118.71 | 51<br>Sb<br>121.76 | 52<br>Te<br>127.60 | 53<br>I<br>126.90 | 54<br>Xe<br>131.29 |
| 55<br>Cs<br>132.91 | 56<br>Ba<br>137.33 | 71<br>Lu<br>174.97 | 72<br>Hf<br>178.49 | 73<br>Ta<br>180.95 | 74<br>W<br>183.84 | 75<br>Re<br>186.21 | 76<br>Os<br>190.23 | 77<br>Ir<br>192.22 | 78<br>Pt<br>195.08 | 79<br>Au<br>196.97 | 80<br>Hg<br>200.59 | 81<br>Tl<br>204.38 | 82<br>Pb<br>207.2 | 83<br>Bi<br>208.98 | 84<br>Po<br>208.98 | 85<br>At<br>209.99 | 86<br>Rn<br>222.02 |
| 87<br>Fr<br>223.02 | 88<br>Ra<br>226.03 | 103<br>Lr<br>262.11 | 104<br>Rf<br>267.12 | 105<br>Db<br>270.13 | 106<br>Sg<br>269.13 | 107<br>Bh<br>270.13 | 108<br>Hs<br>269.13 | 109<br>Mt<br>278.16 | 110<br>Ds<br>281.17 | 111<br>Rg<br>281.17 | 112<br>Cn<br>285.18 | 113<br>Nh<br>286.18 | 114<br>Fl<br>289.19 | 115<br>Mc<br>289.20 | 116<br>Lv<br>293.20 | 117<br>Ts<br>293.21 | 118<br>Og<br>294.21 |
| | | 57<br>La<br>138.91 | 58<br>Ce<br>140.12 | 59<br>Pr<br>140.91 | 60<br>Nd<br>144.24 | 61<br>Pm<br>144.91 | 62<br>Sm<br>150.36 | 63<br>Eu<br>151.96 | 64<br>Gd<br>157.25 | 65<br>Tb<br>158.93 | 66<br>Dy<br>162.50 | 67<br>Ho<br>164.93 | 68<br>Er<br>167.26 | 69<br>Tm<br>168.93 | 70<br>Yb<br>173.05 | | |
| | | 89<br>Ac<br>227.03 | 90<br>Th<br>232.04 | 91<br>Pa<br>231.04 | 92<br>U<br>238.03 | 93<br>Np<br>237.05 | 94<br>Pu<br>244.06 | 95<br>Am<br>243.06 | 96<br>Cm<br>247.07 | 97<br>Bk<br>247.07 | 98<br>Cf<br>251.08 | 99<br>Es<br>252.08 | 100<br>Fm<br>257.10 | 101<br>Md<br>258.10 | 102<br>No<br>259.10 | | |

**Figure 2.2** Periodic table of elements.

Metals are in the blue, red, green and in the lower left orange groups. Metalloids are the elements in the diagonal of the orange square, i.e. Boron, Silicon, Germanium, Arsenic, Antimony, Tellurium.

*Source: www.webelements.com*

## 2.3 Biochemical elements

For the purpose of this study, we are interested in two families of potential risk factors for hypertension: the metal-mixtures and the non-metals. Both families are elements coming from alimentation and/or environmental exposure, and are described in detailed in this section.

### 2.3.1 Metallome

According to the definition by Joanna Szpunar, the *metallome* is "*the entirety of metal and metalloid species* [1] *within a cell or tissue type*" [36]. Although metals and metalloids are toxic for the human body in many cases, they are necessary for physiological processes in some cases. A classical exemple of an essential metal is iron whose deficiency leads to anemia in 2% of the adult population [23]. Exposure can come from water, food, air, medications and even dietary supplements.

Metals are present in the left part of the *periodic table of elements* (see Figure 2.2). They share sharing some characteristics that distinguish them from other elements, such as the as ability to conduct electricity, a metallic shine, malleability and ductility. These four characteristics result from the high mobility of electrons in metals [6].

Some chemical elements between the metals and the nonmetals in the periodic table are classed as metalloids. They have some — but not all — metallic characteristics.

Table 2.2 presents some of the metals and metalloids most relevant to public health.

[1] metal-mixtures: metals and metalloids

**Table 2.2:** Description of metal-mixtures of medical interest.

| Element,<br>*chemical group* | Chemical symbol | Essential versus toxic | Medical interest |
|---|---|---|---|
| Lithium,<br>*metal* | Li | Toxic | - Common drug for bipolar disorder<br>- Renal toxicity and cardiac conduction disorders in high doses |
| Beryllium,<br>*metal* | Be | Toxic | - Professional exposure: processing alloys for high-tech industries<br>- Lung diseases such as chronic granulomatous disease |
| Aluminum,<br>*metal* | Al | Toxic | - Aluminum oxyde hydrate as treatment for dyspepsia (historically)<br>- Nowadays avoided because of the risk of acute dementia and bone fractures<br>- Other sources are food (as preservative and coloring agent) and professional exposure but their contribution in total aluminum exposure is minor [34] |
| Vanadium,<br>*metal* | V | Toxic | - Deficiency could cause metabolic disorders [28]<br>- New candidate drug for diabetes (animal models) [37]<br>- Could be considered as an essential metal-mixture [40] |
| Chromium,<br>*metal* | Cr | Toxic | - Presence in food: yeast, meat, and grain products<br>- Causing lung cancer<br>- Professional exposure (bricklayers): presence in cement causing skin allergy |
| Manganese<br>*metal* | Mn | Essential | - Involved in human enzymes (like manganese superoxide dismutase)<br>- Deficiency causing bone metabolism perturbations |
| Cobalt,<br>*metal* | Co | Essential | - Component of the cobalamine<br>(Vitamin B12)<br>- Vitamin B12 involved in DNA synthesis |
| Nickel,<br>*metal* | Ni | Toxic | - Professional exposure<br>- Skin allergy<br>- Potential carcinogen [19] |
| Copper,<br>*metal* | Cu | Essential | - Presence in many human enzymes<br>- Involved in iron metabolism, energy production and neurotransmitter<br>synthesis [19]<br>- Kidney and liver failure in case of intoxication |
| Zinc,<br>*metal* | Zn | Essential | - Present in many human enzymes<br>- Necessary for fetal growth and embryonic development |
| Arsenic,<br>*metalloid* | As | Toxic | - Historical poison<br>- Professional exposure<br>- Water contamination<br>- Interferes with energy production<br>- Presence in some drugs against leukemia<br>- Can lead to death by organ toxicity and fluid loss [19] |
| Molybdenum,<br>*metal* | Mo | Essential | - Present in human enzymes (sulfite and xanthine oxidase)<br>- Sometimes found in multimineral tablets |

*Continued on next page*

Table 2.3 – *Continued from previous page*

| Element,<br>*chemical group* | Chemical symbol | Essential versus toxic | Medical interest |
|---|---|---|---|
| Palladium,<br>*metal* | Pd | Toxic | - Sources: jewellery and dental restorations<br>- Contact allergy at sites of piercing or oral symptoms if dental restorations [13] |
| Silver,<br>*metal* | Ag | Toxic | |
| Cadmium,<br>*metal* | Cd | Toxic | - Professional exposure (smelting, battery and plastics industries)<br>- Batteries and tobacco |
| Tin,<br>*metal* | Sn | Toxic | |
| Antimony,<br>*metalloid* | Sb | Toxic | |
| Platinum,<br>*metal* | Pt | Toxic | - Genotoxic (study on cells)<br>- Augmentation of platinum group elements (PGE) during the last decades (combustion catalyst for cars)<br>- Other sources: industries using PGE, hospitals and dental laboratories [39] |
| Mercury,<br>*metal* | Hg | Toxic | - Professional exposure (ex. automobile and construction) |
| Thallium,<br>*metal* | Tl | Toxic | - Insecticides, metal alloys and fireworks<br>- Severe poisoning can follow a single dose $> 1$ g |
| Lead,<br>*metal* | Pb | Toxic | - Professional and domestic exposure<br>- Food or water from improperly glazed ceramics |
| Bismuth,<br>*metal* | Bi | Toxic | - Prophylaxis of travelers' diarrhea (approx. 60% effective)<br>- Effective drug for helicobacter pylori infection |

Remark: Of the 22 metal-mixtures, 5 can be considered essential metal-mixtures: manganese, cobalt, copper, zinc and molybdenum.
The main reference for this table is Kasper et al. 2015 [19].

### 2.3.2 Non-metals

In this work, *non-metals* refers simply to biochemical elements that are not metal-mixtures. They include:

- Electrolytes: as defined by the second edition of *Oxford Dictionary of Public Health*, an electrolyte is *"an ion or compound derived from an element, such as sodium, potassium, or carbon, that when dissolved or suspended in a liquid medium transmits an electric current and is deposited on the positive or negative electrode, depending on whether the electrolyte is negatively or positively charged"*[24].

- Metabolites: products of metabolism (energy production).

# Chapter 3

# Data description and preprocessing

## 3.1 Data collection

The main objective of SKIPOGH is to examine the genetic and environmental determinants of hypertension. SKIPOGH is a cross-sectional substudy of a larger European study (EPOGH: European Project on Genes in Hypertension). It is a multi-center study with participants being recruited in the cantons of Bern, Geneva, and Vaud. However, we will use only the SKIPOGH data of the participants recruited in Lausanne. The study design was a simple random sampling without replacement (SRSWOR) from the city of Lausanne's registry. Selected participants where then contacted by mail. The participation rate was 20% in Lausanne. The examination started in December 2009 and ended in April 2012. Inclusion criteria were: a) written informed consent; b) minimum age of 18 years; c) of European descent (defined as having both parents and grandparents born in a restricted list of countries) and d) at least one, and ideally three, first degree family members also willing to participate in the study. Thus, SKIPOGH is a family and individual sample from the city of Lausanne [5].

### 3.1.1 Medical questionnaire, clinical and biological data

Participants were asked to complete a questionnaire at home about their lifestyle habits, medical history and current medication. In particular, they were asked about their age, sex and eventual anti-hypertensive and anti-diabetic treatment.

The participants came to the University Hospital from Lausanne for a clinical examination, after an overnight fast. During the visit, participants' blood pressure values (both SBP and DBP) were measured 5 times every 5 minutes with a sphygmomanometer (device: A&D UM-101; A&D Company, Toshima Ku, Tokyo, Japan). Since the first blood pressure values are often known to be higher than subsequent values due to *white coat hypertension*[1], the mean of both the last four SBP and the last four DBP values were computed. These two mean values are respectively the variables `sbp_mean` and `dbp_mean` in the later described *urine* data that we used. Moreover, participants had to wear a 24-hour ambulatory blood pressure monitoring device on a day chosen for typical weekly activity. This device records blood pressure every 10 minutes during the day and every 30 minutes during the night. The variables for the 24-hour ambulatory blood pressure mean values are `sbp_24` and `dbp_24` for SBP and DBP, respectively. This last method is able to detect white coat hypertension better than blood pressure measured during the visit. Moreover,

[1] White coat hypertension occurs when blood pressure in a office or hospital is higher than blood pressure at home. It is a false positive case of hypertension sometimes believed to be related to the stress a patient feels at the beginning of a doctor's visit [19].

this method has shown to be a much stronger predictor of cardiovascular morbidity and mortality than visit measurements [26].

Biological data (venous blood samples and urine samples) were obtained during the visit. Venous blood glucose, lipid profile and renal function tests as well as serum and urinary electrolytes were analyzed by standard clinical methods. Participants were also asked to collect a 24-hour urine sample for the measurement of urinary volume and concentrations of biochemical elements [5].

24-hour urine excretion ($\hat{Q}$) of a biochemical element was estimated by multiplying its sample concentration ($C$) by 24-hour urine volume ($V$) :

$$\hat{Q} = C \cdot V.$$

### 3.1.2 Defining hypertension

The variables in *urine* data related to hypertension are defined below:

- `sbp_mean`: the mean of the last four measurements of office SBP,
- `dbp_mean`: the mean of the last four measurements of office DBP,
- `d_hta`: treated for hypertension (1: yes, 0: no).

According to the definition of hypertension, a binary variable considering both measured blood pressure values with reference to threshold values and anti-hypertensive treatment was generated from the following pseudo-code where $n$ indicates the number of observations [21, 22]:

- `hta`: hypertension (1: yes, 0: no)

```
# generating hta
for(i in 1:length(hta))
  if(is.na(d_hta[i])==F & is.na(sbp_mean[i])==F
     & is.na(dbp_mean[i])==F)
    { hta[i]<-0
    if(d_hta[i]=="Yes" | sbp_mean[i]>140| dbp_mean[i])>90)
      hta[i]<-1
    }
```

## 3.2 Description of urine data

*Urine* data is a data frame with $n = 1254$ observations and $p = 40$ variables. A description of the data is summarized in Table 3.1.

**Table 3.1:** Description of *urine* data ($n = 1254$, $p = 40$).

| Variable | Type of variable | Taken values (range or levels) | Count of missing values |
|---|---|---|---|
| Participant number | Ordinal | $[11, \ldots, 775]$ | 122 |
| Age in years | Continuous | $[18.0 - 90.0]$ | 126 |
| Sex | Binary | “F” for female and “M” for male | 126 |

*Continued on next page*

Table 3.2 – *Continued from previous page*

| Variable | Type of variable | Taken values (range or levels) | Count of missing values |
|---|---|---|---|
| Diabetes | Binary | 1 for presence and 0 for absence | 126 |
| Anti-hypertensive treatment | Binary | 1 for presence and 0 for absence | 129 |
| Triglycerides in mmol/l | Continuous | $[0.07 - 6.89]$ | 135 |
| Cholesterol in mmol/l | Continuous | $[2.50 - 8.29]$ | 136 |
| Blood glucose in mmol/l | Continuous | $[2.50 - 8.29]$ | 134 |
| Blood insulin in $\mu IU/ml$ | Continuous | $[1 - 6150]$ | 160 |
| Lithium 24h urine excretion in ng | Continuous | $[5771.841 - 1,031,290]$ | 410 |
| Beryllium 24h urine excretion in ng | Continuous | $[0.091 - 105.759]$ | 410 |
| Aluminum 24h urine excretion in ng | Continuous | $[378.061 - 1371,304]$ | 410 |
| Vanadium 24h urine excretion in ng | Continuous | $[120.655 - 7945.755]$ | 410 |
| Chrome 24h urine excretion in ng | Continuous | $[282.013 - 32518.580]$ | 410 |
| Manganese 24h urine excretion in ng | Continuous | $[13.632 - 16948.660]$ | 410 |
| Cobalt 24h urine excretion in ng | Continuous | $[41.007 - 78001.220]$ | 410 |
| Nickel 24h urine excretion in ng | Continuous | $[122.465 - 51160.050]$ | 410 |
| Copper 24h urine excretion in ng | Continuous | $[3680.734 - 1184042]$ | 410 |
| Zinc 24h urine excretion in ng | Continuous | $[44724.680 - 3018641]$ | 410 |
| Arsenic 24h urine excretion in ng | Continuous | $[1683.906 - 3780762]$ | 410 |
| Molybdenum 24h urine excretion in ng | Continuous | $[4487.893 - 736741.300]$ | 410 |
| Palladium 24h urine excretion in ng | Continuous | $[11.671 - 5907.260]$ | 410 |
| Silver 24h urine excretion in ng | Continuous | $[7.395 - 6960.549]$ | 410 |
| Cadmium 24h urine excretion in ng | Continuous | $[58.527 - 3050.796]$ | 410 |
| Tin 24h urine excretion in ng | Continuous | $[57.975 - 938908]$ | 410 |
| Antimony 24h urine excretion in ng | Continuous | $[4.817 - 11076.800]$ | 410 |
| Platinum 24h urine excretion in ng | Continuous | $[3.577 - 74956.460]$ | 410 |
| Mercury 24h urine excretion in ng | Continuous | $[14.910 - 5098.182]$ | 410 |
| Thallium 24h urine excretion in ng | Continuous | $[52.602 - 36763.830]$ | 410 |

*Continued on next page*

Table 3.2 – *Continued from previous page*

| Variable | Type of variable | Taken values (range or levels) | Count of missing values |
|---|---|---|---|
| Lead 24h urine excretion in ng | Continuous | [84.359 – 110999.700] | 410 |
| Bismuth 24h urine excretion in ng | Continuous | [0.375 – 2109.367] | 410 |
| Sodium 24h urine excretion in mmol | Continuous | [10.500 – 410.519] | 152 |
| Potassium 24h urine excretion in mmol | Continuous | [14.853 – 156.070] | 152 |
| Calcium 24h urine excretion in mmol | Continuous | [0.048 – 13.807] | 376 |
| Phosphate 24h urine excretion in mmol | Continuous | [3.000 – 76.370] | 170 |
| Urea 24h urine excretion in mmol | Continuous | [64.218 – 742.300] | 157 |
| Magnesium 24h urine excretion in mmol | Continuous | [0.448 – 10.752] | 158 |
| Systolic blood pressure in mmHg | Continuous | [79.500 – 200.500] | 130 |
| Diastolic blood pressure in mmHg | Continuous | [50.000 – 110.500] | 130 |
| Medical hypertension | Binary | 1 for presence and 0 for absence | 133 |

## 3.3 Data preprocessing

The following steps have been applied:

1. *Removing missing values.* Where a missing value was present in a row of the raw data, the whole row was removed.

2. *Removing exterme outliers.* The row containing the maximum value of blood insulin (6150 microUI/ml) was also removed since such a concentration would probably be incompatible with the fact that the participant is living[2].

3. *Variable transformation.* All continuous variables were graphically checked. Metal-mixtures, glucose, triglycerides and insulin were log-transformed as done by Pang et al. since they showed distributions with right tails [29].

Following this, 608 observations without missing values were retained.

[2] This value represents more than 420 times the daily insulin dose that physicians generally give to a 70-kg diabetic patient (i.e., "*In general, individuals with type 1 DM require 0.5–1 U/kg per day of insulin*" [19]).

# Chapter 4

# Methods

## 4.1 Preliminary remark

It is important to emphasize that principal component analysis and cluster analysis are *exploratory methods* contrary to regression models which predict an outcome. In our work, we separated these two radically different approaches in different chapters. Firstly, we used the exploratory methods in order to summarize the data. Subsequently, we tried to link these findings with hypertension using regression models. The global approach is depicted in Figure 4.1.

## 4.2 Exploratory methods

### 4.2.1 First method: PCA

*Principal component analysis* (PCA) is a statistical tool reducing the dimensions of data. This is particularly interesting when the number of variables $p$ is large in order to summarize the information. Basically, we want to project the $p$-dimensional vectors into a $q$-dimension subspace where $q \ll p$. Our summary will be the projection of the original $p$ vectors on to $q$ directions, the principal components (PCs), which span the subspace. A way of deriving the PCs is to find the projections which maximize the variance. The mathematical demonstration we reported below is taken from the book *Advanced Data Analysis from an Elementary Point of View* by Shalizi [31].

Let $\mathbf{X}$ be a data of size $n \times p$ where $n$ is the number of individuals and $p$ is the number of variables and let $\tilde{\mathbf{X}}_\mathrm{c}$ denote the centered reduced data. Let $\mathbf{x}_\mathrm{i}$ denote the $i$-th row of $\tilde{\mathbf{X}}_\mathrm{c}$. We are looking for the unit vector $\mathbf{w}$ of size $p \times 1$ such that the variance of the projected individuals on it is maximized, i.e. the PC.

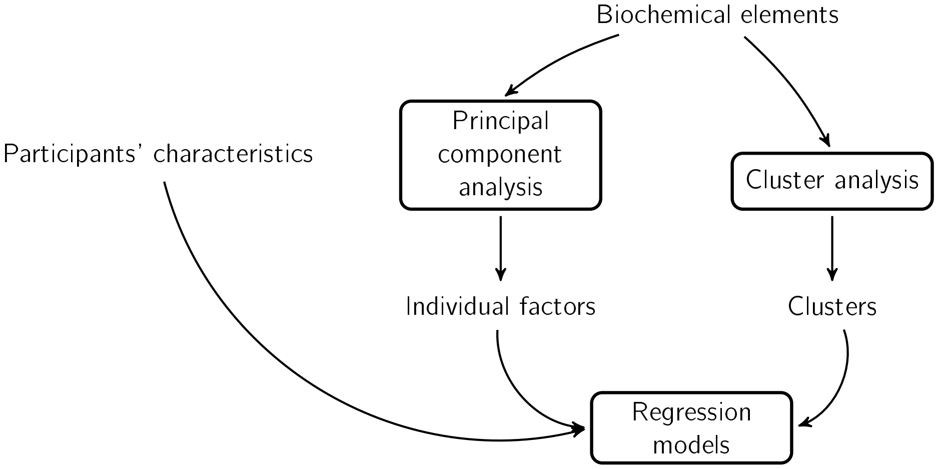


**Figure 4.1** Global approach in our work.

It is important to emphasize that principal component analysis and cluster analysis are *exploratory methods* contrary to regression models which predict an outcome. In our work, we separated these two radically different approaches in different chapters.

The variance is:

$$\begin{aligned}\hat{\sigma^2}(\overrightarrow{w} \cdot \overrightarrow{x_i}) &= \frac{1}{n}\sum_{i=1}^{n}\left(\overrightarrow{x_i} \cdot \overrightarrow{w}\right)^2 \\ &= \frac{1}{n}(\mathbf{Xw})^T(\mathbf{Xw}) \\ &= \frac{1}{n}\mathbf{w}^T\mathbf{X}^T\mathbf{Xw} \\ &= \mathbf{w}^T\frac{\mathbf{X}^T\mathbf{X}}{n}\mathbf{w}.\end{aligned}$$

Thus, we want to maximize $\hat{\sigma^2}(\overrightarrow{w} \cdot \overrightarrow{x_i})$ with the constraint that $\overrightarrow{w} \cdot \overrightarrow{w} = 1$. We will use the Lagrangian function with the Lagrange multiplier $\lambda$:

$$\begin{aligned}\mathcal{L}(\mathbf{w}, \lambda) &= \mathbf{w}^T\frac{\mathbf{X}^T\mathbf{X}}{n}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{w} - 1) \\ \frac{\partial\mathcal{L}}{\partial\lambda} &= \mathbf{w}^T\mathbf{w} - 1 \\ \frac{\partial\mathcal{L}}{\partial w} &= 2\frac{\mathbf{X}^T\mathbf{X}}{n}\mathbf{w} - 2\lambda\mathbf{w}.\end{aligned}$$

Setting the derivatives to zero, we get:

$$\mathbf{w}^T\mathbf{w} = 1$$
$$\frac{\mathbf{X}^T\mathbf{X}}{n}\mathbf{w} = \lambda\mathbf{w}.$$

The solution is given by diagonalization of matrix $\frac{X^TX}{n}$ which is in fact the covariance matrix. We obtain $p$ eigenvectors $\mathbf{w}$ that are the principal components (PCs)[1] and $p$ associated eigenvalues $\lambda$. The maximizing vector will be the one associated with the largest eigenvalue $\lambda$. Eigenvalues are directly proportional to the contribution of their associated PC to the total variance. By nature of diagonalization, the PCs are orthogonal to each other.

All individuals and variables can then be projected on the PCs in order to find some homogeneous patterns of variables or individuals in the data. These projections are called *factors* or *scores*.

Other PCA results are:

- *Cosine*: the correlation between the variable and the factor [38],
- *Squared cosine*: measures the quality of the representation of the variable by a factor [38].

Following the methodology of Pang et al., we performed a PCA on all urine element excretions [29]. Before running the PCA, we standardized the variables to unit variance and zero mean.

Using the Elbow method, we selected the first 3 PCs. The Elbow method consists in finding the eigenvalues that are on the left of the line that separates the rapidly decreasing eigenvalues and the approximately equal eigenvalues (Figure 4.2) [2]. Another method for significant PC selection is to select all principal components with eigenvalues greater than 1 [38]. Using the latter method in this study would mean selecting 7 PCs, which is excessive. As shown by Trevor Hastie et al., we computed the eigenvalues on a normally simulated data $X_{sim}$ of the same dimension as $\tilde{X}_c$, with all $n \cdot p$ cells with unit variance and zero mean [17]. Values from the simulated data are represented by blue crosses in Figure 4.2. Only the first 3 eigenvalues are bigger than the simulated ones, which brings us to the same conclusion as for the Elbow method.

### 4.2.2 Second method: cluster analysis

In this section, we will firstly introduce some general considerations on clustering methods and distance metrics. Afterwards, we will describe the three main steps we followed to construct our clusters:

1. Computing a *distance matrix*: the Manhattan matrix [15],
2. Choosing a clustering algorithm: the PAM algorithm [15],
3. Selecting the number of clusters: the silhouette index [15, 20].

#### General considerations on clustering methods

*Cluster analysis* is a method of finding homogeneous groups in data [20]. These groups are called *clusters*. The number of clusters $K$ is either chosen in function of cluster validation indexes (*parametric clustering*) or itself determined by the clustering method (*nonparametric clustering*) [15]. Clustering methods are

[1] also called *loadings*

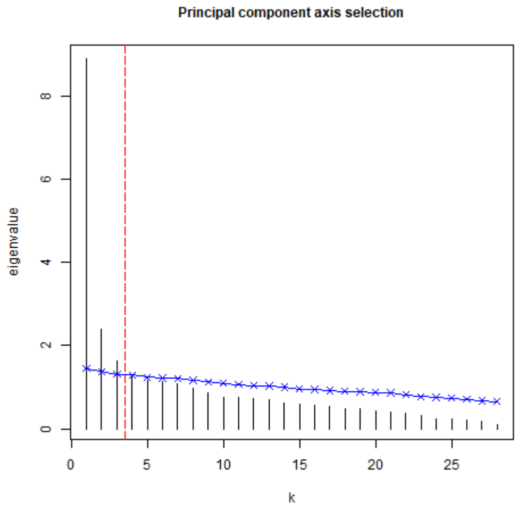


**Figure 4.2** Selecting significant principal components.

Using the Elbow method, we selected the first 3 principal components. The Elbow method consists in finding the eigenvalues that are on the left of the line that separates the rapidly decreasing eigenvalues and the approximately equal eigenvalues (red line). Eigenvalues from normally simulated data are represented by blue crosses. Only the first 3 eigenvalues are bigger than the simulated ones.

unsupervised methods since their aim is not to predict a response variable. After clustering, members assigned to the same cluster are expected to be close to each other in the $p$-dimension space. If the $p$ variables are continuous, this proximity can be defined in terms of Euclidean distance. However other distance metrics exist: the Minkowski-type distance, the Hausdorff distance, the Mahahanobis distance, the Gower dissimilarity, etc [15]. Cluster analysis should not be confused with linear discriminant analysis. Whereas the aim of clustering methods is to establish groups, the aim of linear discriminant analysis is to assign individuals to groups which are already established [15, 38].

Different types of distances are:

1. **Standard Euclidean distance**. If $p$ is the number of variables, the standard Euclidean distance between individuals $x_i$ and $x_j$ and denoted by $d(x_i, x_j)$ is given by:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{p}(x_{il} - x_{jl})^2},$$

where $x_{il}$ and $x_{jl}$ are the values taken in variable $l$ by individuals $x_i$ and $x_j$, respectively.

2. **Manhattan distance**. If $p$ is the number of variables, the Manhattan distance between individuals $x_i$ and $x_j$ and denoted by $d(x_i, x_j)$ is given by:

$$d(x_i, x_j) = \sum_{l=1}^{p} \mid x_{il} - x_{jl} \mid,$$

where $x_{il}$ and $x_{jl}$ are the values taken in variable $l$ by individuals $x_i$ and $x_j$, respectively.

Instead of taking the square of the differences, we are taking the absolute value of the differences. This distance is typically used when adding one unit to $x_{i1}$ and one unit to $x_{i2}$ is the same than adding two units to $x_{i1}$ for any individual $i$ [20].

3. **Minkowski distance**. The Euclidean and the Manhattan distances are special cases of the so-called Minkowski distance where $q = 2$ for Euclidean distance and $q = 1$ for Manhattan distance in the formula below. If $p$ is the number of variables, the Minkowski distance between individuals $x_i$ and $x_j$ and denoted by $d(x_i, x_j)$ is given by:

$$d(x_i, x_j) = \Big( \sum_{l=1}^{p} \mid x_{il} - x_{jl} \mid^q \Big)^{1/q},$$

where $x_{il}$ and $x_{jl}$ are the values taken in variable $l$ by individuals $x_i$ and $x_j$, respectively [15].

4. **Gower dissimilarity**. When the Euclidean and the Manhattan distances are distance metrics for continuous variables, the so-called *Gower dissimilarity* allows to aggregate $p$ variables of mixed type. If the data does not contain any missing value the Gower dissimilarity between individuals $x_i$ and $x_j$ and denoted by $d_G(x_i, x_j)$ is given by:

$$d_G(x_i, x_j) = \frac{\sum_{l=1}^{p} w_j \cdot d_l(x_{il}, x_{jl})}{\sum_{j=1}^{p} w_j},$$

where $d_l(x_{il}, x_{jl})$ is the Manhattan distance between individuals $x_i$ and $x_j$ for variable $l = 1, 2 \dots p$ [15]. As reported by Hennig et al., Gower recommended to use weights $w_l$ to scale the distances for each variable between 0 and 1 such that $w_j \cdot d_l(x_i l, x_j l) \in [0, 1]$ for all possible pairs of $x_j$ and $x_j$ individuals. However, Hennig et al. argued to reserve them only for binary and "very discrete variables" because otherwise the weights could have an excessive high influence on the cluster borders. Hennig et al. also said that weights may be used in order to give more importance to some variables that are considered as main variables [15].

### Clustering procedure

**First step: computing the distance matrix.** The Manhattan distance matrix is computed between all the individuals taking into consideration all standardized urine excretions. The result can then be presented as a $n \times n$ distance matrix. In our case, the distance matrix has dimension $608 \times 608$. The function `daisy()` from the `cluster` packcage in `R` computes the distance matrix. In one-line code we obtained our Manhattan distance matrix:

```
mydaisy<-daisy(...,metric = "manhattan")
```

**Second step: choosing a clustering algorithm (PAM).** The heuristic *partitioning around medoids* (PAM) algorithm belongs to the $k$-medoids clustering methods. The idea behind $K$-medoids is not dissimilar to that of $k$-means. However, whilst we are sampling random points in the $p$-dimension space in $k$-means, we are sampling existing points in $k$-medoids. Compared to $k$-means clustering, this is often a more robust method since it generally does not consider the squared Euclidean distance which is very sensitive to extreme outliers. PAM is an *hard* clustering method. In *hard* clustering, members are assigned to a precise cluster $k$ with null probability of belonging to the $K - 1$ remaining clusters. In contrast, in *soft* clustering every member has a degree of membership to each $K$ cluster [15].

The PAM algorithm follows the procedure described in the book *Handbook of Cluster Analysis* by C. Hennig et al. 2016 [15] in five steps:

1. *Initialize: randomly select (without replacement) $K$ of the $N$ data points as the initial medoids*[2].
2. *Assign each observation to the medoid with which it is closest, where closest is based on a specific distance measure and compute the total cost across all observations, where the cost is the sum of the distance of each observation to its associated medoid.*
3. *For each medoid $k$, for $k = 1,\ldots,K$ consider all $N - K$ nonmedoid, $o$. Swap $k$ and $o$ and recompute the total cost.*
4. *Select the solution with the lowest cost.*
5. *Repeat steps 2-4 until the set of medoids does not change.*

Finding the $K$ medoids is an optimizing problem with a globally optimal solution since we have a finite number of $\binom{N}{K}$ candidate solutions. However, computational cost is often huge in practice. For example, for $N = 500$ and $K = 5$ which is a realistic example in practice, we would have $2 \cdot 10^{11}$ possible sets of medoids and thus $2 \cdot 10^{11}$ costs to compute. With the PAM algoritm, only $K \cdot (N - K)$ costs are computed during the first iteration (2475 in our example).

Let us take a simple example to illustrate the PAM algorithm with $n = 10$ individuals and 2 variables, $x$ and $y$. Our objective will be to construct two clusters ($K = 2$). This data was generated by `R`:

```
set.seed(123)
data<-runif(20)
x<-data[1:10]
y<-data[11:20]
```

The data is presented in Table 4.1. The corresponding Euclidean distance matrix is shown in Table 4.2. The first iteration of a $k$-medoids algorithm is shown in Figure 4.3. In step 1, observations 7 and 9 are randomly chosen as medoids $m_1$ and $m_2$, respectively. In step 2, the sum of the distances is computed in this configuration, assigning the non-medoids to their closest medoid ($\text{cost}_0 = 2.97$). In steps 3 and 4, $m_1$ is swapped with observation 1 as this decreases the sum of the distances by -1.59. Steps 2-4 are repeated until medoid stabilization is achieved. See the *Appendix* for the PAM program used to solve this example.

The PAM algorithm was implemented with SKIPOGH data using the `pam()` function from the `cluster` package in `R` [1].

**Third step: selecting the number of clusters.** To select the optimal number of clusters, we computed the *silhouette index*, also called *silhouette width*, for $K = 1, 2, \ldots 10$. The number of clusters $K$ associated with the biggest silhouette index was selected. The silhouette index $SI_K$ is defined as [20]:

$$SI_K = \frac{1}{n}\sum_{i=1}^{n} \frac{b(i) - a(i)}{max\{a(i), b(i)\}},$$

where $a(i)$ is the average dissimilarity between object $i$ and all objects in its cluster and $b(i)$ is the minimum average dissimilarity of $i$ to all points in any other cluster not including $i$ [15]. In our work we used the Manhattan dissimilarity. The silhouette index is bounded between minus one and one, i.e. $SI_K \in [-1, 1]$.

[2] However, Kaufman et al. recommend to take non-random data points as the initial medoids [20].

**Table 4.1:** Example data for a *k*-medoids algorithm.

| ID | x | y |
|---|---|---|
| 1 | 0.29 | 0.96 |
| 2 | 0.79 | 0.45 |
| 3 | 0.41 | 0.68 |
| 4 | 0.88 | 0.57 |
| 5 | 0.94 | 0.10 |
| 6 | 0.05 | 0.90 |
| 7 | 0.53 | 0.25 |
| 8 | 0.89 | 0.04 |
| 9 | 0.55 | 0.33 |
| 10 | 0.46 | 0.95 |

**Table 4.2:** Distance matrix for the example given in Table 4.1.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 0.71 | | | | | | | | |
| 3 | 0.30 | 0.44 | | | | | | | |
| 4 | 0.71 | 0.15 | 0.49 | | | | | | |
| 5 | 1.07 | 0.38 | 0.78 | 0.47 | | | | | |
| 6 | 0.25 | 0.87 | 0.43 | 0.90 | 1.20 | | | | |
| 7 | 0.75 | 0.33 | 0.45 | 0.48 | 0.44 | 0.81 | | | |
| 8 | 1.10 | 0.42 | 0.80 | 0.53 | 0.08 | 1.21 | 0.42 | | |
| 9 | 0.68 | 0.27 | 0.38 | 0.41 | 0.45 | 0.76 | 0.09 | 0.44 | |
| 10 | 0.17 | 0.60 | 0.28 | 0.57 | 0.98 | 0.41 | 0.71 | 1.01 | 0.63 |

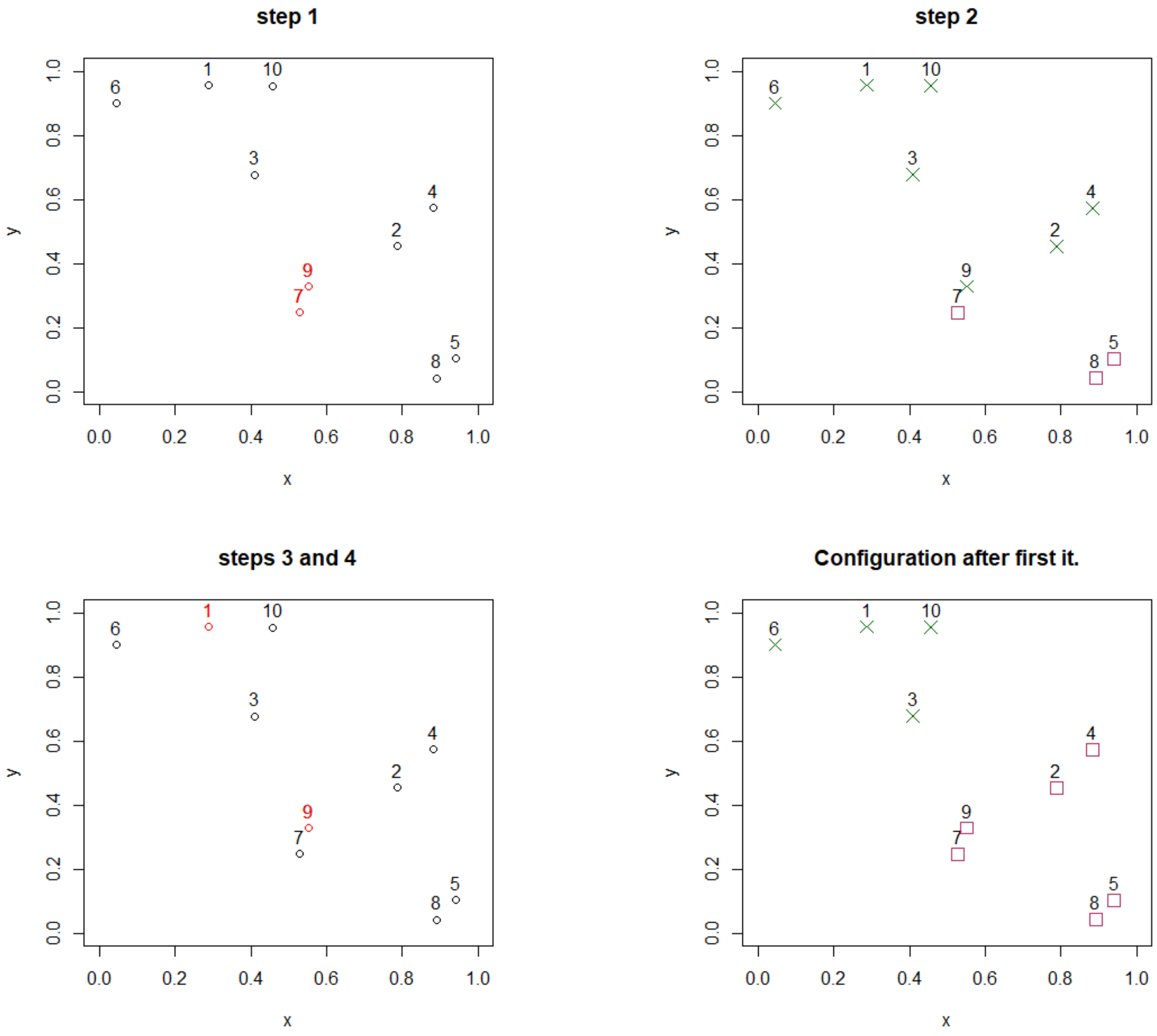


**Figure 4.3** First iteration of the algorithm for the example given in Table 4.1.
In step 1, observations 7 and 9 are randomly chosen as medoids $m_1$ and $m_2$, respectively. In step 2, the sum of the distances is computed in this configuration, assigning the non-medoids to their closest medoid ($\text{cost}_0 = 2.97$). In steps 3 and 4, $m_1$ is swapped with observation 1 as this decreases the sum of the distances by -1.59. Steps 2-4 are repeated until medoid stabilization is achieved.

As reported by Struyf et al. [35], the value $SI_K$ may be interpreted as follows:

- $SI_K \approx 1 \Rightarrow$ *object $i$ is well classified (in $A$),*
- $SI_K \approx 0 \Rightarrow$ *object $i$ lies intermediate between two clusters ($A$ and $B$),*
- $SI_K \approx -1 \Rightarrow$ *object $i$ is badly classified (closer to $B$ than to $A$).*

Thus, the higher the index is, the more distinguishable the clusters are. Conceptually, the silhouette index measures the distance between the clusters whilst taking into account out each cluster is (see examples in Figure 4.4). Not only can the silhouette index be used as a tool to chose the optimal number of clusters, but also to measure the quality of the clustering result and thereby is an index for cluster validation.

## 4.3 Assessing for cluster stability

If we state that our data is a random sample from a finite population, we can consider clustering as a random process. In this scope, clusters are random. However, we would like to construct $K$ clusters from our data that are the more stable as possible. In order to assess for cluster stability, we will summary two methods:

1. **The Rand index** [30]. The *Rand index* is a measure of cluster stability. Firstly, we divide the data of dimension $n \times p$ into three parts: two training sets $S$ and $T$ and one testing set $E$. Then, we perform a clustering on $S$ giving $\mathcal{C}_S$ and on $T$ giving $\mathcal{C}_T$, independently. The purpose is to use points of $E$ to see if there is a difference in cluster attribution if we are using the medoids of $\mathcal{C}_S$ or the medoids of $\mathcal{C}_T$. The rand index is given by:

$$\mathcal{R}(\mathcal{C}_S, \mathcal{C}_T) = \frac{N_{11} + N_{00}}{n(n-1)/2},$$

   where $N_{11}$ is the number of point pairs of $E$ that are in the same cluster with both $\mathcal{C}_S$ and $\mathcal{C}_T$ and where $N_{00}$ is the number of point pairs of $E$ in different clusters with both $\mathcal{C}_S$ and $\mathcal{C}_T$ [15].

   Since Rand index is random from the sampling step, a bootstrap version combining a kind of Rand index and bootstrap techniques was described by Dolnicar and Leisch in 2010 [10].

2. **Bootstrapping average cluster stability** [10]. The technique computes several replicates of Rand indices with the notable difference that $S_i$ and $T_i$ are both bootstrapped samples of size $n$ from the original data that replace $S$ and $T$ in the classical Rand index. The formula for the replicates is:

$$\mathcal{R}_i(\mathcal{C}_{S_i}, \mathcal{C}_{T_i}) = \frac{N_{11_i} + N_{00_i}}{n(n-1)/2},$$

   where $N_{11_i}$ is the number of point pairs of $E$ that are in the same cluster under both $\mathcal{C}_{S_i}$ and $\mathcal{C}_{T_i}$ and where $N_{00_i}$ is the number of point pairs of $E$ in different clusters under both $\mathcal{C}_{S_i}$ and $\mathcal{C}_{T_i}$ [10]. The mean and the standard deviations of bootstrapped replicates are estimates of expectation and standard error of Rand index, respectively.

## 4.4 Checking for assumptions

Statistical test results are not presented if assumptions are violated. For every performed two-sided two-sample Student's test, normality in the two groups was assessed using normal Q-Q plots. Moreover, equal variance assumption was adopted if the $p$-value in the Levene's test was larger than 5%.

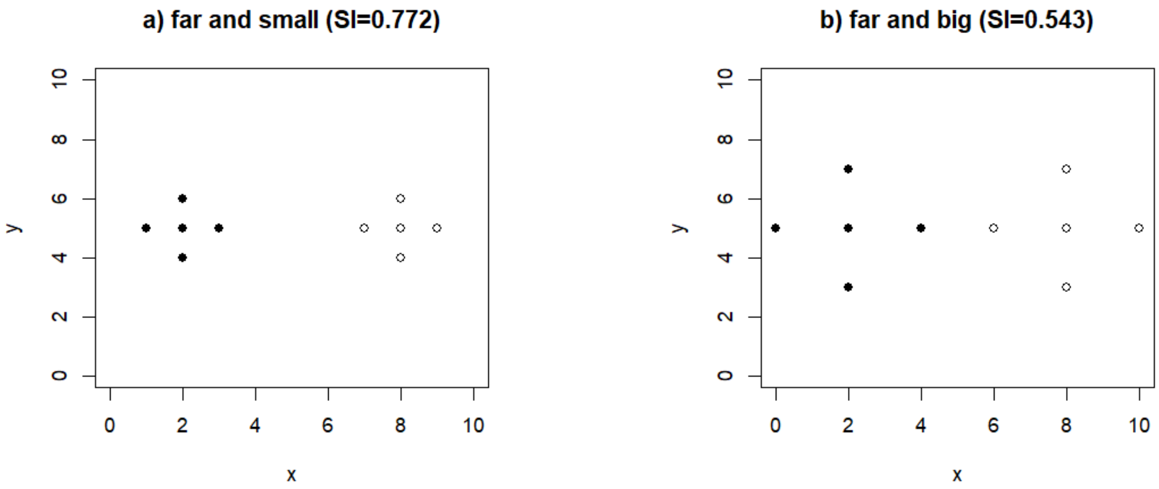


**Figure 4.4** Examples of silhouette index.

These plots show 4 situations concerning 2 clusters (white and black). They illustrate that the silhouette index takes into consideration both the size of the clusters and the distance between each other. Plot a) where the clusters are far and small has the highest silhouette index ($SI_K = 0.772$). Plot b) where the clusters are far and big and plot c) where the clusters are close and small have the same silhouette index ($SI_K = 0.543$). Plot d) where the clusters are close and big has the lowest silhouette index ($SI_K = 0.190$).

# Chapter 5

# Results

In the first section, we describe the data. In the second section, exploratory analyses are presented. Subsequent sections study the link between the biochemical elements and hypertension.

## 5.1 Global data description

For each continuous variable, the mean, the standard deviation (SD), the median, the interquartile range (IQR), the skewness and the kurtosis were given (see Table 5.1). The mean age was 48.1 years (SD=17.8). All metal-mixtures without exception had a positive skewness and a positive kurtosis. This is often seen in variables measuring concentrations since negative concentrations do not exist and extreme values are easy to reach with intoxication and/or impaired physiology. Aluminum, platine, cobalt and thallium kurtosis had extremely high kurtosis, respectively 461, 561, 324 and 576. This is partially due to extreme outliers. Indeed, if we removed the highest thallium value (36763.83 ng), the skewness passed from 23.8 to 1.5, the kurtosis passed from 576 to 6 and the boxplot of the variable completely changed (see Figure 5.1). Sample kurtosis is extremely sensitive to extreme outliers since its formula is spreading the difference between the mean and a value with power 4:

$$kurtosis = \frac{n \cdot (n+1) \cdot (n-1)}{(n-2) \cdot (n-3)} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2}.$$

However, it is possible for a participant to have this thallium value since 36700 ng = 36 mg and severe poisoning begins with values higher than 1000 mg. Thus, we decided to keep this high value. Similarly, other metal-mixture were showing highest values that were often under their lethal dose. Consequently, we decided to keep these high values since we cannot exclude mild intoxication cases.

Discrete variables are reported in Table 5.2 ($n = 608$, 4 variables) where they are simply expressed as counts and proportions. There were 290 men and 318 women. 4.6% of the participants had diabetes. Hypertension concerned 23.3% of the participants. This is slightly lower than the worldwide prevalence estimated to 26.4% (95% CI 26.0 – 26.8%) [21].

## 5.2 Exploratory analysis

In this section, correlation matrix, PCA and cluster analysis are presented.

**Table 5.1:** Description of continuous variables ($n = 608$, 35 variables).

| Variable | Mean | SD | Median | IQR | Skew. | Kurt. |
|---|---:|---:|---:|---:|---:|---:|
| age, years | 48.1 | 17.8 | 48 | 29.6 | 0.1 | -1 |
| triglycerides, mmol/l | 1.1 | 0.7 | 0.9 | 0.7 | 2.9 | 14.6 |
| cholesterol, mmol/l | 5.1 | 1 | 5 | 1.4 | 0.3 | -0.3 |
| blood glucose, mmol/l | 5.3 | 0.7 | 5.2 | 0.7 | 2.4 | 12.2 |
| blood insulin, microIU/ml | 6.3 | 7.5 | 4.7 | 5.6 | 5.5 | 44.7 |
| lithium, ng | 90191.6 | 184560.6 | 53656.4 | 52147.1 | 10.5 | 140.7 |
| beryllium, ng | 4.8 | 6.7 | 3.7 | 3.4 | 11 | 154.4 |
| aluminum, ng | 15452.7 | 58987.4 | 9889.9 | 7914.8 | 20.7 | 460.9 |
| vanadium, ng | 1114.5 | 579.1 | 1018.9 | 668.1 | 3.2 | 31 |
| chrome, ng | 2927.4 | 2011.3 | 2580.2 | 1679.8 | 6.9 | 84.4 |
| manganese, ng | 530.1 | 662.3 | 386.6 | 345.8 | 6.9 | 68 |
| cobalt, ng | 928.6 | 3732.9 | 415.6 | 451.7 | 17 | 324.2 |
| nickel, ng | 4735.6 | 4829.4 | 3562.1 | 2930.3 | 5.1 | 34 |
| copper, ng | 36749.9 | 97315 | 22277.8 | 14475.5 | 8.4 | 75.9 |
| zinc, ng | 616169.5 | 418437.1 | 502317 | 489183.3 | 1.6 | 3.3 |
| arsenic, ng | 86466 | 241222.8 | 33016 | 56860.9 | 10.6 | 142.3 |
| molybdenum, ng | 81900.1 | 66734.4 | 64343.6 | 64918.6 | 2.4 | 10 |
| palladium, ng | 376.6 | 288.5 | 315.8 | 256.8 | 3.7 | 25.5 |
| silver, ng | 291.7 | 385.4 | 140.6 | 325.8 | 3.7 | 22.5 |
| cadmium, ng | 538.6 | 351.1 | 453.8 | 387.2 | 1.9 | 6.4 |
| tin, ng | 1362.9 | 3900.7 | 778.8 | 740.5 | 15.6 | 292.9 |
| antimony, ng | 251.4 | 591.7 | 126.9 | 184.5 | 11.3 | 162.9 |
| platinum, ng | 104 | 395.6 | 67.3 | 59.7 | 23.3 | 560.6 |
| mercury, ng | 819.6 | 713.3 | 593.3 | 736.9 | 1.9 | 4.3 |
| thallium, ng | 453.7 | 1490.9 | 341.9 | 270.3 | 23.8 | 575.6 |
| lead, ng | 2132.2 | 1589.5 | 1717.8 | 1606.9 | 3 | 17.9 |
| bismuth, ng | 71 | 147.1 | 27.4 | 56.5 | 7.1 | 73.6 |
| sodium, mmol | 140.1 | 59.7 | 133.1 | 76.7 | 0.9 | 1.5 |
| potassium, mmol | 63.8 | 22.8 | 61.2 | 29.7 | 0.7 | 1.1 |
| calcium, mmol | 4.2 | 2.3 | 3.8 | 2.9 | 1 | 1.4 |
| phosphate, mmol | 26.4 | 9.5 | 25.3 | 11.6 | 0.8 | 2.3 |
| urea, mmol | 359.3 | 113.8 | 348 | 160 | 0.5 | 0.1 |
| magnesium, mmol | 4.1 | 1.6 | 3.9 | 1.9 | 0.9 | 1.6 |
| systolic blood pressure, mmHg | 118.7 | 17.3 | 116.5 | 21.8 | 1 | 1.5 |
| diastolic blood pressure, mmHg | 74.9 | 9.7 | 74.5 | 13 | 0.4 | 0.1 |

**Table 5.2:** Description of discrete variables ($n = 608$, 4 variables).

| Variable | Reference Category | Reference cases (%) | Non ref. cases (%) |
|---|---|---:|---:|
| Sex | *male* | 290 (47.7%) | 318 (52.3%) |
| Diabetes | *has diabetes* | 28 (4.6%) | 580 (95.4%) |
| Anti-hypertensive drugs | *taking drugs* | 99 (16.3%) | 509 (83.7%) |
| Medical hypertension | *has hypertension* | 142 (23.3%) | 466 (76.7%) |

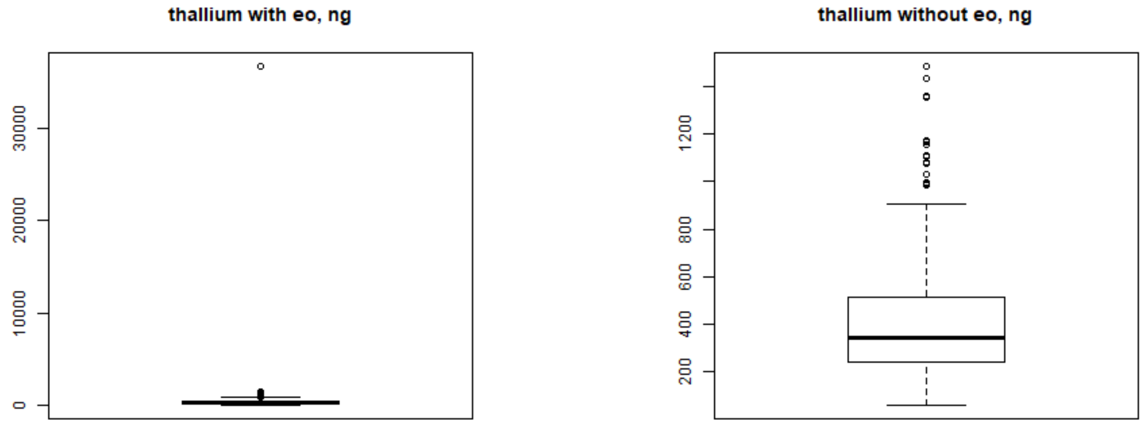


**Figure 5.1** Effect of removing the biggest thallium concentration.
Left: distribution of thallium concentrations in ng from original data. Right: distribution of thallium concentrations in ng after removing the highest value. As we can see, the thallium boxplot completely changed after removing the highest value.

## 5.2.1 Correlation matrix

Correlation matrix was computed using Pearson's method. All metal-mixtures and non-metals are only positively correlated between each other. Moreover, the correlation between vanadium and chrome (0.87, $p$-value $< 0.05$ in the correlation test) and the correlation between phosphates and urea (0.78, $p$-value $< 0.05$ in the correlation test) are the strongest ones (see Figure 5.2). If correlation between vanadium and chrome is difficult to interpret, the correlation between phosphates and urea can be explained by two reasons. Firstly, it can result from the increased level of these biochemical elements often seen in *chronic kidney disease* which concerns at least 6% of the population (US survey) [19]. Secondly, it can be a consequence of a diet rich in meat where both urea and phosphates intakes increase.

## 5.2.2 PCA

### PCA results

Plots showing the variables factor map and individuals factor map from PCA are shown in Figure 5.3 and Figure 5.4, respectively. Factor loadings are presented in Table 5.3. Three PCs explained 46.04% of the total variance. Loadings with absolute value greater than 0.40 were considered as important and reported in bold in Table 5.3. The first PC (explaining 31.71% of the variance) represents the global exposure to metal-mixtures and non-metals since they are all positively correlated with the PC. The second PC (explaining 8.57% of the variance) is almost more interesting than the first PC since it is only correlated with non-metals (sodium, calcium, phosphates, urea and magnesium). Therefore, it can be assumed that this PC underlies the fact that the sources and/or the metabolic pathways are not the same between metal-mixtures and non-metals the data. This difference between metal-mixtures and non-metals can also be seen in Figure 5.3 where metal-mixtures (in maroon) are approximately orthogonal to non-metals (in orange), and therefore linearly independent in the two-dimentional space of the PCA. The third PC (only explaining 5.76% of the

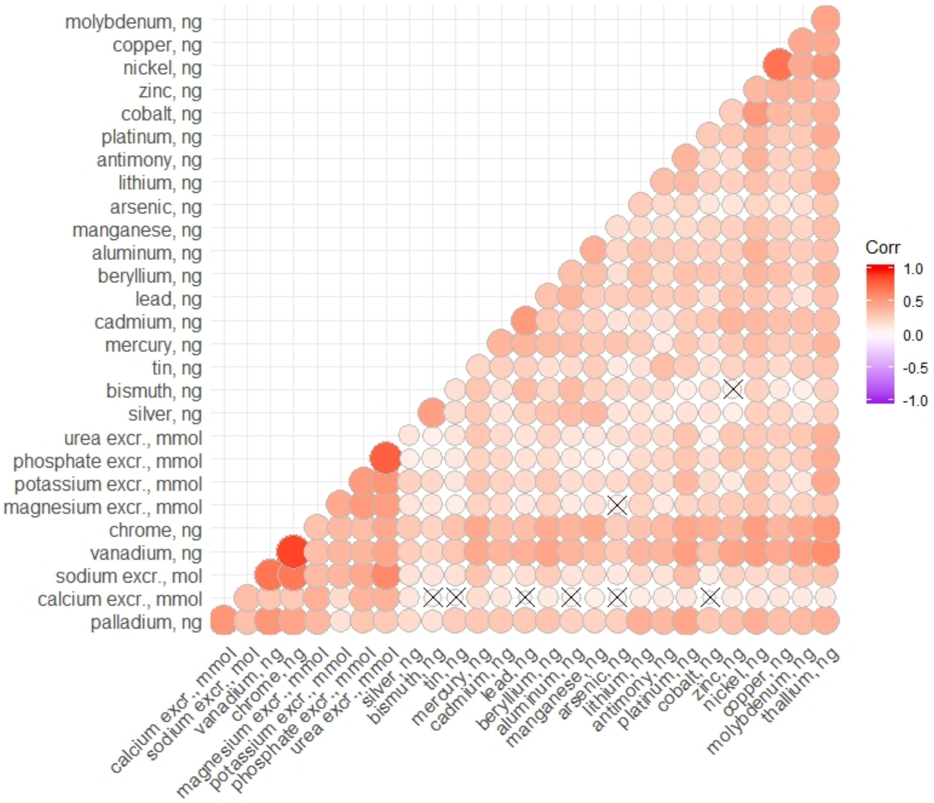


**Figure 5.2** Correlation plot representing correlations between 24-hour urine excretions.

This plot graphically represents correlations between metal-mixtures and non-metals with rounds (Pearson's correlations) with the color intensity directly proportional to the correlation value. Non significant ($p$-value $> 0.05$) correlations are expressed as a crossed round. This plot underlines the fact that all metal-mixtures and non-metals correlate only positively between each other. Moreover, the strongest correlations are between vanadium and chrome (0.87) and between phosphates and urea (0.78).

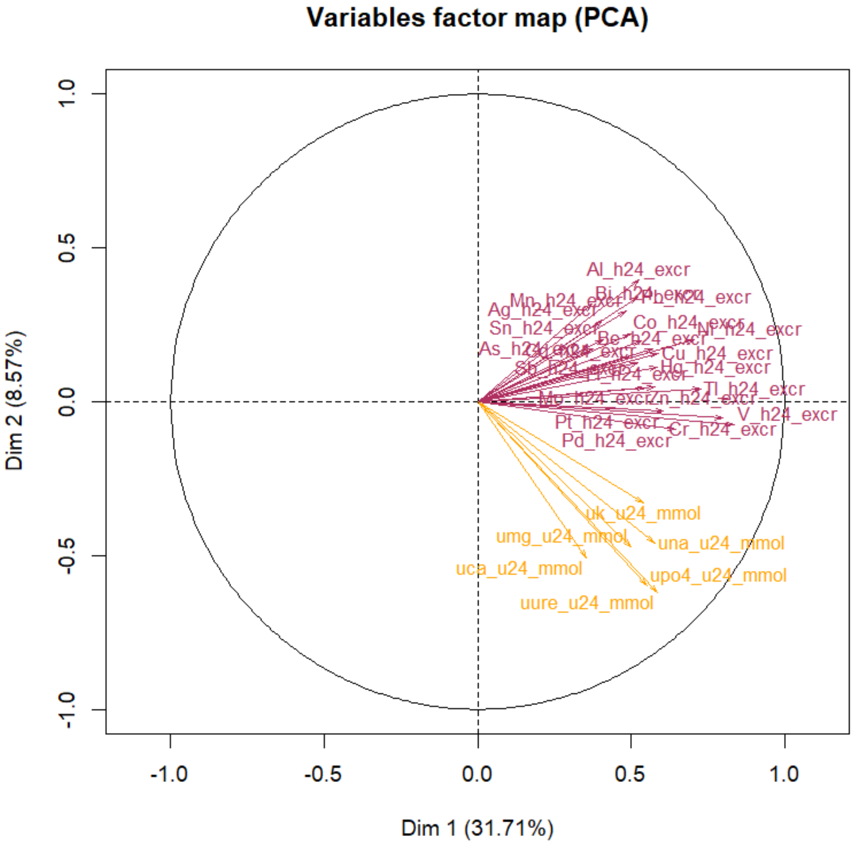


**Figure 5.3** PCA of variables.

Metal-mixtures and non-metals are represented in maroon and orange respectively.

variance) is correlated with silver and bismuth.

A supplementary plot displays the 5 essential metal-mixtures (manganese, cobalt, copper, zinc and molybdenum) in black and the 17 remaining toxic metal-mixtures in gray on the variables factor map (Figure 5.5). Interestingly, this indicates that all the essential metal-mixtures are present in the positive values of the second axis.

Note that we cannot see heterogenous distribution of the individuals with hypertension and those without hypertension in the individuals factor map in Figure 5.4.

### 5.2.3 Cluster analysis

**Clustering results.** The Manhattan distance matrix from SKIPOGH was computed as explained in Chapter 4. It is of a considerable size (608 × 608). Consequently, we only presented it in a graphical way in Figure 5.6 where each point is a pixel in an image of size 608 × 608 picture. Each pixel in this image represents the distance between one individual and another. Pixels are sorted by participant ID. The brighter the pixel

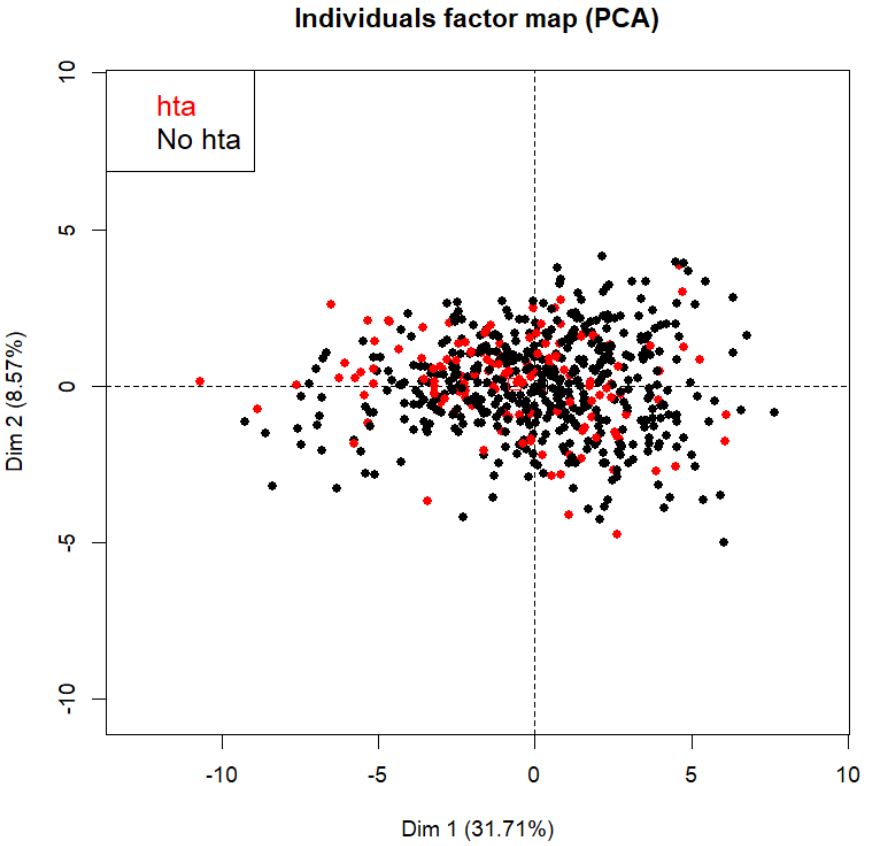


**Figure 5.4** PCA of individuals.

Red dots represent individuals with hypertension (hta). Black dots represent individual with no hypertension (no hta).

**Table 5.3:** Factor loadings in PCA.

| Variable | PC 1 | PC 2 | PC 3 |
|---|---|---|---|
| lithium, ng | **0.521** | 0.126 | -0.023 |
| beryllium, ng | **0.570** | 0.171 | 0.100 |
| aluminum, ng | **0.527** | 0.396 | 0.204 |
| vanadium, ng | **0.836** | -0.075 | -0.039 |
| chrome, ng | **0.799** | -0.054 | 0.010 |
| manganese, ng | **0.485** | 0.295 | 0.217 |
| cobalt, ng | **0.497** | 0.222 | -0.324 |
| nickel, ng | **0.706** | 0.200 | -0.279 |
| copper, ng | **0.588** | 0.158 | -0.357 |
| zinc, ng | **0.548** | 0.045 | -0.322 |
| arsenic, ng | 0.374 | 0.173 | 0.169 |
| molybdenum, ng | **0.578** | 0.048 | -0.387 |
| palladium, ng | **0.645** | -0.091 | -0.166 |
| silver, ng | **0.407** | 0.264 | **0.517** |
| cadmium, ng | **0.531** | 0.199 | -0.100 |
| tin, ng | **0.411** | 0.204 | -0.071 |
| antimony, ng | **0.491** | 0.145 | -0.145 |
| platinum, ng | **0.605** | -0.032 | -0.115 |
| mercury, ng | **0.585** | 0.112 | 0.250 |
| thallium, ng | **0.727** | 0.041 | -0.109 |
| lead, ng | **0.522** | 0.340 | 0.202 |
| bismuth, ng | 0.373 | 0.317 | **0.563** |
| sodium, mmol | **0.575** | **-0.460** | 0.173 |
| potassium, mmol | **0.540** | -0.328 | 0.298 |
| calcium, mmol | 0.354 | **-0.508** | 0.078 |
| phosphate, mmol | **0.552** | **-0.599** | 0.049 |
| urea, mmol | **0.586** | **-0.621** | 0.115 |
| magnesium, mmol | **0.500** | **-0.471** | 0.054 |
| Eigenvalue | 8.878 | 2.401 | 1.613 |
| Total variance (%) | 31.71 | 8.57 | 5.76 |
| Cumulative (%) | 31.71 | 40.28 | 46.04 |

Factor loadings are given in bold if $> 0.40$.

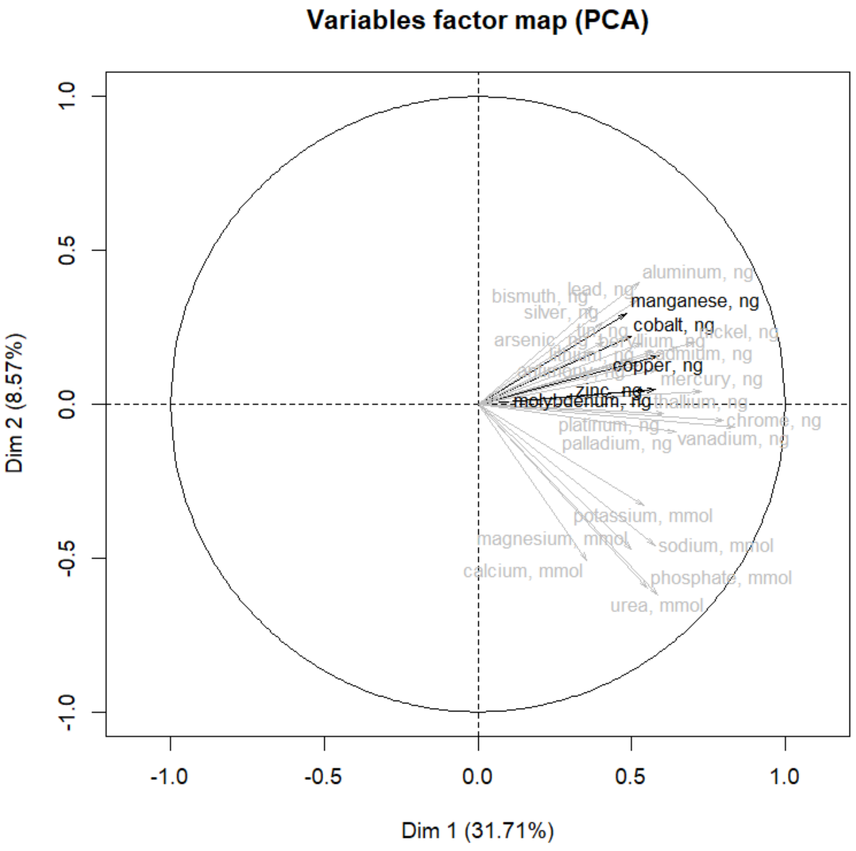


**Figure 5.5** PCA of variables with metal-mixture type.

In black: the 5 essential metal-mixtures (manganese, cobalt, copper, zinc and molybdenum). In gray: the 17 remaining toxic metal-mixtures and the non-metals. All the essential metal-mixtures without exception are in the upper right quadrant.

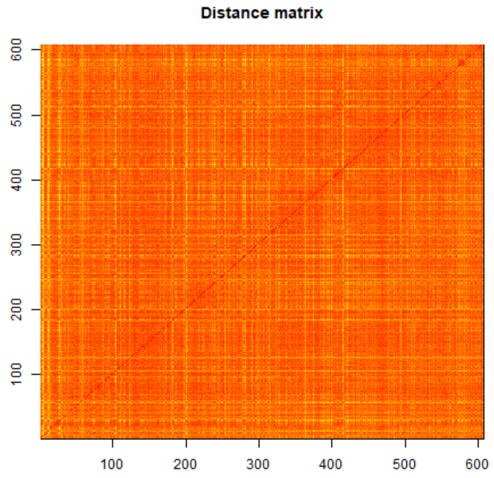


**Figure 5.6** Manhattan distance matrix (608 × 608).

Each pixel in this image represents the distance between one individual and another. Pixels are sorted by participant ID. The brighter the pixel is, the greater the distance is. Naturally, we cannot distinguish patterns in the matrix before the clustering procedure.

**Table 5.4:** Obtained clusters description ($K = 2$).

| Size | maximum diss. | average diss. | diameter | separation |
|---|---|---|---|---|
| 249 | 48.04068 | 21.15747 | 67.54185 | 11.88677 |
| 359 | 38.60081 | 21.89540 | 60.45442 | 11.88677 |

is, the greater the distance is. Naturally, we cannot distinguish patterns in the matrix before the clustering procedure.

From the calculation of silhouette width, the optimal number of clusters was set as 2 (Figure 5.7). Consequently, PAM algorithm was runned with $K = 2$. The cost after the build phase was 22.0 and the cost after the swap phase was 21.6. Silouhette width was 0.1767. The clusters are detailed in Table 5.4. The first and the second clusters had 249 and 359 individuals, respectively. The distance matrix before and after clustering are compared in Figure 5.8.

**Clustered individuals in the space of the PCA.** The two clusters display different biochemical patterns for first two PCs of the PCA as edvidenced by their boxplots and associated two-sided two-sample Student's tests (see Figure 5.9). Interpretation of the first PC of the PCA shows that the second cluster has a higher global exposure to metal-mixtures and non-metals. As for the second PC, the second cluster also has higher concentrations of non-metals.

**Assessing for cluster stability.** Since we performed PAM algorithm from the Manhattan distance matrix, we wanted to compare cluster stability with PAM algorithm performed from a Euclidean distance matrix.

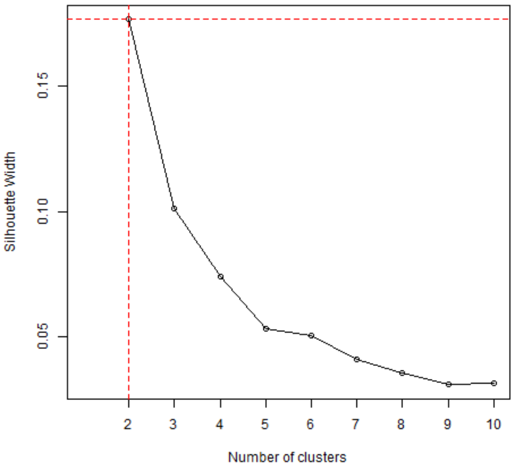


**Figure 5.7** Graphic of silhouette width.

The graph shows that the optimal number of clusters is 2 as this yiels the greatest silhouette width.

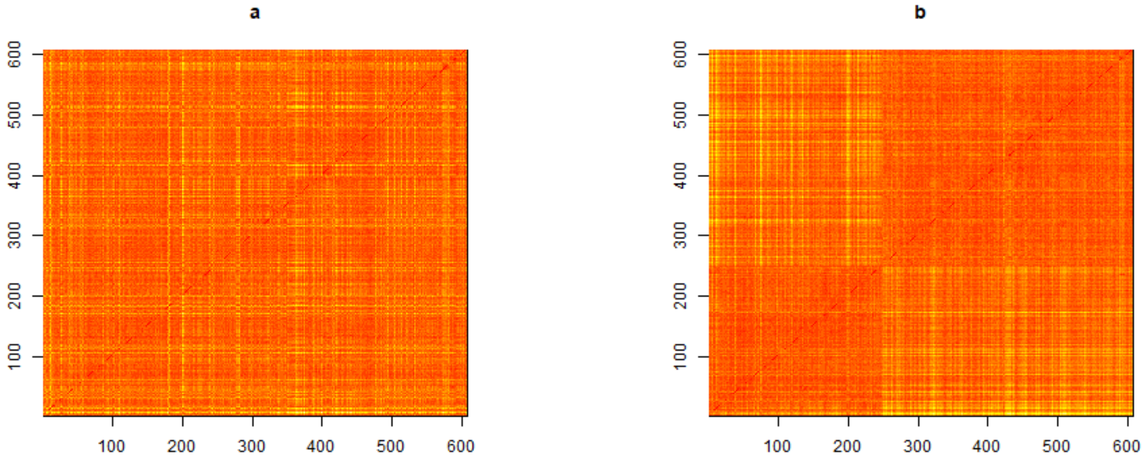


**Figure 5.8** Distance matrix before and after clustering.

a) Distance matrix sorted by participant ID, b) Distance matrix ordered by cluster attribution: the 249 observations on the left belong to cluster 1 and the 359 observations on the right belong to cluster 2. We can see slightly lower distances between individuals belonging to the same cluster.

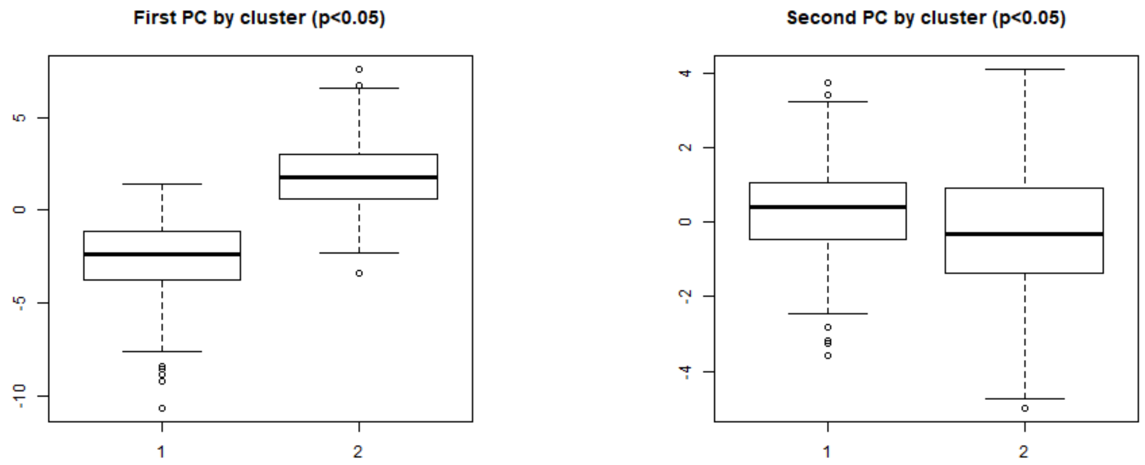


**Figure 5.9** Individual factors (PCA) by cluster attribution.

Left: the second cluster has a higher global exposure to metal-mixtures and non-metals ($p$-value of the Student's test $< 0.05$). Right: the second cluster has higher concentrations of non-metals ($p$-value of the Student's test $< 0.05$).

**Table 5.5:** Bootstrapped replicates of the Rand index.

| Metric | Mean | SD |
|---:|---:|---:|
| Euclidean | 0.084 | 0.149 |
| Manhattan | 0.103 | 0.155 |

We studied the replicates of the Rand index as explained in chapter 4. We fixed at $B = 1000$ the number of replications. Results are shown in Table 5.5 and Figure 5.10. According to the Mann-Whitney test, PAM algorithm from the Manhattan distance matrix provides higher replicates than PAM algorithm from the Euclidean distance matrix, at level $\alpha = 5\%$. Consequently, we kept the Manhattan metric.

## 5.3 Link with the outcome

In this section, previously found individual scores from PCA and clusters from cluster analysis are linked with hypertension as depicted in Figure 4.1. The first part will use simple association analyses and subsequent parts will use regression methods.

### 5.3.1 Simple association analyses

In Figure 5.11, individual factors by hypertension status are summarized in boxplots for dimension 1 and dimension 2, respectively. Two-sided two-sample Student's tests were performed on the two first dimensions to compare the *hypertension* and *no hypertension* groups. The results of the two-sided two-sample Student's tests are presented in Table 5.6. According to the $p$-values from two-sided two-sample Student's tests, mean

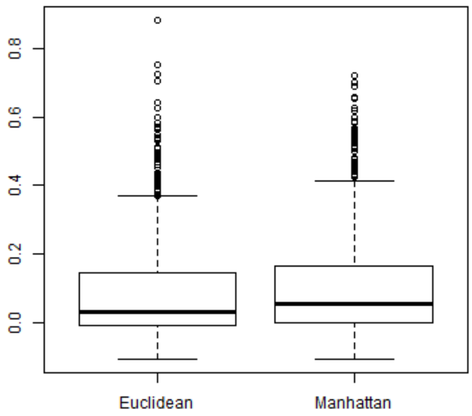


**Figure 5.10** Boxplot of bootstrapped replicates of the Rand index, $B = 1000$.

Boxplots describing the bootstrapped replicates of the Rand index in function of the metric chosen. According to the Mann-Whitney test, PAM algorithm from the Manhattan distance matrix provides higher replicates than PAM algorithm from the Euclidean distance matrix, at level $\alpha = 5\%$. Consequently, we kept the Manhattan metric.

in dimension 1 is lower in hypertensive individuals than in non-hypertensive individuals at level $\alpha = 5\%$ but difference of means between hypertensive and non-hypertensive groups in dimension 2 is not statistically significant at level $\alpha = 5\%$.

Continuous outcomes are SBP and DBP. A Pearson's correlation matrix was computed between SBP, DBP and individual factors. The correlation matrix is shown in Table 5.7. The magnitude of the correlations between SBP, DBP and the individual factors is small, even if they are significant at level $\alpha = 5\%$.

Note that these significant results are only associations between variables and hence do not considerate confounding effects that may also influence blood pressure. In contrast, later used multiple regressions allow this adjustment.

**Table 5.6:** Means of the individual factors by hypertension status.

| Group | Dim 1 | Dim 2 | $p$-value |
|---|---|---|---|
| hta | -0.6466 | 0.0712 | 0.003 |
| no hta | 0.1970 | -0.0216 | 0.532 |

According to the $p$-values from the two-sided two-sample Student's tests, mean in dimension 1 is lower in *hta* individuals than in *no hta* individuals at level $\alpha = 5\%$ but difference of means between *hta* and *no hta* groups in dimension 2 is not statistically significant at level $\alpha = 5\%$.

hta: hypertension. No hta: no hypertension.

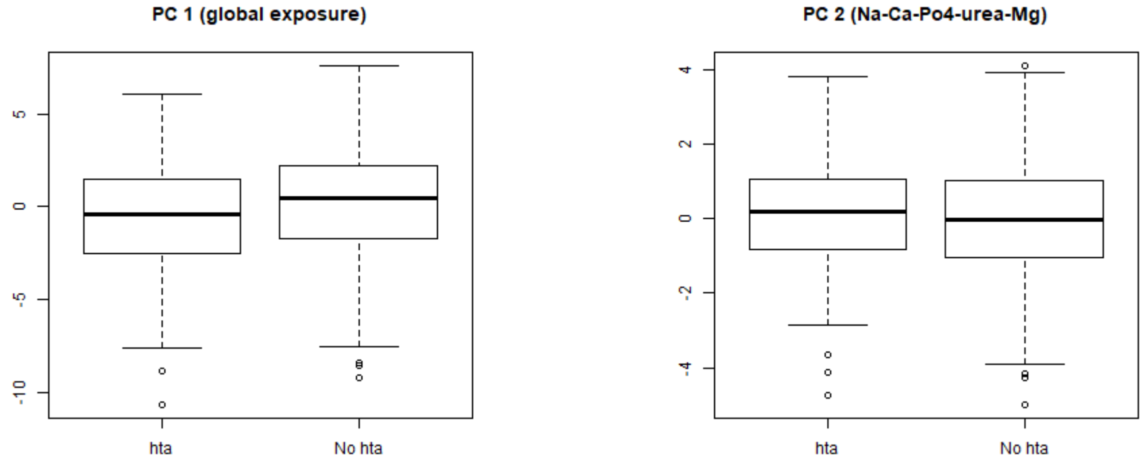


**Figure 5.11** Individual PC 1 and PC 2 scores by hypertension status.

hta: hypertension. No hta: no hypertension.

**Table 5.7:** Correlation matrix between the 1st PC, the 2nd PC, SBP and DBP.

| Variable | 1st PC | 2nd PC | SBP | DBP |
|---|---|---|---|---|
| 1st PC | 1.00 | | | |
| 2nd PC | 0.00 | 1.00 | | |
| SBP | **-0.12** | -0.02 | 1.00 | |
| DBP | 0.03 | **-0.09** | **0.67** | 1.00 |

Significant correlations at level $\alpha = 5\%$ are given in bold.

**Table 5.8:** Estimated regression coefficients from model $M1$.

| Variable | Estimate | Std. error | $t$-value | $P(\|T\| > \|t\|)$ | |
|---|---|---|---|---|---|
| (Intercept) | 78.849436 | 5.631900 | 14.001 | < 2e-16 | * |
| comp1 | -0.225694 | 0.200761 | -1.124 | 0.261383 | |
| comp2 | -0.064776 | 0.395809 | -0.164 | 0.870058 | |
| age | 0.385600 | 0.040689 | 9.477 | < 2e-16 | * |
| sex01M | 4.534503 | 1.342142 | 3.379 | 0.000776 | * |
| diabetes | 2.371990 | 3.379602 | 0.702 | 0.483043 | |
| tg | 1.786916 | 0.946166 | 1.889 | 0.059432 | |
| cho | 0.515524 | 0.657140 | 0.784 | 0.433060 | |
| glu | 2.486982 | 1.046957 | 2.375 | 0.017843 | * |
| ins2s | 0.008919 | 0.089712 | 0.099 | 0.920842 | |
| d_htaYes | 7.934874 | 1.753170 | 4.526 | 7.26e-06 | * |

An asterisk (*) indicates significant coefficients at level $\alpha = 5\%$.

### 5.3.2 Individual factors as independent variables

We performed three models incorporating individual factors as independent variables.

Models $M1$ and $M2$ are multiple linear regression models predicting SBP and DBP, respectively. First, two scores in $COMP1$ and $COMP2$ from PCA were used as independent variables in these two models. Both $M1$ and $M2$ are adjusted for participants' characteristics. Covariates come from the non-standardized log-transformed data as explained in Chapter 3 and models are also adjusted for hypertension treatment $d_{hta}$ since hypertension treatment lowers both SBP and DBP. Thereby, hypertension treatment has a confounding effect.

#### Model M1

The equation of $M1$ modeling SBP is given by:

$$\begin{aligned} SBP_i = {} & \beta_0 + \beta_1 \cdot COMP1_i + \beta_2 \cdot COMP2_i + \beta_3 \cdot age_i + \beta_4 \cdot sex01_i \\ & + \beta_5 \cdot diabetes_i + \beta_6 \cdot tg_i + \beta_7 \cdot cho_i + \beta_8 \cdot glu_i + \beta_9 \cdot ins2s_i \\ & + \beta_{10} \cdot d_{hta,i} + \epsilon_i. \end{aligned}$$

The multiple R-squared of this model is 0.3748 (adjusted R-squared = 0.3644). The estimates of $\beta$ coefficients are presented in Table 5.8.

For abbreviations please refer to *Appendix*. The Fisher test for model $M1$ is significant at level $\alpha = 0.05$ with an $F$-statistic = 35.79. As seen in Table 5.8, the Student's test is significant for variables $age$, $sex01M$, $glu$ and $d_{hta}$ at level $\alpha = 0.05$. Thus, unsurprisingly age and male sex predict a higher SBP prediction as reported in the medical literature. Glucose is also predicting a higher SBP. This can be explained by the fact that both diabetes and hypertension belong to the Metabolic Syndrome as defined in Chapter 2. Less intuitively, having antihypertensive treatment predicts a higher SBP. We can however state that people taking medications are less healthy and consequently more likely to develop hypertension.

**Table 5.9:** Estimated regression coefficients from model $M2$.

| Variable | Estimate | Std. error | $t$-value | $P(\|T\| > \|t\|)$ | |
|---|---|---|---|---|---|
| (Intercept) | 50.60413 | 3.61834 | 13.985 | < 2e-16 | * |
| comp1 | 0.12540 | 0.12898 | 0.972 | 0.331344 | |
| comp2 | -0.30491 | 0.25430 | -1.199 | 0.230996 | |
| age | 0.09867 | 0.02614 | 3.775 | 0.000176 | * |
| sex01M | 3.21195 | 0.86229 | 3.725 | 0.000214 | * |
| diabetes | -6.98053 | 2.17130 | -3.215 | 0.001375 | * |
| tg | 1.06510 | 0.60789 | 1.752 | 0.080263 | |
| cho | 0.97064 | 0.42219 | 2.299 | 0.021848 | * |
| glu | 2.25712 | 0.67264 | 3.356 | 0.000842 | * |
| ins2s | 0.01834 | 0.05764 | 0.318 | 0.750441 | |
| d_htaYes | 1.64266 | 1.12636 | 1.458 | 0.145264 | |

An asterisk (*) indicates significant coefficients at level $\alpha = 5\%$.

### Model M2

The equation of $M2$ modeling DBP is given by:

$$\begin{aligned} DBP_i = \beta_0 + \beta_1 \cdot COMP1_i + \beta_2 \cdot COMP2_i + \beta_3 \cdot age_i + \beta_4 \cdot sex01_i \\ + \beta_5 \cdot diabetes_i + \beta_6 \cdot tg_i + \beta_7 \cdot cho_i + \beta_8 \cdot glu_i + \beta_9 \cdot ins2s_i \\ + \beta_{10} \cdot d_{hta,i} + \epsilon_i. \end{aligned}$$

The multiple R-squared of model $M2$ is 0.1726 (adjusted R-squared 0.1587). Consequently, variability of DBP is less explained by our covariates than variability of SBP in model $M1$. The estimates of $\beta$ coefficients are presented in Table 5.9.

The Fisher test for model $M2$ is significant at level $\alpha = 0.05$ with an $F$-statistic=12.45. Again age and male sex contribute to a higher blood pressure prediction. Here, cholesterol has a significant positive coefficient at level $\alpha = 0.05$. This can be explained by the fact that high cholesterol levels belong to the Metabolic Syndrome. Suprinsingly, the results indicate that diabetes predicts lower DBP. There are two possible explanations:

- Non-significant covariates can make noise,
- Glucose interacts with diabetes. This could be physiologically possible since diabetics with higher glucose have more severe diabetes.

Consequently we computed a reduced version of model $M2$ that we called model $M2r$ with only significant covariates of model $M2$ at level $\alpha = 0.05$ and an interaction between glucose and diabetes. However we kept $COMP1$ and $COMP2$ in this new model since they are the covariates of interest. The equation of model $M2r$ is given by:

$$\begin{aligned} DBP_i = \beta_0 + \beta_1 \cdot COMP1_i + \beta_2 \cdot COMP2_i + \beta_3 \cdot age_i + \beta_4 \cdot sex01_i \\ + \beta_5 \cdot cho_i + \beta_6 \cdot glu_i + \beta_7 \cdot diabetes_i + \beta_8 \cdot glu_i \times diabetes_i \\ + \epsilon_i. \end{aligned}$$

The multiple R-squared of model $M2r$ is similar to multiple R-squared of model $M2$ with 0.1790 (adjusted R-squared 0.1680). The Fisher test is significant at level $\alpha = 0.05$ with an $F$-statistic=16.33. The estimates of $\beta$ coefficients are presented in Table 5.10.

**Table 5.10:** Estimated regression coefficients from model $M2r$.

| Variable | Estimate | Std. error | $t$-value | $P(\|T\| > \|t\|)$ | |
|---|---|---|---|---|---|
| (Intercept) | 42.76614 | 3.96295 | 10.791 | < 2e-16 | * |
| comp1 | 0.10582 | 0.12789 | 0.827 | 0.408329 | |
| comp2 | -0.35875 | 0.25167 | -1.425 | 0.154535 | |
| age | 0.09504 | 0.02457 | 3.868 | 0.000122 | * |
| sex01M | 3.35733 | 0.84546 | 3.971 | 8.03e-05 | * |
| cho | 1.14073 | 0.39600 | 2.881 | 0.004111 | * |
| glu | 3.89406 | 0.76967 | 5.059 | 5.60e-07 | * |
| diabetes | 25.57865 | 9.53225 | 2.683 | 0.007490 | * |
| glu:diabetes | -4.80809 | 1.39674 | -3.442 | 0.000617 | * |

An asterisk (*) indicates significant coefficients at level $\alpha = 5\%$.

It is now less surprising to see that both glucose and diabetes predict higher DBP in this model (positive coefficients). Interestingly, there is a significant negative interaction between glucose and diabetes which signifies that diabetic participants have a smaller increase in DBP by glucose increments. This can be explained by the arterial stiffness in people suffering from diabetes. Arteries in diabetic people are less likely to be elastic and consequently, their volume cannot adapt to pressure variations, leading to a greater difference between DBP and SBP readings. The result is a higher SBP but a lower DBP as measured by *pulse pressure* ($PP$):

$$PP = SBP - DBP.$$

Indeed, pulse pressure ($PP$) is higher in diabetics and has even been showed to increase hospitalization events in diabetics [42]. The negative regression coefficient for this interaction in model $M2r$ could be a result explained by pulse pressure. This effect could be investigated with models predicting $PP$ instead of $SBP$ and $DBP$ in diabetics. However, this is beyond the scope of this study.

Neither the $M1$ nor the $M2$ models show significant coefficients for $COMP1$ and $COMP2$. Classical risk factors for hypertension such as age, sex and diabetes are better predictors.

### Model M3

We tried a third model ($M3$) modeling the binary variable (logistic regression). The equation of model $M3$ is given by:

$$\begin{aligned} log\left(\frac{hta_i}{1 - hta_i}\right) &= \beta_0 + \beta_1 \cdot COMP1_i + \beta_2 \cdot COMP2_i + \beta_3 \cdot age_i + \beta_4 \cdot sex01_i \\ &+ \beta_5 \cdot diabetes_i + \beta_6 \cdot tg_i + \beta_7 \cdot cho_i + \beta_8 \cdot glu_i + \beta_9 \cdot ins2s_i, \end{aligned}$$

where $hta$ is the binary variable defining hypertension status (1: hypertension, 0: no hypertension). Note that model $M3$ is not adjusted for hypertension treatment since $hta$ is defined by $d_{hta}$ ($y$ is assumed to be random).

In model $M3$, null deviance is 660.94 on 607 degrees of freedom and residual deviance is 462.69 on 598 degrees of freedom. The likelihood ratio test of model $M3$ using the Chi-squared test has the $p$-value $< 0.05$, i.e. we reject the null hypothesis $H_0$ that the model with only the intercept is appropriate at level $\alpha = 0.05$. The estimates of $\beta$ coefficients are presented in Table 5.11.

In this logistic model, age and triglycerides have positive estimated coefficients. All other coefficients are non-significant, especially $COMP1$ and $COMP2$.

**Table 5.11:** Estimated Regression coefficients from model $M3$.

| Variable | Estimate | Std. error | $z$-value | $P(\|Z\| > \|z\|)$ | |
|---|---|---|---|---|---|
| (Intercept) | -6.630614 | 1.965623 | -3.373 | 0.000743 | * |
| comp1 | 0.026360 | 0.044178 | 0.597 | 0.550718 | |
| comp2 | 0.033556 | 0.085314 | 0.393 | 0.694077 | |
| age | 0.083365 | 0.009397 | 8.871 | < 2e-16 | * |
| sex01M | 0.171037 | 0.279701 | 0.611 | 0.540869 | |
| diabetes | 0.638388 | 0.558014 | 1.144 | 0.252609 | |
| tg | 0.689512 | 0.272210 | 2.533 | 0.011309 | * |
| cho | -0.161488 | 0.132422 | -1.219 | 0.222658 | |
| glu | 0.781633 | 1.170823 | 0.668 | 0.504393 | |
| ins2s | 0.219532 | 0.149054 | 1.473 | 0.140796 | |

An asterisk (*) indicates significant coefficients at level $\alpha = 5\%$.

### 5.3.3 Cluster ID as an independent variable

We performed multiple regressions with being in cluster 2 (*cluster2* below) as the independent binary variable.

Models $M4$ and $M5$ are linear regression models predicting SBP and DBP, respectively, and are adjusted for participants' characteristics. Covariates come from the non-standardized log-transformed data as explained in Chapter 3 and models are also adjusted for hypertension treatment $d_{hta}$. Model $M6$ is a logistic model predicting the binary variable (1: hypertension, 0: no hypertension).

#### Model M4

The equation of model $M4$ modeling SBP is given by:

$$\begin{aligned} SBP_i &= \beta_0 + \beta_1 \cdot cluster2 + \beta_2 \cdot age_i + \beta_3 \cdot sex01_i \\ &\quad + \beta_4 \cdot diabetes_i + \beta_5 \cdot tg_i + \beta_6 \cdot cho_i + \beta_7 \cdot glu_i + \beta_8 \cdot ins2s_i \\ &\quad + \beta_9 \cdot d_{hta,i} + \epsilon_i. \end{aligned}$$

The multiple R-squared of this model is 0.3898 (adjusted R-squared = 0.3806). The estimates of $\beta$ coefficients are presented in Table 5.12.

The Fisher test for model $M4$ is significant at level $\alpha = 0.05$ with an $F$-statistic = 42.44. As seen in Table 5.12, the Student's test is significantly positive for 6 variables that were already discussed in previous models. However, the estimated coefficient for $cluster2$ is not significant at level $\alpha = 0.05$, thereby suggesting no effect of being in cluster 2 on SBP.

#### Model M5

The equation of $M5$ modeling DBP is given by:

$$\begin{aligned} DBP_i &= \beta_0 + \beta_1 \cdot cluster2 + \beta_2 \cdot age_i + \beta_3 \cdot sex01_i \\ &\quad + \beta_4 \cdot diabetes_i + \beta_5 \cdot tg_i + \beta_6 \cdot cho_i + \beta_7 \cdot glu_i + \beta_8 \cdot ins2s_i \\ &\quad + \beta_9 \cdot d_{hta,i} + \epsilon_i. \end{aligned}$$

The multiple R-squared of this model is 0.1878 (adjusted R-squared = 0.1756). The estimates of $\beta$ coefficients are presented in Table 5.13.

**Table 5.12:** Estimated regression coefficients from model $M4$.

| Variable | Estimate | Std. error | $t$-value | $P(\|T\| > \|t\|)$ | |
|---|---|---|---|---|---|
| (Intercept) | 74.14797 | 9.51703 | 7.791 | 2.95e-14 | * |
| cluster2 | -0.70356 | 1.26813 | -0.555 | 0.57924 | |
| age | 0.38766 | 0.04002 | 9.688 | < 2e-16 | * |
| sex01M | 4.18226 | 1.30988 | 3.193 | 0.00148 | * |
| diabetes | 1.88478 | 3.14829 | 0.599 | 0.54962 | |
| tg | 2.75999 | 1.26919 | 2.175 | 0.03005 | * |
| cho | 0.15756 | 0.66226 | 0.238 | 0.81203 | |
| glu | 12.08047 | 5.87012 | 2.058 | 0.04003 | * |
| ins2s | 1.64967 | 0.69877 | 2.361 | 0.01855 | * |
| d_htaYes | 7.45125 | 1.72414 | 4.322 | 1.81e-05 | * |

An asterisk (*) indicates significant coefficients at level $\alpha = 5\%$.

**Table 5.13:** Estimated regression coefficients from model $M5$.

| Variable | Estimate | Std. error | $t$-value | $P(\|T\| > \|t\|)$ | |
|---|---|---|---|---|---|
| (Intercept) | 42.27783 | 6.13165 | 6.895 | 1.37e-11 | * |
| cluster2 | 1.32460 | 0.81703 | 1.621 | 0.105494 | |
| age | 0.09046 | 0.02578 | 3.509 | 0.000484 | * |
| sex01M | 2.97388 | 0.84393 | 3.524 | 0.000458 | * |
| diabetes | -6.91662 | 2.02838 | -3.410 | 0.000694 | * |
| tg | 1.57705 | 0.81772 | 1.929 | 0.054254 | |
| cho | 0.80436 | 0.42668 | 1.885 | 0.059891 | |
| glu | 13.02165 | 3.78201 | 3.443 | 0.000615 | * |
| ins2s | 0.69425 | 0.45020 | 1.542 | 0.123583 | |
| d_htaYes | 1.39602 | 1.11083 | 1.257 | 0.209340 | |

An asterisk (*) indicates significant coefficients at level $\alpha = 5\%$.

**Table 5.14:** Estimated Regression coefficients from model $M6$.

| Variable | Estimate | Std. error | $z$-value | $P(\|Z\| > \|z\|)$ | |
|---|---|---|---|---|---|
| (Intercept) | -6.807494 | 1.969900 | -3.456 | 0.000549 | * |
| cluster2 | 0.122215 | 0.275414 | 0.444 | 0.657224 | |
| age | 0.082373 | 0.009025 | 9.127 | < 2e-16 | * |
| sex01M | 0.121431 | 0.278675 | 0.436 | 0.663023 | |
| diabetes | 0.623978 | 0.556553 | 1.121 | 0.262225 | |
| tg | 0.683021 | 0.271457 | 2.516 | 0.011865 | * |
| cho | -0.165296 | 0.132238 | -1.250 | 0.211304 | |
| glu | 0.938256 | 1.158391 | 0.810 | 0.417961 | |
| ins2s | 0.196607 | 0.146601 | 1.341 | 0.179889 | |

An asterisk (*) indicates significant coefficients at level $\alpha = 5\%$.

The Fisher test for model $M5$ is significant at level $\alpha = 0.05$ with an $F$-statistic = 15.36. As seen in Table 5.13, the Student's test is significantly positive for 4 variables but the estimated coefficient for $cluster2$ is not significant at level $\alpha = 0.05$, thereby suggesting no effect of being in cluster 2 on DBP.

### Model M6

We tried a sixth model ($M6$) modeling the binary variable (logistic regression). The equation of model $M6$ is given by:

$$
\begin{aligned}
log\left(\frac{hta_i}{1 - hta_i}\right) &= \beta_0 + \beta_1 \cdot cluster2 + \beta_2 \cdot age_i + \beta_3 \cdot sex01_i \\
&+ \beta_4 \cdot diabetes_i + \beta_5 \cdot tg_i + \beta_6 \cdot cho_i + \beta_7 \cdot glu_i + \beta_8 \cdot ins2s_i,
\end{aligned}
$$

where $hta$ is the binary variable defining hypertension status (1: hypertension, 0: no hypertension). In model $M6$, null deviance is 660.94 on 607 degrees of freedom and residual deviance is 463.07 on 599 degrees of freedom. Likelihood ratio test of model $M6$ using the Chi-squared test has the $p$-value $< 0.05$. The estimated of $\beta$ coefficients are presented in Table 5.14.

The Wald test is significat for 2 variables but the estimated coefficient for $cluster2$ is not significant at level $\alpha = 0.05$, thereby suggesting no effect of being in cluster 2 on the binary outcome.

## 5.4 Other regression models

We were also interested in comparing previously found models ($M1$, $M2$, $M3$, $M4$, $M5$ and $M6$) with other multiple regression models. We created the models $MA$ modelling SBP, $MB$ modelling DBP and $MC$ modelling the binary variable $hta$. These models were selected using backward elimination steps with the goal of minimizing the Akaike information criterion (AIC). Full model contained all variables in *urine* data except SBP, DBP and the binary outcome $hta$.

### Model MA

$MA$ models SBP. After the minimizing process, AIC passed from 4921 ($df = 38$) to 4888 ($df = 16$). The multiple R-squared of this model is 0.4264 (adjusted R-squared = 0.4129). The estimates of $\beta$ coefficients

**Table 5.15:** Estimated regression coefficients from model *MA*.

| Variable | Estimate | Std. error | $t$-value | $P(\|T\| > \|t\|)$ | |
|---|---|---|---|---|---|
| (Intercept) | 75.73552 | 15.83771 | 4.782 | 2.19e-06 | * |
| age | 0.39657 | 0.04375 | 9.065 | < 2e-16 | * |
| sex01M | 3.58861 | 1.23375 | 2.909 | 0.00377 | * |
| d_htaYes | 6.81367 | 1.67018 | 4.080 | 5.13e-05 | * |
| tg | 2.96728 | 1.16086 | 2.556 | 0.01083 | * |
| glu | 8.82544 | 5.36468 | 1.645 | 0.10048 | |
| ins2s | 1.51136 | 0.69479 | 2.175 | 0.03000 | * |
| V_h24_excr | 6.37188 | 1.54939 | 4.113 | 4.47e-05 | * |
| Ni_h24_excr | -1.90769 | 1.04873 | -1.819 | 0.06941 | |
| Zn_h24_excr | 2.00974 | 1.02100 | 1.968 | 0.04949 | * |
| Mo_h24_excr | -2.54884 | 1.01466 | -2.512 | 0.01227 | * |
| Cd_h24_excr | -1.85876 | 1.15642 | -1.607 | 0.10851 | |
| Pt_h24_excr | -2.48081 | 0.90931 | -2.728 | 0.00656 | * |
| Bi_h24_excr | 0.73415 | 0.48182 | 1.524 | 0.12812 | |
| umg_u24_mmol | -0.62031 | 0.38567 | -1.608 | 0.10828 | |

An asterisk (*) indicates significant coefficients at level $\alpha = 5\%$.

are presented in Table 5.15. 14 variables were selected using backward elimination steps minimizing the AIC. Over these variables, 7 are metal-mixtures: vanadium, nickel, zinc, molybdenum, cadmium, platinum and bismuth.

For abbreviations please refer to *Appendix*. The Fisher test for model *MA* is significant at level $\alpha = 0.05$ with an $F$-statistic = 31.49. As evidenced in Table 5.15, vanadium and zinc have significant positive coefficients at level $\alpha = 0.05$ and molybdenum and platinum have significant negative coefficients at level $\alpha = 0.05$.

### Model MB

*MB* models DBP. After the minimizing process, AIC passed from 4418 ($df = 38$) to 4379 ($df = 13$). The Multiple R-squared of this model is 0.1948 (adjusted R-squared = 0.1799). This supports the earlier finding which indicates that it is more difficult to model SBP than DBP. The estimates of $\beta$ coefficients are presented in Table 5.16. 11 variables were selected using backward elimination steps minimizing the AIC. Over these variables, only 3 are metal-mixtures: nickel, antimony and bismuth. However, none of them are significant at level $\alpha = 0.05$.

The Fisher test for model *MB* is significant at level $\alpha = 0.05$ with an $F$-statistic = 13.11.

### Model MC

*MC* models the binary outcome. After the minimizing process, AIC passed from 511 ($df = 36$) to 472 ($df = 11$). Null deviance is 660.94 on 607 degrees of freedom and residual deviance is 449.87 on 597 degrees of freedom. The likelihood ratio test of model *MC* using the Chi-squared test has the $p$-value $< 0.05$, i.e. we reject the null hypothesis $H_0$ that the model with only the intercept is appropriate at level $\alpha = 0.05$. The estimates of $\beta$ coefficients are presented in Table 5.17.

In model *MC*, age, triglycerides and zinc are increasing the predicted probablity to have hypertension according to their significant estimated coefficients at level $\alpha = 5\%$.

**Table 5.16:** Estimated regression coefficients from model $MB$.

| Variable | Estimate | Std. error | $t$-value | $P(\|T\| > \|t\|)$ | |
|---|---|---|---|---|---|
| (Intercept) | 47.560254 | 8.113291 | 5.862 | 7.58e-09 | * |
| age | 0.110698 | 0.024187 | 4.577 | 5.75e-06 | * |
| sex01M | 2.923871 | 0.840597 | 3.478 | 0.000541 | * |
| diabetes | -6.314285 | 2.023708 | -3.120 | 0.001895 | * |
| tg | 1.920339 | 0.814985 | 2.356 | 0.018781 | * |
| cho | 0.693786 | 0.423921 | 1.637 | 0.102243 | |
| glu | 10.152343 | 3.913886 | 2.594 | 0.009722 | * |
| ins2s | 0.901545 | 0.469622 | 1.920 | 0.055369 | |
| Ni_h24_excr | -0.980218 | 0.625635 | -1.567 | 0.117702 | |
| Sb_h24_excr | 0.716485 | 0.424483 | 1.688 | 0.091954 | |
| Bi_h24_excr | 0.488656 | 0.315137 | 1.551 | 0.121526 | |
| uure_u24_mmol | 0.006593 | 0.003662 | 1.800 | 0.072326 | |

An asterisk (*) indicates significant coefficients at level $\alpha = 5\%$.

**Table 5.17:** Estimated regression coefficients from model $MC$.

| Variable | Estimate | Std. error | $z$-value | $P(\|Z\| > \|z\|)$ | |
|---|---|---|---|---|---|
| (Intercept) | -11.414821 | 2.878483 | -3.966 | 7.32e-05 | * |
| age | 0.091363 | 0.009714 | 9.405 | $<$ 2e-16 | * |
| diabetes | 0.774399 | 0.520395 | 1.488 | 0.136725 | |
| tg | 0.868417 | 0.251146 | 3.458 | 0.000545 | * |
| cho | -0.192934 | 0.133184 | -1.449 | 0.147441 | |
| Zn_h24_excr | 0.473567 | 0.209705 | 2.258 | 0.023930 | * |
| Ag_h24_excr | -0.192466 | 0.109058 | -1.765 | 0.077598 | |
| Pt_h24_excr | -0.374779 | 0.205700 | -1.822 | 0.068460 | |
| Tl_h24_excr | 0.422767 | 0.268637 | 1.574 | 0.115546 | |
| uk_u24_mmol | 0.011549 | 0.006583 | 1.754 | 0.079376 | |
| umg_u24_mmol | -0.161155 | 0.094318 | -1.709 | 0.087518 | |

An asterisk (*) indicates significant coefficients at level $\alpha = 5\%$.

### Discussion on multiple regressions

Models $MA$, $MB$ and $MC$ are better models than the previous models in term of R-squared for linear regressions and likelihood ratio tests for logistic regressions.

We found several significant coefficients for metal-mixtures at level $\alpha = 5\%$ but none for sodium (una_u24_mmol) that is a component of salt which is a classical risk factor for hypertension. This highlights the fact that metal-mixtures are potential risk factors for hypertension. In model $MA$, vanadium and zinc predicted a higher SBP. Both vanadium and zinc are essential metal-mixtures. Moreover, the predicted probability to have hypertension is higher if zinc concentrations increase according to our logistic model ($MC$). Recently, a Chinese study has shown that participants in the highest quartiles of vanadium and zinc (and some other metal-mixtures) had higher odd ratios for hypertension compared with those in the lowest quartiles [40]. Although normal levels of essential-mixtures are important to maintain a normal blood pressure, the authors caution that excessive accumulation of them can lead to hypertension.

# Chapter 6

# Discussion

In the SKIPOGH data, 23.3% of participants had hypertension. This is slightly lower than the worldwide prevalence estimated to 26.4%. Exploration of variable distributions showed very skewed distributions with extreme outliers. However, outliers were kept in the exploratory analysis since all were nevertheless realistic concentrations. Interestingly, principal component analysis (PCA) could help to distinguish between metal-mixtures and non-metals. Thus, despite competition that often occurs between biochemical elements in physiological processes — notably between elements sharing the same valence — metal-mixtures and non-metals seem to be uncorrelated. Moreover, essential metal-mixtures were all isolated in the upper right quadrant, again suggesting different sources in dietary intakes or different physiological processes compared to other biochemical elements. However, multiple regression incorporating scores of principal components and adjusted for participants characteristics failed to predict hypertension. Then, two clusters were generated from the *k*-medoids algorithm. They displayed different biochemical patterns. However, silhouette width was small and clusters could not predict hypertension in multiple regressions adjusted for participants characteristics.

In supplementary analyses, we performed other regression models using observed covariates with backward elimination of variables minimizing the Akaike information criterion (AIC). These models performed better than those incorporating PCA and cluster information. Moreover, vanadium and zinc predicted higher SBP, which is consistent with the literature. Moreover, the predicted probability to develop hypertension was higher with increments of zinc concentrations in a logistic regression. This underlines the fact that essential metal-mixtures can lead to hypertension when they are in excessive quantities. However, significant estimated coefficients were not found for mercury and cadmium which are toxic metal-mixtures with cardiovascular effects [9].

There are several reasons for finding no significant estimated coefficient for biochemical patterns in the first analyses. Firstly, it is highly likely that main metal-mixtures and non-metals simply contribute little to hypertension compared to classical risk factors for hypertension such as age, male sex and diabetes. Secondly, significant effects can be hidden by non-significant effects in PCA and cluster analysis methods that incorporate all the variables. Restricted choice of variables from medical expertise, adapted regression methods such as minimizing AIC or Lasso and Ridge regressions can deal situations with variables that participate not to the total variance. Moreover, SBP and DBP — that are not independent between each other — could be studied together instead of separately in more complex models such as neural networks allowing multiple outputs. Thirdly, longitudinal studies may be more sensitive to detect an association between metal-mixtures and hypertension.

In conclusion, if heterogeneous biochemical patterns could be distinguished in the SKIPOGH data, they could predict neither systolic, diastolic blood pressure nor hypertension cases using multiple regressions. Restricted variable selection for multiple regression models gave better models.

# Bibliography

[1] R bloggers: clustering mixed data types in R. url: https://www.r-bloggers.com/clustering-mixed-data-types-in-r/, Accessed: 2018-10-29.

[2] Abdi, H., and Williams, L. J. Principal component analysis: Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics 2*, 4 (July 2010), 433–459.

[3] Abhyankar, L. N., Jones, M. R., Guallar, E., and Navas-Acien, A. Arsenic Exposure and Hypertension: A Systematic Review. *Environmental Health Perspectives 120*, 4 (Dec. 2011), 494–500.

[4] Adrogué, H. J., and Madias, N. E. Sodium surfeit and potassium deficit: keys to the pathogenesis of hypertension. *Journal of the American Society of Hypertension: JASH 8*, 3 (Mar. 2014), 203–213.

[5] Alwan, H., Pruijm, M., Ponte, B., Ackermann, D., Guessous, I., Ehret, G., Staessen, J. A., Asayama, K., Vuistiner, P., Younes, S. E., Paccaud, F., Wuerzner, G., Pechere-Bertschi, A., Mohaupt, M., Vogt, B., Martin, P.-Y., Burnier, M., and Bochud, M. Epidemiology of Masked and White-Coat Hypertension: The Family-Based SKIPOGH Study. *PLoS ONE 9*, 3 (Mar. 2014).

[6] Atkins, P. W., Jones, L. L., and Pousse, A. *Chimie: molécules, matière, métamorphoses*. De Boeck, Bruxelles, 2007.

[7] Bondonno, C. P., Liu, A. H., Croft, K. D., Ward, N. C., Shinde, S., Moodley, Y., Lundberg, J. O., Puddey, I. B., Woodman, R. J., and Hodgson, J. M. Absence of an effect of high nitrate intake from beetroot juice on blood pressure in treated hypertensive individuals: a randomized controlled trial. *The American Journal of Clinical Nutrition 102*, 2 (Aug. 2015), 368–375.

[8] Carlström, M., Persson, A. E. G., Larsson, E., Hezel, M., Scheffer, P. G., Teerlink, T., Weitzberg, E., and Lundberg, J. O. Dietary nitrate attenuates oxidative stress, prevents cardiac and renal injuries, and reduces blood pressure in salt-induced hypertension. *Cardiovascular Research 89*, 3 (Feb. 2011), 574–585.

[9] Cosselman, K. E., Navas-Acien, A., and Kaufman, J. D. Environmental factors in cardiovascular disease. *Nature Reviews Cardiology 12*, 11 (Nov. 2015), 627–642.

[10] Dolnicar, S., and Leisch, F. Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Marketing Letters 21*, 1 (Mar. 2010), 83–101.

[11] Elliott, P., Stamler, J., Nichols, R., Dyer, A. R., Stamler, R., Kesteloot, H., and Marmot, M. Intersalt revisited: further analyses of 24 hour sodium excretion and blood pressure within and across populations. Intersalt Cooperative Research Group. *BMJ (Clinical research ed.) 312*, 7041 (May 1996), 1249–1253.

[12] Ezzati, M., Lopez, A. D., Rodgers, A., Vander Hoorn, S., and Murray, C. J. Selected major risk factors and global and regional burden of disease. *The Lancet 360*, 9343 (Nov. 2002), 1347–1360.

[13] Faurschou, A., Menné, T., Johansen, J. D., and Thyssen, J. P. Metal allergen of the 21st century-a review on exposure, epidemiology and clinical manifestations of palladium allergy. *Contact Dermatitis 64*, 4 (Apr. 2011), 185–195.

[14] Galtier-Boissière. *Larousse médical illustré*, The libraire Larousse, Paris ed. 1912.

[15] Hennig, C., Meila, M., Murtagh, F., and Rocci, R., Eds. *Handbook of cluster analysis*. Chapman & Hall/CRC handbooks of modern statistical methods. CRC Press, a Chapman & Hall book, Boca Raton London New York, 2016.

[16] Hutcheon, J. A., Chiolero, A., and Hanley, J. A. Random measurement error and regression dilution bias. *BMJ 340*, jun23 2 (June 2010), c2289–c2289.

[17] James, G., Witten, D., Hastie, T., and Tibshirani, R., Eds. *An introduction to statistical learning: with applications in R*. No. 103 in Springer texts in statistics. Springer, New York, 2013.

[18] Jones, M. R., Tellez-Plaza, M., Sharrett, A. R., Guallar, E., and Navas-Acien, A. Urine Arsenic and Hypertension in US Adults: The 2003–2008 National Health and Nutrition Examination Survey. *Epidemiology 22*, 2 (Mar. 2011), 153–161.

[19] Kasper, D. L., Ed. *Harrison's principles of internal medicine*, 19th edition / editors, Dennis L. Kasper, MD, William Ellery Channing, Professor of Medicine, Professor of Microbiology, Department of Microbiology and Immunobiology, Harvard Medical School, Division of Infectious Diseases, Brigham and Women's Hospital, Boston, Massachusetts [and five others] ed. McGraw Hill Education, New York, 2015.

[20] Kaufman, L., and Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*. Wiley series in probability and mathematical statistics. Wiley, New York, 1990.

[21] Kearney, P. M., Whelton, M., Reynolds, K., Muntner, P., Whelton, P. K., and He, J. Global burden of hypertension: analysis of worldwide data. *The Lancet 365*, 9455 (Jan. 2005), 217–223.

[22] Kearney, P. M., Whelton, M., Reynolds, K., Whelton, P. K., and He, J. Worldwide prevalence of hypertension: a systematic review. *Journal of Hypertension 22*, 1 (Jan. 2004), 11–19.

[23] Killip, S., Bennett, J. M., and Chambers, M. D. Iron deficiency anemia. *American Family Physician 75*, 5 (Mar. 2007), 671–678.

[24] Last, J. M., and Porta, M. *A dictionary of public health*. 2018.

[25] Mente, A., O'Donnell, M. J., Rangarajan, S., McQueen, M. J., Poirier, P., Wielgosz, A., Morrison, H., Li, W., Wang, X., Di, C., Mony, P., Devanath, A., Rosengren, A., Oguz, A., Zatonska, K., Yusufali, A. H., Lopez-Jaramillo, P., Avezum, A., Ismail, N., Lanas, F., Puoane, T., Diaz, R., Kelishadi, R., Iqbal, R., Yusuf, R., Chifamba, J., Khatib, R., Teo, K., and Yusuf, S. Association of Urinary Sodium and Potassium Excretion with Blood Pressure. *New England Journal of Medicine 371*, 7 (Aug. 2014), 601–611.

[26] O'Brien, E., Parati, G., Stergiou, G., Asmar, R., Beilin, L., Bilo, G., Clement, D., de la Sierra, A., de Leeuw, P., Dolan, E., Fagard, R., Graves, J., Head, G. A., Imai, Y., Kario, K., Lurbe, E., Mallion, J.-M., Mancia, G., Mengden, T., Myers, M., Ogedegbe, G., Ohkubo, T., Omboni, S., Palatini, P., Redon, J., Ruilope, L. M., Shennan, A., Staessen, J. A., vanMontfrans, G., Verdecchia, P., Waeber, B., Wang, J., Zanchetti, A., and Zhang, Y. European Society of Hypertension Position Paper on Ambulatory Blood Pressure Monitoring:. *Journal of Hypertension 31*, 9 (Sept. 2013), 1731–1768.

[27] O'Sullivan, A., Gibney, M. J., and Brennan, L. Dietary intake patterns are reflected in metabolomic profiles: potential role in dietary assessment studies. *The American Journal of Clinical Nutrition 93*, 2 (Feb. 2011), 314–321.

[28] Panchal, S. K., Wanyonyi, S., and Brown, L. Selenium, Vanadium, and Chromium as Micronutrients to Improve Metabolic Syndrome. *Current Hypertension Reports 19*, 3 (Mar. 2017), 10.

[29] Pang, Y., Peng, R. D., Jones, M. R., Francesconi, K. A., Goessler, W., Howard, B. V., Umans, J. G., Best, L. G., Guallar, E., Post, W. S., Kaufman, J. D., Vaidya, D., and Navas-Acien, A. Metal mixtures in urban and rural populations in the US: The Multi-Ethnic Study of Atherosclerosis and the Strong Heart Study. *Environmental Research 147* (May 2016), 356–364.

[30] Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association 66*, 336 (Dec. 1971), 846.

[31] Shalizi, C. *Advanced Data Analysis from an Elementary Point of View. Chap. 16. 2018. URL: http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/. Accessed: 2018-12-29.*

[32] Siervo, M., Lara, J., Ogbonmwan, I., and Mathers, J. C. Inorganic Nitrate and Beetroot Juice Supplementation Reduces Blood Pressure in Adults: A Systematic Review and Meta-Analysis. *The Journal of Nutrition 143*, 6 (June 2013), 818–826.

[33] Smallwood, M. J., Ble, A., Melzer, D., Winyard, P. G., Benjamin, N., Shore, A. C., and Gilchrist, M. Relationship Between Urinary Nitrate Excretion and Blood Pressure in the InChianti Cohort. *American Journal of Hypertension 30*, 7 (July 2017), 707–712.

[34] Soni, M. G., White, S. M., Flamm, W. G., and Burdock, G. A. Safety evaluation of dietary aluminum. *Regulatory toxicology and pharmacology: RTP 33*, 1 (Feb. 2001), 66–79.

[35] Struyf, A., Hubert, M., and Rousseeuw, P. Clustering in an object-oriented environment. *Journal of Statistical Software, Articles 1*, 4 (1997), 1–30.

[36] Szpunar, J. Advances in analytical methodology for bioinorganic speciation analysis: metallomics, metalloproteomics and heteroatom-tagged proteomics and metabolomics. *The Analyst 130*, 4 (2005), 442.

[37] Tahrani, A. A., Bailey, C. J., Del Prato, S., and Barnett, A. H. Management of type 2 diabetes: new and future developments in treatment. *The Lancet 378*, 9786 (July 2011), 182–197.

[38] Tillé, Y. *Multivariate Statistics, Course at the University of Neuchâtel*, 2016 ed.

[39] Wiseman, C. L., and Zereini, F. Airborne particulate matter, platinum group elements and human health: A review of recent evidence. *Science of The Total Environment 407*, 8 (Apr. 2009), 2493–2500.

[40] Wu, W., Jiang, S., Zhao, Q., Zhang, K., Wei, X., Zhou, T., Liu, D., Zhou, H., Zeng, Q., Cheng, L., Miao, X., and Lu, Q. Environmental exposure to metals and the risk of hypertension: A cross-sectional study in China. *Environmental Pollution 233* (Feb. 2018), 670–678.

[41] Yang, Q., Liu, T., Kuklina, E. V., Flanders, W. D., Hong, Y., Gillespie, C., Chang, M.-H., Gwinn, M., Dowling, N., Khoury, M. J., and Hu, F. B. Sodium and potassium intake and mortality among US adults: prospective data from the Third National Health and Nutrition Examination Survey. *Archives of Internal Medicine 171*, 13 (July 2011), 1183–1191.

[42] Yu, D., and Simmons, D. Association between pulse pressure and risk of hospital admissions for cardiovascular events among people with Type 2 diabetes: a population-based case-control study. *Diabetic Medicine 32*, 9 (Sept. 2015), 1201–1206.

# Appendices

# Abbreviations

```
cpnbr: participant number
age: age, years
sex01: sex (1=male)
diabetes: diabetes (1=yes)
d_hta: anti-hypertensive drugs, (1=yes)
tg: triglycerides, mmol/l
cho: cholesterol, mmol/l
glu: blood glucose, mmol/l
ins2s: blood insulin, microIU/ml
Li_h24_excr: lithium, ng
Be_h24_excr: beryllium, ng
Al_h24_excr: aluminum, ng
V_h24_excr: vanadium, ng
Cr_h24_excr: chrome, ng
Mn_h24_excr: manganese, ng
Co_h24_excr: cobalt, ng
Ni_h24_excr: nickel, ng
Cu_h24_excr: copper, ng
Zn_h24_excr: zinc, ng
As_h24_excr: arsenic, ng
Mo_h24_excr: molybdenum, ng
Pd_h24_excr: palladium, ng
Ag_h24_excr: silver, ng
Cd_h24_excr: cadmium, ng
Sn_h24_excr: tin, ng
Sb_h24_excr: antimony, ng
Pt_h24_excr: platinum, ng
Hg_h24_excr: mercury, ng
Tl_h24_excr: thallium, ng
Pb_h24_excr: lead, ng
Bi_h24_excr: bismuth, ng
una_u24_mmol: sodium, mmol
uk_u24_mmol: potassium, mmol
uca_u24_mmol: calcium, mmol
upo4_u24_mmol: phosphate, mmol
uure_u24_mmol: urea, mmol
umg_u24_mmol: magnesium, mmol
sbp_mean: systolic blood pressure, mmHg
dbp_mean: diastolic blood pressure, mmHg
hta: medical hypertension (1=yes)
```

# R-code: PAM program

```
# PAM PROGRAM IN R

# I) Enter the inital cluster configuration
set.seed(123)
data<-runif(20)
x<-data[1:10]
y<-data[11:20]
X<-round(cbind(x,y),2)
X<-cbind(X,c(0,0,0,0,0,0,1,0,1,0))
# We added a third column to specify by "1"
# which observations are initial medoids.
colnames(X)[3]<-"m"
X # X is the initial cluster configuration

# II) create a cost function in function of the
#     current cluster configuration X
cost_fun<-function(X){
  N<-nrow(X)
  K<-sum(X[,3])
  mi<-which(X[,3]==1)
  oi<-which(X[,3]==0)
  costs<-rep(NA,time=N)
  D<-as.matrix(dist(X[,-3],method="euclidean",upper = T,diag=T))
  for(i in 1:N){
    costs[i]<-min(D[i,mi])
  }
  result<-sum(costs)
  return(result)
}

# III)  using cost function, return the next cluster
#       configuration in functon of current configuration
one_it_PAM<-function(X){
  N<-nrow(X)
  K<-sum(X[,3])
  mi<-which(X[,3]==1)
  oi<-which(X[,3]==0)
  ci<-cost_fun(X)
  cost<-rep(NA,time=K*(N-K))
  C<-cbind(expand.grid(mi,oi),cost)
  nswap<-nrow(C)
  for(i in 1:nswap){
    X_swap<-X
    X_swap[C[i,2],3]<-1
    X_swap[C[i,1],3]<-0
    C[i,3]<-cost_fun(X_swap)
  }
  best_swap<-which.min(C[,3])
  X_best<-X
  X_best[C[best_swap,2],3]<-1
  X_best[C[best_swap,1],3]<-0
  if(cost_fun(X_best)<cost_fun(X)) return(X_best)
    else return(X)
}

#############################################
#############################################
#############################################

# Example
X
X2<-one_it_PAM(X)
X2
X3<-one_it_PAM(X2)
X3
X4<-one_it_PAM(X3)
X4 # stabilisation after 2 iterations
```

# R-code: whole code

```
#################################################
# MASTER THESIS DENIS DERIAZ 2018               #
#                                               #
# APPLICATION ON BIOCHEMICAL CLUSTER DETECTION  #
# PROFESSOR: ALINA MATEI                        #
#################################################

#Working directory
setwd("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/urine_analysis20181005")

#Packages used
library(psych) # for LaTeX tables
library(ggplot2) # for correlation matrix
library(ggcorrplot) # for correlation matrix
library(FactoMineR) # for PCA
library(cluster) # for gower similarity and pam
library(flexclust) #randIndex() function
library(car) #levene test

########################################################################
########################################################################
####################################################### DATA PREPARATION

#Opening row data: data_urine.txt
XU<-read.table("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/data/20181008_
    SKIPOGH/data_urine1.txt")
#data from "Master_thesis/script00000000.R"

#Missing values
nrow(XU) #1254 obs
missing_count<-XU[1,]
for(i in 1:ncol(missing_count)) missing_count[i]<-length(which(is.na(XU[,i])==T))
t(missing_count)
XUnona<-na.omit(XU)
nrow(XU)     # 1254
nrow(XUnona) # 609

#Histograms of continuous distributions
var_cont<-c(2,6:39)
for(k in var_cont){
  hist(XUnona[,k],main=colnames(XUnona)[k])
}

#Removing max insulin obs
max(XUnona[,9])
```

```
which(XUnona[,9]==max(XUnona[,9])) #obs 218
XUnona<-XUnona[-218,]
dim(XUnona)

#We log-transform graphically positive skewed distr
colnames(XUnona) #10:31
XUnonal<-XUnona
for(i in c(10:31)){
  XUnonal[,i]<-log(XUnona[,i])
}

#Checking again for continuous distributions
var_cont<-c(2,6:39)
for(k in var_cont){
  hist(XUnonal[,k],main=colnames(XUnonal)[k])
}

#Standardizing: mean0 sd1
XUnonalstd<-XUnonal[c(10:37)]
for(i in 1:ncol(XUnonalstd)){
  XUnonalstd[,i]<-(XUnonalstd[,i]-mean(XUnonalstd[,i]))/sd(XUnonalstd[,i])
}
XUnonalstd<-cbind(XUnonalstd,XUnona[,40])
dim(XUnonalstd)

#######################################################################
#######################################################################
############################################## VARIABLE CHARACTERISTICS
var_cont<-c(2,6:39)  # label for continuous var
var_cat<-c(3,4,5,40) # label for categorical var

prettynames<-c("participant’s␣number","age,␣years",
               "sex␣(1=male)", "diabetes␣(1=yes)",
               "anti-hypertensive␣drugs,␣(1=yes)",
               "triglycerides,␣mmol/l","cholesterol,␣mmol/l",
               "blood␣glucose,␣mmol/l","blood␣insulin,␣microIU/ml",
               "lithium,␣ng", "beryllium,␣ng", "aluminum,␣ng",
               "vanadium,␣ng", "chrome,␣ng", "manganese,␣ng",
               "cobalt,␣ng", "nickel,␣ng","copper,␣ng",
               "zinc,␣ng","arsenic,␣ng","molybdenum,␣ng",
               "palladium,␣ng","silver,␣ng","cadmium,␣ng",
               "tin,␣ng", "antimony,␣ng", "platinum,␣ng",
               "mercury,␣ng", "thallium,␣ng", "lead,␣ng", "bismuth,␣ng",
               "sodium,␣mmol", "potassium,␣mmol", "calcium,␣mmol",
               "phosphate,␣mmol", "urea,␣mmol","magnesium,␣mmol",
               "systolic␣blood␣pressure,␣mmHg",
               "diastolic␣blood␣pressure,␣mmHg",
               "medical␣hypertension␣(1=yes)")

#Description table for continuous var
colnames(XUnona)<-prettynames
des_var_cont<-round(describe(XUnona[var_cont],IQR=T),1)[,-c(1,2,6,7,8,9,10,13)]
des_var_cont<-des_var_cont[c(1,2,3,6,4,5)]
#write.table(des_var_cont,"C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis
    /des_cont_var.txt",eol="\\\\n",quote=F,sep="&",
#            row.names = T,col.names = F)

# Illustration kurtosis
#attach(XUnona)
```

```
#png("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis/figures/outlier1.png
     ",width = 404, height=404)
#par(mfrow=c(1,1))
#boxplot(‘thallium, ng‘,main="thallium with eo, ng")
#dev.off()
#png("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis/figures/outlier2.png
     ",width = 404, height=404)
#par(mfrow=c(1,1))
#boxplot(‘thallium, ng‘[-520],main="thallium without eo, ng")
#dev.off()

#Description table for categorical var
for(m in var_cat){
  print(colnames(XUnona)[m])
  print(table(XUnona[,m]))
  print(table(XUnona[,m])/608)
}

# Reasons of hypertension
c_d=0
c_sbp=0
c_dbp=0
for(i in which(XUnona[,40]=="hta")){
  if(XUnona[i,5]=="Yes") c_d=c_d+1
  if(XUnona[i,38]>140) c_sbp=c_sbp+1
  if(XUnona[i,39]>90) c_dbp=c_dbp+1
}
list(nobs_drugs=c_d,nobs_sbp=c_sbp,nobs_dbp=c_dbp)

########################################################################
########################################################################
##################################################### CORRELATION MATRIX
colnames(XUnonal)<-prettynames
CM<-cor(XUnonal[,10:37])
pmatCM<-cor_pmat(XUnonal[,10:37])
ggcorrplot(CM,method=c("circle"),colors=c("purple","white","red"),
           p.mat=pmatCM,hc.order = T, type = "lo")
# width 1000 for output

########################################################################
########################################################################
#################################################################### PCA

#pca objects
pca_ind<-PCA(XUnonalstd,quali.sup = 29)
par(mfrow=c(1,1),cex=0.4)
pca_var<-PCA(XUnonalstd[,-29])
pca_var$eig

#Plot pca var
#png("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis/figures/pca_var.png
     ",width = 790,height = 790)
par(mfrow=c(1,1),cex=1.6)
plot(pca_var,col.var = c(rep("maroon",time=22),rep("orange",time=6)),choix = "var",cex=0.9)
#dev.off()

#Plot pca ind
#png("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis/figures/pca_ind.png
     ",width = 790,height = 790)
```

```
par(mfrow=c(1,1),cex=1.6)
plot(pca_ind,habillage = 29,choix = "ind",col.hab = c("red","black"),
     label = "none",cex=1.4)
#dev.off()

#Plot tox
esslab<-c(0,0,0,0,0,1,1,0,1,1,0,1,rep(0,time=16))
labtox<-rep("gray",time=28)
for(i in 1:28){
  if(esslab[i]==1) labtox[i]<-"black"
}
#png("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis/figures/pca_toxic.
    png",width = 790,height = 790)
#par(mfrow=c(1,1),cex=1.6)
plot(pca_var,col.var = labtox,choix = "var",cex=0.9)
#dev.off()

#comp
comp1<-pca_ind$ind$coord[,1]
comp2<-pca_ind$ind$coord[,2]

#boxplots comp 1 and 2
#png("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis/figures/PC1scores.
    png",width = 404,height = 404)
par(mfrow=c(1,1))
boxplot(comp1~XUnona[,40],main="PC␣1␣(global␣exposure)")
#dev.off()
#png("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis/figures/PC2scores.
    png",width = 404,height = 404)
par(mfrow=c(1,1))
boxplot(comp2~XUnona[,40],main="PC␣2␣(Na-Ca-Po4-urea-Mg)")
#dev.off()
t.test(comp1~XUnona[,40],var.equal=T)
t.test(comp2~XUnona[,40],var.equal=T)

######################################################################
######################################################################
############################################################ Clustering

#SPECIFIC DATA PREPARATION FOR CLUSTERING PROCESS

colnames(XUnonal) # current data

boxplot(log(XUnonal[,9])) # insuline needs to be log transformed
boxplot(XUnonal[,6])
boxplot(log(XUnonal[,6])) # tri needs to be log transformed
boxplot(XUnonal[,8])
boxplot(log(XUnonal[,8])) # glu needs to be log transformed

#We log-transform graphically positive skewed distr
colnames(XUnona) #6,8,9,10:31
XUnonal<-XUnona
for(i in c(6,8,9,10:31)){
  XUnonal[,i]<-log(XUnona[,i])
}

#Preparing XUnonal for daisy function
XUnonalG<-XUnonal
for(i in var_cont){
  XUnonalG[,i]<-as.numeric(XUnonal[,i])
}
for(i in var_cat){
  XUnonalG[,i]<-as.factor(XUnonal[,i])
}
colnames(XUnonalG)
var_cont
for(i in var_cont){
  XUnonalG[,i]<-(XUnonalG[,i]-mean(XUnonalG[,i]))/sd(XUnonalG[,i])
}

# CLUSTERING PROCESS

# I) Distance matrix
mydaisy<-daisy(XUnonalG[,10:37],metric = "manhattan")
range(as.matrix(mydaisy))
# mydaisytoplot needs to be between 0 and 1 (0: black, 1: white)
mydaisytoplot<-as.matrix(mydaisy)*0.999999/99.05336
range(mydaisytoplot)
#png("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis2/figures/
    distancematrix.png")
image(1:608,1:608,mydaisytoplot,
      xlab = "␣",ylab = "␣",main="Distance␣matrix")
#dev.off()

# II and III) number of clusters
# Calculate silhouette width for many k using PAM
sil_width <- c(NA)
for(i in 2:10){
  pam_fit <- pam(mydaisy,
                 diss = T,
                 k = i)
  sil_width[i] <- pam_fit$silinfo$avg.width
}

#Plot of sihouette width
#png("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis2/figures/silhouette.
    png")
par(mfrow=c(1,1))
plot(1:10, sil_width,
     xlab = "Number␣of␣clusters",
     ylab = "Silhouette␣Width",xlim=c(1,10),xaxt='n')
axis(side = 1, at=2:10)
lines(1:10, sil_width)
abline(v=2,lty=2,col="red")
abline(h=sil_width[2],lty=2,col="red")
#dev.off()
#k=3 shows tho greater silhouette width

#Distance matrix befor and after clustering
mycluster<-pam(mydaisy,diss = TRUE,k = 2)
cluster<-as.factor(mycluster$clustering)
c1<-which(cluster==1)
c2<-which(cluster==2)
#png("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis2/figures/dma.png",
    width = 404,height = 404)
par(mfrow=c(1,1))
image(1:608,1:608,mydaisytoplot,
      xlab = "␣",ylab = "␣",main="a")
```

```
#dev.off()
#png("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis2/figures/dmb.png",
    width = 404,height = 404)
par(mfrow=c(1,1))
image(1:608,1:608,mydaisytoplot[c(c1,c2),c(c1,c2)],
      xlab = "␣",ylab = "␣",main="b")
#dev.off()

#PCs by cluster ID
#png("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis2/figures/ttestc1.png
    ",width = 404,height = 404)
#par(mfrow=c(1,1))
boxplot(comp1~cluster,main="First␣PC␣by␣cluster␣(p<0.05)")
#dev.off()
t.test(comp1~cluster,var.equal=T)
#png("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis2/figures/ttestc2.png
    ",width = 404,height = 404)
#par(mfrow=c(1,1))
boxplot(comp2~cluster,main="Second␣PC␣by␣cluster␣(p<0.05)")
#dev.off()
t.test(comp2~cluster,var.equal=T)

#Cluster stability (functions from resampling_methods.R)
Xcat<-XUnonal[,var_cat]
Xcont<-XUnonal[,var_cont]
Xcat[,1]<-as.numeric(Xcat[,1])-1 # 1:M
Xcat[,2]<-as.numeric(Xcat[,2])   # 1:diabetes
Xcat[,3]<-as.numeric(Xcat[,3])-1 # 1:anti-hta drug
Xcat[,4]<-2-as.numeric(Xcat[,4]) # 1:hta
XUnonal2<-XUnonal
XUnonal2[,var_cat]<-Xcat
XUnonal2[,var_cont]<-Xcont
XUnonal2std<-XUnonal2
for(i in 1:ncol(XUnonal2std)){
  XUnonal2std[,i]<-(XUnonal2[,i]-mean(XUnonal2[,i]))/sd(XUnonal2[,i])
}
dim(XUnonal2std)

ri_pam_bo<-function(data,k,metric="gower"){
  n<-nrow(data)
  # in the bootstrap version S,D and E have size n
  indS<-sample(x=c(1:n),size = n,replace = T)
  indD<-sample(x=c(1:n),size = n,replace = T)
  S<-data[indS,]          # training set S
  D<-data[indD,]          # training set D
  E<-data                 # evaluation set E
  ne<-nrow(E)
  dissS<-daisy(S,metric =metric,warnType = F)
  dissD<-daisy(D,metric =metric,warnType = F)
  Cs<-pam(dissS,diss=T,k=k)
  Cd<-pam(dissD,diss=T,k=k)
  distmedCs<-as.matrix(daisy(rbind(E,data[Cs$id.med,]),
                          metric = metric,warnType = F))[c(1:ne),-c(1:ne)]
  distmedCd<-as.matrix(daisy(rbind(E,data[Cd$id.med,]),
                          metric = metric,warnType = F))[c(1:ne),-c(1:ne)]
  # assign each point of E to the closest meloid of Cs and Cd
  indCs<-rep(NA,time=ne)
  indCd<-rep(NA,time=ne)
  for(j in 1:ne){
    indCs[j]<-which.min(distmedCs[j,])
    indCd[j]<-which.min(distmedCd[j,])
  }
  # compare output and measure stability with rand index
  # corrected for agreement by chance
  s<-randIndex(table(indCs,indCd),correct = T)
  return(s)
}

k=2
B=10 #put B=1000
M<-matrix(0,B,ncol=2)
for(i in 1:B){
  M[i,1]<-ri_pam_bo(XUnonal2std[,10:37],k,"euclidean")
  M[i,2]<-ri_pam_bo(XUnonal2std[,10:37],k,"manhattan")
}

#png("C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis2/figures/stabmetric.
    png",width = 404,height = 404)
par(mfrow=c(1,1))
boxplot(M,main="␣",names=c("Euclidean","Manhattan"))
#dev.off()

#bootstrapped replicates
round(colMeans(M),3)
round(c(sd(M[,1]),sd(M[,2])),3)
wilcox.test(x=M[,1],y=M[,2],exact=T,alternative = "two.sided")

######################################################################
######################################################################
############################################################ REGRESSIONS

# REGRESSIONS FROM PC SCORES

colnames(XUnonal)<-colnames(XU)
m1<-lm(sbp_mean~comp1+comp2+age+sex01+diabetes+tg+cho+glu+ins2s+d_hta,data=XUnonal)
m2<-lm(dbp_mean~comp1+comp2+age+sex01+diabetes+tg+cho+glu+ins2s+d_hta,data=XUnonal)
m2r<-lm(dbp_mean~comp1+comp2+age+sex01+cho+glu*diabetes,data=XUnonal)
summary(m1)
summary(m2)
summary(m2r)

# logistic
colnames(XUnonal)<-colnames(XU)
m3<-glm(hta~comp1+comp2+age+sex01+diabetes+tg+cho+glu+ins2s,data=XUnonal,family="binomial")
summary(m3) # --> but predicting "no hta" !
levels(XUnonal$hta)
hta01<-factor(XUnonal$hta,levels = c("No␣hta","hta"))
m3<-glm(hta01~comp1+comp2+age+sex01+diabetes+tg+cho+glu+ins2s,data=XUnonal,family="binomial")
summary(m3) #--> predicting "hta"

# likelihood ratio test for the logistic model
mnull<-glm(hta01~1,family = "binomial",data=XUnonal)
anova(mnull,m3,test="Chisq")

# REGRESSIONS FROM CLUSTER ID

#pred with reg
m4<-lm(sbp_mean~cluster+age+sex01+diabetes+tg+cho+glu+ins2s+d_hta,data = XUnonal)
```

```
m5<-lm(dbp_mean~cluster+age+sex01+diabetes+tg+cho+glu+ins2s+d_hta,data = XUnonal)
m6<-glm(hta01~cluster+age+sex01+diabetes+tg+cho+glu+ins2s,family = "binomial",data = XUnonal)
summary(m4) #no pred
summary(m5) #no pred
par(mfrow=c(2,3),pty="s")
plot(m4,which = 1:6)
summary(m6)

# likelihood ratio test for the logistic model
mnull<-glm(hta01~1,family = "binomial",data=XUnonal)
anova(mnull,m6,test="Chisq")

#odd ratio c3
coef(m6)
or<-exp(coef(m6)[3])
se<-sqrt(vcov(m6)[3,3])
#95 CI
or
c(exp(log(or)-1.96*se),exp(log(or)+1.96*se))

table(hta01[which(cluster==3)])
111/114

# OTHER REGRESSIONS

# MA
MA<-step(lm(sbp_mean~.,data=XUnonal[,-c(1,39,40)]),direction = "backward")
MB<-step(lm(dbp_mean~.,data=XUnonal[,-c(1,38,40)]),direction = "backward")
MC<-step(glm(hta01~.,data = XUnonal[,-c(1,38,39,5,40)],family = "binomial"),direction = "
    backward")
MAfull<-lm(sbp_mean~.,data=XUnonal[,-c(1,39,40)])
MBfull<-lm(dbp_mean~.,data = XUnonal[,-c(1,38,40)])
MCfull<-glm(hta01~.,data=XUnonal[,-c(1,5,38,39,40)],family="binomial")

AIC(MA,MAfull)
AIC(MB,MBfull)
AIC(MC,MCfull)

# likelihood ratio test for the logistic model
mnull<-glm(hta01~1,family = "binomial")
anova(mnull,MC,test="Chisq")

########################################################################
########################################################################
######################################################## Abbrev table

abbrev<-cbind(colnames(XU),prettynames)
colnames(abbrev)<-c("abbreviation","description")

#write.table(abbrev,"C:/Users/erabl/Documents/01_Statistiques_UNINE/Master_thesis/thesis/des_
    abbrev.txt",eol="\n",quote=F,sep=": ",
#           row.names = F,col.names = F)
```