A statistical analysis on factors that contributed/slowed down the spread of COVID-19

Project for the exam: Machine learning, statistical learning, deep learning and artificial intelligence

Marzio De Corato

September 3, 2020

"Text"

CONTENTS

LIST OF FIGURES

ABSTRACT

1 INTRODUCTION

The recent pandemic diffusion of the virus *SARS-CoV-2* connected to the *COronaVIrus Disease 19* raised up a scientific issue about the different spread ratio among the different population clusters such as cities,regions and states. This behaviour can be ascribed to the fact that the features of these clusters, such as the densities, the public transports as well as the individual mean incomes are very heterogeneous [3, 4]. Furthermore the individual policies undertaken by the decision maker of these clusters have a large effect on the spread of the virus [5]. A good tool for an analysis about the correlations of the spread of the virus with the cluster feature is the Principal Component Analysis (PCA): this approach allows to visualize in to a single 2D plot the main statistical correlation of a data set. It is worth nothing that this type of analysis represent only the first step for the identification of causal correlation between the features: the findings obtained via PCA should be validated with a model; however this second step is out from the aim of this work. In this work three different type of cluster were considered: the Italian provinces (*province*), the Italian regions (*regioni*), and 150 countries. The choice of the of Italian provinces was motivated by the fact that this the smallest cluster for which the daily cases of COVID19 are easily available, on the other hand the regions provide aggregate data (in particular about the sanitary system) that are not easily retrievable for the provinces. Finally the inclusion of the states is aimed to see if the statistical correlation obtained for the smaller cluster were still valid in a macroscopic scale. For this last cluster the lock-down policy and its effect was not considered, because the data collection of the policies for each 150 countries was too time consuming, however this analysis may be an outlook for a future update of this work.

2 DATA DESCRIPTION

In this section a brief overview about the data analysed is provided: in particular, for each cluster,a brief description is provided about the data gathering and assembling.

2.1 Provinces

For this cluster the cumulative number of cases of COVID19 up to 24/08/2020 was obtained from the github website of *Protezione Civile* [6]. The overall number of total cases was than divided (normalized) by the population number of its province. The population number

(2019) was retrieved from Istituto Nazionale di Statistica (ISTAT) [7]. The size of the province was also considered in order to evaluate the density. Since the diffusion of a virus is causally correlated to connectivity and public mobility services [9, 8] we considered also a set of economic indices in order to evaluate the industrialization and the wealth of each province: these are the mean income for person (2019), the public transport (2012 measured as demand for resident), the private transport (2012 measured as cars for resident), the pollution and the unemployment (2019). The first one was taken from *Ministero dell'Economia e delle Finanze* as reported in the following website [10] while the other ones were obtained from the ISTAT databases [7].

2.2 Regions

As for the provinces also for this cluster the cumulative number of cases of COVID19 was taken from from [6] up to 24/08/2020. In this case the same source provides the overall number of tests. The unemployment rate as well the mean income were also available from ISTAT [7] (2019). Since the Italian constitution delegates partially the management of health services to the regions, new features connected to the latter are available. Here we considered the following ones: the number of resident citizens for general practitioner (*medico di base*, the normalized number of structures for hospitalization, the normalized number of medical guards (*guardie mediche*) multiplied by 10^5 and their normalized number of visit performed multiplied by 10^5. These data were retrieved from the statistical yearbook of national health service 2017 (*Annuario Statistico del sistema sanitario nazionale 2017* published on the Italian minister of health [11].Finally we also included the marks for the essential assistance levels *livelli essenziali di assistenza* as provided by the Italian minister of health for 2017 [13] . These represent an evaluation of the health services for each region according to the Italian Government. A full description of the feature analysed can be found here [12]

2.3 Countries

For this type of cluster the overall number of cases and deaths was taken from from the World Health Organization [14] up to 27/08/2020. These number were normalized with the data about the total population for state provided by World Bank [15] up to 2019. The GDP (up to 2019) and Universal Health Coverage Index [16] (up to 2017) for each country was also taken from the latter source [15] and than normalized. Finally the National Health Expenditure (NHA) (normalized) up to 2017, the Traffic Morality (normalized and scaled with a factor 10^5) up to 2016, the Pollution Mortality (normalized), the

PM 2.5 concentration ($\mu g/m^3$) up to 2016 were taken from the WHO databases [17]. It is worth nothing that selection for the counties considered in this work were the ones for which all these indicators were available.

3 THEORETICAL BACKGROUND

As pointed out by Mackay [1] the basic idea of unsupervised learning is to mime human behaviour is to find regularities in data and group them. Among the different techniques of unsupervised learning [2] here we considered the PCA: in this section the basic ideas of this tool will be focused following the approach proposed by James et al. [2]. The aim of PCA is to plot n observations with p features in only one 2D plot,with the least possible loss of information, instead of $\binom{p}{2}$ 2D plots. To do so the PCA chooses, among the possible axis formed by the linear normalized combination of the features, the two associated with the largest variance, given that these two axis are orthogonal. Thus, supposing that the means of the features are null, the problem is to find the coefficients (usually called loadings) ϕ_{ij} that solves the following optimization problem:

$$\max \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \tag{1}$$

with the constraint

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1 \tag{2}$$

The loadings define the principal components (that have to be orthogonal) :

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + ... + \phi_{p1} x_{ip} \tag{3}$$

If the all the principal components are taken in to account no information is lost, however the complexity of it due to high dimensionality is the same of the starting one; however if we sacrifice the p-2 component, that have a lower variance with respect to the first two, the complexity is dramatically reduced to a 2D plot. In this case the toll paid is the lost of the information contained in the $p-2$ component. From a computational point of view this task can be performed with the standard techniques used for solving a eigenvalue problem. Once the two principal component are found, the data can plotted into the new coordinate set; the advantage of this new representation

lies on the fact that if the first two principal component loading vectors are plotted a bird's-eye view of the statistical correlation between the features is obtained: basically the cosine of the angle between the loadings approximates the statistical correlation, while the position of the data with respect to the loadings will plot its feature. In this way an intuitive representation of the data, their features and the correlation among them is provided. Finally this can be accompanied by a plot of the correlation matrix as it is done in this work.

4 RESULTS AND DISCUSSION

For each cluster the following procedure was followed: first the correlation matrix was calculated together with its heath-map, then the PCA was performed on the scaled data. For this procedure an histogram about the importance of the components, a 2D plot of the loading vector normalized on unitary circle, and the full 2D PCA plot with the data were reported.

4.1 Provinces

From the inspection of the plots 1,3 and 4 it is possible to point out that the public transport, the density, the normalized cumulative cases and the mean income are positively correlated. This statistical correlations are causally validated by different previous publications [22, 21, 20, 19]. The argument that explains this behaviour is that a higher density, a higher public transport demand and a higher income increase the rate of contact between the individuals: the first one since people are closer and so a contact between has a higher probability with respect to a low density areas (such as a rural context), the second one since when individuals use the public transport are very close each other (indeed there is no correlation between the cases and private transport), the third one because rich people can spend more money for social events or perhaps in to travels. This fact is corroborated by the negative statistical correlation between cases and unemployment (in this case also the fact that a worker has a higher mobility due to the fact that he/she has to reach the place of work must be considered). One can be at first confused by the fact that the air quality have a strong correlation with the private transport and the public transport has a different direction: this can ascribed to the fact that in rural provinces, were the air quality is higher and the public transport services are reduced, people are forced to own a private vehicle, while in urban provinces they can consider to use only the public transport. Finally note that looking to the plot 2 the

PCA analysis reported in plots 3 and 4 captures almost the 40 % of variance.

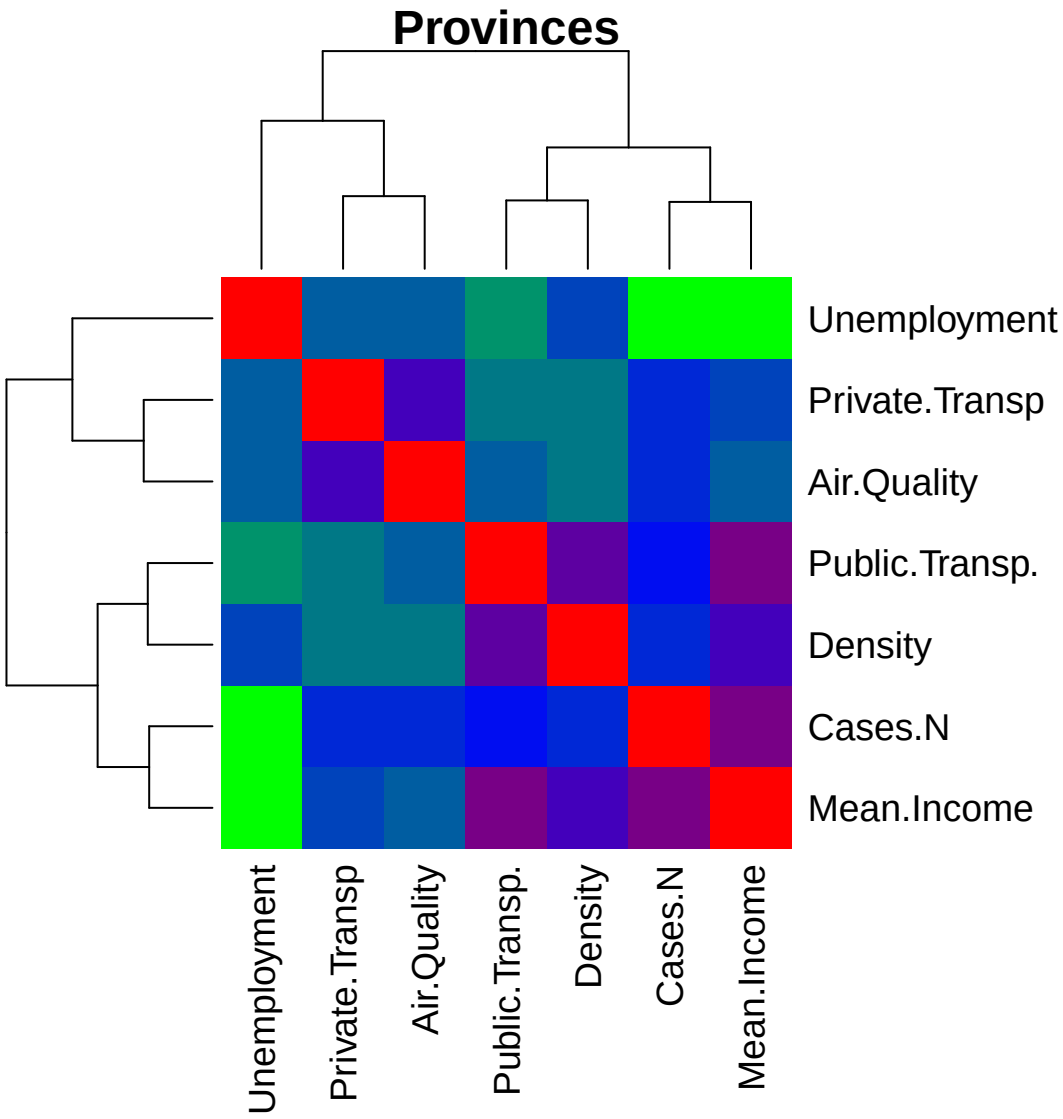

Figure 1: The Correlation matrix for the Italian provinces accompanied by its heat-map as generated by the following dataset: local unemployment (2019), local private transport (2012 number of cars for 1000 residents), the air quality (2012) ,the local public transport (2012 measured as the demand for resident), the density (2019 measured as resident for Km^2, the cumulative cases up to 26/08/2020 and the mean income (2019 measured in kEUR). Note that the public transport, the denisty the cumulative cases and the mean income are highly correlated. On the other hand the unemployment is negatively correlated to this first cluster

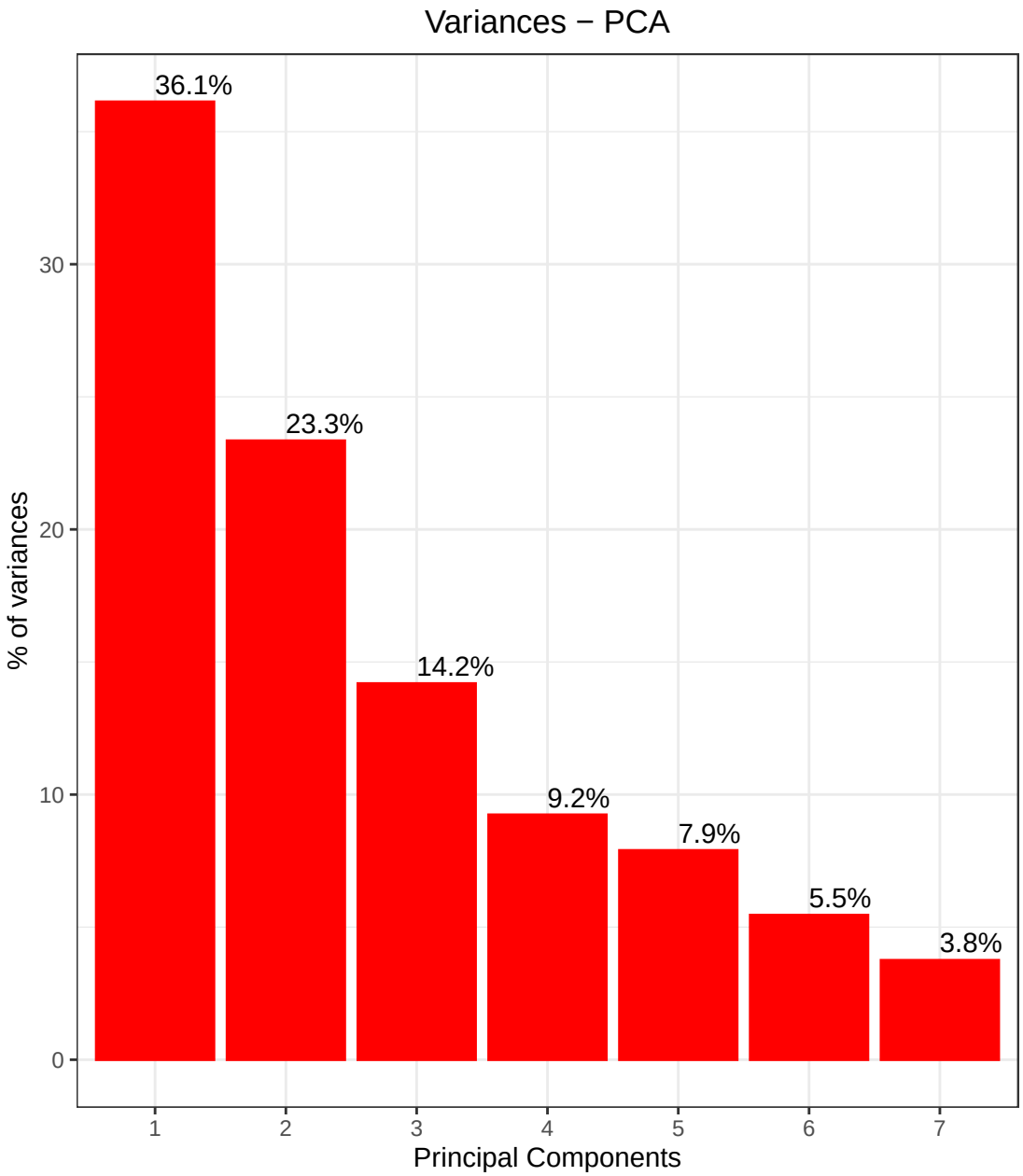

Figure 2: Percentage of variance for each component of the dataset concerning the provinces. Note that using the first two, almost the 60 % of variance is captured, as consequence the information lost is almost the 40 %

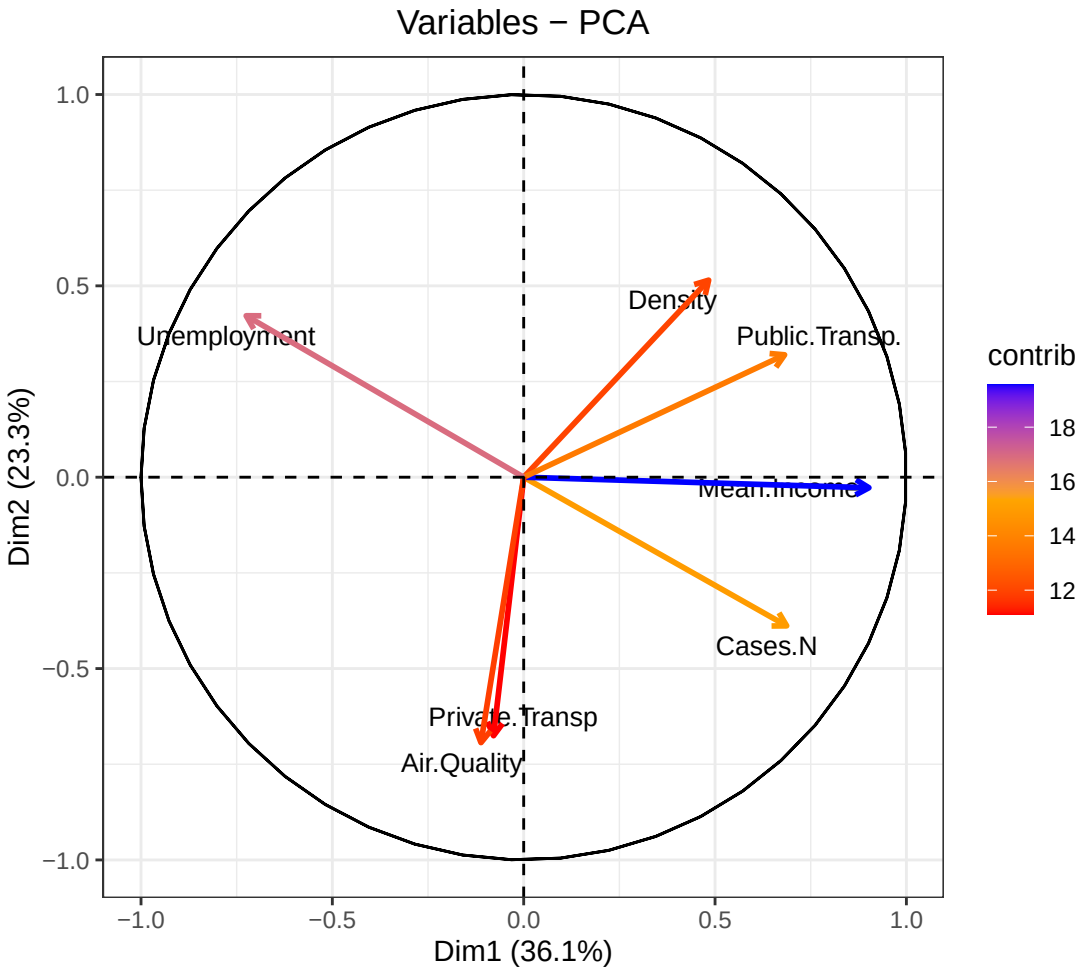

Figure 3: 2D plot of the loading vector for the dataset of provinces normalized on the unitary circle. The features, the date of their collection and unit of measure are the same of Fig. 3

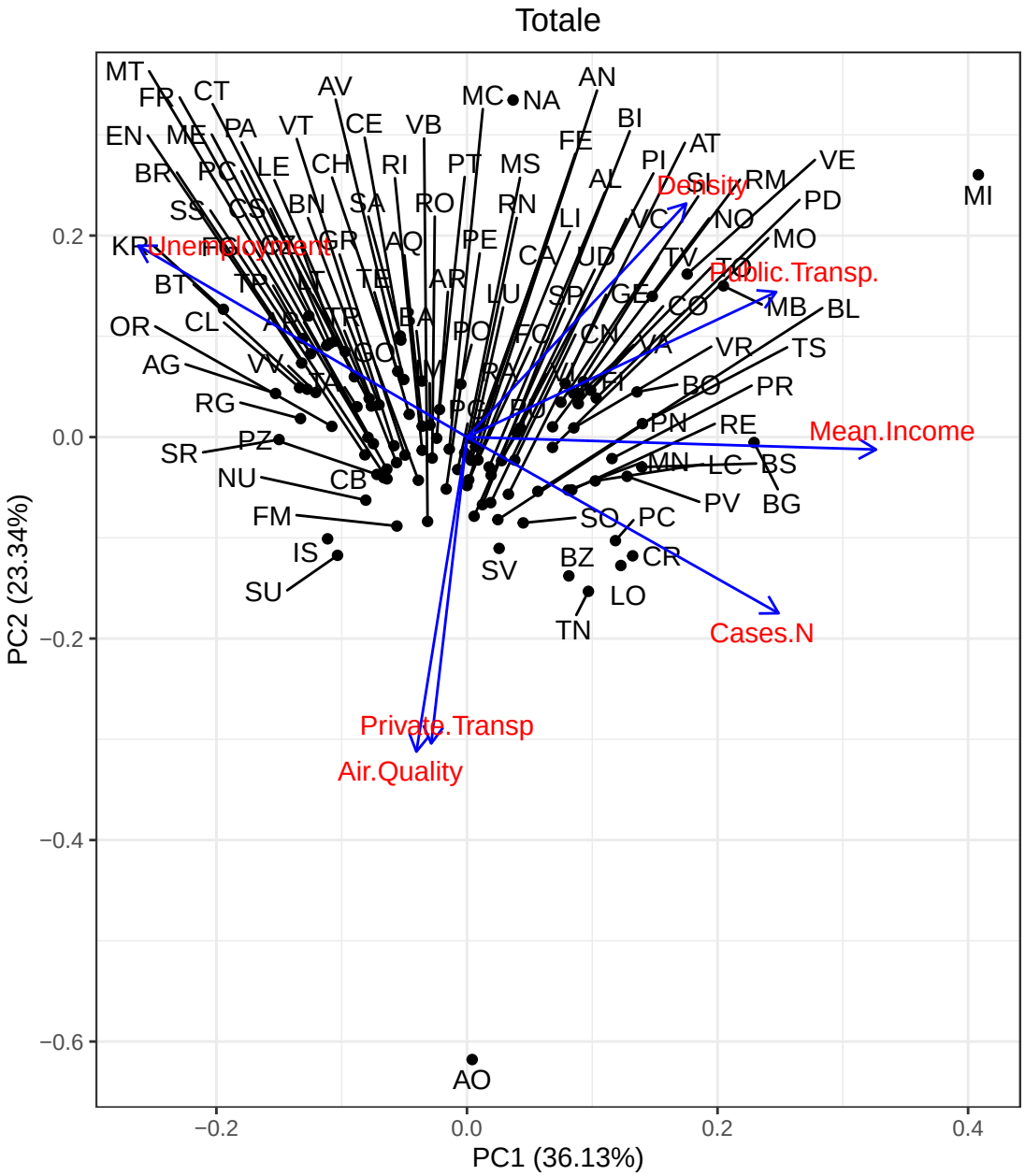

Figure 4: The dataset of provinces plotted on the two principal components that capture the 60 % of variance. The provinces are labelled with their vehicle registration abbreviation. The features, the date of their collection and unit of measure are the same of Fig. 1

4.2 Regions

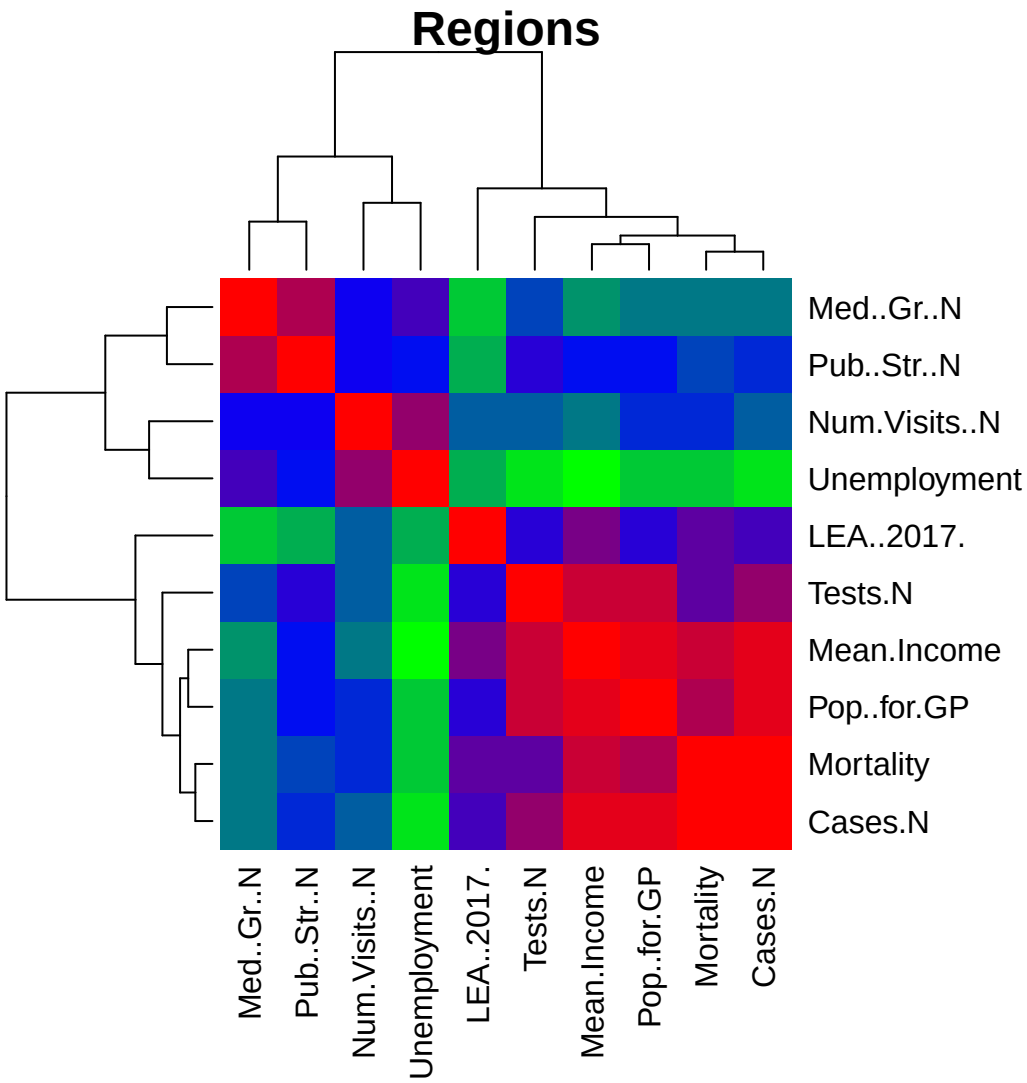

Figure 5: The Correlation matrix for the Italian regions accompanied by its heat-map as generated by the following dataset: the normalized number of medical guards (2017), the normalized number of structures for hospitalization (2017), the normalized number of visit for medical guards (2017), the unemployment ratio (2019), the essential assistance levels (LEA - 2017, the number of normalized tests (24/08/2020), the mean income (2019), the population for general practitioner (2017), the normalized deaths for COVID19 (24/08/2020), and the normalized number of cases (24/08/2020)

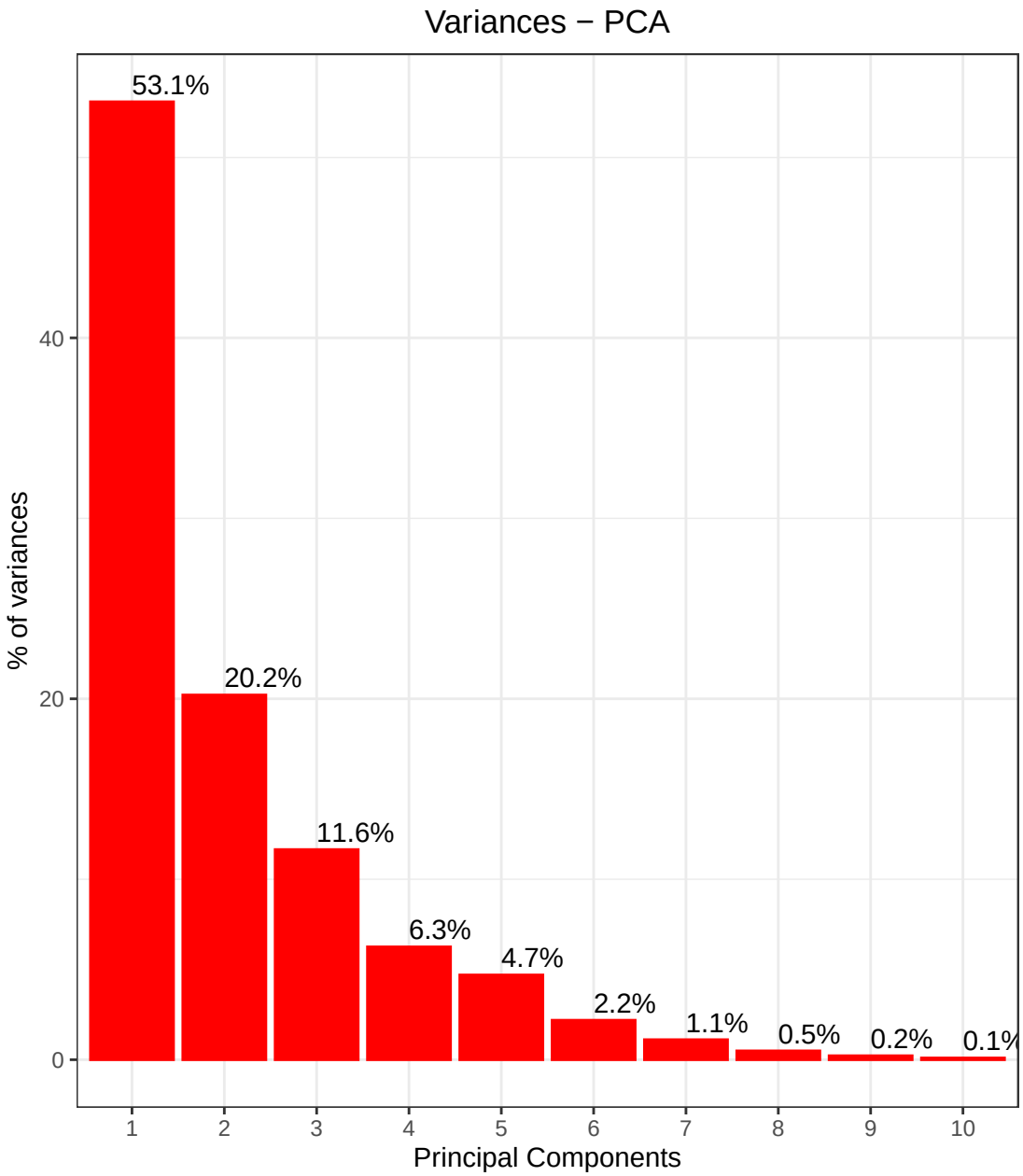

Figure 6: Percentage of variance for each component of the dataset concerning the provinces. Note that using the first two, the 73.3 % of variance is captured, as consequence the information lost is almost the 26.7 %

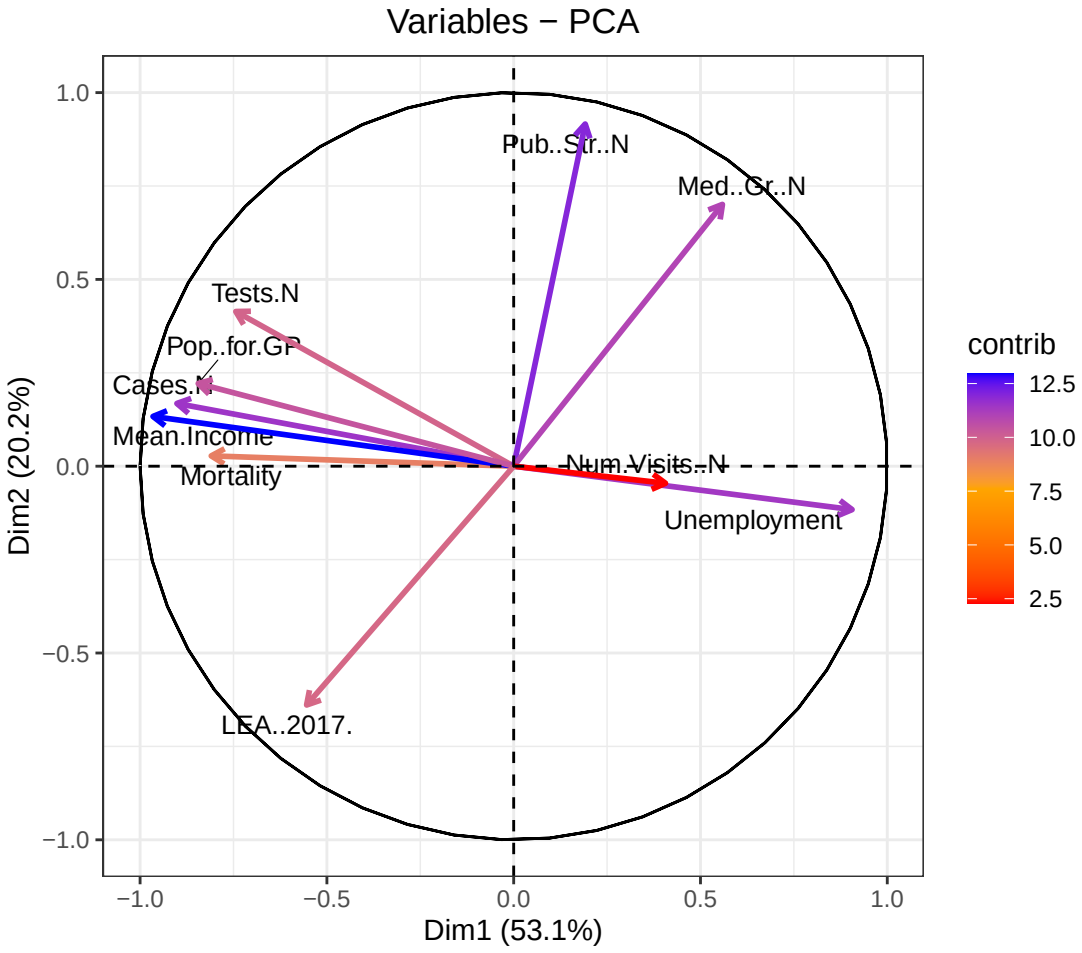

Figure 7: 2D plot of the loading vector for the dataset of regions normalized on the unitary circle. The features, the date of their collection and unit of measure are the same of Fig. 5

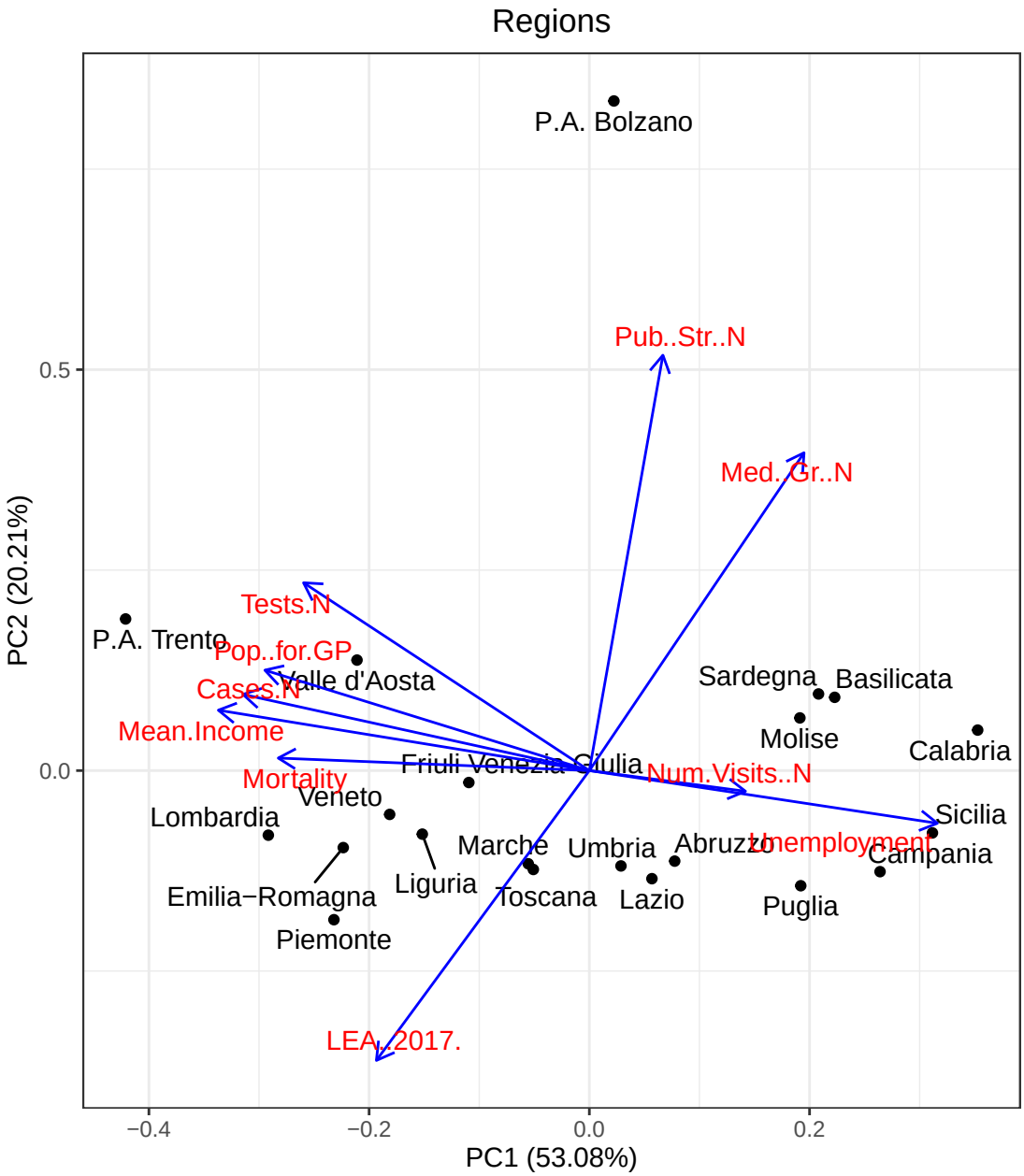

Figure 8: The dataset of regions plotted on the two principal components that capture the 73.3 % of variance. The features, the date of their collection and unit of measure are the same of Fig. 1

REFERENCES

[1] David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[2] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[3] Abiel Sebhatu, Karl Wennberg, Stefan Arora-Jonsson, and Staffan I Lindberg. Explaining the homogeneous diffusion of covid-19 nonpharmaceutical interventions across heterogeneous countries. *Proceedings of the National Academy of Sciences*, 2020.

[4] Piotr Skórka, Beata Grzywacz, Dawid Moroń, and Magdalena Lenda. The macroecology of the covid-19 pandemic in the anthropocene. *PloS one*, 15(7):e0236856, 2020.

[5] Per Block, Marion Hoffman, Isabel J Raabe, Jennifer Beam Dowd, Charles Rahal, Ridhi Kashyap, and Melinda C Mills. Social network-based distancing strategies to flatten the covid-19 curve in a post-lockdown world. *Nature Human Behaviour*, pages 1–9, 2020.

[6] Protezione Civile. https://github.com/pcm-dpc/COVID-19.

[7] Istituto nazionale di Statistica (ISTAT). https://www.istat.it/it/dati-analisi-e-prodotti/banche-dati/statbase.

[8] Moritz UG Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Louis Du Plessis, Nuno R Faria, Ruoran Li, William P Hanage, John S. Brownstein, Maylis Layan, Alessandro Vespignani, Huaiyu Tian, Christopher Dye, Oliver G. Pybus, and Samuel V. Scarpino. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497, 2020.

[9] Alun L Lloyd and Robert M May. How viruses spread among computers and people. *Science*, 292(5520):1316–1317, 2001.

[10] Ministero dell'Economia e delle Finanze. https://www.idealista.it/news/finanza/economia/2019/04/02/130615-la-mappa-del-reddito-pro-capite-nelle-province-italiane, 2019.

[11] Ministero della salute. http://www.salute.gov.it/portale/documentazione/p6_2_2_1.jsp?lingua=italiano&id=2879, 2017.

[12] Ministero della salute. http://www.salute.gov.it/portale/lea/dettaglioContenutiLea.jsp?lingua=italiano&id=1300&area=Lea&menu=leaEssn.

[13] Ministero della salute. http://www.salute.gov.it/portale/lea/dettaglioContenutiLea.jsp?lingua=italiano&id=4747&area=Lea&menu=monitoraggioLea.

[14] World Health Organization. https://covid19.who.int/table.

[15] World Bank. https://data.worldbank.org/indicator.

[16] World Helth Organization. https://www.who.int/data/gho/indicator-metadata-registry/imr-details/4834.

[17] World Helth Organization. https://www.who.int/data/gho.

[18] Noah C Peeri, Nistha Shrestha, Md Siddikur Rahman, Rafdzah Zaki, Zhengqi Tan, Saana Bibi, Mahdi Baghbanzadeh, Nasrin Aghamohammadi, Wenyi Zhang, and Ubydul Haque. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? *International Journal of Epidemiology*, 49(3):717–726, 02 2020.

[19] Simone Weyers, Nico Dragano, Susanne Möbus, Eva-Maria Beck, Andreas Stang, Stephan Möhlenkamp, Karl Heinz Jöckel, Raimund Erbel, and Johannes Siegrist. Low socio-economic position is associated with poor social networks and social support: results from the heinz nixdorf recall study. *International Journal for Equity in Health*, 7(1):13, 2008.

[20] Sebastiano Gangemi, Lucia Billeci, and Alessandro Tonacci. Rich at risk: socio-economic drivers of covid-19 pandemic spread. *Clinical and Molecular Allergy*, 18(1):1–3, 2020.

[21] World Health Organization. Centre for Health Development and World Health Organization. *Hidden cities: unmasking and overcoming health inequities in urban settings*. World Health Organization, 2010.

[22] Carl-Johan Neiderud. How urbanization affects the epidemiology of emerging infectious diseases. *Infection ecology & epidemiology*, 5(1):27060, 2015.