Is there evidence of a North-South “sustainability gap”?

A preliminary analysis based on statistical learning techniques

June 13, 2022

Carlo Cignarella, Margherita Lanini, Marco Zeppi

Abstract

The aim of this report is trying to detect emerging patterns across the Italian provinces in achieving the Sustainable Development Goals, as defined by the United Nations. To achieve this, we apply some supervised and unsupervised statistical learning techniques to the ISTAT dataset which accompanies 2021 annual report on Italian performance towards *Agenda 2030*. Our preliminary results suggest the presence of a “sustainability gap” between northern and southern areas.

Contents

1 Introduction

Every year since 2018, the Italian National Institute of Statistics (ISTAT) publishes an annual report on Italian performance in achieving Sustainable Development Goals (SDGs) (ISTAT, 2021). SDGs were defined by the United Nations (UN) in 2015 as key indicators of the plan *Transforming our world: the 2030 Agenda for Sustainable Development.* Within the attempts to go beyond GDP as measure of growth, development and prosperity, SDGs balance all the three dimensions of sustainable development: economic, social and environmental. SDGs aim to stimulate action over the years towards sustainability consistently with the perspective such that "you can't solve a problem if you can't monitor your progress in solving it". The present report makes a contribute to such efforts towards sustainability trying to detect emerging patterns at provincial level inside the Italian scenario. Indeed, its ultimate goal is to check the existence of a "sustainability gap" among different Italian territories, in particular between northern and southern areas. The reminder of this paper is organized as follows. In section 2 we provide a brief description of the data and an overview of the statistical learning tools (as described by James et al. (2021)) we applied to them. Section 3 contains our results. Section 4 reports a discussion and our final remarks. Lastly, the appendix A collects some additional material and our `R` code to perform data pre-processing.

2 Methods

2.1 Data description and selection

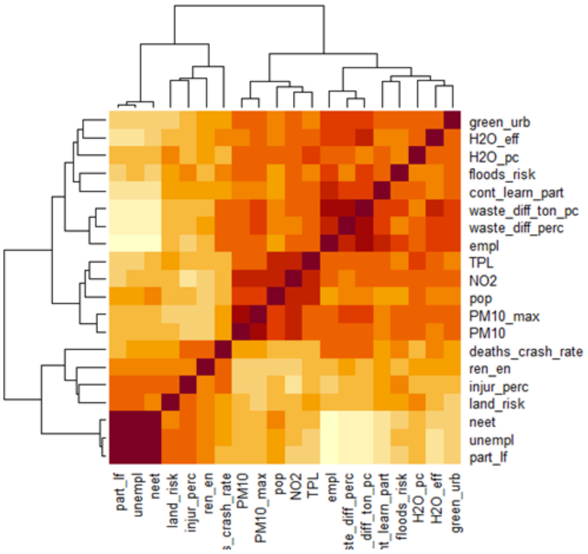
The ISTAT dataset on which the Italian annual report on SDGs is built is freely available on ISTAT web-page. It covers 2004-2020 period and reports Italian performance in achieving all 17 SDGs, the associated 169 targets and 232 indicators. We can see targets and indicators as more and more specific measures that we can aggregate to form a SDG which is inherently very general. We focus on the provincial level and therefore we select all available 20 indicators for all 107 Italian provinces.[1] Moreover, we choose 2015 as the reference year since it is the one with the lowest number of missing values. To deal with NAs we decide to substitute the missing values of a province with its average value over the period 2013-2017 or, if such option was not

[1] There are just 20 indicators for Italian provinces since different indicators are available for different levels of aggregation. Other possible observation units are the country as a whole, macro-regions, regions, main cities but also age or gender groups. We pick provinces because such aggregation level let us to obtain the higher number of observations as possible.

possible, with the average value of all other provinces of the same region in 2015.[2]

However, the available provincial 20 indicators capture very different phenomena given that they refer to all the three dimensions of sustainable development: economic, social and environmental. For this reason, it is crucial to transform such variables to be able to compare and select them in order to perform the following steps of our analysis. First of all, we transform the skewed variables by applying a logarithmic transformation. Then, we weight scale sensitive variables by provinces' population. Finally, we normalize all variables to avoid problems due to different units of measure. After having carried out the above-mentioned steps, we investigate the relationships between such variables by looking at their correlation and at the presence of possible clusters. Indeed, even if such indicators capture different phenomena, some of them are very similar. We perform this last step by running a hierarchical clustering with the complete method among all 20 transformed variables. Results are reported in figure 1.

Figure 1: Hierarchical clustering and correlation among all 20 provincial indicators

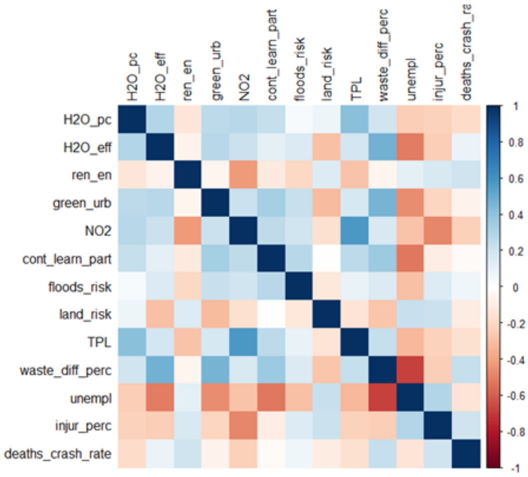


With these results in mind, we avoid to select highly correlated variables inside the same cluster. For instance, we select the unemployment rate in place of NEET, non-participation rate and employment rate. At the end, we remain with 13 indicators out of 20 which can be grouped according to sustainable development dimensions as economic, social and environmental indicators. All selected and not selected indicators, divided into three categories according to the dimension to which they refer, are reported in table 5 in the appendix A.1. To conclude this section, we can see from figure 2 that the variables we are left with do not seem to display strong

[2] As a last resort, for some special provinces such as *Aosta* which is the only one of its region, we substitute its missing values with the closest observed value within 2013-2017 period.

or problematic (from a intuitive point of view) correlations.

Figure 2: Correlation among all 13 obtained provincial indicators



2.2 Clustering

Before performing cluster analysis, we plot pairwise scatter-plot matrices with points colored according to the territorial partition of provinces (“North”, “Center”, “South”). In figure 11 we observe that points are not well separated, no clear pattern can therefore be identified *ex-ante*. We decide to perform two different clustering approaches, the agglomerative hierarchical clustering and the K-means. We first define an Euclidean distance between the points and compute cluster distances using four different linkage functions. The complete linkage seems to fit our data structure better than the others, due to its capacity to detect spherical clusters. Results of all different applications are reported in figure 12. We start from the Hierarchical approach investigating the optimal number of clusters. Among the possible indicators to assess the optimal number we look at the Total Within Sum of Squares (Elbow Method) and at the Silhouette. Secondly, we implement the k-means algorithm to compare results, we set random initialization and then refine the attempt setting as initialization points the centroids found with the Hierarchical approach.

2.3 PCA

We perform Principal Component Analysis to spot the indicators that explain the highest share of variability in our data. Observing the first two principal components we try to interpret results in order to analyze whether certain macro-categories of variables characterize the two components. To this end we focus on loading vectors, trying to observe which variables weight more in the two components. The PCA biplot and the final components matrix are useful representations for the analysis.

2.4 Classification and Cross-validation

We then shift to supervised techniques. Two different classification methods are applied: logistic regression and Linear Discriminant Analysis. The purpose of the former is to assess whether the variables included in the dataset can predict a province being excellent in waste sorting, since this is one of the most relevant features in explaining data variability, according to PCA. Instead, with LDA we try to predict the macro-area where the province lies ("North", "Center" or "South") using waste sorting performance and unemployment rate as predictors, as they are the two that stand out from the First Principal Component (see also section 3.2).

We first create a dummy for recycling "champions", assigning 1 if the value for the province is above the third quartile of the total distribution, 0 otherwise. The reference variable is always `waste_diff_perc`, the percentage of waste sorted over the total produced. Twenty-seven provinces are champions, eighty are not. We estimate the logit model, including in the predictors the five variables related to environmental standards or behaviors: `green_urb`, `TPL`, `ren_en`, `NO2`, `H2O_eff`.

As for the Linear Discriminant Analysis, the two predictors should be distributed normally according to the algorithm assumptions. We do not run any specific test, but we do plot the two distributions, which do not look pretty much gaussian (the mean of the scaled variables is not 0). Notwithstanding, we decide to proceed with the analysis and keep this detail in mind when evaluating the results.

Our strategy involves splitting the data among training set (65%) and test set (35%). Both algorithms are run twice: first, the coefficients or classifiers are estimated on training data, then the effective performance is tested on the unused records. The partition is done randomly by the software, but we set a seed in order to obtain reproducible results.

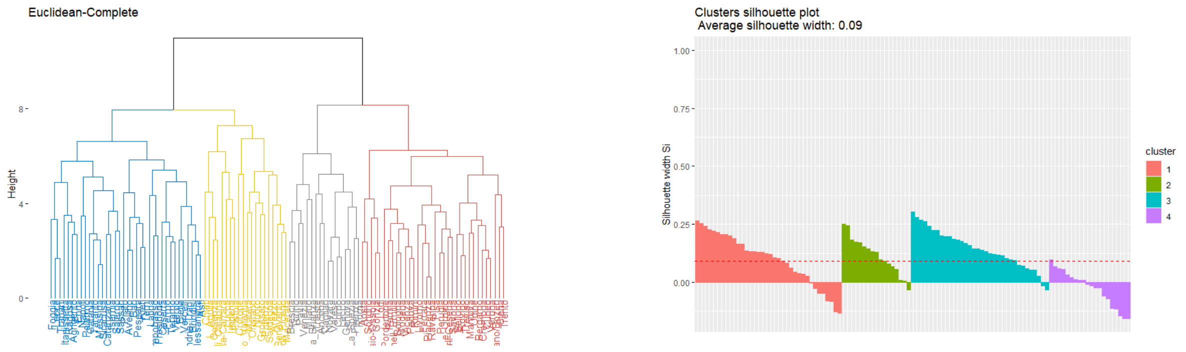
As the final step of our analysis, we try to validate the results of the supervised classification using the Leave-One-Out Cross-validation approach, which we consider the best strategy as we have 107 provinces and 13 variables.

3 Results

3.1 Clustering

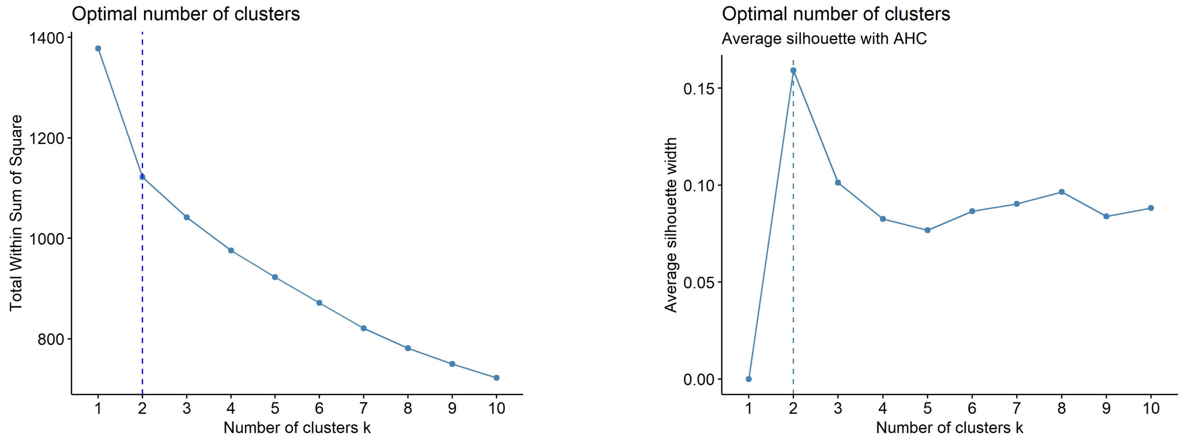
We report in figure 3 the principal results from our clustering analysis. As for the Hierachical clustering we set first 4 clusters, but the results in term of goodness suggest that the best solution may be to cut the dendogram at 2 clusters. In fact, the 4 clusters solution shows poor results in terms of average silhouette.

Figure 3: Hierarchical Clustering with 4 clusters



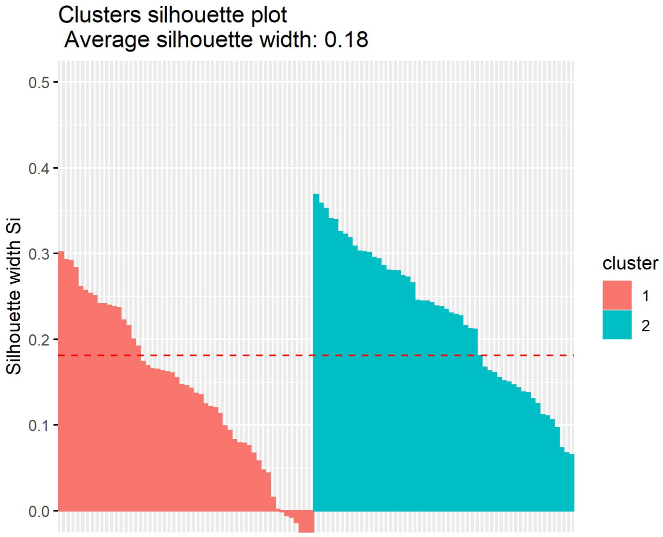
The optimality of the solution with two clusters is confirmed by the Elbow method and by the improvements in terms of average silhouette reported in figure 4.

Figure 4: Hierarchical Clustering measures of goodness



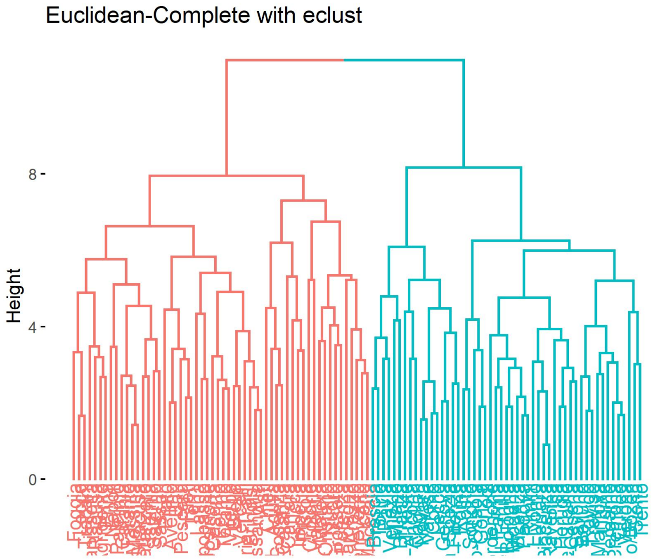
Coming to the K-means attempt we observe a result similar to the Hierarchical one. In particular the best result is obtained initializing the algorithm with the centroids found in the Hierarchilcal analysis, indeed we observe a non negligible improvement in average silhouette up to 0.18, as we observe in figure 5.

Figure 5: K-means silhouette with centroids initialization from the Hierarchical attempt



The final solution with two clusters presents an economically interpretable evidence as we can see in figure 6. Indeed, we can observe a clear divide between cities form the South of Italy and cities form the Center and the North. This evidence suggests that different trends SDGs performances depend on geographical location.

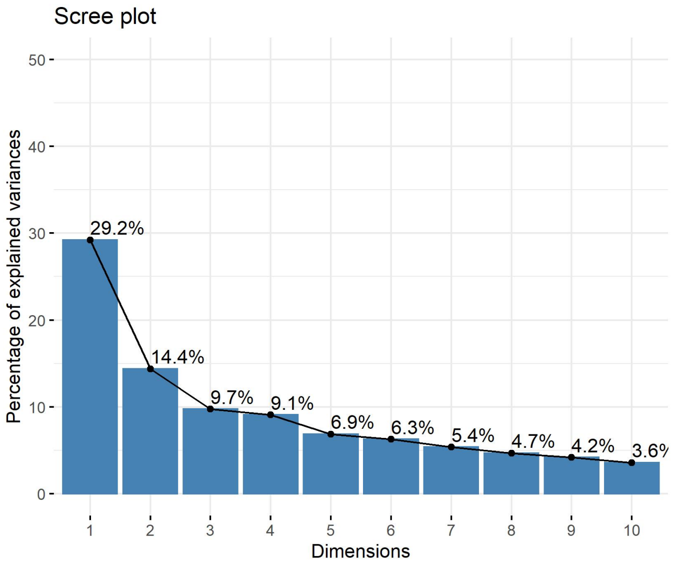
Figure 6: Hierarchical attempt with two clusters



3.2 PCA

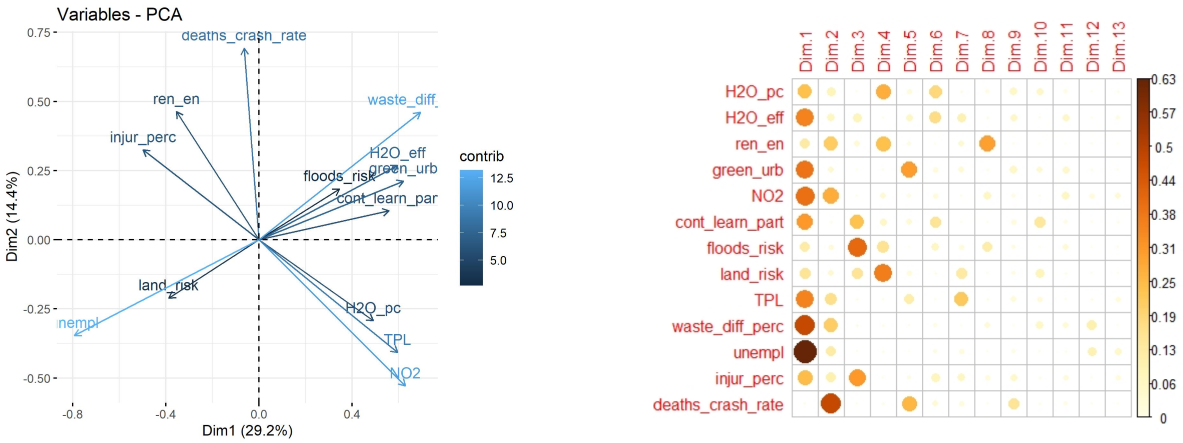
Looking at the screeplot in figure 7 we observe that the first two components account for around 45% of total data variability.

Figure 7: PCA - Screeplot



From figure 8 we can observe which are the most relevant variables. As regard the first component the variable that accounts for the majority of variability is unemployment. In addition we notice mostly environmental variables. This implies that the first component apart from unemployment is driven mostly by the environmental performance. Instead, in the second one the rate of deaths in car crashes dominates on the other variables.

Figure 8: Biplot and correlation matrix



Finally, it can be interesting to observe the biplot with the location of each province in figure

13. Indeed we notice that around unemployment are located mostly cities from the South, while cities from the Center and the North lie in proximity of environmental variables that indicate air pollution or efficiency in waste recycling.

3.3 Classification

The results of the LOGIT regression are summarised in Table 1. A couple of variables are highly significant and have a positive impact on the probability for a province to be a waste champion. The coefficients refer to the model run on train data. We then try to classify data in the test set. If p>0.5 according to the model,the predicted response is 1. Overall accuracy on test data is 0.757, which is a quite good result. Clearly, this is the value of accuracy for one of the possible test samples.

Before performing LDA, we look at the scaled distribution of the variable *waste_ diff_ perc*, one of the two predictors, and label each data point (province) according to the matching macro territorial area (Figure 9). One can notice than northern provinces (red points) perform better than the average, while provinces located in the South are concentrated at the bottom of the picture. There are fewer and more scattered observations for the center.

Figure 10 reports some bar plots. The x-axis shows the value of the line defined by the coefficient of linear classifier for LDA. Waste performance yields a much more well-separated distribution than unemployment. Interestingly, such classifier does not take into account the sign (or the direction) of the distribution: the center is grouped around the mean, as expected, but the South for instance occupy the positive semiaxis.

Table 2 yields a somehow puzzling output of the LDA on test data. The overall accuracy is high, above 75%, but the sensitivity for "center" is really low, meaning that provinces in the center are constantly misclassified. The confusion matrix suggests that 6 out of 7 central provinces are actually predicted to lie somewhere else. Eyeballing the partition plot (figure 14, see the additional material) it stands out, again, that classification of central provinces often fails.

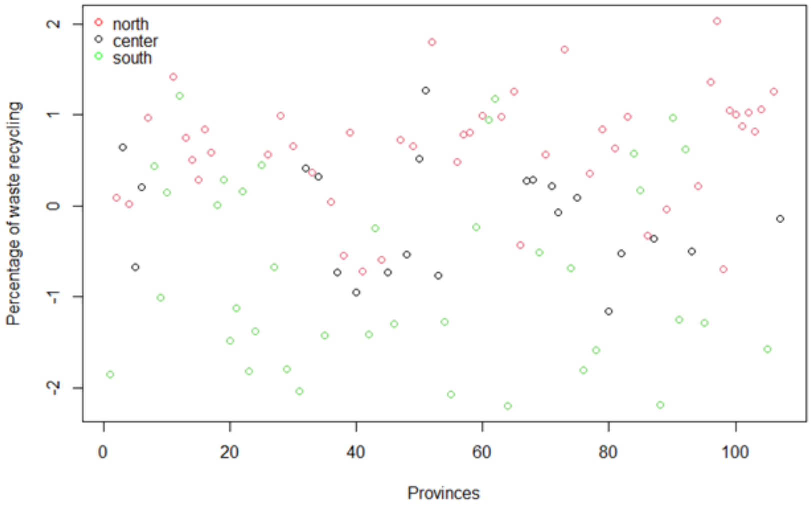
Cross-validation The Leave-One-Out Cross-Validation approach allows to assess the validity of the results obtained so far in our supervised analysis.

Table 3 reports a summary of the LOOCV algorithm run on the LOGIT model. Admittedly, this approach fails as the main measures suggest a very poor performance. RMSE and MAE are too high for an accurate classification, given that the response should be between 0 and 1. In

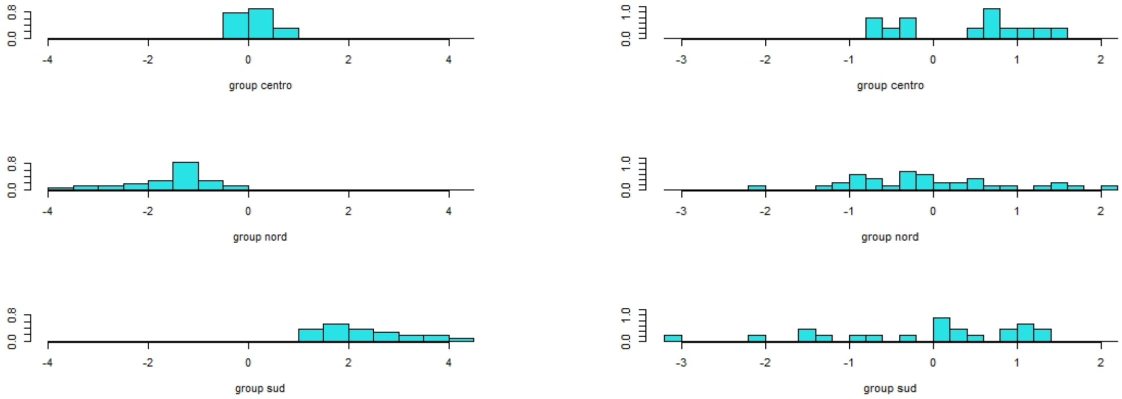
Table 1: LOGIT results

Estimates on training data	waste_champion
H2O_eff	0.802^{**}
	(0.385)
ren_en	-0.513
	(0.480)
green_urb	1.030^{***}
	(0.376)
NO2	-0.859^{*}
	(0.522)
TPL	0.224
	(0.406)
Constant	-1.421^{***}
	(0.391)
Observations	70
Log Likelihood	-30.325
Accuracy on test data	0.757
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 9: Distribution of waste_diff_perc by territorial area



Note: Data are scaled and centered around mean = 0. The x-axis reports provinces by their index from 1 to 107.



(a) waste_diff_perc

(b) unempl

Figure 10: LDA results on train data: bar plots

Table 2: LDA results on test data: confusion matrix and statistics

CONFUSION MATRIX			reference	
		Center	North	South
prediction	Center	1	1	1
	North	5	14	1
	South	1	0	13
STATISTICS BY CLASS		Center	North	South
	sensitivity	0.1429	0.9333	0.8667
	specificity	0.9333	0.7273	0.9545
OVERALL STATISTICS	accuracy	95% C.I.		
	0.7568	(0.588; 0.8823)		

contrast, the Rsquared is too low.

LOOCV results for LDA are instead summarised in Table 4. These are pretty satisfactory, as the overall accuracy improved, up to near 80%: the two predictors correctly classify the macro area of a province 8 times out of 10. This figure, however, must be taken with a pinch of salt, because, we stress in the previous section, the sensitivity is unbalanced across the three possible outcomes.

Table 3: LOOCV results: LOGIT

Model:	Generalized Linear Model (LOGIT)
samples	107
predictors	5
Root Mean Squared Error	0.4196
Rsquared	0.0828
Mean Absolute Error	0.3410

Table 4: LOOCV results: LDA

Model:	Linear Discriminant Analysis
samples	107
predictors	2
classes	North, Center, South
Accuracy	0.7850

4 Discussion and final remarks

The main evidence coming from the unsupervised analysis is the gap between North and South, not only economically but also in terms of environmental performance. This point reflects the asymmetric growth and development that has always characterized the two macro-regions over time: a "sustainability gap" seems to emerge. A few remarks about the supervised analysis can be mentioned as well. In principle, trying to predict the performance in waste sorting based on other environmental and and green behavior parameters is not a trivial approach. However, the analysis we propose failed, probably because the few measures available at a provincial level are not the most meaningful ones and show overall a low variability. The second strategy, instead, yields significant results: an environmental parameter and a socio-economic indicator, combined, can predict the macro area rather confidently. This suggests that a lot has still to be done to ensure a more equitable development and to ensure the progressive catching up of the territorial divide.

There are indeed some limitations to the analysis that we carried out. They mainly concern the fact that, from the original dataset containing hundreds of statistical measures, we were left with 13 only after selecting the provincial dimension and filtering. To increase the number of observations and thus to allow the application of more refined techniques, city-level data may be collected. To the best of our knowledge, such collection is still missing, apart from the 14 Metropolitan cities. Furthermore, we chose one year as a reference, because it was the one with the lowest non-observed data points. A comparison with data retrieved from more recent surveys may help to assess the validity of the conclusions presented here and to better define the extent of the phenomena. Finally, as the structure of the dataset should encourage to adopt a temporal approach, a possible line of research is to isolate a balanced panel and detect any patterns in the evolution of the SDGs over time.

A Appendix

A.1 Additional material

Table 5: Selected and not selected variables' names & codes

variables' names	variables' code	selected	type
Green areas over total areas	green_urb	yes	env
Drinking water distribution network efficiency	H2O_eff	yes	env
Provided water per capita	H2O_pc	yes	env
Exposed population to the risk of flooding	floods_risk	yes	env
Lifelong learning	cont_learn_part	yes	soc
Recycled waste in tons per capita	waste_diff_ton_pc	no	env
Percentage of recycled waste over total waste	waste_diff_perc	yes	env
Employment rate	empl	no	eco
Seats per km offered by local public transport	TPL	yes	soc
Nitrogen concentration	NO2	yes	env
Population	pop	no	soc
PM10 by day	PM10_max	no	env
PM10 by year	PM10	no	env
Traffic fatalities	deaths_crash_rate	yes	soc
Renewable energy	ren_en	yes	env
Injury rate	injur_perc	yes	soc
Exposed population to the landslide risk	land_risk	yes	env
NEET	neet	no	eco
Unemployment rate	unempl	yes	eco
Failure to work participation rate	part_lf	no	eco

The type of an indicator reflects the dimension to which it refers: eco=economic, env=environmental, soc=social.

Figure 11: Pairwise scatterplot

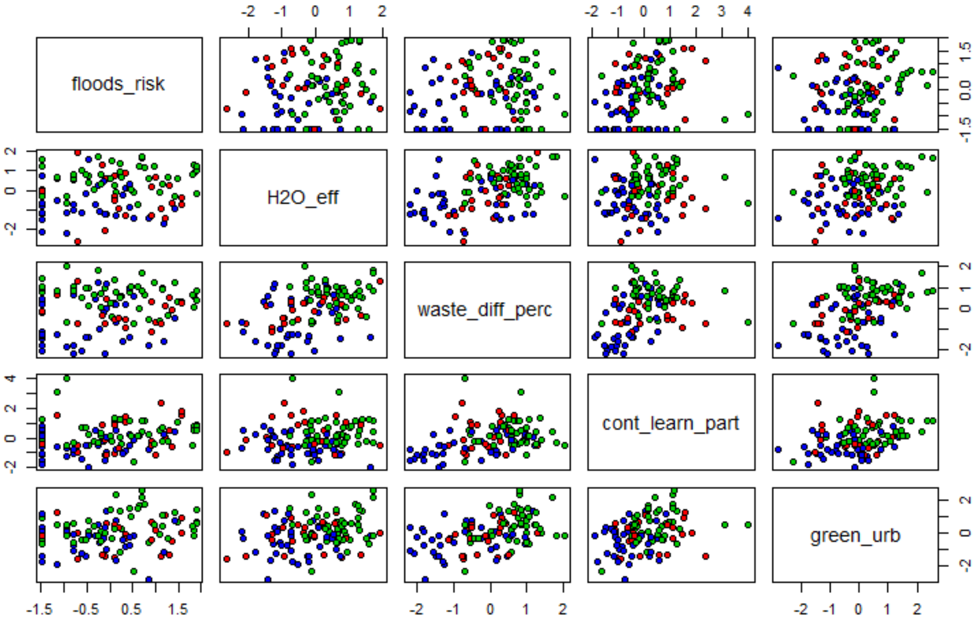


Figure 12: Hierarchical dendograms with different linkage functions

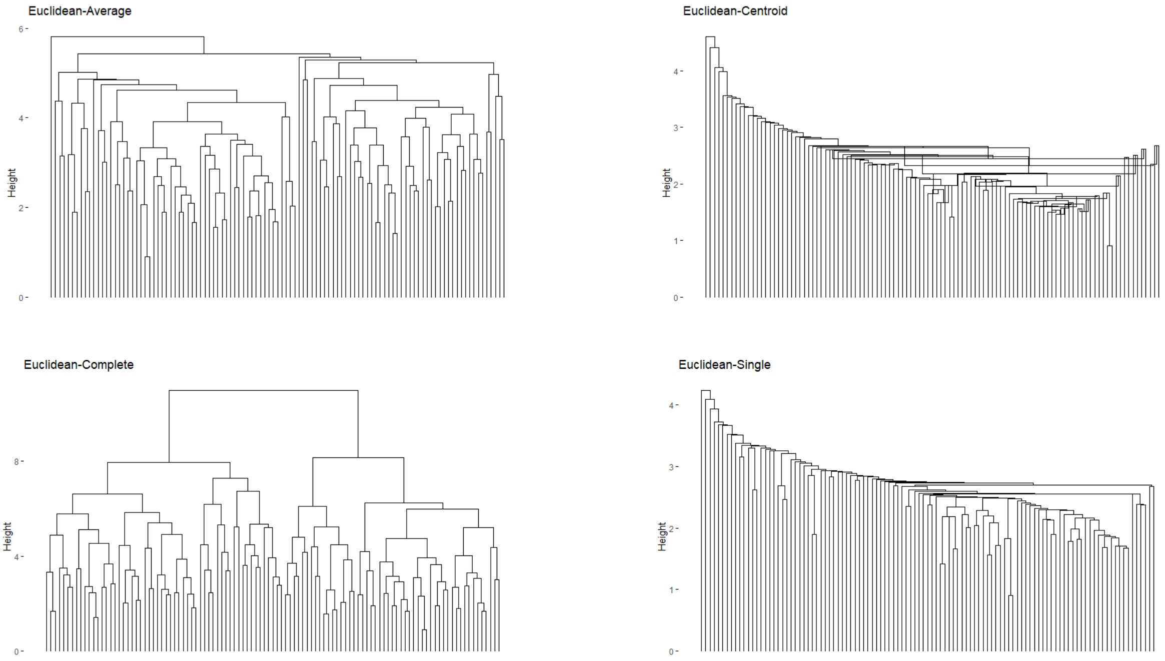


Figure 13: PCA biplot with loadings and data points

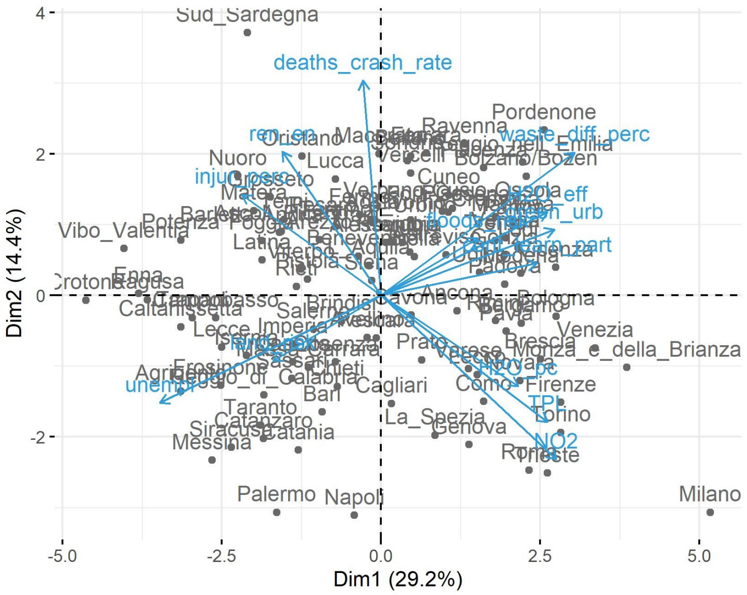
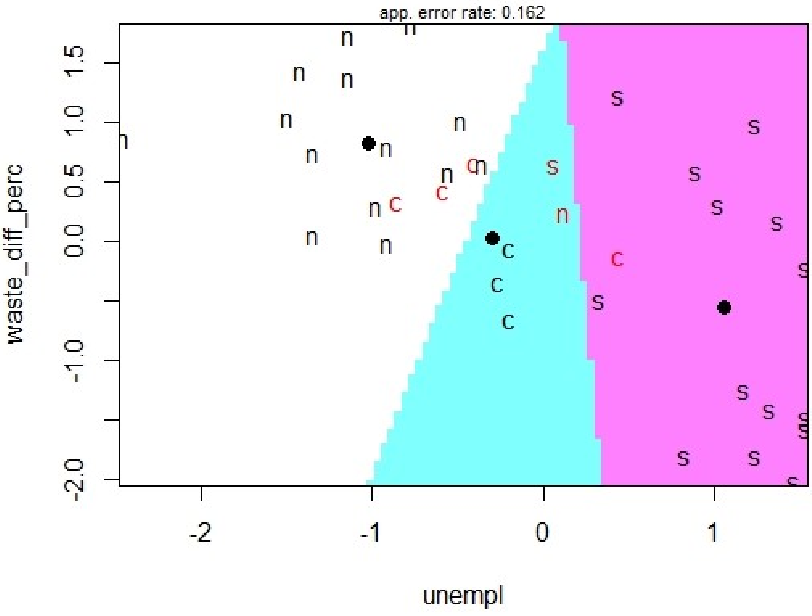


Figure 14: LDA partition plot on test data



Note: n stands for “North”, c for “Center” and s for “South”

.

A.2 R script

The present section aims to report all pre-processing steps we made to obtain our dataset from the original ISTAT one.

```r
#We use two basic libraries to perform the following steps:
library(dplyr)
library(writexl)
#Let’s import the dataset as a csv file and assign some NAs to harmonize
  #different symbols:
setwd("C:/Users/SuperChiara/Desktop/SLLD/Progetto")
d <- read.csv("Misure_statistiche_2004_2022.csv", sep = ";", dec = ",",
              na = c("....", "", " ", "-", "*", "..", "..."))
#FILTER BY AND RENAME PROVINCES
#The units of observation will be the Italian provinces, therefore we filter
  #first by DIMENSION "Territorio" and then by CLASS, excluding both the regions
  #and the macro geographical areas.
d <- d[d$DIMENSIONE == "Territorio",]
d <- d[!(d$CLASSE %in% c("Valle_d_Aosta/Vallee_d_Aoste", "Piemonte", "Liguria",
                         "Lombardia","Trentino-Alto_Adige/Sudtirol",
                         "Friuli-Venezia_Giulia", "Veneto", "Emilia-Romagna",
                         "Toscana", "Umbria", "Marche", "Lazio", "Abruzzo",
                         "Campania","Molise", "Calabria","Puglia", "Basilicata",
                         "Sicilia","Sardegna", "Italia",
                         "Totale_comuni_capoluogo", "Fiumi_Interregionali",
                         "Globali", "Dati_non_indicati", "EXTRA-REGIO_NUTS_1",
                         "Provincia_Autonoma_di_Trento",
                         "Provincia_Autonoma_di_Bolzano/Bozen",
                         "Nord", "Nord-ovest", "Nord-est", "Centro",
                         "Mezzogiorno", "Sud","Isole")),]
#We thereby rename some provinces to overcome data redundancy, which is due to
  # the several territorial reforms that occured in Italy in the last decade
d$CLASSE[d$CLASSE == "Forli"] <- "Forli-Cesena"
d$CLASSE[d$CLASSE %in% c("Andria", "Barletta","Trani")]<-"Barletta-Andria-Trani"
d$CLASSE[d$CLASSE == "Massa"] <- "Massa-Carrara"
d$CLASSE[d$CLASSE == "Monza"] <- "Monza_e_della_Brianza"
d$CLASSE[d$CLASSE == "Pesaro"] <- "Pesaro_e_Urbino"
d$CLASSE[d$CLASSE == "Verbania"] <- "Verbano-Cusio-Ossola"
d$CLASSE[d$CLASSE %in% c("Carbonia", "Carbonia-Iglesias",
                         "Medio_Campidano")] <- "Sud_Sardegna"
d$CLASSE[d$CLASSE == "Ogliastra"] <- "Nuoro"
```

```
d <- d[!(d$CLASSE == "Olbia-Tempio"),]
#REMOVE DUPLICATES
#We observe that two different statistical surveys are linked to most measures.
  #22 duplicate measures and 107 provinces
#We decide to compute the average of the numbers provided by the different
  #surveys and assign each observation this mean value.
d_duplicates <- d %>% group_by(CLASSE, MISURA_STATISTICA) %>%
  count() %>%
  filter(n != 1)
duplicate_measures<-c("Acqua_erogata_pro_capite",
                      "Efficienza_delle_reti_di_distribuzione_dell_acqua_potabile",
                      "Popolazione_esposta_al_rischio_di_alluvioni",
                      "Popolazione_esposta_al_rischio_di_frane",
                      "Incidenza_delle_aree_di_verde_urbano_sulla_superficie_urbanizzata_delle_citta",
                      "NO2_Biossido_di_azoto._Concentrazione_media_annuale_nei_comuni_capoluogo_di_provincia/citta_metropolitana",
                      "O3_Ozono._Numero_di_giorni_di_superamento_dell_obiettivo_nei_comuni_capoluogo_di_provincia/citta_metropolitana",
                      "PM10_Concentrazione_media_annuale_nei_comuni_capoluogo_di_provincia/citta_metropolitana",
                      "PM2.5_Concentrazione_media_annuale_nei_comuni_capoluogo_di_provincia/citta_metropolitana",
                      "Superamenti_del_valore_limite_giornaliero_previsto_per_il_PM10_nei_comuni_capoluogo_di_provincia/citta_metropolitana",
                      "Ammontare_di_rifiuti_urbani_oggetto_di_raccolta_differenziata",
                      "Energia_elettrica_da_fonti_rinnovabili",
                      "Giovani_che_non_lavorano_e_non_studiano_(NEET)",
                      "Numero_morti_in_incidente_stradale",
                      "Partecipazione_alla_formazione_continua",
                      "Posti-km_offerti_dal_Tpl",
                      "Raccolta_differenziata_dei_rifiuti_urbani",
                      "Tasso_di_disoccupazione",
                      "Tasso_di_infortuni_mortali_e_inabilita_permanente",
                      "Tasso_di_mancata_partecipazione_al_lavoro",
                      "Tasso_di_mortalita_per_incidente_stradale",
                      "Tasso_di_occupazione_(20-64_anni)")
duplicate_prov <- c("Agrigento", "Alessandria", "Ancona", "Aosta", "Arezzo",
                    "Ascoli_Piceno" ,"Asti", "Avellino", "Bari",
                    "Barletta-Andria-Trani", "Belluno", "Benevento", "Bergamo",
                    "Biella", "Bologna", "Bolzano/Bozen", "Brescia",
                    "Brindisi", "Cagliari", "Caltanissetta", "Campobasso",
                    "Caserta", "Catania", "Catanzaro", "Chieti",
                    "Como", "Cosenza", "Cremona", "Crotone","Cuneo", "Enna",
                    "Fermo", "Ferrara","Firenze", "Foggia", "Forli-Cesena",
                    "Frosinone","Genova", "Gorizia",  "Grosseto", "Imperia",
                    "Isernia", "L_Aquila", "La_Spezia","Latina","Lecce","Lecco",
                    "Livorno", "Lodi","Lucca", "Macerata", "Mantova",
                    "Massa-Carrara","Matera", "Messina", "Milano", "Modena",
                    "Monza_e_della_Brianza", "Napoli", "Novara", "Nuoro",
                    "Oristano", "Padova", "Palermo", "Parma","Pavia", "Perugia",
                    "Pesaro_e_Urbino", "Pescara","Piacenza", "Pisa", "Pistoia",
                    "Pordenone","Potenza", "Prato", "Ragusa", "Ravenna",
                    "Reggio_di_Calabria", "Reggio_nell_Emilia", "Rieti","Rimini"
    ,
                    "Roma", "Rovigo", "Salerno", "Sassari","Savona", "Siena",
                    "Siracusa", "Sondrio","Sud_Sardegna", "Taranto", "Teramo",
                    "Terni", "Torino", "Trapani", "Trento", "Treviso",
                    "Trieste", "Udine","Varese","Venezia","Verbano-Cusio-Ossola"
    ,
                    "Vercelli", "Verona", "Vibo_Valentia","Vicenza", "Viterbo")
years <- c("X2004", "X2005", "X2006",
           "X2007", "X2008", "X2009", "X2010", "X2011", "X2012",
           "X2013", "X2014", "X2015", "X2016", "X2017", "X2018",
           "X2019", "X2020", "X2021", "X2022")
#We perform this strategy by iterating with for loops over provinces, measures
  #and years and updating the original dataset at each iteration.
for (prov in duplicate_prov){
    for (mis in duplicate_measures){
    a<- d[d$CLASSE == prov & d$MISURA_STATISTICA == mis, ]
    if(nrow(a) == 0){
      next
    }
    for (y in years){
      sumcol <- 0
      n_rows <- 0
      for (row in 1:nrow(a)){
        if (is.na(a[row, y])){
          next
        }
        sumcol <- sumcol + as.numeric(sub(",", ".", a[row, y], fixed = TRUE))
        #decimali con virgola --> punto
        n_rows <- n_rows + 1
      }
      m <- sumcol/n_rows
      a[, y] <- m
    }
    d[d$CLASSE == prov & d$MISURA_STATISTICA == mis, ] <- a
  }
}
#Let's count the number of observations for each statistical measure.
#We remove four of them which have a much lower occurency with respect to the
  #others, since they are reported in the dataset for the 24 biggest cities only.
d %>% group_by(MISURA_STATISTICA) %>% count() %>% View()
low_occurrency <- c("Razionamento_dell_erogazione_dell_acqua_per_uso_domestico_per_parte_o_tutto_il_territorio_comunale",
                    "Numero_di_Giorni_estivi_(anomalie_rispetto_alla_normale_climatologica_1971-2000_nei_capoluoghi_di_Regione_e_citta_metropolitane)",
                    "Numero_di_Notti_tropicali_(anomalie_rispetto_alla_normale_climatologica_1971-2000_nei_capoluoghi_di_Regione_e_citta_metropolitane)",
                    "Numero_di_giorni_senza_pioggia_(anomalie_rispetto_alla_normale_climatologica_1971-2000_nei_capoluoghi_di_Regione_e_citta_metropolitane)")
d <- d[!(d$MISURA_STATISTICA %in% low_occurrency),]
#We remove some columns which are not of main interest in order to cut duplicate
  #rows using distinct().
colnames(d)[1] <- "GOAL"
d <- d %>% select(-c("GOAL", "GLOBAL_INDICATOR", "NOTE")) %>% distinct()
#Two Sardinian provinces still have some duplicates. We run two more loops to
  #remove them.
d %>% filter(COD_MISURA == "SDG-90") %>% group_by(CLASSE) %>% count() %>% View()
d_SudS <- d[d$CLASSE == "Sud_Sardegna" & d$COD_MISURA == "SDG-90",]
d_Nuoro <- d[d$CLASSE == "Nuoro" & d$COD_MISURA == "SDG-90",]
#fix Sud Sardegna
for (y in years){
  sumcol <- 0
  n_rows <- 0
  for (row in 1:nrow(d_SudS)){
    if (is.na(d_SudS[row, y])){
      next
```

```r
    }
    sumcol <- sumcol + as.numeric(sub(",", ".", d_SudS[row, y], fixed = TRUE))
    n_rows <- n_rows + 1
  }
  m <- sumcol/n_rows
  d_SudS[, y] <- m
}
d[d$CLASSE == "Sud_Sardegna" & d$COD_MISURA == "SDG-90",] <- d_SudS
#fix Nuoro
for (y in years){
  sumcol <- 0
  n_rows <- 0
  for (row in 1:nrow(d_Nuoro)){
    if (is.na(d_Nuoro[row, y])){
      next
    }
    sumcol <- sumcol + as.numeric(sub(",", ".", d_Nuoro[row, y], fixed = TRUE))
    n_rows <- n_rows + 1
  }
  m <- sumcol/n_rows
  d_Nuoro[, y] <- m
}
d[d$CLASSE == "Nuoro" & d$COD_MISURA == "SDG-90",] <- d_Nuoro
#We check that each statistical measure appears exactly 107 times, which is the
  #number of provinces.
d <- d %>% distinct()
d %>% group_by(MISURA_STATISTICA) %>% count() %>% View()
#SELECT THE BENCHMARK YEAR
#We count the number of NAs that occur for each year.
#Our rule-of-thumb is to pick 2015 as the reference year, since it has by far
  #the lowest NAs occurency, 77 (out of 2354 rows).
d[ d == "NaN"] <- NA
sapply(d, function (x) sum(is.na(x))) %>% View()
#Let's be wary and save a copy of the dataframe with all years, before selecting
  #some columns of interest.
d_allyears <- d
#DEALING WITH NAs#We clean our dataset further, leaving out two measures which have more than 20
  #NAs each.
d <- d[!(d$MISURA_STATISTICA %in% c("PM2.5_Concentrazione_media_annuale_nei_comuni_capoluogo_di_provincia/citta_metropolitana",
                             "O3_Ozono._Numero_di_giorni_di_superamento_dell_obiettivo_nei_comuni_capoluogo_di_provincia/citta_metropolitana")), ]
#We are interested in filling in the remaining NAs of year 2015. Let's first
  #vizualize them in a dataframe:
NAs_2015 <- d %>% filter(is.na(X2015))
#We use a mixed approach at this step. First, for each NA we observe whether
  #there are values reported for the immediately precedent or subsequent years
  #(at least two), in which case we assign 2015 the average value for the
  #statistical measure in these years.
years_close <- c("X2013", "X2014", "X2016", "X2017")
for (row in 1:nrow(d)){
  if(is.na(d[row, 20])){
    NAs_temporal <- is.na(d[row, 18]) + is.na(d[row, 19]) +
                        is.na(d[row, 21]) + is.na(d[row, 22])
    if(NAs_temporal > 2){
      next
    } else{
      sum <- 0
      obs <- 0
      for (y in years_close){
        if(is.na(d[row, y])){
          next
        }
        sum <- sum + as.numeric(d[row, y])
        obs <- obs + 1
      }
      m <- sum / obs
    }
    d[row, 20] <- m
  }
}
#There are still 20 NAs (we only filled in 9 with the last trick):
NAs_2015_new <- d %>% filter(is.na(X2015))
#We assign them the average value of the administrative region the province
  #belongs to for the specific statistical measure.
for (row in 1:nrow(d)){
  if(is.na(d[row, 20])){
    misure_regione <- d[d$MISURA_STATISTICA == d[row, 2] &
                          d$REGIONE == d[row, 8],]
    sum <- 0
    obs <- 0
```

```r
    for(r in 1:nrow(misure_regione)){
      if(is.na(misure_regione[r, 20])){
        next
      }
      sum <- sum + as.numeric(misure_regione[r, 20])
      obs <- obs + 1
    }
    m <- sum/obs
    d[row, 20] <- m
  }
}
#Still some missing values for Sardinia, were it is not possible to get a
  #regional average. Let's get back to the diacronic approach and select the
  #closest year as a reference
d[ d == "NaN"] <- NA
for (row in 1:nrow(d)){
  if(is.na(d[row, 20])){
    NAs_temporal <- is.na(d[row, 18]) + is.na(d[row, 19]) +
      is.na(d[row, 21]) + is.na(d[row, 22])
    sum <- 0
    obs <- 0
    for (y in years_close){
      if(is.na(d[row, y])){
        next
      }
      sum <- sum + as.numeric(d[row, y])
      obs <- obs + 1
      m <- sum / obs
    }
    d[row, 20] <- m
  }
}
#no more missing values for 2015!
d[ d == "NaN"] <- NA
d %>% filter(is.na(X2015)) %>% View()
any(is.na(d$X2015))
#INVERT DATASET
#We shall now turn our dataframe upside down, keeping provinces as rows and
  #statistical measures as columns
d <- d[order(d$CLASSE),]
territorio <- c(unique(d$CLASSE))
```

```
d <- d[order(d$MISURA_STATISTICA),]
misure <- c(unique(d$MISURA_STATISTICA))
d <- d[order(d$MISURA_STATISTICA),]
values_2015 <- as.numeric(c(d$X2015))
matrice_trasposta <- c(territorio,values_2015)
d_new <- matrix(matrice_trasposta, nrow = 107, ncol = 21)
d_new <- data.frame(d_new, row.names = 1)
#rename colums
colnames(d_new)[1:20] <- misure
#We convert the class of the numbers in the dataframe into numbers
for (m in misure){
  d_new[, m] <- as.numeric(d_new[, m])
}
#EXPORT EXCEL SHEET
any(is.na(d_new))
#we eventually add a column with the names of the provinces, otherwhise they
  #would be discarded.
d_new <- d_new %>% mutate(Provincia = c(territorio))
write_xlsx(d_new,
           "C:/Users/SuperChiara/Desktop/SLLD/Progetto/SDGs_project_pp.xlsx")
```

References

ISTAT (2021). *Rapporto SDGs 2021. Informazioni statistiche per l'Agenda 2030 in Italia.*

James, G., D. Witten, T. Hastie, and R. Tibshirani (2021). *An introduction to statistical learning.* New York: Springer.