ISTAT 2021 report on SDGs

SLLD Module 1
May 23rd 2022

Carlo Cignarella, Margherita Lanini, Marco Zeppi

Motivation and background

- Every year ISTAT publishes an annual report on Italian performance in achieving Sustainable Development Goals (SDGs).
- SDGs were defined by UN in 2015 as key indicators of the 2030 Agenda. They balance all the three dimensions of sustainable development: economic, social and environmental.
- Our research question: detecting emerging patterns at provincial level.

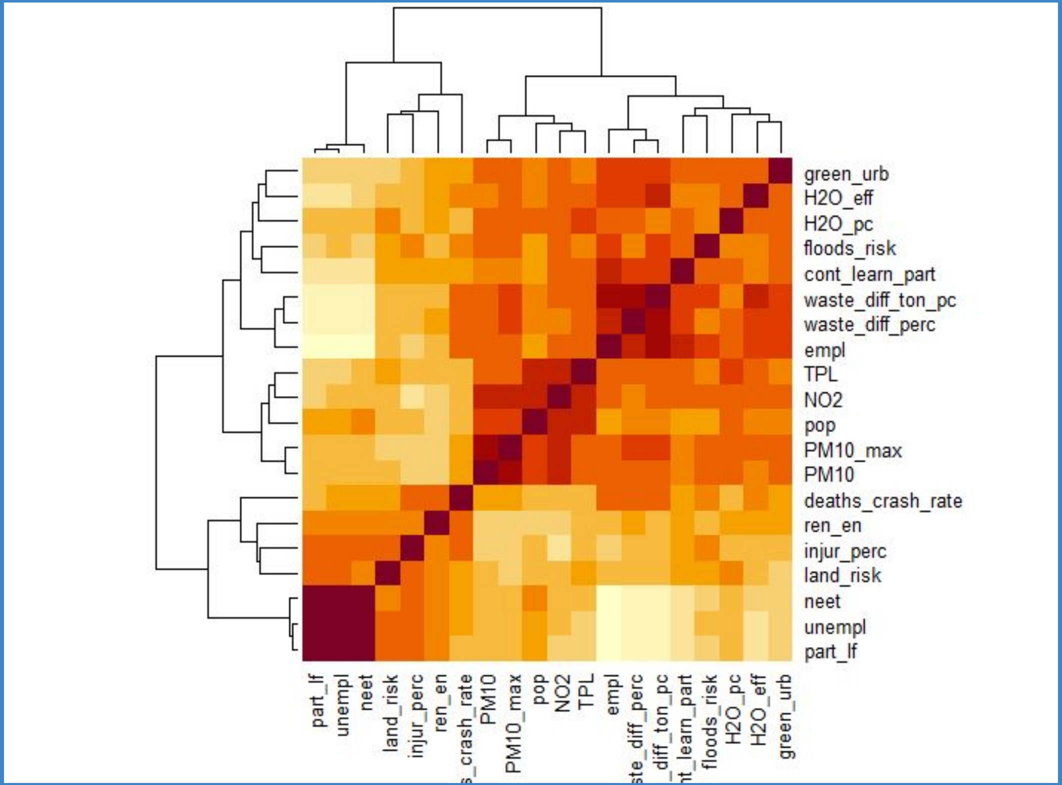
Outline

- Data description and selection
- Statistical and computational methods:
    - Clustering
    - Principal Component Analysis
    - Classification & cross-validation
- Further research

Data description

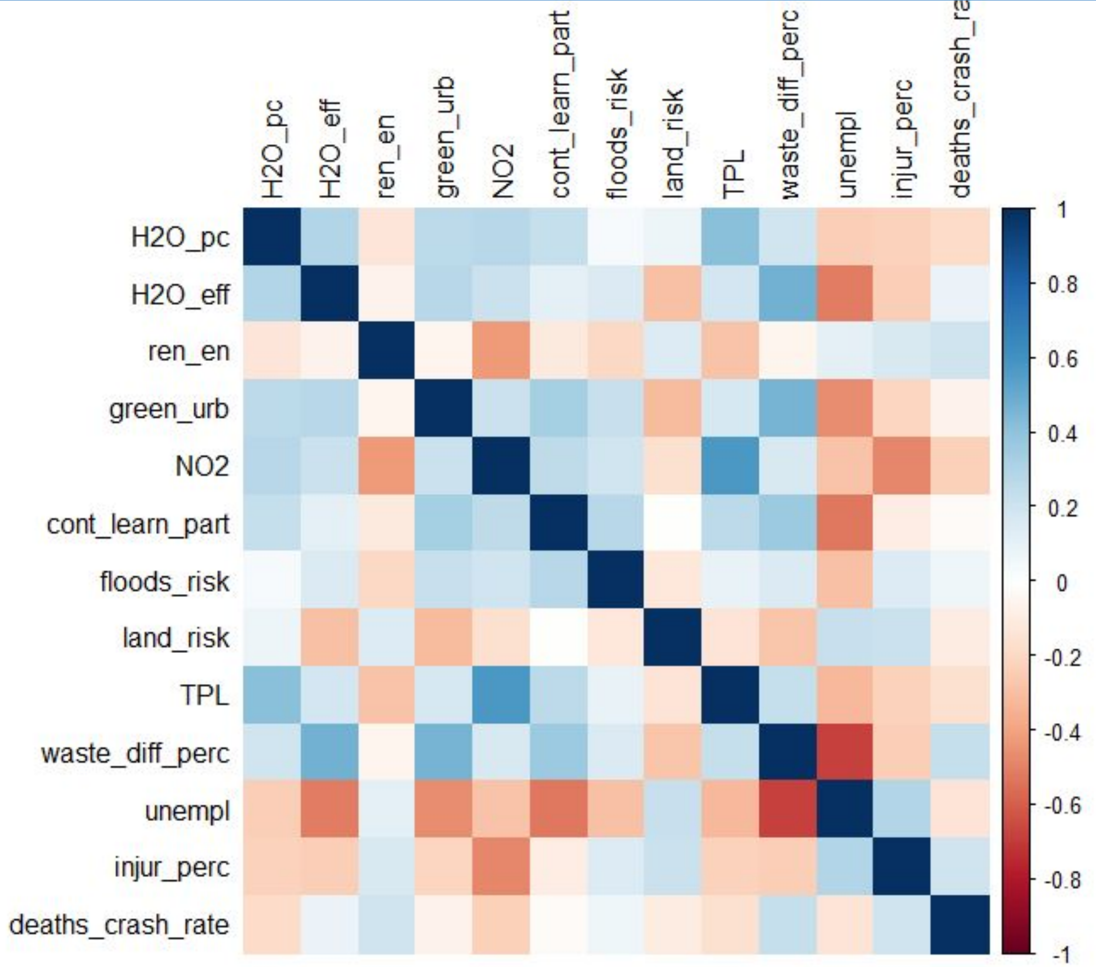
- The original dataset covers 2004-2020 period.
- It reports all 17 SDGs, the associated 169 targets and 232 indicators.
- We select all available 20 indicators for the 107 Italian provinces in 2015.
- We deal with NAs for 2015:
  - average 2013-2017;
  - average same region in 2015;
  - closest value within 2013-2017.

Data selection

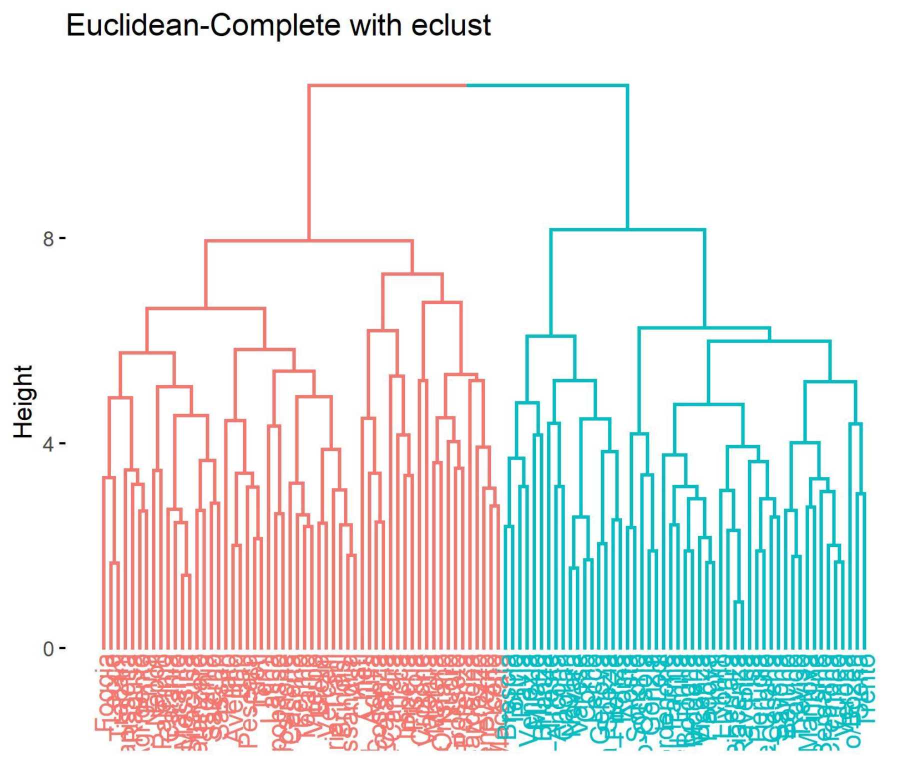


- Transformation, weighting and scaling

- Unemployment rate
  - NEET
  - non-participation rate
  - employment rate (20-64)
- NO2(air quality)
  - PM10
  - PM10 max
- Perc. of waste recycling
  - waste recycling (in tons)
  - population

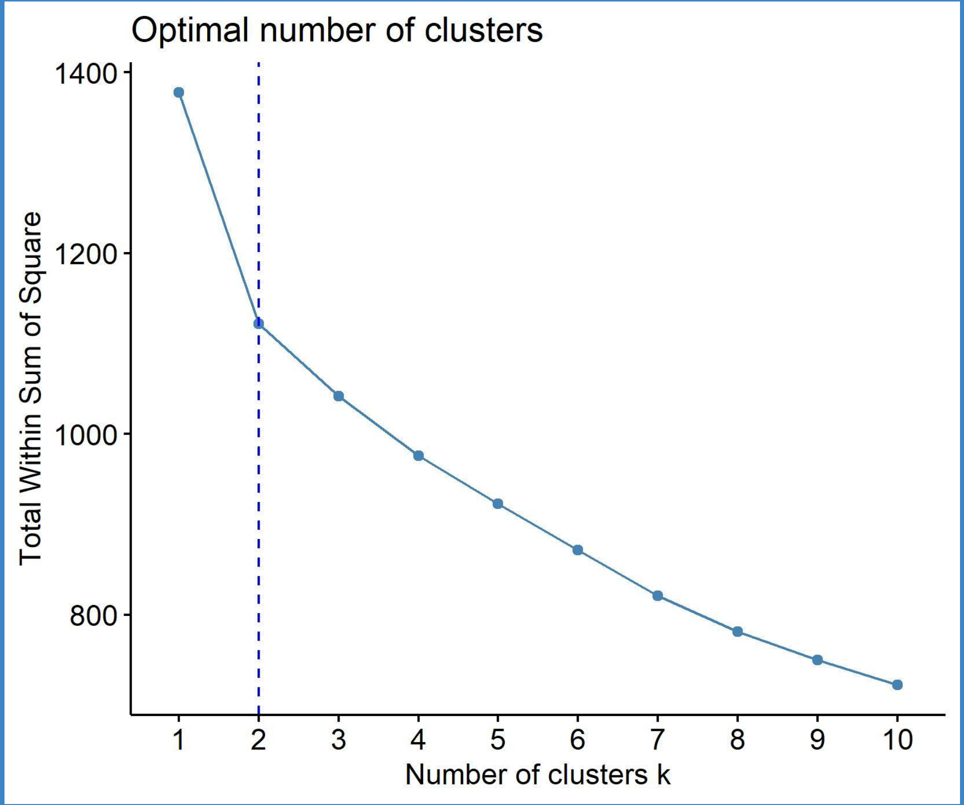

H2O_pc
H2O_eff
ren_en
green_urb
NO2
cont_learn_part
floods_risk
land_risk
TPL
waste_diff_perc
unempl
injur_perc
deaths_crash_ra
deaths_crash_rate
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

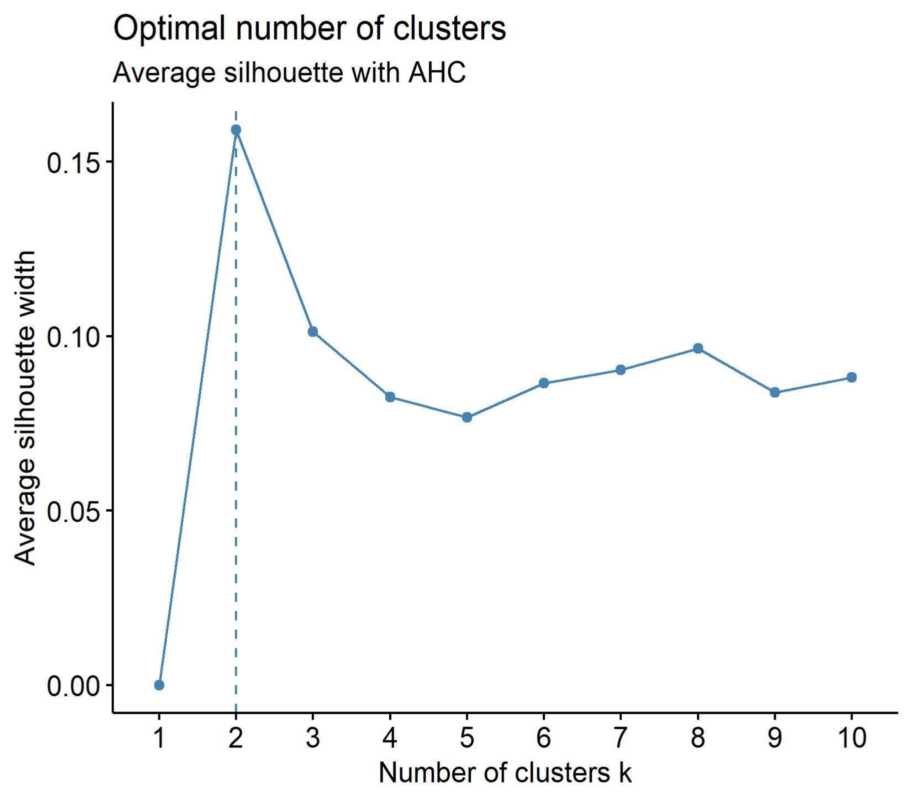
Clustering: an approach



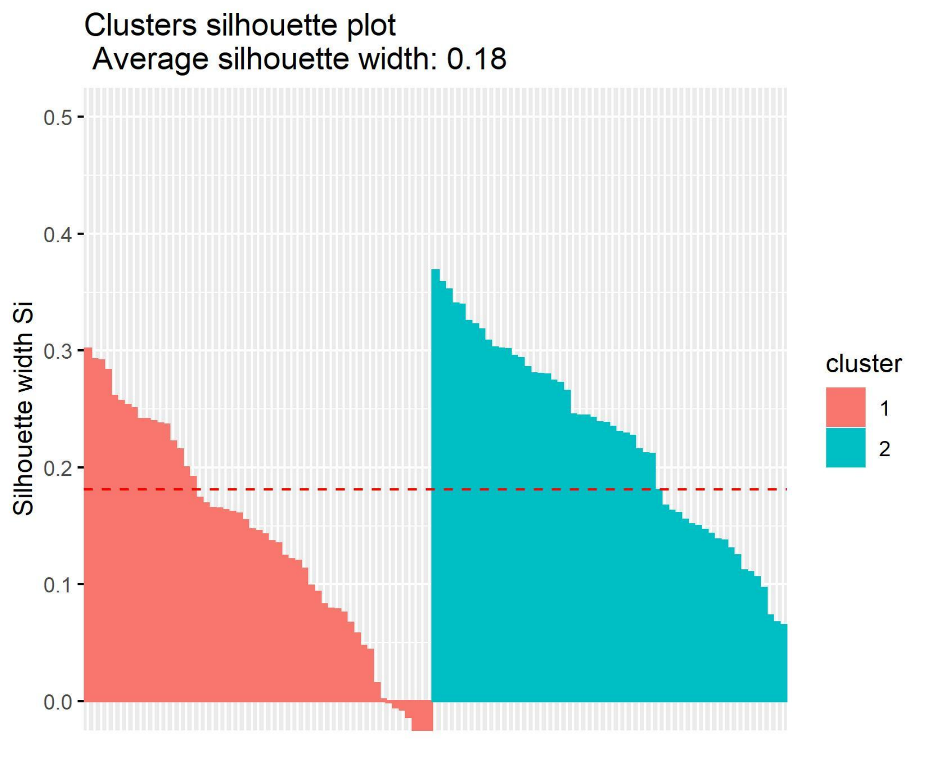
- Agglomerative hierarchical clustering
- Euclidean method
- Complete linkage function better suits the shape of our data
- Broad pattern: northern and central vs southern cities
- We did try k-means, obtaining similar results

Clustering: optimal k



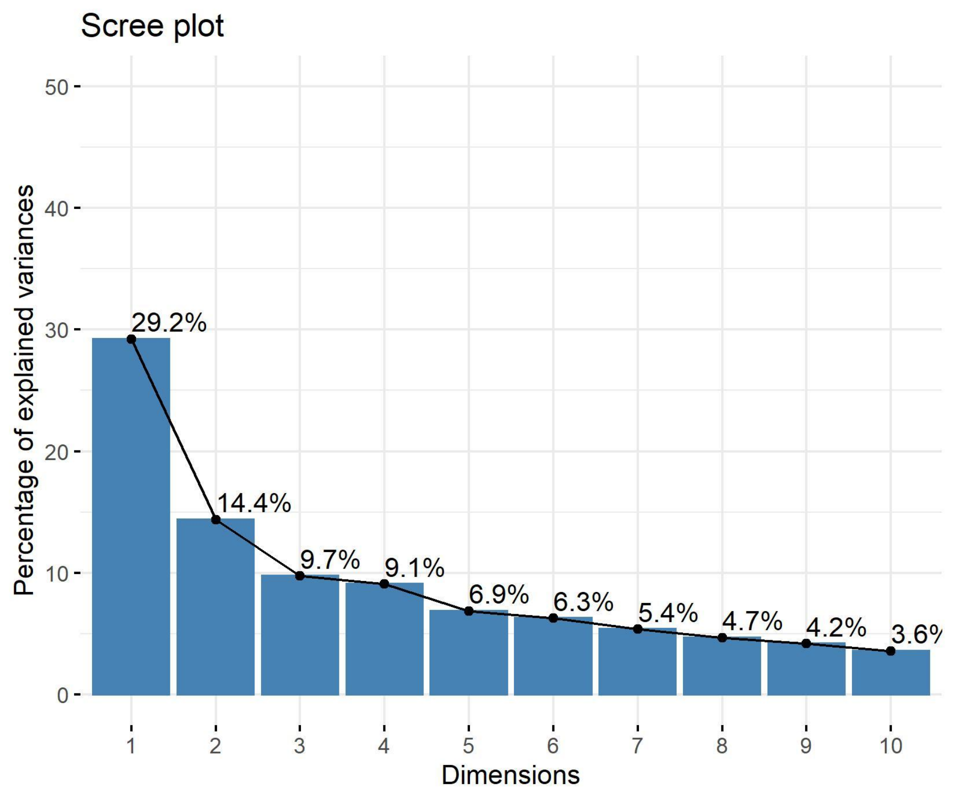


Clustering, refined



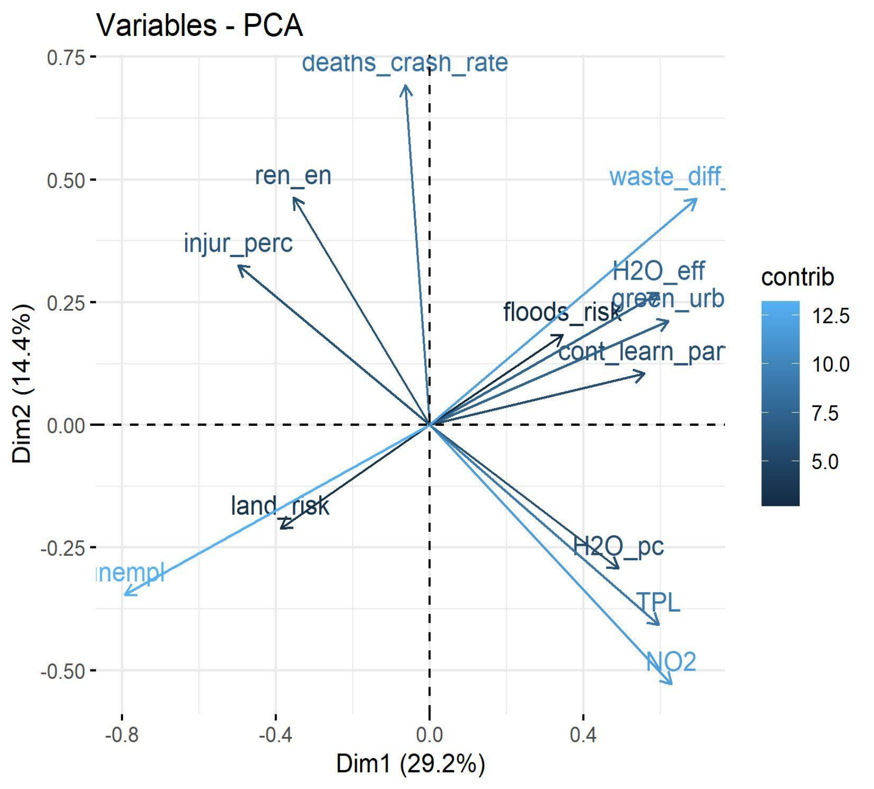
- Mixed approach: HC centroids used to initialize K-means
- Slight improvement in evaluation measures (see silhouette)
- Overall, quite poor performance in clustering

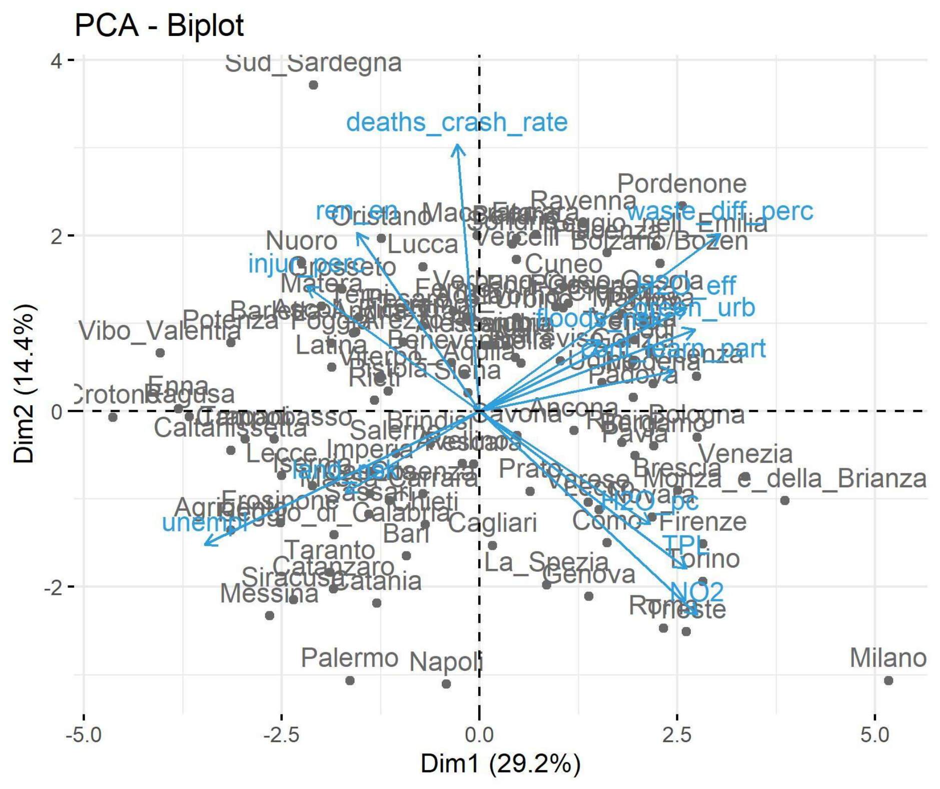
Principal Component Analysis



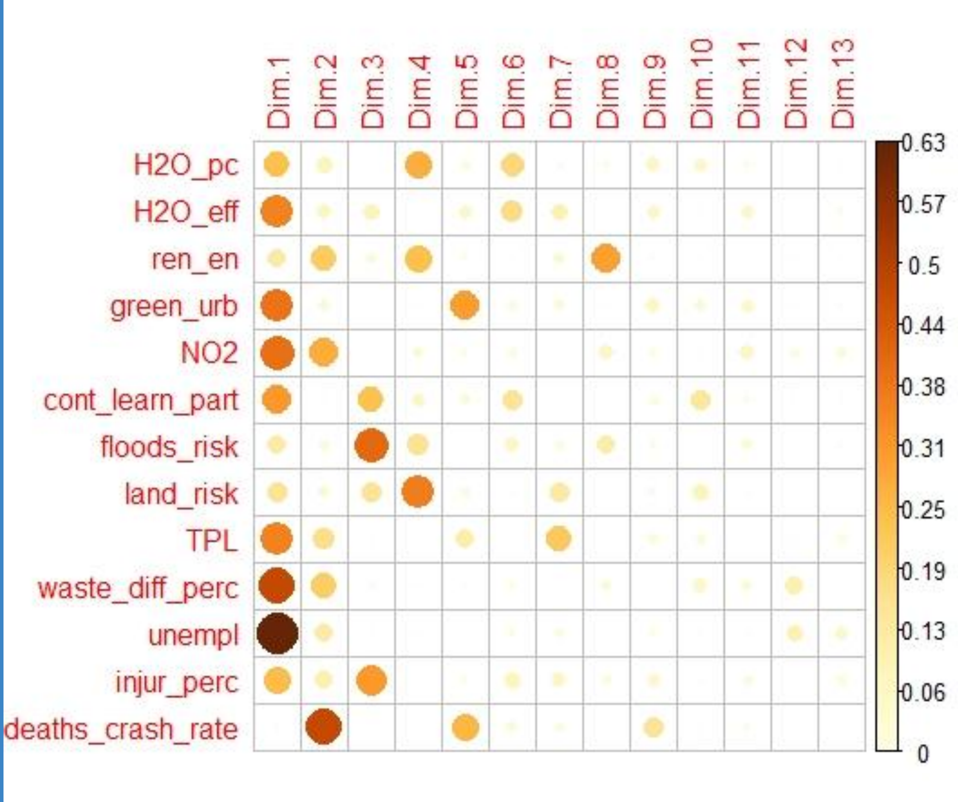
- The scree plot suggests the first two dimensions account for about 45% of data variance
- Less than desirable, but still not negligible
- Which variables weight more?

PCA: biplots


Variables - PCA
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.8
-0.4
0.0
0.4
Dim1 (29.2%)
Dim2 (14.4%)
deaths_crash_rate
ren_en
injur_perc
waste_diff_
H2O_eff
green_urb
floods_risk
cont_learn_par
land_risk
unempl
H2O_pc
TPL
NO2
contrib
12.5
10.0
7.5
5.0


PCA - Biplot
4
2
0
-2
-5.0
-2.5
0.0
2.5
5.0
Dim1 (29.2%)
Dim2 (14.4%)
Sud_Sardegna
deaths_crash_rate
Pordenone
Ravenna
waste_diff_perc
Reggio_nell_Emilia
Bolzano/Bozen
ren_en
Oristano
Lucca
Vercelli
Nuoro
Cuneo
injur_perc
Grosseto
Matera
H2O_eff
green_urb
Potenza
Foggia
Vibo_Valentia
floods_risk
Latina
Benevento
L'Aquila
Viterbo
Pistoia
Siena
Rieti
cont_learn_part
Modena
Padova
Crotone
Enna
Ragusa
Campobasso
Caltanissetta
Ancona
Savona
Bologna
Bergamo
Pavia
Salerno
Brindisi
Lecce
Imperia
Venezia
land_risk
Prato
Varese
Brescia
Monza_e_della_Brianza
Massa-Carrara
Sassari
Novara
H2O_pc
Frosinone
Chieti
Agrigento
Reggio_di_Calabria
unempl
Cagliari
Como
Firenze
Bari
TPL
Taranto
Torino
La_Spezia
Catanzaro
Genova
Siracusa
Catania
NO2
Messina
Roma
Trieste
Palermo
Napoli
Milano

PCA: corrplot



- First component:
  - unemployment, waste_diff_perc
  - H2O_eff, green_urb
- Second component:
  - deaths_crash_rate
- Both:
  - NO2, TPL

Classification: framework

- First strategy: LOGIT
  - predict waste sorting “champions” (dummy = 1 if above 3rd quartile of waste_diff_perc)
  - predictors: all 5 vars. related to environmental standards
  - 65% training, 35% test (seed is set to replicate results)
- Second strategy: Linear Discriminant Analysis
  - predict the macro territorial area (Nord-Centro-Sud) based on unemployment and waste sorting performance
  - same partition

Classification : logit results

- Table of coefficients: logit results on training data
    - urban green and water efficiency are highly significant

- Accuracy on test data = 0.757

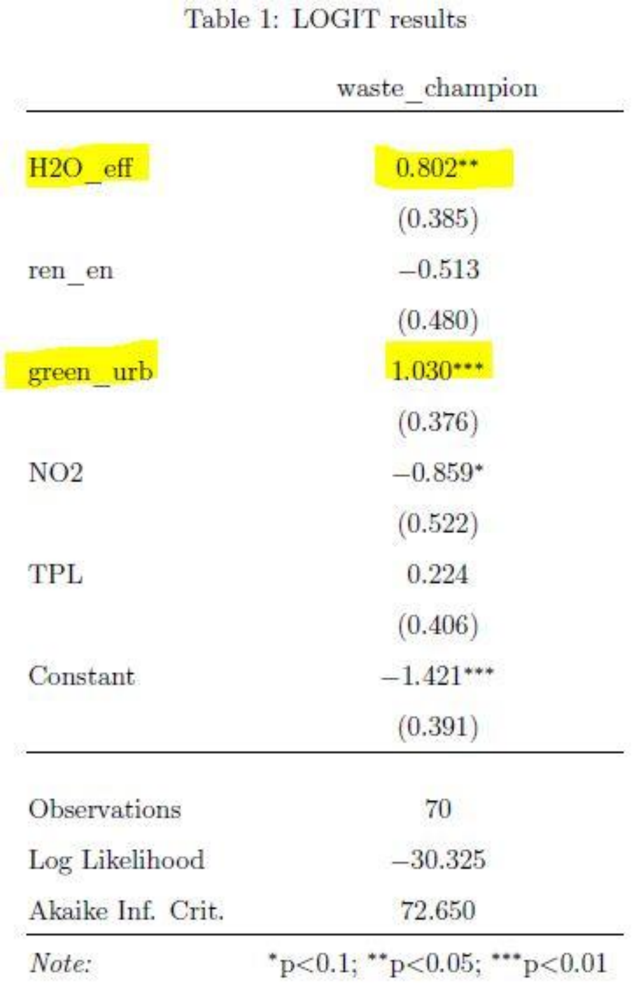
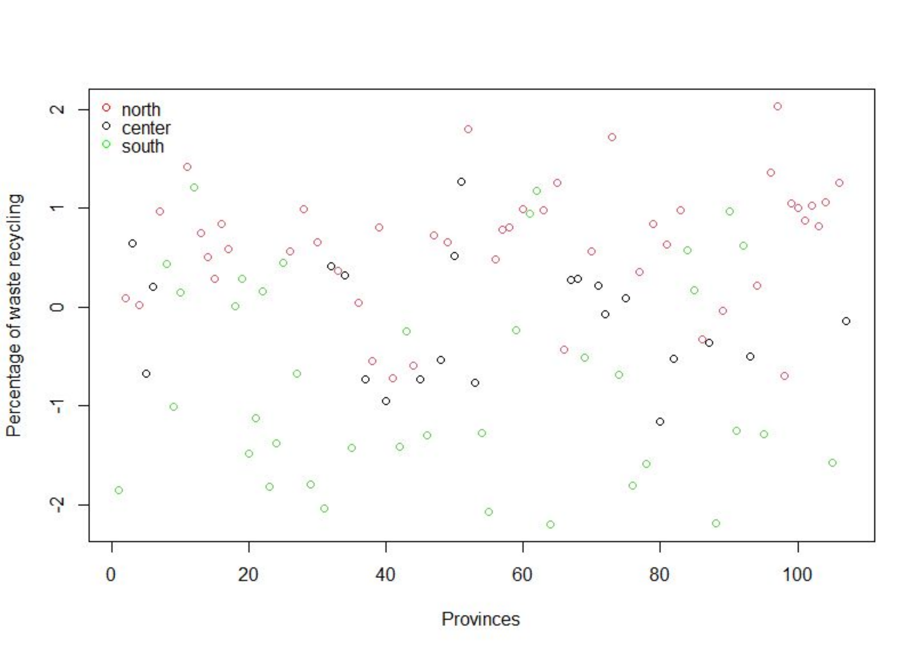


Table 1: LOGIT results

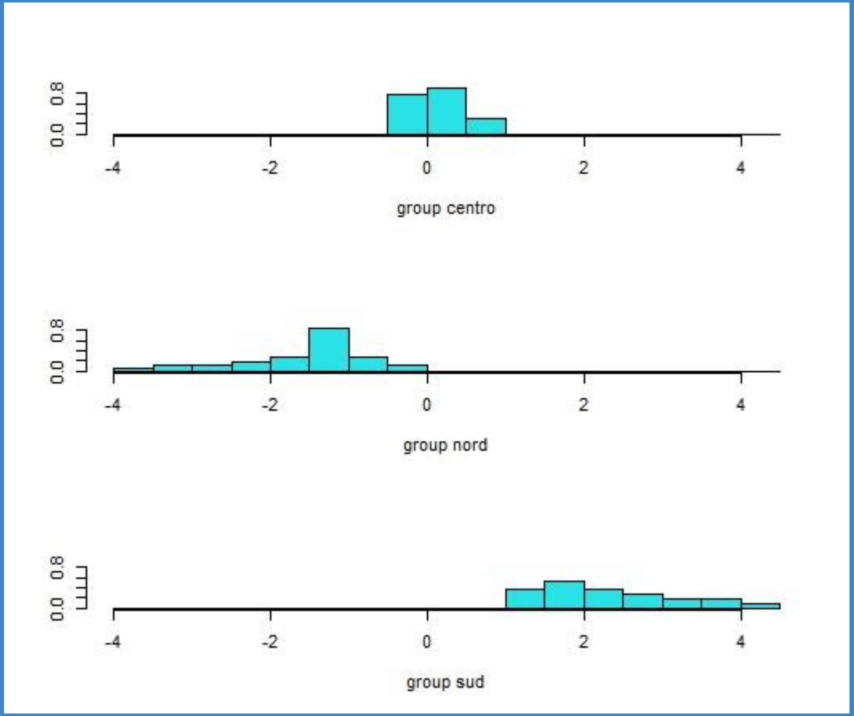
	waste_champion
H2O_eff	0.802^{**}
	(0.385)
ren_en	−0.513
	(0.480)
green_urb	1.030^{***}
	(0.376)
NO2	-0.859^{*}
	(0.522)
TPL	0.224
	(0.406)
Constant	-1.421^{***}
	(0.391)
Observations	70
Log Likelihood	−30.325
Akaike Inf. Crit.	72.650
Note:	*p<0.1; **p<0.05; ***p<0.01

Classification: LDA

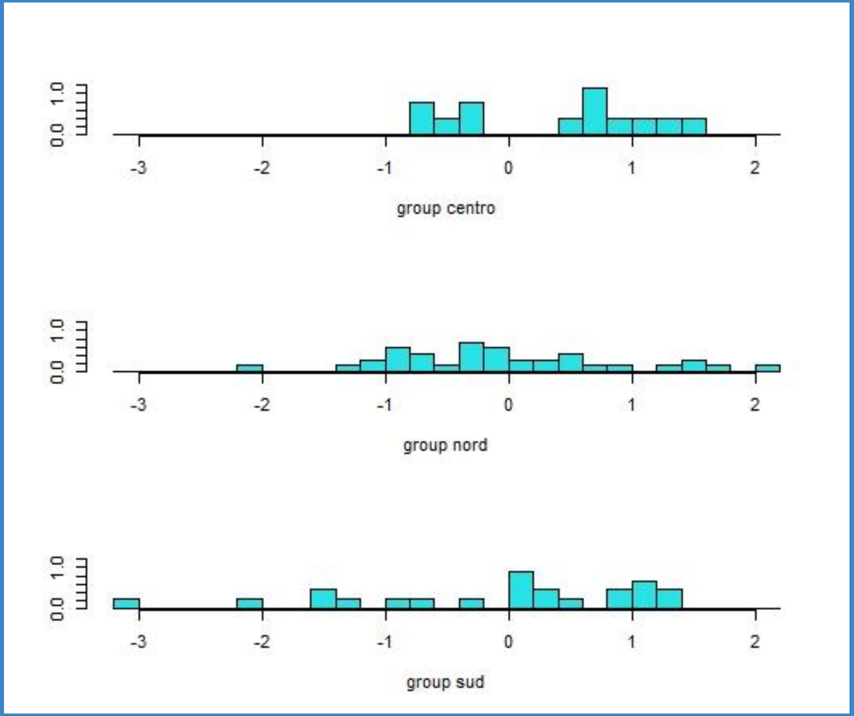


- ex ante distribution of outcome variable
- territorial labels have a pattern
  - northern provinces perform better than the average
  - higher dispersion for the south
  - fewer observations for the center

Classification: LDA results (train)

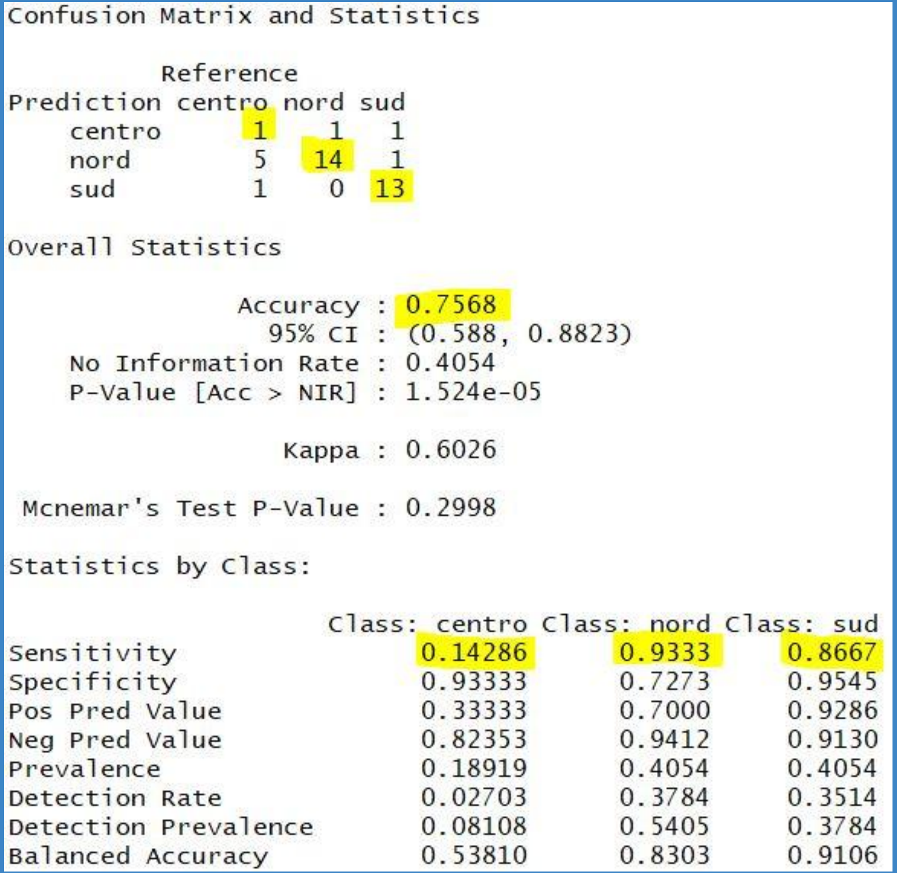


Percentage of waste sorted



Unemployment

Classification: LDA results (test)

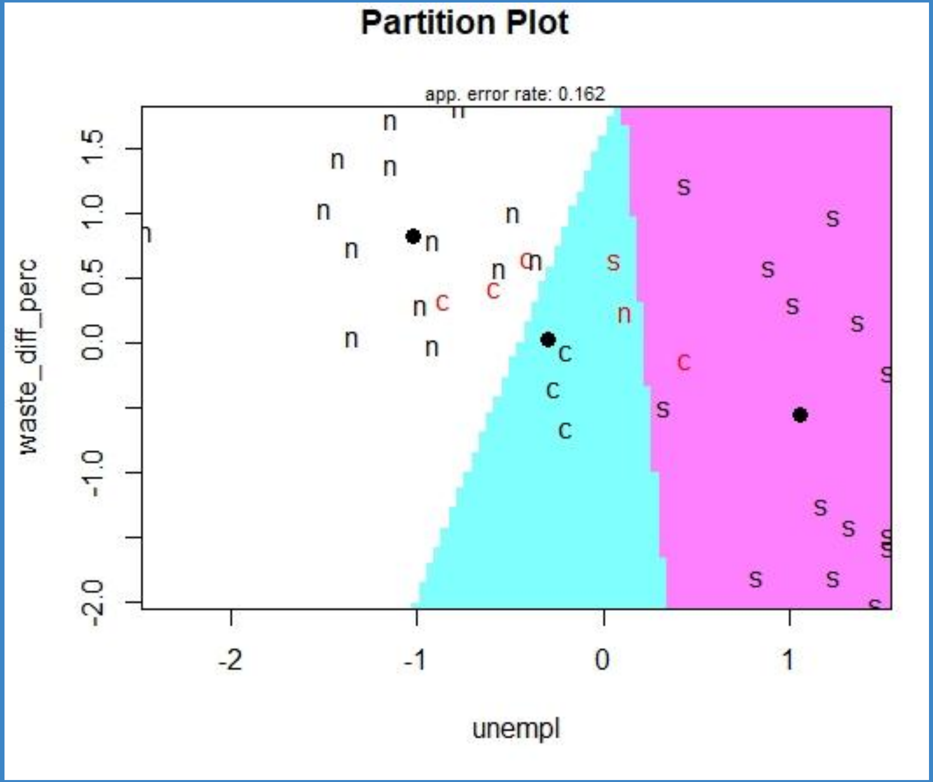


```
Confusion Matrix and Statistics

          Reference
Prediction centro nord sud
    centro      1    1   1
    nord        5   14   1
    sud         1    0  13

Overall Statistics

               Accuracy : 0.7568
                 95% CI : (0.588, 0.8823)
    No Information Rate : 0.4054
    P-Value [Acc > NIR] : 1.524e-05

                  Kappa : 0.6026

 Mcnemar's Test P-Value : 0.2998

Statistics by Class:

                     Class: centro Class: nord Class: sud
Sensitivity                0.14286      0.9333     0.8667
Specificity                0.93333      0.7273     0.9545
Pos Pred Value             0.33333      0.7000     0.9286
Neg Pred Value             0.82353      0.9412     0.9130
Prevalence                 0.18919      0.4054     0.4054
Detection Rate             0.02703      0.3784     0.3514
Detection Prevalence       0.08108      0.5405     0.3784
Balanced Accuracy          0.53810      0.8303     0.9106
```

Confusion matrix:

- (only) 27 test records
- good overall accuracy
- provinces in the center are widely misclassified

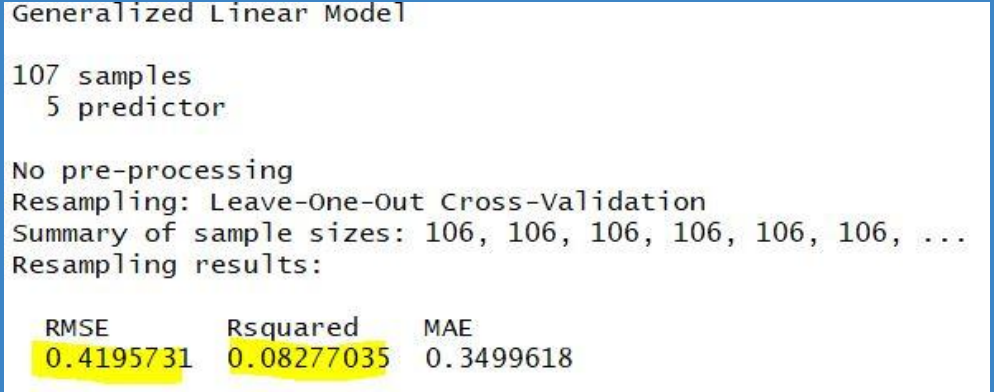
Classification: LDA results (test)



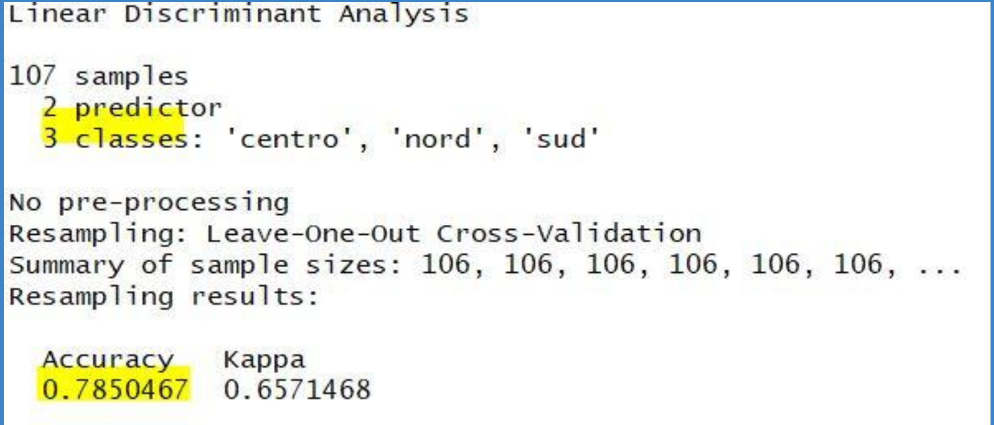
Cross-validation: framework

- Recall: 107 records (provinces), 13 variables
- LOOCV approach
- Performed to validate both (supervised) classification techniques

Cross-validation: results



```
Generalized Linear Model

107 samples
  5 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 106, 106, 106, 106, 106, 106, ...
Resampling results:

  RMSE       Rsquared    MAE
  0.4195731  0.08277035  0.3499618
```



```
Linear Discriminant Analysis

107 samples
  2 predictor
  3 classes: 'centro', 'nord', 'sud'

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 106, 106, 106, 106, 106, 106, ...
Resampling results:

  Accuracy   Kappa
  0.7850467  0.6571468
```

- logit
  - ○ High RMSE and MAE
  - ○ Rsquared low
- LDA
  - ○ Accuracy: 0.785!
  - ○ but…
- Overall, the latter classification strategy performs much better than the former.

Limitations and further research

LIMITATION:

- from the original dataset containing hundreds of statistical measures, we were left with 13 only after selecting the provincial dimension and filtering.

STRATEGIES:

- Possible strategy: change the level of observation, i.e. city-level focus;
- Select a different year to make a comparison;
- Panel approach to detect any diachronic pattern.

References

- ISTAT (2021). *Rapporto SDGs 2021. Informazioni statistiche per l'Agenda 2030 in Italia.*
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer Texts in Statistics. Springer, New York, NY.