Countries Socio-Economic Development Analysis

Statistical Learning Techniques

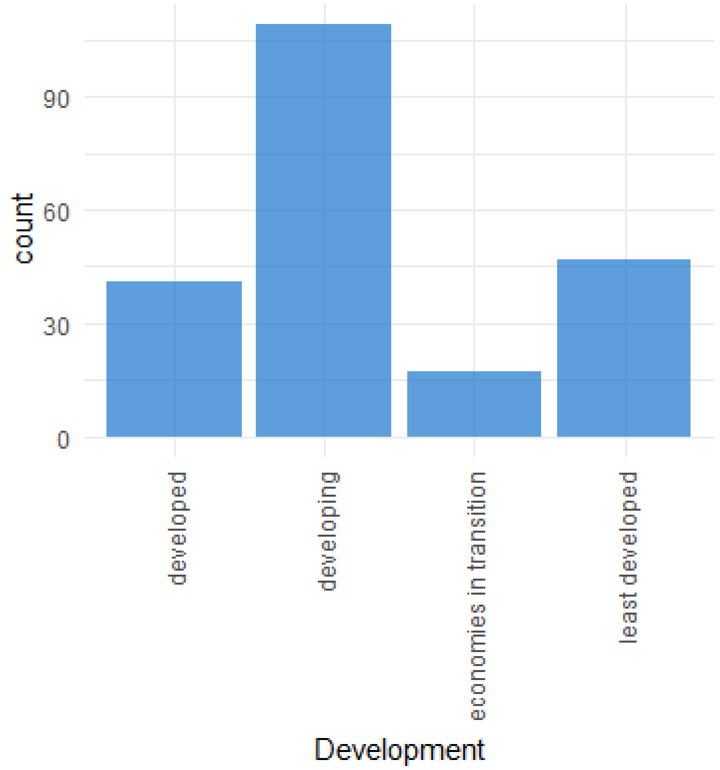
Angelina Khatiwada,

MSc Data Science and Economics, UNIMI

Nov, 2021

Dataset Exploratory Analysis

- **UNData countries dataset**
- **229 countries** (regions and development level)
- **52 key statistical indicators**:
  - GDP per capita (current US$)
  - Unemployment (% of labour force)
  - Food production index (2004-2006=100)
  - International trade: Imports (million US$)
  - Population age distribution (60+ years, %)
  - Health: Total expenditure (% of GDP)
- Columns and rows with > 50% of NAs removed: 214 countries and 51 indicators
- **Missing values imputation** with MissForest
- Data standardization, zero variance check



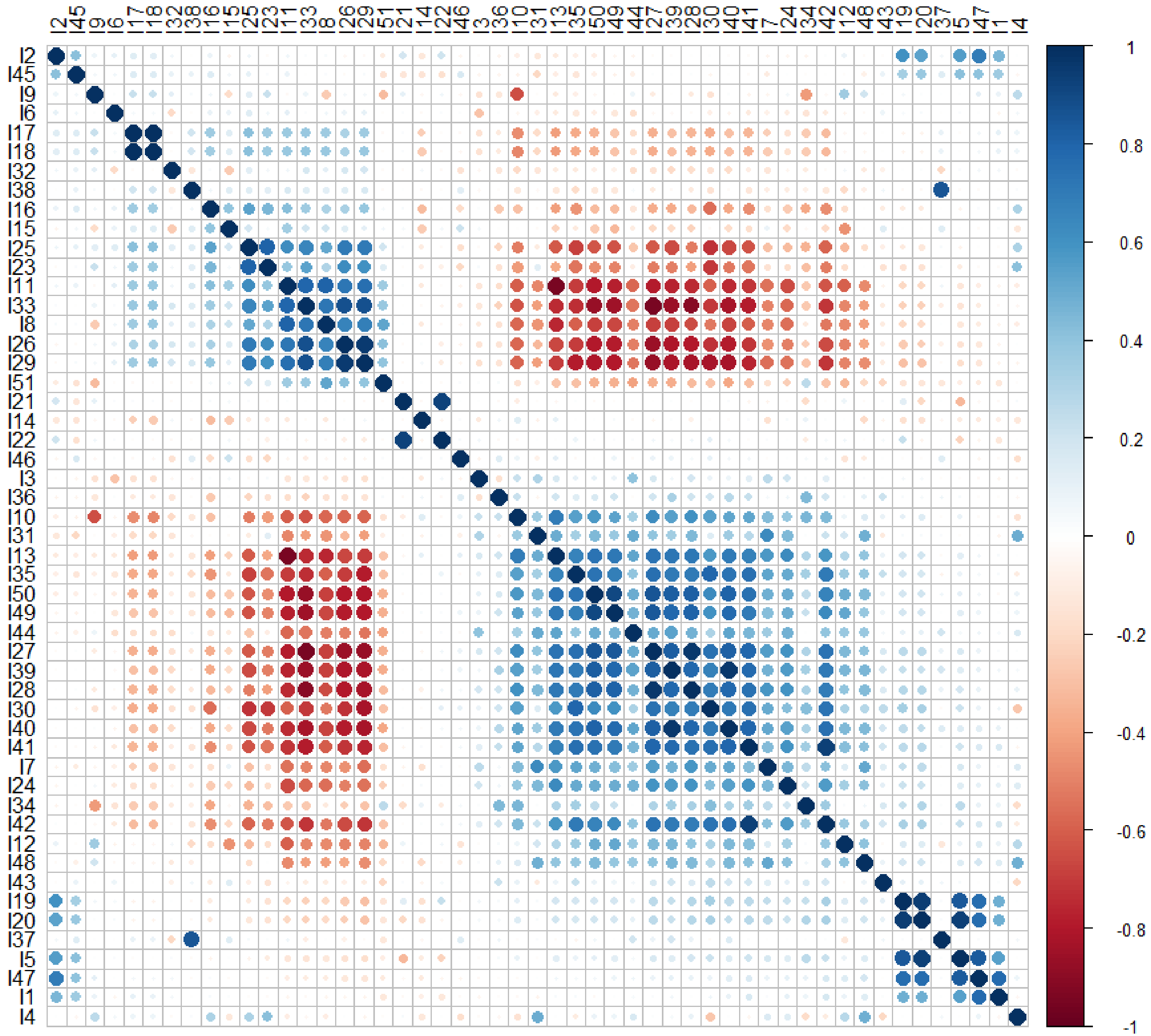
UNSUPERVISED LEARNING:

PCA, K-Means, Hierarchical Clustering

UNIVERSITÀ DEGLI STUDI DI MILANO

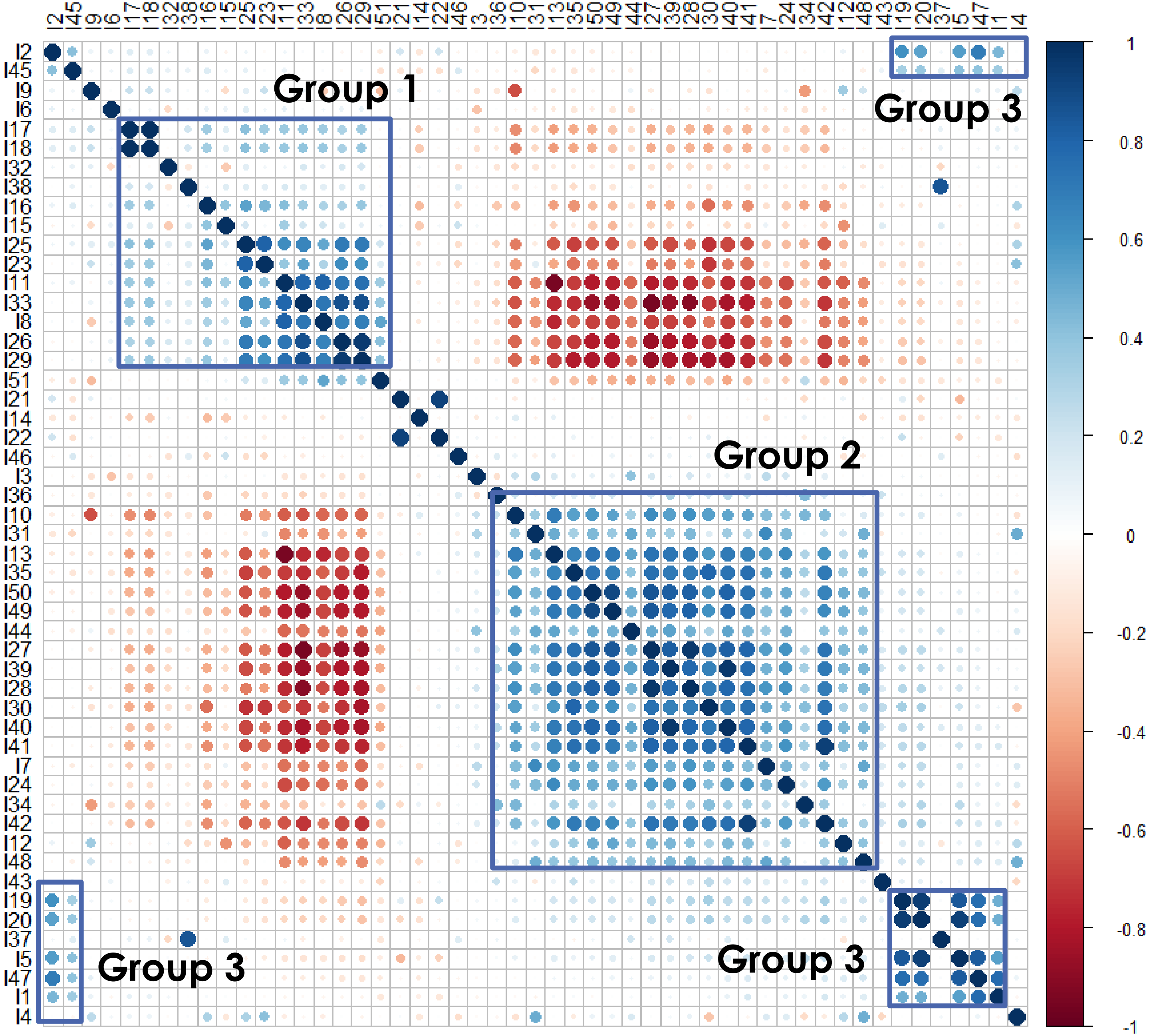
Correlation plot

- **3 groups of correlated indicators**

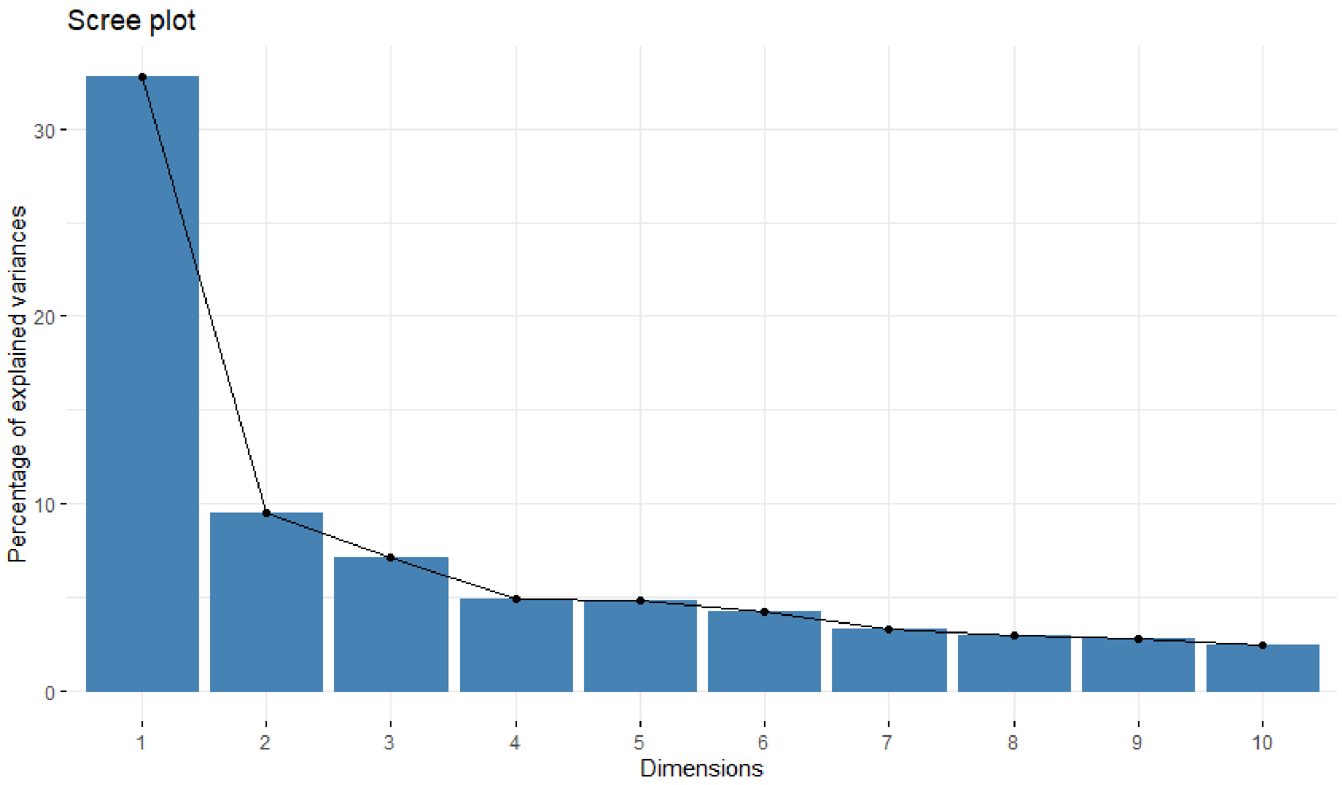


Correlation plot

- **3 groups of correlated indicators**
- **Group 2 as an example:**
    - I10 - Economy: Services and other activity (% of GVA)
    - I28 - Life expectancy at birth (males, years)
    - I30 - Population age distribution (60+ years, %)
    - I35 - Health: Physicians (per 1000 pop.)



Principal Component Analysis

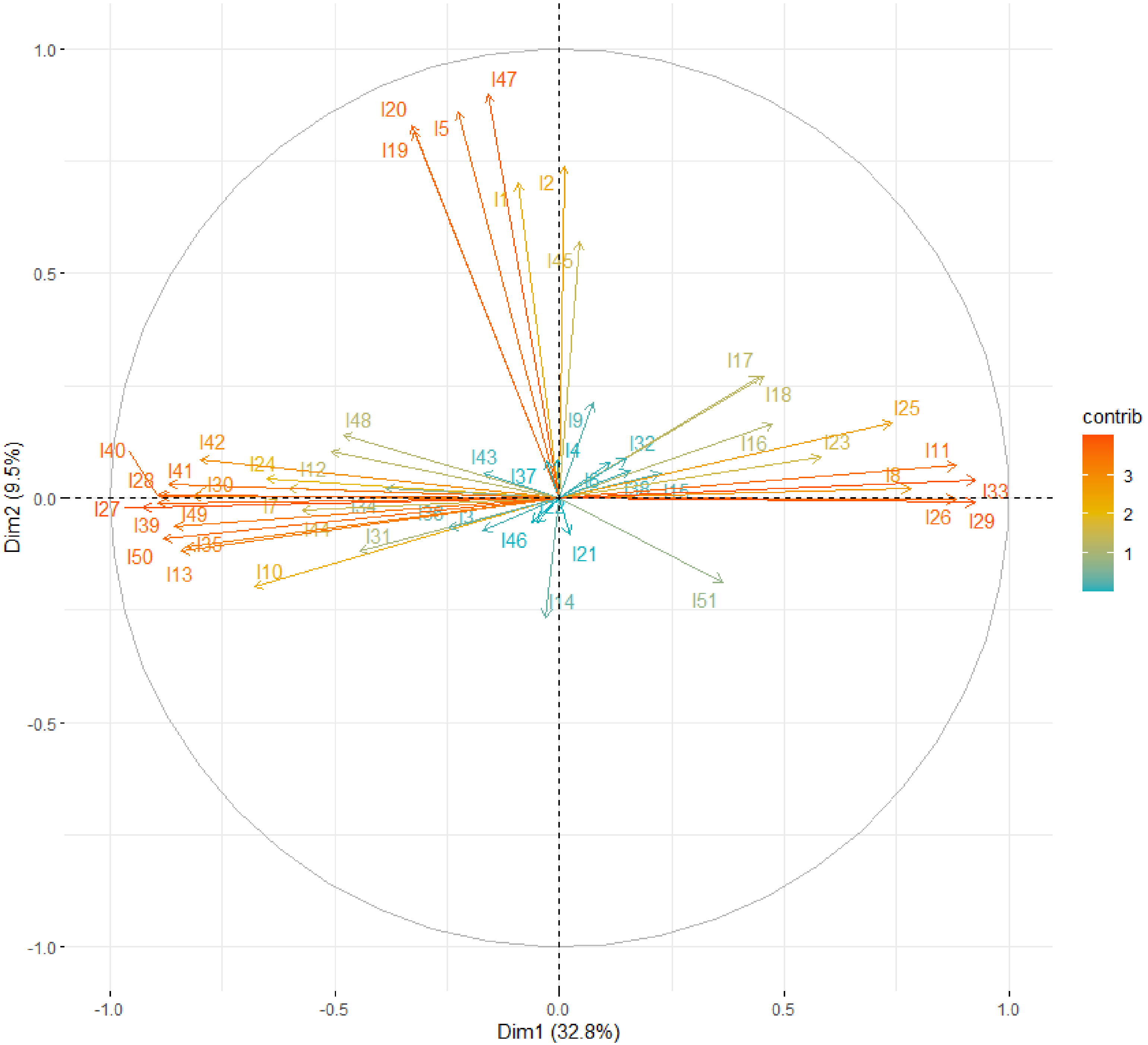


12 components to perform a dimensionality reduction (80% of variance)

PC	Eigenvalue	Variance percentage	Cumulative Variance Percentage
Dim.1	16.7087	32.7622	32.7622
Dim.2	4.8317	9.474	42.2362
Dim.3	3.6102	7.0788	49.315
Dim.4	2.5058	4.9133	54.2283
Dim.5	2.4781	4.8591	59.0874
Dim.6	2.1483	4.2124	63.2998
Dim.7	1.6692	3.2728	66.5726
Dim.8	1.5158	2.9721	69.5448
Dim.9	1.4004	2.7459	72.2907
Dim.10	1.2441	2.4395	74.7302
Dim.11	1.0936	2.1444	76.8746
Dim.12	1.0816	2.1208	78.9954
Dim.13	0.9431	1.8492	80.8446
Dim.14	0.9016	1.7679	82.6125
Dim.15	0.8327	1.6327	84.2452
Dim.16	0.7536	1.4777	85.7229
Dim.17	0.6152	1.2063	86.9291
Dim.18	0.6036	1.1834	88.1126
Dim.19	0.5271	1.0335	89.1461
Dim.20	0.5171	1.014	90.16
Dim.21	0.4926	0.966	91.126
Dim.22	0.4661	0.9139	92.04
Dim.23	0.4118	0.8075	92.8475
Dim.24	0.3804	0.7459	93.5934
Dim.25	0.3516	0.6893	94.2827
Dim.26	0.3369	0.6606	94.9433
Dim.27	0.2983	0.5849	95.5281
Dim.28	0.2903	0.5692	96.0973
Dim.29	0.2529	0.4959	96.5932
Dim.30	0.2306	0.4522	97.0454
Dim.31	0.2111	0.4138	97.4592
Dim.32	0.1991	0.3904	97.8496
Dim.33	0.1749	0.343	98.1926
Dim.34	0.1491	0.2923	98.4848
Dim.35	0.1416	0.2776	98.7624
Dim.36	0.1157	0.2269	98.9894
Dim.37	0.1022	0.2005	99.1898
Dim.38	0.0931	0.1826	99.3725
Dim.39	0.0771	0.1512	99.5237
Dim.40	0.0493	0.0967	99.6204
Dim.41	0.0473	0.0927	99.7131
Dim.42	0.0397	0.0779	99.791
Dim.43	0.0363	0.0713	99.8622
Dim.44	0.0236	0.0463	99.9085
Dim.45	0.0157	0.0307	99.9392
Dim.46	0.0137	0.0269	99.9661
Dim.47	0.0101	0.0198	99.9859
Dim.48	0.0038	0.0075	99.9933
Dim.49	0.0029	0.0058	99.9991
Dim.50	0.0004	0.0009	99.9999
Dim.51	0	0.0001	100

PCA

- Original variables plotted against PC1 and PC2
- Color - contributions of the variables to the PCs (loadings)
- Distance between variables and the origin - the quality of the representation by PCs



PCA

- **PC1 explained:**

  positively correlated with:

  - I8 - Economy: Agriculture (% of GVA);
  - I11 - Employment: Agriculture (% of employed);
  - I26 - Fertility rate, total (live births per woman);
  - I29 - Population age distribution (0-14 years, %);
  - I33 - Infant mortality rate (per 1000 live births).

  negatively correlated with:

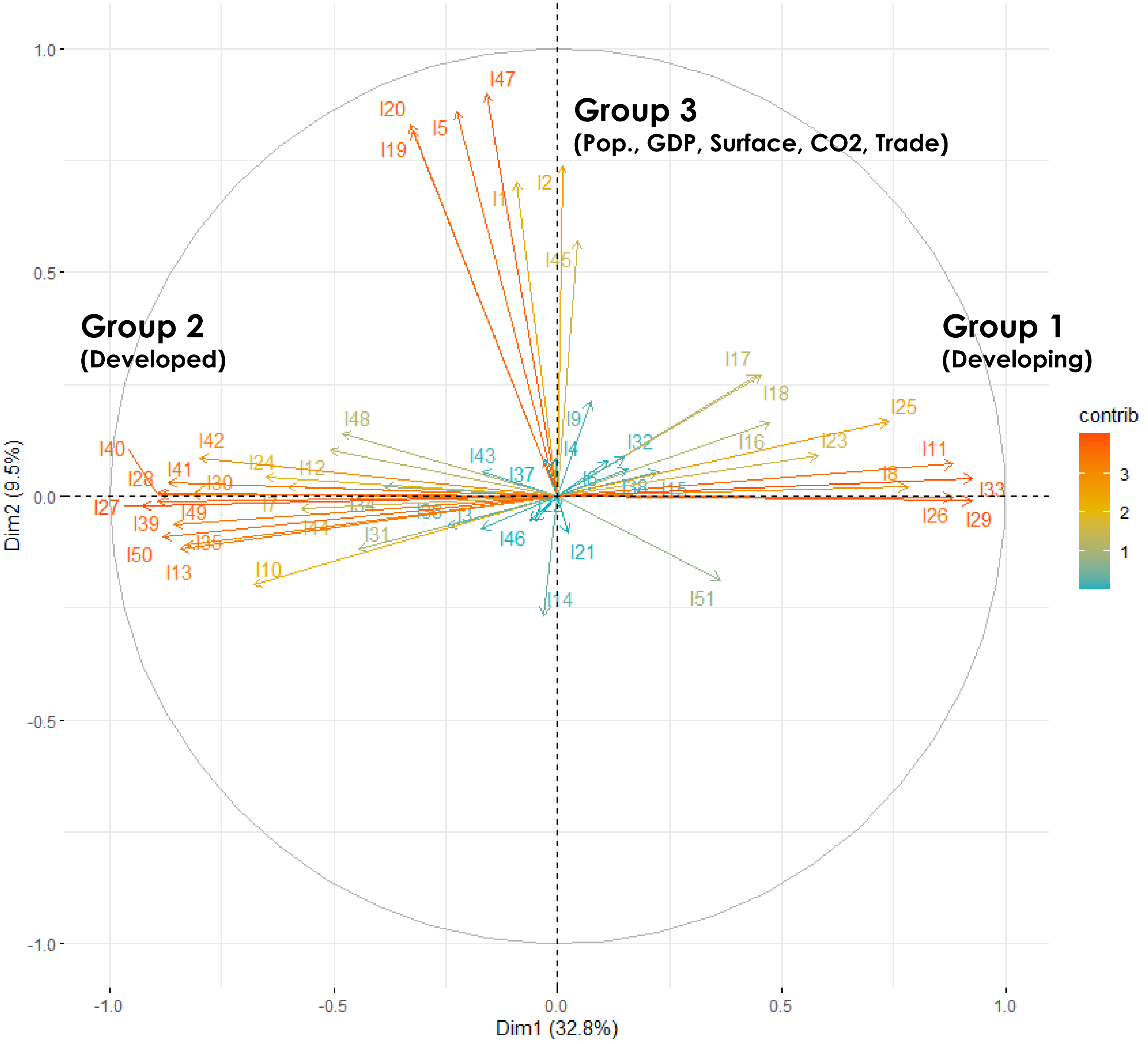
  - I49 - Pop.using improved drinking water (urban,%);
  - I30 - Population age distribution (60+ years,%);
  - I27 - Life expectancy at birth (females, years);
  - I13 - Employment: Services (% of employed).

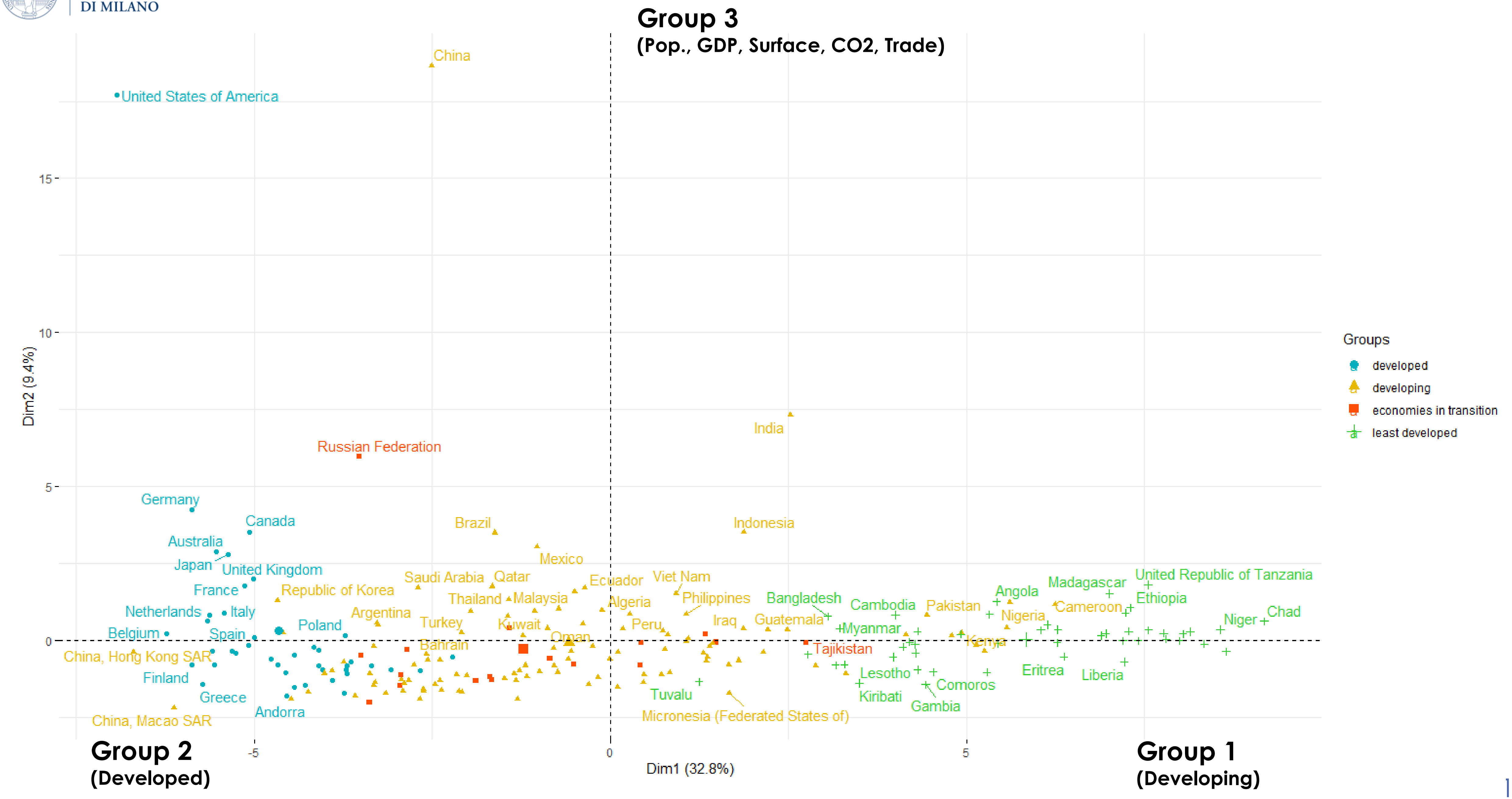
- **PC2 explained:**
  - I1 - Surface area (km2);
  - I2 - Population in thousands (2017);
  - I5 - GDP: Gross domestic product (million current US$);
  - I19 - International trade: Exports (million US$);
  - I20 - International trade: Imports million US$);
  - I47 - CO2 emission estimates (million tons/tons per capita).

- **PC3 explained:**
  - I4 - Sex ratio (m per 100 f, 2017);
  - I9 - Economy: Industry (% of GVA);
  - I16 - Labour force participation (male %);
  - I31 – International migrant stock (% of total pop.)

PCA

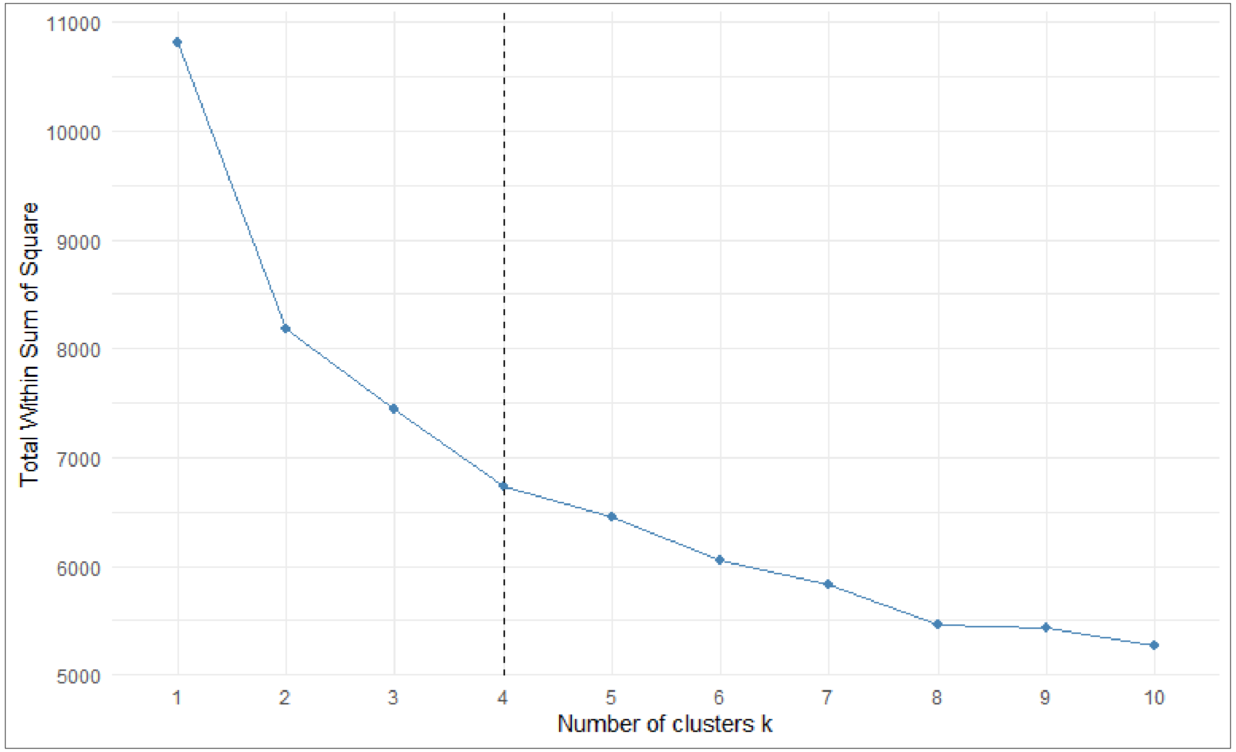
- Original variables plotted against PC1 and PC2
- Color - contributions of the variables to the PCs (loadings)
- Distance between variables and the origin - the quality of the representation by PCs
- I10, I28, I30, I35 are positively correlated



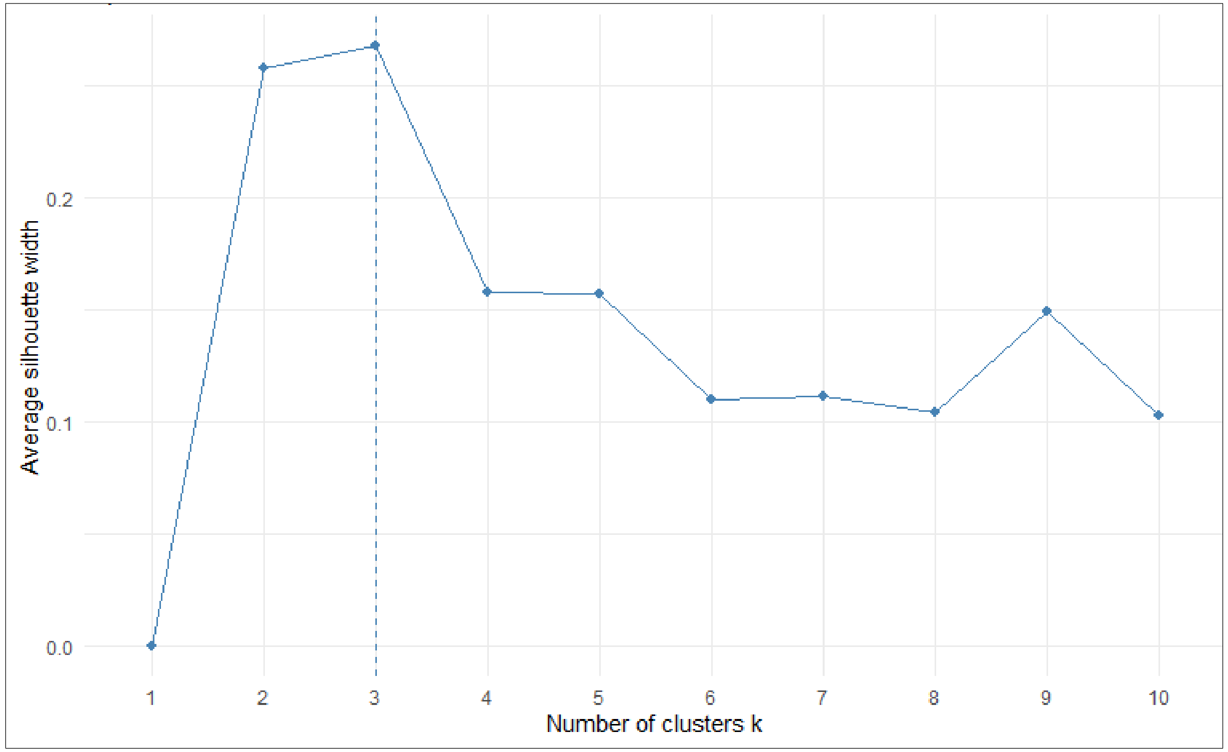

Group 3
(Pop., GDP, Surface, CO2, Trade)
Group 2
(Developed)
Group 1
(Developing)
Dim1 (32.8%)
Dim2 (9.4%)
Groups
developed
developing
economies in transition
least developed
United States of America
China
Russian Federation
India
Indonesia
Germany
Canada
Australia
Japan
United Kingdom
France
Netherlands
Italy
Belgium
Spain
Poland
China, Hong Kong SAR
Finland
Greece
Andorra
China, Macao SAR
Brazil
Mexico
Saudi Arabia
Qatar
Republic of Korea
Thailand
Malaysia
Argentina
Turkey
Kuwait
Bahrain
Oman
Ecuador
Algeria
Peru
Viet Nam
Philippines
Iraq
Guatemala
Bangladesh
Cambodia
Pakistan
Myanmar
Tajikistan
Tuvalu
Micronesia (Federated States of)
Lesotho
Kiribati
Comoros
Gambia
Angola
Nigeria
Kenya
Eritrea
Madagascar
Cameroon
Liberia
United Republic of Tanzania
Ethiopia
Niger
Chad

K-Means Clustering

Optimal number of cluster k = 4. Euclidean distance applied

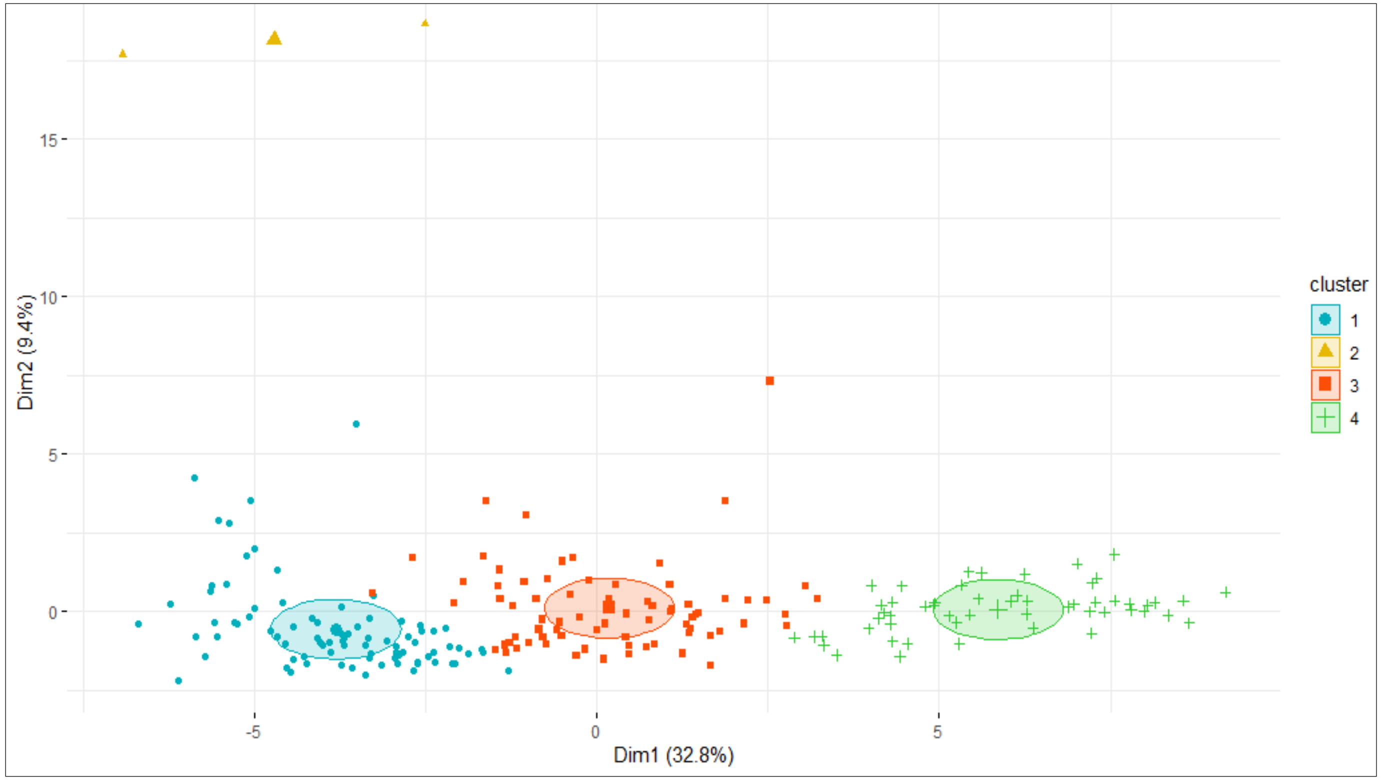


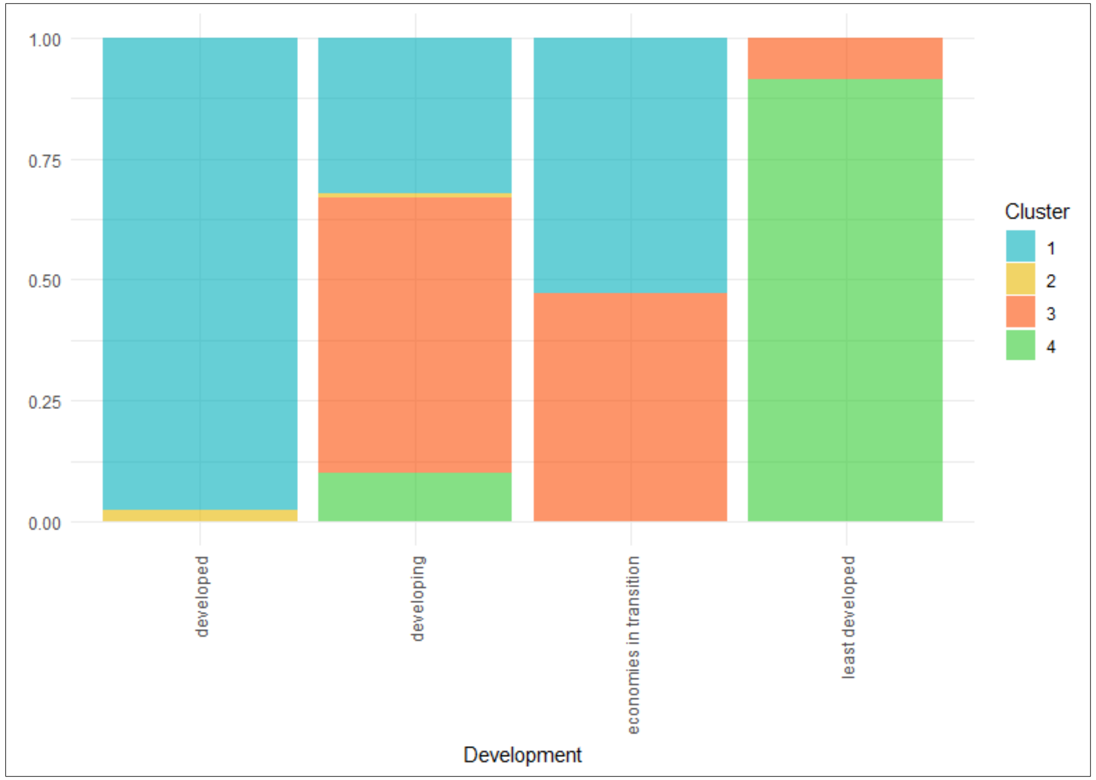
Elbow method



Silhoette method

K-Means Clustering



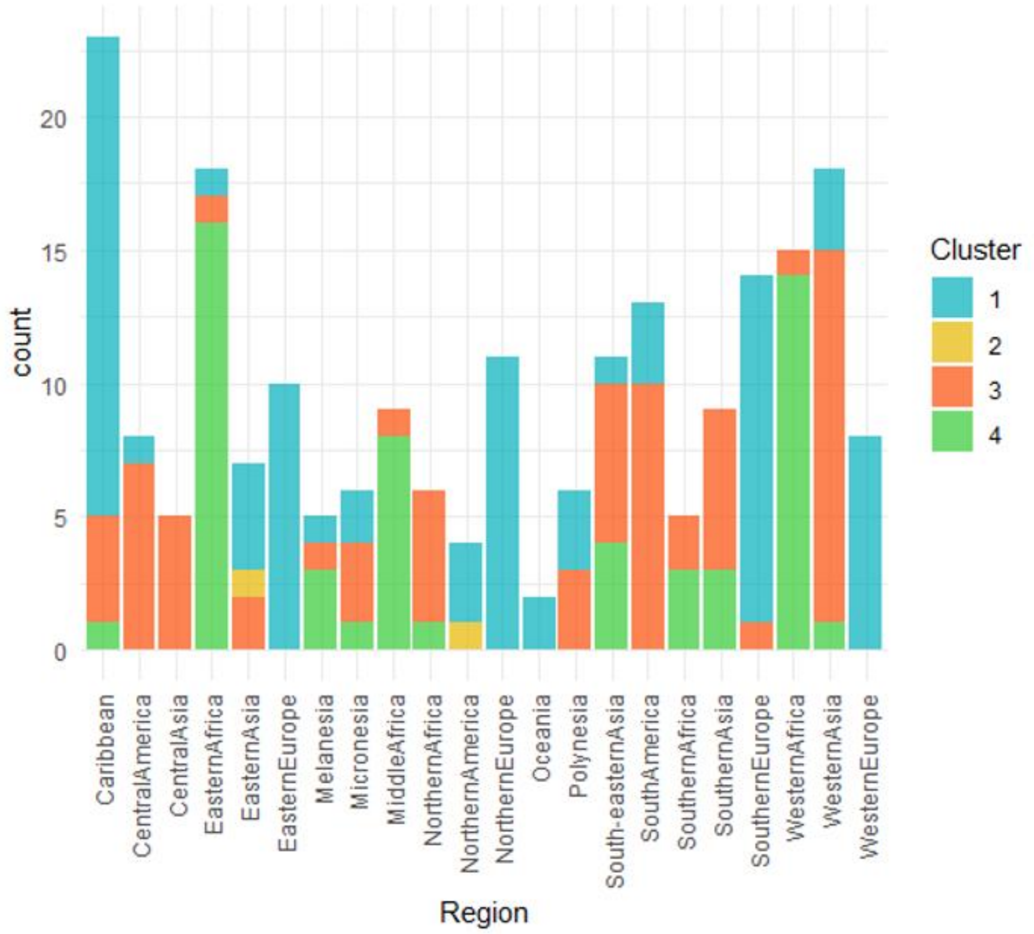


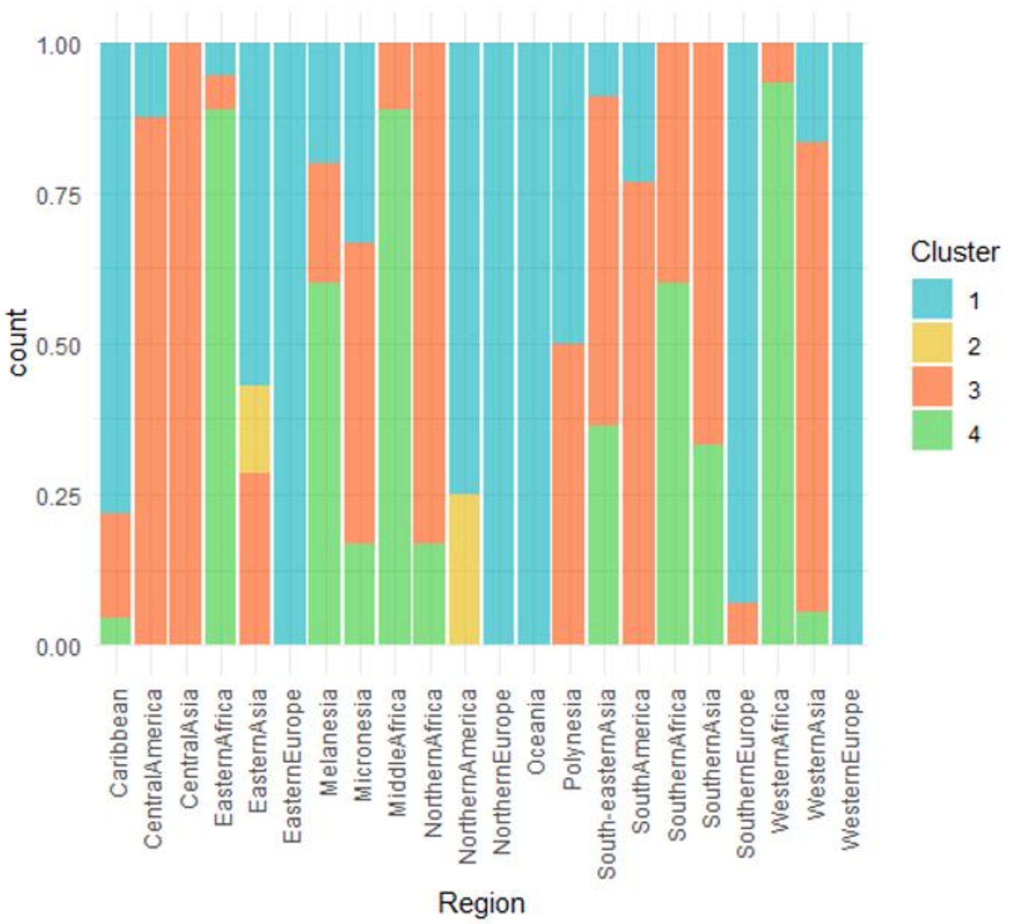
The result of K-means clustering of the countries. K = 4

Association between clusters and the countries socio-economic development classes

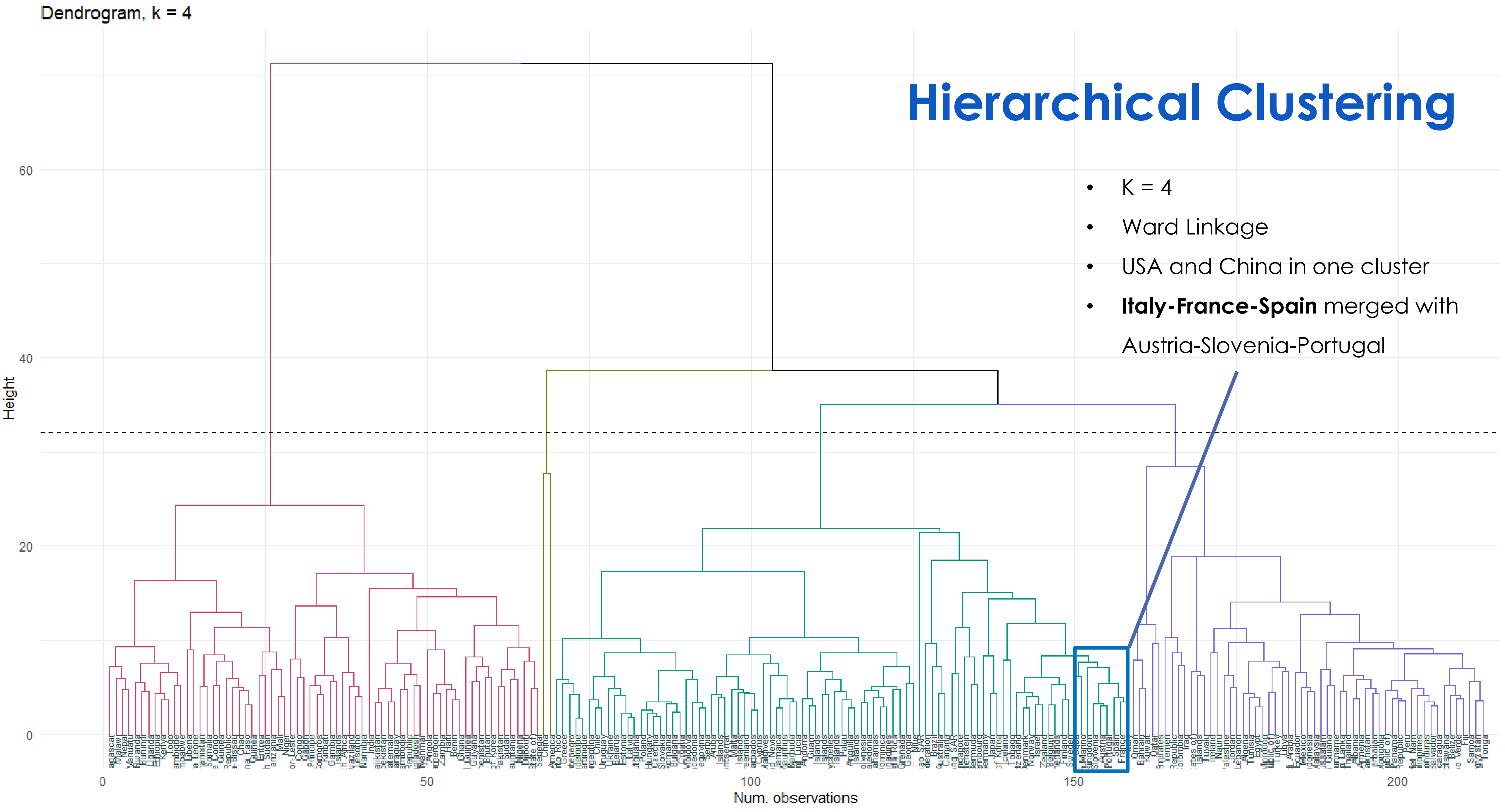
UNIVERSITÀ
DEGLI STUDI
DI MILANO

K-Means Clustering





Association between clusters and the countries region

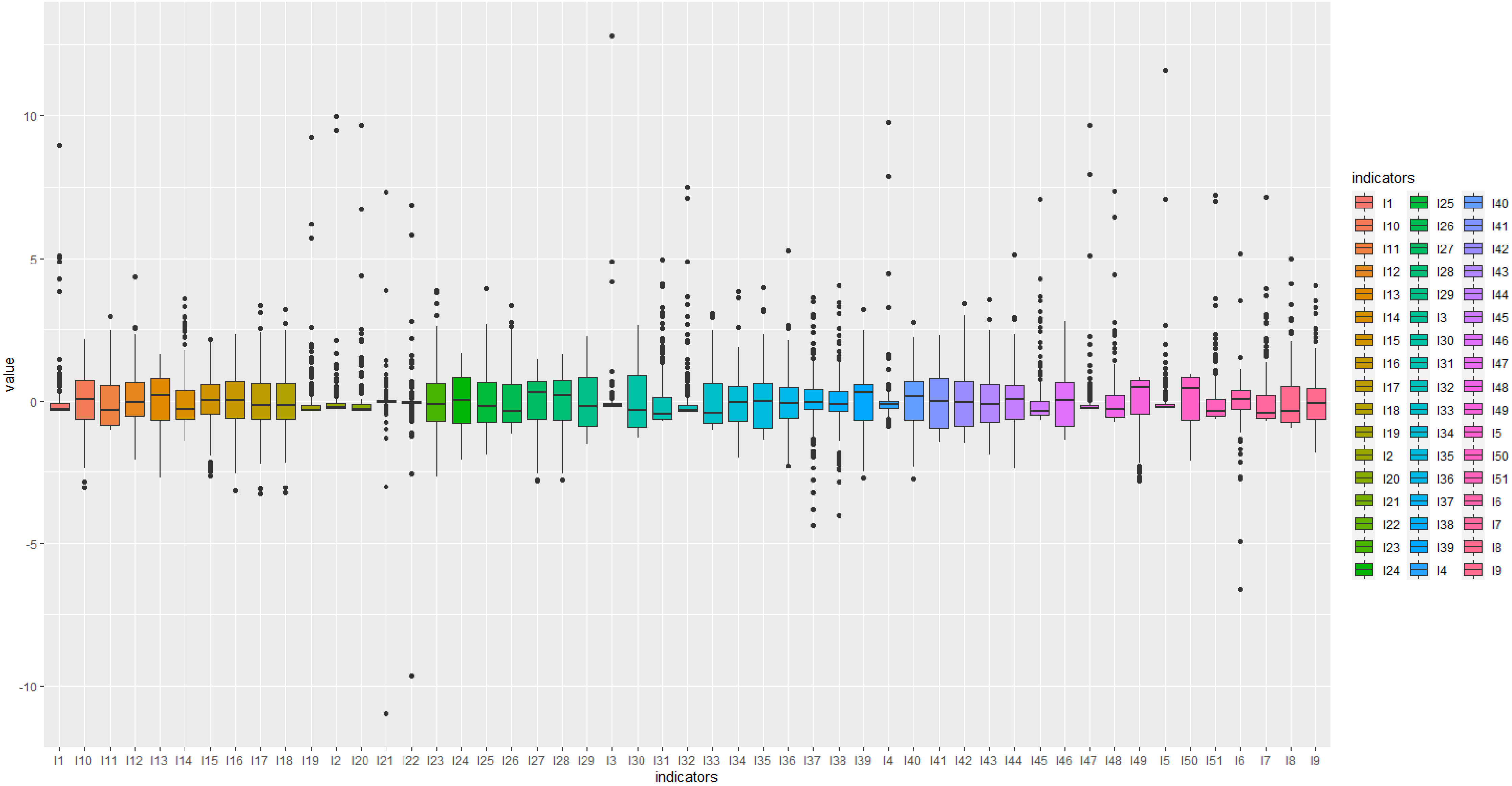


Hierarchical Clustering

- K = 4
- Ward Linkage
- USA and China in one cluster
- **Italy-France-Spain** merged with Austria-Slovenia-Portugal

Outliers Detection

Univariate method: **Interquartile range**



Outliers Detection

- Multivariate methods:
  - **Isolation Forest** (threshold = 0.60)
  - **PCA inversion** (12 PCs, MSE between original data and reconstructed, threshold = 0.30)

Country	Region	Development	Iforest anomaly score	PCA reconstruction loss
Russian Federation	EasternEurope	economies in transition	0.6182	1.2975
Colombia	SouthAmerica	developing	0.6050	1.0553
China, Macao SAR	EasternAsia	developing	0.6171	0.7633
Angola	MiddleAfrica	least developed	0.6005	0.5935
Syrian Arab Republic	WesternAsia	developing	0.6112	0.5903
Timor-Leste	South-easternAsia	least developed	0.6103	0.5197
United Arab Emirates	WesternAsia	developing	0.6061	0.4524
Bermuda	NorthernAmerica	developing	0.6063	0.4169
Cuba	Caribbean	developing	0.6069	0.4000
Qatar	WesternAsia	developing	0.6188	0.3921
Equatorial Guinea	MiddleAfrica	developing	0.6119	0.3727
Nigeria	WesternAfrica	developing	0.6041	0.3275
Greece	SouthernEurope	developed	0.6267	0.3211
Madagascar	EasternAfrica	least developed	0.6068	0.3072

Outliers identified with both Isolation Forest and PCA inversion methods

SUPERVISED LEARNING:

Random Forest, Logistic Regression, K-NN, Neural Network

Target variable - Development level:

developed

developing

economies in transition

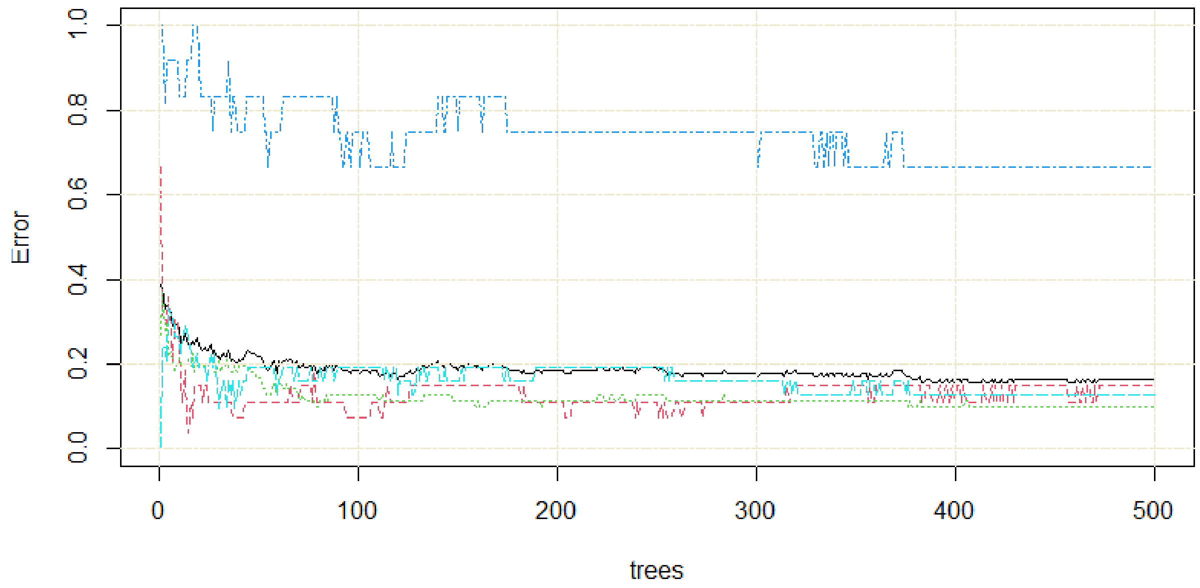
least developed

Random Forest

Train/Test split (p = 0.65)

RF Model 1:

- 139 countries – train set
- 74 – test set
- 51 standardized indicators
- 500 trees
- 7 random predictors
- Bootstrap resampling



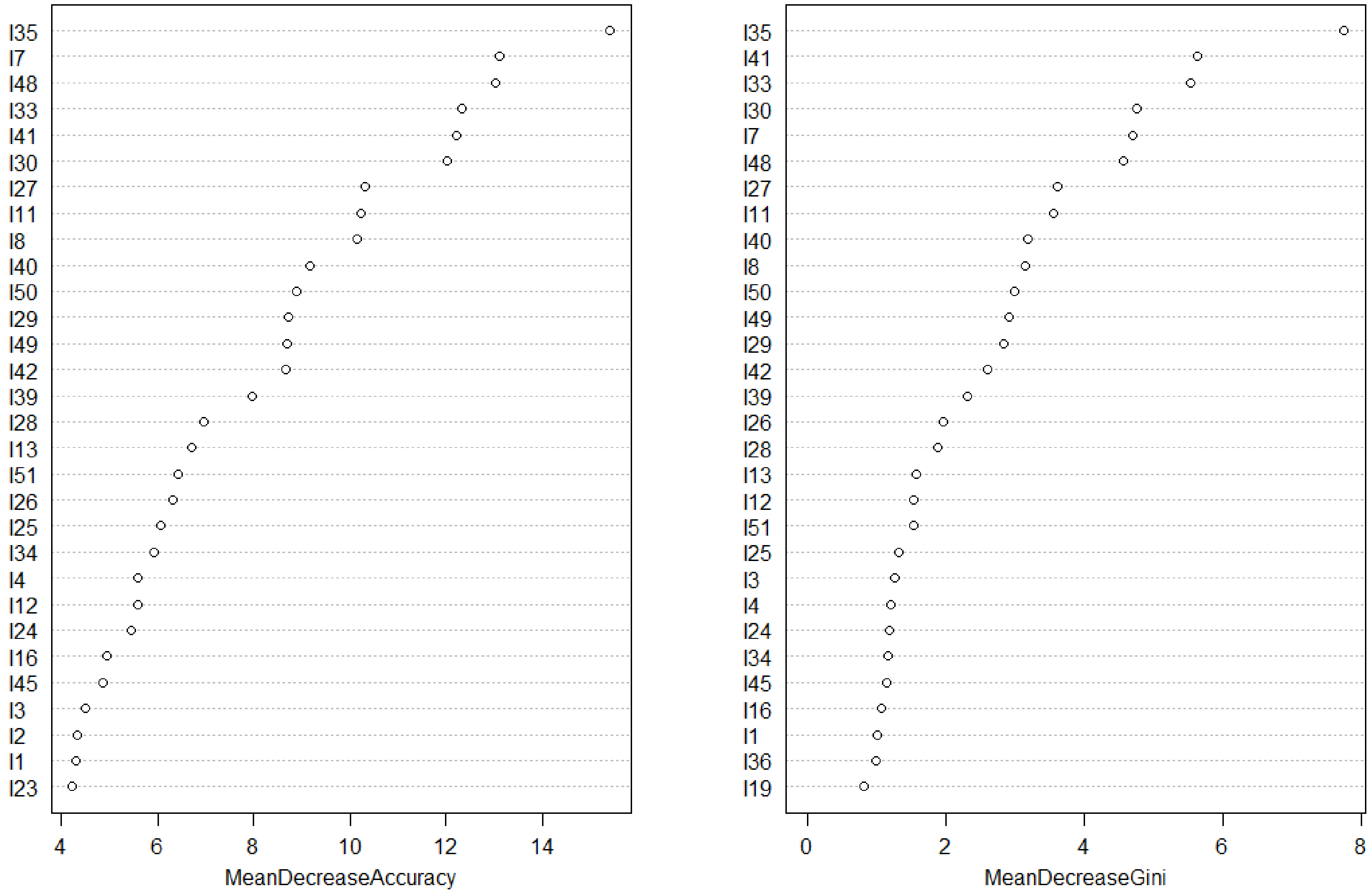

```
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 7

        OOB estimate of  error rate: 18.44%
Confusion matrix:
                        developed developing economies in transition least developed class.error
developed                      22          5                        0               0   0.1851852
developing                      4         62                        1               4   0.1267606
economies in transition         1          6                        4               1   0.6666667
least developed                 0          4                        0              27   0.1290323
```

Random Forest

```
Overall Statistics

               Accuracy : 0.8219
                 95% CI : (0.7147, 0.9016)
    No Information Rate : 0.5205
    P-Value [Acc > NIR] : 7.658e-08

                  Kappa : 0.7306

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: developed Class: developing Class: economies in transition Class: least developed
Sensitivity                    0.8571            0.7895                        0.60000                0.9375
Specificity                    0.9322            0.9143                        0.97059                0.9298
Pos Pred Value                 0.7500            0.9091                        0.60000                0.7895
Neg Pred Value                 0.9649            0.8000                        0.97059                0.9815
Prevalence                     0.1918            0.5205                        0.06849                0.2192
Detection Rate                 0.1644            0.4110                        0.04110                0.2055
Detection Prevalence           0.2192            0.4521                        0.06849                0.2603
Balanced Accuracy              0.8947            0.8519                        0.78529                0.9337
```

Performance on the test set

Random Forest



Importance of the indicators in Random Forest model with the original parameters

RF Model 2 (16.82% OOB):

- 214 countries
- 51 indicators

RF Model 3 (23.36% OOB) – multicollinearity

- 214 countries
- 12 PCs

RF Model 3 (17.5% OOB) – outliers

- 200 countries
- 51 indicators

Random Forest and other models

- **Multinomial Logistic Regression**: default parameters
- **K-NN:** Euclidean Distance, k = 8
- **Neural Network:**
  - softmax activation function,
  - two hidden layers (10, 4)
  - Resilient backpropagation with weight backtracking

Model	Train set accuracy	Test set accuracy
Random Forest	0.82	0.82
Multinomial Logistic Regression	1.00	0.64
K-NN (k = 8)	0.78	0.81
Neural Network	0.96	0.66

Challenges: small dataset

Challenges:

- Few observations in a high-dimensional space
- Overfitting: a low bias and a high variance models
- Underfitting: a high bias and a low models
- Low prediction power
- Imbalanced dataset

Techniques to improve the modelling:

- Relevant features selection
- A simple model with a small number of parameters
- Outliers removal
- Augmenting the dataset with synthetic samples
- Adding information from other sources.

References

1. Kaggle: UN countries dataset 2017. https://www.kaggle.com/sudalairajkumar/undata-country-profiles. [Online; Accessed: 2021-11-15].
2. UNdata, international statistical database. http://data.un.org/Host.aspx?Content=About. [Online; Accessed: 2021-11-15].
3. UN ESCAP, UN ECA, UN ECE, UN ESCWA, UN ECLAC, et al. World economic situation and prospects 2017. 2017.
4. Stefan Fritsch, Frauke Guenther, andMaintainer Frauke Guenther. Package ‘neuralnet’. Training of Neural Networks, 2019.
5. Paulene Govender and Venkataraman Sivakumar. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmospheric Pollution Research, 11(1):40–56, 2020.
6. Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A:Mathematical, Physical and Engineering Sciences, 374(2065):20150202, 2016.
7. Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(1):1–39, 2012.
8. Victor Francisco Rodriguez-Galiano, Bardan Ghimire, John Rogan,Mario Chica-Olmo, and Juan Pedro Rigol-Sanchez. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67:93–104, 2012.