SAPIENZA
UNIVERSITÀ DI ROMA

Adv. in Data Analysis & Statistical Modelling
24/25

STATISTICAL MODEL-BASED COMPOSITE INDICATORS FOR TRACKING COHERENT POLICY CONCLUSIONS

ALEYNA KANDEMIR FRANCESCO NATALI
ANDI PRASETYA LEONARDO AGATE

OVERVIEW OF THE PRESENTATION

SAPIENZA
UNIVERSITÀ DI ROMA

01 THEORETICAL FOUNDATION

Model -Based Composite Indicator (CI)

Definition

CI is formed when manifest indicators (MIs) are compiled into a single index, based on an underlying model of the multi-dimensional concept being measured. CI is usually computed starting from fewer manifest variables (e.g., MPI and HDI). In contrast, it is more and more necessary to have tools able to handle a large number of manifest indicators.

Pros and Cons of CIs

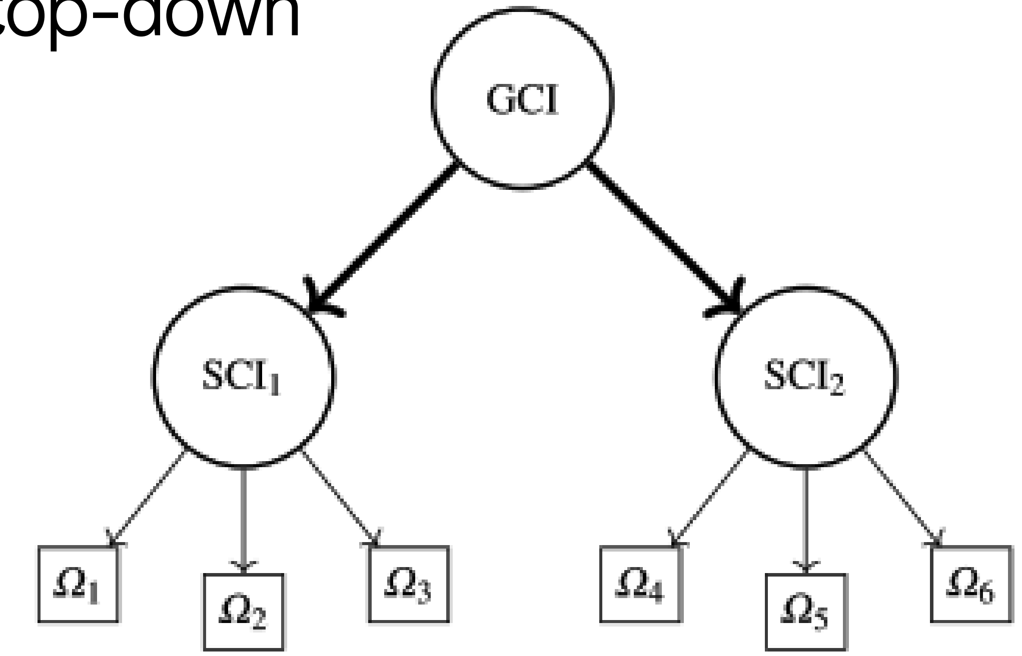
- **Pros:**
  - Simplify complex multidimensional data.
  - Support benchmarking and policy analysis.
- **Cons:**
  - Sensitive to subjective choices (weights, normalization).
  - Risk of oversimplification.

Measurement Model in Model-Based CI

Two different approaches might be distinguished with respect to the nature of the relationships between MIs and latent constructs (e.g., SCIs and GCIs) that formally describe the measurement model and thus define the direction of these relationships.

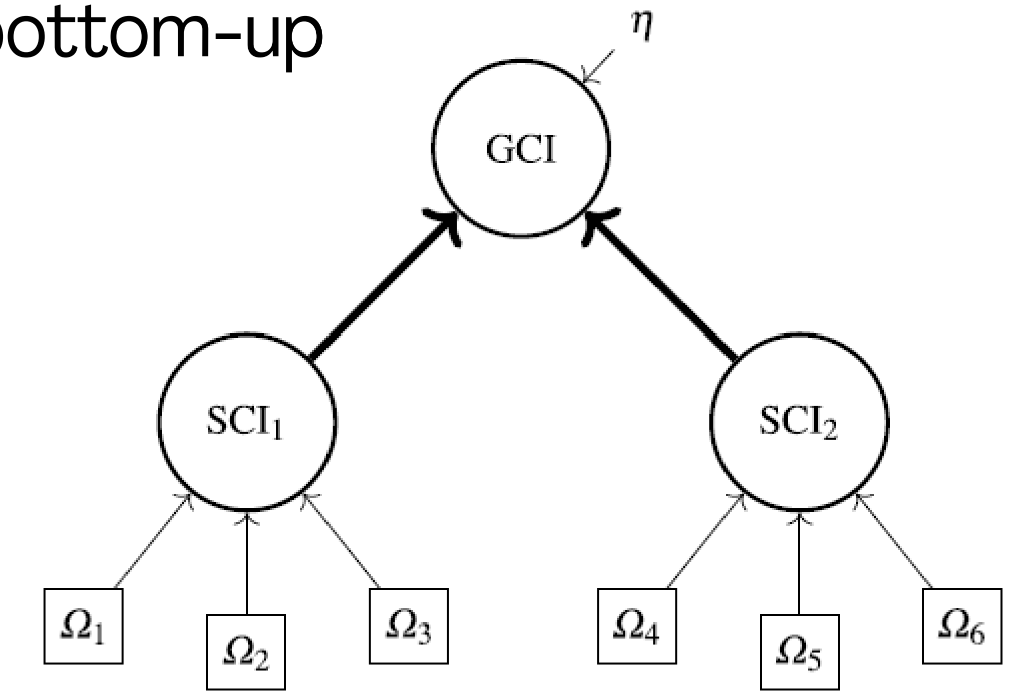
Reflective Construct

- Latent variables cause observed indicators.
- CI → MI
- top-down



Formative Construct

- Observed indicators form latent variables.
- MI → CI
- bottom-up



SAPIENZA
UNIVERSITÀ DI ROMA

Correlation Matrix of the Model

Analyzes relationships between indicators and latent variables using:

$$\Sigma = \Phi \Psi \Phi^T + \Theta$$

Where:

- Σ(Sigma): Covariance matrix of indicators.
- Φ(Phi): Loading matrix.
- Ψ(Psi): Covariance matrix of latent variables.
- Θ(Theta): Residual matrix.

Goodness of Fit (GoF)

Measures how well the model explains the data:

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

Where:

- SSresidual: Unexplained variance.
- SStotal: Total variance.

A high value indicates that the composite indicator explains most of the variance in the observed data

SAPIENZA
UNIVERSITÀ DI ROMA

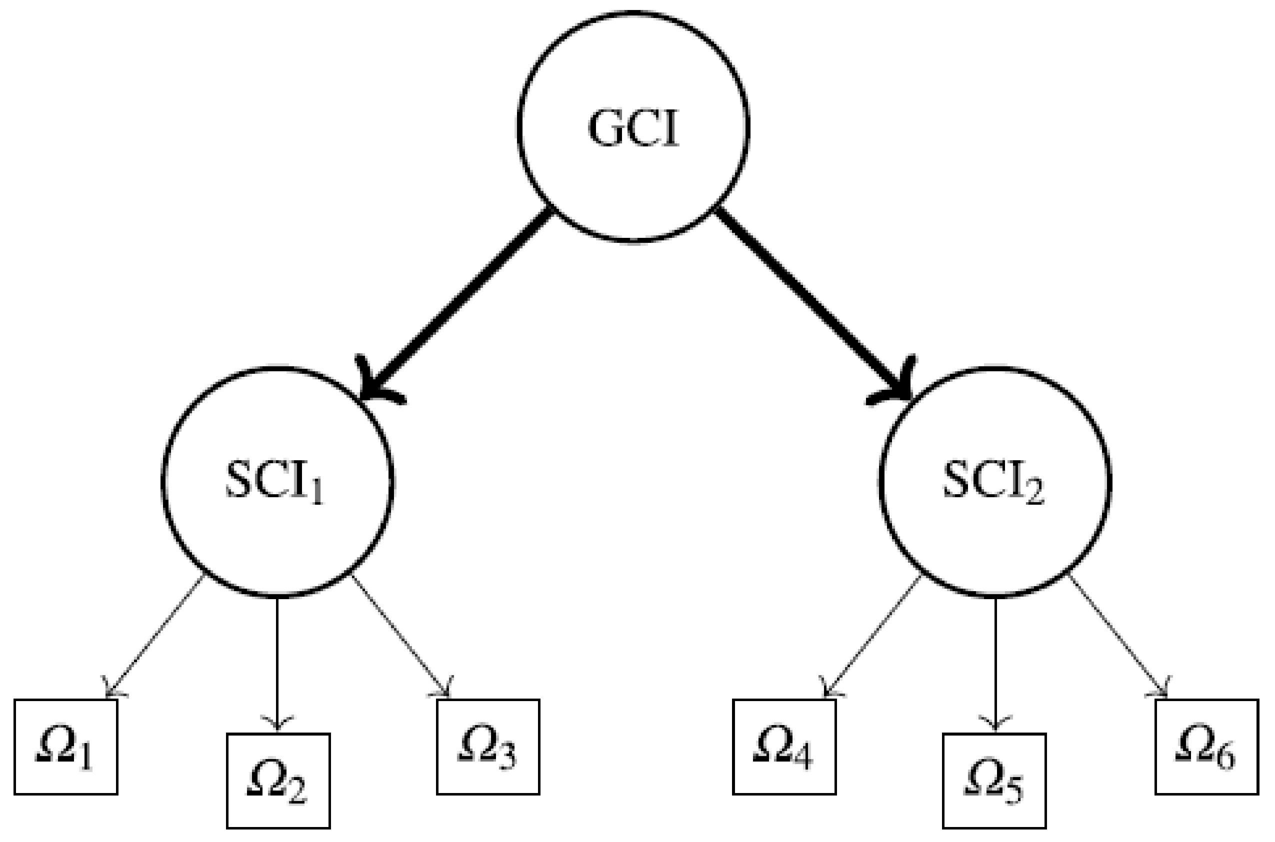
Model Selection

Confirmatory Model

A theoretical framework allows researchers to hypothesize:

- The MIs used to define SCIs.
- The SCIs necessary for the GCI.
- The relationships (reflective or formative) describing the phenomenon.

The confirmatory model is fully defined and specified, enabling predictions based on the model. Empirical observation of the MIs tests whether the phenomenon aligns with hypotheses. In this approach, MIs are preselected, relationships between MIs and CIs (SCI/GCI) are predefined, and the relationship types are established.



Confermatory and Reflective model-based CI.

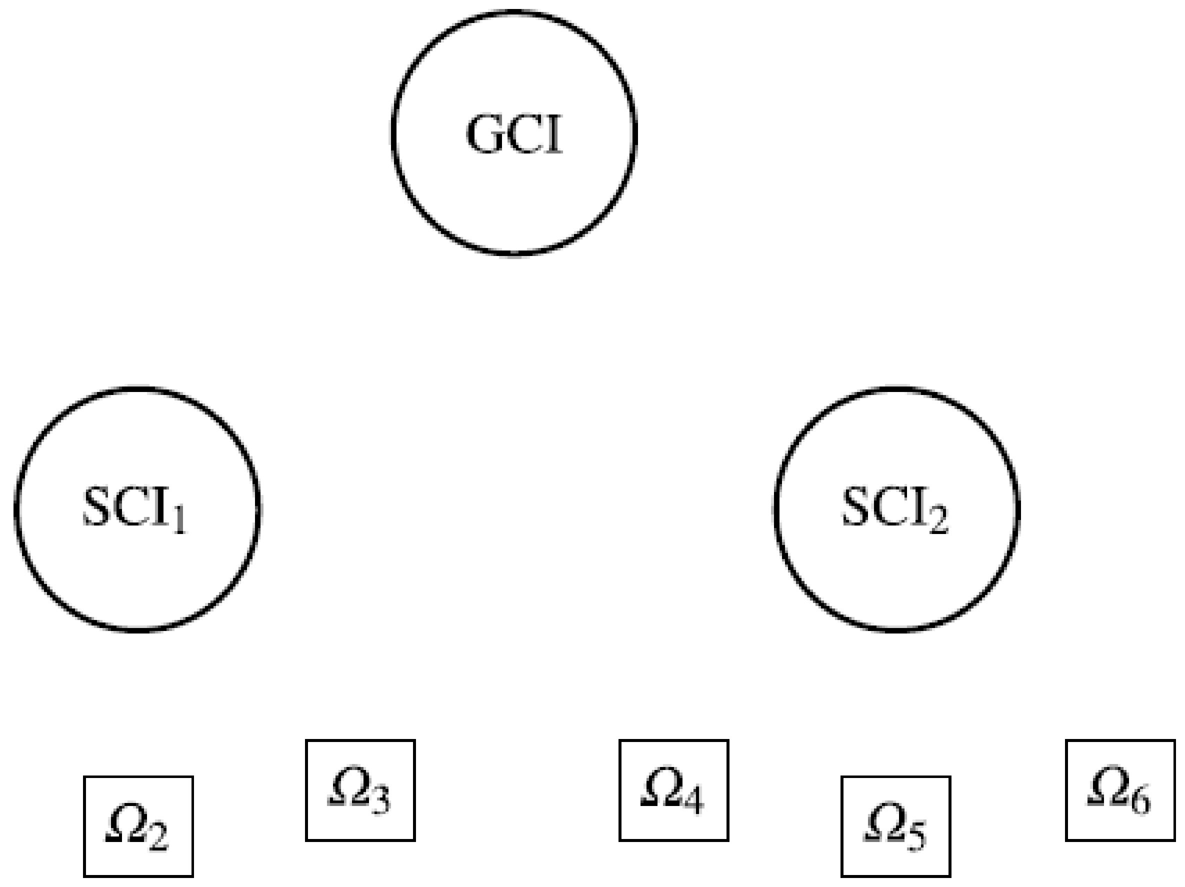
Model Selection

Exploratory Model

When a theoretical framework is unavailable or unconfirmed, an exploratory approach is used. In this approach:

- MIs for SCIs are unidentified, so a broader range is considered.
- SCIs for GCI are not predetermined.
- Relationship typologies describing the phenomenon are unknown.

Theory-based approaches differ from model-based and data-driven methods, which follow an exploratory path to identify the optimal synthesis of MIs. The contrast between confirmatory and exploratory approaches is often explained through Factor Analysis, showcasing the differences between CFA and EFA.



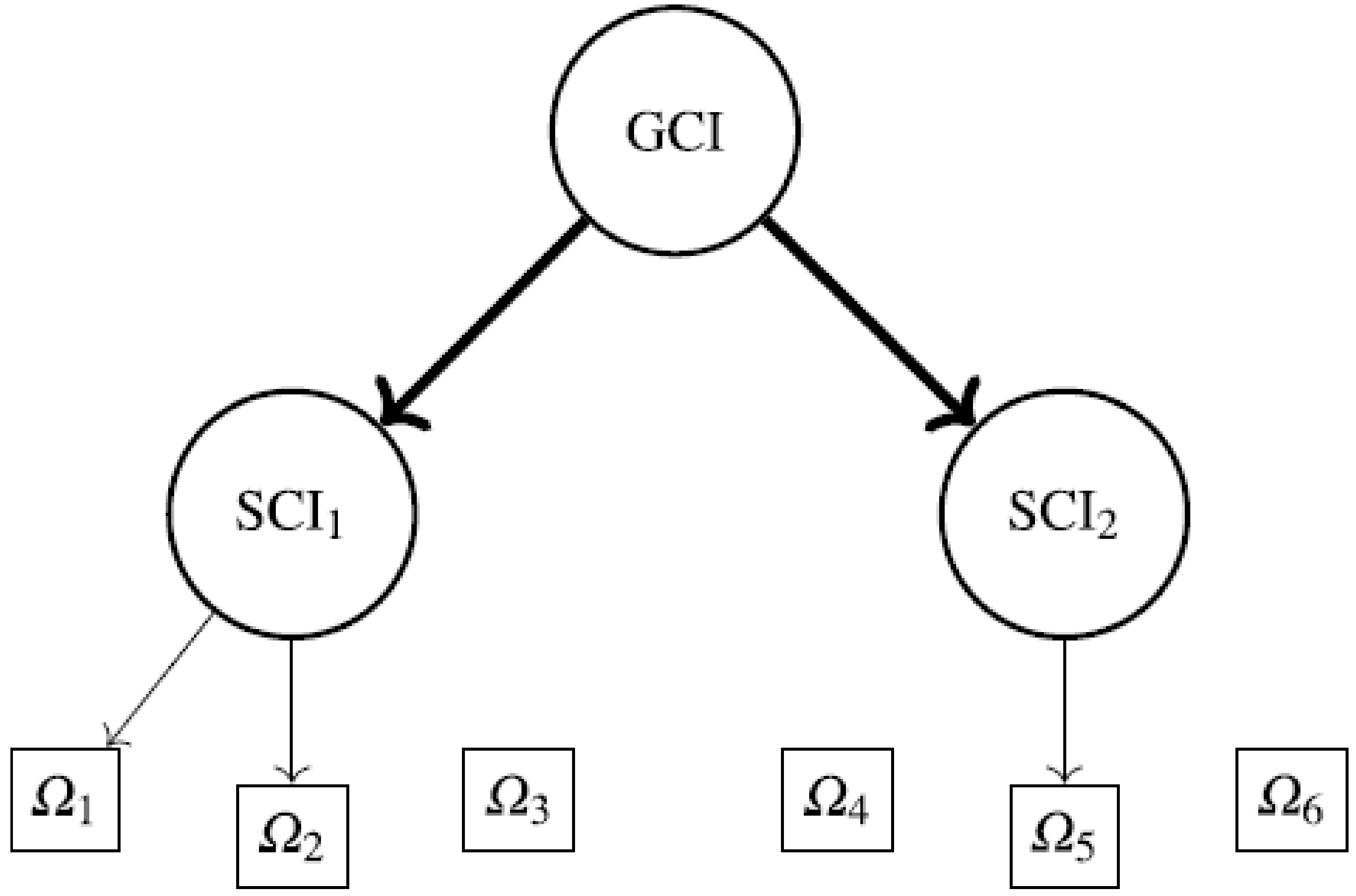
Exploratory model-based CI.

Sapienza Università di Roma

Model Selection

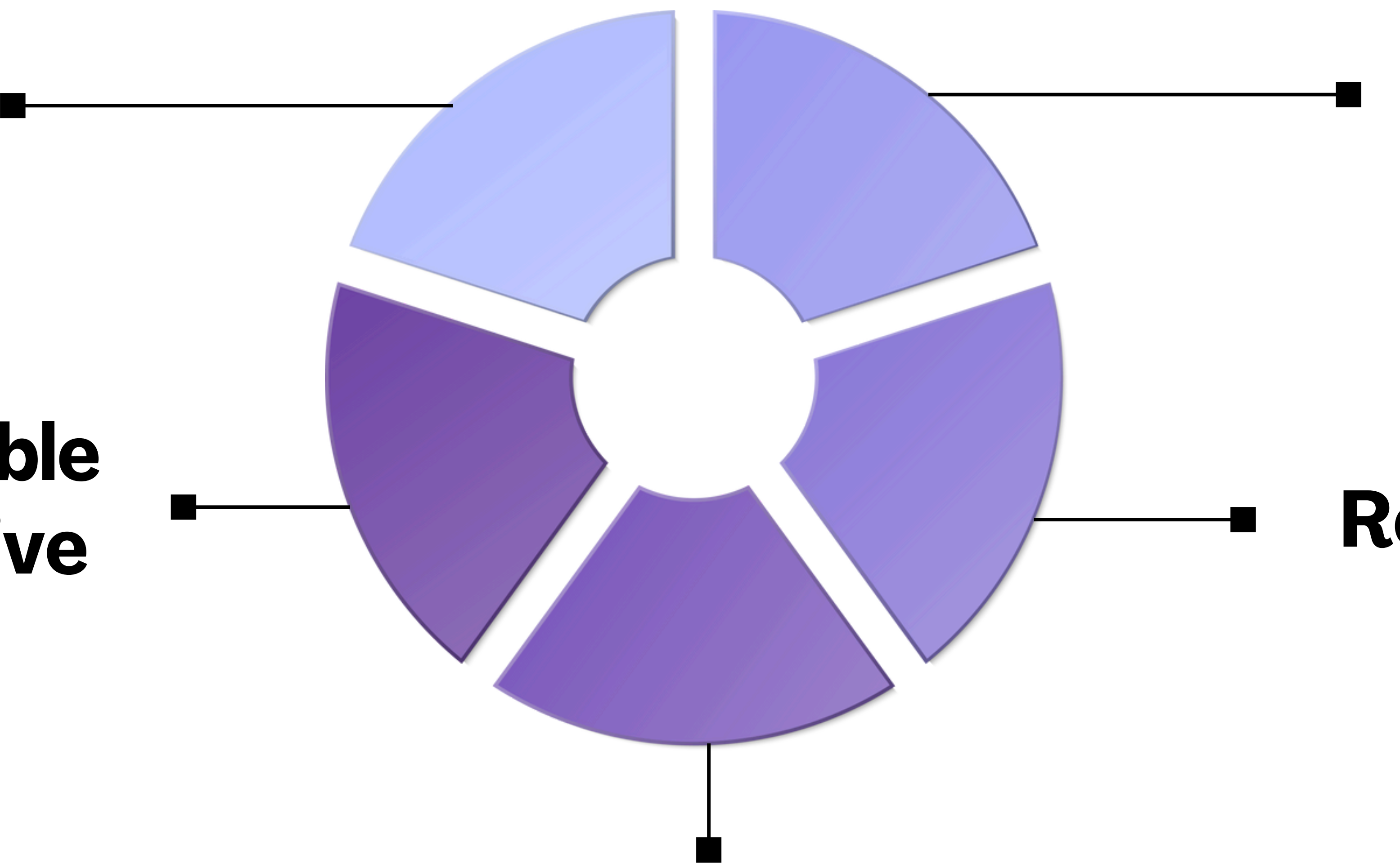
Mixed Model

When the theoretical framework is partially available or confirmed, a mixed approach—combining confirmatory and exploratory methods—can be used to avoid suboptimal model adjustments. Model selection requires both the researcher's expertise and a systematic focus on the most relevant indicators and relationships.



Mixed Confermatory/Exploratory model-based CI.

Properties of Model-Based CI



Scale-Invariant Model-Based CI and Data Transformation

Purpose of Normalization:

- Eliminate the influence of measurement units from Manifest Indicators (MIs).
- Allow comparison and combination of MIs into Specific Composite Indicators (SCIs) and the General Composite Indicator (GCI).

Normalization Methods (Linear Transformations):

Standardization	Generic element of $\mathbf{Z}$	Characteristics of $\mathbf{Z}$
$\mathbf{Z} = \mathbf{J}\mathbf{X}(\mathrm{dg}(\Sigma))^{-\frac{1}{2}}$ with $\mathbf{J} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}'_n$	$z_{ij} = \frac{x_{ij}-\mu_j}{\sigma_j}$	Mean zero and unitary variance
Min–max normalization (unit-based)		
$\mathbf{Z} = \frac{\mathbf{X}-\mathbf{1}_n \min \mathbf{X}.}{\mathbf{1}_n \max \mathbf{X}-\mathbf{1}_n \min \mathbf{X}}$	$z_{ij} = \frac{x_{ij}-\min(\mathbf{x}_j)}{\max(\mathbf{x}_j)-\min(\mathbf{x}_j)}$	Values are between 0 and 1
Normalized dispersion		
$\mathbf{Z} = \mathbf{J}\mathbf{X}\mathrm{diag}(\mu_{\mathbf{X}})^{-1}$ with $\mathbf{J} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}'_n$	$z_{ij} = (x_{ij} - \mu_j)/\mu_j$	Mean zero and standard deviation equal to the coefficient of variation

Non-compensable and Non-negative Model-Based CI

Non-Compensability:

- Positive relations among MIs are not compensated by negative ones. All MIs are concordantly related to the CI, where increments in CI correspond to increments in MIs, and vice-versa.

Ensuring Non-Compensability:

- Constrain all weights to be strictly positive.
- Reverse all MIs with negative weights.
- Positive loadings in FA ensure concordance between rankings of SCIs and the GCI.

Importance of Non-Negative Weights:

- Weights must be positive to ensure proper interpretation. Negative weights imply the MI reflects negatively on the latent construct and must be reversed.

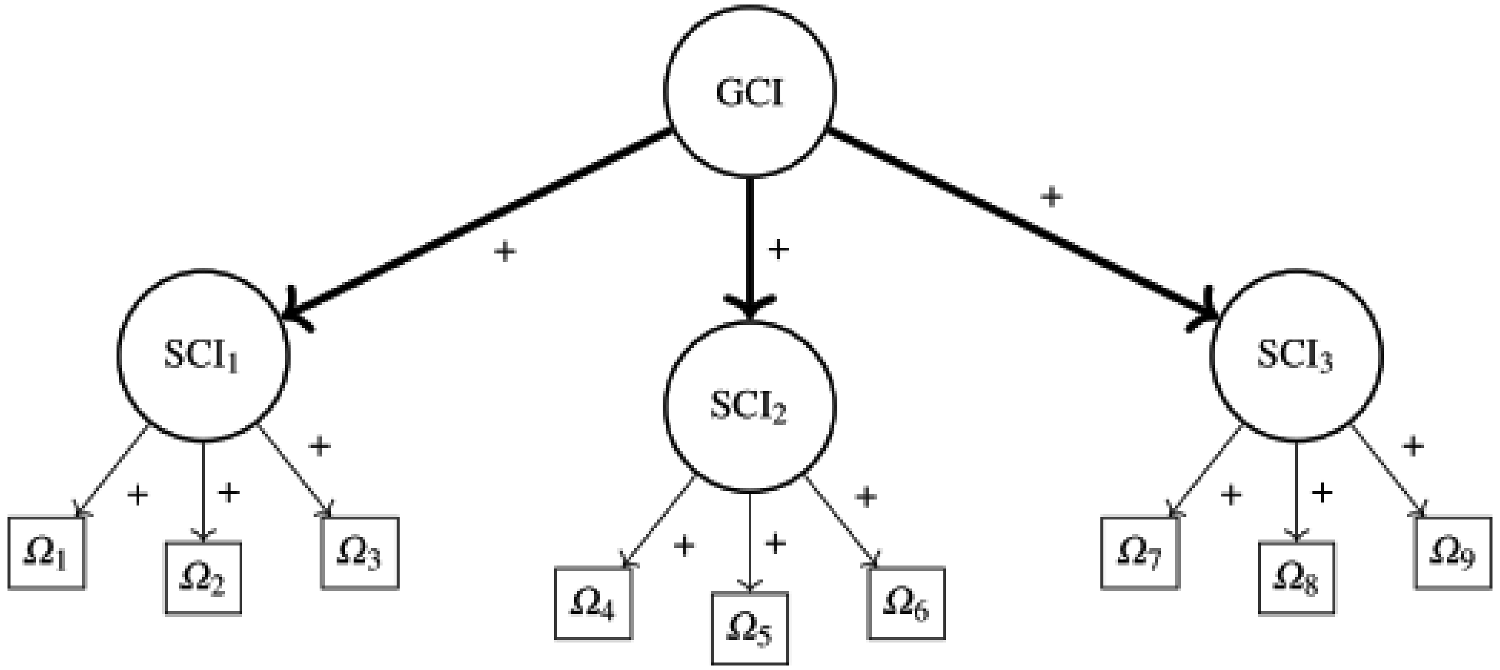
Polarity

Importance of Polarity:

- Determines the correlation structure among MIs and between MIs and SCIs.
- Ensures the composite indicator is non-compensable by clustering MIs with positive correlations.

Changing Polarity:

For example, we use normalization with the Min-max method in all MI and we use a factor analysis model, we can change the polarity of the “negative” MI using: $x_{ij} = 1 - x_{ij}$.



SAPIENZA
UNIVERSITÀ DI ROMA

Reliability of CI

The reliability of a CI is the global consistency of MIs based on the correlations between different MIs related to the same CI (latent construct). In the model-based CI, the correlation matrix is key to ensuring a comprehensive representation of how MIs relate to each other and their latent constructs.

$$\Sigma_{\mathbf{X}} = \frac{1}{n}\mathbf{X}'\mathbf{J}\mathbf{X} = \mathbf{B}\mathbf{V}\left[\mathbf{c}\left(\frac{1}{n}\mathbf{g}'\mathbf{J}\mathbf{g}\right)\mathbf{c}' - \frac{1}{n}\mathbf{E}'_{\mathbf{Y}}\mathbf{E}_{\mathbf{Y}}\right]\mathbf{V}'\mathbf{B} + \frac{1}{n}\mathbf{E}'_{\mathbf{X}}\mathbf{E}_{\mathbf{X}} = \mathbf{B}\mathbf{V}\Sigma_{\mathbf{Y}}\mathbf{V}'\mathbf{B} + \Psi_{\mathbf{X}}$$

where:

$$\Sigma_{\mathbf{Y}} = \mathbf{c}\mathbf{c}' + \Psi_{\mathbf{Y}}$$

with:

J is an idempotent centering matrix

$\mathbf{g} \sim N(0, 1)$

$\mathbf{E}_{\mathbf{Y}} \sim N_H(\mathbf{0}, \Psi_{\mathbf{Y}})$

$\Sigma_{\mathbf{E}_{\mathbf{Y}}} = \Psi_{\mathbf{Y}}$ is the diagonal positive definite variance-covariance matrix of the error of SCIs

$\mathbf{E}_{\mathbf{X}} \sim N_J(\mathbf{0}, \Psi_{\mathbf{X}})$, where $\Sigma_{\mathbf{E}_{\mathbf{X}}} = \Psi_{\mathbf{X}}$

X is matrix of MIs

V is membership matrix between MIs and SCIs

Unidimensionality and Presence of a General Factor

Definition of Unidimensionality:

- Evaluates the extent to which a single latent indicator (SCI) is measured by a cluster of MIs.

Kaiser Rule for Unidimensionality Check:

- The first eigenvalue of the correlation matrix for the cluster must be >1.
- All other eigenvalues must be <1.
- SCI is considered unidimensional if the variance of the second component is <1.

02 METHOD ALGORITHM

Exploratory Factor Analysis (EFA)

Definition

EFA is a dimensionality reduction technique used to identify linear combinations of variables in a dataset that are correlated with each other. These combinations are referred to as factors.

The steps involved in performing an Exploratory Factor Analysis (EFA) are as follows:

- Calculate the correlation matrix,
- Extract the factors,
- Rotate the factor loadings,
- Analyze the factor loadings.

Mathematical Model of EFA

In EFA, the observation vector xi (of size J×1, where J is the number of variables) can be approximated by a random factor vector yi (of size H×1, where H is the number of factors) as follows:

$$\mathbf{x}_i - \boldsymbol{\mu} = A\mathbf{y}_i + \mathbf{e}_i$$

where:

$x_i = N_J\left(\boldsymbol{\mu}, \sum_x\right)$ represents the observed variables

$\boldsymbol{e}_i = N_J\left(0_J, \psi\right)$ is the error vector, with zero mean and a diagonal covariance matrix

$\boldsymbol{\mu} = \left[\mu_1, \mu_2, \ldots, \mu_J\right]^T$ is the mean vector of the variables

A is the factor loading matrix

$y_i = N_H\left(0_H, I_H\right)$ is the factor scores vector, distributed normally with zero mean and identity covariance matrix

SAPIENZA
UNIVERSITÀ DI ROMA

Disjoint Factor Analysis (DFA)

Disjoint Factor Analysis (DFA) is a variant of Factor Analysis where the factors are assumed to explain distinct, non-overlapping subsets of observed variables.

$$\mathbf{X} = \mathbf{Y}\mathbf{V}'\mathbf{B} + \mathbf{E}_{\mathbf{X}}$$

Main considerations:

- Restriction: Each variable is associated with only one latent factor.
- Variance-Covariance Structure: Block diagonal format.
- Matrix Constraints: A=BV, where:

  V is the membership matrix that groups variables by factor.

  B is a diagonal matrix of factor loadings.

SAPIENZA
UNIVERSITÀ DI ROMA

Aggregating Factors into a General Index

Weighting Scheme:

Assign weights to factors based on

- **Subjective criteria**: Expert judgment or policy priorities.
- **Statistical methods**: Proportional to factor variance or eigenvalues from FA.
- **Equal weighting**: All factors contribute equally.

Compute the general index as a weighted sum of the factors:

$$\text{General Index} = \sum_{i=1}^{n} w_i \cdot F_i$$

03 APPLICATION ON MATLAB

Dataset Description

The data described socio-economic conditions in communities within the United States. The data combines from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR.

Communities and Crime

Donated on 7/12/2009

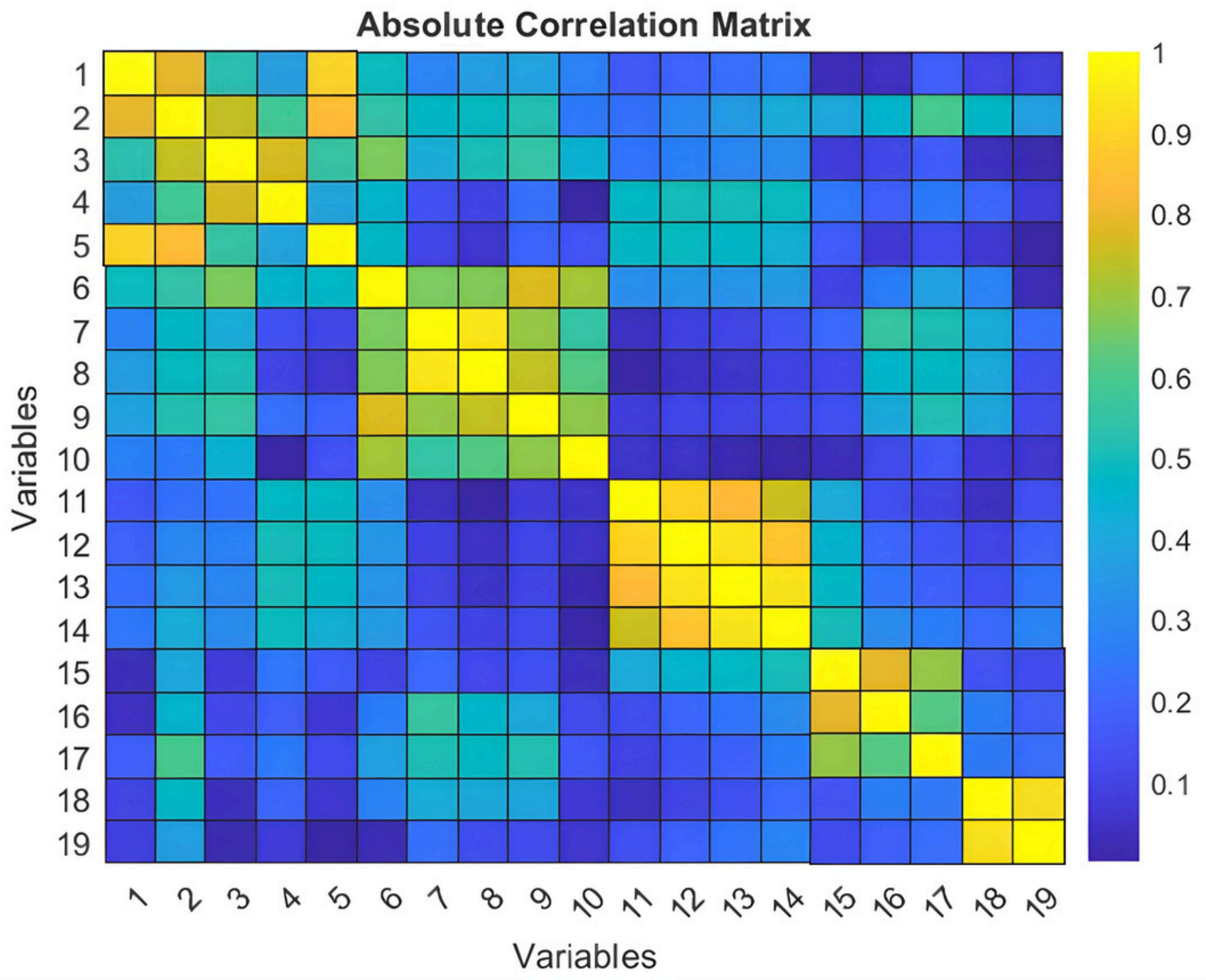
Communities within the United States. The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Social Science	Regression
Feature Type	**# Instances**	**# Features**
Real	1994	127

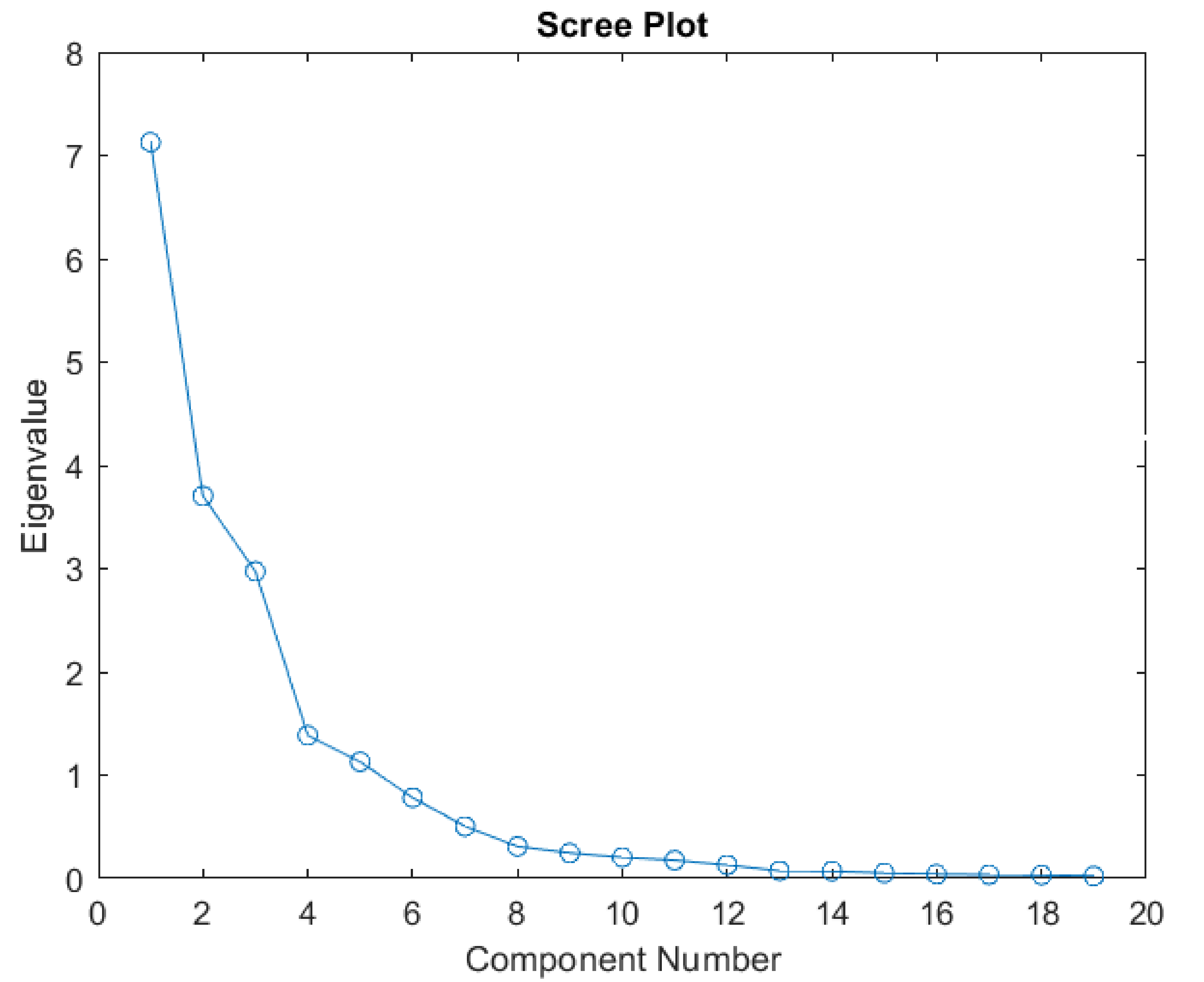
Access on dataset: https://archive.ics.uci.edu/dataset/183/communities+and+crime

Initial Correlation Matrix

- Standardization of data matrix
- Computation of correlation matrix:



Results of EFA

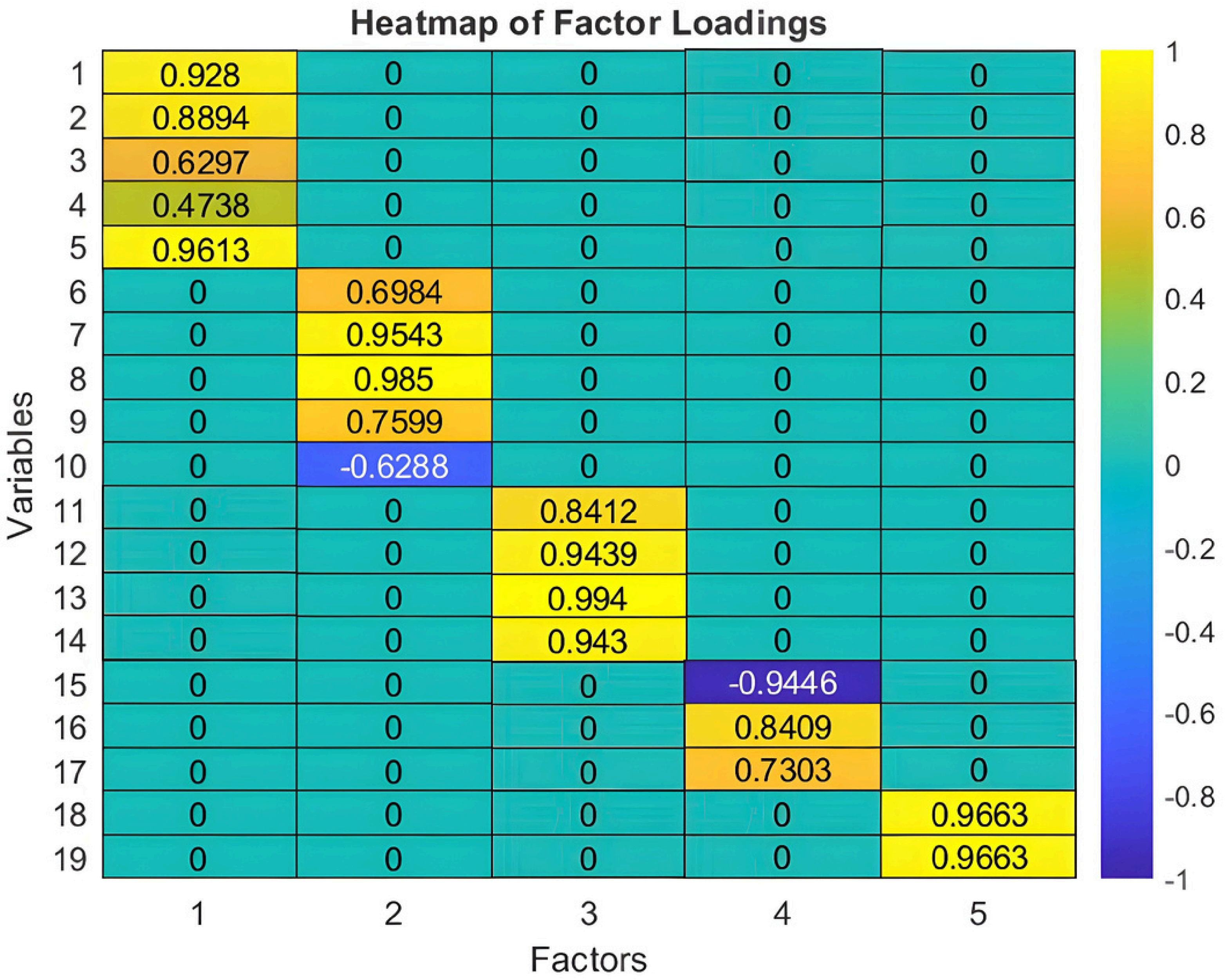


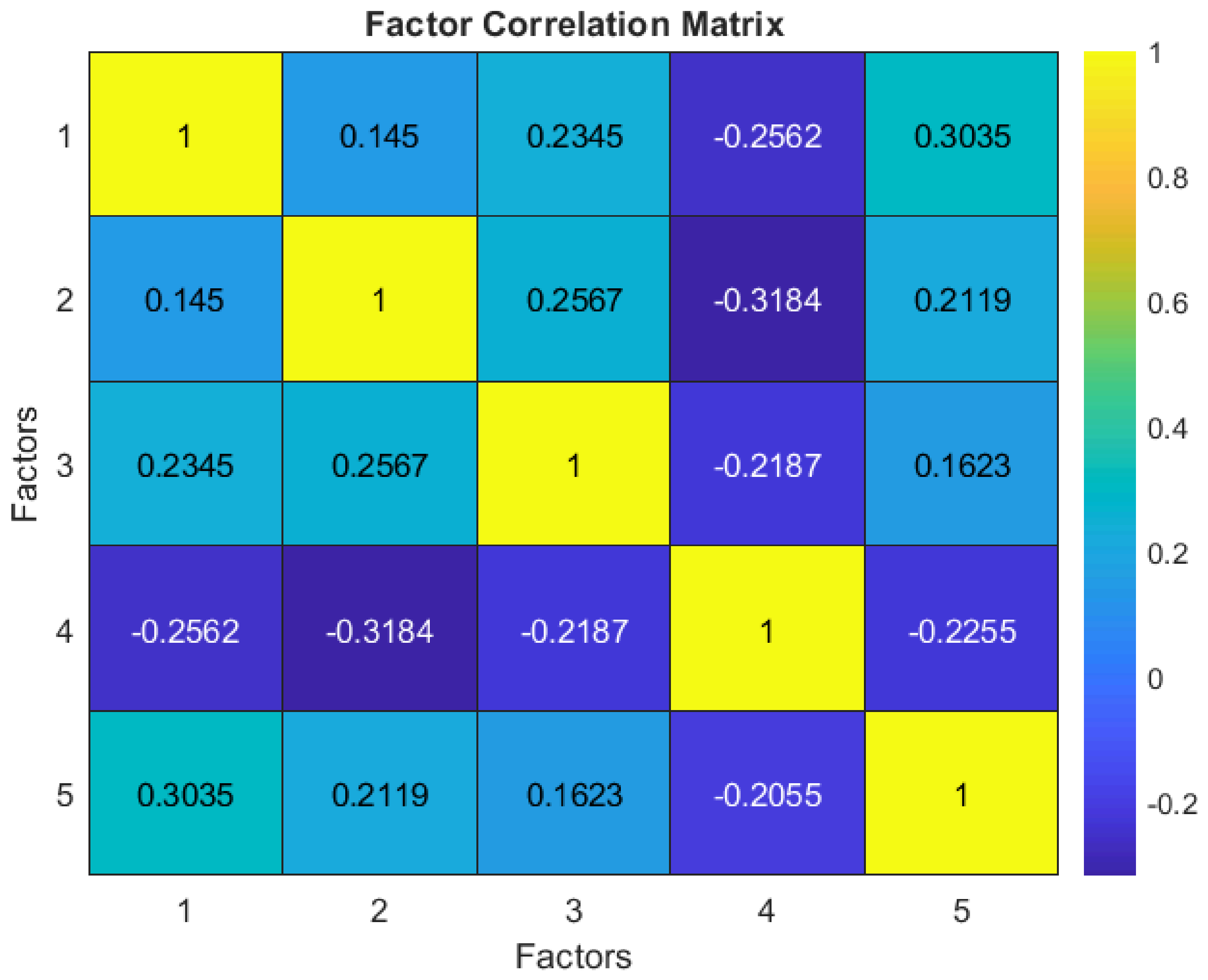
Eigen Values
7.1281
3.7010
2.9724
1.3855
1.1296
0.7822
0.5044
0.3102
0.2465
0.2047
0.1766
0.1305
0.0736
0.0695
0.0521
0.0410
00367
0.0324
0.0229

k = 5 factors

(Kaiser's criterion)

Results of DFA with 5 factors






```
explained variance for 5 factors: 0.73802
```

SAPIENZA
UNIVERSITÀ DI ROMA

Specific Composite Indicator (SCI)

Cultural integration

- PctRecentImmig: percent of population who have immigrated within the last 3 years (numeric - decimal)
- PctNotSpeakEnglWell: percent of people who do not speak English well (numeric - decimal)
- PctLargHouseFam: percent of family households that are large (6 or more) (numeric - decimal)
- PersPerRentOccHous: mean persons per rental household (numeric - decimal)
- PctForeignBorn: percent of people foreign born (numeric - decimal)

Socioeconomic condition

- PctPopUnderPov: percentage of people under the poverty level (numeric - decimal)
- PctLess9thGrade: percentage of people 25 and over with less than a 9th grade education (numeric - decimal)
- PctNotHSGrad: percentage of people 25 and over that are not high school graduates (numeric - decimal)
- PctUnemployed: percentage of people 16 and over, in the labor force, and unemployed (numeric - decimal).
- medFamInc: median family income (differs from household income for non-family households) (numeric - decimal)

Recent immigration

- PctRecImmig5: percent of _population_ who have immigrated within the last 5 years (numeric - decimal)
- PctImmigRec5: percentage of immigrants who immigrated within last 5 years (numeric - decimal)
- NumInShelters: number of people in homeless shelters (numeric - decimal)
- PctSameCity5: percent of people living in the same city as 5 years before (numeric - decimal)

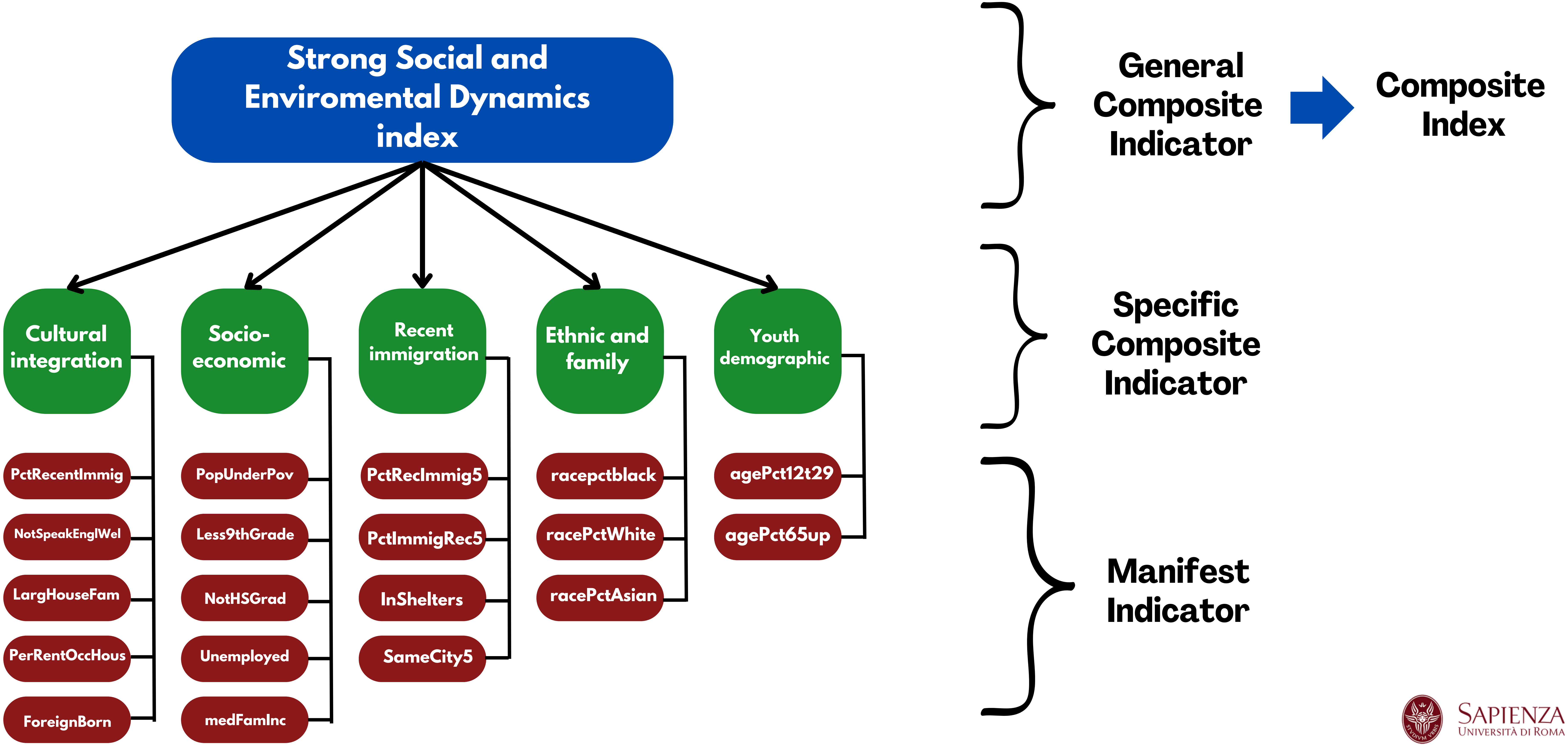
Ethnic and Family composition

- racepctblack: percentage of population that is african american (numeric - decimal)
- racePctWhite: percentage of population that is caucasian (numeric - decimal)
- racePctAsian: percentage of population that is of asian heritage (numeric - decimal)

Youth demographics

- agePct12t29: percentage of population that is 12-29 in age (numeric - decimal)
- agePct65up: percentage of population that is 65 and over in age (numeric - decimal)

SAPIENZA
UNIVERSITÀ DI ROMA



SAPIENZA
UNIVERSITÀ DI ROMA

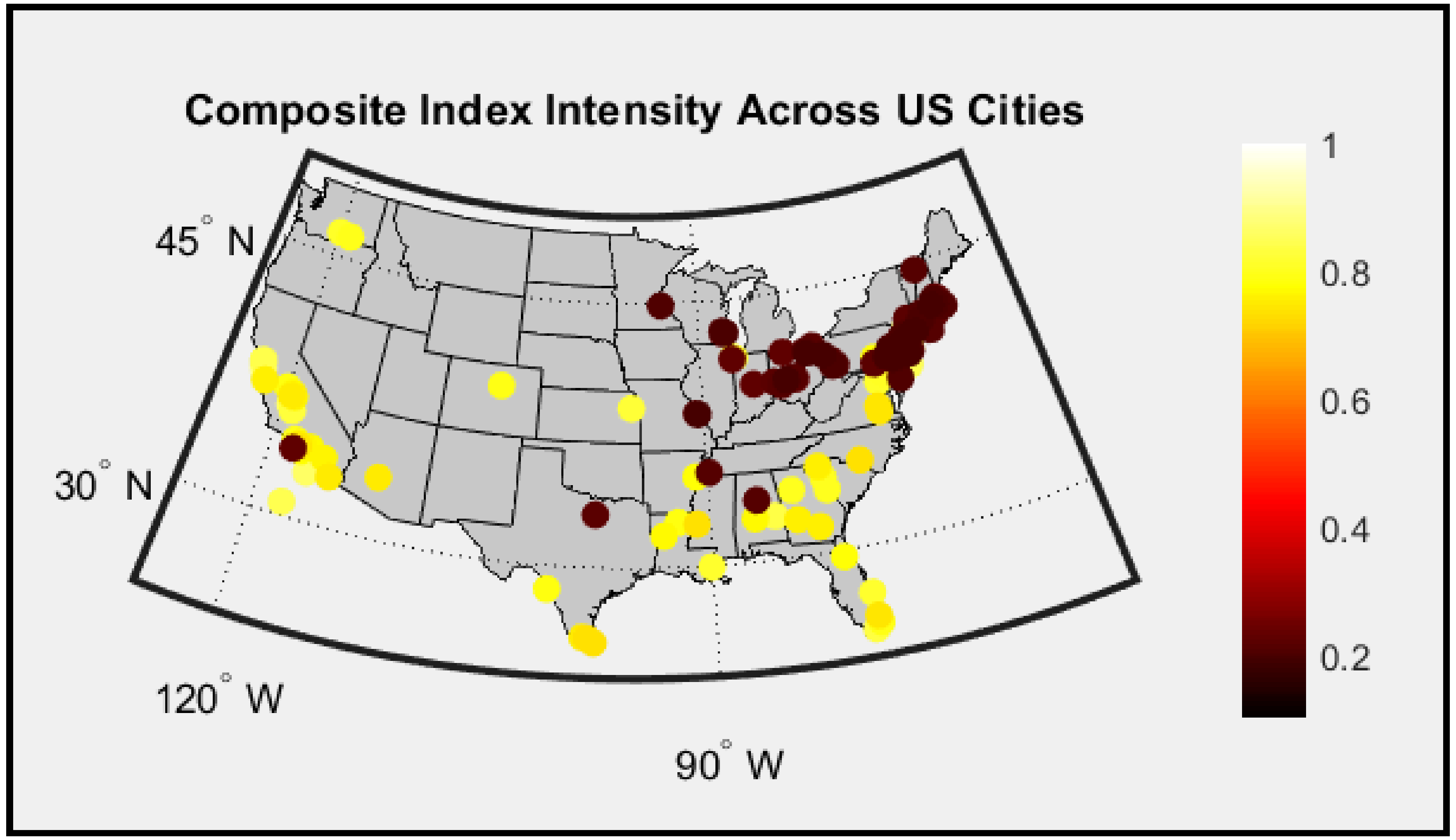
Comparison between Highest and Lowest Composite Index

Community name	Composite Index
Newark City	1
Camden City	0.9636
Miami City	0.9251
Compton City	0.9195
Lynwood City	0.8559
Cudahy City	0.8506
Huntington City	0.8493
Bell City	0.8419
Bell Gardens City	0.8382
Hartford Town	0.8297

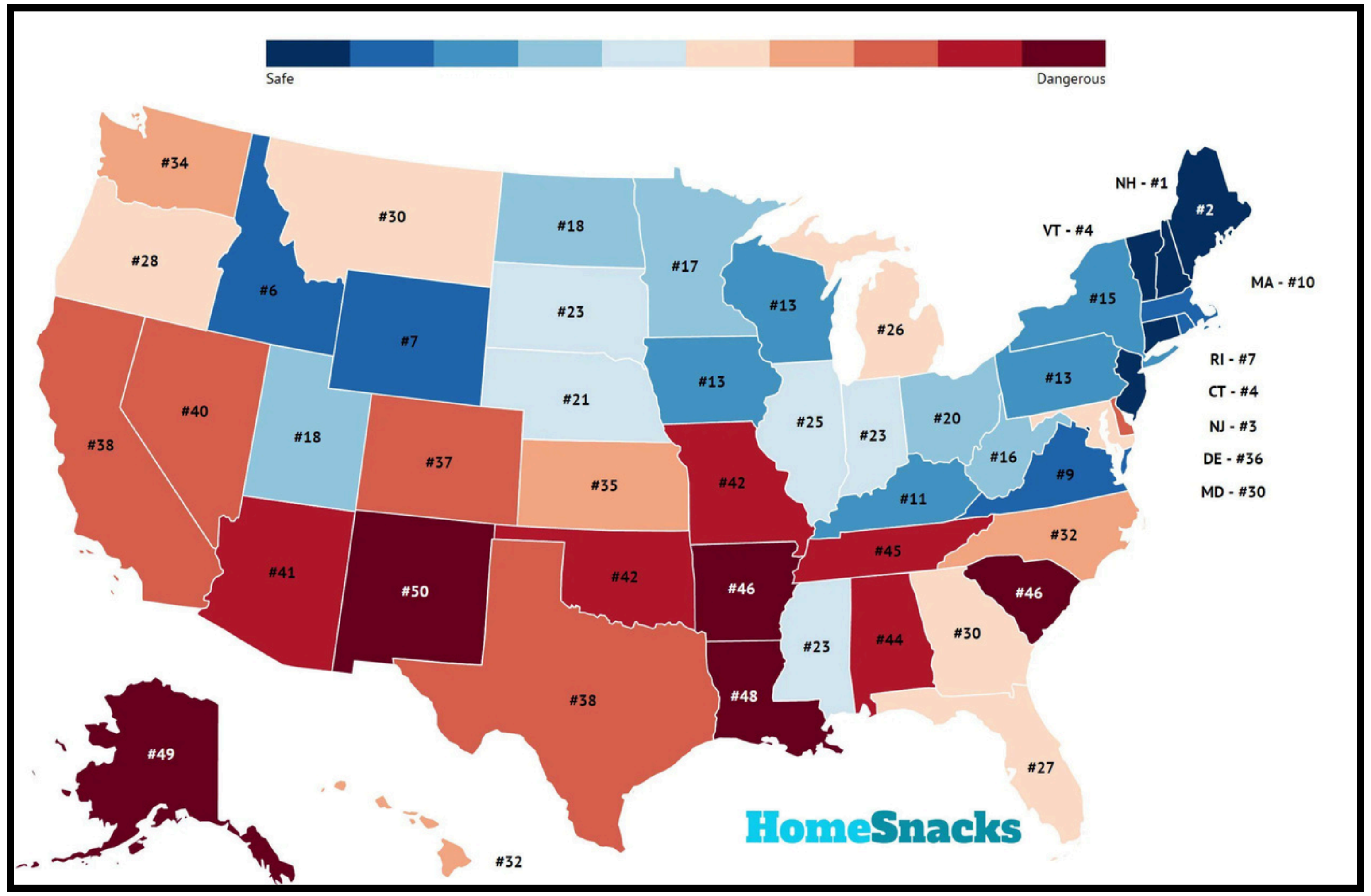
Community name	Composite Index
Paradise Valley Town	0.0246
Mequon City	0.0244
Hopewell Township	0.0209
Brentwood City	0.0152
Colleyville City	0.0144
German Town City	0.0111
Bedford Town	0.0106
Dublin City	0.0061
Mountain Brook City	0.0028
Sudbury Town	0

Comparison of Results with Real-World Data

Top 100 and Bottom 100 Cities by Index Value



Data comes from the FBI Crime Explorer



SAPIENZA
UNIVERSITÀ DI ROMA

THANK YOU !