

Building indicators through a statistical ap- proach based on principal component analysis (PCA)

Fernando Santa
June 12th, 2018

Outline

- Definitions
- Statistical considerations
- Assumptions

Definitions

- Variable: a property that is measurable or observable in an object of study

x

- Indicator: a variable, or a value derived from variables, which gives information about the state of a phenomenon with a meaning

z

Methods

- Non statistical approach: multicriteria evaluation
- Statistical approach: PCA, Empirical orthogonal functions (EOF), Independent Component Analysis (ICA)

Statistical considerations

- PCA is an exploratory multivariate technique
- Reduce dimensionality of the problems through building a new set of “latent” variables as linear combinations of the original variables

$$z = a_1 x_1 + a_2 x_2 + \dots + a_k x_k$$

- It looks for keeping the maximum amount of the variability of the original data in a space with less dimension

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

- Its performance is based on the assumptions of classic multivariate statistics

Assumptions

- Univariate symmetrical distributions
- Linear relationship between pair of variables
- Statistical units are independent

Procedure

- Explore univariate distributions to identify outliers and skewed distributions
- Explore bivariate distributions to identify possible non-linear relationship
- Build the PCA doing the spectral decomposition of the correlation matrix

Univariate distributions

- Histograms
- Measures (CV, Skewness, Kurtosis)

Bivariate distributions

- Scatterplots
- Pearson's correlation coefficient

Observations

- Use non-linear transformations such as Box-Cox transformations to reduce skewed and non constant variance

$$y_i = \begin{cases} \frac{(x_i+m)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log x_i + m & \text{if } \lambda = 0 \end{cases}$$

- Build the PCA with the set of variables after transformation and evaluate the effect