Functional archetype and archetypoid analysis

Irene Epifanio
Dpt. Matemàtiques, Univ. Jaume I (SPAIN)
epifanio@uji.es; http://www3.uji.es/~epifanio
UNIVERSITAT JAUME•I

Outline

- Archetypal analysis

- AA and ADA for multivariate data
- AA and ADA for functional data
- Application: Human development around the world over the last 50 years

- Conclusions and future work

Archetypes

- Archetype (Wikipedia): from Greek, ἀρχή, archē, "beginning, origin", and τύπος, tupos, "pattern," "model," or "type"; original pattern from which copies are made.

UNIVERSITAT JAUME•I

Archetypes in Star Wars

Wise Old Man

Damsel in distress

Female warrior

Reluctant hero

Epic hero

Evil figure

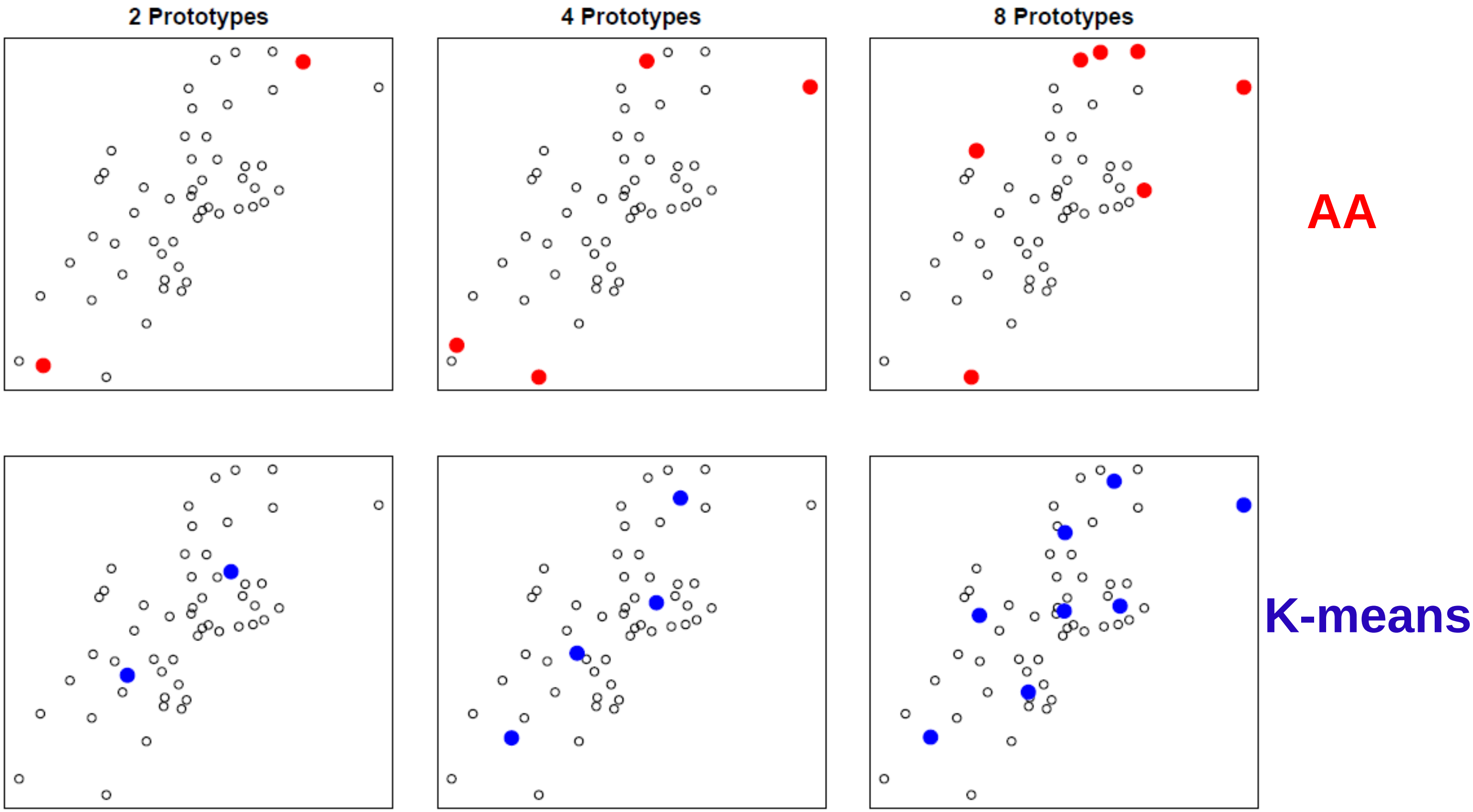
Archetype analysis (AA)

- **Archetype concept in Statistics is the same as in life common.**
- **Objetive (Cutler and Breiman, 1994): to find a few, not necessarily observed, extremal cases or pure types (the archetypes) such that:**

1. all the observations are approximated by convex combinations of the archetypes, and

2. all the archetypes are convex combinations of the observations.

Archetypes in 2D



Archetypoid analysis (ADA)

- Objetive (Vinué, Epifanio, Alemany, 2015): to find a few, observed, extremal cases or pure types (the archetypoids) such that:

  1. all the observations are approximated by convex combinations of the archetypoids, and

  2. all the archetypoids are real observations.

UNIVERSITAT JAUME•I

AA for multivariate data

Let X be an *n×m* matrix with n observations and m variables. The objective of AA is to find the matrix Z of *k* *m*-dimensional archetypes. AA computes two matrices α and β which minimize the residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2, \quad (1)$$

under the constraints

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$ and

2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ for $j = 1, \ldots, k$.

UNIVERSITAT JAUME•I

ADA for multivariate data

The objective of ADA is to find the matrix Z of *k* *m*-dimensional archetypoids (real cases). ADA computes two matrices α and β which minimize the residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2, \quad \textbf{(2)}$$

under the constraints

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$ and

2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \in \{0, 1\}$ and $j = 1, \ldots, k$. ←

UNIVERSITAT JAUME•I

AA solution

- Cutler and Breiman (1994) proposed an alternating minimizing algorithm.
- Implemented in R by Eugster and Leisch (2009):

package archetypes.

- To solve the convex least squares problems, they used a penalized version of the non-negative least squares algorithm by Lawson and Hanson (1974).

UNIVERSITAT JAUME•I

ADA solution

- Vinué, Epifanio, Alemany (2015) proposed an algorithm.
- It consists of two phases, a BUILD step and a SWAP step.
  - An initial set of archetypoids is computed in the BUILD phase.
  - The SWAP step seeks to improve the set of archetypoids by exchanging chosen observations for unselected cases and by checking if these replacements reduce the RSS.
- Implemented in R by Vinué et al. (2015b):

  package Anthropometry.

UNIVERSITAT JAUME•I

AA and ADA for functional data

Objective of functional AA (FAA): to find *k* archetype functions (mixture of the data),

Objective of functional ADA (FADA): to find *k* functions of the sample (archetypoids),

in such a way that our functional data sample can be approximated by mixtures of those archetypal functions.

- The vector norms are replaced by functional norms (L^2 norm, $\|f\|^2 = < f, f > = \int_a^b f(t)^2 dt$) in equation 1 and 2; the vectors x_i and z_j correspond to the functions x_i and z_j.
- The meaning of α and β in the functional case is identical to the multivariate case.

UNIVERSITAT JAUME•I

Computational details: basis approach

- Each function x_i is expressed as a linear combination of known basis functions B_h with *h* = 1, ..., *m*:

$$x_i(t) = \sum_{h=1}^{m} b_i^h B_h(t) = \mathrm{b}_i' \mathrm{B}$$

- b_i : the vector of the coefficients
- B the functional vector whose elements are the basis functions.

UNIVERSITAT JAUME•I

Computational details: basis approach

$$RSS = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{k} \alpha_{ij} z_j\|^2 = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} x_l\|^2 =$$

$$\sum_{i=1}^{n} \|\mathrm{b'}_i \mathrm{B} - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathrm{b'}_l \mathrm{B}\|^2 = \sum_{i=1}^{n} \|(\mathrm{b'}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathrm{b'}_l)\mathrm{B}\|^2 = \quad (3)$$

$$\sum_{i=1}^{n} \|\mathrm{a'}_i \mathrm{B}\|^2 = \sum_{i=1}^{n} < \mathrm{a'}_i \mathrm{B}, \mathrm{a'}_i \mathrm{B} >= \sum_{i=1}^{n} \mathrm{a'}_i \mathrm{W} \mathrm{a}_i,$$

where: $\mathrm{a'}_i = \mathrm{b'}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathrm{b'}_l$ **and W is the matrix containing the inner products of the pairs of basis functions** $w_{m_1,m_2} = \int B_{m_1} B_{m_2}$

Constraints for α and β identical as the multivariate case.

UNIVERSITAT JAUME•I

Computational details: basis approach

- In the case of an orthonormal basis such as Fourier, W is the order *m* identity matrix, and FAA (FADA, respectively) is reduced to AA (ADA, respectively) of the basis coefficients.
- But, in other cases, we may have to resort to numerical integration to evaluate W, but once W is computed, no more numerical integrations are necessary.

UNIVERSITAT JAUME•I

Multivariate FAA and FADA

- Key: to define an inner product between bivariate functions, which is computed simply as the sum of the inner products of the two components.
- FAA or FADA for M multivariate functions is equivalent to M independent FAA or FADA, respectively, with shared parameters α and β.

UNIVERSITAT JAUME•I

Multivariate FAA and FADA computation

- Let $f_i(t) = (x_i(t), y_i(t))$ be a bivariate function. Its squared norm: $\|f_i\|^2 = \int_a^b x_i(t)^2 dt + \int_a^b y_i(t)^2 dt$
- The coefficients for x_i and y_i respectively for the basis functions B_h are b^x_i and b^y_i

$$RSS = \sum_{i=1}^{n} \|f_i - \sum_{j=1}^{k} \alpha_{ij} z_j\|^2 = \sum_{i=1}^{n} \|f_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} f_l\|^2 =$$

$$\sum_{i=1}^{n} \|x_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} x_l\|^2 + \sum_{i=1}^{n} \|y_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} y_l\|^2 = \quad (4)$$

$$\sum_{i=1}^{n} \mathrm{a^x}'_i \mathrm{W} \mathrm{a^x}_i + \sum_{i=1}^{n} \mathrm{a^y}'_i \mathrm{W} \mathrm{a^y}_i,$$

where $\mathrm{a^x}'_i = \mathrm{b^x}'_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathrm{b^x}'_l$ and $\mathrm{a^y}'_i = \mathrm{b^y}'_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathrm{b^y}'_l$

UNIVERSITAT JAUME•I

Application

- Two indicators of World Bank Open Data:
  - Total fertility rate (TFR): no. children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with current age-specific fertility rates.
  - Life expectancy at birth (LEB): the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.

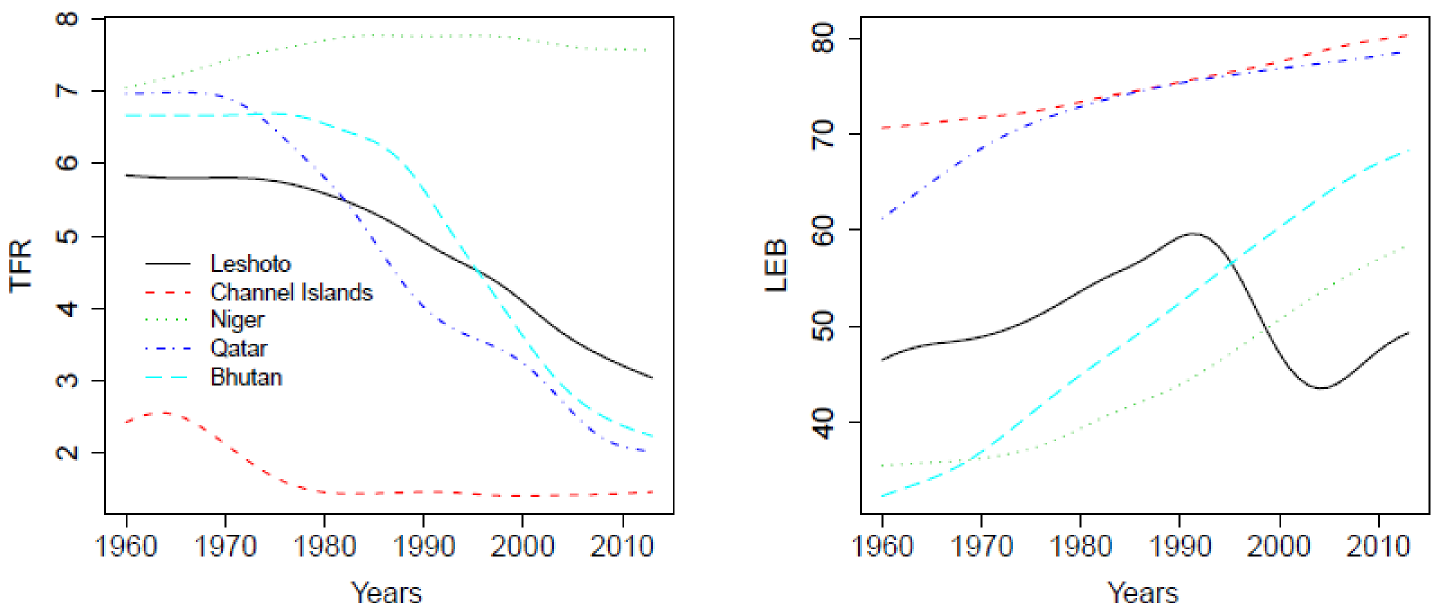
- The series of each country goes from 1960 to 2013.

Application

- 190 countries considered.
- All functions (those with or without missing years) are expressed with 32 B-spline basis functions of order 4 (cubic splines) from 1960 to 2013, with equally spaced knots.
- TFR and LEB are measured in non-compatible units, so each functional variable should be standardized.
- Bivariate FADA with $k = 5$ archetypoids.

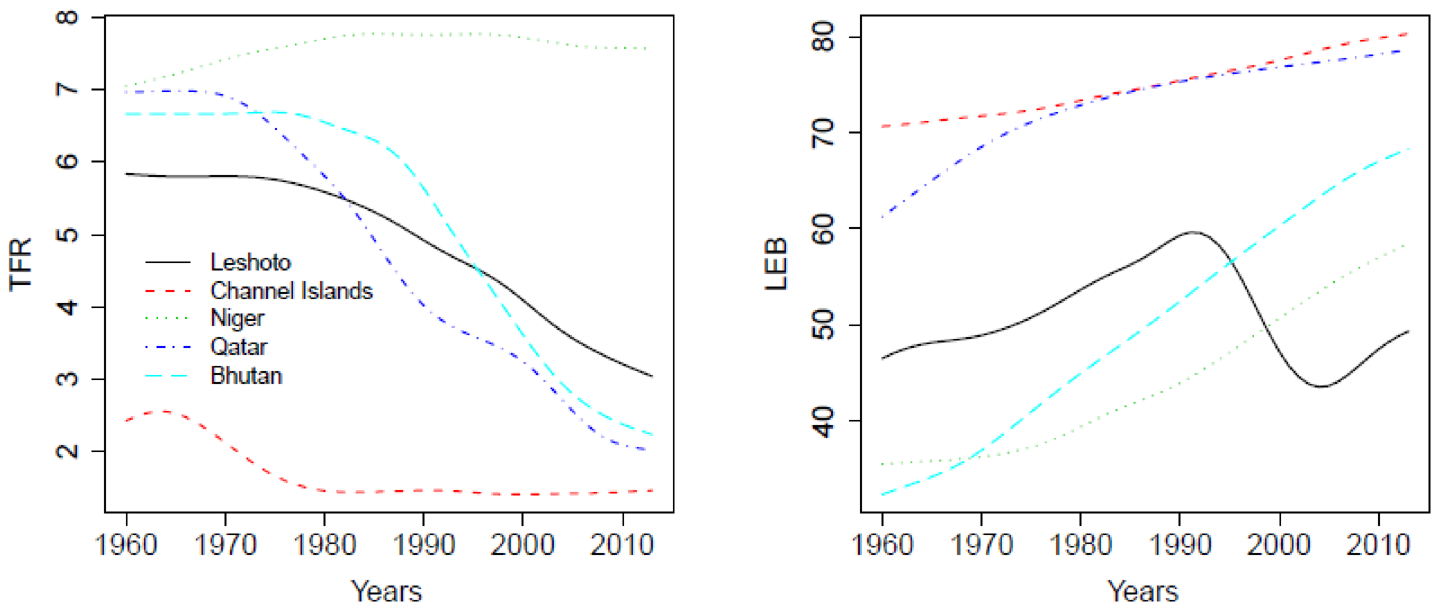
UNIVERSITAT JAUME•I

5 functional archetypoids



UNIVERSITAT JAUME•I

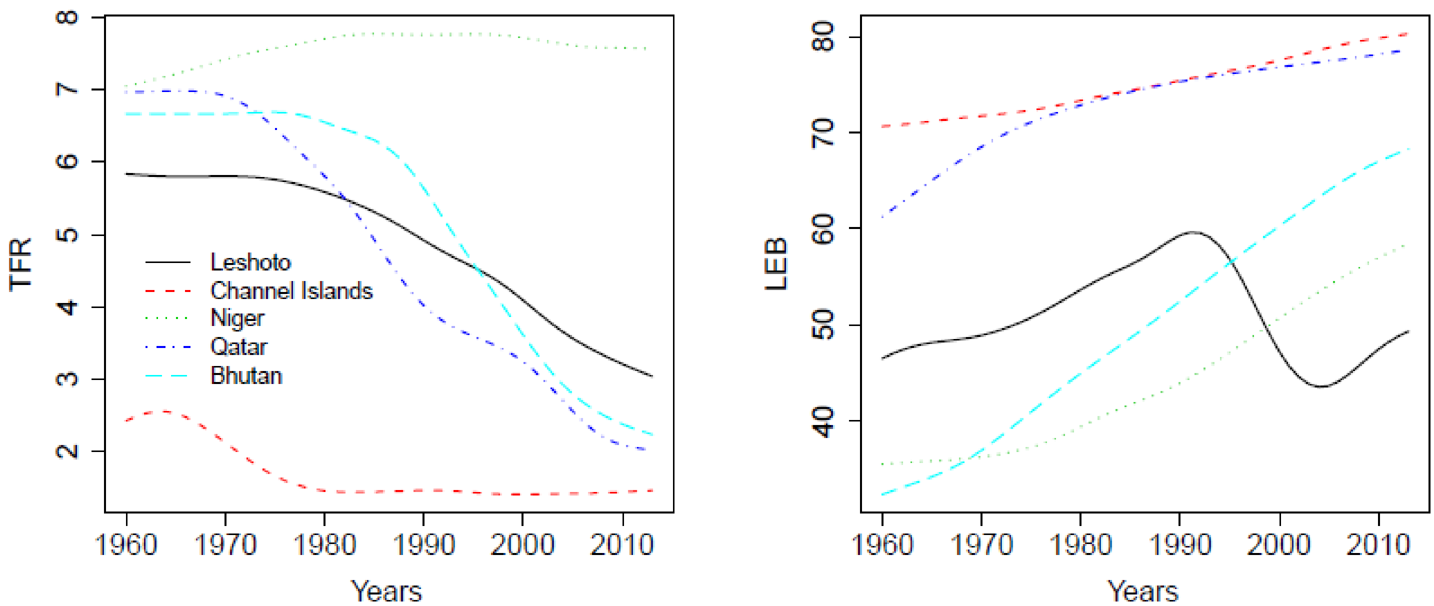
Leshoto



- TFR has decreased from nearly 6 children in 1960 to 3.
- LEB curve reflects a significant problem in Southern Africa: HIV/AIDS.

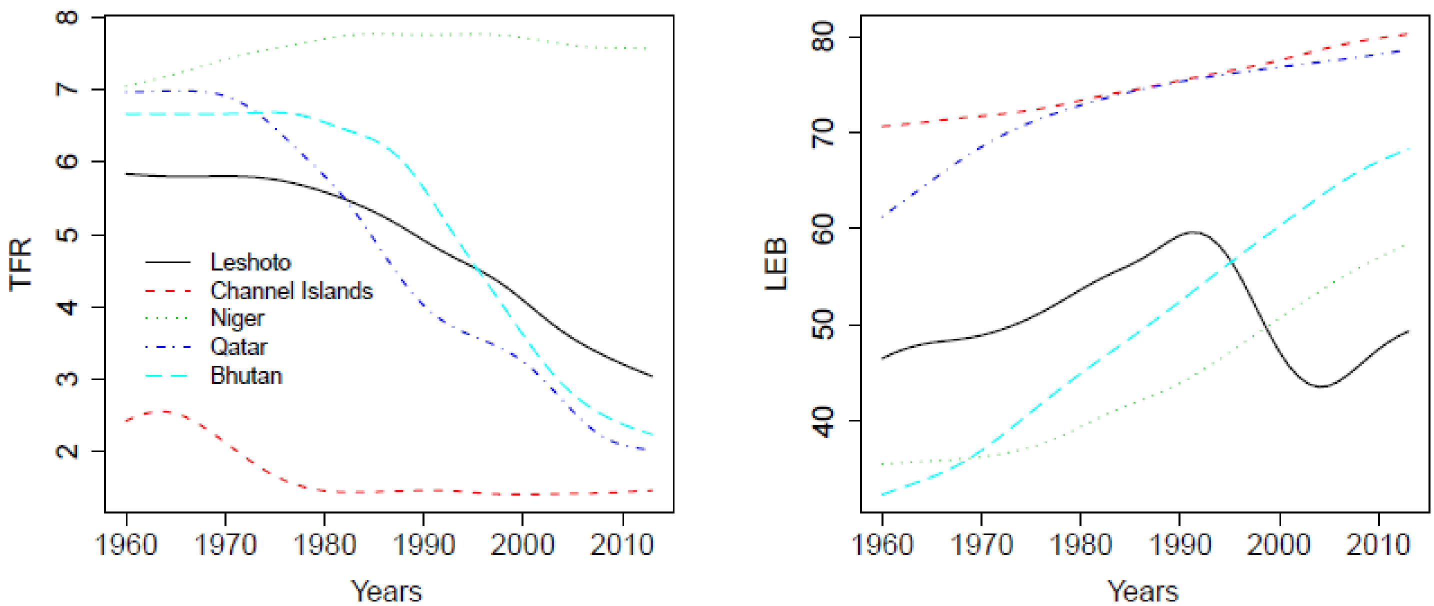
UNIVERSITAT JAUME•I

Channel Islands



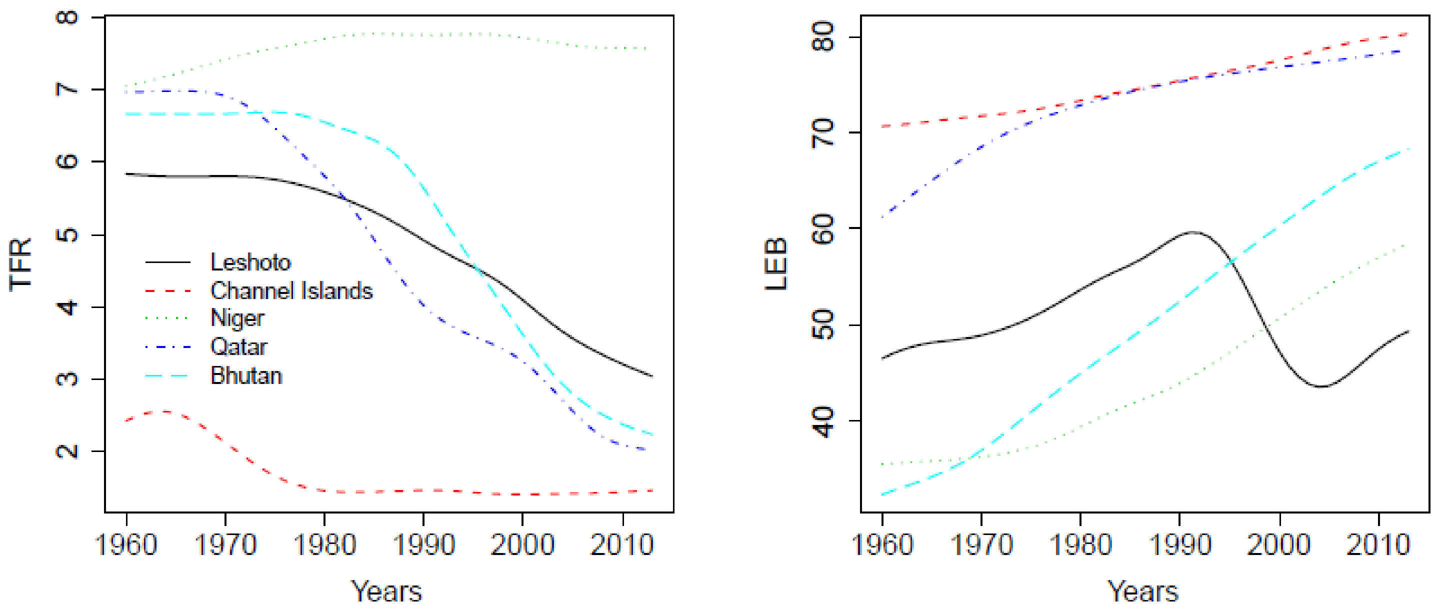
- **Representative of countries with low TFR and high LEB over the years.**

Niger



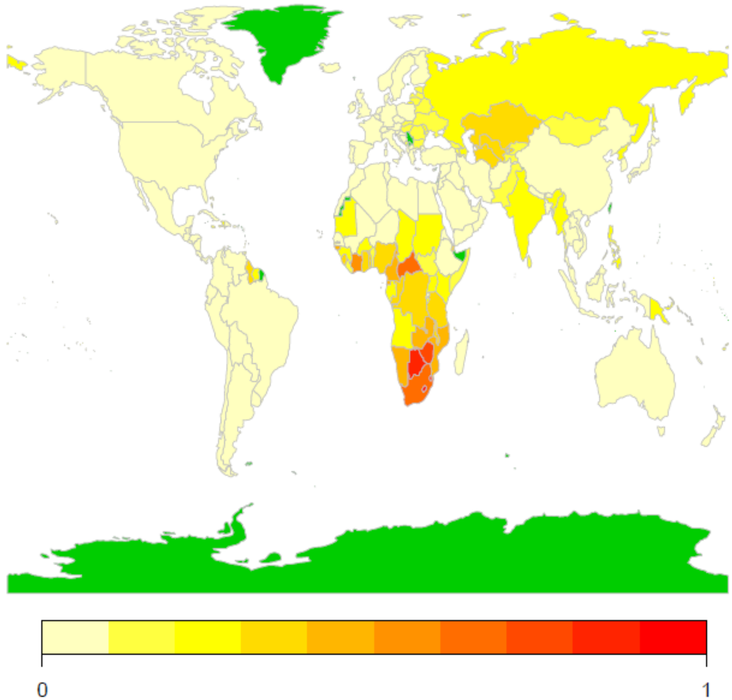
- **Representative of countries with high TFR over the years, but low LEB (36 years) in the 1960s, which has increased to nearly 60 years nowadays.**

Qatar and Bhutan



- TFR has decreased spectacularly, from nearly 7 in the 1960s to 2 nowadays. But, this decrease has taken place at different times.
- LEB has increased considerably.

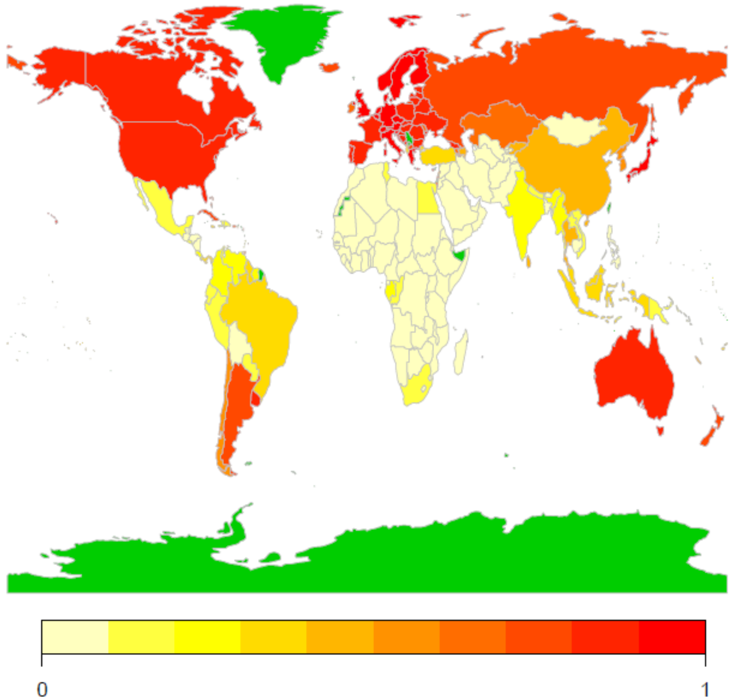
Abundance map for Leshoto



- Countries with indicator curves similar to Lesotho are their neighboring countries, which are the countries most affected by HIV/AIDS.

UNIVERSITAT JAUME•I

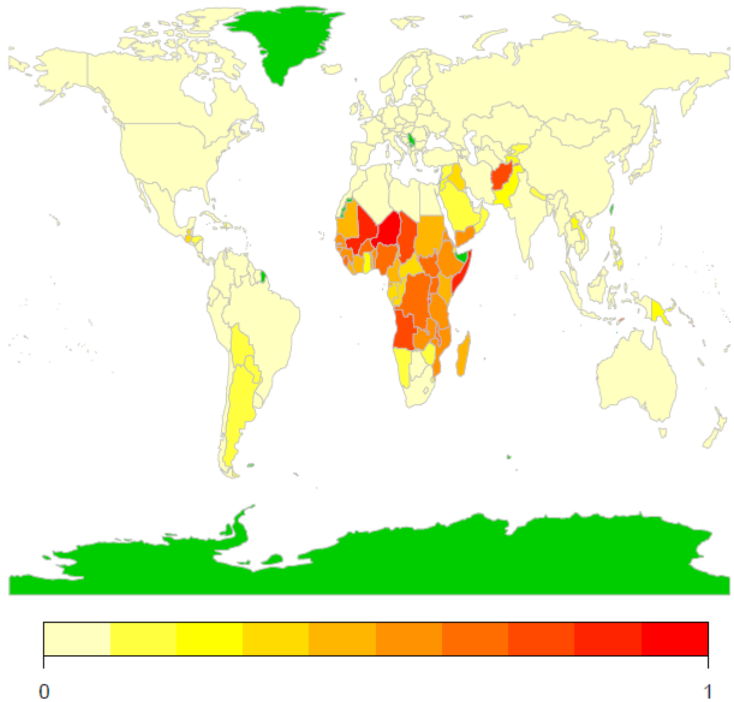
Abundance map for Channel Islands



- The countries whose indicator functions coincide with those of the Channel Islands are Japan, Australia, North America and most European countries, and to a lesser extent, countries such as Russia and Argentina.

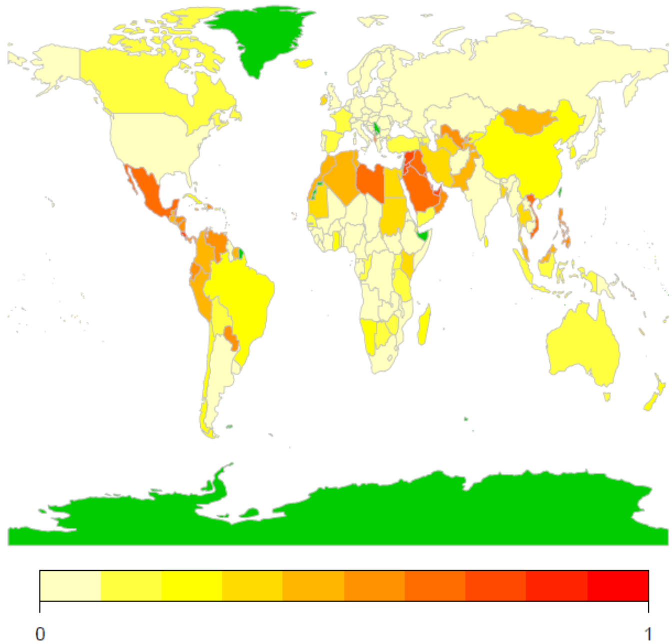
UNIVERSITAT JAUME•I

Abundance map for Niger



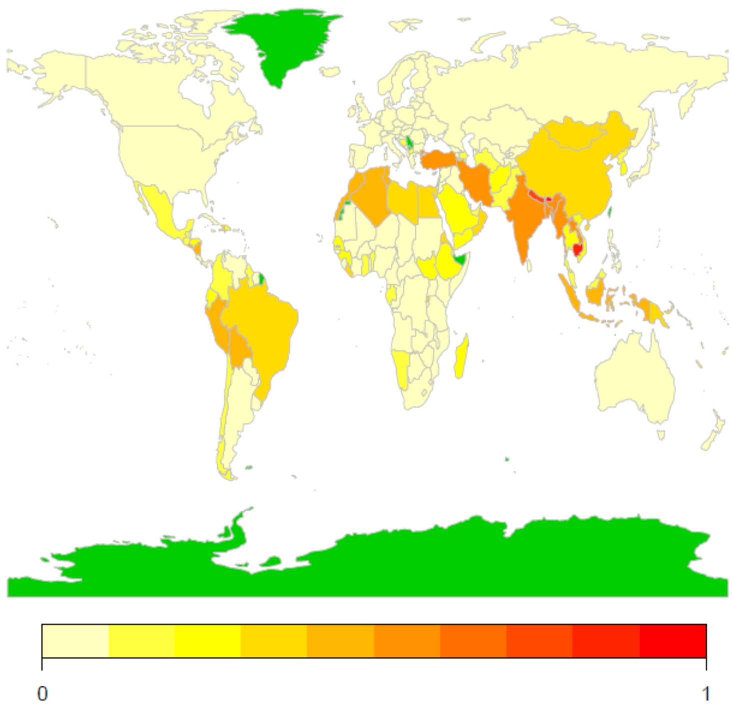
- **Countries which mainly share their indicator functions are those in Central Africa and Afghanistan.**

Abundance map for Qatar



- Countries with a similar behavior: majority of countries in the Arabian peninsula and neighboring countries, many countries in Central America and several in South America, several in Asia and countries in North Africa, although, those North African countries also share characteristics with Bhutan.

Abundance map for Bhutan



- Morocco, Algeria and Tunisia are a mixture of Qatar and Bhutan. Other countries are also a mixture of two or three profiles. For example, Turkey is a mixture between 30% the Channel Islands, 20% Qatar and 50% Bhutan.

Conclusions

- FAA and FADA are introduced.
- Computational methods are proposed based on the coefficients of a basis.
- Multivariate FAA and FADA are also introduced.
- Bivariate FADA applied to the study of human development around the world over the last 50 years.
- The use of FAA and FADA can be an interesting tool for making data easier to interpret, since they are based on the principle of opposites which accommodates human cognition.

UNIVERSITAT JAUME•I

Future work

- Weighted and robust functional versions.
- AA and ADA for mixed data (functional and vector parts).
- FAA and FADA when multivariate arguments are involved.
- Other techniques for non-negative matrix factorization could be extended to the functional case.
- Applications: AA and ADA, and FDA are quite new, and therefore there is no doubt plenty of scope for combining them. In fact, …

UNIVERSITAT JAUME•I

Future work

- Applications: AA and ADA, and FDA are quite new, and therefore there is no doubt plenty of scope for combining them. In fact, … FADA has recently applied in sports analytics, to find archetypoids in NBA.

UNIVERSITAT JAUME•I

- Guillermo Vinué and Irene Epifanio. Archetypoid Analysis for Sports Analytics. Data Mining and Knowledge Discovery

Reference

- I. Epifanio. Functional archetype and archetypoid analysis. Computational Statistics & Data Analysis, 64 (3), 757-765, 2016.

- Pre-print version and code are available at http://www3.uji.es/~epifanio

UNIVERSITAT JAUME•I

Thank very much for your attention

http://www3.uji.es/~epifanio

UNIVERSITAT JAUME•I