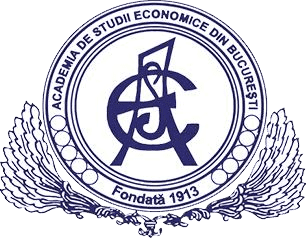
**ACADEMIA DE STUDII ECONOMICE DIN BUCUREȘTI, FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ**



**PROIECT DEZVOLTARE SOFTWARE PENTRU ANALIZA DATELOR**

**Scurtă analiză asupra speranței de viață**

Prof Coordonator: Vinte Claudiu

Studenti:

Radu Mihaela

Radu Elena

Pirvulescu Alice-Maria

Maierean Alex-Gabriel

### 

1. **Sursa de date**

In urma analizei pe care o propun in acest proiect putem determina care sunt factorii cu o influenta sporita asupra sperantei de viata.

Datele utilizate în acest proiect provin din: [<https://data.worldbank.org/>] și au fost colectate în contextul:[<https://www.kaggle.com/kacperk77/life-expectancy/version/2?select=lifeexpectancy1.csv>]. Setul de date oferă informații relevante despre Speranta de viata, careeste durata medie a vietii unei persoane. Detalii despre setul de date pot fi găsite la: [<https://en.wikipedia.org/wiki/Life_expectancy>].

1. **Referinta de timp**

Seturile de date pe care le-am colectat se referă la situații din întreaga lume pentru anul 2014. Acestea fac referire la mai multe tari din regiuni diferite.

1. **Descrierea variabilelor**

### În vederea realizării unei analize, am folosit 63 variabile, reprezentate de cele 63 țări mentionate,și 18 variabile reprezentand astfel:

* **literacyrate:** Rata de alfabetizare, total adulți (% din populația cu vârsta de 15 ani și peste) în 2014
* **homicidiesper100k:** Omucideri la 100.000 de locuitori în 2014
* **electricity:** Consum de energie electrică (kWh per capita) în 2014
* **Schooling:** Numărul de ani de școlarizare (ani) în 2014
* **HIV.AIDS:** Decese la 1.000 de nașteri vii din cauza HIV/SIDA (0-4 ani) în 2014
* **wateraccess:** Acces la surse de apă îmbunătățite (% din populația totală cu acces) în 2014
* **tuberculosis:** Incidența tuberculozei (la 100.000 de persoane) în 2014
* **inflation:** Inflație, prețuri de consum (anual %) în 2014
* **healthexppercapita:** Cheltuieli medii de sănătate pe cap de locuitor, PPC 2005-2014
* **fertilityrate:** Rata de fertilitate, total nașteri (per femeie) în 2014
* **lifeexp:** Speranța de viață la naștere, total (ani) în 2014
* **internet:** Persoanele care utilizează Internetul (% din populație) în 2014
* **gdppercapita:** PIB pe cap de locuitor, PPC (dolari internaționali curenți) în 2014
* **CO2:** Emisii medii de CO2 (tone metrici pe cap de locuitor) 2005-2014
* **forest:** Suprafața forestieră (% din suprafața terenului) în 2014
* **urbanpop:** Populația urbană în 2014
* **urbanpopgrowth**: Creșterea medie a populației urbane (anual %) în perioada 2005-2014
* **leastdeveloped:** 1 = țara este considerată cel puțin dezvoltată; 0 = țara este considerată în curs de dezvoltare sau dezvoltată

1. **Descrierea observațiilor**

Orice gurven este preocupat de speranta de viata a cetatenilor sai si are ca scop intretinerea sau cresterea sperantei de viata. In urma analizei pe care o propun in acest proiect putem determina care sunt factorii cu o influenta sporita asupra sperantei de viata.

Analiza efectuata in proiect urmareste sa raspunda la urmatoarele intrebari cheie:

* Care sunt principalii factorii economici care influenteaza speranta de viata?
* Care sunt factorii sociali care influenteaza cel mai mult speranta de viata?
* Daca statutul economic al tarii este o influenta majora asupra.
* Ce alti factori ar putea avea o influenta substantiala asupra sperantei de viata?

1. **Abordarea privind analiza datelor**

Analiza datelor va implica utilizarea mai multor metode statistice și tehnici de analiză pentru a obține o înțelegere detaliată a setului de date. În primul rând, vom efectua o Analiză Factorială (AF) pentru a identifica posibilele relații și corelații între variabilele observate. Această tehnică ne va ajuta să reducem dimensiunea setului de date și să identificăm factorii subiacenți care pot influența variabilitatea observațiilor.

Ulterior, vom aplica Analiza Componentelor Principale (ACP) pentru a obține o perspectivă asupra structurii de bază a datelor, evidențiind contribuția fiecărei variabile la variabilitatea totală. Această metodă ne va permite să identificăm principalele componente care explică variația din setul de date.

### **Motivația de a opta pentru o anumită metodă de analiză a datelor**

### Am optat pentru utilizarea Analizei Factoriale (AF) și a Analizei Componentelor Principale (ACP) în cadrul acestui proiect, având în vedere natura complexă a setului de date și obiectivele noastre specifice. Iată câteva argumente care susțin alegerea noastră:

### **Dimensiunea Redusă a Setului de Date:**

### Datorită numărului semnificativ de variabile în setul nostru de date, AF și ACP sunt alegeri potrivite pentru a reduce dimensiunea setului și a identifica factorii principali care contribuie la variabilitate.

### **Explorarea Relațiilor Complexe:**

### Analiza Factorială permite investigarea relațiilor complexe între variabile, oferind o perspectivă asupra structurii ascunse a datelor. Acest aspect este esențial în contextul nostru, unde interacțiunile între diferite aspecte ale dezvoltării socio-economice pot fi intricate.

### **Identificarea Componentelor Principale:**

### Analiza Componentelor Principale ne va ajuta să identificăm variabilele-cheie care contribuie semnificativ la variabilitatea totală. Această informație este crucială pentru a evidenția factorii principali care influențează fenomenele studiate.

### **Interpretare și Comunicare Eficientă:**

### ACP oferă rezultate interpretabile și ușor de comunicat, ceea ce va facilita prezentarea concluziilor noastre către diverse părți interesate, inclusiv colegi de proiect, decidenți și publicul larg.

### **Prezentarea rezultatelor (** Interpretare, Discuții și Concluzii)

## 

## **Analiza variabilitatii datelor**

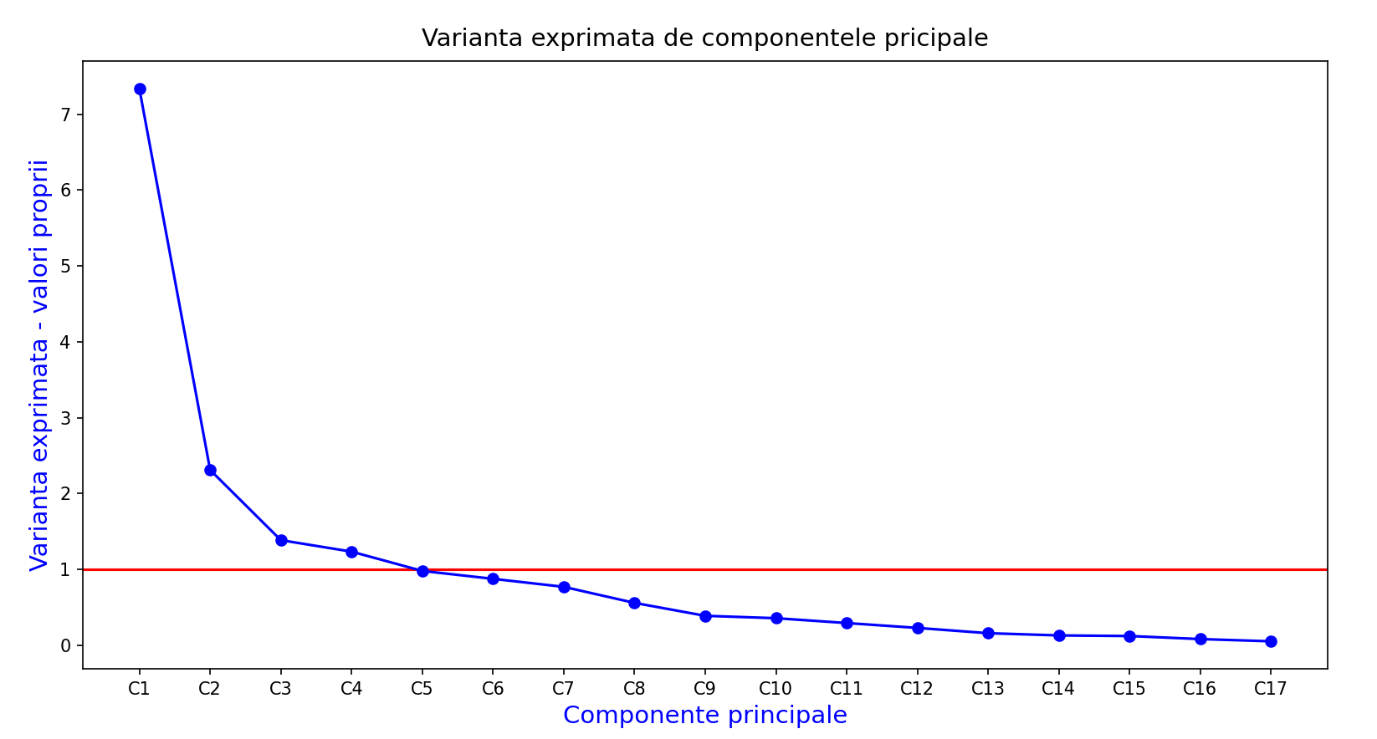


Fig.1: Varianta exprimata de componentele principale

Analiza in componente principale pleaca de la calculul vectorilor si valorilor proprii ale matricei de corelatie. Valorile proprii reprezinta variantele componentelor, iar vectorii proprii reprezinta coeficientii legaturii liniare dintre variabilele observate si componentele principale.

Conform criteriilor Cattel și Kaiser, sunt semnificative doar primele patru componente, deoarece au o varianta mai mare decat 1.

## **Corelatii intre variabile si componentele principale**

Aceste corelatii, numite si corelatii factoriale se calculeaza pronind de la vectorii matricei de corelatie. Acestea reprezinta coeficientii de corelatie liniara dintre componentele principare si variabilele studiate. Fiind calculate din vectroii proprii ai matricei, corelatiile reflecta contributia fiecarei variabile studiate la crearea unei componente principale. In urmatoarea imagine putem observa corelatiile factoriale:

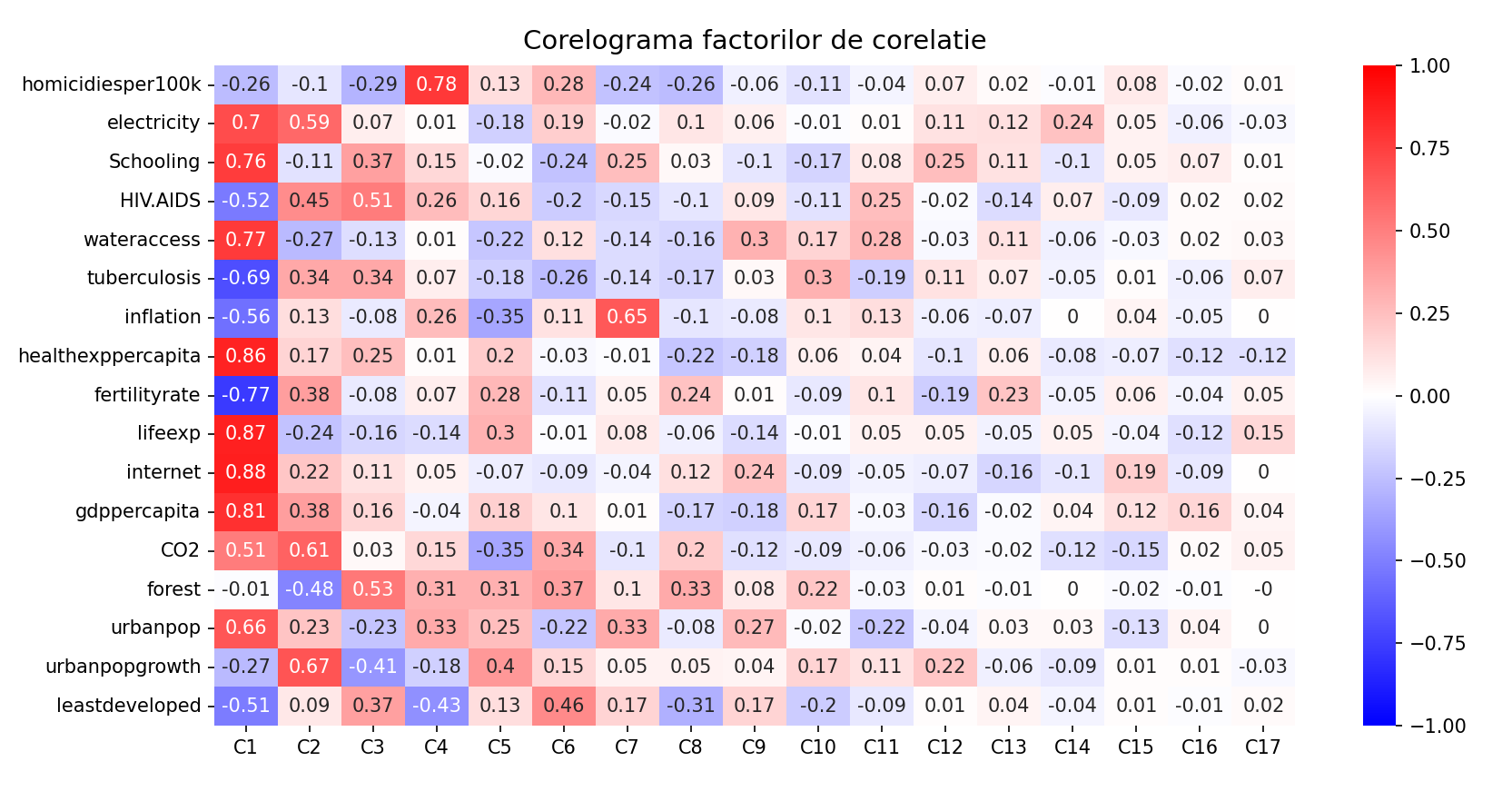


Fig.2: Corelograma factorilor de corelatie

Corelatiile sunt indispensabile pentru a incerca sa denominam componentele principale. Atribuie o semnificatie acestor componente principale. Denominarea mai este numita si etichetare, etichetele fiind legate de aspectele corelatiilor puternice dintre variabile si componentele respective.

Din corelograma de mai sus putem observa ca C1 este influentata in mod direct de costurile cu sanatatea, speranta de viata, internetul si gdp per capita, iar in mod indirect de fertilitate si tuberculoza. Dupa cele analizate putem atribui acestei componente eticheta “Cheltuieli”. Pentru C2, care este influentata cel mai mult de emisiile de dioxid de carbon, electricitate si de cresterea populatiei urbane o putem eticheta cu “Urban”.

Avem mai jos si cercul corelatiilor pentru primele doua componentele principale:

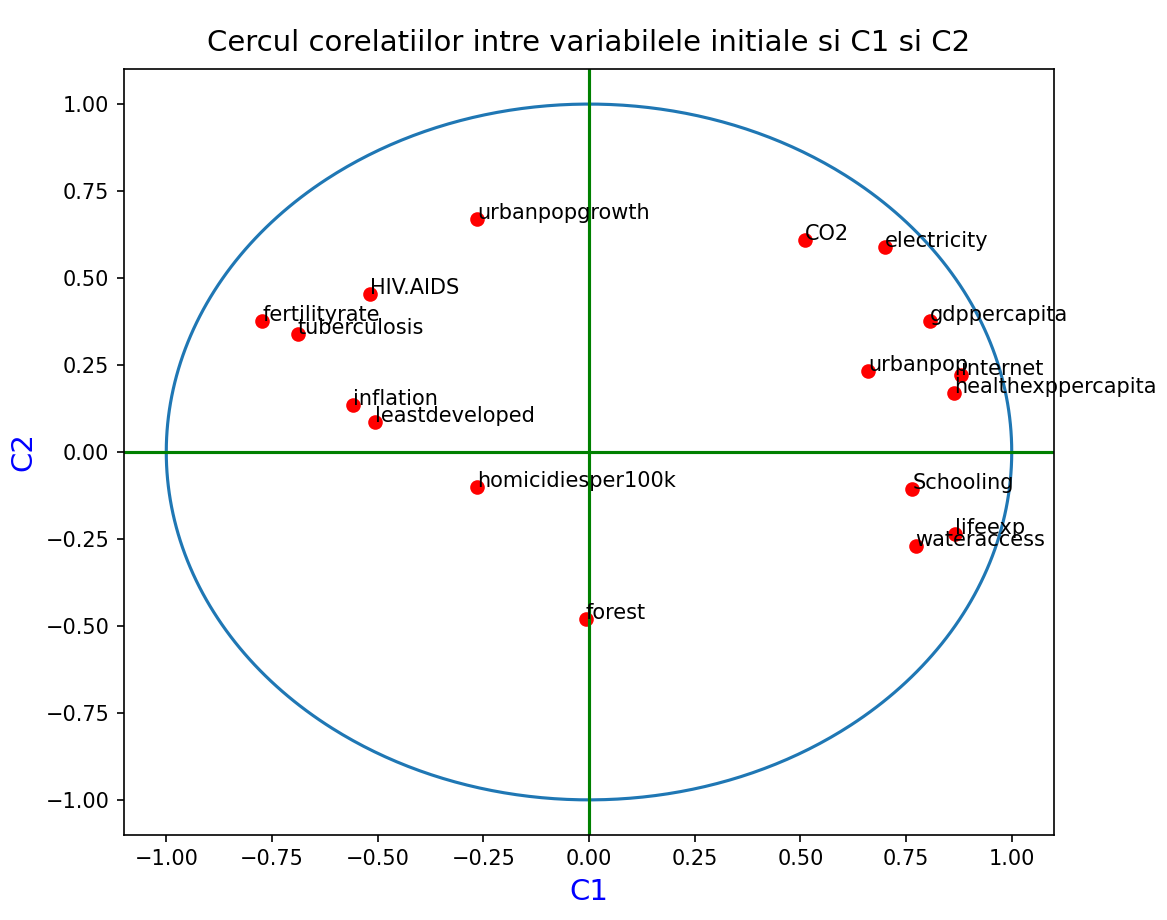


Fig.3: Cercul Corelatiilor

## **Calcul scoruri:**

Scorurile reprezinta componentele principale standardizate. Putem observa ca tara care are cea mai mare speranta de viata dupa primele doua componente principale este Singapore, urmata de Emiratele Arabe UniteSingapore, urmata de Emiratele Arabe Unite.

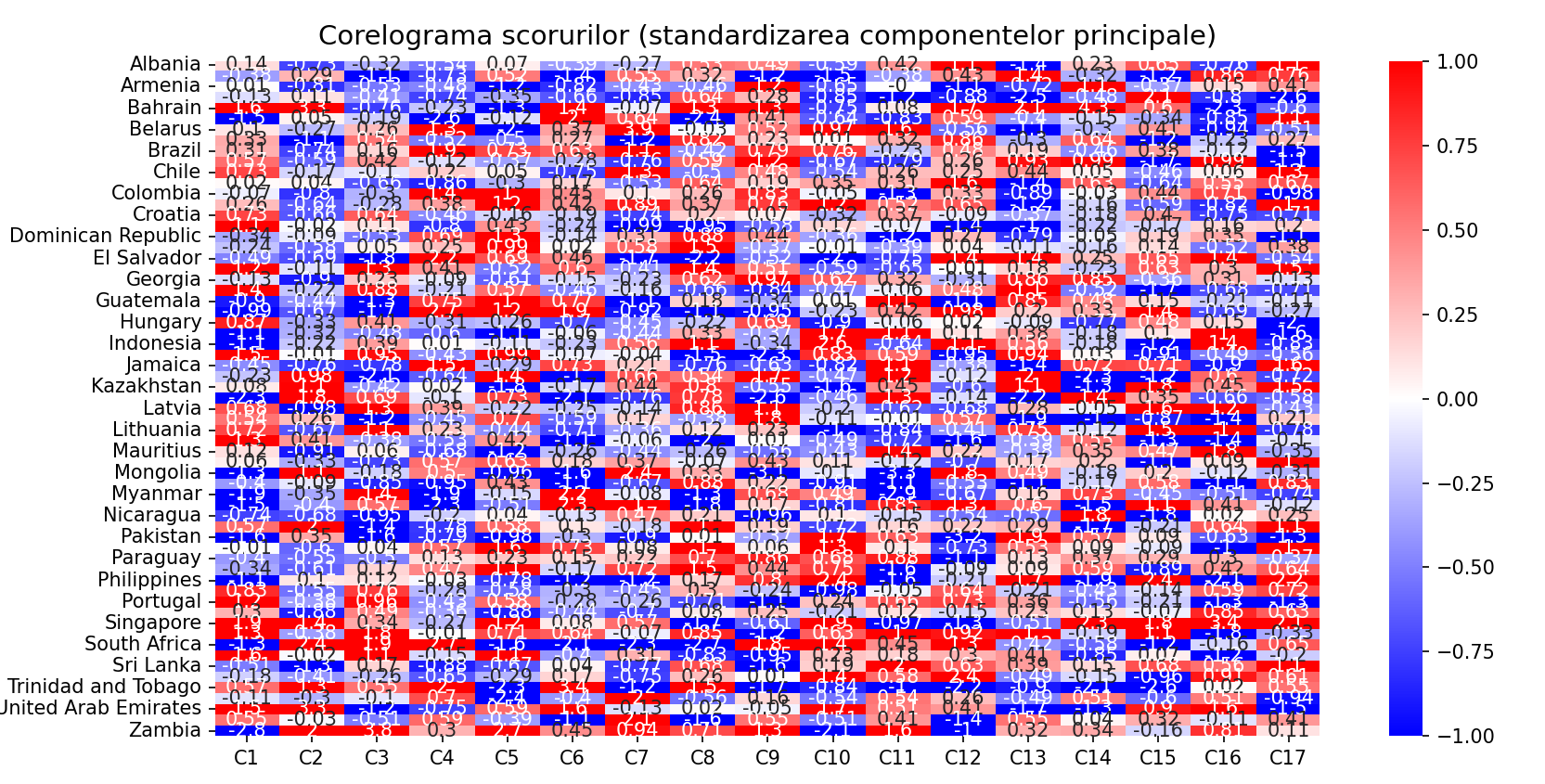


Fig.4: Corelograma scorurilor

## **Comunalitati**

Comunalitatile pun in evidenta transferul treptat de informatie de la variabilele studiate la componentele principale. Acestea sunt reprezentate prin corelograma atasata mai jos. Putem observa variabile cu legaturi puternice cu componente nu atat de semnificative.

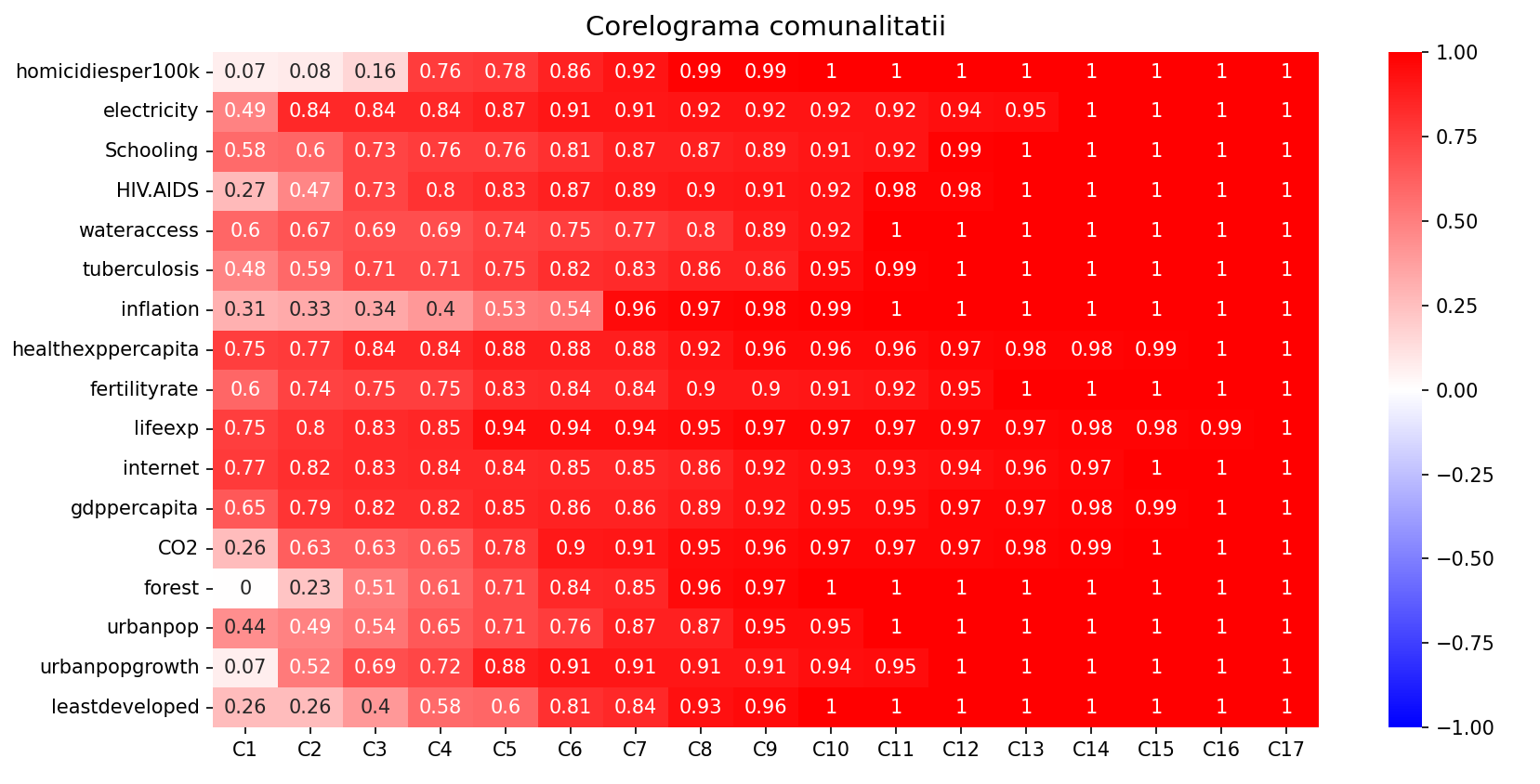
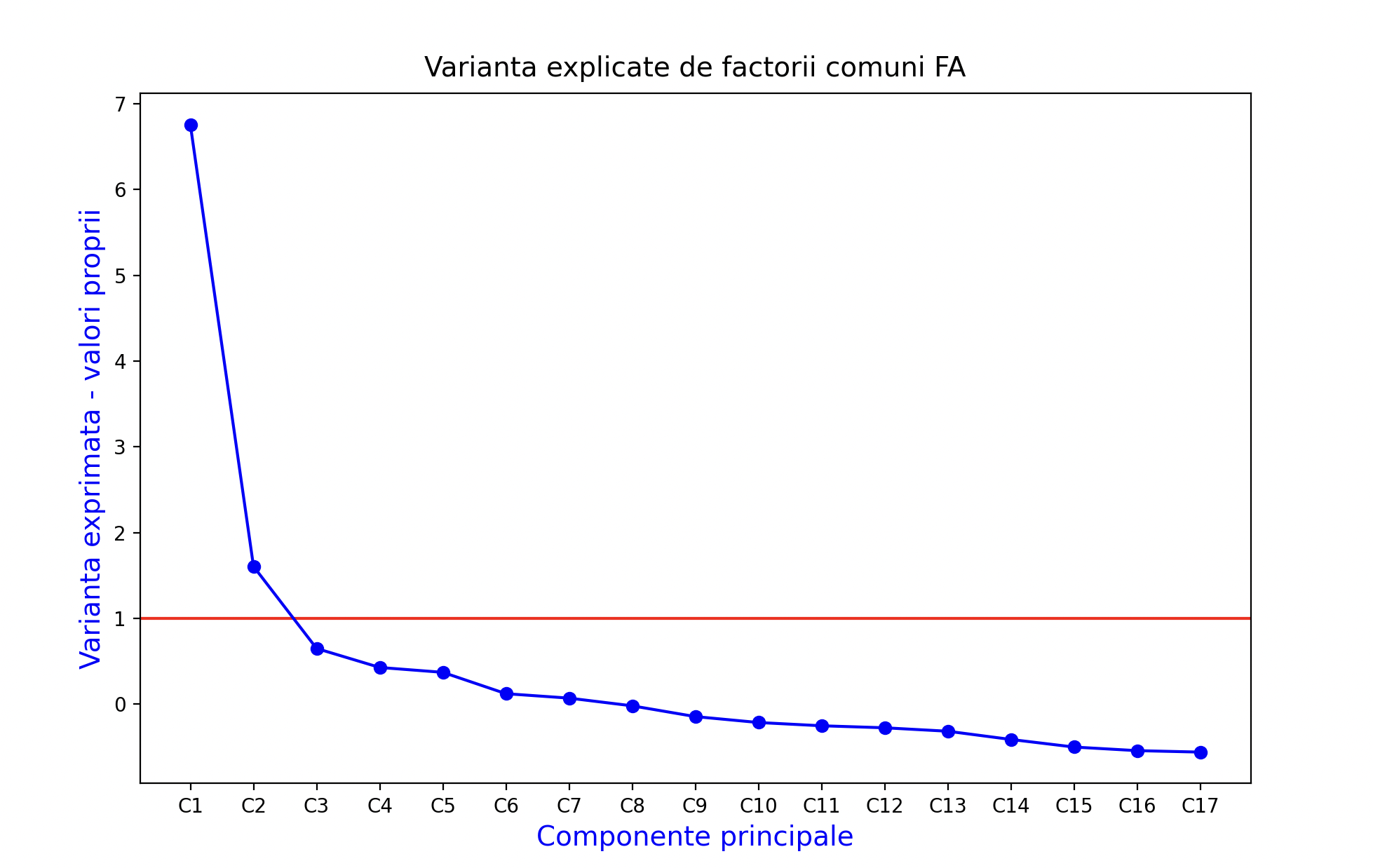


Fig.5: Corelograma comunalitatii

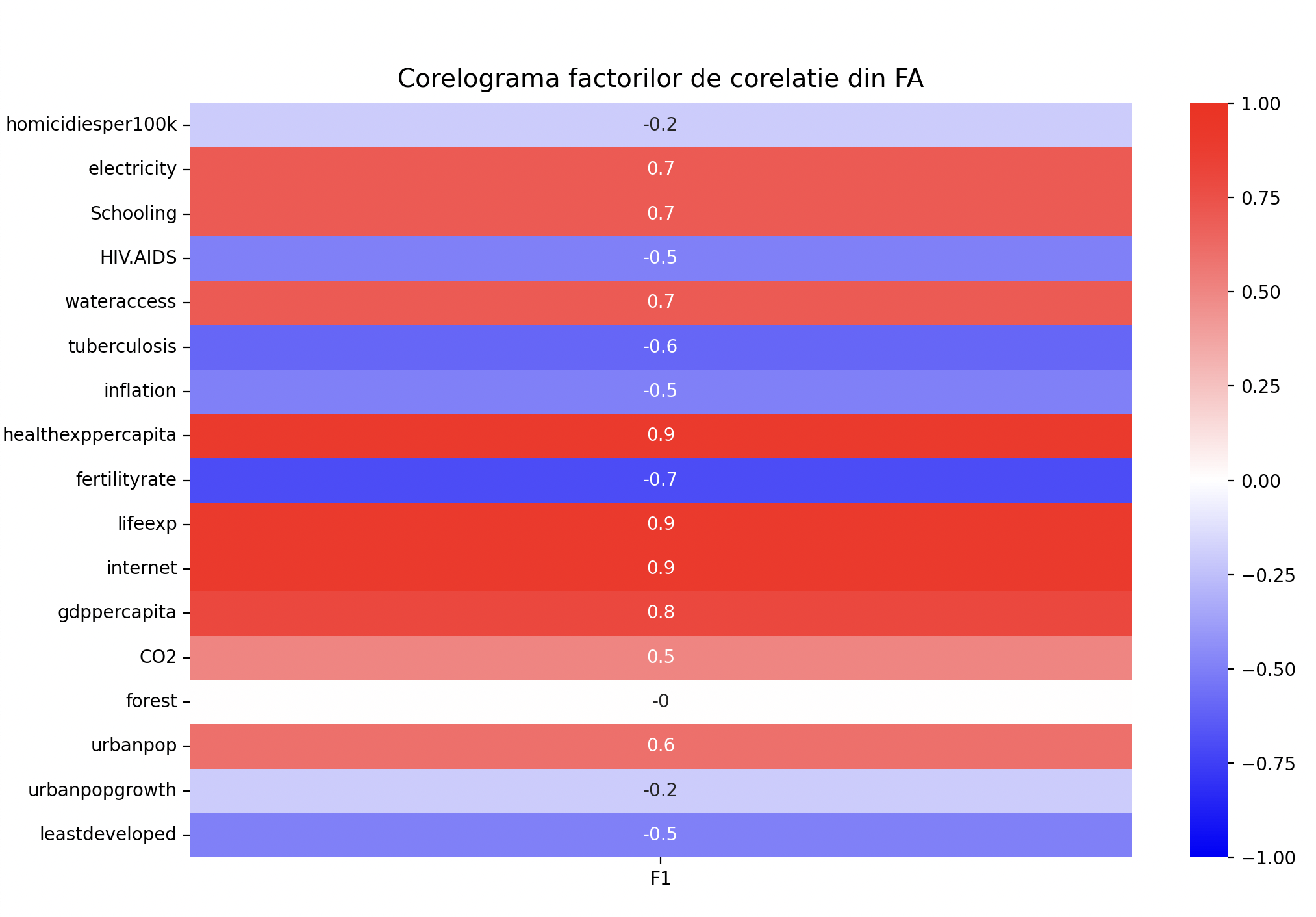
În procesul de analiză a datelor, am aplicat Analiza Factorială (FA) și Analiza Factorială de Echivalență (AEF) pentru a extrage informații semnificative și a identifica structuri subiacente ale setului de date.

* **Graficul Variantei Explicate de Factorii Comuni FA:**
  + Acest grafic evidențiază contribuția fiecărui factor comun la variația totală a datelor. Valorile mai mari indică influența semnificativă a factorilor în explicarea variației observate.



**Corelograma Factorilor de Corelație din FA:**

* + Diagrama prezintă corelațiile dintre variabilele inițiale și factorii comuni identificați prin Analiza Factorială. Acest aspect furnizează o perspectivă asupra relațiilor și influențelor dintre variabile.



**Corelograma Indicilor KMO (Kaiser-Meyer-Olkin):**

* + Această corelogramă ilustrează distribuția indicilor KMO pentru variabilele observate. Valorile ridicate indică adecvarea datelor pentru analiza factorială.

