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



**Proiect Dezvoltare Software pentru Analiza Datelor**

Student: Matei Adrian

**București, 2021**

Cuprins

[1. Sursa de date 3](#_Toc92672787)

[2. Referință de timp privind datele 3](#_Toc92672788)

[3. Descrierea variabilelor 3](#_Toc92672789)

[4. Descrierea observatiilor 5](#_Toc92672790)

[5. Abordarea privind analiza datelor 5](#_Toc92672791)

[6. Motivația de a opta pentru o anumită metodă de analiză a datelor 5](#_Toc92672792)

[7. Prezentarea rezultatelor și interpretarea acestora 6](#_Toc92672793)

[8. Discuții si concluzii 14](#_Toc92672794)

**Figuri**

[Figură 1. Varianța explicată de componentele principale 6](#_Toc92673217)

[Figură 2. Corelograma factorilor de corelație 7](#_Toc92673218)

[Figură 3. Corelograma scorurilor (componentele principale standardizate) 8](#_Toc92673219)

[Figură 4. Corelograma a calității reprezentării observațiilor pe axele componentelor principale 9](#_Toc92673220)

[Figură 5. Corelograma contribuției observațiilor la varianta axelor componetelor principale 10](#_Toc92673221)

[Figură 6. Corelograma comunități (regăsirea componentelor principale în variabilele inițiale) 11](#_Toc92673222)

[Figură 7. Cercul corelațiilor între variabilele inițiale și componentele C1 și C2 12](#_Toc92673223)

[Figură 8. Distribuția observațiilor în spațiul componentelor C1 și C2 13](#_Toc92673224)

# Sursa de date

Sursa de date este The World Bank, unde am filtrat variabilele în funcție de țările Uniunii Europene, indicatorii și perioada de timp dorită.

Link: [World Development Indicators | DataBank (worldbank.org)](https://databank.worldbank.org/reports.aspx?source=2&series=SP.POP.TOTL&country=)

# Referință de timp privind datele

Setul de date conține date din anul 2019, fiind cel mai recent an cu date complete pentru toate statele Uniunii Europene disponibile la sursa de date de mai sus. De asemenea, consider că 2019 este un an seminificativ, deoarece este anul dinaintea pandemiei, în care indicatorii selectați pentru analiză aveau valori neafectate de ceea ce avea să urmeze. Astfel, în viitor, odată cu finalul pandemiei, aceste date vor putea fi comparate cu datele de la acel moment pentru vizualizarea impactului pandemiei asupra idicatorilor selectați.

# Descrierea variabilelor

Am ales acești indicatori pentru că am considerat că sunt semnificativi în descrierea evoluției țărilor Uniunii Europene, fiind și în legătură strânsă unul cu celălalt prin prisma faptului că în opinia mea, descriu evoluția țărilor respective. De asemenea, este interesant de observat diferențele dintre țările din vestul și țările din estul UE. Consider că aceștia au o influență puternică unul asupra celuilalt, reușind să descrie poate motivele pentru care anumite țări se află la un anumit nivel față de celelalte.

1. ChEdu

* Cheltuieli pentru educație (milioane dolari)
* Cheltuielile cu educația se referă la cheltuielile curente de funcționare în educație, inclusiv salariile și excluzând investițiile de capital în clădiri și echipamente.

1. PIB

* PIB pe cap de locuitor (dolari)
* PIB-ul pe cap de locuitor este produsul intern brut împărțit la populație. PIB-ul este suma valorii adăugate brute de către toți producătorii rezidenți din economie plus orice taxe pe produse și minus orice subvenții care nu sunt incluse în valoarea produselor. Se calculează fără deduceri pentru deprecierea bunurilor fabricate sau pentru epuizarea și degradarea resurselor naturale.

1. PersNet

* Persoane care folosesc internetul (procent din populație)
* Utilizatorii de internet sunt persoane fizice care au folosit internetul (din orice locație) în ultimele 3 luni. Internetul poate fi utilizat prin intermediul unui computer, telefon mobil, asistent digital personal, aparat de jocuri, TV digital etc.

1. AngStudSup

* Forța de muncă cu studii superioare (procent din populația activă cu studii superioare)
* Raportul dintre forța de muncă cu studii avansate și populația în vârstă de muncă cu studii avansate. Învățământul avansat cuprinde învățământul terțiar cu ciclu scurt, o diplomă de licență sau un nivel de educație echivalent, o diplomă de master sau un nivel de educație echivalent sau o diplomă de doctorat sau un nivel de educație echivalent conform Clasificării Standarde Internaționale a Educației 2011 (ISCED 2011).

1. SV

* Speranța de viață (ani)
* Speranța de viață la naștere indică numărul de ani pe care un nou-născut ar trăi dacă tiparele predominante de mortalitate la momentul nașterii ar rămâne aceleași pe tot parcursul vieții.

1. RS

* Rata șomaj (procent din totalul forței de muncă)
* Șomajul se referă la ponderea forței de muncă fără muncă, dar disponibilă și în căutarea unui loc de muncă. Definițiile forței de muncă și ale șomajului diferă în funcție de țară.

1. PopUrb

* Populația urbană (persoane)
* Populația urbană se referă la persoanele care locuiesc în zonele urbane, așa cum sunt definite de birourile naționale de statistică. Acesta este calculat folosind estimările populației Băncii Mondiale și ratele urbane din Perspectivele de urbanizare mondială ale Națiunilor Unite.

1. RSinucid

* Rata mortalității prin sinucidere la 100.000 de locuitori
* Rata mortalității prin sinucidere este numărul de decese prin sinucidere într-un an la 100.000 de locuitori. Rata brută de sinucidere (neajustată în funcție de vârstă).

1. Pop

* Populația (persoane)
* Populația totală se bazează pe definiția de facto a populației, care numără toți rezidenții, indiferent de statutul juridic sau de cetățenie. Valorile afișate sunt estimări la mijlocul anului.

1. PopServSanBaz

* Persoane care utilizează cel puțin servicii de salubritate de bază (procent din populație)
* Procentul de persoane care utilizează cel puțin servicii de salubritate de bază, adică instalații de salubritate îmbunătățite care nu sunt partajate cu alte gospodării. Acest indicator include atât persoanele care utilizează servicii de salubritate de bază, cât și pe cei care utilizează servicii de salubrizare gestionate în siguranță. Instalațiile de salubritate îmbunătățite includ spălarea/turnarea în sistemele de canalizare, fose septice sau latrine cu gropi; latrine de groapă ventilate îmbunătățite, toalete compozite sau latrine de groapă cu plăci.

# Descrierea observatiilor

În ceea ce privesc observațiile, am ales cele 27 de țări ale Uniunii Europene, deoarece consider că este o zonă semnificativă din punct de vedere al evoluției umanității, dar și pentru că România face parte din UE, iar mare parte din țările de pe continentul Europa, pe care ne aflăm și noi, fac parte de asemenea din UE. Acestea sunt întâlnite în cadrul csv-ului sub forma denumirii complete a acestora.

# Abordarea privind analiza datelor

În cadrul analizării datelor îmi propun să observ în ce măsură există o diferență între țările din estul Uniunii Europene și cele din vestul UE.

De asemenea, voi analiza și modul în care educația influențează traiul oamenilor și evoluția unor indicatori economici precum PIB-ul pe cap de locuitor sau demografici precum populația, în special cea urbană versus cea rurală, ori rata șomajului și în general dezvoltarea unei țări.

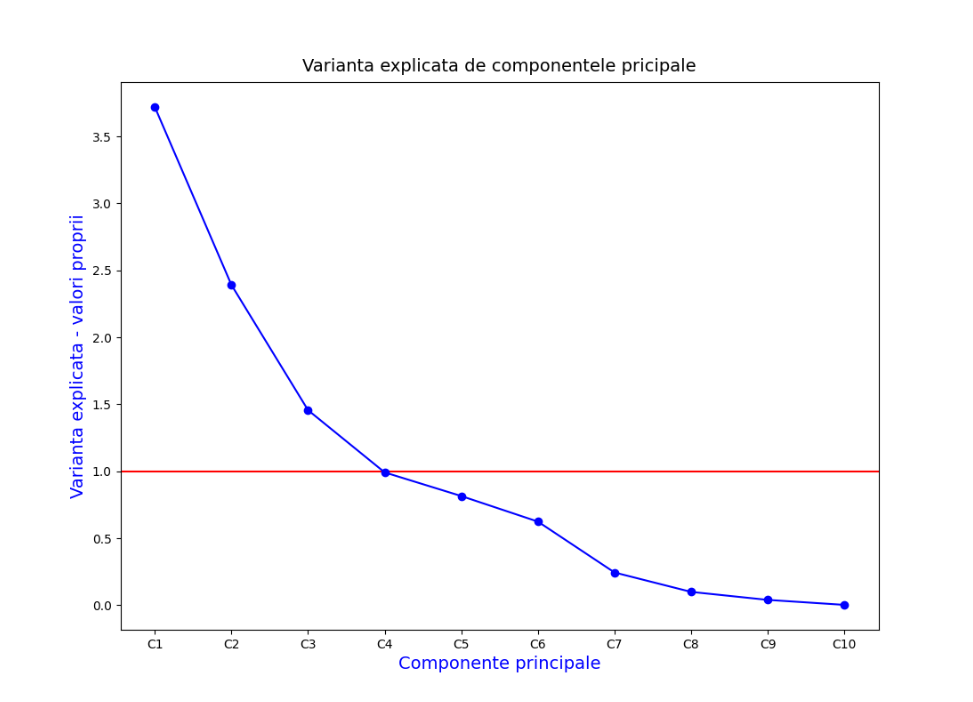
Totodata, este interesant de observat felul în care indicatorii economici influențează alți indicatori precum rata sinuciderilor sau a angajaților cu studii superiorare și a populației cu acces la cel putin salubritate de bază.

Pe de alta parte, voi avea în vedere și modul în care persoanele cu acces la internet se află în număr mare în zonele în care cheltuieli cu educația sunt mai mari sau în cele în care acestea sunt mai mici, astfel poate reieșind felul în care oamenii încearcă să se auto-educe sau poate doar să aibă o ocupație pentru a scăpa de griji.

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

Am optat pentru Analiza componentelor principale ca metodă de analiză a datelor, deoarece, pe lânga faptul că este metoda care mi-a atras cel mai mult atenția în cadrul seminarelor, mi se pare o modalitate interesantă de a analiza datele. Poate una dintre piedicile pe care le avem în momentul analizei unui set de date este dimensiunea si complexitatea acestuia, astfel, prin metoda ACP setul de date își reduce dimensiunea și permite o mai bună vizualizare a datelor, toate acestea realizându-se fără a pierde informații și a surprinde cât mai bine datele semnificative.

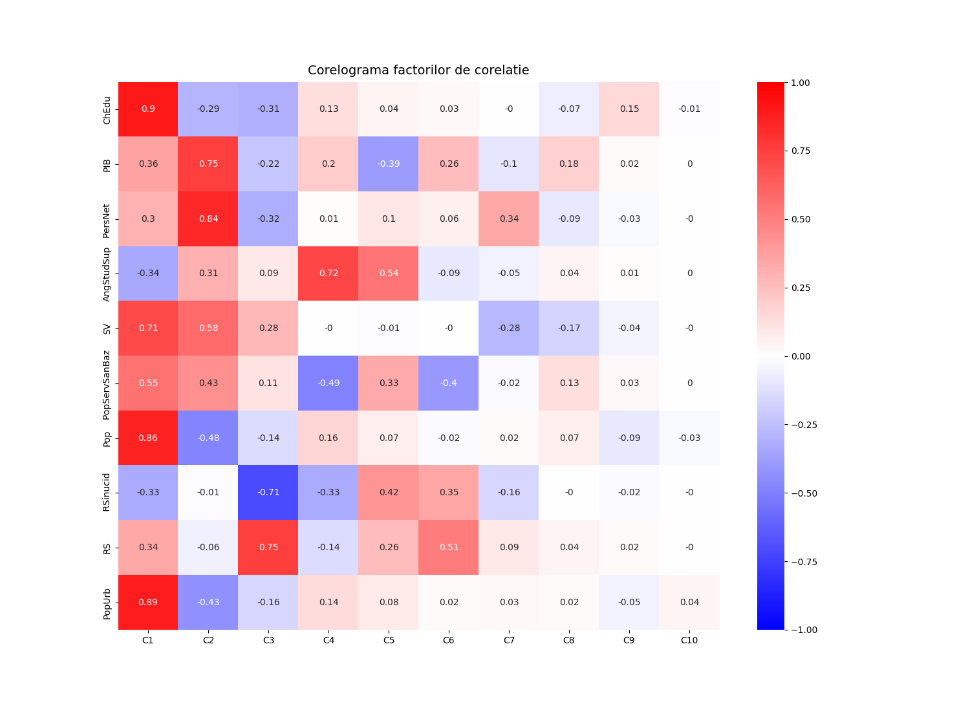
# Prezentarea rezultatelor și interpretarea acestora



Figură 1. Varianța explicată de componentele principale

În plot-ul de deasupra este vorba de valorile proprii reprezentate de varianța explicată de componentele principale. Componentele principale sunt construite sub formă de cobinație liniară de variabile inițiale, concentrând o cât mai mare parte din varianță. Astfel, prima componentă preia maximul din varianța variabilelor inițiale, a doua prea maximul de varianța rămasă excluzând-o pe prima, a treia preia maximul de varianță rămasa excluzându-le pe primele două, etc.

De asemenea, în urma obținerii componentelor și a varianței explicate de acestea, conform criteriului lui Kaiser, sunt semnificative doar componentele care au o varianță mai mare decât 1, din figură reieșind că este vorba despre primele 3 componente (C1, C2, C3).



Figură 2. Corelograma factorilor de corelație

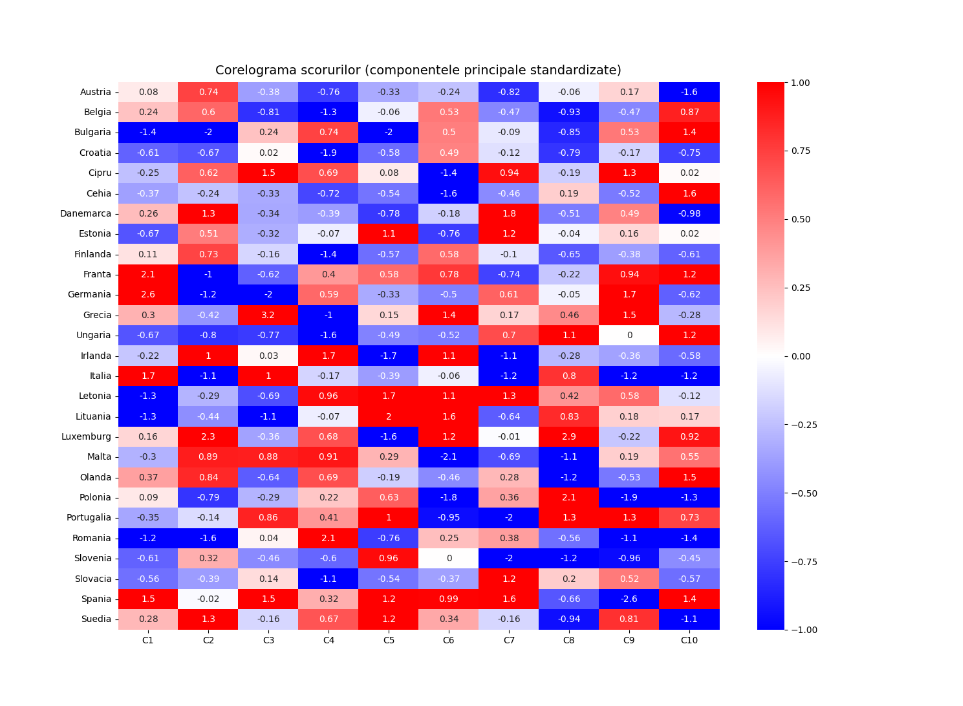
Matricea factorilor de corelație reprezentați de ceoficienții de corelație liniară dintre variabilele observate și componentele principale este una pătratică formată din 10 linii și 10 coloane. Numărul liniilor este reprezentat de numărul variabilelor observate, iar numărul coloanelor de numarul componentelor principale rezultate din acestea.

Factorii de corelație reprezentați în corelograma de mai sus ajută la etichetarea componentelor principale prin observarea a ceea ce au în comun variabilele puternic corelate cu fiecare componenta în parte.

În cazul primei componente, aceasta este puternic influențată pozitiv și în mod direct de cheltuielile cu educația, populația, populația urbana, speranța de viață și populația minim cu acces la servicii sanitare de baza, acestea având cele mai mari corelații și în mod indirect de rata sinuciderilor. Putem să-i atribuim eticheta de nivelul populatiei, în special a celei urbane, și a cheltuielilor cu educația, a speranței de viață și a populației cu acces la servicii sanitare de bază.

În cazul celei de a doua componente, PIB-ul pe cap de locuitor, numărul persoanelor cu abonament la internet și speranța de viață o influențează pozitiv, această componentă putând semnifica nivelul PIB-ului pe cap de locuitor, a numărului persoanelor cu abonament la internet și a speranței de viață a acestora.

A treia componentă este reprezentată de rata șomajului, deci poate desemna nivelul ratei șomajului din diferite țări.

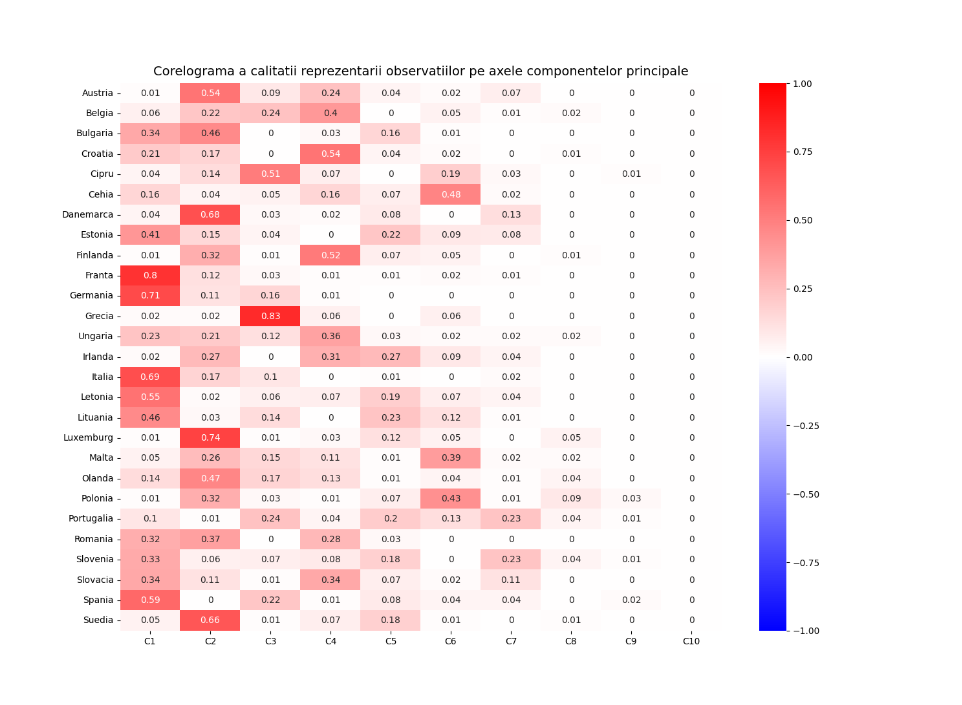


Figură 3. Corelograma scorurilor (componentele principale standardizate)

Scorurile reprezintă perspectiva componentelor principale în relația cu observațiile. Din corelograma de mai sus reiese faptul că țări precum Spania, Italia, Germania sau Franța au un nivel ridicat al populației, în special cea urbană și minim cu acces la servicii sanitare de bază, a cheltuielilor cu educația și în consecință a speranței de viață, la polul opus fiind țări precum România, Slovenia, Lituania, Letonia, Croația sau Bulgaria.

Țări precum Suedia, Olanda, Malta, Luxemburg, Irlanda, Finlanda, Danemarca sau Austria au un nivel ridicat al PIB-ului pe cap de locuitor, a numărului persoanelor cu abonament la internet și a speranței de viață, fiind de asemenea și țări cu populația mai redusă fața de cele din componenta 1.

De asemenea, Spania, Portugalia, Malta, Italia, Grecia și Cipru sunt țări cu un nivel ridicat în ceea ce privește rata șomajului.



Figură 4. Corelograma a calității reprezentării observațiilor pe axele componentelor principale

Calitatea reprezentării observațiilor pe axele componentelor principale reprezintă nivelul reprezentării observațiilor în noile axe de coordonate. Astfel, corelograma de mai sus reflectă nivelul repectării interpretării datelor pentru componentele principale în raport cu celelalte țări prezente.

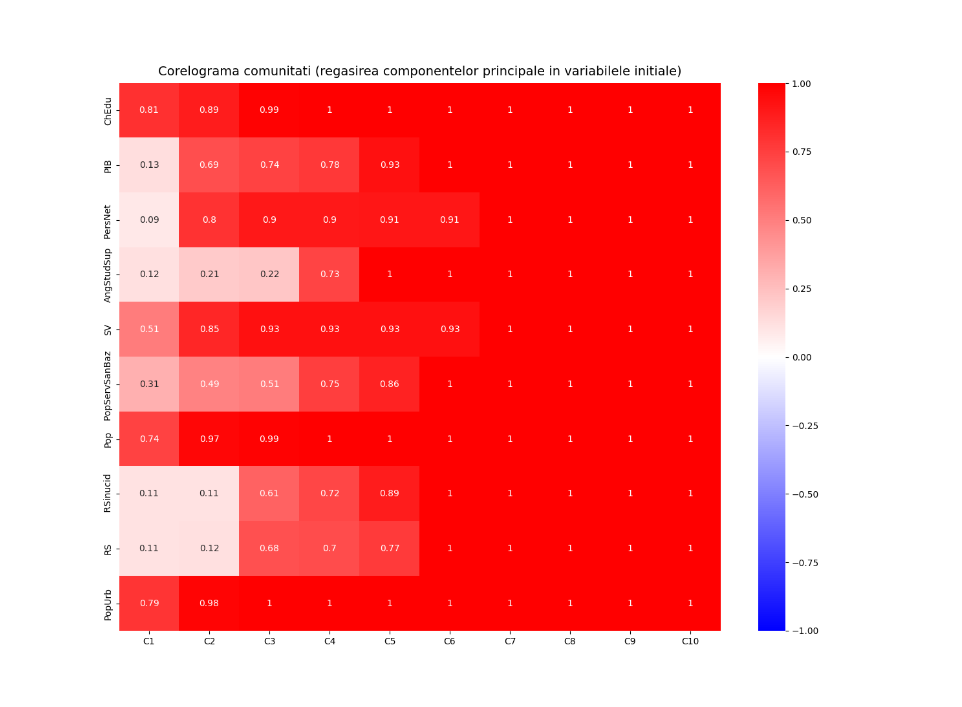
Din corelogramă reiese faptul că țări precum Spania, Italia, Germania sau Franța au un nivel ridicat al populației, în special cea urbană și minim cu acces la servicii sanitare de bază, a cheltuielilor cu educația și în consecință a speranței de viață, țările ca Suedia, Luxemburg, Danemarca sau Austria un nivel ridicat al PIB-ului pe cap de locuitor, a numărului persoanelor cu abonament la internet și a speranței de viață, iar Grecia sau Cipru au un nivel ridicat în ceea ce privește rata șomajului.



Figură 5. Corelograma contribuției observațiilor la varianta axelor componetelor principale

Contribuția observațiilor la varianța axelor componentelor principale arată importanța unei țări pentru o anumită componentă în raport cu celelalte observații.

Din corelograma de mai sus reiese faptul că pentru prima componentă principala Germania, Franța, Italia au cea mai mare contribuție, pentru cea de a doua componentă Luxemburg, și Bulgaria, iar pentru cea de a treia Grecia și Germania.

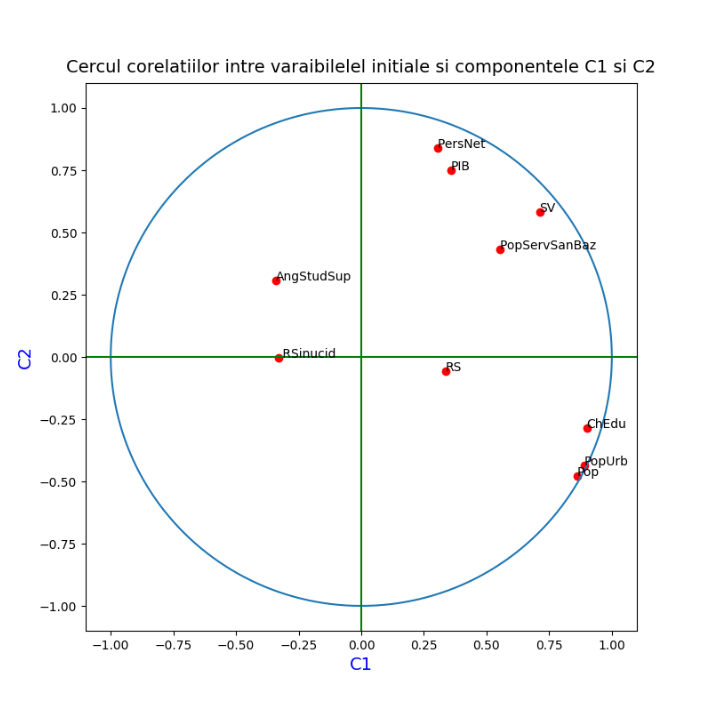


Figură 6. Corelograma comunități (regăsirea componentelor principale în variabilele inițiale)

Comunitățile reprezintă comunalitățile între variabilele inițiale și componentele principale. Din corelograma de mai sus reies componetele principale care se regăsesc în variabilele principale în functie de cantitatea de varianță explicată a fiecărei componente.

Astfel, în cazul primei componente este explicată în special de populația urbană, speranța de viața și cheltuielile cu educația. A doua componentă este explicată de populația urbană, populație, speranța de viață, numărul persoanelor cu abonament la internet, PIB ul pe cap de locuitor și cheltuielile cu educația. Cea de a treia componentă este explicată de populația urbanî, rata șomajului, rata sinuciderilor, populația, speranța de viața, PIB-ul pe cap de locuitor, și cheltuielile cu educația.

De asemenea, se pot observa și variabilele care au o legătură mai puternică cu componentele mai puțin semnificative precum rata șomajului, rata sinuciderii sau numărul angajaților cu studii superioare.



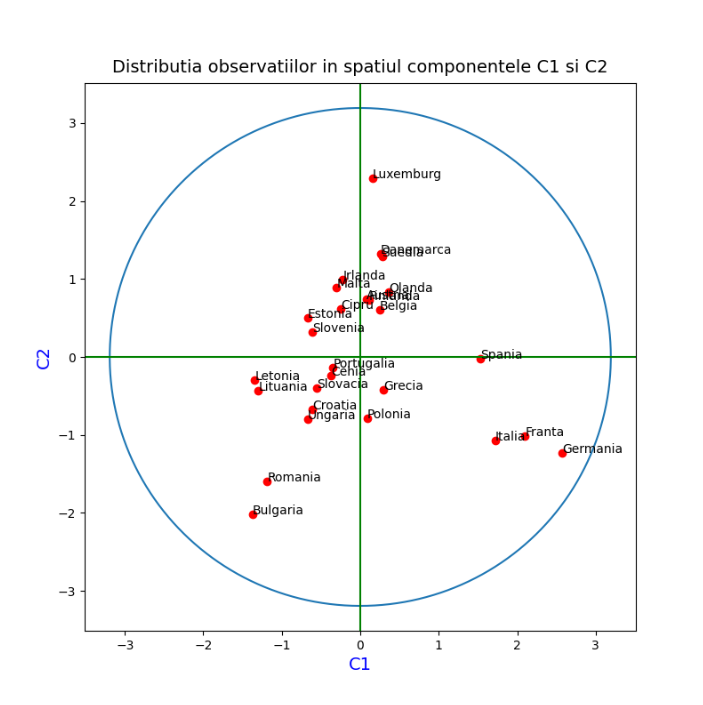
Figură 7. Cercul corelațiilor între variabilele inițiale și componentele C1 și C2

În cazul cercului corelațiilor este utilizată matricea de corelație, figura de mai sus realizându-se în cazul primelor două componente: C1 (nivelul populației, în special a celei urbane, și a cheltuielilor cu educația, a speranței de viață și a populației minim cu acces la servicii sanitare de bază) și C2 (nivelul PIB-ului pe cap de locuitor, a numărului persoanelor cu abonament la internet și a speranței de viață a acestora).

Astfel, numărul persoanelor care au acces la internet, PIB-ul pe cap de locuitor, speranța de viață și populația minim cu acces la servicii sanitare de bază sunt puternic corelate atât cu C1, cât și cu C2.

Rata șomajului, cheltuielile cu educația, populația urbană și populația sunt puternic corelate cu C2, dar au valori negative în ceea ce privește corelația cu C1.

În cazul ratei sinuciderilor atât corelația cu C2, cât si cea cu C1 are valoare nagativă, fiind totuși mai puternică față de C1, iar numărul angajaților cu studii superioare este puternic corelat cu C1, având valoare negativă față de C2.



Figură 8. Distribuția observațiilor în spațiul componentelor C1 și C2

Distribuția observațiilor în spațiul componentelor principale reprezintă legatura dintre țările a căror variabile au fost analizate și componentele principale, în cazul figurii de mai sus fiind reprezentată legătura cu primele două componente.

Astfel, țările precum Luxemburg, Danemarca, Suedia, Olanda, Austria, Belgia și Finlanda sunt asociate pozitiv ambelor componente principale C1 (nivelul populației, în special a celei urbane, și a cheltuielilor cu educația, a speranței de viață și a populației minim cu acces la servicii sanitare de bază), respectiv C2 (nivelul PIB-ului pe cap de locuitor, a numărului persoanelor cu abonament la internet și a speranței de viață a acestora).

În cazul Spaniei, Italiei, Franței, Germaniei, Greciei și Poloniei, sunt asociate pozitiv față de C1, dar negativ față de C2.

Irlanda, Malta, Cipru, Estonia, Slovenia sunt asociate pozitiv cu C1, dar negativ cu C2, deși valorile sunt puțin sub axa lui C1.

Nu în ultimul rand, țări precum Portugalia, Cehia, Slovacia, Ungaria, Letonia, Lituania, România și Bulgaria sunt asociate negativ atât cu C1, cât și cu C2, nivelul asocierii scăzând conform enumerării anterioare.

# Discuții si concluzii

În concluzie, deși în cazul țărilor din vest puternic dezvoltate, nu toți indicatorii cu importanță în dezvoltare au o influență mare sau valori ridicate. Se poate observa că majoritatea acestora se bucură de un nivel al populației crescut, inclusiv a celei urbane, al cheltuielilor cu educația, a speranței de viață și a populație minim cu acces la servicii sanitare de bază, inclusiv țările mai mici, cu o populație mai mică bucurându-se de un nivel al PIB-ului pe cap de locuitor ridicat, a numărului de abonamente la internet ridicat și de asemenea speranță de viață ridicatș.

În schimb, țări precum România, Bulgaria, Ungaria, Letonia, Lituania, Slovacia, țări din estul europei și foste țări comuniste sunt asociate în mod negativ cu cele mai semnificative componente principale, în special cu primele 2 așa cum reiese și din figura distribuției observațiilor în spațiul componentelor principale.

De asemenea, țările care au fost puternic asociate cu componentele semnificative au un nivel ridicat al speranței de viață, ceea ce poate să însemne că dezvoltarea acestora influențează semnificativ viețile oamenilor, aceștia ducând probabil un trai mai bun și reușind să-și sporească speranța de viață.

Astfel, în urma analizei efectuate, reiese faptul că ipoteza inițială este în mare parte confirmată.