TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Creació d'un indicador sintètic de desenvolupament mitjançant anàlisis factorial exploratori

Autor: Gemma Vendrell Tarrés

Director: Javier Sierra Martínez

Departament: Econometria, Estadística i Economia aplicada

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUM

L'objectiu principal del treball consisteix en la creació d'un indicador sintètic que resumeixi la informació continguda en diferents indicadors parcials. Aquest indicador servirà per observar el posicionament dels països Europeus pel que fa al seu benestar.

ABSTRACT

The main objective of the work consists in the creation of a synthetic indicator that summarizes the information contained in diferent partial indicators. This indicator will be used to observe the position of European countries in terms of their well-being

PARAULES CLAU

- Indicador: mesura o variable utilitzada per avaluar o quantificar una determinada condició, fenomen o situació

- Variància: mesura que quantifica la dispersió o variabilitat d'un conjunt de dades respecte la seva mitjana.

- Homoscedasticitat: concepte estadístic que indica que la variància dels errors o residus en un model de regressió és constant en tots els nivells de les variables independents

- Correlació: mesura que indica el grau de relació entre les variables i proporciona informació sobre la seva direcció i força.

- Clúster: es refereix a un grup o agrupament de casos, objectes o elements similars o relacionats entre si.

CLASSIFICACIÓ AMS

- 62-09 Graphical methods
- 62H20 Measures of association (correlation, canonical correlation, etc.)
- 62H25 Factor analysis and principal components; correspondence analysis
- 62H30 Classification and discrimination; cluster analysis

ÍNDEX

1. Introducció 4
2. Metodologia 6
   2.1. Tècniques estadístiques emprades 6
   2.2. Procedència de les dades 6
   2.3. Recursos informàtics 6
3. Marc teòric 8
   3.1. Què és un indicador? 8
   3.2. Indicador sintètic 10
      3.2.1. Avantatges 10
      3.2.2. Desavantatges 10
      3.2.3. Mètodes 11
   3.3. Metodologia de creació d'un indicador sintètic 15
4. Desenvolupament d'un marc conceptual 16
5. Selecció de les variables 17
6. Anàlisi descriptiva 20
7. Imputació de dades 24
8. Normalització de les dades 26
9. Anàlisi multivariant 28
   9.1. Anàlisi clúster 28
   9.2. Anàlisi factorial exploratori 31
10. Agregació de la informació 44
11. Conclusions 49
12. Referències bibliogràfiques 51
13. Annex 52

ÍNDEX IL·LUSTRACIONS

Il·lustració 1: Matriu de correlació Font: Elaboració pròpia 21
Il·lustració 2: Boxplot distàncies de Mahalanobis Font: Elaboració pròpia 22
Il·lustració 3: K-MEANS Font: Elaboració pròpia 30
Il·lustració 4: Mapa dels clústers Font: Elaboració pròpia 31
Il·lustració 5: Gràfic matriu de correlació Font: Elaboració pròpia 34
Il·lustració 6: Scree plot Font: Elaboració pròpia 37
Il·lustració 7: Parallel analysis scree plot Font: Elaboració pròpia 38
Il·lustració 8: Factor exploratori ortogonal 38
Il·lustració 9: Factor exploratori obliqua 39
Il·lustració 10: Gràfic biplot sense rotació Font: Elaboració pròpia 40
Il·lustració 11: Gràfic biplot rotació varimax Font: Elaboració pròpia 41
Il·lustració 12: Gràfic biplot rotació varimax Font: Elaboració pròpia 41
Il·lustració 13: Diagrama dels factors Font: Elaboració pròpia 43
Il·lustració 14: Mapa Indicador Font: Elaboració pròpia 48

ÍNDEX TAULES

Taula 1: Països Inclosos a L'estudi Font: Elaboració pròpia 19
Taula 2: Anàlisi descriptiu Font: Elaboració pròpia 20
Taula 3: Distàncies de Mahalanobis Font: Elaboració pròpia 23
Taula 4: Resum dels NA's per variable Font: Elaboració pròpia 25
Taula 5: Comparació anàlisi descriptiu Font: Elaboració pròpia 26
Taula 6: Gràfic mètode del colze Font: Elaboració pròpia 30
Taula 7: Resum ponderacions Font: Elaboració pròpia 46
Taula 8: Resum indicador Font:Elaboració pròpia 47

1. Introducció

En el context de la creixent disponibilitat de dades i la necessitat d'avaluar i treure'n informació, els indicadors sintètics s'han convertit en eines molt valuoses per capturar la complexitat i multidimensionalitat dels fenòmens d'estudi, aquests indicadors proporcionen una visió més completa i comprensiva en poder combinar informació de diferents fonts i variables, per aquest motiu em va semblar interessant realitzar el meu treball de fi de grau sobre aquest tema.

Aquest treball es centra en el desenvolupament d'un indicador sintètic que permeti avaluar el benestar en un país. Concretament he triat comparar l'indicador per països d'Europa ja que penso que en ser una regió bastant gran hi ha força varietat de països, des dels més petits als més grans, dels més pobres als més rics i així em permetrà aconseguir bons resultats.
Dins del procés de creació d'un indicador ens trobem amb la necessitat d'emprar eines estadístiques com per exemple l'anàlisi factorial, amb la finalitat d'identificar i seleccionar els factors que ens permeten simplificar la informació inicial, o l'anàlisi clúster per identificar diferents patrons i perfils dins de la mostra ajudant així a comprendre les característiques i el comportament entre els diferents països. A més a més d'aquestes tècniques també s'ha hagut d'utilitzar processos de normalització i de tractament de valors absents, per així garantir la qualitat dels resultats.

En resum, aquest treball s'enfoca en la construcció d'un indicador sintètic mitjançant la integració d'indicadors parcials per països pertanyents a Europa, amb la finalitat d'aprendre profundament el seu procediment de construcció i cada una de les eines utilitzades. Els objectius principals són els següents:

1) Creació d'un indicador sintètic que resumeixi la informació que continguin altres indicadors parcials, amb l'objectiu que aquest indicador serveixi per observar tant l'evolució com el posicionament dels països de la Unió Europea pel que fa al benestar.
2) Profunditzar en diferents eines estadístiques coneixent així la seva utilització i el seu funcionament.
3) Presentar les conclusions derivades de l'estudi per la seva anàlisis i discussió.
4) Familiaritzar-se amb les funcionalitats i característiques que ens ofereix el llenguatge R.

En una primera etapa el treball consta d'un marc més teòric on s'explica què és un indicador sintètic, de quins tipus n'hi ha i els seus avantatges i inconvenients. Seguidament s'exposa quina és la metodologia a seguir per la seva creació i es duu a terme la creació de l'indicador amb les dades escollides, al final s'hi afegiran les referències consultades i a l'annex el codi en R escrit durant la realització del treball.

2. Metodologia

2.1. Tècniques estadístiques emprades

Per assolir els objectius esmentats s'utilitzaran diverses tècniques d'anàlisi multivariant, principalment un model factorial que ens servirà per reduir la quantitat de dades i l'anàlisi clúster per comprendre si hi ha alguns països amb característiques similars. També s'han utilitzat tècniques de normalització i d'imputació de dades, totes les eines esmentades s'expliquen de forma més detallada al llarg de l'informe.

2.2. Procedència de les dades

Les dades utilitzades en aquest treball s'han extret de la pàgina web oficial de l'Euroestat, l'oficina d'estadística de la Unió Europea. Euroestat és una font fiable i reconeguda per proporcionar estadístiques oficials sobre diversos aspectes socioeconòmics i demogràfics de la Unió Europea i els seus estats. Per accedir a les dades, s'ha utilitzat l'eina d'extracció i descàrrega de dades disponible a la mateixa pàgina web i s'han fet servir eines d'Excel i R per tal de poder obtenir una base de dades adequada per l'estudi.

Cal destacar que les dades proporcionades per l'Eurostat són recopilades directament de fonts oficials, com ara institucions nacionals d'estadística, organitzacions internacionals i altres fonts fiables. Aquestes dades són recopilades amb mètodes estandarditzats i subjectes a procediments de control de qualitat per assegurar la seva fiabilitat i precisió. Per dur a terme el treball s'ha pres precaució per assegurar que les dades seleccionades i utilitzades són les més actualitzades i representen de forma precisa el fenòmen estudiat.

2.3. Recursos informàtics

Per l'elaboració del treball s'han utilitzat els següents recursos informàtics:

- L'Excel, que es una aplicació desenvolupada per Microsoft que s'utilitza per a la gestió de dades, càlculs i anàlisis de dades mitjançant fulls de càlcul, permet a l'usuari organitzar i manipular les dades de manera estructurada, és molt útil gràcies a la quantitat de funcions que té i que permet fer de manera àgil, eficaç i senzilla.

- El Word, també és una aplicació de Microsoft i ens permet crear documents com per exemple informes, cartes, currículums... amb un ventall molt ampli d' opcions de format i organització del contingut del document, s'hi poden inserir taules, imatges i

gràfics de manera molt fàcil i per aquest motiu s'ha tornat indispensable per als professionals i els estudiants.

- R, és un llenguatge de programació per a l'anàlisi estadístic i la manipulació de dades. Aquest programa ens proporciona una àmplia gamma de funcions i paquets que permeten la importació, manipulació, visualització i modelització de dades.

3. Marc teòric

3.1. Què és un indicador?

Un indicador econòmic és un tipus de dada econòmica de caràcter estadístic (Coll Morales, 2020). Els indicadors són eines clau per mesurar i avaluar el desenvolupament econòmic d'un país, regió o empresa. Aquestes mesures ajuden als analistes a prendre decisions estratègies. En l'àmbit de l'economia ens ajuden a mesurar el seu creixement, n'és un exemple el PIB que mesura la producció econòmica d'un país i la seva capacitat de generar ingressos. Aquests indicadors ajuden a mesurar el creixement d'un país al llarg dels anys i a comparar-lo amb altres països. Una altra raó important dels indicadors és avaluar l'estabilitat financera, per exemple l'IPC i la taxa d'atur ens ajuden a avaluar la capacitat que té el país de mantenir una economia estable. També ens poden ajudar a predir tendències futures en l'economia, així, un creixement en la taxa d'ocupació ens pot indicar una expansió econòmica i un creixement de la taxa d'inflació pot indicar una possible recessió.

En funció del que mesura un indicador podem realitzar diverses classificacions, podem classificar els indicadors en funció del temps de reacció, la tendència i l'àmbit econòmic.
En funció del seu temps de reacció els podem classificar de la següent forma:

- Indicador econòmic endarrerit: Reflexa els canvis en l'economia després que hagin ocorregut. Això significa que els canvis econòmics que s'observen ja han succeït en el passat i per tant no serveixen per fer prediccions futures. Un exemple seria la taxa de desocupació, es mesura un cop els empleats ja han perdut la seva feina. En general aquests indicadors són útils per avaluar la salut econòmica d'un país en el passat, però no son tan útils per predir canvis en el futur.

- Indicador econòmic de cicle o coincident: És aquell indicador que pateix modificacions en el seu valor al mateix temps que ho fa l'economia, s'utilitza per plasmar l'estat actual de l'economia i veure com està funcionant en temps real. Per exemple ho seria la producció industrial que mesura la producció de béns i serveis en un país, aquest indicador mostra l'activitat econòmica present i la seva tendència.

- Indicador econòmic avançat: És aquell indicador que experimenta modificacions abans que aquests s'hagin materialitzat en l'economia. Aquest tipus d'indicadors ens permeten predir els canvis futurs en l'economia. Seria un exemple d'aquest indicador l'índex de preus de les accions, és un indicador avançat que ens indica les expectatives dels inversors sobre l'activitat futura de les empreses i l'economia.

Segons la seva tendència els podem classificar en tres tipus:

- Indicador acíclic: No existeix correlació entre la seva evolució i l'evolució de l'economia, és a dir no està relacionat amb el cicle econòmic d'expansió i contracció d'una economia. Aquests tipus d'indicadors es mantenen mes o menys constants independentment de l'estat de l'economia. Un exemple seria la població en edat de treballar.

- Indicador contracíclic: Són indicadors que es mouen en sentit contrari a la tendència econòmica general. Si l'economia decreix aquest indicador experimentarà una tendència a l'alça, manté una correlació inversa.

- Indicador procíclic: El seu comportament va en línia amb el cicle econòmic. Existeix una estreta correlació entre la evolució i l'indicador, en serien algun exemple la producció industrial, l'ocupació… durant una fase d'expansió econòmica la producció industrial augmenta, es creen més llocs de treball...

En funció de l'àmbit econòmic podem classificar els indicadors de la següent forma:

- Indicadors del mercat de treball: com es el cas de la taxa de desocupació, la població activa...
- Indicadors de la situació econòmica i el creixement econòmic: com per exemple el producte interior brut (PIB).
- Indicador de preus i poder adquisitiu: n'és un exemple l'índex de preus al consumidor (IPC), la inflació...
- Indicadors financers i d'estat de comptes: el ROE, el ROI, el TIR en són alguns exemples.
- Indicador de les operacions comercials amb l'exterior: la balança comercial, la balança de pagaments...

3.2. Indicador sintètic

L'objectiu d'un indicador sintètic és resumir un concepte multidimensional en un únic índex simple en base a un model conceptual. Es pot definir com una funció d'una o més variables que mesuren una característica o un atribut dels individus d'estudi (Schuschny y Soto, 2009). Pot ser tant de forma quantitativa com qualitativa.

En aquest treball es definirà un indicador sintètic al que es construeix com una funció de dos o més variables. Abans de poder construir un indicador d'aquestes característiques s'ha de saber quina és la definició que es vol descriure, en el nostre cas es vol crear un indicador de qualitat de vida. Aquests indicadors normalment es creen amb l'objectiu de mesurar el desenvolupament d'una unitat d'anàlisi. La unitat d'anàlisi fa referència als diferents països de la unió europea. És necessari tenir clar que podem obtenir informació oficial i confiable per fer el nostre estudi, per fer aquest treball s'han utilitzat un seguit de variables la informació de les quals s'ha extret de la pàgina oficial de l'Eurostat i per tant podem confirmar que és verídica.

3.2.1. Avantatges

Un punt a favor d'utilitzar aquest tipus d'indicadors es que permeten reduir la informació que prové d'un concepte multidimensional i que d'alguna altra forma podria ser difícil d'interpretar. Serveixen per integrar i resumir diferents dimensions sobre un tema, a més a més són fàcils d'interpretar per la seva capacitat de síntesi i es poden comparar de forma senzilla entre ells i la seva evolució.

3.2.2. Desavantatges

Si l'indicador està mal construït o mal interpretat podem generar informació no robusta i confusa, per això durant la seva construcció s'ha de considerar fer anàlisis de sensibilitat i robustesa. També en reduir un tema complexe en un sol valor pot donar lloc a biaixos o simplificació excessiva. Per aquests motius una manera per intentar evitar aquestes limitacions és calcular sub-indicadors que representin el comportament dels diferents subsistemes dels quals està composta la representació que volem estudiar. A més d'això s'ha de disposar de les dades necessàries pel nostre estudi i que els seus temps de mesura siguin els adequats.

3.2.3. Mètodes

Els procediments per obtenir els indicadors sintètics es diferencien principalment per la manera en la que es ponderen i s'agreguen els indicadors al sistema inicial. Hi ha diferents maneres d'obtenir un indicador sintètic, en veurem alguns exemples (Domínguez Serrano et al., 2011):

- Agregacions simples:

Una part important dels treballs sobre la construcció d'indicadors sintètics utilitza projeccions lineals unidimensionals, que generen mitjanes ponderades dels indicadors simples, i només es diferencien per la manera com s'han normalitzat els subindicadors. Es considera que tots els subindicadors utilitzen la mateixa unitat de mesura. Les diferències en les tècniques de normalització, la manera de ponderar aquest subindicadors donen lloc a diferents mètodes.
La metodologia més aplicada en investigacions empíriques és aquella que proposa el mateix pes per a tots els subindicadors, la seva dificultat és baixa i molt fàcil d'interpretar.
Consisteix en assignar el mateix pes a tots els indicadors i agregar la informació en forma de suma.
Així l'indicador per una unitat i es defineix com :

$$IS_i = w \cdot IN_{i1} + w \cdot IN_{i2} + \cdots + w \cdot IN_{ik} = \sum_{j=1}^{k} w \times IN_{ij}$$

on:

- w és el pes assignat als indicadors, normalment $w = \frac{1}{k}$, sent k el nombre de subindicadors
- IN_{ij} és el valor normalitzat de l'indicador *j* per la unitat *i*

Aquest procediment té alguns inconvenients que s'han de tenir en compte a l'hora d'analitzar. Primer de tot el grau de complexitat es concentra en la forma en la que els indicadors s'agrupen. A més a més imaginant el cas que tinguéssim un conjunt d'indicadors inicials *k* que es divideix en dos grans dimensions, k_1 i k_2, amb $k_1 > k_2$, de manera que realitzem una agregació en dues fases, en assignar el mateix pes a cada fase d'agregació, el pes final que obtindrien els indicadors de cada dimensió serien diferents:

$$w_1 = \frac{1}{2} \times \frac{1}{k_1}, w_2 = \frac{1}{2} \times \frac{1}{k_2}, per\ tant\ w_1 < w_2$$

D'aquesta forma els indicadors que tindran un pes menor en l'indicador sintètic final seran aquells del grup *k* on hi hagi major número d'indicadors.

- Mètodes participatius:

En aquesta metodologia s'obtenen sumes ponderades a partir de valoracions subjectives sobre els diferents aspectes que s'han d'avaluar que ens proporcionen un conjunt d'individus de referència. Aquest tipus de metodologia ens ajuda a avaluar conceptes que no es poden definir a partir d'indicadors quantitatius.

Algun exemple és el mètode del panel d'experts (Tsaur et al., 2006),(Ugwu et al., 2006) i el mètode d'opinió pública(Cottrell et al., 2009), en el dos l'assignació de les ponderacions es basa en la opinió. Per mostrar les seves opinions a cada individu li corresponen N punts que ha de distribuir entre els indicadors, assignant més punts a aquells que representen aspectes que cregui que són de major importància. L'assignació dels punts l'ha de fer cada individu de forma independent sense influència de ningú més per no repercutir en el resultat. Quan tots els participants ja han fet aquest exercici s'agafa la puntuació mitjana que ha sortit a cada indicador. De forma que al pes assignat a l' indicador I_j correspon a:

$$w_j = \frac{q_j}{\sum_{s=1}^{m} q_s}$$

On:

- w_j és el pes final assignat a l'indicador I_j
- q_j és la puntuació mitjana assignada a l'indicador I_j
- q_s la puntuació mitjana assignada a l'indicador I_s de la dimensió *s*

Una vegada ja tenim els pesos per a cada indicador, obtenim l'indicador sintètic fent una agregació mitjançant una suma ponderada dels valors normalitzats dels indicadors. En el cas que els indicadors del sistema siguin de tipus qualitatiu, l'indicador sintètic s'obté directament com una suma de les ponderacions. Algun dels inconvenients d'aquest mètode és la fiabilitat de les ponderacions ja que cada individu té una experiència i un nivell d'estudis diferent, cosa que pot influir molt en el resultat.

- Tècniques d'anàlisi multivariant:

i. Anàlisis de components principals:

És una tècnica d'estadística multivariant que s'utilitza per reduir el número inicial de variables en un anàlisi, desenvolupada per Hotelling (1933), consisteix en explicar la quantitat màxima de variabilitat de la mostra amb un número de variables inferior a l'inicial, les variables escollides per aquest mètode es diuen component principals.

L'ACP permet obtenir mesures sintètiques que contenen el màxim número d'informació possible, per utilitzar-la es necessita que hi hagi correlació entre les variables inicials. Degut a la seva capacitat de reduir un sistema compost per variables és un mètode molt utilitzat a l'hora de crear un indicador sintètic. L'aplicació de l'ACP ens proporciona un conjunt de noves variables no correlacionades, amb mitjana aritmètica igual a zero, amb la màxima variància i definides com combinacions lineals dels indicadors inicials. Podem identificar dos tipus de procediments que utilitzen ACP per obtenir els valors d'un indicadors sintètic.

- En primer lloc tenim aquell que es basa en la formulació d'una escala additiva. En aquest cas, primer s'han d'identificar el que diem com indicadors suplents, són aquells indicadors d'una determinada component que tenen una major correlació amb els valors obtinguts per aquesta component. La seva identificació fa que sigui més fàcil interpretar els components principals i permet poder seleccionar els indicadors del sistema inicial més representatius i descartant aquells que ens donen informació més secundària. Un cop seleccionats els indicadors suplents s'ha de definir una variable representativa de cada component principal a partir d'una combinació lineal dels seus indicadors suplents. Per acabar, l'indicador sintètic es construirà mitjançant una suma ponderada de les variables representatives assignant el mateix pes a cada component.

- Una altra opció per crear un indicador sintètic és utilitzar els valors obtinguts de les components principals seleccionades. Depenent de com es defineixi l'indicador sintètic en podem trobar dos grups: El primer utilitza la primera component principal per crear l'indicador. El segon grup utilitza els valors de totes les components principals seleccionades.

Si tenim un conjunt de p variables $X_1, X_{2,\dots,} X_p$ que contenen les característiques d'un conjunt de n objectes, podem definir la matriu X_{nxp} com aquella que te n objectes a les files i p columnes que són les diferents variables, cada un dels elements de la matriu x_{ij} és el valor que pren la variable j en l'objecte i. Per saber quines són les millors variables per l'anàlisi busquem aquelles que tinguin màxima variabilitat, i que siguin redundants entre elles, es a dir amb correlació zero (Riba & Satorra, 2000). Aquestes variables les podem aconseguir a partir de la matriu de variàncies i covariàncies, matriu S:

$$\begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix}$$

Els elements de la diagonal són les variàncies i els factors de fora la diagonal les covariàncies. Quan les variables estan estandarditzades la seva desviació típica és 1, llavors la matriu S correspon a una matriu de correlacions.

Matriu de correlacions R:

$$\begin{pmatrix} 1 & r & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

On: $r_{ij} = \frac{s_{ij}}{s_i s_j} = s_{ij}$

L'objectiu de l'anàlisi de components principals és construir p noves variables $Z_1, Z_2, \ldots, Z_p$, anomenades components principals, que continguin la mateixa informació que les variables originals, que siguin incorrelacionades i de variància 1 i ordenades de major a menor importància. Les variables originals han de ser una mitjana ponderada de les components principals, on les variables més importants tinguin major pes:

$$X_j = w_{ji}Z_1 + w_{j2}Z_2 + \cdots + w_{jp}Z_p \qquad \text{per i =1,2,...p}$$

L'element més important és la matriu de pesos W que permet expressar les variables originals com la combinació lineal de les noves variables. Com que hem construït les variables $Z_1, Z_2, \ldots, Z_p$ de manera que no tinguessin correlació es verifica que la suma dels quadrats dels pesos de la fila i és igual a la variància de X_i:

$$w_{i1}^2 + w_{12}^2 + \cdots + w_{ip}^2 = s_i^2$$

També es demostra que la suma dels quadrats dels pesos de la columna j és la variància total captada per Z_j, que l'anomenarem variància explicada per la component j i la denotarem λ_j:

$$w_{ij}^2 + w_{2j}^2 + \cdots + w_{pj}^2 = \lambda_j$$

I $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ ja que hem dit que les variables han d'estar ordenades per ordre d'importància. Finalment, com que la suma dels quadrats de tots els elements de la matriu dels pesos ha de ser la mateixa tant per files com per columnes, les components principals contenen la mateixa informació que les dades originals i la suma dels seus valors propis coincideixen amb la variància total de les variables, tenim que:

$$\lambda_1 + \lambda_2 + \cdots + \lambda_p = s_1^2 + s_2^2 + .. + s_p^2 = Variancia\ Total$$

I per tant, el percentatge de la variació del model que està continguda en el component j és igual a:

$$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} \times 100$$

ii. Anàlisis factorial:

L'anàlisi factorial és una tècnica estadística utilitzada per identificar patrons en dades multivariants i reduir la seva complexitat, és a dir, s'ha d'identificar les variables que estan relacionades entre si i després agrupar-les amb factors (Teodoro Luque Martínez, 2000).

El seu objectiu consisteix a buscar el número mínim de dimensions capaces d'explicar el màxim d'informació que ens aporten les dades. Quan es recull un número elevat de variables, en un qüestionari per exemple, de forma simultània ens pot interessar saber si les preguntes s'agrupen d'alguna manera característica. En aquest cas si apliquem un anàlisis factorial a les respostes podem trobar grups de variables amb el mateix significat i gràcies a això aconseguir reduir les dimensions necessàries per explicar les respostes dels individus. L'aplicació d'aquesta tècnica implica varis passos. Primer, s'han de seleccionar un conjunt de variables per incloure a l'anàlisi. A continuació es realitza l'anàlisi factorial, això implica l'estimació d'una matriu de correlacions entre les variables. Aquesta matriu s'utilitza per identificar els factors que expliquen la major part de la variància de les dades. La construcció d'un indicador sintètic a partir de subindicadors es pot obtenir mitjançant la reducció d'aquests subindicadors en una sèrie de factors bàsics, tot i que això només serà possible si existeixen subindicadors que donen informació addicional que pot ser obviada, o sigui hi haurà d'haver correlació entre ells, ja que en cas que no n'hi hagués tots aportarien informació substancial i el número de factors no podria ser inferior al d'indicadors originals. Més endavant es mostrarà en més detall l'expressió matemàtica.

3.3. Metodologia de creació d'un indicador sintètic

Per a la creació de l'indicador, seguint l'experiència realitzada per (Nardo et al., 2005) utilitzarem una seguit d'etapes que s'expliquen a continuació:

1) Desenvolupament d'un marc conceptual:
   És necessària la creació d'un marc conceptual que serveixi per justificar la construcció de l'indicador.
2) Selecció de les variables:
   Avaluar quines variables són les necessàries per crear l'indicador.
3) Imputació de dades:
   En cas que ens falti informació podem imputar alguna dada per tal que no sigui un problema més endavant.

4) Normalització de les dades:
   Segurament les variables estaran en diferents escales per això és necessari normalitzar-les perquè així es puguin comparar.
5) Anàlisi multivariant:
   Quan ja s'han escollit les variables que formaran part de l'indicador, es pot fer un anàlisi exploratori per tal d'avaluar si realment són útils i concorden amb les idees donades anteriorment, en aquest punt és quan ens podem adonar de la manca d'informació i això podria ser un problema més endavant.
6) Ponderació i agregació de les dades:
   Un cop tinguem els indicadors parcials s'ha de decidir quina ponderació ha de tenir cada un d'ells en la creació de l'indicador compost. Després ja es pot crear l'indicador i s'ha de presentar la informació de manera clara i senzilla utilitzant taules i gràfics.
7) Anàlisi de robustesa i sensibilitat:
   Finalment s'ha de fer una validació final mitjançant un anàlisi de sensibilitat per mirar si petites variacions en les dades dels indicadors o variables utilitzats provoquen petites variacions en el valor de l'indicador compost.

4. Desenvolupament d'un marc conceptual

Els indicadors són eines importants a l'hora de prendre decisions ja que ens aporten informació científica i tècnica. Són també importants per avaluar i predir tendències de la situació d'una regió o una localitat pel que fa qüestions econòmiques i socials, així per poder avaluar el compliment dels objectius fixats en les polítiques d'un govern.
Un indicador econòmic és una dada estadística que ens permet entendre com estava la economia d'un país en el passat, el present i el futur, i també ens permet comparar-la amb la d'altres països per tenir més perspectiva.
Un indicador ha de tenir les següents característiques:

- Específic
- Explicatiu, amb el seu nom s'ha de saber si es tracta d'un valor absolut, relatiu o una taxa, o a quin grup de població es refereix...
- Ha d'estar disponible per diversos anys
- Fàcil d'interpretar, que no hi hagi dubte del seu significat

Escollir un número correcte d'indicadors socials que siguin comprensibles i permetin sintetitzar be la informació és una tasca complicada, per fer aquest treball s'han triat quatre àrees temàtiques: Població, Salut, Educació, Economia.

5. Selecció de les variables

La selecció de les variables es fa segons la qualitat de la informació que es té, la freqüència en la que està mesurada i la disponibilitat al públic.

S'ha de tenir present que hi ha variables que depenen del tamany de la població, com per exemple el PIB, o de la superfície del país, per poder comparar entre països de manera realista és necessari que en aquests casos s'ajusti l'escala i es treballi amb valors relatius que es poden expressar en unitats, per càpita, per hectàrees. També pot passar que es donin situacions on manqui informació, per això caldrà plantejar-se d' utilitzar un mètode d'imputació de dades.
Per l'elaboració de l'indicador s'han triat les següents variables (Tota la informació es pot trobar a la pagina web oficial de l'Eurostat (https://ec.europa.eu/eurostat/web/main/data/database):

Àmbit	Variable	Descripció	Mesura	Any	Freqüència
Economia	PIB	Principals agregats del PIB per càpita	euro per càpita	2021	Anual
	Ingressos govern	Ingressos totals del govern	Milers d'Euros	2021	Anual
	Despesa govern	Despesa total del govern	Milers d'Euros	2021	Anual
	Mitjana salarial	Salari mitja per empleats a temps complet	Mitjana	2021	Anual
Salut	Esperança de vida	Esperança de vida al néixer	Anys	2020	Anual
	Mortalitat infantil	Taxa de mortalitat infantil	Taxa	2020	Anual
	No fumadors	Percentatge de no fumadors	Percentatge	2019	Anual
	Llits	Quantitat de llits de propietat pública	Per cent mil habitants	2020	Anual
	Escàners	Quantitat d'escàners de tomografia computada	Per cent mil habitants	2020	Anual
	Accidents laborals	Percentatge de treballadors que han patit un	Percentatge de persones que	2020	Annual

		accident laboral dels 15 als 64 anys	treballen ho treballaven fins als últims 12 mesos		
Població	Risc pobresa	Taxa de risc de pobresa (punt de tall: 60% de la renda mitjana equivalent després de les transferències socials)	Taxa	2021	Anual
	Atur	Atur de la població dels 20 als 64 anys	Percentatge	2020	Anual
	Inserció laboral	Temps mitjà entre l'abandonament de l'educació i l'inici del primer treball, en mesos	Mitjana	2009	Anual
Educació	professors	Proporció d'alumnes i estudiants respecte de professors i personal acadèmic	Taxa	2020	Annual
	Graduats	Nombre de graduats en educació terciària	Nombre	2020	Anual
	Nens a la llar d'infants	Alumnes de primera infància com a % de la població d'edat corresponent	Percentatge	2020	Anual
	No poden estudiar	Població que vol participar en educació i formació però no poden pel seu cost econòmic	Percentatge	2016	Anual

Finalment, d'aquestes setze variables se n'han escollit dotze ja que són les que ens serveixen per crear el nostre indicador:

- Mitjana salarial
- PIB/càpita
- Ingressos govern
- Despeses govern
- Inserció laboral
- Professors
- Graduats
- Mortalitat infantil
- Risc pobresa
- Atur
- Esperança de vida
- Llits

Totes aquestes variables s'extreuen per als països que ens interessen, en aquest cas els corresponents al continent Europeu, actualment la regió Europea esta composta per 46 països i 2 territoris dependents. Seguidament observem el llistat de països que inclourem a l'estudi ja que se n'ha pogut extreure informació suficient:

ID	Països Europeus	ID	Països Europeus
1	Albània	17	Irlanda
2	Alemanya	18	Islàndia
3	Àustria	19	Itàlia
4	Bèlgica	20	Letònia
5	Bulgària	21	Lituània
6	Xipre	22	Luxemburg
7	Croàcia	23	Malta
8	Dinamarca	24	Noruega
9	Eslovàquia	25	Països baixos
10	Eslovènia	26	Polònia
11	Espanya	27	Portugal
12	Estònia	28	República txeca
13	Finlàndia	29	Romania
14	França	30	Suècia
15	Grècia	31	Suïssa
16	Hongria		

TAULA 1: PAÏSOS INCLOSOS A L'ESTUDI FONT: ELABORACIÓ PRÒPIA

6. Anàlisi descriptiva

Primer de tot es comença per fer una anàlisi descriptiva de totes les variables, la següent taula mostra el valor mínim, el primer quartil, la mediana, la mitjana, el tercer quartil, el màxim i la quantitat de valors absents que tenim per cada variable.

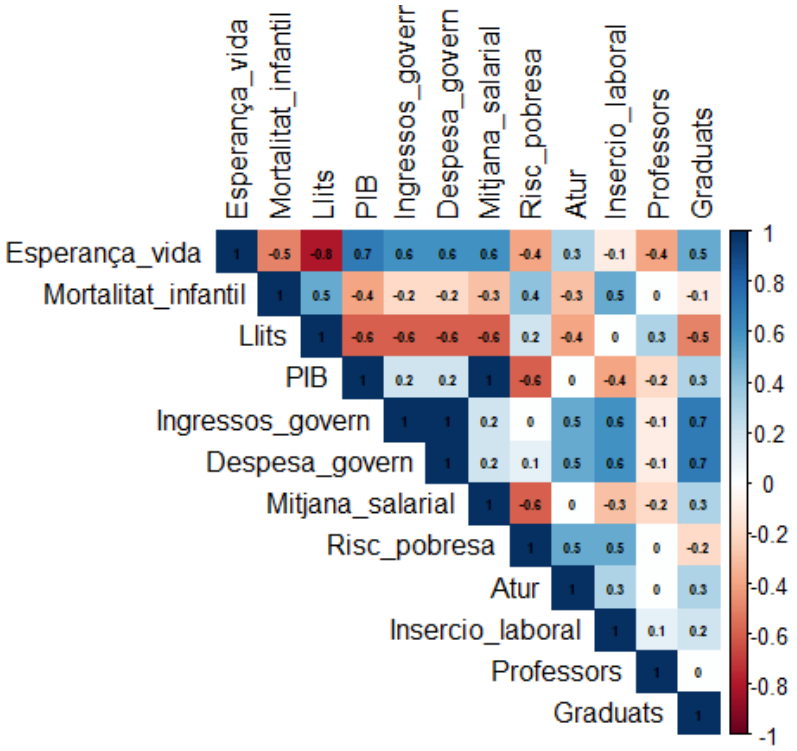
	Esperança_vida	Mortalitat_infantil	Llits	PIB	Ingressos_govern	Despesa_govern
Mínim	75,1	1,4	144,92	10330	5412,4	6542,4
25%	78,8	2,4	255,74	18582,5	26269,85	28425,075
Mediana	81,9	3,2	383,52	27800	97495,15	105091,1
Mitjana	80,71935484	3,406451613	382,3504545	37555,33333	243183,1467	264424,6567
75%	82,85	3,7	489,305	47977,5	243453,7	252285,175
Màxim	84	10	660,58	112780	1711747	1845999
NA's	0	0	9	1	1	1
	Mitjana_salarial	Risc_pobresa	Atur	Insercio_laboral	professors	Graduats
Mínim	10345	8,6	2,5	3,3	7,1	273
25%	16674	12,85	4,8	4,5	10,175	1216,5
Mediana	25034	15,25	5,35	5,6	11,1	5673
Mitjana	30482,73077	16,24642857	6,41	6,133333333	11,51153846	8095,285714
75%	44101,25	19,7	7,325	6,9	12,775	11471
Màxim	72247	23,4	16,4	13,1	16,5	52431
NA's	5	3	1	4	5	3

TAULA 2: ANÀLISI DESCRIPTIU FONT: ELABORACIÓ PRÒPIA

Aquesta taula ens aporta informació general, fent una anàlisi de tots els països junts, per exemple, observem que l'esperança de vida mitjana és de 80,7 i que la diferència entre el país amb l'esperança de vida més baixa i el més alt és d'aproximadament 9 anys, aquests valors corresponent als països de Bulgària i Espanya respectivament. En mortalitat infantil observem una mitjana de 3,4 i també observem molta diferència entre el país amb pitjor taxa i el millor, amb una diferència de 9 punts i que correspon als països d' Estònia (amb taxa màxima) i Albània (amb la taxa mínima). Si ens fixem amb la mediana sobre la variable Llits veiem que el 50% de països disposen d'un mínim de 383 llits per cada cent habitants, pel mateix percentatge el salari mitjà és de com a màxim 25034€ l'any, en l'altra meitat de països és superior a aquest valor. Letònia és el país amb més risc de pobresa a diferència de la República Txeca que és el país amb millor taxa. S'ha de prestar especial atenció en les variables que tenim valors buits (fila NA'S) ja que pot suposar un problema a l'hora de crear l'indicador, per això s'haurà de mirar d'imputar els valors necessaris.

Finalment, comentar que amb el PIB es pot comparar el benestar econòmic entre països, i ens podem fer una idea de en quins països els seus habitants són més rics que en la resta, a les nostres dades trobem que Luxemburg té el PIB per càpita més alt i Bulgària el més baix, tot i que aquest indicador no ens reflecteix la distribució de la riquesa entre els seus habitants i hauríem de comparar amb altres valors per poder assegurar que Luxemburg és el país on, en promig, i viuen els habitants més rics.

Un altre aspecte que interessa saber és si hi ha correlació entre les variables, s' avalua si hi ha tendència creixent o decreixent entre dades, dues variables estan associades si una d'elles dona informació sobre l'altra, per observar-ho es pot fer una matriu de correlacions, cal destacar que no es pot calcular l'estadístic de Pearson si hi ha valors nuls per aquest motiu cal eliminar les files amb dades absents per poder visualitzar-ho, un cop s'han imputat aquestes dades ja es pot assegurar que la correlació no ha variat significativament. Un valor proper a 1 indica que hi ha correlació positiva, és a dir que les variables es relacionen directament, per altra banda, si el valor és proper a -1 la relació és negativa i per tant es relacionen inversament, finalment si r = 0 no hi ha relació lineal entre variables.



IL·LUSTRACIÓ 1: MATRIU DE CORRELACIÓ FONT: ELABORACIÓ PRÒPIA

Si s'observa aquest gràfic es veu com hi ha bastanta correlació entre variables, destaca el cas del PIB per càpita amb la Mitjana salarial on tenen una forta associació lineal i directa d'igual

forma succeeix amb els ingressos del govern i les despeses aspecte totalment lògic. Pel que fa el risc de pobresa s'observa que quan la mitjana salarial és més elevada aquell país té menys risc de pobresa.

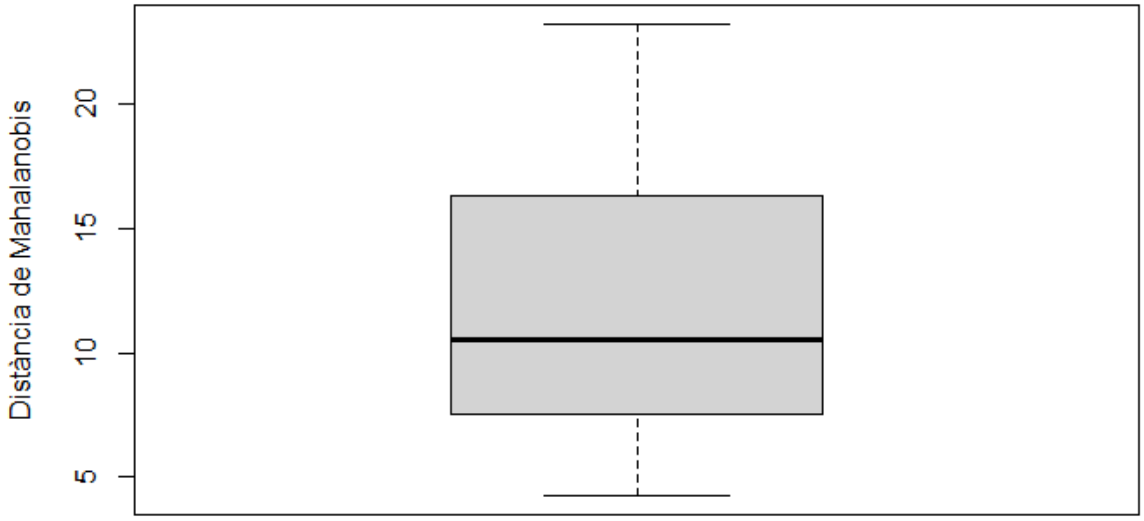
Per acabar l'anàlisi descriptiu ens interessa veure si hi ha presència de valors atípics, normalment s'utilitza la distància de Mahalanobis al quadrat per detectar atípics a nivell multidimensional, la seva expressió és:

$$D^2 = (X_i - \bar{X})' S^{-1} (X_i - \bar{X})$$

On:

- X és un vector columna amb els valors de totes les variables per l'observació i-éssima.
- $\bar{X}$ és un vector columna de les mitjanes mostrals
- S és la matriu de variàncies i covariàncies de les variables: $S = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_{ij} - \bar{X}_j\right)\left(x_{ij'} - \bar{X}_{j'}\right)$

Un cop calculada la distància podem fer un *boxplot* per visualitzar la presència d'algun valor atípic:



IL·LUSTRACIÓ 2: BOXPLOT DISTÀNCIES DE MAHALANOBIS FONT: ELABORACIÓ PRÒPIA

Tot i que aparentment no sembla que hi hagi cap distància fora del normal també es pot identificar si un valor es atípic amb el supòsit que les distàncies es distribueixen mitjançant una Chi-quadrat amb graus de llibertat igual al número de variables (assumim k=12):

$$D^2 \sim \chi_k^2$$

Podem calcular la significació de cada una de les distàncies amb la hipòtesi nul·la que el cas i no és atípic, si fem una taula resum dels resultats utilitzant una significació de l'1%:

Països	Distància	Valor crític	Significació	Conclusió
Albania	20,996667	26,2169673	0,05042928	No Atípic
Austria	8,54224362	26,2169673	0,74145132	No Atípic
Belgium	9,01862399	26,2169673	0,70133888	No Atípic
Bulgaria	7,86753128	26,2169673	0,79539479	No Atípic
Croatia	5,83393309	26,2169673	0,92422071	No Atípic
Cyprus	4,26950436	26,2169673	0,97806757	No Atípic
Czechia	11,2614863	26,2169673	0,50665221	No Atípic
Denmark	12,7532863	26,2169673	0,38722506	No Atípic
Estonia	7,10053467	26,2169673	0,85089886	No Atípic
Finland	5,69940621	26,2169673	0,93047031	No Atípic
France	22,030793	26,2169673	0,03717614	No Atípic
Germany	23,2553684	26,2169673	0,02563454	No Atípic
Greece	19,0884304	26,2169673	0,0864166	No Atípic
Hungary	10,9371129	26,2169673	0,5343157	No Atípic
Iceland	6,59449198	26,2169673	0,88320785	No Atípic
Ireland	8,68503125	26,2169673	0,72957143	No Atípic
Italy	14,9904157	26,2169673	0,24196101	No Atípic
Latvia	8,23581086	26,2169673	0,76644169	No Atípic
Lithuania	10,5671431	26,2169673	0,56633251	No Atípic
Luxembourg	18,8581348	26,2169673	0,09201007	No Atípic
Malta	11,8407772	26,2169673	0,45855038	No Atípic
Netherlands	10,5372805	26,2169673	0,56893327	No Atípic
Norway	18,7199032	26,2169673	0,09551595	No Atípic
Poland	5,35241041	26,2169673	0,94516112	No Atípic
Portugal	7,33687051	26,2169673	0,83457273	No Atípic
Romania	17,6462638	26,2169673	0,1268609	No Atípic
Slovakia	8,43211245	26,2169673	0,75051667	No Atípic
Slovenia	6,73502274	26,2169673	0,87461612	No Atípic
Spain	17,8218855	26,2169673	0,12120669	No Atípic
Sweden	7,72994189	26,2169673	0,80586245	No Atípic
Switzerland	11,2615828	26,2169673	0,50664405	No Atípic

TAULA 3: DISTÀNCIES DE MAHALANOBIS FONT: ELABORACIÓ PRÒPIA

Com que no es troba cap distància significativa podem concloure que no hi ha presència de valors atípics.

7. Imputació de dades

És molt freqüent que no es disposi de totes les dades necessàries a l'hora de crear un indicador, sobretot si ens centrem amb un conjunt de països. Hi ha tres maneres de solucionar aquest problema:

1) Eliminar la informació: en aquest cas s'omet el registre faltant de l'anàlisi, en aquest cas s'ha de tenir en compte que podria perjudicar en el biaix o augmentar la dispersió. En el cas que la unitat d'anàlisi sigui país hauríem d'eliminar el registre sencer i per tant eliminar el país podria no ser d'utilitat.

2) Eliminar la variable: es podria decidir no utilitzar la variable que contingui alguna dada que falta, es considera que si una variable té menys del 5% de dades perdudes respecte el total no es convenient eliminar-la.

3) Fer una imputació simple, per exemple utilitzant la mitjana, mediana o regressions si disposem de la informació necessària. També podem fer una imputació múltiple mitjançant algoritmes de Monte Carlo o cadenes de Markov.

Assignar un valor a les dades perdudes ajuda a reduir el biaix i es realitza l'anàlisi sobre un conjunt complet d'informació. Per les dades utilitzades en aquest estudi es farà servir algun mètode d'imputació, ja que no es vol eliminar els països amb alguna dada faltant de l'anàlisi, ja que quedaria amb poca mostra. Podem seguir un dels següents mètodes:

Imputació simple de dades:

1) Modelització implícita: S'utilitzen mecanismes d'assignació basats en suposicions implícites. Consisteix en omplir els valors buits gràcies a unitats que es comportin de la mateixa manera i amb les mateixes característiques. Es pot reemplaçar els valors absents per valors d'altres fonts externes.

2) Modelització explícita: Es realitza considerant un model d'estadística que utilitza suposicions concretes i explícites. Es pot realitzar de diverses formes:

   - Utilitzant la mitjana. Es tracta d'utilitzar la mitjana de la resta de la mostra. S'ha de tenir en compte que els valors imputats seran estimadors esbiaixats per la mitjana

poblacional i la variància resultat de l'indicador estarà subestimada ja que no considerarà la incertesa de les dades absents.

- Fer ús de regressions lineals.
- Aplicar l'algoritme de expectació i maximització, es basa en la idea de maximitzar una funció d'esperança respecte als paràmetres del model. En la primera fase es calcula la funció d'esperança condicional dels valors perduts, assumint unes estimacions inicials del model. En la segona part s'utilitza la funció d'esperança calculada anteriorment per actualitzar els paràmetres del model per aconseguir una millor estimació. Es repeteix el procés fins assolir convergència.
- Imputació múltiple de dades perdudes: Consisteix en generar múltiples conjunts de dades completades a través d'un procés iteratiu. Una de les principals tècniques utilitzades es la de Monte Carlo i cadenes de Markov. Aquesta metodologia assigna valors a cada dada perduda a partir d'una distribució de dades estimada amb la finalitat de representar la incertesa conseqüent de la informació no disponible. Un cop tenim aquest conjunt de dades s'analitzen estadísticament amb la finalitat d'obtenir estimadors dels valors que seran utilitzats en la imputació.

Seguidament fem una taula on ens mostra quants valors nuls hi ha per cada variable:

	NA's
Esperança_vida	0
Mortalitat_infantil	0
Llits	9
PIB	1
Ingressos_govern	1
Despesa_govern	1
Mitjana_salarial	5
Risc_pobresa	3
Atur	1
Insercio_laboral	4
Professors	5
Graduats	3

TAULA 4: RESUM DELS NA'S PER VARIABLE FONT: ELABORACIÓ PRÒPIA

Per imputar les dades absents es farà ús d'un mètode simple, concretament s'assigna als valors nuls la mitjana de la variable, d'aquesta forma un cop fet el canvi es pot tornar a fer un taula resum i observar que ja no hi ha valors NA per a cap variable, la mitjana és la mateixa i canvia els valors dels diferents percentils.

	Esperança_vida		Mortalitat_infantil		Llits		PIB	
	Abans	Després	Abans	Després	Abans	Després	Abans	Després
Mínim	75,1	75,1	1,4	1,4	144,92	144,92	10330	10330
25%	78,8	78,8	2,4	2,4	255,74	276,76	18582,5	19055
Mediana	81,9	81,9	3,2	3,2	383,52	382,35	27800	28920
Mitjana	80,719	80,72	3,41	3,41	382,35	382,35	37555,33	37555,33
75%	82,85	82,85	3,7	3,7	489,305	448,94	47977,5	47115
Màxim	84	84	10	10	660,58	660,58	112780	112780
NA's	0	0	0	0	9	0	1	0

	Ingressos_govern		Despesa_govern		Mitjana_salarial		Risc_pobresa	
	Abans	després	Abans	després	Abans	després	Abans	després
Mínim	5412,4	5412,4	6542,4	6542,4	10345	10345	8,6	8,6
25%	26269,85	26436,5	28425,075	28573,75	16674	18448	12,85	13,35
Mediana	97495,15	98668,8	105091,1	105814,2	25034	28765	15,25	16
Mitjana	243183,15	243183,15	264424,66	264424,66	30482,73	30482,73	16,25	16,25
75%	243453,7	243370,02	252285,18	259269,33	44101,25	41664	19,7	19,4
Màxim	1711747	1711747	1845999	1845999	72247	72247	23,4	23,4
NA's	1	0	1	0	5	0	3	0

	Atur		Insercio_laboral		Professors		Graduats	
	Abans	després	Abans	després	Abans	després	Abans	després
Mínim	2,5	2,5	3,3	3,3	7,1	7,1	273	273
25%	4,8	4,8	4,5	4,75	10,18	10,4	1216,5	1264,5
Mediana	5,35	5,4	5,6	5,6	11,1	11,51	5673	6763
Mitjana	6,41	6,41	6,13	6,13	11,51	11,51	8095,29	8095,29
75%	7,33	7,25	6,9	6,17	12,8	12,6	11471	10814
Màxim	16,4	16,4	13,1	13,1	16,5	16,5	52431	52431
NA's	1	0	4	0	5	0	3	0

TAULA 5: COMPARACIÓ ANÀLISI DESCRIPTIU FONT: ELABORACIÓ PRÒPIA

8. Normalització de les dades

Normalment les variables que es trien per crear un indicador no estan expressades amb la mateixa unitat, per exemple tenim algunes variables mesurades com a taxa, algunes com a percentatge i d'altres com a número, per aquest motiu abans de procedir a agregar les

variables serà necessari normalitzar-les per evitar que hi hagi correlació entre variables amb diferents unitats i arribar a un anàlisi incorrecte.

Alguns dels mètodes per normalitzar variables són:

- Utilitzar taxes o percentatges de variació. Quan s'utilitza informació per diversos anys podem utilitzar la taxa de variació que es calcula de la següent manera:

$$y_t = \frac{x_t - x_{t-1}}{x_t} \times 100$$

  D'aquesta forma s'obté un estimador sense dimensions.

- La manera més simple de normalitzar variables i fer-les comparables és establir un rànquing dels seus valors, d'aquesta manera les dades queden independitzades dels possibles valors atípics.

- Estandardització: Com que es pot calcular per cada variable la mitjana i la desviació típica es pot calcular fàcilment els valors estàndard, de la següent manera:

$$y_t = \frac{x_t - \mu_t}{\sigma_t}$$

- Re-escalament: Consisteix en transformar els nivells de les variables per definir-les entre l'interval [0,1] mitjançant la distància entre els valors màxims i mínims que adopta la variable, el seu càlcul és:

$$y_t = \frac{x_t - (x_t)}{((x_t) - (x_t)\ )} \in [0,1]$$

  Així doncs la unitat amb més força tindrà un 1 i el de menys un 0. S'ha de tenir en compte que en cas que hi hagi valors atípics pot distorsionar significativament el resultat ja que els valors normals els concentraria en un petit rang.

- Distància a una unitat d'anàlisi de referència: en aquest cas la nova variable es calcula de la següent manera:

$$y_t = \frac{x_t}{x_{t0}^R}$$

  On R és una unitat de referència, per tant quedaran els resultats referents a aquesta unitat.

Per normalitzar les dades d'aquest treball s'utilitza el mètode d'estandardització i s'aplica a les variables que no estan mesurades en taxa ni en percentatge, aquestes són: Ingressos govern, despesa govern, mitjana salarial, inserció laboral i graduats.

9. Anàlisi multivariant

9.1. Anàlisi clúster

L'anàlisi clúster és una tècnica d'aprenentatge no supervisat que s'utilitza per agrupar un conjunt d'objectes en un subconjunt més petit de clústers on els objectes hauran de tenir un cert grau d' homogeneïtat interna i heterogeneïtat externa, és a dir, hauran de tenir característiques similars entre els objectes del mateix clúster i diferenciar-se de la resta de clústers. L'objectiu d'aquest anàlisi és descobrir patrons interns en les dades i agrupar-les mitjançant similituds o distàncies entres els objectes (Parra Francisco, 2017).

Podem agrupar els algoritmes de formació de conglomerats en dues categories:

- Algoritmes de partició (no jeràrquics), on dividim un conjunt d'observacions en k conglomerats amb k definit inicialment.
- Algoritmes jeràrquics, el mètode proporciona una jerarquia de divisions del conjunt d'elements.

Ens centrarem en el mètode de k-means que correspon als algoritmes no jeràrquics, el seu objectiu és agrupar observacions similars amb un número fix de clústers. L'algoritme es pot resumir com (Delgado Ronald, 2018):

a) Definir k centroides a l'atzar
b) Calcular les distàncies de cada un dels punts d'entrada als k centroides i assignar cada punt al centroide al qual la seva distància sigui menor.

c) Actualitzar la posició dels k centroides calculant la posició mitja de tots els punts que pertanyen a cada classe.
d) Repetir els passos a) i b) fins que els centroides no canviïn de posició i que per tant les assignacions a cada grup tampoc canviïn.

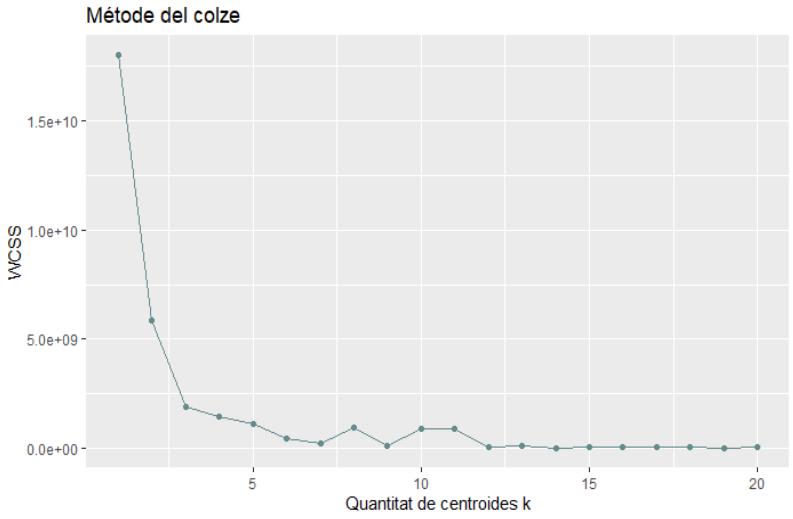
L'algoritme de k-means és un problema d'optimització on la funció a minimitzar és la suma de les distàncies quadràtiques de cada objecte al centroide del seu clúster, el problema es pot formular de la següent forma:

$$\min E(\mu_i) = min \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_j||^2$$

On S és el conjunt de dades dels quals els seus elements són x_j representat per vectors i on cada un dels seus elements representa un atribut. μ_i és el centroide de cada grup k. Un dels seus inconvenients és que l'analista hagi d'escollir el número de k, és fàcil cometre algun error, també hem de tenir en compte que l'algoritme es sensible als *outliers* és a dir que pot ser que afecti als clústers, tot i que també es pot utilitzar per detectar anomalies (Sanz Francisco, s.d.).

A vegades no és fàcil conèixer la quantitat òptima de centroides k a utilitzar, podem aplicar la tècnica del Colze (Elbow Method) que busca seleccionar la quantitat ideal de grups a partir de la optimització de la WCSS (Within Cluster Summed Squares) que mesura la dispersió interna de les dades dins de cada clúster. El nombre òptim de clústers el podem trobar si fem un gràfic de WCSS en funció del nombre de clúster, en representar-ho s'observa un canvi significatiu en la pendent de la corba, de manera que recorda a un colze, el número òptim de clústers el marcarà el punt d'aquest colze.

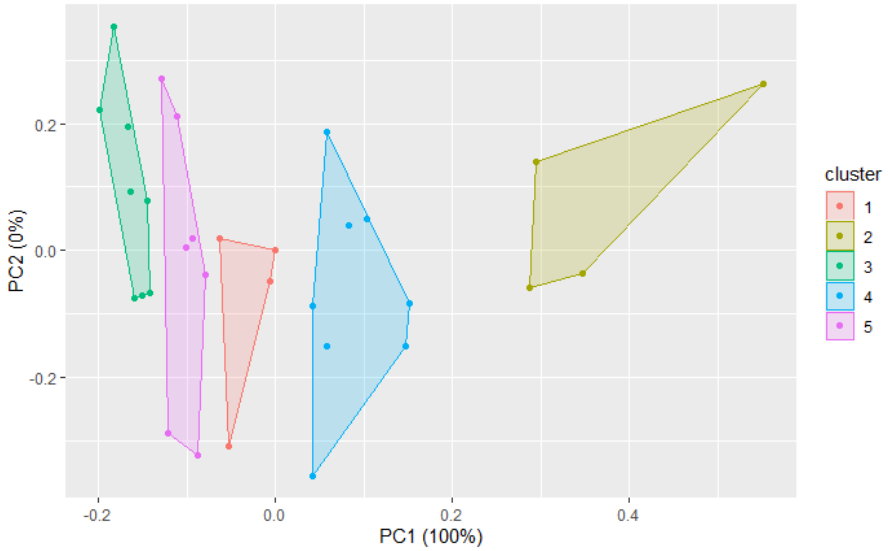
Si fem aquest gràfic en les dades del treball obtenim:



TAULA 6: GRÀFIC MÈTODE DEL COLZE FONT: ELABORACIÓ PRÒPIA

A partir de la corba obtinguda podem veure com a mesura que utilitzem una k més gran disminueix la variació de WCSS, visualitzant la gràfica un nombre adequat de clústers seria entre 5 i 6 ja que després ja no hi ha variacions importants. Seguidament apliquem l'algoritme de k-means mitjançant la funció *kmeans()* que ens proporciona R amb k = 5.

Podem veure els resultats de l'agrupament:



IL·LUSTRACIÓ 3: K-MEANS FONT: ELABORACIÓ PRÒPIA

Per poder visualitzar millor quins són els països que tenim a cada clúster podem fer un mapa Europeu i agrupar d'un mateix color els països que pertanyen al mateix grup, d'entrada caldria esperar que en fer aquesta anàlisi els països agrupats siguin pròxims segons la seva disposició geogràfica, amb la funció *hcmap* d'R podem visualitzar els grups en un mapa Europeu:

IL·LUSTRACIÓ 4: MAPA DELS CLÚSTERS FONT: ELABORACIÓ PRÒPIA

En el mapa anterior podem visualitzar l'anàlisi resultant, els països que estan dins del mateix grup representa que son els més similars entre ells i tenen diferències amb la resta de grups, tenim:

- Grup 1: Albània, França, Itàlia, Malta
- Grup 2: Irlanda, Luxemburg, Noruega, Suïssa
- Grup 3: Bulgària, Croàcia, Grècia, Letònia, Hongria, Polònia, Romania, Eslovàquia
- Grup 4: Àustria, Bèlgica, Dinamarca, Finlàndia, Islàndia, Alemanya, Països Baixos, Suècia
- Grup 5: Xipre, República Txeca, Estònia, Lituània, Portugal, Eslovènia, Espanya

Fent l'anàlisi s'observa que els clústers han separat els països més occidentals i els nòrdics en diferents grups, per exemple dins el grup 3 (groc) la majoria de països són de la part d'occident, el grup número 4 (lila) ajunta els països més nòrdics i centrals; per tant s'arriba a la conclusió que hi ha diferències entre aquests tipus de països.

9.2. Anàlisi factorial exploratori

L'anàlisi factorial exploratori s'utilitza per descobrir l'estructura interna d' un nombre gran de variables partint de la base que es creu que hi ha uns factors associats a grups de variables. Esquema d'anàlisi factorial (de la Fuente Fernández, s.d.):

Siguin (X_1, X_2, ..., X_p) un conjunt de p variables tipificades que s'inclouen a l'anàlisi i tenint la informació sobre n subjectes obtenim la matriu:

Subjecte	X_1	X_2	...	X_p
1	x_{11}	x_{12}	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2p}
...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{np}

El model de l'anàlisi factorial ve donat per les equacions:

$$X_1 = a_{11}\mathbf{F_1} + a_{12}\mathbf{F_2} + \ldots + a_{1k}\mathbf{F_k} + \mathbf{u_1}$$
$$X_2 = a_{21}\,\mathbf{F_1} + a_{22}\mathbf{F_2} + \ldots + a_{2k}\mathbf{F_k} + \mathbf{u_2}$$
...
$$X_p = a_{p1}\mathbf{F_1} + a_{p2}\mathbf{F_2} + \ldots + a_{pk}\mathbf{F_k} + \mathbf{u_p}$$

On:

- ($\mathbf{F_1}$, $\mathbf{F_2}$,..., $\mathbf{F_k}$) (k<p) són els factors comuns, se suposa que E(F_i) = 0 i var(F_i) = 1
- ($\mathbf{u_1}$,$\mathbf{u_2}$,...,$\mathbf{u_p}$) són els factors únics o específics, se suposa que tenen mitjana 0 i no estan correlacionats [E(u_i) = 0; Cov(u_i, u_j) = 0 sí i≠j; (i, j = 1, … ,p)]. A més a més se suposa que tampoc hi ha correlacions entre aquests dos tipus de factors Cov(F_i, u_j)=0, ∀i=1,..,k; j=1, …, p.

- I els coeficients (a_{ij}) {i=1,...,p; j=1,...,k} les càrregues factorials

Si es compleix que [Cov(Fi, Fj) = 0 si i≠j; j, i=1,…,k] estem davant d'un model amb factors ortogonals. Si no es compleix es diu que el model és de factors oblics.
Expressat en forma matricial:

x = A f + u ⇔ X = FA' + U

On X és la matriu de dades, A és la matriu de càrregues factorials, F = matriu de puntuacions factorials:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}, \qquad f = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_p \end{pmatrix}, \qquad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix}, \qquad A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pk} \end{pmatrix},$$

$$F = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1k} \\ f_{21} & f_{22} & \cdots & f_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ f_{p1} & f_{p2} & \cdots & f_{pk} \end{pmatrix}$$

Utilitzant les hipòtesis anteriors, tenim:

$$VAR(X_i) = \sum_{j=1}^{k} a_{ij}^2 + \psi_i = h_i^2 + \psi_i, \qquad i = \{1, .., p\}$$

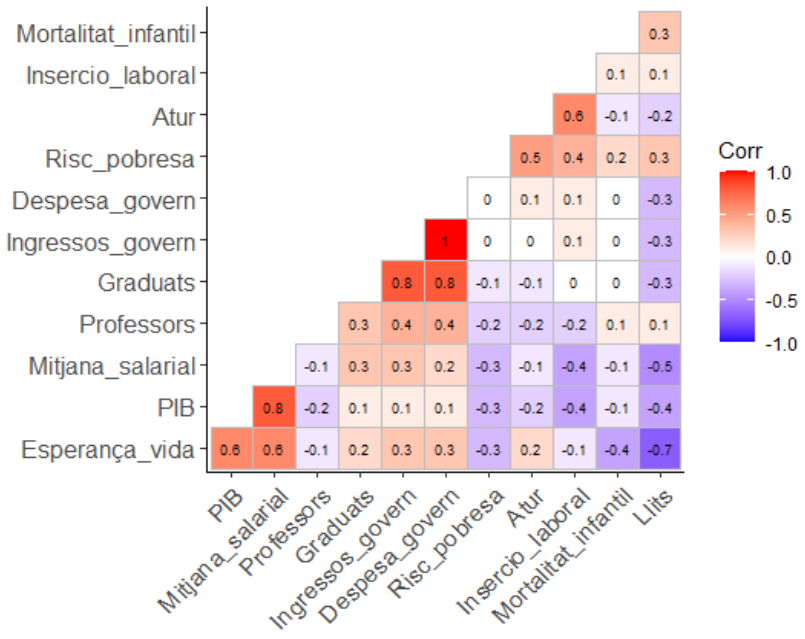
On $h_i^2 = Var(\sum_{j=1}^{k} a_{ij} F_j)$ i $\psi_i = Var(u_i)$ són la comunalitat i especificitat de la variable X_i. Com a conseqüència la variància de cada una de les variables analitzades es pot descompondre en dos parts: la comunalitat h_i^2 que representa la variància explicada pels factors comuns i la especificitat ψ_i que representa la part de la variància específica de cada variable. A més a més tenim que:

$$Cov(X_i, X_i) = Cov\left(\sum_{j=1}^{k} a_{ij} F_j, \ \sum_{j=1}^{k} a_{ij} F_j\right) = \sum_{j=1}^{k} a_{ij} \times a_{ij} \ \ \forall i \neq j$$

Per tant són els factors comuns els que expliquen les relacions entre les variables, per això els factors comuns tenen interès i son susceptibles d'interpretació experimental. Els factors únics s'inclouen en el model ja que no poden expressar p variables en funció d'un número més reduït de factors.

Anàlisis factorial exploratori en R (Bolaños Luis, 2020):

1) Calcular la matriu de correlació: En R es pot fer servir la funció *cor()* que s'utilitza per realitzar una matriu de correlacions. Per graficar la matriu una opció és fer servir la funció gg*corrplot()* del paquet gg*corrplot*.



IL·LUSTRACIÓ 5: GRÀFIC MATRIU DE CORRELACIÓ FONT: ELABORACIÓ PRÒPIA

Després de calcular la matriu de correlació s'ha de verificar si la matriu de dades és factoritzable mitjançant la prova d'esfericitat de Barlett i la prova de Kaiser-Meyer-Oklin.

1.1) Verificar que la matriu és factoritzable:

Ens interessa comprovar que hi ha correlació suficient entre les variables per poder efectuar l'anàlisi factorial.

La prova de Bartlett s'utilitza per provar la hipòtesi nul·la que les variables analitzades no estan correlacionades, o sigui, que la matriu de correlació és la identitat. En R la funció *cortest.bartlett* ens permet realitzar aquesta prova, si l'apliquem a les nostres dades s'obté:

```
mat_cor <- round(cor(data[,-1]),1
p_esf <- cortest.bartlett(mat_cor, n =31, diag = TRUE)
p_esf$p
[1] 9.323724e-34
```

Un p valor < 0.05 ens permet rebutjar la hipòtesi nul·la i per tant sí que hi ha correlació entre variables i es pot seguir endavant amb l'anàlisi factorial.

Una altra prova que és adient aplicar és el criteri de Kaiser-Meyer-Okin, és una mesura que ens indica la proporció de variància en les variables que poden ser causades per factors subjacents, serveix per saber si les dades són aptes per realitzar l'anàlisi factorial. Com a referència tenim que:

- 0.00 a 0.49 → inacceptable
- 0.50 a 0.59 → miserable
- 0.60 a 0.69 → mediocre
- 0.70 a 0.79 → mitjà
- 0.80 a 0.89 → meritori
- 0.90 a 1.00 → meravellós

Per realitzar aquest test s'utilitza la funció *KMO()* d'R, aplicant-la a les dades del treball:

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = mat_cor)
Overall MSA =  0.65
MSA for each item =
Esperança_vida  Mortalitat_infantil  Llits
0,7             0,48                 0,83
PIB  Ingressos_govern  Despesa_govern  Mitjana_salarial
0,57 0,59              0,59            0,69
Risc_pobresa  Atur  Insercio_laboral  Professors  Graduats
0,59          0,52  0,69              0,69        0,84
```

El resultat és de 0.65 que ens indica que no és perfecte però seguirem amb l'anàlisi factorial.

2) Escollir un mètode per extreure els factors
Alguns exemples que es poden escollir són:

- Mètode de mínims quadrats no ponderats → Es busca trobar una solució que minimitzi la suma dels errors quadrats entre les variables observades i les variables estimades a partir de factors.

- Mètode de Mínims quadrats generalitzats → Minimitza la suma dels quadrats de les diferències entre la matriu de correlacions observada i reproduïda. Les correlacions es ponderen per l'invers de la seva exclusivitat, de manera que les variables amb un alt valor d'exclusivitat tindran una ponderació menor d'aquelles que tinguin un valor baix d'exclusivitat.

- Mètode de màxima versemblança → L'objectiu és trobar els valors dels factors i les càrregues que maximitzin la versemblança de les dades observades, proporciona les

estimacions dels paràmetres que amb major probabilitat ha produït la matriu de correlacions observada, si la mostra prové d'una distribució normal multivariada.

- Factorització de eixos principals → Parteix de la matriu de correlacions entre les variables que estem analitzant. S'agafen els coeficients de correlació múltiple entre les variables i s'eleven al quadrat, aquests valors es posen a la diagonal principal de la matriu com estimacions inicials de les comunalitats de les variables. Es calculen les càrregues factorials utilitzant les estimacions inicials de les comunalitat, aquestes càrregues factorials resultants s'utilitzen per estimar de nou les comunalitats que reemplacen a les estimacions prèvies de comunalitat a la diagonal. Les iteracions continuen fins que el canvi en les comunalitats, d'una iteració a la següent satisfan el criteri de convergència per la extracció.

- Mètode Alfa → Tracta de calcular el coeficient alfa de Cronbach que ens indica la correlació mitjana entre les variables incloses en un factor. L'objectiu és maximitzar aquest coeficient.

- Factorització imatge → Mètode per la extracció de factors, basat en la teoria de les imatges. La part comú d'una variable, anomenada imatge parcial, es defineix com la regressió lineal sobre la resta de les variables, en lloc de ser una funció dels factors hipotètics.

  Amb R podem utilitzar la funció *fa()* per realitza un anàlisis factorial i podem utilitzar les següents comandes per aplicar un mètode o un altre:

  - minres: residu mínim
  - mle: màxima versemblança
  - paf: mètode d'eixos principals
  - alpha: alfa
  - minchi: mínims quadrats
  - minrak: rang mínim

  En el nostre anàlisi utilitzarem el mètode d'eixos principals.

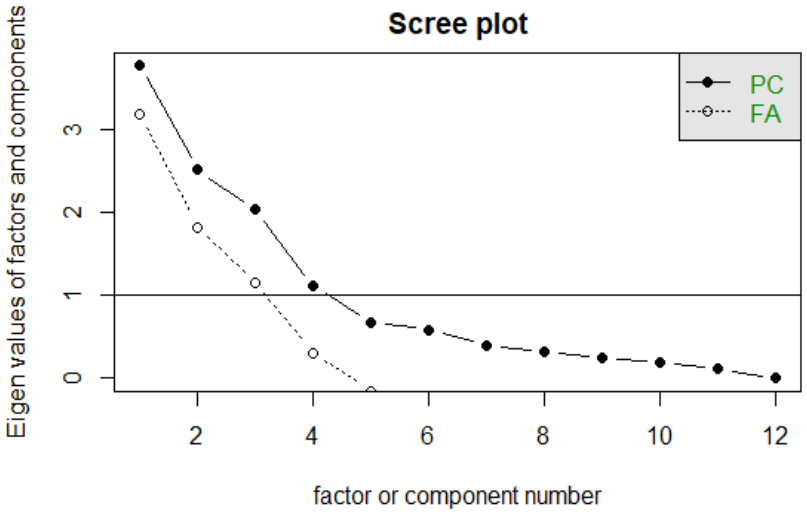
3) Determinar el nombre de factors:
   - Kaiseer Criterion (Guttman, 1954): aquesta regla es basa en la idea que tots els factors amb valors propis (*eigenvalues*) més grans que 1 expliquen més variància que la variància mitjana d'una variable original, és a dir, ens indica que s'han de conservar tots aquells factors que tinguin un *eigenvalue* de 1 o més gran.

   - Anàlisi de Scree Plot (Cattell, 1966): aquest mètode és una representació gràfica que mostra la quantitat de variància explicada per cada un dels components. Es pretén seleccionar nombre reduït de factors que té *eigenvalues* significativament superiors a

la resta. S'identifica el punt d'inflexió en la curva del scree plot, lloc a partir del qual la curva es transforma en una línia recta.

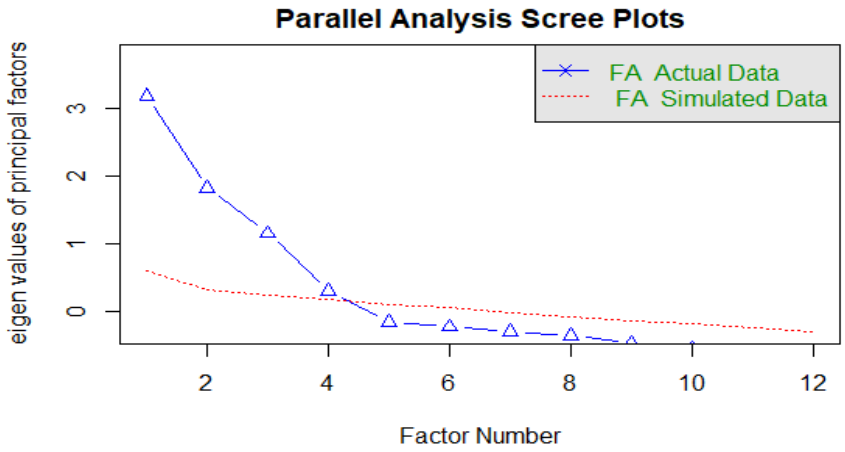
- Análisi paral·lel (Horn, 1965): Aquesta regla complementa les anteriors quan tenim un número elevat de variables inicials i de factors resultants. El procediment es basa en la comparació de matrius de correlació observada i matrius de correlació aleatòries, la idea és identificar les correlacions que són significativament més altes en la matriu observada que en les matrius aleatòries, el que fa el procediment és ordenar les observacions de manera aleatòria entre cada variable i es tornen a calcular els *eigenvalues* a partir d'aquesta nova base de dades ordenada de forma aleatòria. Els factors amb valors *eigenvalues* més grans als valors aleatoris els mantenim per la interpretació.

  En les nostres dades si realitzem l'Scree plot, obtenim:



IL·LUSTRACIÓ 6: SCREE PLOT FONT: ELABORACIÓ PRÒPIA

Aquest gràfic indica que per realitzar una anàlisi factorial es necessiten quatre factors, es fa una anàlisi paral·lela per comparar el resultat i s'obté que ens convé utilitzar el mateix nombre de factors, o sigui, quatre.



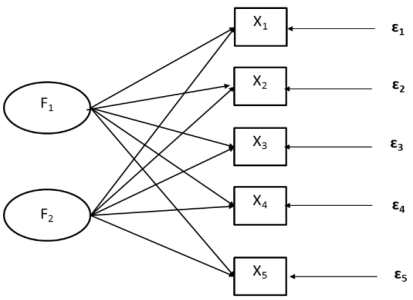
IL·LUSTRACIÓ 7: PARALLEL ANALYSIS SCREE PLOT FONT: ELABORACIÓ PRÒPIA

4) Rotar la matriu

L'obtenció de la matriu factorial és només un primer pas, normalment la matriu que s'obté no ens dona uns factors interpretables. Interessa aconseguir una estructura simple de factors, així doncs, es volen unes saturacions que siguin altes en comparació a les altres, que seran baixes, per així destacar la influència dels factors comuns sobre les variables observables. Hi ha dues formes de realitzar la rotació de factors, la ortogonal i la obliqua (IBM, s.d.):

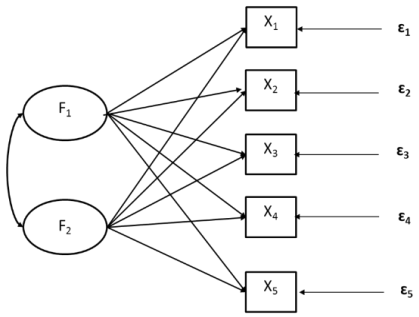
a) Rotació ortogonal: Els eixos es giren de tal manera que els factors conserven la incorrelació entre ells, els factors es queden perpendiculars entre ells.

b) Rotació obliqua: Els eixos ja no són ortogonals, és a dir, els factors tindran una certa correlació, els eixos dels factors ja no són perpendiculars, és útil quan es preveu que hi hagi interacció o correlació entre factors.



IL·LUSTRACIÓ 8: FACTOR EXPLORATORI ORTOGONAL

FONT: (JOAQUIM ALDÁS & EZEQUIEL URIEL, 2017)



IL·LUSTRACIÓ 9: FACTOR EXPLORATORI OBLIQUA

FONT: (JOAQUIM ALDÁS & EZEQUIEL URIEL, 2017)

Tenir en compte que la rotació no afecta la bondat de l'ajust, el que canvia es la variància explicada per cada factor.

- Varimax: Mètode de rotació ortogonal que minimitza el número de variables que tenen càrregues altes a cada factor. Simplifica la interpretació dels factors.

- Criteri Oblimin directe: Mètode per a la rotació obliqua. Per aplicar aquest mètode és necessari el valor delta que servirà per ajustar els eixos en funció de les saturacions buscant una millor aproximació, però considerant que la variància es distribueix entre tots els factors.

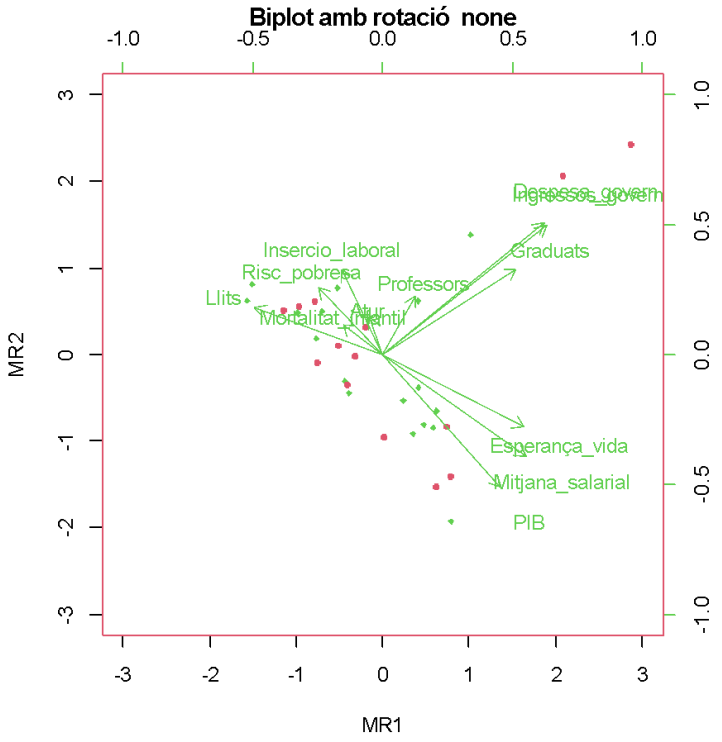
- Mètode quartimax: Mètode que minimitza el número de factors necessaris per explicar cada variable.

- Mètode equamax: Mètode de rotació que és combinació del mètode varimax i quartimax, que simplifiquen els factors i les variables. Minimitza tan el nímero de variables com el número de factors necessaris per explicar una variable.

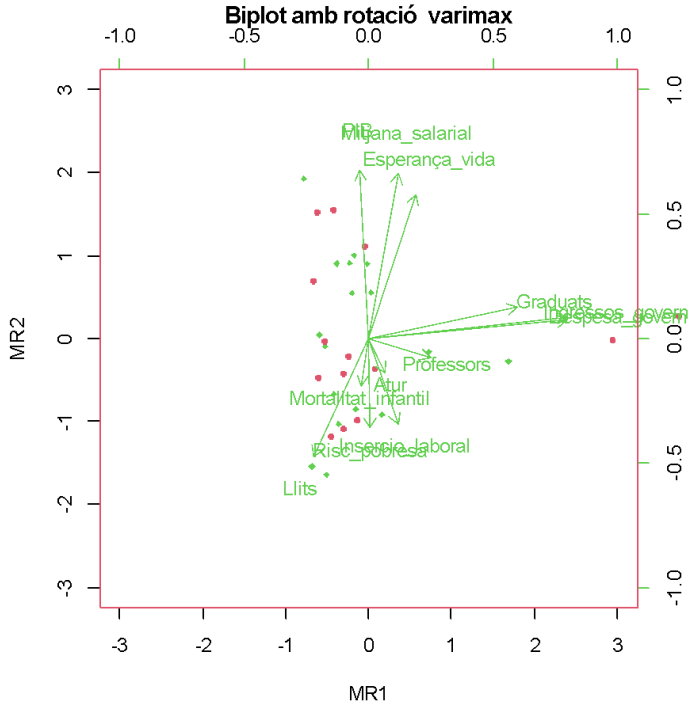
- Rotació Promax: Rotació obliqua que permet que els factors estiguin correlacionats. Aquesta rotació es pot calcular més ràpidament que una rotació oblimin directa, és útil quan tenim un conjunt de dades grans.

La funció *biplot.psych()* del paquet *psych* d'R s'utilitza per crear *biplots* (gràfic utilitzat per visualitzar la relació entre dos conjunts de variables) dels resultats dels anàlisis factorials, el *biplot* que mostra la relació entre les observacions i els factors.

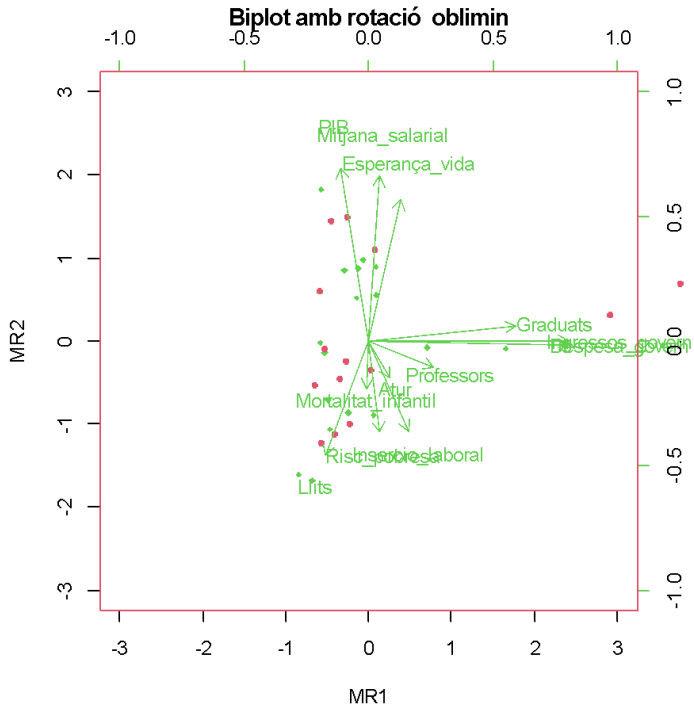
Per les dades a analitzar farem una visualització de les rotacions varimax (ortogonal), oblimin (oblicua):



IL·LUSTRACIÓ 10: GRÀFIC BIPLOT SENSE ROTACIÓ FONT: ELABORACIÓ PRÒPIA



IL·LUSTRACIÓ 11: GRÀFIC BIPLOT ROTACIÓ VARIMAX FONT: ELABORACIÓ PRÒPIA



IL·LUSTRACIÓ 12: GRÀFIC BIPLOT ROTACIÓ VARIMAX FONT: ELABORACIÓ PRÒPIA

5) Interpretació

Finalment es fa l'anàlisi factorial amb quatre factors i rotació varimax (ortogonal) i el model d'eixos principals, per fer-ho utilitzarem la funció *fa()* indicant els factors amb l'argument nfactors=, *rotate=* per indicar que farem servir una rotació *varimax* i l'argument *fm=* per indicar que utilitzarem el model d'eixos principals. S'obtenen els següents resultats:

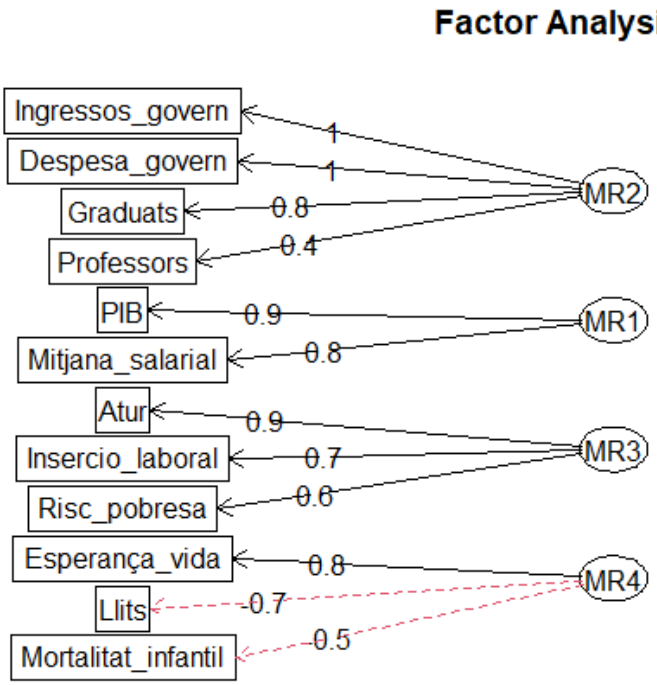
```
Factor Analysis using method =  pa
Call: fa(r = data[, -1], nfactors = 4, rotate = "varimax", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
                          PA2   PA1   PA3   PA4   h2     u2     com
Esperança_vida           0.18  0.51  0.05  0.75 0.87  0.1349 1.9
Mortalitat_infantil      0.01  0.00  0.04 -0.46 0.21  0.7864 1.0
Llits                   -0.23 -0.38 -0.05 -0.65 0.62  0.3767 1.9
PIB                     -0.01  0.93 -0.21  0.15 0.93  0.0705 1.2
Ingressos_govern         1.00  0.08  0.05  0.10 1.01 -0.0142 1.0
Despesa_govern           0.99  0.07  0.07  0.10 1.01 -0.0059 1.0
Mitjana_salarial         0.18  0.80 -0.18  0.24 0.76  0.2432 1.4
Risc_pobresa            -0.03 -0.10  0.64 -0.35 0.55  0.4548 1.6
Atur                    -0.02 -0.06  0.79  0.24 0.69  0.3125 1.2
Insercio_laboral         0.08 -0.25  0.70 -0.04 0.56  0.4365 1.3
Professors               0.39 -0.28 -0.39 -0.05 0.39  0.6109 2.8
Graduats                 0.76  0.10 -0.07  0.08 0.60  0.4016 1.1

                       PA2  PA1  PA3  PA4
SS loadings           2.84 2.09 1.78 1.49
Proportion Var        0.24 0.17 0.15 0.12
Cumulative Var        0.24 0.41 0.56 0.68
Proportion Explained  0.35 0.25 0.22 0.18
Cumulative Proportion 0.35 0.60 0.82 1.00

Mean item complexity =  1.5
Test of the hypothesis that 4 factors are sufficient.
```

En les columnes PA1, PA2, PA3, PA4 es poden observar les càrregues factorials per cada variable, expressen la magnitud de la correlació de la variable amb el factor, també s'expressa el valor de les comunalitats i unicitat. En la taula de sota es troba la proporció de variància explicada per cada factor i la total que correspon a un 68%. Seguidament ens indica el valor mig de la complexitat de l'item que com més petit millor i s'observa que utilitzar 4 factors és suficient.

Es pot fer un gràfic d'arbre amb la funció *diagrama()* d'R, aquest diagrama ens ajuda a visualitzar l'estructura de l'anàlisi factorial, la relació entre les seves variables observades i els factors:



IL·LUSTRACIÓ 13: DIAGRAMA DELS FACTORS FONT: ELABORACIÓ PRÒPIA

Representat amb quadrats tenim les variables utilitzades, a la dreta del gràfic (encerclats) hi ha els quatre factors i amb les fletxes s'observa la relació dels factors amb les variables i la distància que hi ha entre ells la qual ens indica la seva força, una variable més propera a un factor tendeix a estar més correlacionada amb aquest factor.

Es poden identificar els quatre factors de la següent manera:

- Factor 1: Economia, inclou el PIB per càpita i la mitjana salarial.
- Factor 2: Economia i educació que engloba els ingressos i despeses del país i la quantitat de professors i graduats que hi ha.
- Factor 3: Laboral, on hi trobem les variables atur, inserció laboral i risc de pobresa.
- Factor 4: Salut, que inclou l'Esperança de vida, la quantitat de llits als hospitals i la mortalitat infantil.

10. Agregació de la informació

Finalment s'arriba a un dels punts més importants del procediment, en el qual s'han d'agregar els indicadors i les variables seleccionades en l'indicador compost pròpiament dit. Això implica la necessitat d'incorporar la informació mitjançant l'aplicació de diferents factors de ponderació que reflecteixin la importància relativa de cada subindicador. La metodologia d'agregació ha d'estar explicada amb claredat i que sigui fàcilment reproduïble, garantint transparència en el procés. S'ha de tenir en compte que no existeix una metodologia objectiva per establir els pesos de les variables, per aquest motiu se sol recórrer a l'opinió experta i la recerca de consensos amb grups d'interès que sintetitzin les prioritats i els diferents punts de vista. És raonable assignar més visibilitat a aquelles variables que tinguin més fiabilitat, és a dir, aquelles que la quantitat de dades perdudes sigui mínima o que estiguin mesurades amb criteris estàndards. En molts casos es pot aplicar ponderacions equi-probables sobretot si no tenim arguments que indiquin la necessitat de ponderar en diferents pesos cada una de les variables. Per últim, cal destacar que s'han de tenir en compte les possibles correlacions entre variables per evitar una doble comptabilitat que es podria produir en el cas que dos indicadors estiguessin explicant el mateix fenomen, per aquest motiu s'ha explicat la funció d'una anàlisi exploratòria, ja que en el cas que dues variables col·lineals s'estiguessin introduint en el l'indicador compost s'estaria duplicant el pes de la dimensió que representen. A continuació es revisen alguns exemples dels principals procediments que es poden seguir per agregar la informació (Schuschny & Soto, s.d.):

- Establir pesos equiproporcionals: Aquest criteri facilita el càlcul i funciona bé quan totes les dimensions de l'anàlisi tenen la mateixa prioritat i són representades per una quantitat similar de sub-indicadors.

- Mètodes participatius: En la primera part del treball ja s'ha comentat com funciona aquest procediment, en resum, s'utilitza l' opinió d'experts o de la població perquè indiquin quines creuen que són les variables més importants.

- Ponderació a través del càlcul de la distància: Per cada variable, com més lluny de l'objectiu estigui més gran serà la prioritat per arribar-hi. Es pot considerar com a factor de ponderació el quocient entre el valor de la variable i l'objectiu on s'ha d'arribar. Per definir els objectius ens podem basar en les metes polítiques que s'apliquen, els nivells de sostenibilitat acceptables, etc. També podem elegir com a valors de referència els valors mínims, màxims,

mitjans,...Tot i així en alguns casos podria ser que definir un objectiu no fos viable o que la comparació entre unitats fos dificultosa.

- Ponderació mitjançant càlculs de regressió: Els models de regressió lineals poden ser molt útils ja que ens proporcionen informació sobre el vincle entre un conjunt nombrós de variables i una variable dependent. Si suposem que les variables independents del model lineal són les que hem triat per fer l'indicador i la variable dependent representa un objectiu global, tenim:

$$\hat{Y} = \hat{\alpha} + \widehat{\beta_1} y^{1j} + \cdots + \widehat{\beta_p} y^{pj} \qquad \forall\, 1 \leq j \leq N$$

  Sobre aquesta representació els valors dels coeficients estimats $\hat{\beta}$ poden ser considerats com a factors de ponderació de las diverses variables. S'ha de tenir en compte que les variables han de ser independents entre elles ja que si hi ha multicol·linealitat l'anàlisi es torna deficient. Per això s'aconsella fer una anàlisi exploratòria com la que s'ha explicat anteriorment.

- Anàlisi envoltant de dades (Data Envelopment Analysis): Permet identificar aquelles unitats que tenen un millor rendiment i a partir d'aquí establir un indicador global que a partir del qual s'avaluïn la resta d'unitats. Considerem que les variables seleccionades estan normalitzades de tal manera que a mesura que els valors d'aquestes variables augmenten el rendiment és millor.

- Models de components no observades: La idea, basada en Hall i Jones (1999), se suposa que les variables que construeixen l'indicador sintètic depenen d'una variable no observada més un terme d'error, a través de l'estimació de la variable no observada es pot conèixer les relacions que hi pot haver entre l'indicador i les seves variables, els pesos seran aquells que minimitzin el terme d'error. Aquest procediment és semblant al que s'utilitza en una regressió, la diferència és que en aquest cas no es coneix la variable dependent.

Una vegada s'ha determinat els pesos de cada variable arriba el moment d'agrupar totes les variables en un indicador sintètic, també trobem diferents tècniques per agregar la informació.

El mètode més simple d'agregació de tota la informació és sumar per cada unitat el rànquing que pertany cada una de les p variables, és a dir:

$$I_t^j = \sum_{i=1}^{p} Ranquing_{y_t^{it}} \; \forall \; 1 \leq j \leq N$$

Aquest és un mètode simple i que va bé en els casos que hi hagi valors atípics, tot i que també s'ha de tenir en compte la possible pèrdua d'informació en el valor absolut de les variables que componen l'indicador.

Una altra opció seria comptar el número d'indicadors que estan per sota o per sobre de valors de referència prèviament establerts, seria calcular:

$$I_t^j = \sum_{i=1}^{p} sgn[\frac{y_t^{ij}}{E(y_t^i)} - (1+\delta)] \quad on\ \delta\ és\ un\ valor\ umbral \quad \forall\ 1 \leq j \leq N$$

El valor de δ es determina un cop feta l'anàlisi exploratòria i se sap aproximadament els valors que adquireixen les variables, aquest model tampoc es veu afectat per la presència d'atípics però també es pot perdre informació de la magnitud que poden prendre les variables.

També es pot optar per utilitzar la mitjana aritmètica ponderada que es calcularia com:

$$I_t^i = \sum_{i=1}^{p} w^i y_t^{ij}$$

O de forma similar es pot utilitzar la mitjana geomètrica ponderada:

$$I_t^j = \prod_{i=1}^{p} \left(y_t^{ij}\right)^{w^i}$$

Arribat aquest punt i veient totes les possibles opcions, s'ha escollit utilitzar una mitjana aritmètica ponderada on s'utilitzarà com a ponderació els "*SS loadings*" del model que és la suma de totes les càrregues factorials elevades al quadrat, representen la quantitat de variància de cada variable que s'explica pels factors obtinguts en l'anàlisi factorial, així doncs les ponderacions per cada factor seran les següents:

Variable	PONDERACIÓ
PA1	2,84
PA2	2,09
PA3	1,78
PA4	1,49

TAULA 7: RESUM PONDERACIONS FONT: ELABORACIÓ PRÒPIA

Finalment la mitjana aritmètica ponderada es calcula de la següent manera:

$$\overline{x_p} = \frac{\sum_{i=1}^{N} x_i \cdot w_i}{\sum_{i=1}^{N} w_i}$$

En calcular la mitjana aritmètica ponderada per cada país obtenim l'indicador sintètic, per tal que la seva interpretació sigui més senzilla s'ha escalat el resultat utilitzant la funció *scale()* d'R perquè l'indicador prengui valors entre 0 i 100, la següent taula ens mostra el resultat final ordenat de més gran (valor més favorable) a més petit:

Pais	Indicador	Escalat	Portugal	-0,09	31,62
Germany	1,27	100,00	Iceland	-0,12	30,11
France	1,01	86,85	Cyprus	-0,12	29,72
Italy	0,91	81,83	Netherlands	-0,13	29,38
Spain	0,78	75,29	Malta	-0,23	24,19
Luxembourg	0,49	60,97	Slovenia	-0,33	19,28
Greece	0,49	60,94	Croatia	-0,34	18,69
Norway	0,31	51,86	Latvia	-0,38	16,84
Switzerland	0,22	47,31	Estonia	-0,39	16,26
Sweden	0,20	46,08	Lithuania	-0,40	15,61
Belgium	0,13	42,78	Poland	-0,42	14,74
Ireland	0,13	42,40	Slovakia	-0,53	9,31
Denmark	0,09	40,57	Bulgaria	-0,57	7,12
Austria	-0,02	35,13	Hungary	-0,59	6,37
Finland	-0,02	35,09	Romania	-0,61	4,99
Albania	-0,03	34,67	Czechia	-0,71	0,00

TAULA 8: RESUM INDICADOR FONT:ELABORACIÓ PRÒPIA

Per visualitzar-ho més bé es pot crear una mapa amb la funció *hcmap()* d'R on amb una escala de colors marca amb blau més fort aquelles regions on l'indicador és proper a 100 i amb blau més fluix les properes a 0:

IL·LUSTRACIÓ 14: MAPA INDICADOR FONT: ELABORACIÓ PRÒPIA

És fàcil veure com Alemanya és el país amb un valor més alt, per tant es pot pensar que en aquest país tenen un bon nivell de vida, seguit de França (86,85 punts) i d'Itàlia (81,83); els països amb un indicador més baix, i que per tant la qualitat de vida no és tan bona, són Hongria, Romania i República Txeca, per sota dels 7 punts. Cal destacar que Espanya es trobaria en la posició número 4 amb un valor de 75,29. En general, podem arribar a la conclusió que els països centrals tenen un valor més elevat i els que estan situats a l'Europa oriental tenen valors més baixos.

11. Conclusions

A partir del desenvolupament d'un indicador sintètic creat específicament per aquesta investigació s'ha pogut explorar i analitzar patrons i tendències interessants que reflecteixen les disparitats entre regions. L'anàlisi descriptiu ha revelat que països com Bulgària, on hi ha l'esperança de vida més baixa, Albània, país amb la pitjor taxa de mortalitat infantil, i Letònia amb el risc més alt de pobresa, semblen exhibir nivells més baixos de benestar i un major grau de pobresa en comparació amb països com Espanya que té l'esperança de vida més alta, República Txeca amb un risc baix de patir pobresa i Luxemburg que té un PIB per càpita elevat en comparació amb la resta de països estudiats. També s' observa com hi ha força correlació entre les variables d'estudi, per exemple un PIB per càpita elevat ens indica que el risc de pobresa en aquell país serà baix i la mitjana salarial serà alta, els països on els ingressos i les despeses del govern són elevades hi ha una millor inserció laboral i una esperança de vida més elevada.

Amb l'objectiu d'identificar agrupacions entre països que tenen característiques similars s'ha realitzat una anàlisi clúster amb la qual s'ha observat que els països ubicats a la part oriental d'Europa (Grècia, Bulgària, Romania, Croàcia, Hongria, Polònia i Eslovàquia) tenen similituds entre ells i es diferencien per exemple dels països més nòrdics com Islàndia, Finlàndia, Suècia, i Dinamarca i de l'Europa central on hi trobem Alemanya, Països Baixos, Bèlgica i Àustria.

A través de l'anàlisi factorial s'ha pogut reduir la quantitat de variables i s'aconsegueix explicar, per exemple, la mateixa informació que obtenim de les variables atur, la inserció laboral i el risc de pobresa amb un sol factor, ajudant a fer l'indicador final de manera més fàcil. Per acabar, mitjançant una regressió lineal i utilitzant les estimacions de les betes com a factors de ponderació s'ha pogut obtenir l'indicador desitjat i s'ha pogut observar com els països que amb l'anàlisi clúster pertanyien al mateix grup obtenen un valor similar de l'indicador, cal destacar que els països on l'indicador ha obtingut valors més elevats són Romania, Bulgària, Lituània i Letònia, i els valors més baixos els tenen països com Irlanda, Islàndia i Noruega.

En aquest treball final de grau he pogut assolir els meus objectius de la següent manera:

1) He desenvolupat amb èxit un indicador sintètic que resumeix la informació continguda en altres indicadors parcials. Aquest indicador té com a finalitat observar tant l'evolució com el posicionament dels països de la Unió Europea analitzant el seu benestar.

2) Durant el procés he pogut aprofundir en diverses eines estadístiques i he adquirit coneixements sobre la seva utilització i el seu funcionament. Aquesta comprensió m'ha permès treballar amb eficàcia en l'anàlisi de dades i en la creació de l'indicador.
3) Un dels punts clau del meu treball era poder arribar a algunes conclusions derivades de l'estudi realitzat i establir relacions i patrons relacionats amb el benestar entre països de la UE, cosa que crec que he assolit.
4) En el transcurs de la meva investigació m'he familiaritzat amb les funcionalitats i característiques del llenguatge de programació R. Mitjançant aquesta eina he pogut processar i analitzar les dades de manera eficient així com crear visualitzacions gràfiques que ajuden a comprendre millor les tendències i els resultats obtinguts.

Així doncs he aconseguit els objectius de la meva recerca, aquest treball ha estat una oportunitat per demostrar la meva habilitat per treballar amb les dades i poder aplicar eines estadístiques apreses durant aquests quatre anys del grau d'Estadística.

12. Referències bibliogràfiques

Bolaños Luis. (2020). *RPubs - Análisis Factorial*. https://rpubs.com/luis_bolanos/FA

Coll Morales, F. (s.d.). *Indicador económico - Qué es, definición y concepto | 2023 | Economipedia*. Recuperat 5 juny 2023, de https://economipedia.com/definiciones/indicador-economico.html

Cottrell, S., Van Der Duim, R., Ankersmid, P., & Kelder, L. (2009). Measuring the Sustainability of Tourism in Manuel Antonio and Texel: A Tourist Perspective. *https://doi.org/10.1080/09669580408667247*, *12*(5), 409-431. https://doi.org/10.1080/09669580408667247

de la Fuente Fernández, S. (s.d.). *Análisis Factorial Santiago de la Fuente Fernández*.

Delgado Ronald. (2018). *RPubs - Introducción a los Modelos de Agrupamiento en R*. https://rpubs.com/rdelgado/399475

Domínguez Serrano, M., Blancas Peral, F. J., Guerrero Casas, F. M., & González Lozano, M. (2011). *Una revisión crítica para la construcción de indicadores sintéticos*.

IBM. (s.d.). *Análisis factorial: Rotación - Documentación de IBM*. Recuperat 10 juny 2023, de https://www.ibm.com/docs/es/spss-statistics/saas?topic=analysis-factor-rotation

Nardo, M., Saisana, M., Saltelli, A., & Tarantola, S. (2005). *Tools for Composite Indicators Building*. http://europa.eu.int

Parra Francisco. (2017). *RPubs - Estadística y Machine Learning con R*. https://rpubs.com/PacoParra/293407

Riba, C., & Satorra, A. (2000). *Mètodes Estadístics Aplicats a les Ciències Polítiques i de l'Administració Anàlisi de Components Principals*.

Sanz Francisco. (s.d.). *Algoritmo K-Means - Clustering y cómo funciona*. Recuperat 14 juny 2023, de https://www.themachinelearners.com/k-means/#Que_es_el_Clustering

Schuschny, A., & Soto, H. (s.d.). *Guía metodológica Diseño de indicadores compuestos de desarrollo sostenible*.

Teodoro Luque Martínez. (2000). *Técnicas de análisis de datos en investigación de mercados*. Pirámide.

Tsaur, S. H., Lin, Y. C., & Lin, J. H. (2006). Evaluating ecotourism sustainability from the integrated perspective of resource, community and tourism. *Tourism Management*, *27*(4), 640-653. https://doi.org/10.1016/J.TOURMAN.2005.02.006

Ugwu, O. O., Kumaraswamy, M. M., Wong, A., & Ng, S. T. (2006). Sustainability appraisal in infrastructure projects (SUSAIP): Part 1. Development of indicators and computational methods. *Automation in Construction*, *15*(2), 251. https://doi.org/10.1016/J.AUTCON.2005.05.006

13. Annex

CODI R:

```
## Library

library(readxl)

library(openxlsx)

library(sqldf)

library(openxlsx)

library(corrplot)

library(corrplot)

library(graphics)

library(mapdata)

library(ggplot2)

library(maps)

library(ggrepel)

library(tidyverse)

library(magrittr)

library(ggfortify)

library(dplyr)

library(GPArotation)

library(psych)

library(polycor)

library(ggcorrplot)

library(highcharter)

library(magrittr)

library(dplyr)

## Lectura de la base de dades

dades1 <- read_excel("basedades.xlsx", sheet=1)

dades2 <- read_excel("basedades.xlsx", sheet=2)

dades3 <- read_excel("basedades.xlsx", sheet=3)

dades4 <- read_excel("basedades.xlsx", sheet=4)

#A continuació posem tots els tres fitxer en un de sol amb l'ajuda d'SQL:

dades12 <- merge(dades2,dades1, by = 'Paisos', all.x = TRUE)

dades123 <- merge(dades12,dades3, by = 'Paisos', all.x = TRUE)

dades <- merge(dades123,dades4, by = 'Paisos', all.x = TRUE)

attach(dades)

#Podem començar fent un 'summary' de les nostres variabeles

colnames <- colnames(dades)

resum <- data.frame()

matriu <- matrix(0,7,12)

for (i in 2:ncol(dades)){

  matriu[1,i-1] <- min(dades[,i],na.rm=T)

  matriu[2,i-1] <- quantile(dades[,i], 0.25, na.rm = T)

  matriu[3,i-1] <- median(dades[,i], na.rm = T)

  matriu[4,i-1] <- mean(dades[,i], na.rm = T)

  matriu[5,i-1] <- quantile(dades[,i], 0.75, na.rm = T)

  matriu[6,i-1] <- max(dades[,i],na.rm = T)

  matriu[7,i-1] <- sum(is.na(dades[,i]))


}

colnames(matriu) <- colnames(dades[-1])

rownames(matriu) <- c("Mínim", "25%", "Mediana", "Mitjana", "75%", "Màxim", "NA's")

resum <- as.data.frame(matriu)

# Creem un document de Word i afegim la taula com a contingut

#write.xlsx(as.data.frame(matriu), "resum.xlsx", rownames = T)

#Mirem quin pais te major i menor esperança de vida


Paisos[which.min(Esperança_vida)]

Paisos[which.max(Esperança_vida)]


Paisos[which.min(Mortalitat_infantil)]

Paisos[which.max(Mortalitat_infantil)]
```

```
Paisos[which.min(Risc_pobresa)]

Paisos[which.max(Risc_pobresa)]


Paisos[which.min(PIB)]

Paisos[which.max(PIB)]

#Anàlisis de les correlacions

dadesna <- na.omit(dades)

correlacion<-round(cor(dadesna[-1]), 1)

# windows(width = 20, height = 10)

# par(mar = c(2, 2, 2, 2))

corrplot(correlacion, method="color", type="upper", tl.col = "black",addCoef.col = 'black',number.cex = 0.5)

#Quants NA's tenim per variable?

colnames <- colnames(dades)

resum <- data.frame()

matriu <- matrix(0,1,12)

for (i in 2:ncol(dades)){

  matriu[1,i-1] <- sum(is.na(dades[,i]))


}

colnames(matriu) <- colnames(dades[-1])

rownames(matriu) <- c("NA's")

resum <- as.data.frame(matriu)

#Imputacio de dades

for(i in 2:length(dades)){

dades[which(is.na(dades[,i])),i] <- mean(dades[,i],na.rm=T)

}


#attach(dades)

## Distància de mahalanobis

mahalanobis <- sort(mahalanobis(dades[,-1], colMeans(dades[,-1]), cov(dades[,-1])), decreasing = TRUE)

boxplot(mahalanobis, ylab = "Distància de Mahalanobis")


x <- dades[,-1]

m <- colMeans(x)

xc <- scale(x,center=T,scale=F)

s <- cov(x)

D2 <- diag(xc%*%solve(s)%*%t(xc))

print(D2)


vc <- qchisq(.99, df = 12)

sg <- pchisq(D2, df =12, lower.tail = FALSE)

print(vc)

print(sg)

conclusion <- ifelse(sg>0.01, "No Atípic", "Atípic")

res <- data.frame(D2,vc,sg,conclusion)

print(res)

#write.xlsx(res, "Distancia mahalanobis.xlsx", rownames=T)

## Tornem a fer una descriptiva de les dades

colnames <- colnames(dades)

resum <- data.frame()

matriu <- matrix(0,7,12)

for (i in 2:ncol(dades)){

  matriu[1,i-1] <- min(dades[,i],na.rm=T)

  matriu[2,i-1] <- quantile(dades[,i], 0.25, na.rm = T)

  matriu[3,i-1] <- median(dades[,i], na.rm = T)

  matriu[4,i-1] <- mean(dades[,i], na.rm = T)

  matriu[5,i-1] <- quantile(dades[,i], 0.75, na.rm = T)

  matriu[6,i-1] <- max(dades[,i],na.rm = T)

  matriu[7,i-1] <- sum(is.na(dades[,i]))


}

colnames(matriu) <- colnames(dades[-1])

rownames(matriu) <- c("Mínim", "25%", "Mediana", "Mitjana", "75%", "Màxim", "NA's")

resum <- as.data.frame(matriu)

# Creem un document de Word i afegim la taula com a contingut

#write.xlsx(as.data.frame(matriu),
```

```
"resum_despues.xlsx", rownames = T)

## correlation plot

dades <- na.omit(dades)

correlacion<- round(cor(dades[-1]), 1)

corrplot(correlacion, method="color", type="upper", tl.col = "black",addCoef.col = 'black',number.cex = 0.5)

## Normalització de les dades

dades_norm <- dades

n <- c(6,7,8,11,13)

for (i in n){

  dades_norm[,i] <- (dades[,i]- mean(dades[,i]))/sd(dades[,i])


}

summary(dades_norm[,c(6,7,8,11,13)])

sd(dades_norm[,8])

data <- dades_norm

## Anàlisi clúster

datat <- data

rownames(datat) <- datat[,1]

rownames(datat)[[12]] <- "Germany"

datat <- datat[,-1]

hc <- hclust(dist(datat), method = "centroid")

plot(hc)

rect.hclust(hc,8,border = "red")

cluster <- cutree(hc, 8)  # 'num_clusters' es el número de clústeres deseados

table(cluster)

# Añadimos la variable grupo a la base de datos

datat$Cluster <- with(datat,cluster)

#head(datat)

## kmeans


# Realizar el algoritmo de k-means

set.seed(1234)

k <- 5  # Número de clústeres deseado

resultado <- kmeans(datat, centers = k)

# Obtener los resultados del algoritmo

centroides <- resultado$centers  # Coordenadas de los centroides de los clústeres

asignaciones <- resultado$cluster  # Asignaciones de los elementos a los clústeres

datat$Cluster_kmeans <- with(datat,asignaciones)

set.seed(1234)

wcss <- vector()

for(i in 1:20){

  wcss[i] <- sum(kmeans(datat, i)$withinss)

}

ggplot() + geom_point(aes(x = 1:20, y = wcss), color = '#668B8B') +

geom_line(aes(x = 1:20, y = wcss), color = '#668B8B') +

  ggtitle("Métode del colze") +

  xlab('Quantitat de centroides k') +

  ylab('WCSS')



ggplot() + geom_point(aes(x = datat$Esperança_vida, y = datat$PIB, color = Cluster_kmeans), data = datat, size = 2) +


scale_colour_gradientn(colours=rainbow(4)) +

  geom_point(aes(x = resultado$centers[, 1], y = resultado$centers[, 2]), color = 'black', size = 3) +

  ggtitle('K-means amb 5 clústers') +

  xlab('X') + ylab('Y')

autoplot(resultado, datat, frame=TRUE)

## Creació d' un mapa

#head(datat)

dades_mapa <- data.frame(rownames(datat), datat$Cluster)

colnames(dades_mapa) <- c("Pais", "Cluster")

dades_mapa$Pais[7] <- "Czech Republic"

#dades_mapa

# hcmap("custom/europe", showInLegend = FALSE) %>%

#   hc_title(text = "Mapa dels clústers")
```

```
mapdata <-
get_data_from_map(downl
oad_map_data("custom/eu
rope"))

glimpse(mapdata)

valores <- c(1,2,3,4,5,6,7,8)

colores <- c("coral",
"indianred1", "gold2",
"lemonchiffon",
"aquamarine4", "yellow2",
"deepskyblue3",
"lightblue4")

data_colores <-
cbind(valores, colores)

hcmap(

  map = "custom/europe",

  data = dades_mapa,

  joinBy = c("name","Pais"),

  name = "Clusters",

  value = "Cluster",

  # tooltip = list(pointFormat
= "{point.name} {point.tz}"),

  #dataLabels = list(enabled
= TRUE, format =
"{point.country}")

) %>%

  hc_colorAxis(

    dataClassColor =
"category",

    # dataClasses = dta_clss,

    min = min(valores),

    max = max(valores),

    stops =
color_stops(length(valores),
colores),

    labels = list(enabled =
FALSE)

  ) %>%

  hc_legend(title = list(text =
"Valores"), enabled = TRUE,
layout = "vertical",

           align = "right",
verticalAlign = "middle",
floating = TRUE,

           valueDecimals = 0,
valueSuffix = " país",
symbolHeight = 12,
symbolWidth = 24)%>%

  hc_title(text = "Clusters")

## Mapa k means


#head(datat)

dades_mapa_kmeans <-
data.frame(rownames(datat
), datat$Cluster_kmeans)

colnames(dades_mapa_km
eans) <- c("Pais", "Cluster")

dades_mapa_kmeans$Pais[
7] <- "Czech Republic"

#dades_mapa


# hcmap("custom/europe",
showInLegend = FALSE)
%>%

#   hc_title(text = "Mapa
dels clústers")


mapdata <-
get_data_from_map(downl
oad_map_data("custom/eu
rope"))

glimpse(mapdata)

valores <- c(1,2,3,4,5)

colores <- c("cadetblue3",
"coral", "gold2",
"mediumpurple3",
"darkolivegreen3")

data_colores <-
cbind(valores, colores)

hcmap(

  map = "custom/europe",

  data =
dades_mapa_kmeans,

  joinBy = c("name","Pais"),

  name = "Clusters",

  value = "Cluster",

  # tooltip = list(pointFormat
= "{point.name} {point.tz}"),

  #dataLabels = list(enabled
= TRUE, format =
"{point.country}")

) %>%

  hc_colorAxis(

    dataClassColor =
"category",

    # dataClasses = dta_clss,

    min = min(valores),

    max = max(valores),

    stops =
color_stops(length(valores),
colores),

    labels = list(enabled =
FALSE)

  ) %>%

  hc_legend(title = list(text =
"Valores"), enabled = TRUE,
layout = "vertical",

           align = "right",
verticalAlign = "middle",
floating = TRUE,

           valueDecimals = 0,
valueSuffix = " país",
symbolHeight = 12,
symbolWidth = 24)%>%

  hc_title(text = "Clusters")


##Anàlisis factorial

Paso 1:Calcular la matriz de
correlación policorica
```

```
mat_cor <- round(cor(data[,-1]),1) #matriu de correlació

ggcorrplot(mat_cor,type="lower",ggtheme = theme_classic, hc.order = T, lab = T, lab_size=2.5)


#Verificar que la matriu es factorizable:


#mat_cor <- hetcor(data[,-1])$correlations

p_esf <- cortest.bartlett(mat_cor, n= 31, diag = TRUE)

p_esf$p


kmo <- KMO(data[,-1])

kmo

#cortest.bartlett(data[,-1])

#Anàlisis factorial:


modelo2<-fa(mat_cor,

        nfactors = 3,

        rotate = "none",

        fm="minres") # modelo minimo residuo


######comparando las comunalidades

c1 <- sort(modelo1$communality,decreasing = T)

c2 <- sort(modelo2$communality,decreasing = T)

head(cbind(c1,c2))


####comparacion de las unicidades

sort(modelo1$uniquenesses,decreasing = T)->u1

sort(modelo2$uniquenesses,decreasing = T)->u2

head(cbind(u1,u2))


#Mirar quants factors són necessaris


scree(mat_cor)

fa.parallel(mat_cor,n.obs=200,fa="fa",fm="minres")

#Rotació

#Rotacio

rot<-c("none", "varimax", "oblimin","equamax")

bi_mod<-function(tipo){

  biplot.psych(fa(data[,-1],nfactors = 2,fm="minres",rotate = tipo),main = paste("Biplot amb rotació ",tipo),col=c(2,3,4),pch = c(21,18),group = bfi[,"gender"])

}

sapply(rot,bi_mod)


modelo_varimax<- fa(mat_cor,nfactors = 4,rotate = "varimax",

                    fa="pa")

fa.diagram(modelo_varimax)

print(modelo_varimax$loadings,cut=0)


#Anàlisi final


model<-fa(data[,-1],

        nfactors = 4,

        rotate = "varimax",

        fm="pa")

df <- data.frame(data, model$scores)

df_factors <- df[,c(1,14,15,16,17)]

## Agregacio

names(model)

ponderacions <- c(2.087, 2.840, 1.776, 1.489)


indicador <- c()

mitjana_ponderada <- c()

for( i in 1:31){

  indicador[i] <- ponderacions[1]*df_eigen$PA1[i] + ponderacions[2]*df_eigen$PA2[i] + ponderacions[3]*df_eigen$PA3[i] + ponderacions[4]*df_eigen$PA4[i]

  mitjana_ponderada[i] <- indicador[i]/sum(ponderacions)

}

df_indicador <- data.frame(data[,1], mitjana_ponderada)

colnames(df_indicador) <- c("Pais", "Indicador")

df_indicador$Escalat <- rescale(df_indicador$Indicador, to = c(0,100))
```

```
#write.xlsx(df_indicador,
"indicador final.xlsx",
rownames = T)


summary(df_indicador$Esca
lat)


mapdata <-
get_data_from_map(downl
oad_map_data("custom/eu
rope"))
df_indicador$Pais[7] <-
"Czech Republic"
df_indicador$Pais[12] <-
"Germany"


hcmap(
  map = "custom/europe",
  data = df_indicador,
  joinBy = c("name","Pais"),
  name = "Indicador",
  value = "Escalat",
  tooltip = list(pointFormat =
"{point.name} {point.tz}"),
  dataLabels = list(enabled =
TRUE, format =
"{point.country}")
) %>%
  hc_colorAxis(
    dataClassColor =
"category"
    #dataClasses = dta_clss
  ) %>%
  hc_title(text = "Indicador")
```