Analiza datelor. Tehnici de analiză

Proiectul învățământului superior din Moldova – DATASEA

Răzvan Rughiniș

razvan.rughinis@upb.ro

Concepte principale

Analiza factorială

Analiza cluster

Analiza de regresie

Analiza de rețea

Serii de timp

Obiective ale tehnicilor de analiză

- Explorarea datelor
  - Frecvențe, medii și corelații
- Măsurare: analiza factorială / reducerea dimensionalității
  - Măsurarea constructelor latente prin indicatori vizibili
  - Identificarea dimensiunilor personalității pe baza acțiunilor și preferințelor
- Clasificare: analiza cluster
  - Clasificarea indivizilor în tipuri: „Spune-mi cu cin’ te-nsoțești, ca să-ți spun cine ești”
- Explicare prin factori externi: analiza de regresie
- Explicare prin contagiune: analiza de rețea
- Extrapolare: seriile de timp
  - Explicarea acțiunilor prin factori externi și patternuri temporale
  - Tendințe inerțiale, sezoniere, variații aleatorii

https://tinyurl.com/sondaj-relatii-2023

Condițiile unei relații de succes

Indicatorii

Aici sunt câteva aspecte despre care unii oameni cred că sunt importante pentru o căsătorie de succes. Te rog, pentru fiecare, spune-mi cum crezi că este, pentru succesul unei căsătorii: (2=foarte important, 1=destul de important, 0=nu prea important)

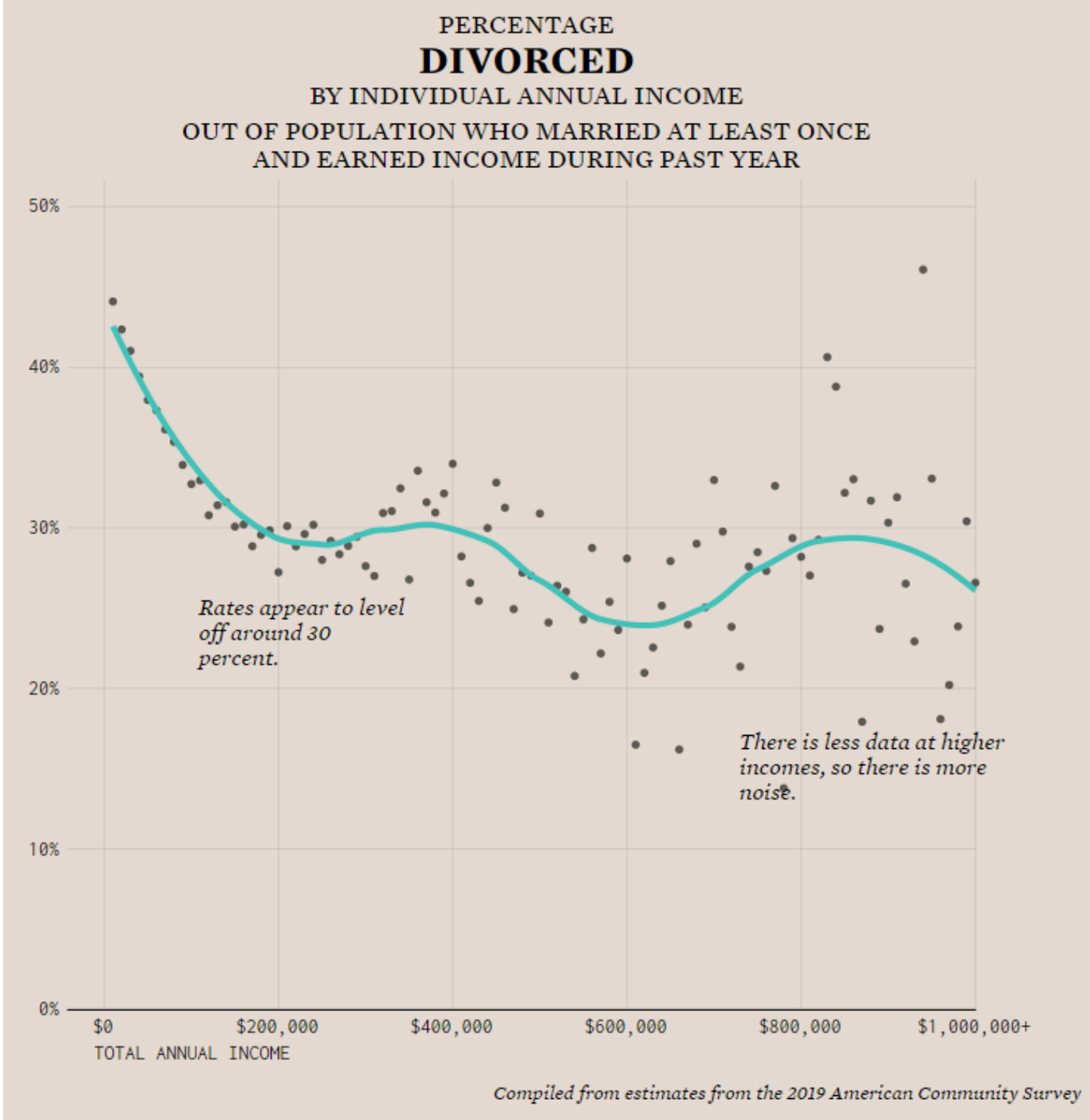
- Fidelitatea
- Partenerii să fie din același mediu social
- Să aibă aceeași religie
- Să aibă o locuință bună
- Un venit adecvat
- Să fie de acord în chestiunile politice
- Să trăiască separat de socri
- Să aibă o relație sexuală fericită
- Să împartă treburile domestice
- Să aibă copii

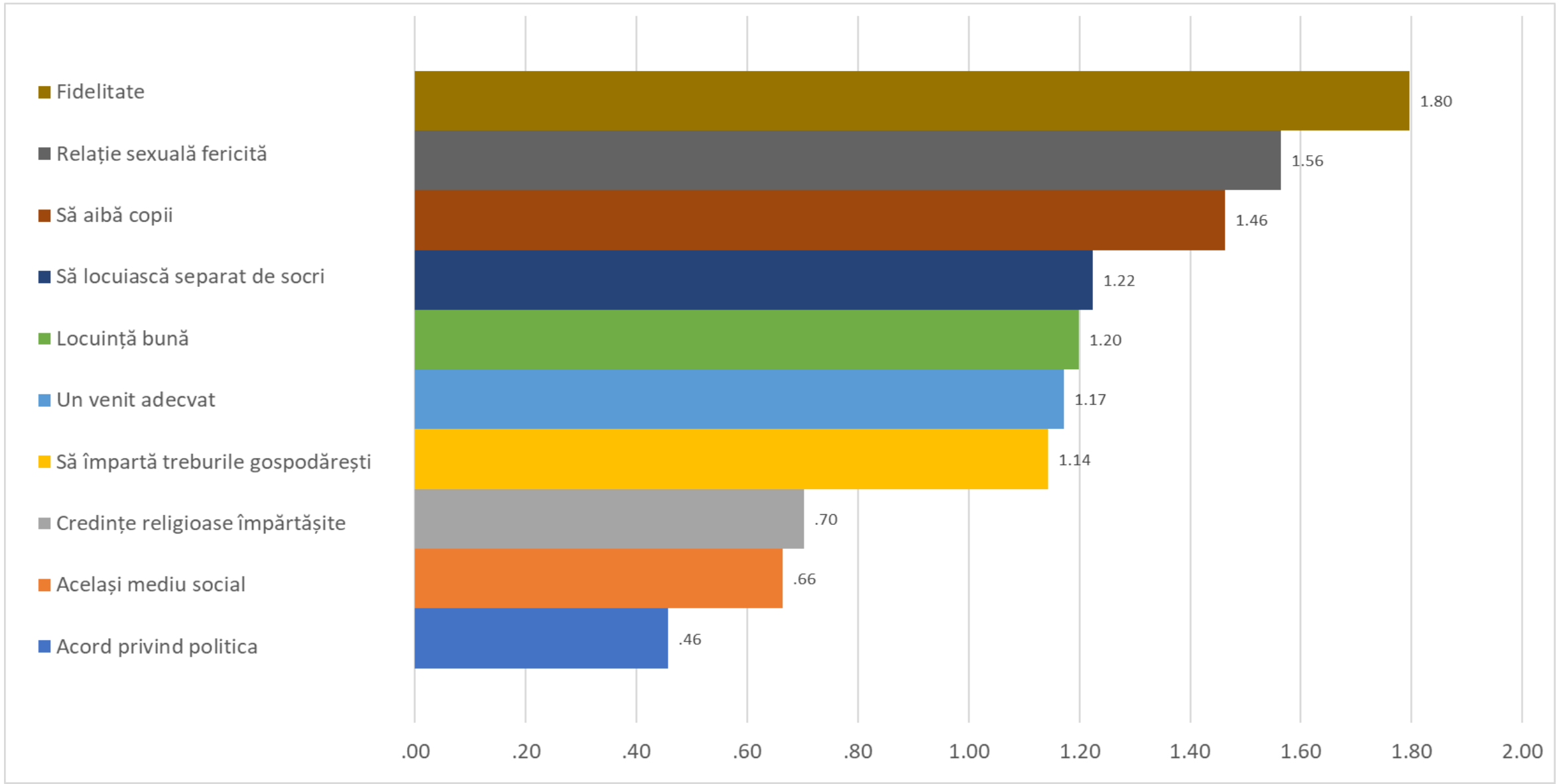
Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

Percepție vs. realitate

- O relație clară între venit și divorțialitate
  - Cel puțin până la USD 200,000/an
- Deși categoriile de venituri mari au o variabilitate mult mai mare
  - Au mai puțini indivizi

Sursa





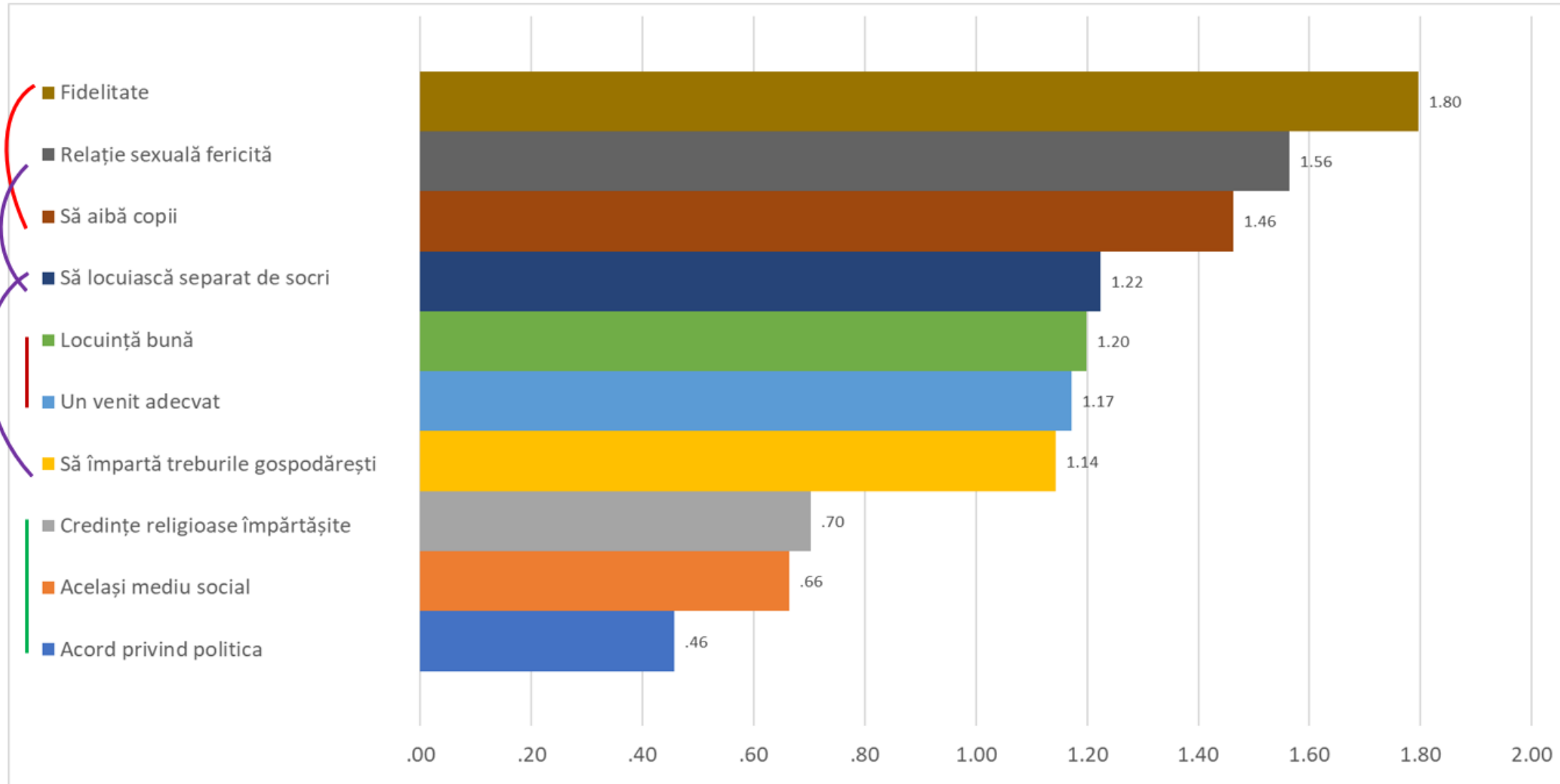
Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

Tabelul de corelații între indicatori

	Fidelitate	Copii	Venit	Casă	Același mediu	Aceeași religie	Acord în politică	Locuiesc singuri	Relație sexuală	Împart treburile
Fidelitate	1	.223	.079	.088	.058	.160	.028	-0.021	.041	.085
Copii	.223	1	.138	.177	.102	.192	.058	.034	.111	.175
Venit	.079	.138	1	.454	.298	.146	.143	.090	.134	.096
Casă	.088	.177	.454	1	.252	.191	.186	.116	.144	.185
Același mediu	.058	.102	.298	.252	1	.355	.301	.079	.027	.063
Aceeași religie	.160	.192	.146	.191	.355	1	.316	-.013	-.027	.074
Acord în politică	.028	.058	.143	.186	.301	.316	1	.065	.051	.122
Locuiesc singuri	-.021	.034	.090	.116	.079	-.013*	.065	1	.247	.137
Relație sexuală	.041	.111	.134	.144	.027	-.027	.051	.247	1	.233
Împart treburile	.085	.175	.096	.185	.063	.074	.122	.137	.233	1

Coeficienți de corelație Bravais-Pearson

Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

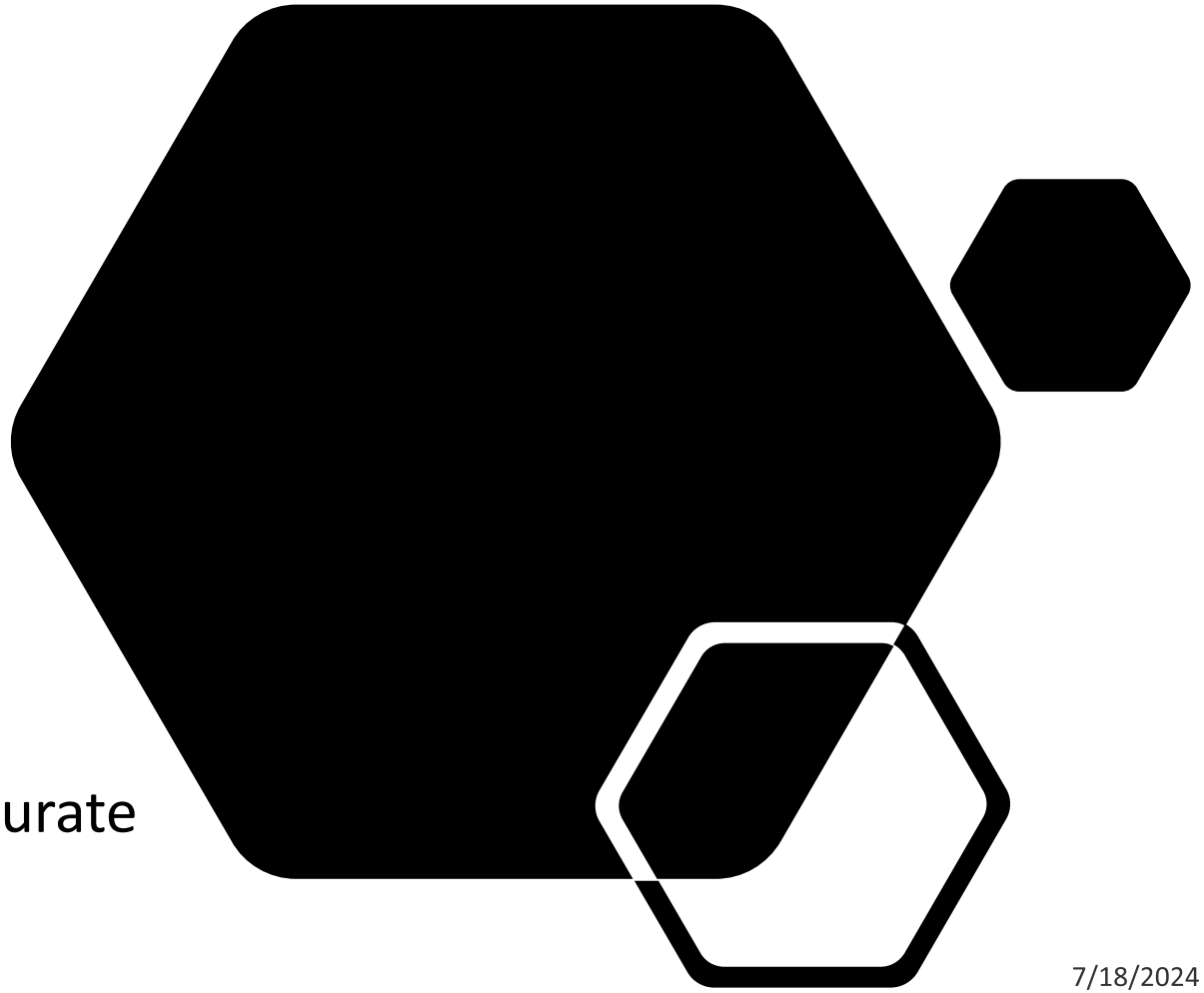


Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

Analiza factorială

Reducerea dimensionalității

Trecerea de la indicatori la constructele măsurate (dimensiuni)

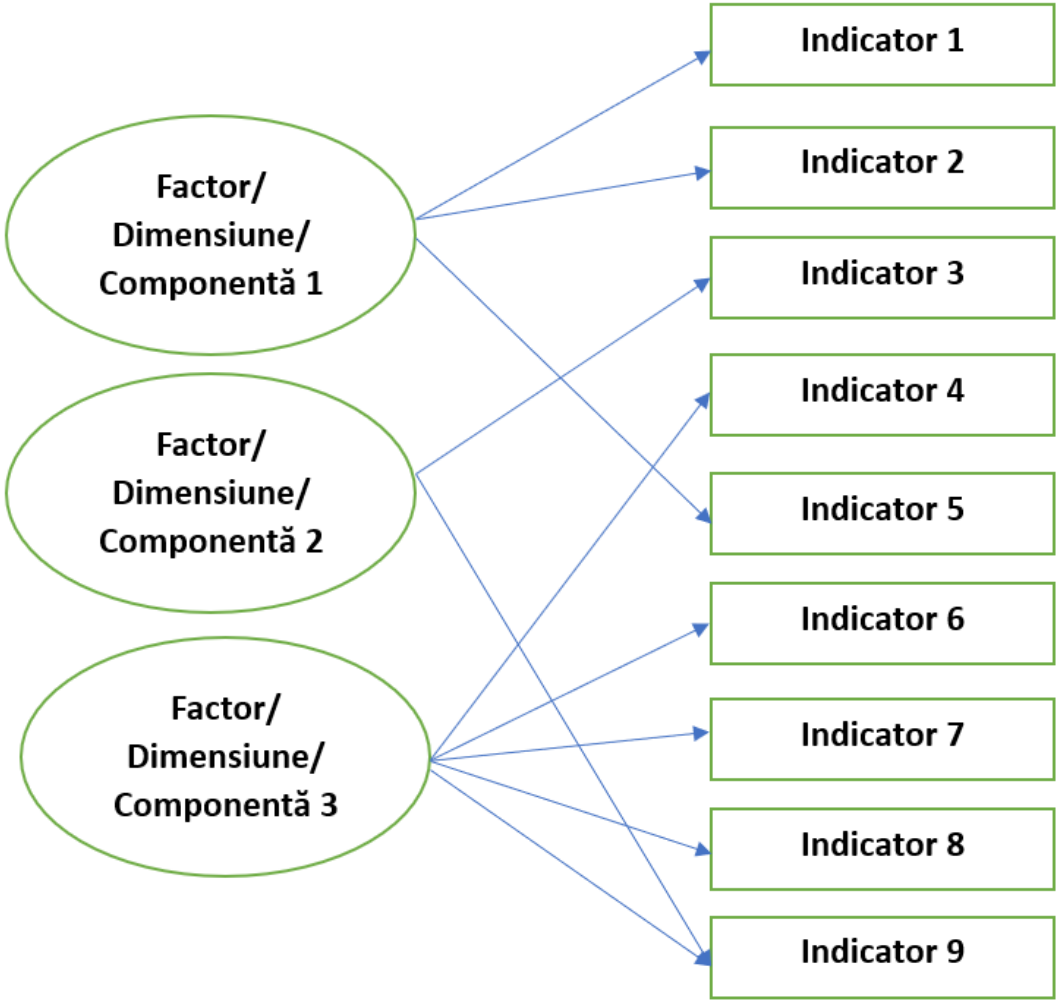


Indicatori vs. factori, dimensiuni sau componente

- Model reflectiv

- Indicatorii măsoară factorii
- Factorii se mai numesc dimensiuni sau componente

- Indicatorii sunt observabili direct
- Factorii sunt latenți sau neobservați

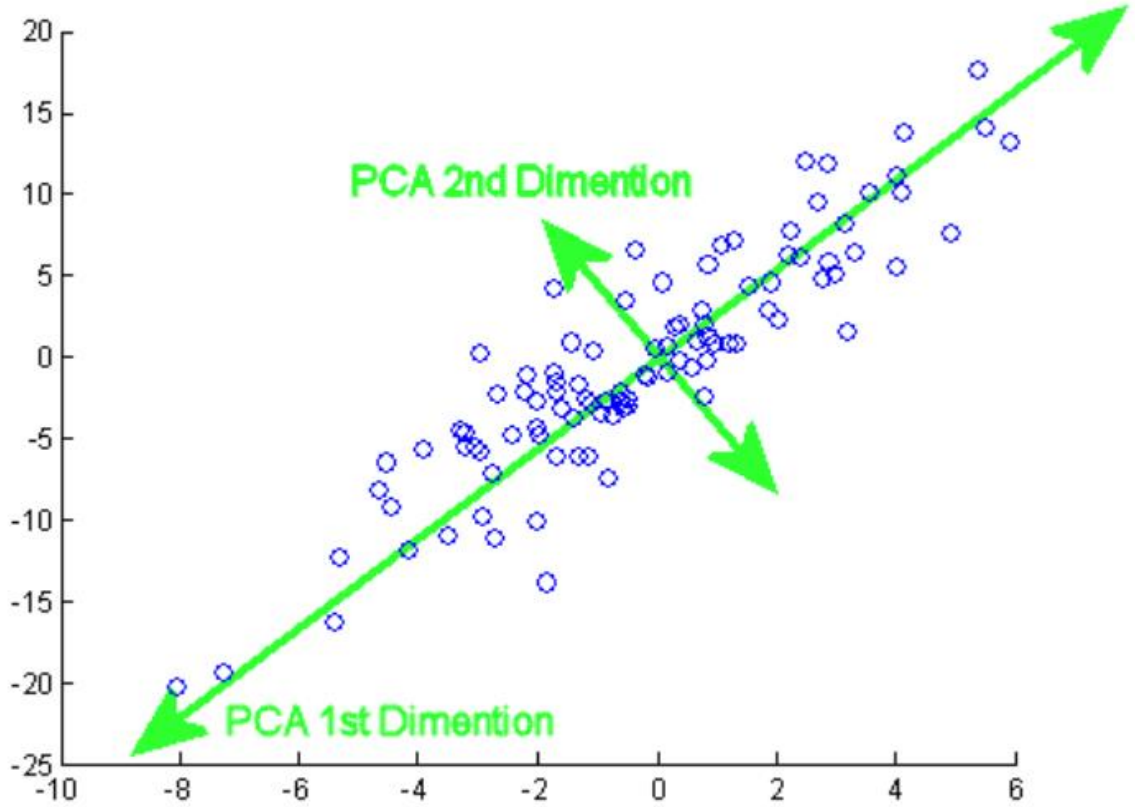


Analiza factorială a ***n*** indicatori

- Estimează...
  - **Forțele** care se află în spatele unei diversități de simptome corelate
    - Acestea sunt factori / dimensiuni
    - Interpretăm semnificația factorilor în funcție de simptomele asociate
  - **Conceptele măsurate** de ***n*** indicatori corelați
- O analiză factorială **generează *n* variabile noi**
  - Fiecare factor e o nouă variabilă, cu valori pentru toate cazurile analizate
    - „Scorul factorial” are valori continue, reale
    - Extragem mai puțini factori decât indicatorii inițiali
- Fiecare individ analizat primește valori estimate pentru fiecare factor extras

Reducerea dimensionalității

- Care sunt cele mai importate dimensiuni ale datelor mele?
  - Compresia datelor
  - Descoperirea structurilor fundamentale
  - Eliminarea erorilor de măsurare
  - Vizualizarea volumelor uriașe de date
  - Extragerea de date pentru antrenarea altor modele
- Exemplu grafic
  - Doi factori / două dimensiuni
  - Dimensiunea 1 este mai importantă



Sursă

Etapele analizei factoriale

- Selectăm factorii cei mai relevanți
- Interpretăm factorii în funcție de corelațiile lor cu indicatorii
  - Etichetăm factorii în baza de date

Patru factori identificați pentru 10 itemi: cum îi interpretăm?

Fiecare factor este o nouă variabilă

Fiecare respondent are o valoare pentru fiecare factor

Similaritate

Intimitate

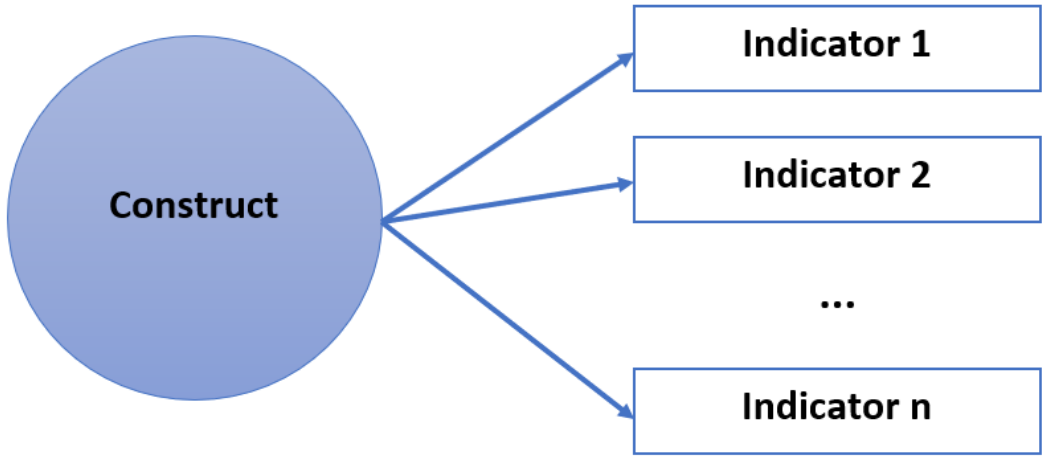
Familia tradițională

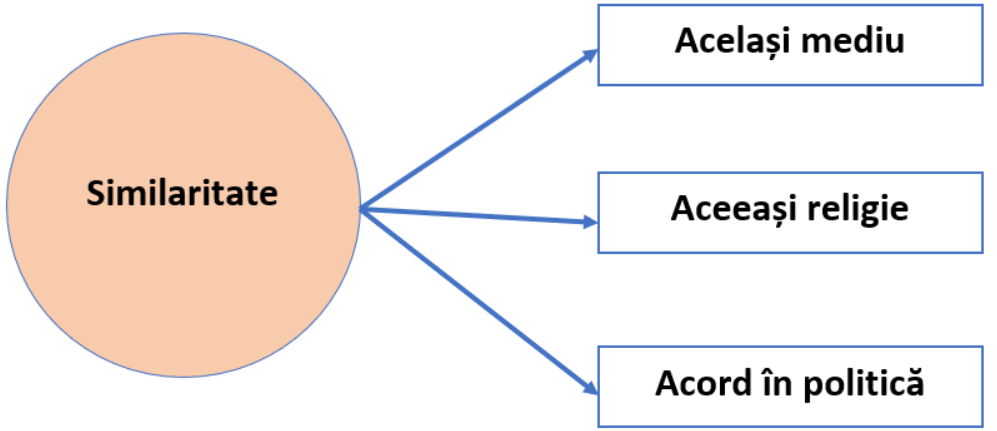
Banii NU contează

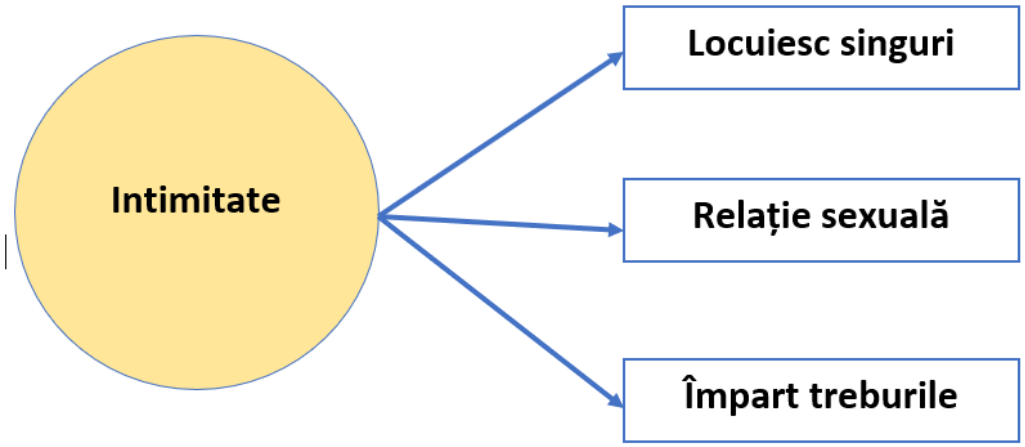
	Factor			
	1	2	3	4
Fidelitate	-.002	-.070	.766	.005
Copii	.016	.118	.701	-.100
Venit	-.022	-.024	.003	-.875
Locuință bună	.038	.081	.077	-.771
Același mediu	.619	-.058	-.080	-.316
Aceeași religie	.737	-.121	.244	.028
Acord în politică	.792	.140	-.113	.078
Locuiesc singuri	.053	.680	-.224	-.035
Relație sexuală	-.122	.726	.053	-.087
Împart treburile	.095	.606	.270	.068

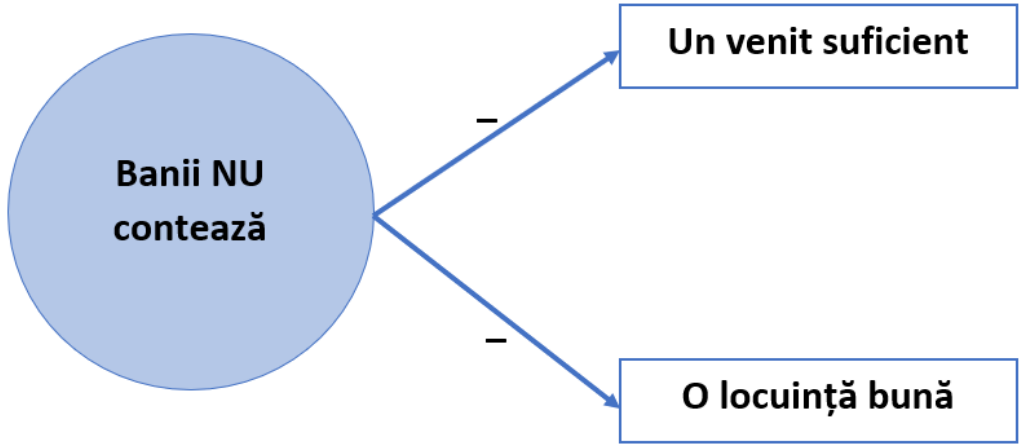
Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

Modelul reflectiv de măsurare



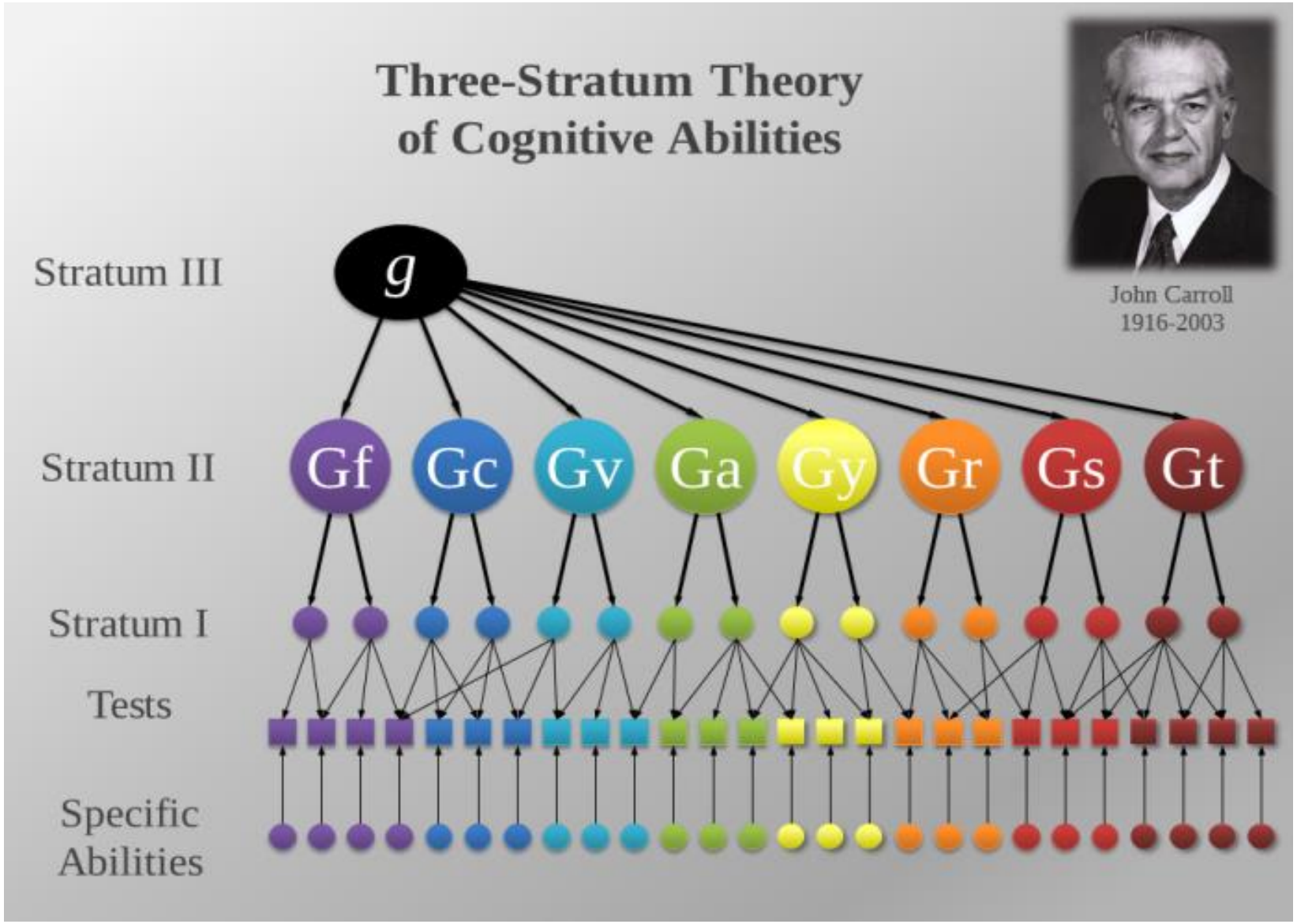






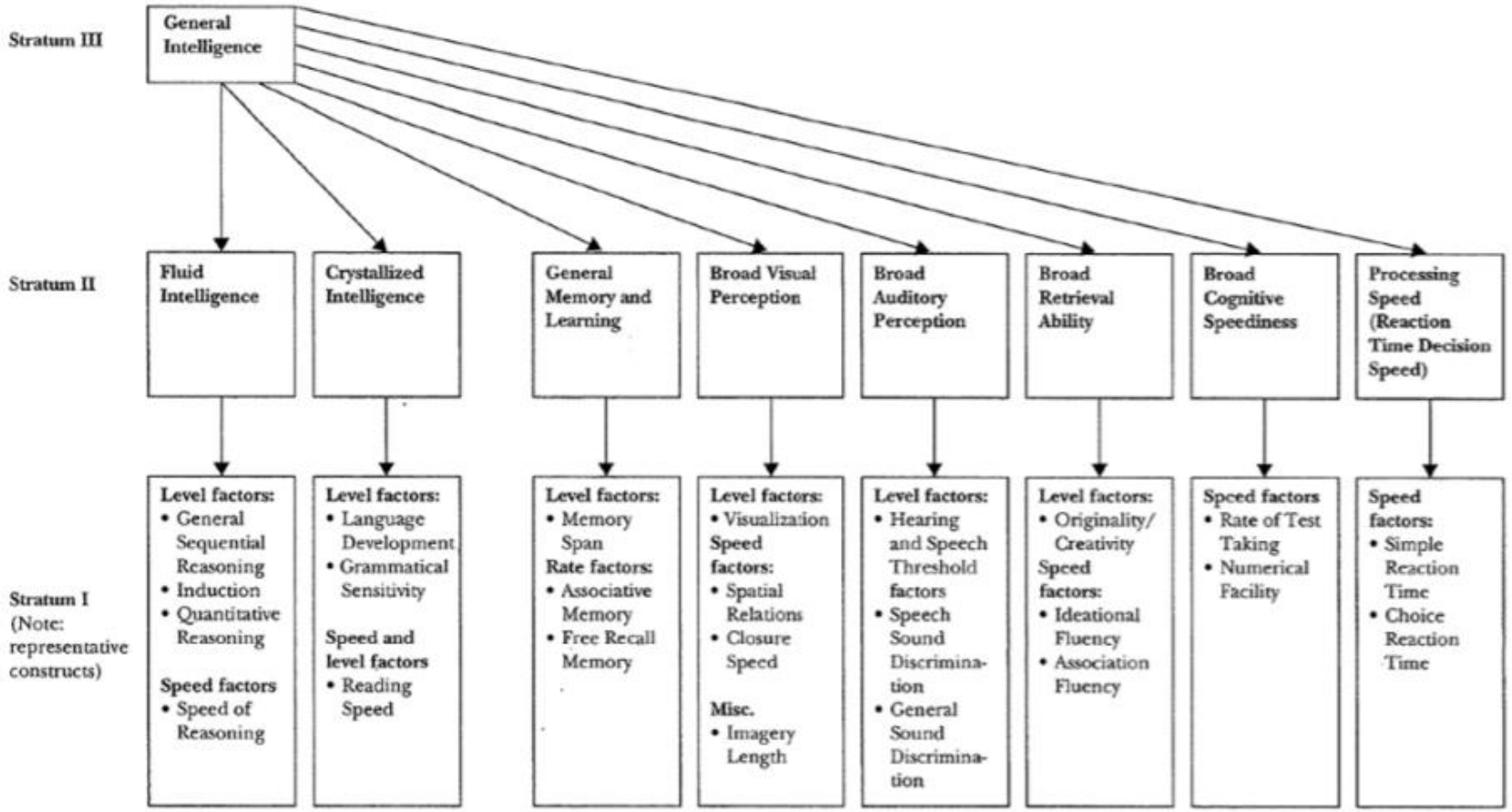
Măsurarea inteligenței & IQ

- 3 straturi



J. B. Carroll (1993), *Human cognitive abilities: A survey of factor-analytic studies*, Cambridge University Press, New York, NY, USA. [Grafic]

Carroll's Three-Stratum Theory of Cognitive Ability

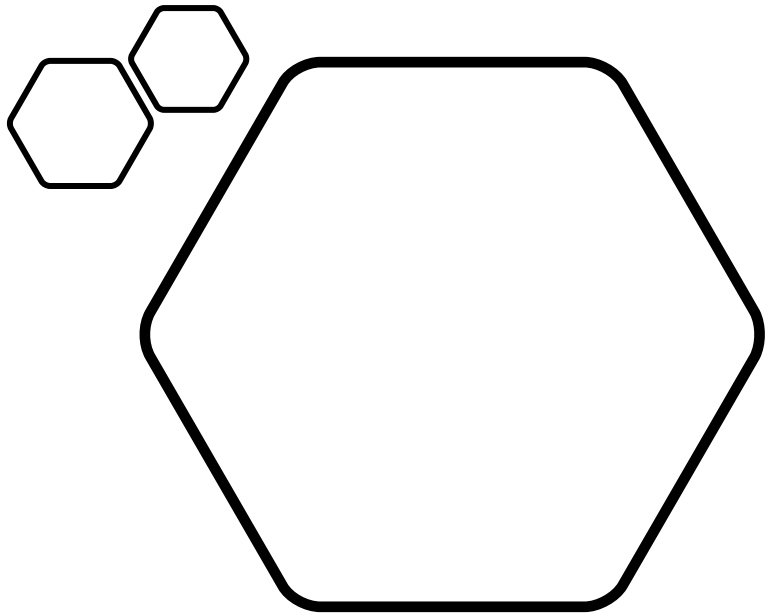


Adapted with permission from Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York: Cambridge University Press. (p. 626).

6

McDaniel, 2018

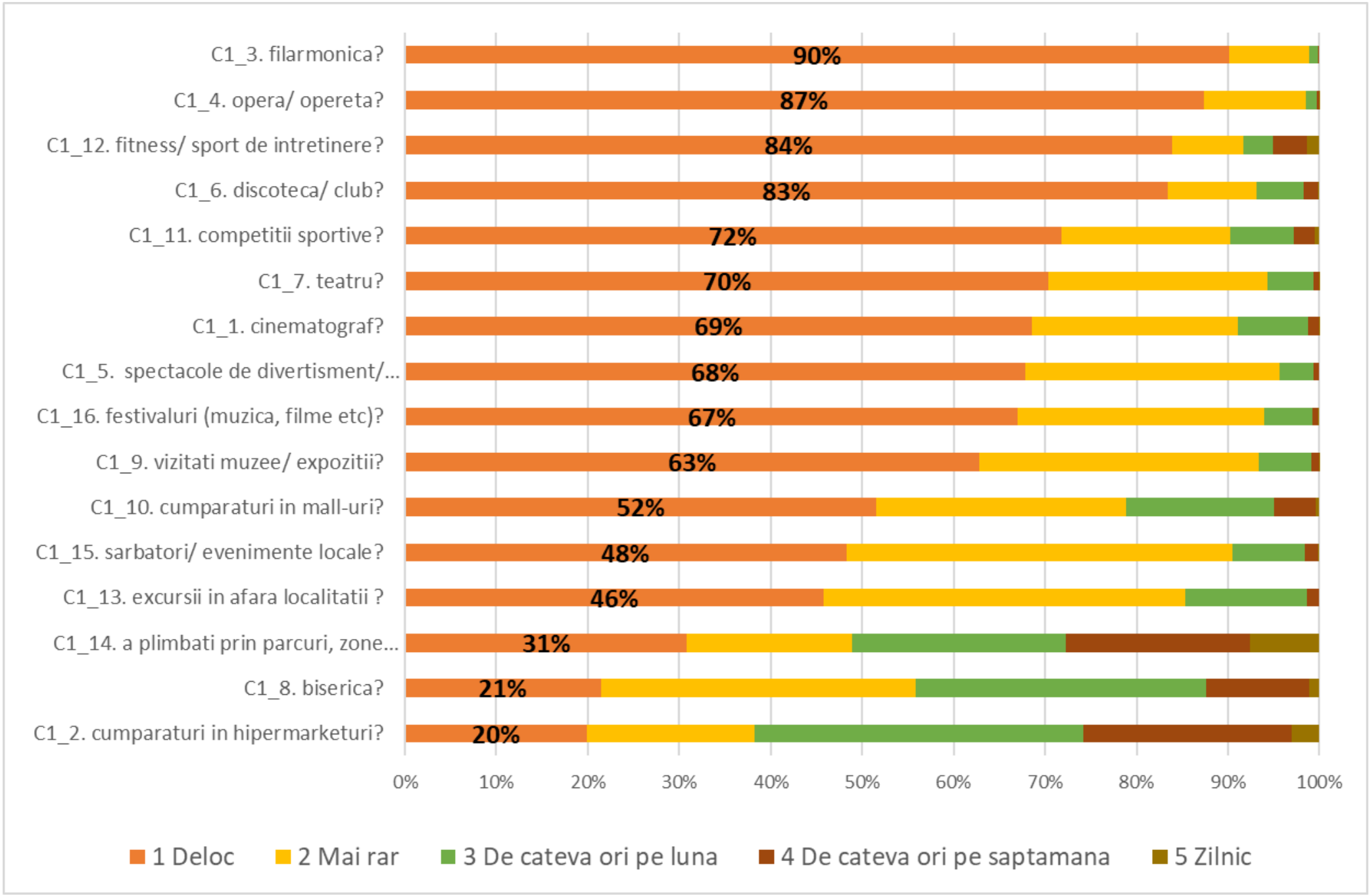
Studiu de caz



Comportamentul de consum cultural

Sursă: Barometrul de Consum Cultural - BCC 2012

Cât de des mergeți la...



Etapa 1- Tabelul de corelații între indicatori

De obicei, cat de des mergeti la...	C1_1	C1_3	C1_4	C1_5	C1_8	C1_14	C1_2	C1_10	C1_6	C1_7	C1_9	C1_11	C1_12	C1_13
C1_1. cinematograf	1.00	.286	.340	.434	-.082	.319	.290	.406	.447	.440	.410	.273	.358	.372
C1_3. filarmonica		1.00	.647	.284	.055	.198	.138	.197	.125	.407	.326	.156	.128	.207
C1_4. opera			1.00	.340	.045	.248	.177	.231	.118	.450	.374	.140	.247	.258
C1_5. spectacole de divertisment/ muzica				1.00	0.02	.302	.299	.357	.378	.431	.449	.295	.309	.379
C1_8. biserica					1.00	.086	.070	.000	-.079	.000	.057	.030	-.010	.020
C1_14. va plimbati prin parcuri, zone verzi						1.00	.371	.367	.167	.315	.336	.225	.264	.424
C1_2. cumparaturi in hipermarketuri							1.00	.424	.201	.240	.296	.229	.156	.379
C1_10. cumparaturi in mall-uri								1.00	.316	.324	.385	.274	.278	.401
C1_6. discoteca/ club									1.00	.224	.209	.242	.326	.282
C1_7. teatru										1.00	.463	.219	.285	.338
C1_9. muzee/ expozitii											1.00	.305	.296	.434
C1_11. asistati la competitii sportive												1.00	.262	.309
C1_12. faceti fitness													1.00	.354
C1_13. excursii in afara localitatii ?														1.00

Etapa 1- Tabelul de corelații între indicatori

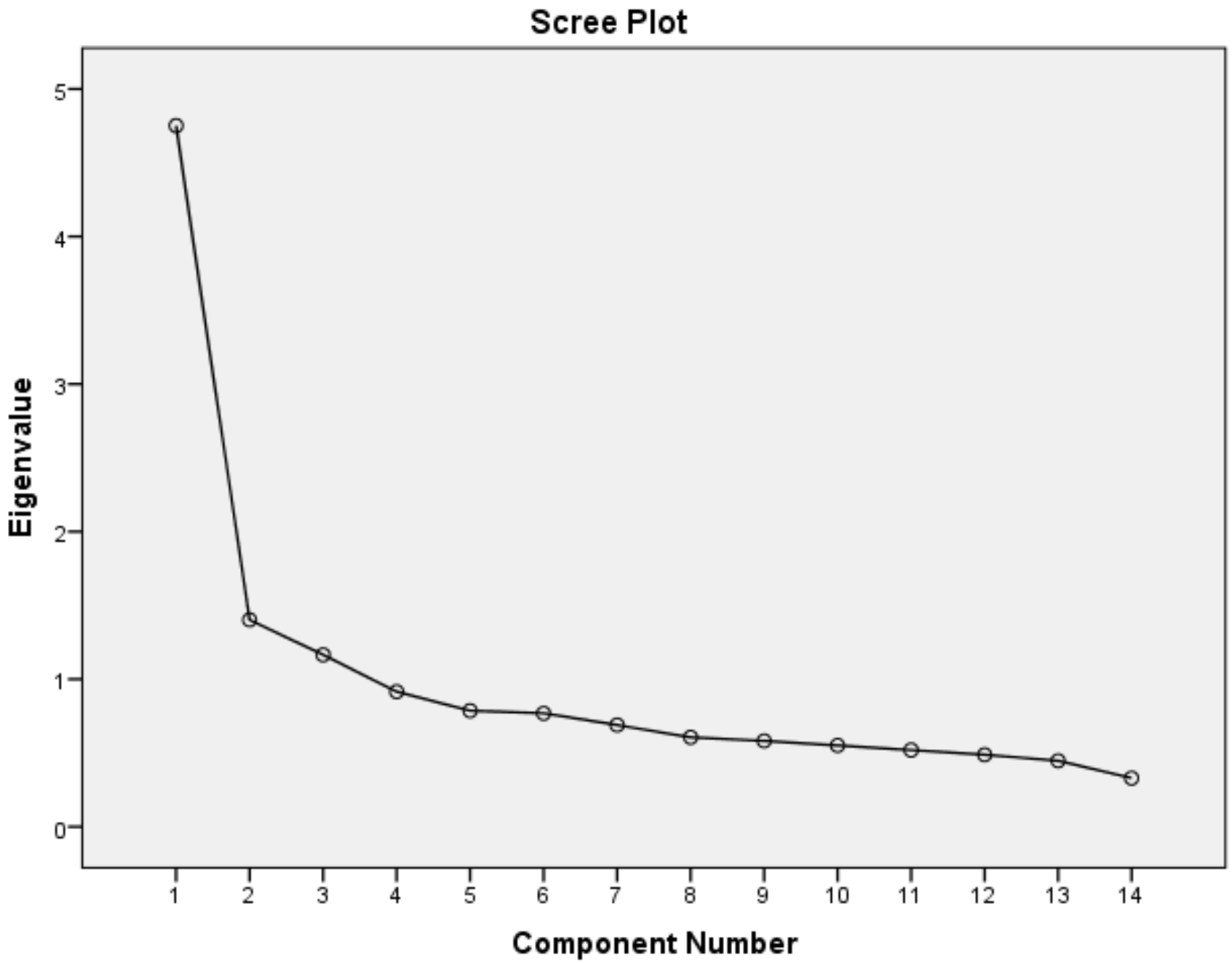
De obicei, cat de des mergeti la...	C1_1	C1_3	C1_4	C1_5	C1_8	C1_14	C1_2	C1_10	C1_6	C1_7	C1_9	C1_11	C1_12	C1_13
C1_1. cinematograf	1.00	.286	.340	.434	-.082	.319	.290	.406	.447	.440	.410	.273	.358	.372
C1_3. filarmonica		1.00	.647	.284	.055	.198	.138	.197	.125	.407	.326	.156	.128	.207
C1_4. opera			1.00	.340	.045	.248	.177	.231	.118	.450	.374	.140	.247	.258
C1_5. spectacole de divertisment/ muzica				1.00	0.02	.302	.299	.357	.378	.431	.449	.295	.309	.379
C1_8. biserica					1.00	.086	.070	.000	-.079	.000	.057	.030	-.010	.020
C1_14. va plimbati prin parcuri, zone verzi						1.00	.371	.367	.167	.315	.336	.225	.264	.424
C1_2. cumparaturi in hipermarketuri							1.00	.424	.201	.240	.296	.229	.156	.379
C1_10. cumparaturi in mall-uri								1.00	.316	.324	.385	.274	.278	.401
C1_6. discoteca/ club									1.00	.224	.209	.242	.326	.282
C1_7. teatru										1.00	.463	.219	.285	.338
C1_9. muzee/ expozitii											1.00	.305	.296	.434
C1_11. asistati la competitii sportive												1.00	.262	.309
C1_12. faceti fitness													1.00	.354
C1_13. excursii in afara localitatii ?														1.00

Etapa 2 – Selecția factorilor relevanți

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.751	33.937	33.937	4.751	33.937	33.937
2	1.402	10.015	43.952	1.402	10.015	43.952
3	1.165	8.323	52.275	1.165	8.323	52.275
4	.915	6.537	58.812			
5	.786	5.615	64.427			
6	.769	5.491	69.919			
7	.689	4.923	74.842			
8	.606	4.328	79.170			
9	.581	4.153	83.323			
10	.551	3.934	87.257			
11	.520	3.711	90.969			
12	.488	3.484	94.453			
13	.447	3.194	97.646			
14	.330	2.354	100.000			

Extraction Method: Principal Component Analysis.


Scree Plot
Eigenvalue
0
1
2
3
4
5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Component Number

Etapa 3 – Etichetarea și interpretarea factorilor

Component Matrix[a]

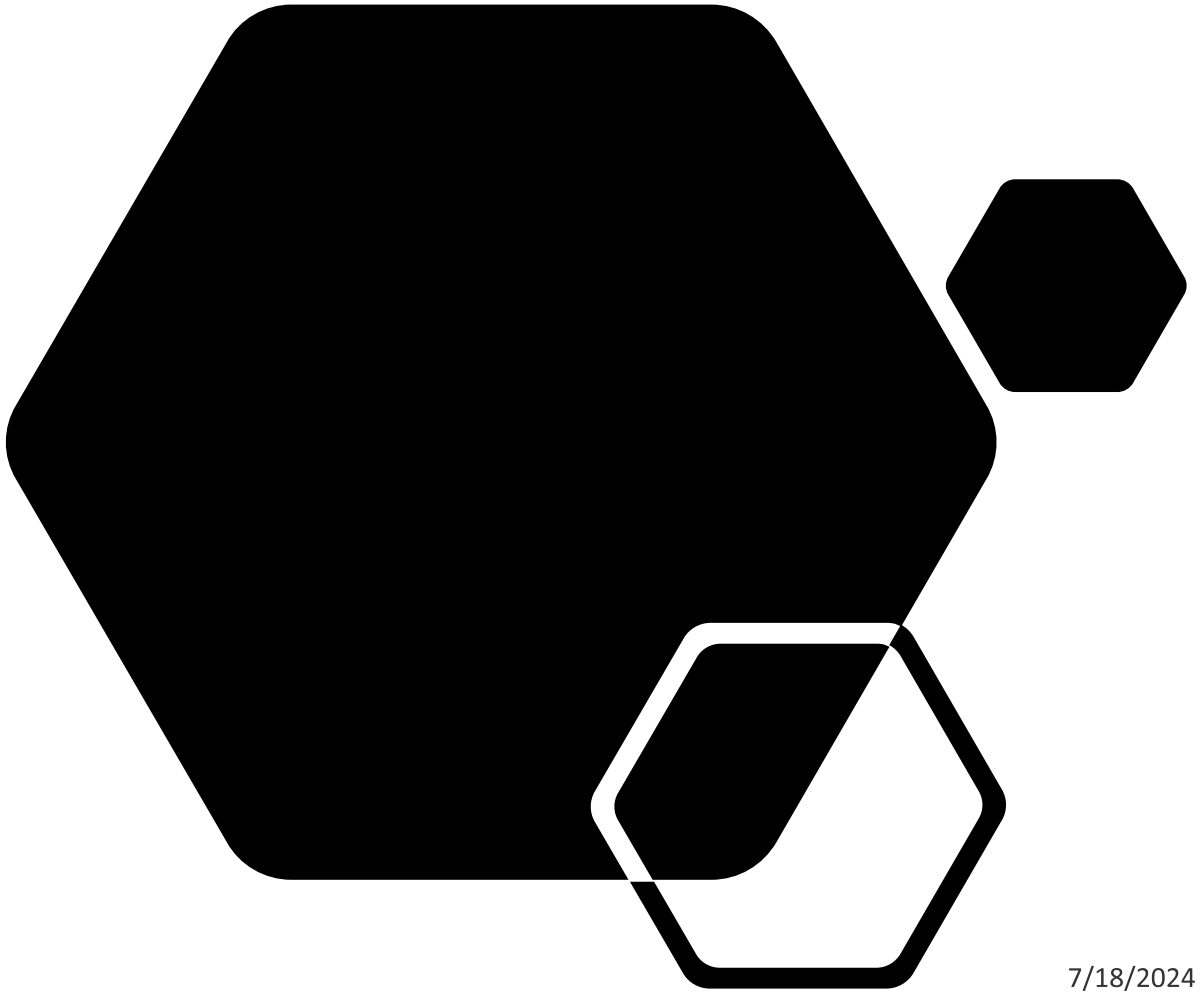
	Component		
	1	2	3
C1_1. cinematograf	**.697**	-.088	-.292
C1_3. filarmonica	**.518**	**.678**	-.054
C1_4. opera/ opereta	**.581**	**.642**	-.060
C1_5. spectacole de divertisment/ muzica	**.686**	-.017	-.112
C1_8. biserica	.045	.143	**.706**
C1_14. va plimbati prin parcuri, zone verzi	**.582**	-.128	**.368**
C1_2. cumparaturi in hipermarketuri	**.535**	-.255	**.394**
C1_10. cumparaturi in mall-uri	**.645**	-.236	.127
C1_6. discoteca/ club	**.505**	-.366	-.408
C1_7. teatru	**.667**	.281	-.102
C1_9. muzee/ expozitii	**.692**	.099	.089
C1_11. sa asistati la competitii sportive	**.486**	-.231	.040
C1_12. faceti fitness/ sport de intretinere	**.536**	-.196	-.207
C1_13. in excursii in afara localitatii	**.666**	-.223	.163

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Analiza cluster

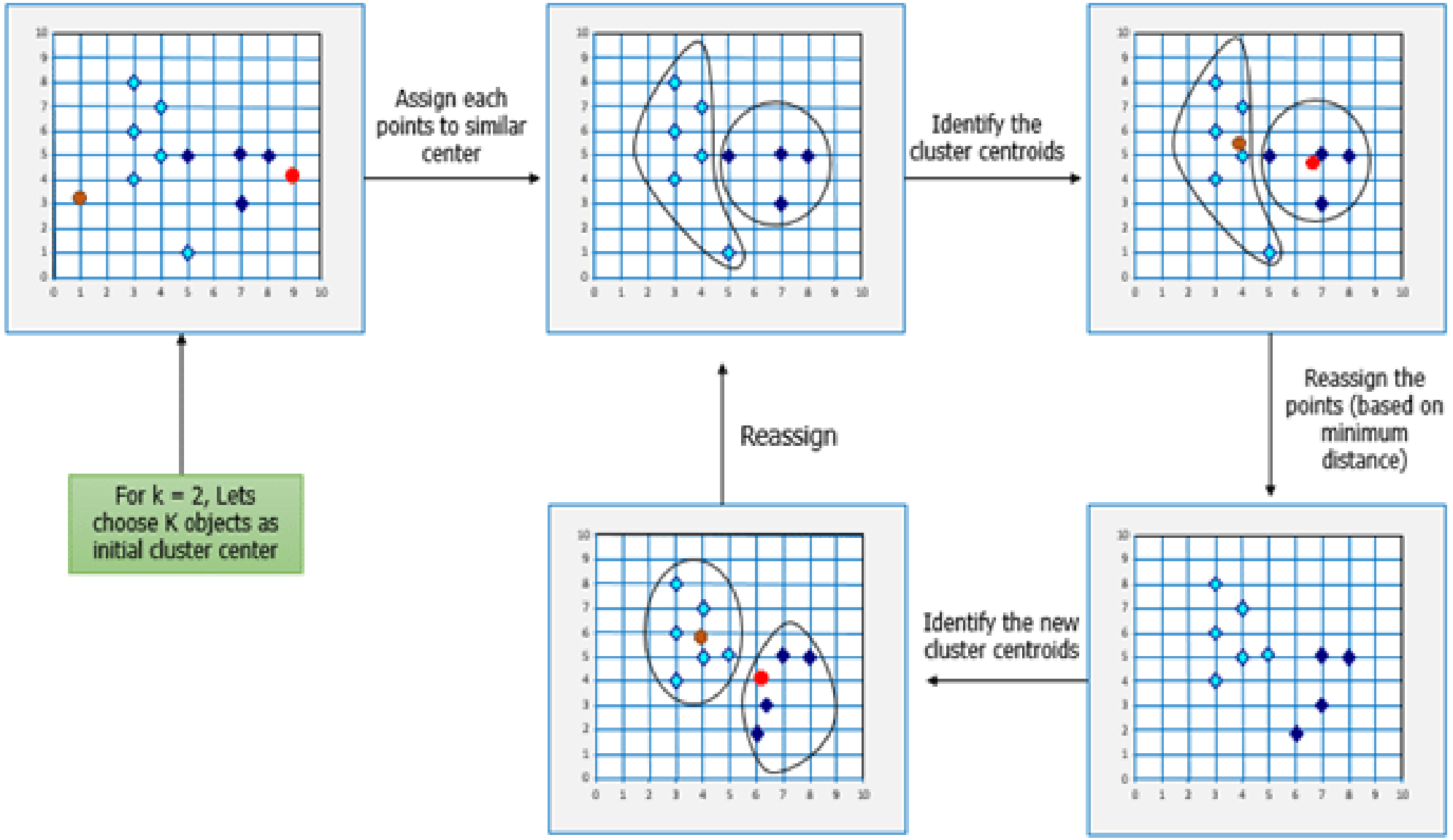
Identificarea unor tipologii de indivizi



Analiza cluster

- Identifică...
  - **Tipurile** de indivizi în funcție de o serie de criterii
  - Cluster = tip
  - Interpretăm semnificația tipurilor în funcție de profilul lor pe criteriile analizate
  - Putem explora clasificări cu mai multe sau mai puține tipuri
- O analiză cluster generează **o singură variabilă cu n valori**
  - Valori naturale, pozitive (de la 1 la n)
  - Fiecare valoare este un tip
- Fiecare caz este inclus într-un tip și numai unul

Analiza cluster K-means



Cuplu cu copii, familie extinsă

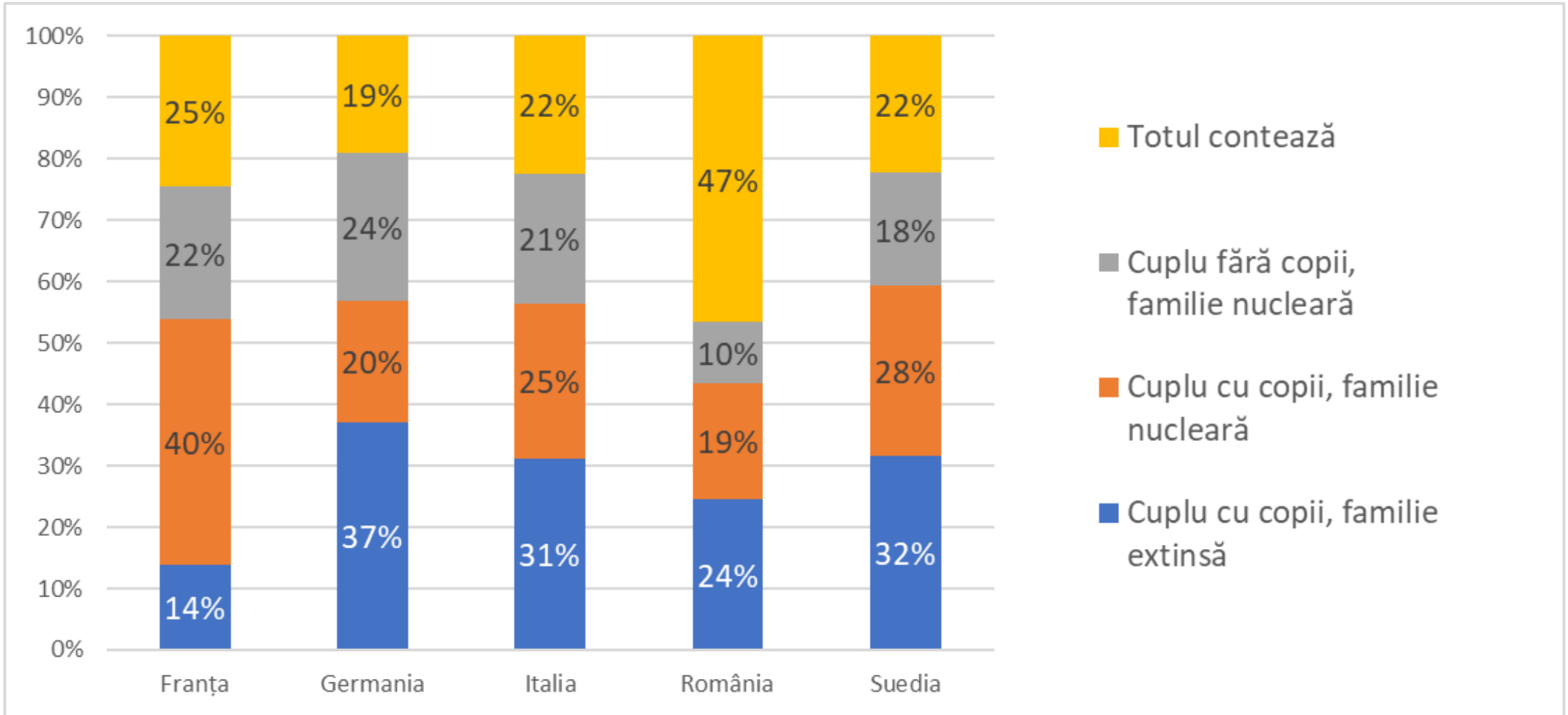
Cuplu cu copii, familie nucleară

Cuplu fără copii, familie nucleară

Totul contează

	Cluster			
	1	2	3	4
Fidelitate	1.81	1.86	1.58	1.91
Copii	1.42	1.87	**.64**	1.77
Relație sexuală	1.25	1.79	1.58	1.69
Locuință bună	.92	1.25	1.00	1.64
Venit	.90	1.15	1.05	1.60
Împart treburile	.85	1.46	.86	1.40
Aceeași religie	.63	**.31**	**.30**	1.49
Același mediu	**.41**	**.31**	**.55**	1.41
Locuiesc singuri	**.32**	1.71	1.65	1.41
Acord în politică	**.30**	**.25**	**.33**	.96

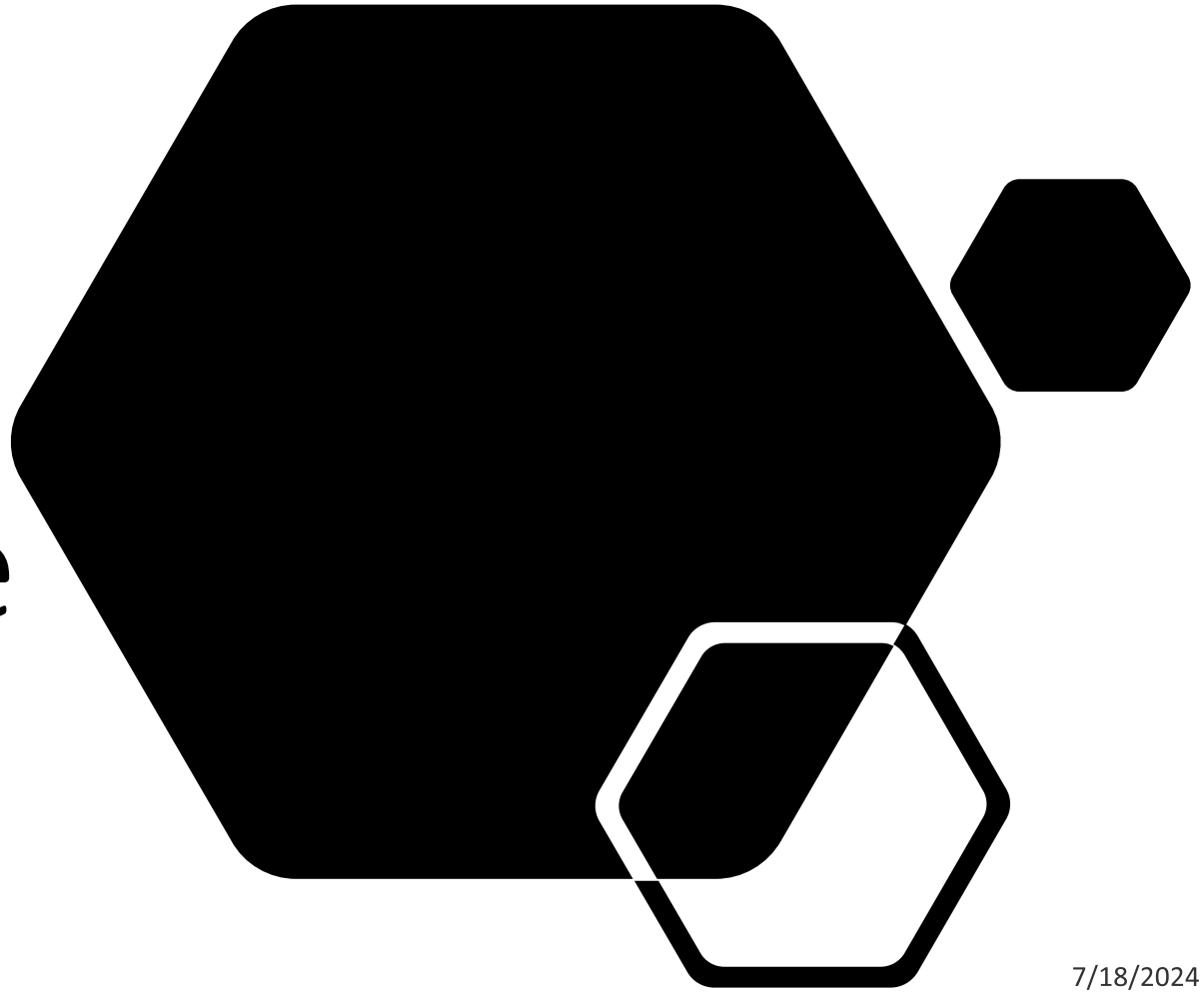
Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia



Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

Analiza de regresie

Explicații prin factori externi



Analiza de regresie

- Modelează variația unui efect ca funcție a mai multor predictori
  - Funcție liniară, polinomială, exponențială etc
  - Regresie liniară: $Y = a + b_1*x_1 + b_2*x_2 +...+b_n*x_n$
- Corelație: relații bivariate
- Regresie: modele multivariate
  - Relațiile dintre Y și cei n predictori sunt estimate simultan
  - Dacă se adaugă / scoate un X, se modifică toate estimările
  - Coeficientul b_i sau $beta_i$ arată relevanța predictivă a lui x_i când restul de x sunt „ținuți sub control”, adică sunt constanți
  - Coeficienții beta lucrează cu unități de măsură standardizate (abateri standard) și devin comparabili

Regresia ca bază a predicției

- O regresie estimată permite anticiparea unor valori Y necunoscute pentru care știm valorile X_i
  - Cu cât crește proporția de premii Nobel în funcție de consumul de ciocolată, vin și lapte

Analiza de regresie: teoriile căsătoriei de succes

	Similaritate R Square: 23%		**Intimitate R Square: 21%**		**Familia tradițională R Square: 17%**		**Banii nu contează R Square: 15%**	
	Beta	Sig.	**Beta**	Sig.	**Beta**	Sig.	**Beta**	Sig.
Sex - Feminin	**.065**	.000	**-.012**	.137	**.081**	.000	**.014**	.077
Vârstă	**.220**	.000	**-.195**	.000	**.105**	.000	**-.143**	.000
Mărimea localității	**-.003**	.677	**.060**	.000	**-.102**	.000	**.026**	.001

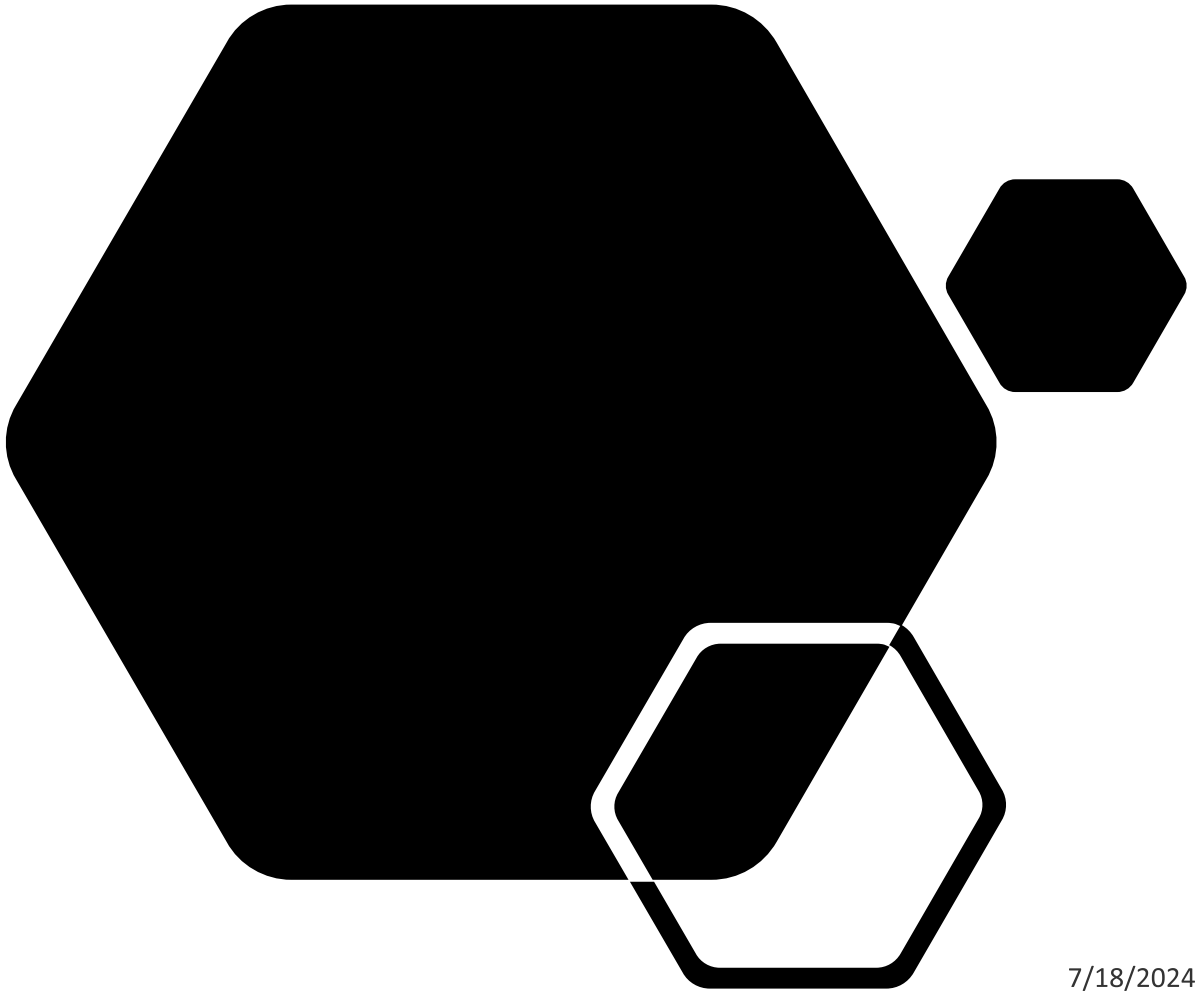
Când controlăm genul și mărimea localității, vârsta este cel mai puternic predictor pentru teoriile privind succesul în căsătorie

- Relații pozitive cu teoriile similarității, familiei tradiționale
- Relații negative cu teoriile intimității și „banii nu contează”

Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

Analiza de rețea

Explicații prin proximitate, contagiune



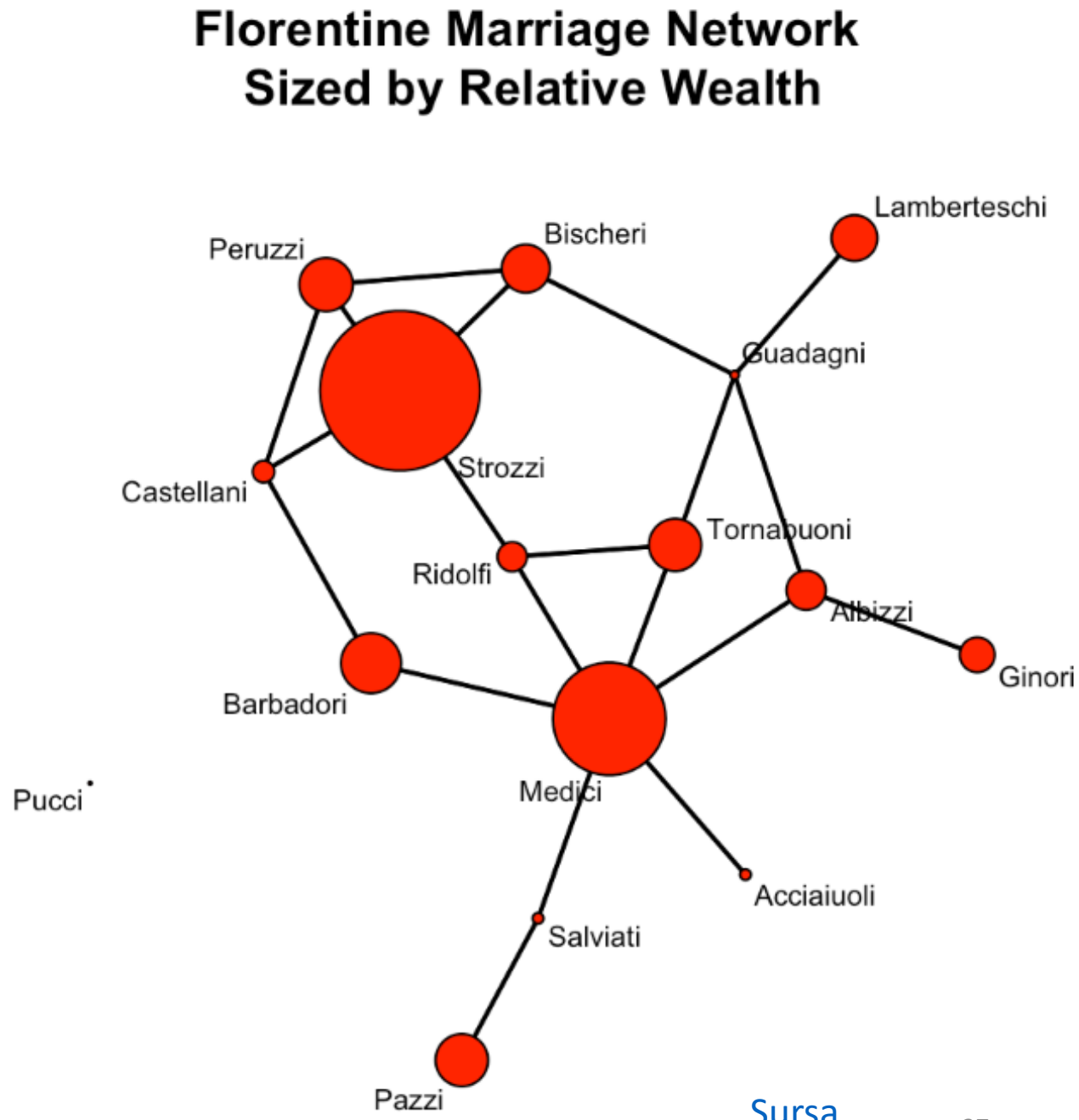
Analiza de rețea

- Cartografiază rețele pe baza **conexiunilor** dintre noduri
- Identifică proprietățile emergente ale rețelelor
  - De ex.: autocorelația sau homofilia de rețea; densitatea
    - „Cine se-aseamănă se-adună”/ „Birds of a feather fly together”
    - Filter bubble
- Identifică proprietățile nodurilor în rețea
  - De ex. centralitatea
- Identifică clustere de noduri bazate pe interconectivitate

Autocorelația

În ce măsură suntem influențați de cei similari cu noi – sau de noi înșine în trecut?

- Cine se-aseamănă se-adună
- Așchia nu sare departe de trunchi
- Orașele mai bogate se învecinează cu orașe mai bogate sau mai sărace?
- Familiile mai bogate se căsătoresc cu familii mai bogate sau mai sărace?
- Tendințe inerțiale – serii de timp

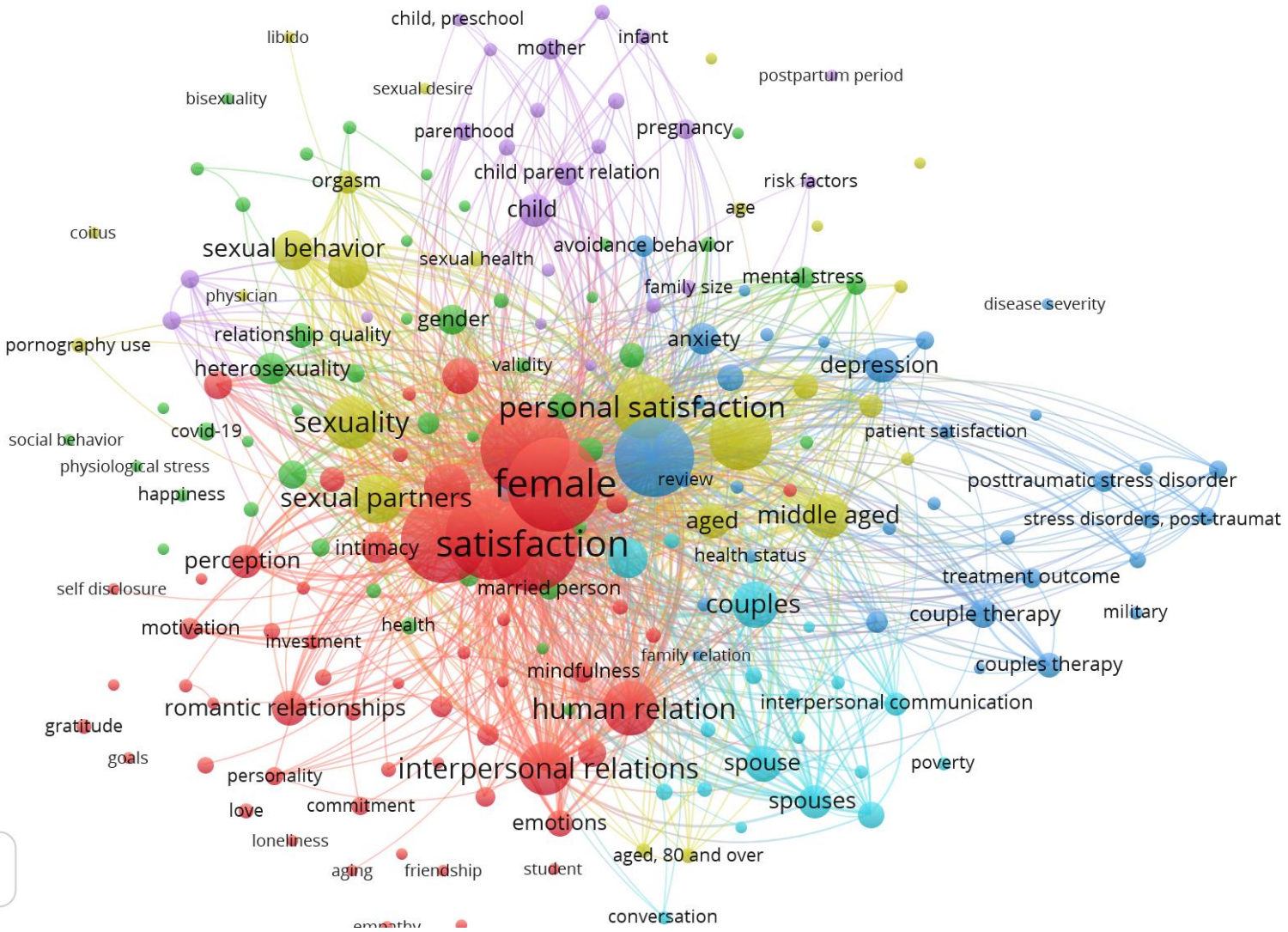


Sursa

Co-ocurența

VOSviewer: hărți bibliometrice de co-ocurență a cuvintelor cheie

- Relationship satisfaction
- Scopus
- Ultimii 5 ani
- Social sciences și Psychology



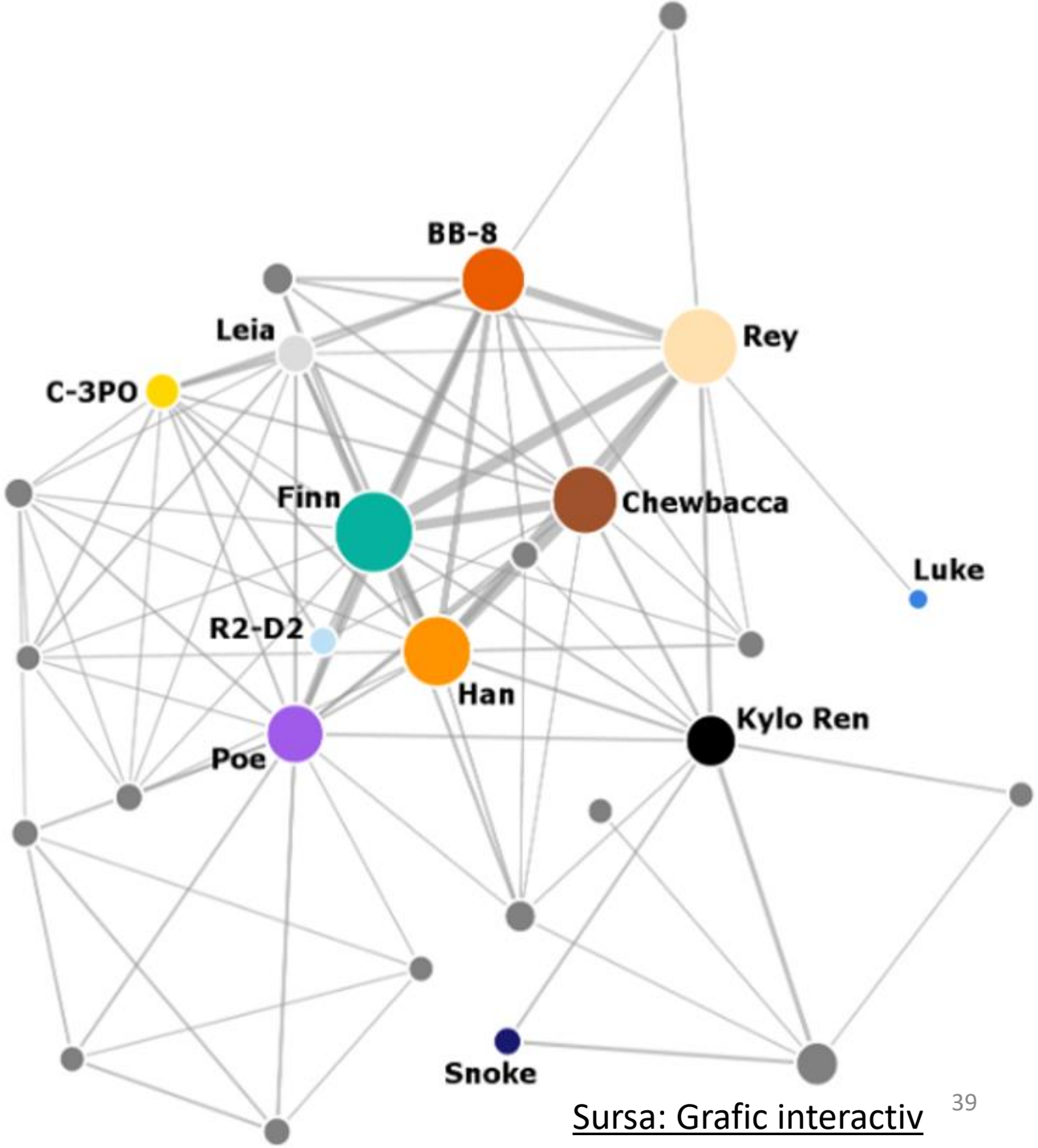
Vizualizare a rețelei

Centralitatea

Cât de central / periferic este un nod în rețea?

Măsuri ale influenței / relevanței

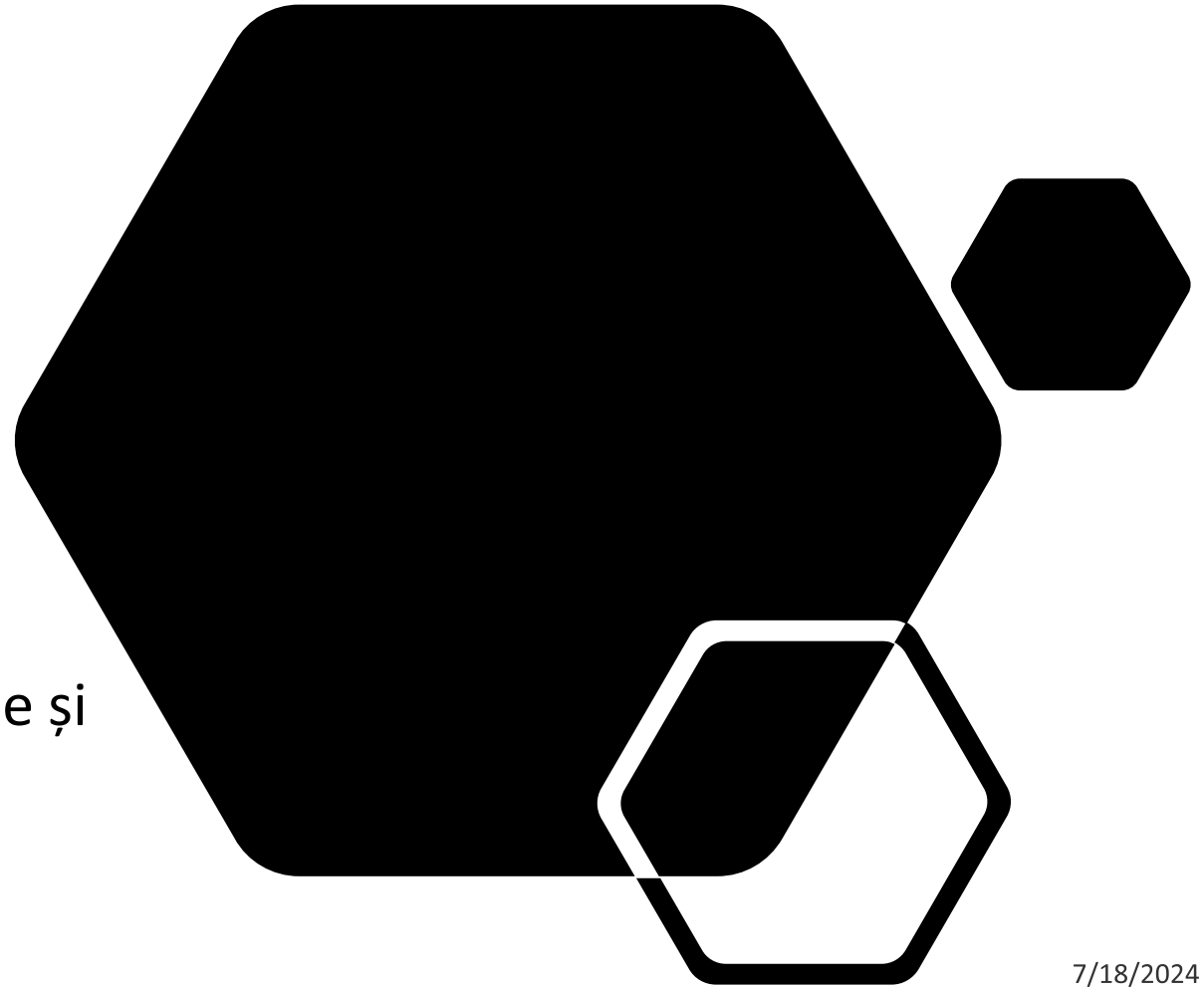
Ex: Analiza de rețea a personajelor din Star Wars ep VII



Sursa: Grafic interactiv

Serii de timp

Extrapolări și explicații prin tendințe temporale și factori externi



7/18/2024

Serii de timp

- Modelează variația unui fenomen ca funcție temporală
- Izolează și estimează tipuri de variații în timp
  - Tendințe, ciclicități, variații aleatorii
  - Alți factori externi influenți cu care fenomenul co-variază

Serii de timp

- Prezicem valori numerice continue (necunoscute) pe baza evoluției lor până acum
  - Câte grade vor fi mâine?
  - Care va fi prețul unui Bitcoin peste două zile?
- Combină
  - Proprietăți temporale intrinseci: cicluri, tendințe
  - Modele explicative extrinseci: cauze externe

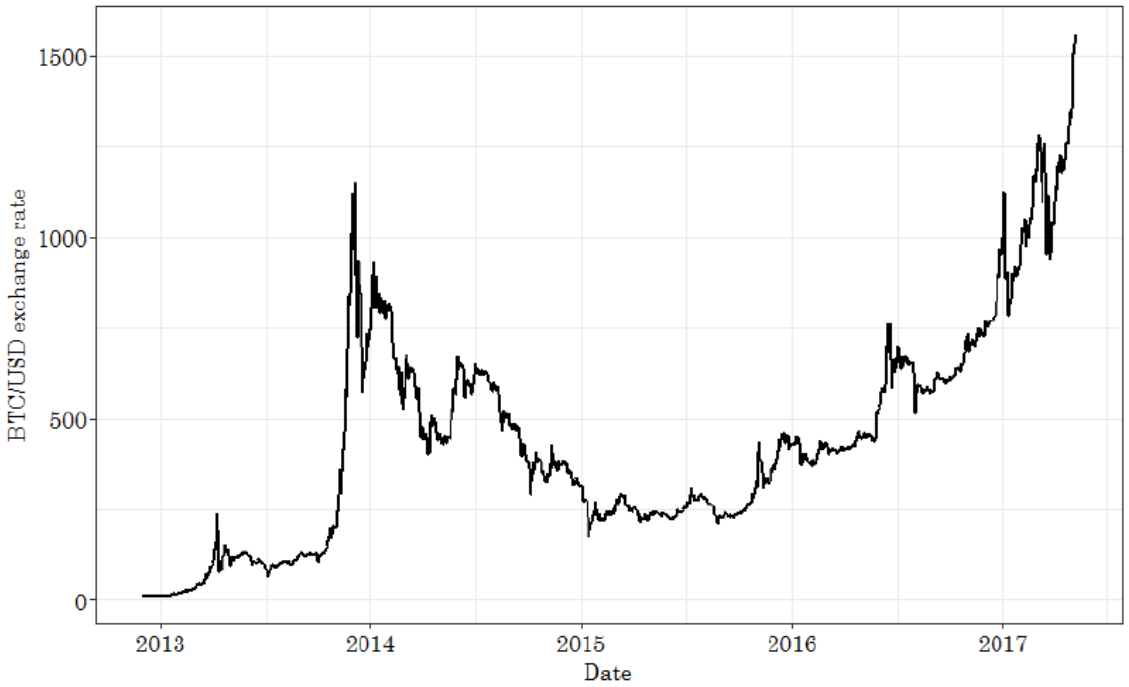


Figure 2: Bitcoin exchange rate with USD

Poyser 2018

Descompunerea în componente

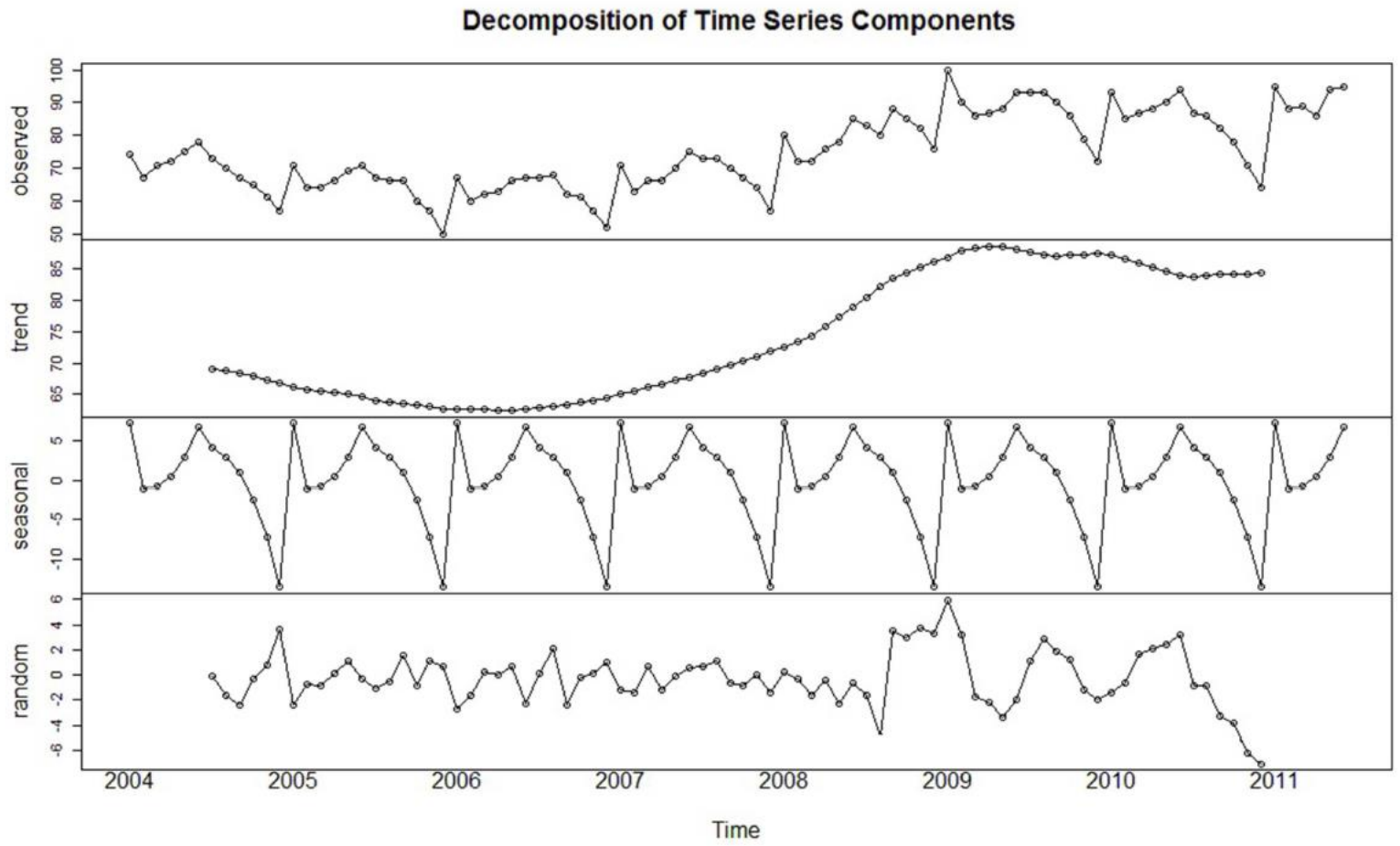
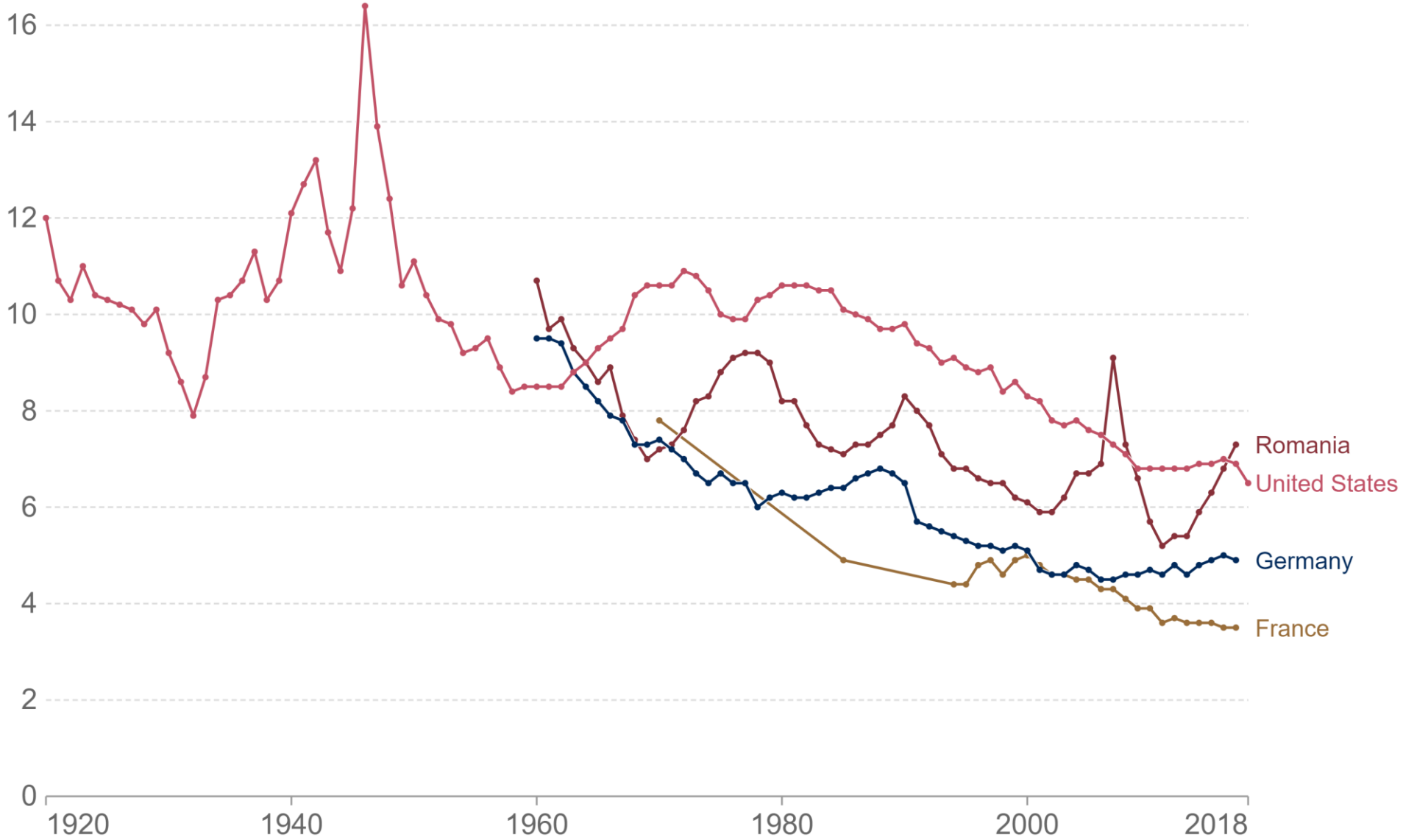


FIGURE 2 | The original time series decomposed into its trend, seasonal, and irregular (i.e., random) components. Cyclical effects are not present within this series.

Sursa

Marriages per 1,000 people

Number of marriages in each year per 1,000 people in the population

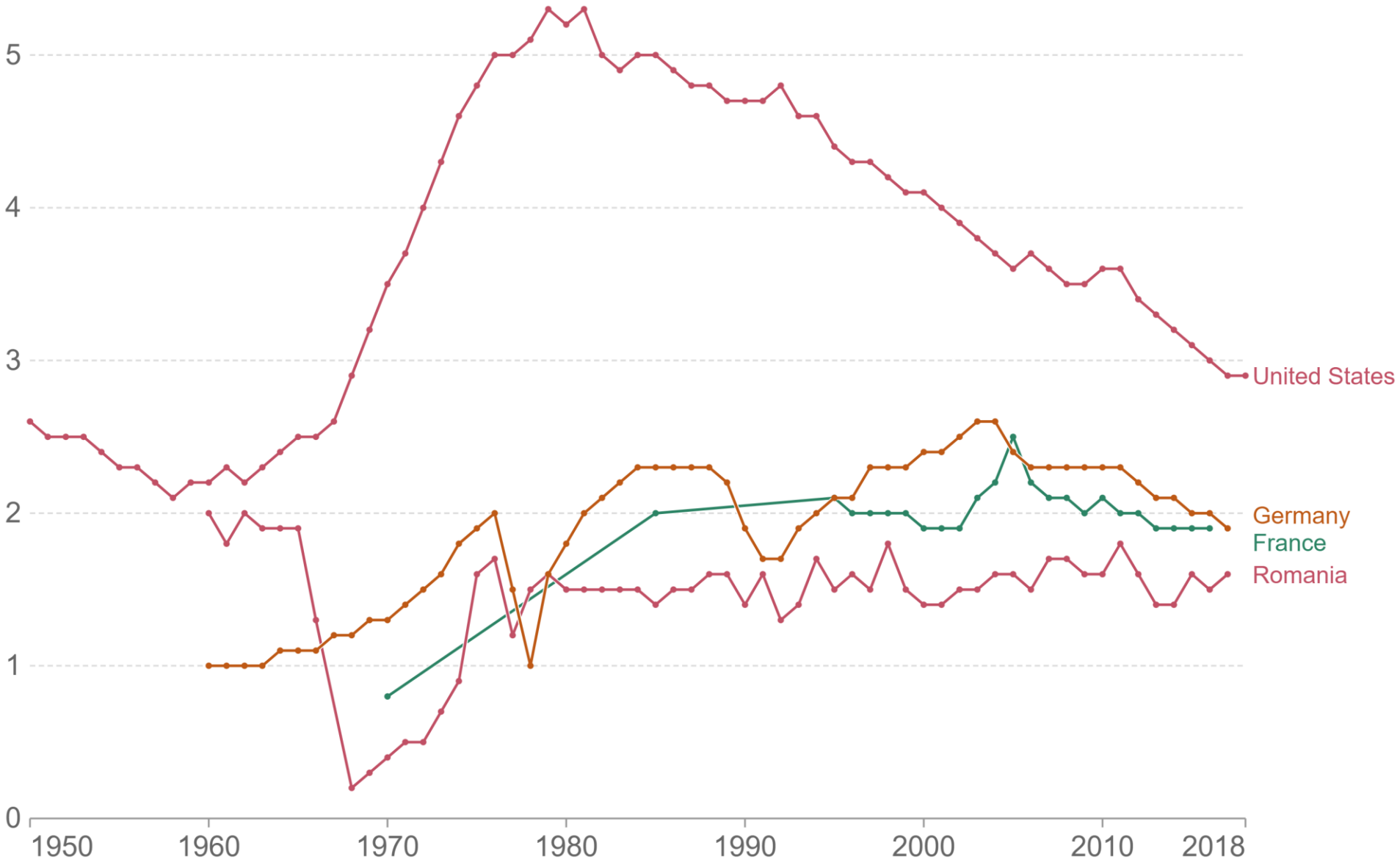


Source: OWID based on UN, OECD, Eurostat and others

OurWorldInData.org/marriages-and-divorces • CC BY

Sursa

Divorces per 1,000 people

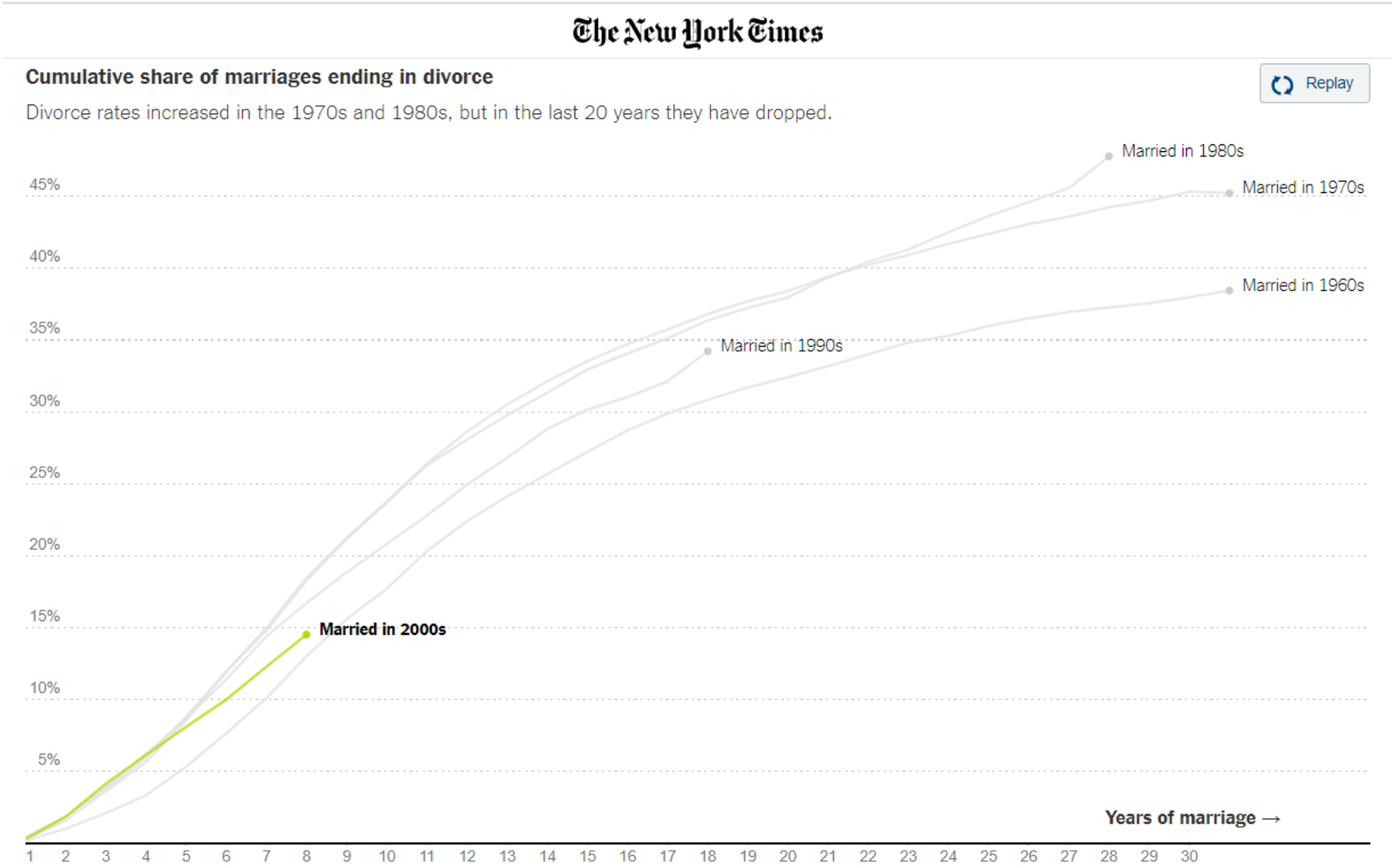


Source: OWID based on UN, OECD, Eurostat and other sources

OurWorldInData.org/marriages-and-divorces • CC BY

Sursa

Ratele cumulative sunt mai stabile



Sursa

Concluzii

- Tehnicile de analiză reduc complexitatea datelor empirice
- Identificarea patternurilor relevante
- Scopuri
  - Explorarea: frecvențe, medii, corelații
  - Măsurare: reducerea dimensionalității
  - Clasificarea: analiza cluster
  - Explicarea: analiza de regresie, analiza de rețea
  - Extrapolarea: serii de timp