Analiza Componentelor Principale pentru date Categoriale (CATPCA)

Cristian Opariuc-Dan[1]

Abstract

In many cases, the basic assumptions of parametric exploratory factor analysis are not met and yet even in these cases, the technique is used. There is the risk of inducing significant errors that could invalidate the factor analysis model. To avoid such situations, we can use another technique, available for ordinal or even nominal variables, less known and used, called "Principal Components Analysis for Categorical Data". This article aims at an introduction to these methods of data analysis. The article begins with a fictional example illustrating the configuration and analysis of results provided by SPSS for Windows for CATPCA (Categorical Principal Components Analysis).

Keywords: non-parametric analysis, nonparametric factor analysis, principal components

Résumé

Dans de nombreux cas, les hypothèses de base de l'analyse factorielle exploratoire paramétrique ne sont pas remplies, et pourtant, même dans ces cas, la technique est utilisée. Il y a le risque d'induire des erreurs importantes qui pourraient invalider le modèle d'analyse factorielle. Pour éviter de telles situations, nous pouvons utiliser une autre technique, disponibles pour les variables ordinales ou même nominale, moins connu et utilisé, appelé "Analyse en Composantes Principales pour les données catégorielles". Cet article vise à une introduction à ces méthodes d'analyse des données.L'article commence par un exemple fictif illustrant la configuration et l'analyse des résultats fournis par SPSS pour Windows pour CATPCA (Analyse en Composantes Principales pour les données catégorielles).

Mots-clés: analyse non-paramétrique, l'analyse factorielle non paramétrique, les composants principaux.

Rezumat

În foarte multe situații, asumpțiile de bază ale analizei factoriale exploratorii parametrice nu sunt îndeplinite și, totuși, chiar în aceste cazuri, tehnica se folosește. Există astfel riscul inducerii unor erori semnificative care pot invalida modelul de analiză factorială. Pentru a evita asemenea situații, vom putea utiliza o altă tehnică, disponibilă în cazul variabilelor situate la nivel ordinal sau nominal, mai puțin cunoscută și utilizată, numită „Analiza Componentelor Principale pentru date Categoriale”. Prezentul articol urmărește o introducere în aceste procedee de analiză a datelor. Articolul pornește de la un exemplu fictiv și ilustrează configurarea și analiza rezultatelor furnizate de SPSS for Windows în cazul CATPCA (Categorical Principal Components Analysis).

Cuvinte cheie: analize neparametrice, analiza factorială neparametrică, componente principale

[1] Universitatea Ovidius din Constanța
Adresa de corespondență: copariuc@gmail.com

Analiza Componentelor Principale pentru date Categoriale (CATPCA)

Analiza factorială exploratorie în varianta extragerii componentelor principale reprezintă o tehnică parametrică intens utilizată în domeniul psihologiei, mai ales în procesul de asigurare a validității constructului psihologic măsurat, alături de alte procedee specifice de analiză a datelor. Bazată pe corelații parametrice, analiza factorială de acest tip va avea o putere statistică foarte mare în condițiile îndeplinirii adecvate a asumpțiilor de bază ale acesteia:

Nivelul de măsură– în sensul strict al tehnicii, trebui să respecte *minimum scala de interval*. Deși se poate utiliza în cazul scalelor Likert, spre exemplu, este greu de presupus că acestea îndeplinesc strict criteriul intervalelor egale. Mai mult, numeroase instrumente psihologice folosesc itemi dihotomici, fără posibilitatea stabilirii unor relații de ordine între variantele de răspuns. În această situație, utilizarea analizei factoriale este discutabilă. Prin forțarea acestei asumpții este acceptată și scala ordinală în analiza factorială, însă, sub aspect pur statistic, utilizarea analizei factoriale parametrice pentru date ordinale nu este indicată.

Corelațiile liniare – reprezintă o a doua asumpție importantă în analiza factorială parametrică. Alături de faptul că toate variabilele supuse analizei factoriale parametrice trebuie să covarieze, postulatul indică și tipul de corelație – cea liniară. Se știe că pot exista corelații între variabile care nu au un caracter liniar (vezi exemplul corelației dintre motivație și performanță). În aceste situații, analiza factorială parametrică nu se poate folosi. De aceea, înaintea includerii variabilelor în analiza factorială va trebui identificată natura relațiilor dintre acestea.

Distribuții univariate și multivariate normale – probabil cea mai problematică asumpție. Analiza factorială parametrică presupune existența normalității distribuției pentru fiecare dintre variabilele supuse acestui procedeu. Este dificil, dacă nu imposibil, să asigurăm normalitatea distribuției fiecărui item, neluând în discuție faptul că unii itemi nici nu pot fi analizați sub acest aspect (de exemplu itemii dihotomici sau cei pur categoriali). Cei mai mulți analiști pur și simplu ignoră această asumpție, însă dacă ne gândim că întregul proces al analizei factoriale are la bază corelațiile, acestea fiind puternic afectate de lipsa de omogenitate a variațelor, atunci ne putem face o imagine legată de modul în care va fi afectat modelul general al tehnicii.

Mărimea lotului de cercetare – reprezintă un alt criteriu important. Raportul optim dintre numărul de variabile incluse în analiza factorială parametrică și numărul de subiecți necesari este de 1:20. Pentru fiecare variabilă inclusă în analiza factorială sunt necesari circa 20 de subiecți. Astfel, spre exemplu, pentru un chestionar cu 20 de itemi sunt necesari peste 400 de subiecți pentru ca tehnica să prezinte relevanță.

Aceste patru cerințe de bază vor trebui îndeplinite simultan pentru a putea aplica principiile analizei factoriale parametrice. În realitate, sunt destul de puține situațiile în care o asemenea analiză se realizează în concordanță cu aceste asumpții. Cel mai frecvent asistăm la încălcarea principiului normalității distribuției și a celui al nivelului de măsură. Cu toate că analiza factorială parametrică este destul de puternică pentru a compensa aceste scăpări procedurale, considerăm că, în anumite cazuri, utilizarea variantei alternative – analiza componentelor principale pentru date categoriale – este de preferat sau se poate folosi ca o confirmare sau infirmare a modelului de analiză factorială parametrică.

Analize categoriale

Analizele categoriale reprezintă o suită de tehnici statistice de procesare a datelor cu caracter neparametric, menite să compenseze situații în care, din motive de nerespectare a asumpțiilor, tehnicile parametrice nu se pot utiliza. Cu toate că au o putere statistică mai redusă în comparație cu procedeele neparametrice, sunt mult mai relevante în situațiile menționate mai sus. Aceste tehnici sunt numite și tehnici de scalare optimală și pot include regresiile categoriale (CATREG), corelațiile canonice neliniare (OVERALS), scalarea multidimensională (PROXSCAL) și, inversul acesteia, descompunerea multidimensională (PREFSCAL), analiza de corespondență și, desigur, analiza categorială a componentelor principale (CATPCA).

Când se utilizează tehnicile de scalare optimală?

Înainte de a putea răspunde la această întrebare va trebui să avem în vedere momentul

în care datele presupuse ca fiind parametrice nu au acest caracter. Sunt două mari categorii de situații la care putem face referire: datele sunt situate natural la un nivel de măsură neparametric și datele devin neparametrice din cauza nerespectării asumpțiilor de bază.

Prima situație se referă la variabilele de tip nominal sau ordinal. În acest caz, variabile precum „genul biologic", „culoarea părului", „culoarea ochilor" sau „gradul didactic", „gradul militar", „grupa de vârstă" se situează la un nivel de măsură nominal, respectiv ordinal și sunt natural la nivel neparametric. Ultima variabilă dată ca exemplu, „grupa de vârstă", poate induce confuzii. Avem în vedere cazul în care grupa de vârstă se referă la „sub 20 de ani", „între 20 și 40 de ani" etc. De aceea, recomandăm ca întotdeauna să se folosească variabile la cel mai înalt nivel de măsură posibil (de exemplu să se înregistreze vârsta efectivă – variabilă scalară – și nu grupa de vârstă – variabilă ordinală). O variabilă scalară poate fi transformată foarte ușor într-una ordinală, invers este imposibil (Opariuc-Dan, 2009).

A doua situație se referă la variabile scalare care nu îndeplinesc condițiile unor analize parametrice din următoarele motive: distribuții care se abat de la distribuția normală, număr insuficient de cazuri, număr suficient de cazuri, însă insuficient sub raportul număr de variabile – număr de subiecți, neîndeplinirea cerințelor proprii ale tehnicii de analiză (sub aspectul variațelor, a coliniarității, corelației reziduurilor etc.).

Analizele parametrice sunt pretențioase. Atunci când datele nu respectă criteriile, consecințele pot fi destul de neplăcute sub aspect științific. Din fericire, există și tehnici neparametrice care ne pot oferi analize adecvate sau care pot, în ultimă instanță, verifica metodele parametrice.

Principiile analizei componentelor principale pentru date categoriale

Principiile nu diferă semnificativ în comparație cu cele utilizate la analiza factorială clasică. Urmărim extragerea unui factori latenți, comuni unui set de variabile, și identificarea modului în care variabilele pot explica factorul latent. Fiecare variabilă prezintă o variație proprie dar și o varianță comună. Inițial ele se prezintă ca un număr de factori independenți, tot atâția câte variabile sunt incluse în analiză. Ideea analizei factoriale este aceea de a reduce acești factori, pe baza variațelor comune, până la identificarea numărului minim de factori care pot explica variațele variabilelor inițiale.

Nu vom insista asupra acestor principii, presupunând că sunt bine cunoscute. Vom încerca, practic, să furnizăm o serie de repere în vederea realizării efective a CATPCA folosind aplicația SPSS for Windows.

Configurarea analizei categoriale a componentelor principale

Această tehnică se poate folosi pentru orice fel de date: nominale, ordinale, de interval sau de raport, ajustându-se în funcție de tipul acestora. Pentru a se putea efectua aceste analize, avem nevoie de o licență separată pentru modulul „***Categories***" al SPSS pentru Windows. În acest moment, va deveni disponibilă opțiunea „***Optimal Scaling***" din cadrul sub-meniului „***Dimension Reduction***" al meniului „***Analyze***" din SPSS for Windows. Acționând această opțiune, vom avea posibilitatea de a defini modelul de analiză categorială dorit. Există două secțiuni în figura 1. Prima secțiune (***Optimal scaling level***) se referă la tipul de variabile de care dispunem. Alegând „***All variables are multiple nominal***" confirmăm un tip de analiză pur categorial și, prin urmare, vom avea la dispoziție doar analiza de corespondență (***Multiple Correspondence Analysis***) sau corelațiile canonice neliniare (***Nonlinear Canonical Correlations***). Dacă alegem „***Some variable(s) are not multiple nominal***" vom comunica faptul că am inclus în analiză și variabile situate la nivel ordinal sau scalar, prin urmare, pe lângă corelațiile canonice, devine disponibilă și analiza categorială a componentelor principale (***Categorical Principal Components***). Cea de-a doua secțiune (***Number of Sets of Variables***) se referă la existența sau inexistența seturilor de variabile, adică a unei variabile independente care să împartă baza de date (de exemplu genul biologic). Putem alege între „***One set***", un singur set de variabile, fără influența unei variabile independente, caz în care se poate efectua analiza componentelor principale sau analiza corespondenței multiple sau „***Multiple sets***", situație în care singura analiza validă poate fi reprezentată de corelațiile canonice neliniare. Toate modelele de analiză, în

funcție de selecție, sunt marcate în secțiunea „***Selected Analysis***”. Cu ajutorul butonului „***Define***” putem trece la definirea propriu-zisă a analizei componentelor principale.

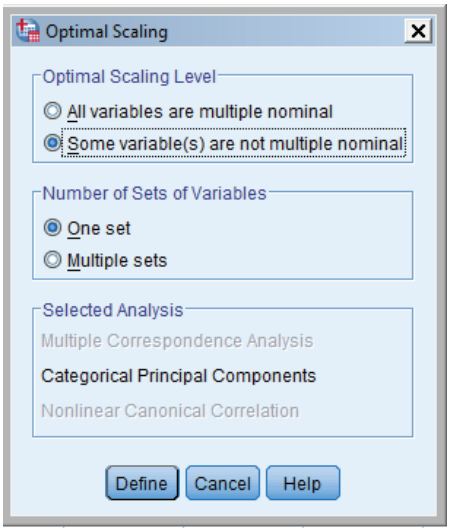

Figura 1 Definirea modelului de analiză

Fereastra de configurare a analizei categoriale a componentelor principale are o multitudine de opțiuni care presupun explicații suplimentare pentru a putea înțelege conceptul.

Să presupunem că lucrăm la validarea constructului pentru inventarul de personalitate BigFive Plus, varianta Iași, Ticu Constantin. Știm că dimensiunea „Extraversiune” este compusă dintr-un număr de 6 factori (afectivitate, sociabilitate, asertivitate, activitate, excitabilitate și veselie), fiecare factor fiind măsurat printr-un număr de 8 itemi dihotomici, categoriali. Amplitudinea teoretică a fiecărui factor este cuprinsă între valoarea minimă 0 puncte și valoarea maximă 8 puncte. Am putea folosi analiza factorială clasică, însă aceste variabile nu se distribuie normal, existând riscul ca modelul de analiză factorială să nu fie unul valid.

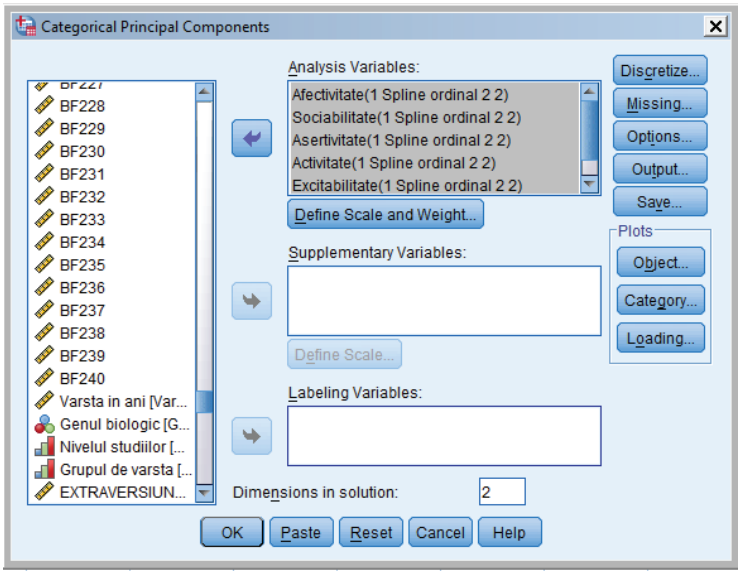

Figura 2 Configurarea analizei categoriale pe componente principale

Vom introduce toate cele șase variabile în lista „***Analysis Variables***” în vederea realizării analizei componentelor principale. În cazul în care suspectăm existența unor variabile care pot covaria cu cele șase, le putem introduce și defini în lista „***Supplementary Variables***”. SPSS nu le va lua în considerare la construcția modelului principal, însă va identifica efectul lor asupra modelului. Lista „***Labeling Variables***” permite includerea unei variabile independente care va marca pe grafice situația scorurilor (spre exemplu variabila gen biologic va marca pe grafice bărbații și femeile). Analiza categorială a componentelor principale nu identifică automat numărul dimensiunilor extrase. Prin urmare, trebuie să pornim de la un model teoretic. Este normal să presupunem că cei șase factori vizează o singură dimensiune, extraversiunea, și nu altele. Vom alege în caseta „***Dimensions in solution***” să identifice 2 dimensiuni și nu una

singură. Noi presupunem că există o singură dimensiune comună a celor șase variabile, dar dacă nu este așa? Din acest motiv, vom construi un model care va avea cu cel puțin o dimensiune mai mult decât numărul de dimensiuni stipulate teoretic.

Acest tip de analize se bazează pe date întregi și pozitive. Valorile nule, negative sau fracționare vor fi convertite în mod specific folosind opțiunile oferite de butonul „***Discretize***".

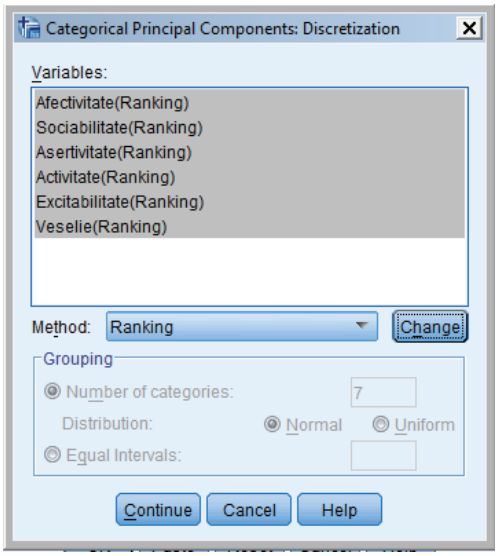

Figura 3 Discretizarea variabilelor nevalide

Există mai multe posibilități de transformare a valorilor nevalide în valori întregi și pozitive (Meulman, Heiser, & SPSS, Inc., 2007). Implicit, variabilele nu sunt supuse procesului de transformare în variabile discrete (***Unspecified***), fiind tratate ca atare. Opțiunea „***Grouping***" implică gruparea scorurilor într-un număr specificat de categorii și recodificarea acestora pe baza acestor grupări în intervale. Gruparea în intervale este similară celei utilizate în statisticile univariate la crearea etaloanelor (opțiunea „***Number of categories***") în care se specifică numărul de clase și tipul de standardizare (normală sau uniformă) sau se grupează în intervale egale, nestandardizate, în genul cuantilării (Opariuc-Dan, 2009). SPSS nu va mai lua în considerare datele efective ci aceste intervale nou create, asimilate etaloanelor. Opțiunea „***Ranking***" transformă toate variabilele în ranguri și folosește rangurile în locul valorilor efective. Este utilă, în special în cazul itemilor construiți pe scale Likert. Opțiunea „***Multiplying***" efectuează operațiuni de normalizare a distribuției (de standardizare), apoi valorile sunt multiplicate cu 10, rotunjite la valori întregi și se adaugă o constantă pentru toate scorurile astfel încât cea mai mică valoare să fie 1 (Meulman, Heiser, & SPSS, Inc., 2007). Acest tip de discretizare se folosește numai în cazul variabilelor scalare care nu prezintă o distribuție normală. Ca repere de lucru, dacă variabila are puține categorii, categoria minimă fiind 1 (cum ar fi cazul scalelor Likert), putem lăsa aceste variabile nemodificate. Dacă variabilele sunt scalare (cum ar fi, de exemplu coeficientul de inteligență) nedistribuite normal, putem folosi multiplicarea în vederea normalizării. Variabilele nominale se pot grupa iar cele ordinale pot fi transformate în ranguri.

În situația noastră, amplitudinea distribuției este situată între 0 și 8 puncte. Valoarea nulă va crea probleme, iar scalele provin din itemi dihotomici categoriali. Prin urmare, este greu de asimilat cele șase variabile ca fiind variabile situate la un nivel de interval. Le vom trata ca variabile ordinale și le vom transforma în ranguri. SPSS permite, pentru fiecare tip de variabilă, alegerea unei metode de discretizare. Noi vom selecta toate variabilele, vom alege metoda de transformare în ranguri și vom apăsa butonul „***Change***" pentru a aplica această opțiune. Vom reveni apoi la fereastra inițială folosind butonul „***Continue***".

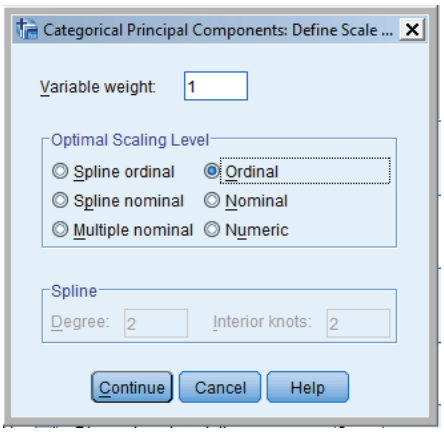

Figura 4 Definirea scalelor

Următoarea etapă implică definirea tipului de scală prin accesarea butonului „***Define Scale andWeight***". Și în acest caz, putem defini tipul scalei pentru fiecare variabilă. În situația noastră vom selecta toate variabilele înainte de a accesa butonul. În figura 4 sunt prezentate nivelurile de scalare optimală. Implicit, SPSS presupune nivelul „***Spline ordinal***" – în care ordinea

scorurilor se menține în cadrul variabilei analizate, toate fiind tratate la nivelul scalei de măsură inițiale. Se poate folosi în cazul scalelor Likert, scorurile fiind ajustate în jurul categoriilor scalei respective. Mai mult, se poate defini scala precizându-se numărul nodurilor interioare (categoriile interioare de răspuns) și numărul efectiv al categoriilor. Astfel, pentru o scală Likert cu valori între 1 și 5, vom putea defini în secțiunea „***Spline***" un număr de 5 niveluri (***Degree***) și 3 noduri interioare (***Interior knots***). SPSS va ajusta datele observate acestei scale definite, având ca rezultat o estimare apropiată de modelul teoretic. Opțiunea „***Spline nominal***" se referă la date categoriale. În acest caz, SPSS grupează scorurile în categorii și construiește o scală nominală, neordonată, pe baza opțiunilor furnizate în secțiunea „Spline". Comportamentul este similar opțiunii anterioare, singura diferență fiind cea de nivel de măsură. Opțiunea „***Multiple nominal***" se referă tot la date nominale. De data aceasta categoriile de grupare sunt derivate din date și nu specificate explicit. Este opțiunea care poate furniza seturi diferite de indicatori pentru fiecare dimensiune deoarece gruparea nu este una fixă, unică, ci variază. Opțiunea „***Ordinal***" se referă la date ordinale natural care nu presupun ajustarea pe o scală implicit declarată, similară opțiunii „***Nominal***" care se referă la un alt nivel de măsură. Aceste două opțiuni grupează categoriile folosind datele reale și nu efectuează o ajustare la nivelul scalei teoretice ca în cazul alternativelor de tip spline. În sfârșit, opțiunea „***Numeric***" stabilește nivelul de măsură scalar pentru variabile. Categoriile sunt tratate ordonat și având intervale egale. Se menține atât ordinea categoriilor cât și egalitatea intervalelor. Dacă variabilele sunt continui, prelucrarea este similară analizei factoriale clasice.

Există o legătură între modul de discretizare al variabilelor și alegerea scalei. În situația noastră am optat pentru ***ranguri***. Vom putea folosi opțiunile de tip ordinal, cu sau fără specificarea scalei teoretice. Noi am ales opțiunea „Ordinal" deoarece poate produce estimări mai bune, chiar dacă nu atât de fin ajustate ca în cazul în care am fi folosit opțiunea „Spline ordinal".

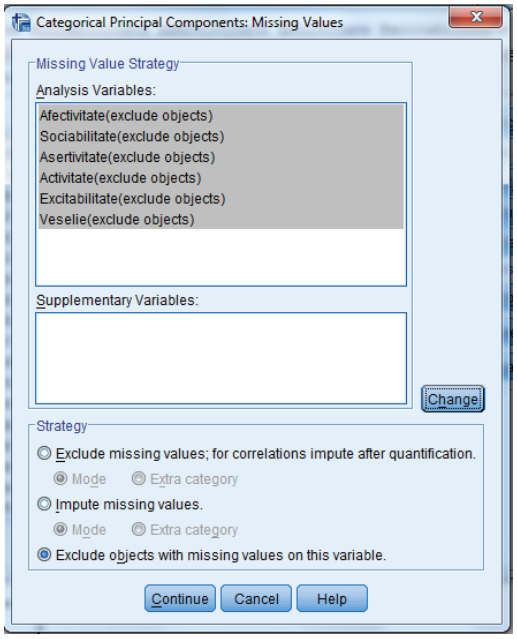

Figura 5 Tratamentul cazurilor cu date lipsă

Analiza categorială este foarte sensibilă la cazurile lipsă și impune modul de tratare a acestora. Vom stabili aceste elemente folosind butonul „***Missing***". Tratamentul poate fi aplica atât variabilelor inițiale cât și variabilelor suplimentare, care presupunem că influențează modelul. Ca și în celelalte situații, tratamentul poate fi aplicat fiecărei variabile în parte. Vom selecta toate variabilele și vom alege opțiunea „***Exclude objects with missing values on this variable***." Apoi vom apăsa butonul „***Change***". În condițiile în care există un număr suficient de date, se recomandă utilizarea acestei opțiuni. Dacă eliminând cazurile cu date lipsă observăm că numărul subiecților valizi este foarte mic, este bine să utilizăm una dintre celelalte două opțiuni disponibile. Opțiunea „***Exclude missingvalues for correlations impute afterquantification***" vizează un tratament pasiv al cazurilor lipsă. Acestea nu vor fi selectate la analiza variabilei. Dacă toate variabilele au date lipsă pentru subiectul respectiv, acestea vor fi considerate variabile suplimentare. După analiza inițială, dacă se dorește calculul corelațiilor, cazurile lipsă vor fi înlocuite cu valoarea modală a variabile scalate (***Mode***) sau cu valoarea cuantificată a acesteia (***Extra category***). Opțiunea „***Impute missingvariable***" presupune un tratament activ al cazurilor lipsă. Similar opțiunii anterioare, cazurile lipsă se înlocuiesc în funcție de modalitatea precizată, apoi sunt incluse și în analiza inițială.

Butonul „***Options***” specifică modalitatea de realizare a analizei componentelor principale pentru date categoriale. Secțiunea „***Supplementary Objects***” permite introducerea cazurilor care vor fi ignorate în timpul analizei. Putem introduce un interval (***Range of cases***), spre exemplu de la subiectul 80 la subiectul 96 sau un caz individual (***Single case***), de exemplu doar subiectul 56. Această opțiune se dovedește utilă atunci când avem scoruri extreme care ar putea influența analiza datelor. În secțiunea „***Normalization method***” vom putea alege metoda de extragere a componentelor principale. Cea mai folosită metodă este „***Variable Principal***” prin care se optimizează asocierea dintre variabile. Coordonatele variabilelor în spațiul determinat de scoruri sunt saturațiile în factor latent, accentul cade pe corelațiile dintre variabile. Metoda „***Object Principal***” optimizează distanța dintre scoruri. Accentul cade pe diferențele dintre variabile și modul în care acestea se grupează în spațiul determinat de scoruri. Metoda „***Symmetrical***” se axează pe relațiile existente între scoruri și variabile, pe măsura în care scorurile saturează fiecare dintre variabile. Opțiunea „***Independent***” are în vedere analiza distanțelor dintre scoruri pe de o parte și a corelațiilor între variabile pe de altă parte. Este un combinație a primelor două metode. În fine, opțiunea „***Custom***” permite specificarea unei valori cuprinsă între 1 (corespunzătoare metodei „Object Principal”) și -1 (corespunzătoare

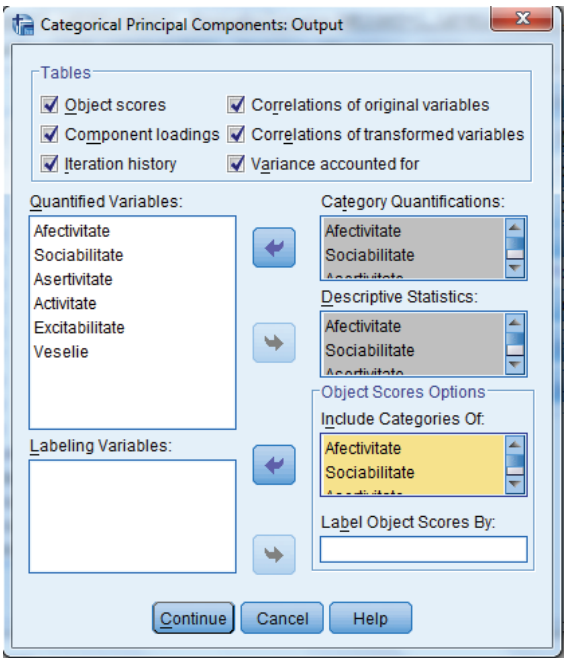

Figura6 Alegerea informațiilor care vor fi afișate în urma analizei

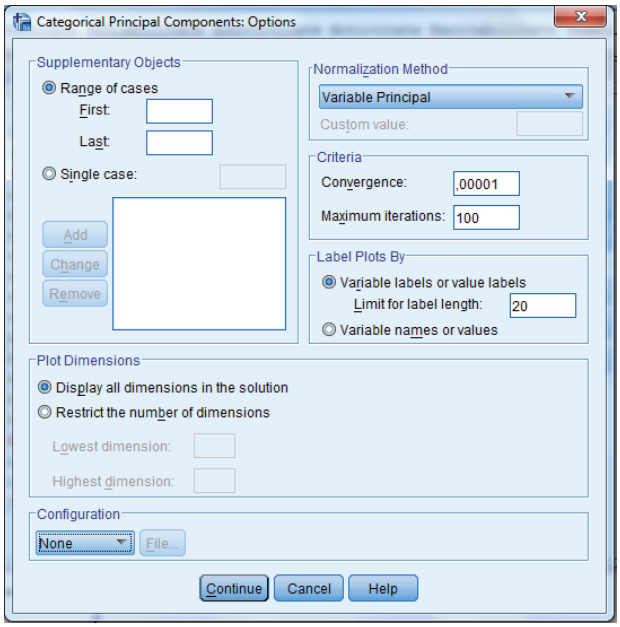

Figura 7 Opțiuni privind analiza categorială pe componente principale

metodei „Variable Principal”) trecând prin 0 (corespunzătoare metodei „Symmetrical”) la care să se raporteze metoda de normalizare. Astfel poate fi modificată rădăcina matricei de corelații (eigenvalue) atât la nivelul scorurilor, cât și la nivelul variabilelor.

Secțiunea „***Criteria***” permite specificarea numărului maxim de iterații în vederea identificării unui model (***Maximum Iterations***), rareori fiind nevoie de modificarea acestei valori, și a pragului de convergență a matricei de corelații în vederea identificării unui model complet (***Convergence***). Concret, analiza se va opri în cazul în care, pentru ultimele iterații, diferența dintre ele se situează sub pragul de convergență.

Secțiunea „***Labelplotsby***” vizează modul de marcare a graficelor. Se pot afișa etichetele variabilelor sau valorile acestora sau numele variabilelor sau valorile acestora. Opțiunile au relevanță doar la nivelul graficelor generate de CATPCA. În fine, secțiunea „***Plot dimensions***” controlează numărul de dimensiuni care vor fi afișate grafic (factori). Se poate alege reprezentarea tuturor factorilor în cazul în care numărul de dimensiuni este relativ mic (3 sau 4 dimensiuni) sau se pot specifica dimensiunile care vor fi reprezentate grafic (în general cele mai importante).

Cu ajutorul butonului „***Output***” putem controla ce date se vor afișa în foaia de rezultate (Output). În secțiunea „***Tables***” putem preciza tabelele care vor fi construite. Alegând „***Object***

scores" vom putea afișa scorurilor subiecților din baza de date pentru variabilele selectate în lista „***Include Categories Of***". De asemenea, aceste scoruri vor putea fi etichetate ca aparținând variabilei introduse în caseta „***LabelObjectScoresBy***" (spre exemplu genul biologic). Alegerea acestei opțiuni poate genera tabele foarte mari deoarece acestea conțin toți subiecții din baza de date. În mod normal, SPSS are o limită de afișare a datelor în Output (de 100 de rânduri), dar care poate fi modificată. Opțiunea „***Component loadings***" va afișa modul în care fiecare dintre variabile saturează factorii latenți identificați iar alegerea opțiunii „***Iterationhistory***" va afișa întregul ciclu de iterații. În cazul în care pentru identificarea modelului sunt necesare numeroase iterații, bifarea acestei opțiuni poate genera, de asemenea, tabele mari. În mod normal, SPSS afișează doar prima și ultima iterație. Alegerea opțiunii „***Correlation of original variables***" permite afișarea matricei de corelații dintre variabile, dar și a rădăcinilor acesteia (eigenvalue) pentru fiecare variabilă în parte. „***Correlations of transformedvariables***" va prezenta o altă matrice de corelații a variabilelor, similară celei anterioare, însă după ce variabilele au fost normalizate prin metoda de normalizare precizată anterior.

Bifarea casetei „***Varianceaccounted for***" afișează cantitatea de varianță explicată per variabile și per dimensiuni, sub aspectul variației totale, al coordonatelor vectoriale și centroide.

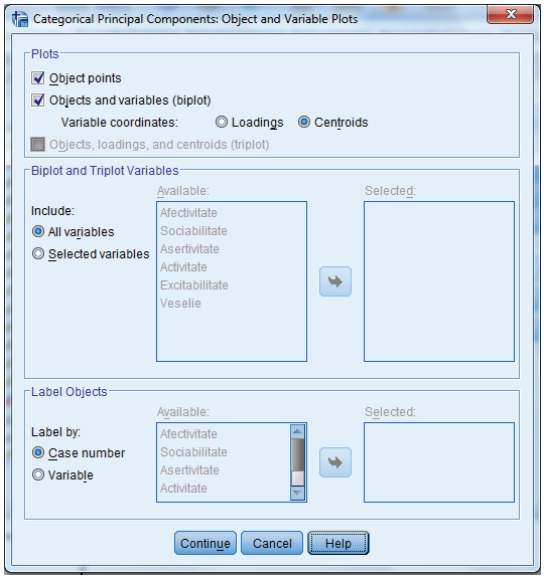

Figura 8 Grafice referitoare la scoruri

Coordonatele vectoriale sunt bazate pe proiecția modelului de analiză și sunt determinate de cele două spații ale analizei: pe de o parte spațiul scorurilor, pe de cealaltă parte spațiul variabilelor. Categoriile de răspuns sunt reprezentate printr-o linie dreaptă între două dimensiuni (factori latenți) iar o coordonată vectorială se referă la coordonatele fiecărei categorii de răspuns de pe această axă.

Coordonatele centroide nu mai implică spațiul variabilelor ci doar pe cel al scorurilor și indică poziția pe care o obține fiecare categorie de răspuns (determinată de scorurile acesteia) în spațiul celor două dimensiuni.

Lista „***CategoryQuantification***" furnizează informații legate de modul de cuantificare al fiecărei categorii, precum și coordonatele acesteia, pentru fiecare dintre variabilele analizate iar lista „***Descriptive Statistics***" oferă statistici descriptive, univariate, ale variabilelor.

În secțiunea „***Plots***" se pot configura graficele acestei analize. Deoarece CATPCA se bazează destul de mult pe interpretarea grafică, vom acorda atenție și opțiunilor corespunzătoare. Astfel, butonul „***Object***" permite desenarea de grafice referitoare la scoruri. Alegerea opțiunii „***Objectpoint***" permite desenarea norului de puncte al scorurilor, repartiția acestora între două dimensiuni. Opțiunea „***Objectsandvariables (biplot)***" desenează scorurile în raport cu coordonatele variabilelor – saturația în factori (***Loadings***) sau coordonatele centroide (***Centroids***). Secțiunea „***Biplot and Triplot Variables***" permite alegerea variabilelor pentru care se vor desena scorurile, saturația în factori și coordonatele centroide. Este un grafic complet și foarte util în analiză. Secțiunea „***LabelObjects***" permite marcarea punctelor pe grafic. În general se folosește opțiunea implicită, cea prin care se marchează numărul înregistrării din baza de date. Este utilă, mai ales la identificarea cazurilor extreme.

Butonul „***Category***" permite afișarea graficelor privind scalele rezultate. Lista „***CategoryPlots***" permite desenarea coordonatelor centroide și vectoriale pentru fiecare variabilă introdusă (câte un grafic separat pentru fiecare dintre variabile) iar lista „***JointCategoryPlots***" permite afișarea acelorași coordonate, însă pe un grafic unic, pentru toate variabilele introduse. Acesta din urmă este un element foarte important în identificarea comportamentului variabilelor în raport cu factorii latenți identificați.

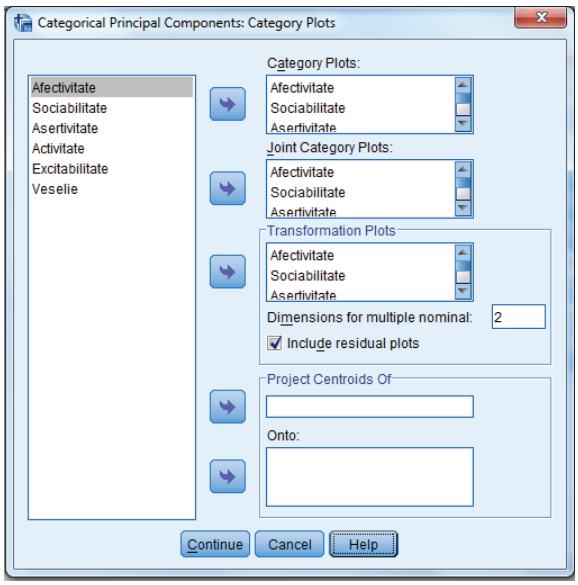

Figura 9 Grafice referitoare la scale

În lista „***TransformationPlots***" putem include variabile pentru care dorim să studiem modul de cuantificare în raport cu indicatorii originali ai datelor. De asemenea, se pot include și grafice ale reziduurilor (distanțe între datele originale și datele cuantificate) și se poate specifica numărul de dimensiuni de referință în cazul metodelor nominal multiple. În secțiunea „***Project Centroids Of***" se poate alege o variabilă și se poate urmări modul în care coordonatele centroide ale acesteia se proiectează pe alte variabile specificate. În cazul în care comportamentul unei variabile este atipic, se va folosi această opțiune pentru a se urmări distanța dintre variabila aleasă și celelalte variabile de referință.

Butonul „***Loadings***" permite desenarea graficelor referitoare la saturația factorilor. Bifarea casetei „***Display component loadings***" va permite afișarea graficului privind saturarea fiecărei dimensiuni cu variabilele corespunzătoare. De asemenea, se poate alege între introducerea tuturor variabilelor sau selectarea anumitelor variabile care să fie reprezentate grafic. Bifarea „***Include centroids***" va permite și reprezentarea coordonatelor centroide a tuturor variabilelor sau a variabilelor selectate.

Tabelul 1 Sumarul cazurilor analizate

Case Processing Summary

Valid Active Cases	4441
Active Cases with Missing Values [a]	206
Supplementary Cases	0
Total	4647
Cases Used in Analysis	4441

a. Excluded cases (first 30 are shown): 16 44 48 72 100 103 243 349 371 382 434 456 461 542 551 554 575 604 640 748 759 776 789 800 806 832 837 890 902 912.

Acestea au fost principalele opțiuni privind configurarea analizei categoriale pe componente principale. Demersul efectuat a presupus discretizarea variabilelor prin transformarea în ranguri, stabilirea nivelului scalei, a metodei de normalizare, a rezultatelor și graficelor afișate. Analiza se inițiază prin apăsarea butonului ***„OK"*** și poate dura o perioadă, mai ales în cazul computerelor slabe.

Primul tabel face un inventar al situației cazurilor analizate. Putem constata că analiza s-a desfășurat pe un număr de 4441 de subiecți (***Valid Active Cases***), 206 subiecți fiind excluși din motive de lipsă a datelor (***Active CaseswithMissingValues***). În subsolul tabelului (vezi tabelul 1) au fost listați subiecții excluși, conform poziției înregistrărilor din baza de date. Deoarece am ales ca situațiile în care există date lipsă să fie excluse din analiză, nu există cazuri suplimentare (***SupplementaryCases***). În concluzie, dintr-un total de 4647 de subiecți, au fost selectați 4441 de subiecți în vederea analizei (***Cases Used in Analysis***).

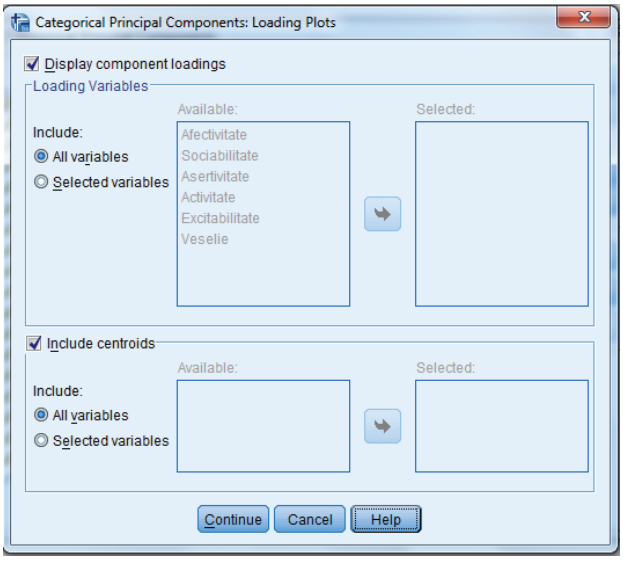

Figura 10 Grafice referitoare la saturația factorilor

Tabelul 2 Statistici descriptive ale variabilelor

Sociabilitate[a]

		Category after Discretization[c]	Frequency	
			Original Data	Analyzed Data
Valid	0	1	77	73
	1	2	479	466
	2	3	540	523
	3	4	613	590
	4	5	739	713
	5[b]	6	764	731
	6	7	762	736
	7	8	582	566
	8	9	43	43
	Total		4599	4441
Missing[d]	System		48	
	Total		48	
Total			4647	4441

a. Optimal Scaling Level: Ordinal.

b. Mode.

c. Ranking

d. Strategy for missing values: Exclude objects with missing values.

Următoarele tabele conțin, pentru fiecare variabilă, statisticile descriptive asociate acesteia. Astfel, pentru variabila „Sociabilitate" scorurile variau inițial între 0 și 8, existând un număr total de 4599 de cazuri valide (***Total***) și de 48 de cazuri excluse din analiză deoarece nu există date (***Missing***). Metoda de optimizare a fost cea ordinală, valoarea modală este 5 pe scala originală iar discretizarea s-a realizat prin calculul rangurilor. În urma acestui proces, au rezultat un număr de 9 categorii (***Category of Discretization***), afișându-se frecvențele absolute (***Frequency***) atât pentru datele originale, nediscretizate (***Original Data***), cât și pentru datele discretizate (***Analyzed Data***). De asemenea, aflăm că strategia de lucru în cazul datelor lipsă a fost aceea de a le elimina din analiză.

Tabelul 3 Tabelul istoricului iterațiilor

Iteration History

	Variance Accounted For		Loss		
Iteration Number	Total	Increase	Total	Centroid Coordinates	Restriction of Centroid to Vector Coordinates
0[a]	4,067135	,000462	7,932865	7,904300	,028565
1	4,077058	,009924	7,922942	7,904300	,018642
2	4,083277	,006219	7,916723	7,898839	,017884
3	4,084812	,001535	7,915188	7,897087	,018101
4	4,085367	,000555	7,914633	7,896226	,018407
5	4,085602	,000234	7,914398	7,895740	,018659
6	4,085709	,000107	7,914291	7,895436	,018855
7	4,085764	,000056	7,914236	7,895245	,018991
8	4,085795	,000031	7,914205	7,895125	,019079
9	4,085813	,000018	7,914187	7,895049	,019138
10	4,085824	,000011	7,914176	7,895000	,019176
11[b]	4,085830	,000007	7,914170	7,894968	,019202

a. Iteration 0 displays the statistics of the solution with all variables, except variables with optimal scaling level Multiple Nominal, treated as numerical.

b. The iteration process stopped because the convergence test value was reached.

Unul dintre cele mai importante tabele se referă la istoricul iterațiilor (tabelul 3). În mod normal, SPSS ar fi afișat doar prima iterație (0) și ultima iterație (11). Specificând în analiză afișarea tuturor iterațiilor, a rezultat un tabel semnificativ mai voluminos.

Tabel 4 Sumarul modelului bidimensional

Model Summary

		Variance Accounted For	
Dimension	Cronbach's Alpha	Total (Eigenvalue)	% of Variance
1	,842	3,355	55,910
2	-,441	,731	12,187
Total	,906[a]	4,086	68,097

a. Total Cronbach's Alpha is based on the total Eigenvalue.

Constatăm că soluția s-a găsit după un număr de 11 iterații, criteriul de convergență fiind atins, creșterea varianței nemaifiind semnificativă. Cele două coloane ale varianței dobândite (Variance Accounted For) indică varianța totală (Total) și cantitatea cu care a crescut varianța între iterații (Increase). Observăm că eigenvalue a crescut de la 4,06 la 4,08 între cele 11 iterații, cele șase variabile saturând de la 67,78% la 68,09% modelul bidimensional analizat. Gradul de saturație exprimat procentual se obține împărțind eigenvalue la numărul variabilelor analizate. Distanțele între saturația explicată de variabile și procentul de 100% al modelului bidimensional se analizează prin pierderi (Loss). Analiza acestora nu este complicată și poate fi dedusă intuitiv. Cert este că cele șase variabile, raportate la modelul bidimensional reușesc să explice un procent de 68,09% din variația factorului latent. Diferența se datorează, probabil, unor alte variabile.

În tabelul 4 se prezintă sumarul modelului bidimensional conceput inițial.Prima dimensiune presupusă (extraversiunea) este acoperită de cele șase variabile în proporție de 55,91%. Este o acoperire foarte bună, iar variabilele sunt consistente (Alpha Cronbach=0,842). Cea de-a doua dimensiune, necunoscută, este acoperită de 12,18% de cele șase variabile. Și în acest caz variabilele au o consistență acceptabilă (Alpha Cronbach=0,441), valoarea negativă arătând probleme legate de sensul în care variabilele saturează acest factor. Totuși, eigenvalue este subunitar (0,731), fapt care ne poate determina să respingem existența, în realitate, a celei de-a doua dimensiuni, reținând doar factorii cu eigenvalue supraunitar.

Până în acest moment am decis existența unui model unidimensional corespunzător celor șase variabile, factorul latent fiind cel presupus (extraversiunea), model cu o putere explicativă de 55,91% și cu o bună consistență (0,842).

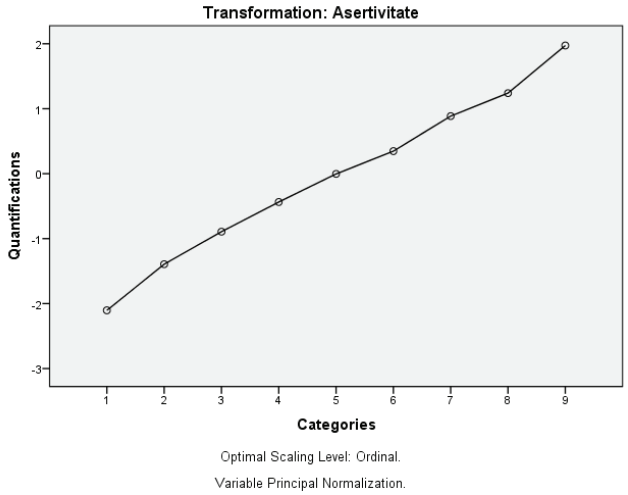

Figura 11 Evoluția variabilei în urma cuantificării

Tabelul 5 Date privind cuantificarea variabilelor

Activitate[a]

Category	Frequency	Quantification	Centroid Coordinates Dimension 1	Centroid Coordinates Dimension 2	Vector Coordinates Dimension 1	Vector Coordinates Dimension 2
0	103	-2,253	-1,753	-,843	-1,693	-,950
1	203	-1,898	-1,402	-,844	-1,427	-,801
2	374	-1,292	-,955	-,573	-,971	-,545
3	507	-,840	-,626	-,363	-,631	-,354
4	726	-,352	-,288	-,105	-,264	-,148
5	820	,054	,049	,008	,040	,023
6	724	,553	,432	,204	,416	,233
7	624	1,151	,848	,516	,865	,485
8	360	1,720	1,299	,713	1,292	,725

Variable Principal Normalization.

a. Optimal Scaling Level: Ordinal.

Metoda de normalizare a fost cea ordinală simplă, iar în tabelul 5 se arată, pentru fiecare variabilă, cum a decurs acest proces. Inițial sunt prezentate categoriile variabilei (Category) și frecvențele absolute (Frequency). Apoi, în urma procesului de normalizare, se prezintă notele standardizate ale fiecărei categorii conform distribuției normale (Quantification). Aflăm că răspunsurile din categoria 0 se situează la 2,25 abateri standard în stânga mediei, cele din categoria 1 la 1,89 abateri standard în stânga mediei, cele din categoria 7 la 1,15 abateri standard în dreapta mediei și așa mai departe. Problema care se pune este dacă această transformare este una liniară, dacă datele pot fi tratate ca date parametrice. În definitiv acesta este scopul normalizării.

În figura 11 este ilustrat graficul acestei transformări. Caracterul liniar se păstrează oarecum pentru scorurile mici, însă se abate semnificativ în cazul scorurilor mari. Variabila poate fi cu greu acceptată ca parametrică, o conformare a utilității CATPCA.

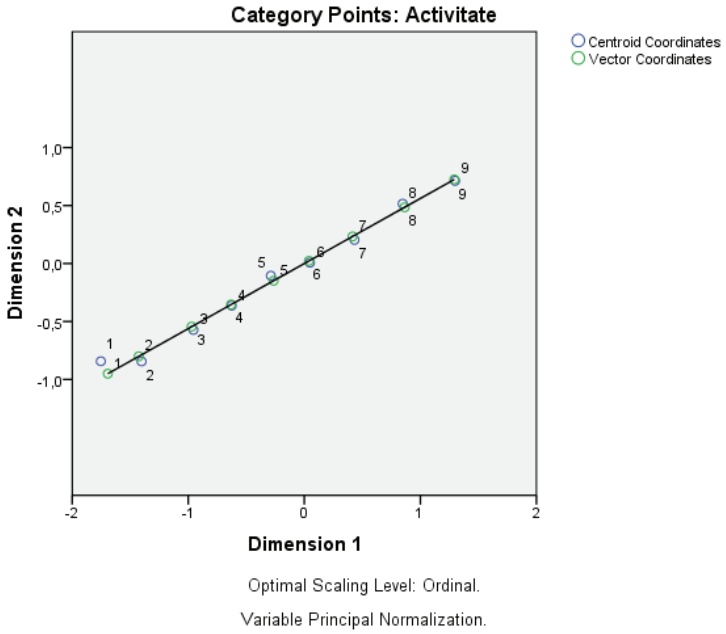

Figura 12 Coordonatele centroide și proiecția coordonatelor vectoriale

De asemenea, sunt ilustrate coordonatele centroide și vectoriale ale fiecărei categorii de răspunsuri în raport cu cele două dimensiuni. Cu toate că cele două concepte au un caracter mai abstract, sensul acestora transpare cu ușurință din analiza figurii 12.

Tabelul 6 Tabelul evoluției variației

Variance Accounted For

	Centroid Coordinates			Total (Vector Coordinates)		
	Dimension			Dimension		
	1	2	Mean	1	2	Total
Afectivitate	,559	,193	,376	,558	,190	,748
Sociabilitate	,641	,024	,332	,640	,023	,663
Asertivitate	,533	,013	,273	,533	,011	,544
Activitate	,565	,179	,372	,565	,178	,742
Excitabilitate	,522	,198	,360	,521	,196	,717
Veselie	,539	,140	,339	,537	,133	,670
Active Total	3,359	,746	2,053	3,355	,731	4,086
% of Variance	55,977	12,440	34,209	55,910	12,187	68,097

Coordonatele vectoriale reprezintă linia oblică pe care sunt reprezentate categoriile de răspuns. Astfel, categoria 0 se află la -1,69 pentru prima dimensiune și -0,95 pentru a doua dimensiune. Cu alte cuvinte, subiecții cu răspunsuri în această categorie se situează în mod cert în zona introvertiților (se află la aproape două abateri standard în stânga mediei pentru prima dimensiune) și în zona medie pentru cea de-a doua dimensiune. Dacă vom observa cu atenție coordonatele vectoriale, vom constata că pentru cea de-a doua dimensiune acestea se situează între -0,95 și 0,75, adică exact în zona scorurilor medii, nediferențiind subiecții așa cum o face prima dimensiune. De aici putem trage concluzia că variabila „activitate" reprezintă bine prima dimensiune și nesatisfăcător pe cea de-a doua. Raportul dintre coordonatele centroide și cele vectoriale nu pune în evidență decât o ușoară problemă la prima categorie (0) și la dimensiunea 2. În realitate scorurile mici tind mai mult spre media dimensiunii 2 față de cum estimează coordonatele vectoriale.

Tabelul evoluției variației (tabelul 6), un nume destul de nefericit ales, deoarece se referă mai puțin la variață și mai mult la coordonatele fiecărei variabile în raport cu fiecare dimensiune, reperul fiind intersecția mediei celor două dimensiuni (punctul de coordonate 0, 0). Se observă că în cazul tuturor variabilelor coordonatele primei dimensiuni sunt mult mai mari în comparație cu coordonatele celei de-a doua dimensiuni, un motiv în plus să considerăm că prima dimensiune este cea relevantă. Ar fi câteva lucruri de spus aici, ca repere pentru analiză. În primul rând, mediile coordonatelor centroide trebuie să fie relativ mari. Variabilele pentru care aceste medii sunt mici (în general sub 0,10) nu au relevanță în cadrul modelului de analiză. În al doilea rând, analizând totalul variației pentru coordonatele vectoriale avem o imagine asupra celor mai importanți factori care explică varianța criteriului. În cazul nostru, extraversiunea este explicată mai ales de afectivitate, activitate și excitabilitate, cu toate că și celelalte componente joacă un rol foarte important.

Tabelul 7 Corelațiile dintre variabile, înainte și după transformare

Correlations Original Variables

	Afectivitate	Sociabilitate	Asertivitate	Activitate	Excitabilitate	Veselie
Afectivitate	1,000	,581	,440	,415	,383	,512
Sociabilitate	,581	1,000	,447	,541	,452	,506
Asertivitate	,440	,447	1,000	,481	,443	,465
Activitate	,415	,541	,481	1,000	,532	,381
Excitabilitate	,383	,452	,443	,532	1,000	,429
Veselie	,512	,506	,465	,381	,429	1,000
Dimension	1	2	3	4	5	6
Eigenvalue	3,340	,728	,587	,537	,437	,371

Correlations Transformed Variables

	Afectivitate	Sociabilitate	Asertivitate	Activitate	Excitabilitate	Veselie
Afectivitate	1,000	,585	,441	,416	,386	,515
Sociabilitate	,585	1,000	,455	,544	,460	,510
Asertivitate	,441	,455	1,000	,484	,447	,462
Activitate	,416	,544	,484	1,000	,538	,383
Excitabilitate	,386	,460	,447	,538	1,000	,427
Veselie	,515	,510	,462	,383	,427	1,000
Dimension	1	2	3	4	5	6
Eigenvalue	3,355	,731	,578	,532	,435	,369

Următoarele două tabele se referă la corelațiile dintre variabilele incluse în analiză. Nu sunt necesare explicații suplimentare, lucrurile transpar foarte clar din tabelul 7. În primul tabel este prezentată matricea inițială de corelații, cu variabilele originale, netransformate. De asemenea, pe ultimul rând sunt prezentate și rădăcinile matricei pentru fiecare variabilă, explicând, din nou, varianța comună. În general, după transformare, corelațiile cresc, fapt care ne

arată modul în care analiza a fost optimizată. În situația în care după optimizare corelațiile scad semnificativ, înseamnă că metoda de transformare folosită nu este adecvată și va trebui înlocuită.

Tabelul 8 Identificarea dimensiunilor în cazul fiecărui subiect

Object Scores

Case Number	Dimension 1	Dimension 2	Afectivitate	Sociabilitate	Asertivitate	Activitate	Excitabilitate	Veselie
1	,661	1,865	3	5	4	7	8	6
2	,414	-,677	6	3	6	5	4	6
3	1,644	-,061	6	7	6	7	7	8

În secțiunea „***Objects***” vom regăsi un tabel imens care conține toți subiecții din baza de date. Pentru fiecare dintre subiecți avem în vedere modul în care sunt reprezentate dimensiunile extrase. Astfel, primul subiect este reprezentat mai bine de a doua dimensiune în comparație cu prima dimensiune, pentru al doilea subiect lucrurile stau la fel, însă cele două dimensiuni sunt antagonice și așa mai departe. De asemenea, programul prezintă și scorurile efective ale subiecților la fiecare dintre variabilele supuse analizei. Tabelul se poate folosi pentru inspectarea de finețe a datelor și în vederea identificării modului în care lucrează concret dimensiunile.

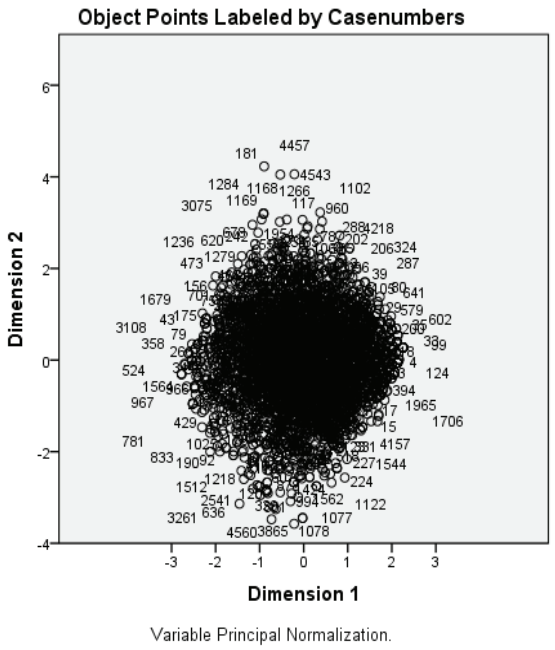

Figura 13 Norul de puncte al scorurilor asociat celor două dimensiuni extrase

Asociat acestui tabel vom regăsi și norul de puncte al scorurilor asociat celor două dimensiuni extrase. Putem remarca o serie de scoruri deplasate, în partea de sus a graficului, subiecți care sunt reprezentați mediu la nivelul primei dimensiuni, însă puternic la nivelul celei de-a doua dimensiuni. Inspectarea acestui grafic ne confirmă supozițiile numerice. Prima dimensiune este mai „grupată”, reprezentând mai bine subiecții în comparație cu a doua dimensiune, la care constatăm un grad mai ridicat de dezorganizare. Graficul din figura 13 nu reprezintă altceva decât transpunerea într-un sistem de coordonate bidimensionale a datelor din tabelul prezentat mai sus.

Tabelul 9 Saturația în factori

Component Loadings

	Dimension 1	Dimension 2
Afectivitate	,747	-,436
Sociabilitate	,800	-,151
Asertivitate	,730	,106
Activitate	,751	,422
Excitabilitate	,722	,443
Veselie	,733	-,365

Variable Principal Normalization.

Poate cel mai important tabel este tabelul saturației în factori (***Component Loadings***). Similar analizei factoriale clasice, CATPCA indică proporția de varianță a fiecărei dimensiuni cu care contribuie fiecare dintre variabile. Într-adevăr, prima dimensiune este cea relevantă, aici fiind cei mai ridicați coeficienți de saturație. Cele șase variabile introduse în model contribuie la explicarea extraversiunii, așa cum s-a constatat anterior. Nu putem să ignorăm contribuțiile variabilelor la cea de-a doua dimensiune, mai ales că, în anumite cazuri, acestea sunt destul de ridicate. Putem oare să presupune existența unei a doua dimensiuni importante?

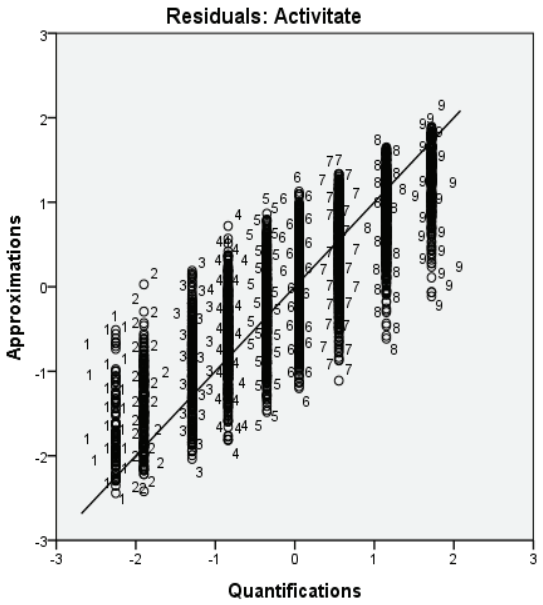

Figura 14 Analiza reziduurilor

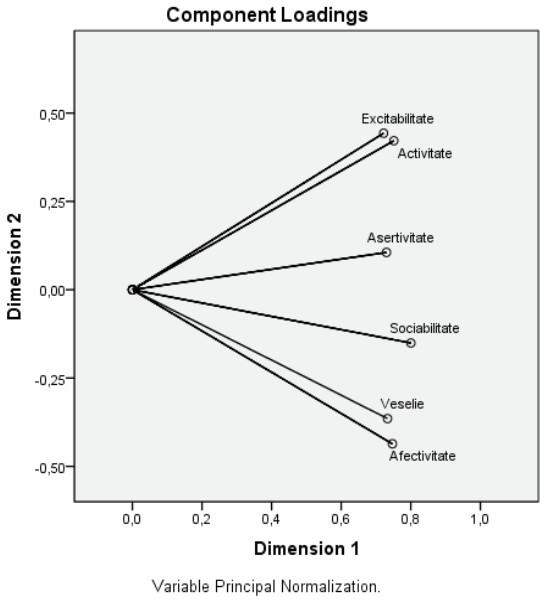

Figura 15 Coordonatele saturației în factori

Răspunsul îl găsim analizând graficul din figura 14, graficul coordonatelor saturației în factori. Se observă trei zone distinct marcate în raport cu cea ce-a doua dimensiune, prima dimensiune nepunând probleme. Prima zonă identifică Excitabilitatea și activitatea. Aici, scorurile mari la extraversiune se asociază cu scoruri mari la a doua dimensiune. Următoarea zonă cuprinde asertivitatea și sociabilitate, în care se păstrează oarecum echilibrul dintre cele două dimensiuni. În fine, ultima zonă se referă la veselie și afectivitate, unde scorurile la extraversiune se asociază cu scoruri mici la a doua dimensiune. Putem trage de aici concluzia că există mai multe tipuri de extravertiți. Avem de a face cu extravertiții afectivi, veseli, care „se bagă în suflet", te copleșesc cu atenție, glumesc mai tot timpul și extravertiții activi, pragmatici, puși pe treabă, entuziaști. Analiza acestui grafic poate furniza informații suplimentare legate de dimensiunile analizate. Iată că, cea de-a doua dimensiune se referă la o componentă structurală a extraversiunii pe care o putem numi axa pragmatism-afecțiune a extraversiunii.

Analiza reziduurilor permite aprecierea distanțelor la care se situează categoriile variabilelor în comparație cu dreapta de regresie normală. În cazul variabilei „activitate" putem constata că scorurile mici (0, 1, 2 și chiar 3) supraestimează repartiția normală. Iar scorurile mari (8, 7 și 6) o subestimează. În raport cu distribuția normală ar fi trebuit să avem mai puți subiecți cu răspunsuri orientate către scoruri mici și mai mulți subiecți cu răspunsuri orientate către scoruri mari. Un motiv în plus în favoarea renunțării la analiza factorială clasică și a abordării tehnicilor de tip categorial. În realitate, distribuțiile nu au un caracter normal și mai curând unul logistic.

Scurte concluzii

Analiza categorială pe componente principale se poate folosi cu succes în situațiile în care variabilele nu pot fi supuse analizei factoriale clasice, fie ca urmare a naturii acestora, fie din cauza nerespectării asumpțiilor. Mai mult, datorită bogăției de informații și a fineței analizei, este de dorit ca analiza factorială clasică să fie completată cu CATPCA, rezultând astfel un tablou complet al dimensiunilor extrase.

Bibliografie

Meulman, J., Heiser, W., & SPSS, Inc. (2007). *PASW Categories 18.* Illinois: SPSS Inc.

Opariuc-Dan, C. (2009). *Statistică aplicată în științele socio-umane. Noțiuni de bază. Statistici univariate.* Cluj-Napoca: ASCR.