Eksploracijska analiza podataka

Dragoš, Matea

SVEUČILIŠTE U ZAGREBU
FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE
SVEUČILIŠNI PREDDIPLOMSKI STUDIJ

Matea Dragoš

EKSPLORACIJSKA ANALIZA PODATAKA

ZAVRŠNI RAD

Zagreb, rujan 2015.

SVEUČILIŠTE U ZAGREBU
FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE
SVEUČILIŠNI PREDDIPLOMSKI STUDIJ

Matea Dragoš

EKSPLORACIJSKA ANALIZA PODATAKA

ZAVRŠNI RAD

Voditelj rada: Prof. dr.sc. Tomislav Bolanča

Članovi povjerenstva:

Prof. dr. sc. Tomislav Bolanča

Doc. dr. sc. Šime Ukić

Izv. prof. dr. sc. Juraj Šipušić

Zagreb, rujan 2015.

ZAHVALA

Zahvaljujem se mentoru prof. dr. sc. Tomislavu Bolači na pomoći i razumijevanju koje mi je pružio pri izradi ovog rada.

Najviše od svega zahvaljujem roditeljima i prijateljima bez kojih sve ovo ne bih uspjela.

SAŽETAK

Ovaj završni rad obrađuje temu eksploracijska analiza podataka. U prvih nekoliko poglavlja rad se fokusira na definiranje eksploracije, multivarijatne analize podataka, te kombinacija tih dviju analiza. Kombinacijom eksploracijske i multivarijatne analize podataka dobivamo nove mogućnosti za rješavanje složenih znanstvenih procesa. U središnjem poglavlju opisani su napoznatiji modeli multivarijatne analize podataka, među kojima je najpoznatiji model analize glavnih komponenata koji nudi način prikaza podataka u svrhu pronalaženja njihovih sličnosti i različitosti. U posljednjem poglavlju je opisan primjer PCA modela u statističkom programu Statistica.

Ključne riječi: eksploracija, PCA, multivarijatna analiza podataka, program Statistica

ABSTRACT

This thesis describes exploratory data analysis. In first few chapters, the thesis focuses on defining term exploration and multivariate data analysis along with the combination of that two analysis. By combinating explorational and multivariate data analysis, we get new features for solving complex science processes. The central chapter describes the most famous models of multivariate data analysis including the famous model of principal component analysis which offers a way to display data for the purpose of finding their similarities and differences. The last chapter describes the example of PCA model using programme Statistica.

Keywords: exploration, PCA, multivariate data analysis, programme Statistica

SADRŽAJ

1. UVOD

Proces eksploracije sastoji se od slijeda postupaka koje treba provesti kako bi se došlo do pouzdanih informacija što nam pomaže u boljem donošenju odluka o planiranim intervencijama.

Ti se zadaci svrstavaju u nekoliko faza:

- Definiranje problema i ciljeva istraživanja
- Postavljanje hipoteza
- Određivanje izvora podataka
- Određivanje metoda
- Određivanje vrste uzorka
- Analiza podataka i interpretacija rezultata
- Sastavljanje izvještaja

Povećana količina složenih podataka zahtijeva učinkovitije alate za rudarenje podataka i obrnuto, te inženjera koji će prvo istražiti veze, povezanosti i skupine, a potom pomoći analitičaru za postavljanje hipoteza. Moderne znanosti i industrije trebaju brza rješenja za složene biološke probleme koje zahtijevaju svestrane i nesvakidašnje metode za uspješno rukovanje složenim uvijetima mjerenja. Stoga je potrebno prevesti parametre iz više ili manje apstraktnih matematičkih modela u stvarnom svijetu jezika izražavajući čimbenike za donošenje novih i preciznijih odluka. Slijedom toga, alati su potrebni kako bi se olakšalo optimalno iskorištavanje ogromne količine informacija raznolikih instrumenata. Zato ovdje veliku ulogu ima istraživačka analiza podataka kombinirajući multivarijantne matematičke alate i istraživačke pristupe za analizu podataka postavljene s koreliranim varijablama.

Eksploracijska tehnologija koja se pojavila zajedno sa matematičkim zbivanjima rezultira svestranost multivarijantne analize koja čini ovu tehnologiju kao nezamijenjiv alat za istraživačke svrhe. Dvije temeljne značajke multivarijatne analize podataka su znanstveni proces koji vodi iz multivarijatnog promatranja do informacija i matematički razvoj s opisom kao neke od najvažnijih svestranih multivarijantnih modela koji se koriste u kemometriji.

Posljednih godina znanstvena disciplina poznata kao kemometrija postigla je veliki razvitak. Taj razvitak inicirao je napretku laboratorijskoj automatizaciji te inteligentnim instrumentima, korištenjem močnih računala i softvera. Dakle, kemometrija je postala alat u svim dijelovima kvantitativne kemije, no posebno u području analitičke kemije u kojem su analitičari tijekom rada sve više suočeni s potrebom korištenja matematičkih i statističkih metoda.

Franck i Kowalski dali su poznatu definiciju kemometrije: „Kemometrija se može definirati kao kemijska disciplina koja koristi kemijske, statističke, i druge metode temeljene na logičnim objašnjenjima:

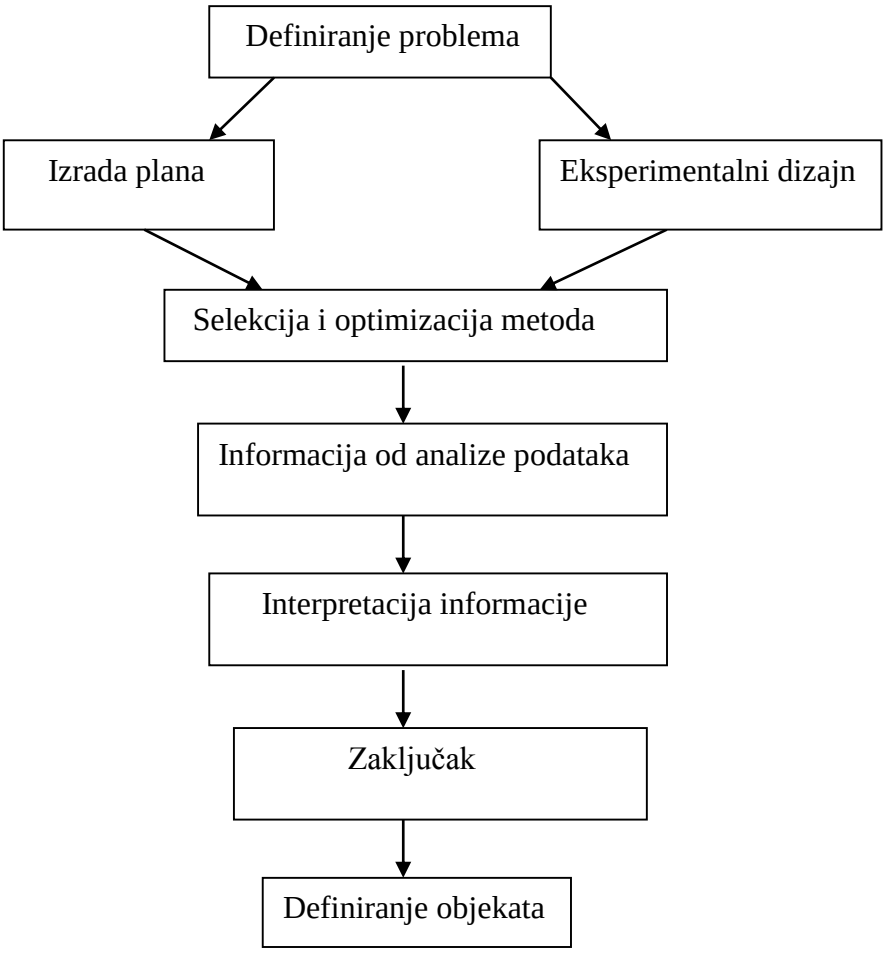
- za dizajniranje ili odabir optimalnih mjerenih postupaka i eksperimenata
- za osiguravanje maksimalnih kemijskih informacija analizirajući kemijske podatke.“ [1]

Još jedna definicija koji su dali Franck i Kowalski glasi: „ Kemijski alati su vozila koja mogu pomoći kemičaru da se učinkovitije presele na putu mjerenja do poznatih informacija.“ [2]

Kemometrija je zajednički nazivnik za sve moguće primijenjene alate kako bi se racionalnije analizirala kemijska mjerenja. Korištenje pojma kemometrija ima važnu svrhu gdje cijeli problem treba promatrati, analizirati i tumačiti na direktan ili neizravan kemijski kontekst. Kemometrija nam nudi značajno nove mogućnosti u pristupu multivarijantnog problema koji savršeno nadopunjuje klasično znanstvene metodologije, čime se osigurava tehnologija u smislu da je holističko rješenje kombinirajući strategije i alate u samom središtu primjene. Holističke metode u svakom koraku mogu pomoći u rješavanju problema.

Dakle, holistička strategija (kao što je prikazana na slici 1.) s različitim metodama kemometrije u svakom od svojih koraka može pomoći u rješavanju problema u stvarnom svijetu kemije i drugih znanstvenih disciplina.

[1] J. W. Einax, H. W. Zwanziger, S. Geiß, Chemometrics in enviromental analysis, str.19.
[2] J. W. Einax, H. W. Zwanziger, S. Geiß, Chemometrics in enviromental analysis, str.20.



Slika 1. *Holistički prikaz rješavanja kemometrijskog problema*

Eksploracijsko multivarijatna analiza podataka, kemometrija i multivarijatna analiza su tehnologije, a ne discipline, jer to su alati koji rade u kontekstu različitih disciplina prilagođeni okolnostima, poput kemije, biologije ili psihologije.

2. EKSPLORACIJSKA ANALIZA PODATAKA

Pojam eksploracijska analiza podataka prvi put je korištena u psihološkim znanostima.
Formalna definicija eksploracijske analize podataka može se naći, ali možda najbližu i najizravniju daje Hoaglin: "Ukratko, eksploracijska analiza podataka naglašava fleksibilnost u potrazi za tragovima i dokazima, dok je potvrdna analiza podataka najpreciznije vrjednovanje dostupnih dokaza." [3]

Eksploracijska analiza bavi se istraživačkim pristupom koji je fokusiran na izradi analize podataka, ocjenjivanjem prikladnih modela i njihovih podataka, i ako je potrebno, modificiranje modela i/ ili temelja podataka. Na svakom koraku, novi uvid u smislu korelacija između objekata ili varijabli, udaljavanjem uzoraka ili učinaka do potrebnih zaključaka. Eksploracijski pristupi omogućuju rezultate pomoću kojih analitičar može definirati te pronaći kombinacije uvijeta analize koje pružaju optimalno razumijevanje podataka.
Eksploracijska analiza podataka nikad ne ispriča cijelu priču, ali ništa drugo ne može poslužiti kao kamen temeljac, kao prvi korak.[4]

Zbog sve ubrzanog razvitka računala, sada je moguće provesti novu istraživačku analizu u čistom elektroničkom obliku korištenjem digitalnih prikaza u kombinaciji s interaktivnom grafičkim sučeljem između zapažanja i parametra modela. Imajući te podatke u računalu omogućen je pregled objekata i varijabli, te za primjenu različitih metoda među mnoštvom mogućih matematičkih tretmana. Dakle, nakon što su podaci predstavljeni u računalu omogućuju opsežne pokuse podataka u istom smislu kao kada kemičar provodi eksperimente u laboratoriju.

[3] Claus A. Andersson, M. Sc., Chew Eng, Exploratory Multivariate Data Analysis with Applications in Food Technology, str.5.
[4] Claus A. Andersson, M. Sc., Chew Eng, Exploratory Multivariate Data Analysis with Applications in Food Technology, str.5.

Visokom izvedbom i učinkovitom matematičkom okruženju za obavljanje takvih pokusa može se izvoditi na svakom modernom računalu, čime se otvara put za znanstvenike da preuzmu odgovornost za istraživanje strukture njihovih vlastitih podatka ne uzimajući u obzir izmjerene podatke, već u vezi njih kao simboli dinamičkih sustava koji se mogu rekombinirati i pretvoriti u smislenu informaciju korištenjem multivarijantne kemometrije

2.1. Multivarijatna analiza podataka

Multivarijatna analiza podataka definira se kao skup statističko-matematičkih postupaka pogodnih za analizu podataka promatranjem više varijabli pri čemu je svaka na svoj način važna, te su zbog toga ti podaci shvačeni kao multivarijantni ili multidimenzionalni analitički postupci. Primjenjuje se u svim situacijama gdje se u istraživanjima promatra velik broj varijabli koje su u međusobnim odnosima, te gdje se zahtijeva utvrđivanje osnovnih odnosa među podacima.

Najčešći razlozi korištenja multivarijatne analize podataka je dostupnost računalnih programa koji su danas sve više u upotrebi, potrebnih za složena izračunavanja. Postoje različita istraživanja za ostvarivanje što boljih ciljeva multivarijantne analize podataka.
Oni glase:

1. Prikupljanje podataka bez gubitka važnih informacija jer svaka informacija nam nešto govori
2. Grupiranje podataka na osnovi mjernih karakteristika
3. Razlike između varijablama
4. Predviđanje
5. Formiranje hipoteza i testova zbog potvrđivanja/odbacivanja postavljenih pretpostavki

Zahvaljujući multivarijatnoj analizi podataka i tehnikama unutar te iste analize, danas se svaka hipoteza može klasificirati koja se mora prethodno formulirati u razumljivom obliku, kako bi se mogli primijeniti postupci analize. Jedno od temeljnih znanstvenih istraživanja je načelo o međusobnoj povezanosti pojava, te kao takva istraživanja zadatak je utvrditi što točniju vrstu i prirodu takve povezanosti. Utvrđivanjem povezanosti moguće je predvidjeti promjene zbivanja u pojavama koje su s prvom povezane. Usporedno s tim, ostvaruje se i drugi cilj znanstvenog istraživanja, a to su razumijevanje i objašnjenje pojava.

Da bi se što točnije utvrdile veze između raznih pojava, kao i prirodu takvih veza, danas se sve više pristupa poboljšanim tehnikama i stvaranju novih tehnika multivarijatne analize. Tome pridonose i razni računalni softveri.

Podaci dobiveni mjerenjem promjena na pojavama koje se promatraju nazivaju se varijable. Varijable reflektiraju promjene na pojavama koje se proučavaju. Kako su promjene međusobno povezane, upravo se ta povezanost reflektira i na varijablama. Međusobna povezanost između varijabli izražava s nekim od brojnih indikatora kao stupanj takve povezanosti. Takvi indikatori zovu se koeficijenti korelacije. Utvrđivanjem koeficijenata korelacije između promatranih varijabli utvrđujemo stupanj međusobne povezanosti između promatranim pojavama.[5]

Upravo zato je vrlo važan postupak utvrđivanje korelacije među varijablama. Znanstvena istraživanja nastoje obuhvatiti što veći broj pojava, posebice kada su te pojave višestruko povezane. Takav pristup rezultira većim brojem varijabli, te proučavanjem povezanosti između tih varijabli. To znači da u rezultatu imamo i velik broj koeficijenata korelacije. Ako promatrane pojave iskazuju međusobnu povezanost, samim time postoji mogućnost utvrđivanje veličine korelacije svake promatrane pojave sa svakom drugom pojavom. Prema tome u rezultatu će takva proučavanja dobiti matricu s koeficijentima korelacije iz kojeg znamo da su promatrane pojave u međusobnoj vezi, ali o uzorcima se još ne zna ništa. Ovi analitički postupci veoma su složeni, te ih je još teže interpretirati. Da bi se što točnije utvrdila priroda tih veza, pristupa se što poboljšanijim tehnikama multivarijatne analize. Tome pridonose i razni statistički softveri za obradu podataka.

Iz svega navedenog u matematičkoj statistici razvilo se posebno područje istraživanja čiji je cilj proučavanje i tumačenje složenih veza. Zbog toga je proizišla potreba za upoznavanjem postupaka i metoda analize koji će omogućiti bolje razumijevanje odnosa između raznovrsnih pojava. Upravo takve mogućnosti nam pruža multivarijatna analiza podataka i njoj pripadajući modeli.

[5] J. W. Einax, H. W. Zwanziger, S. Geiß, Chemometrics in enviromental analysis, str.101.

2.2. Eksploracijska multivarijatna analiza podataka

Eksploracijsko multivarijatna analiza podataka je ujedinjenje istraživačkih analiza podataka i analiza multivarijatnih podataka. Kombinirajući ova dva različita pristupa nudi nam nove mogućnosti i puteve za rješavanje složenih znanstvenih i industrijskih problema. Statistika datira od nekoliko stotina godina u vrijeme kada je samo teorija bila razvijena, te kada su numerički i računalni alati bili iznimno ograničeni u odnosu na danas. Dakle, matematička apstrakcija od bilo koje vrste problema je morala biti predstavljena na vrlo jednostavan matematički odnos koji je zahtijevao procjenu od samo nekoliko parametara. Prije računala, statističari su bili prisiljeni prepisivati svoje probleme u rješive oblike, čime se ograničavanje domene sastojalo od jednog modela.

Parametri viših brojeva su se mogli lako procijeniti, no time se zahtjevalo korištenje multivarijatnih modela koji sada pružaju snalažljiv i klasičan pristup, primjerice hipoteza ispitivanja i analiza varijance. Tri glavna aspekta istraživačkog multivarijantnog pristupa u smislu istraživačke analize glavnih sastavnica koja nudi značajne i nove mogućnosti koja su od temeljne važnosti na način provođenja istraživanja. Prvi aspekt između opažanja su iskorišteni za poboljšanje razumijevanja složenih sustava utvrđivanjem latentnih faktora promatranja. To se obavlja s namanjim pretpostavkama i modelima koji su optimalni kada se primjenjuju na nepoznate razdiobe.

Drugo, izravna primjena na primjer analize glavnih komponenata koja pomaže analitičaru suziti važne čimbenike, te konačno, metode, su prikladne za primjenu u kontekstu rješavanja problema uvođenjem matematičkih metafora koje odgovaraju stvarnim životnim čimbenicima čime smanjuju stvaran matematički ili kemijski problem. Neki modeli, poput PCA, neuronske mreže i genetski algoritmi su unaprijed stvoreni zbog matematičkih modela na kojima se temelje, odnosno linearnih odnosa, eksponencijalnih funkcija i logičkih operatora, te je njihova prilagodljivost zbog velikog broja parametara uključena u mnogim baznim funkcijama. Novi eksploracijski multivarijatni alati za analizu hipoteza zahtijevanju manje znanja na početku novog istraživačkog projekata.

U multivarijatnoj analizi izvođenoj iz eksperimentalne matematike, nema početne hipoteze kao što se zahtjevala prije, osim da se u analizi eksperimentalni dio analizira. Prema tome, poznate značajke sustava ispituju se na način kako bi se osigurala valjanost uspoređujući s poznatim varijacijama i grupacijama u podacima kao i vanjskim znanjem. Zbog različitih razloga, tijekom zadnjih stoljeća opća potražnja za novim znanjima promijenila se od onih općih, egzistencijalnih i temeljnih pitanja prema rješavanju specifičnih problema.

Cilj formulacija je učinkovit i precizan pokazatelj za uspjeh. Velik dio javnih znanstvenih istraživanja usmjeren je na popravljanje trenutnih problema, kao što su identificiranje mehanizama virusa, poboljšanje hranjive vrijednosti hrane ili pronalaženje lijeka za rak. Ciljevi za rješavanje ovih problema danas nisu toliko značajni o prirodi problema, niti dobitak temeljnih znanja, nego u kratkom vremenu riješiti probleme koji su trenutno u oku kritičnog društva. U skladu s ovim zahtjev za dobivanje brzih predmetnih popravaka, zasad još nepoznatih, područja koriste brze, osjetljive i ponovljive holističke metode koje su potrebne da se u kratkom vremenu i pri pristupačnim troškovima može donositeljima pružiti odluka na svim razinama s visokokvalitetnim podacima koji mogu ocijeniti multivarijatne metode te pružiti uvid u znanje kako bi se omogućilo upravljanje i kontrola raznih vrsta problema. Kvaliteta podataka je važan aspekt u poboljšanju ishoda multivarijatnih istraživanja.

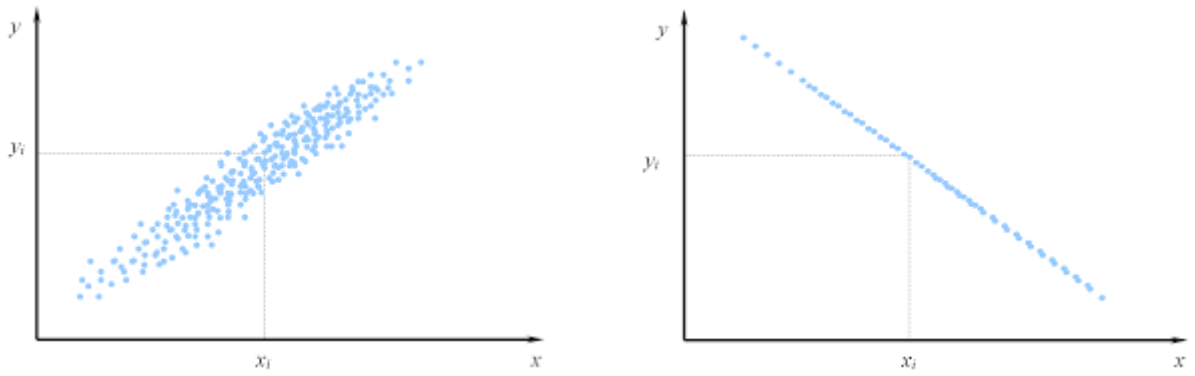
Prema tome, analitičar bi trebao biti u mogućnosti identificirati poznate uzorke ili grupirati u odnosu na njegovo/njezino znanje uzoraka iz sustava pod prismotrom i nadalje utvrditi fizičko/kemijski karakter temeljne glavne komponente odabrane od strane multivarijatnog modela. Prema tome, ovaj korak je sastavni dio uspješnosti jer se koristi za ispitivanje značajnih sustavnih odstupanja.

3. EKSPLORACIJSKO MULTIVARIJATNE OBRADE PODATAKA

U društevenim i prirodnim znanostima veliku važnost pridaje se eksploracijsko multivarijatnim metodama obrade podataka. Razne pojave koje se istražuju u tim znanostima vrlo su složene. Multivarijatne metode moraju izmjeriti relevantne podatke te iz tih podataka izvući pravi smisao. Mjerenja su često podložna pogreškama, stoga se proces mjerenja često sastoji od ponavljanih pokušaja mjerenja istog problema na različite načine. U znanosti, jedan od temeljnih ciljeva je utvrđivanje povezanosti između pojava, bilo da se radi o utvrđivanju uzročno-posljedičnih odnosa ili samo korelacija. U multivarijatnom pristupu analizira se međuodnos više od dviju varijabli. Problemi koji se istražuju su složeni, tako da zahtijevaju ispitivanje većeg broja ispitanika da bi se moglo odgovoriti na problem istraživanja. Upravo je i to glavni razlog primjene eksploracijsko multivarijatne obrade podataka.

3.1. Korelacija

Korelacija ili povezanost je temeljna pojava u znanosti jer utvrđivanje povezanosti između pojava jedan je od temeljnih ciljeva. Povezanost između dviju varijabli može biti različitog stupnja, odnosno visine te može biti pozitivna i negativna. Koeficijent korelacije opisuje stupanj povezanosti između dvije varijable i ide u rasponu od -1, 0, pa sve do 1. Povezanost između dviju varijabli za pozitivnu i negativnu korelaciju prikazano je na slici 2.



a) b)

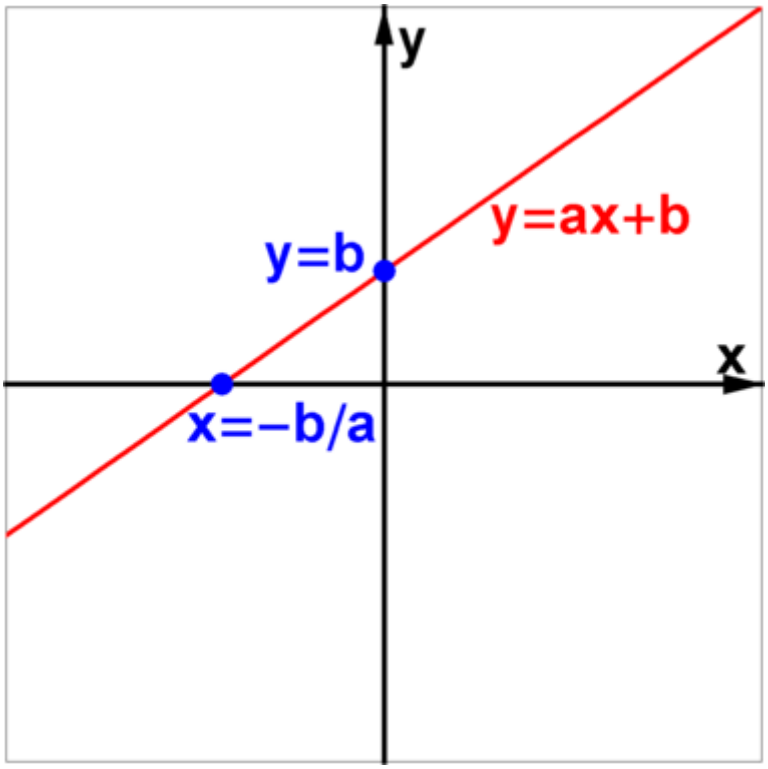
Slika 2. *Prikaz povezanosti između varijabli za pozitivnu a) i negativnu korelaciju b)*

Koeficijent korelacije prvi je predložio Francis Galton, a računski ga je usavršio Karl Pearson. Pearsonov koeficijent korelacije je linearni koeficijent korelacije. Korelacije prikazane na slikama su linearne, no isto tako međusobni odnos varijabli može biti i nelinearan.
Kružići na slikama predstavljaju položaj ispitanika u prostoru koji je definiran proučavanim varijablama. Upravo njihov raspored kružića dozvoljava da se njihov položaj opiše pravcem.

Kada međusobni odnos varijabli nije linearan, taj odnos je složeniji i teže opisljiv. Najjednostavnija jednadžba pravca u matematici koja se opisuje na linearnom području je jednadžba pravca:

$$y = ax + b \qquad (1)$$

U toj jednadžbi (1) koeficijent a predstavlja udaljenost od ishodišta koordinatnog sustava do mjesta na kojem pravac siječe koordinatu os y, a koeficijent b predstavlja nagib pravca. Koeficijent b je jednak tangensu kuta koji zatvara pravac sa koordinatom osi x. (slika 3)



Slika 3. *Prikaz jednadžbe pravca*

Linearni model utemeljen na jednadžbi pravca vrlo dobro opisuje pojave u raznim znanostima i upotrebljava se kao glavni model. Uz linearni model, temelj multivarijatnih metoda obrade podataka još je i multivarijatna normalna raspodjela. Prema ovome, temeljne postavke multivarijatnih metoda obrade podataka su linearnost i multivarijatna normalna raspodjela varijabli. Naravno, postoje i mnogi statistički postupci te računalni programi za normalizaciju varijabli. Normalizacija varijabli se obavlja samo kada rezultati znatno odstupaju od normalne raspodjele.

3.2. Faktorska analiza

Faktorska analiza je skup statističko-matematičkih postupaka kojima se, polazeći od većeg skupa varijabli, utvrđuje manji skup temeljnih varijabli ili faktora (Fulgosi, 1984).[6]
Faktorska analiza je temelj multivarijatnih metoda iz kojih se sve ostale multivarijatne metode mogu smatrati slučajevima faktorske analize. Uz smanjenje broja početnih varijabli, glavni cilj je utvrditi strukturu zadanog područja tj. prostora, odnosno što mjeri pojedini skup mjernih instrumenata, te se tada može govoriti o klasifikaciji varijabla.

Faktorska analiza je pojam za velik broj statističko-matematičkih postupaka kojima se na temelju analize korelacija početnih varijabli utvrđuju faktori. Početne varijable predstavljaju nešto očigledno, te su to najčešće mjerni instrumenti. Ova analiza se može prihvatiti kao opća znanstvena metoda. Thomson je 1951. pisao o praktičnom i teorijskom cilju faktorske analize, te je rekao: „Praktična želja je sažeti opis ljudskog intelekta na komparativno manji broj navoda umjesto glomaznog zapisa testovnih rezultata s namjerom davanja profesionalnog ili pedagoškog savjeta.“ [7]
Analiza je prvobitno bila razvijena u području psihologije, ali je kasnije bila prihvaćena i u drugim znanstvenim disciplinama poput pedagogije, sociologije, ekonomije i drugih. Usmjerena je na proučavanje povezanosti početnih varijabli koje predstavljaju pojave, a utvrđivanje povezanosti pojava i uzorka temeljni su ciljevi znanosti.

Konačan produkt faktorske analize su faktori koji su linearne kombinacije početnih varijabli, te se nazivaju latentnim što znači prikrivenim varijablama jer ih tek treba otkriti ovom analizom. Faktori se mogu smatrati uzorkom početnih varijabli jer oni daju objašnjenje za povezanost pojava koje ispituju početne varijable. Prema tome, faktori otkrivaju potencijalne uzorke povezanosti pojava.

[6] Aleksandar Halmi, Temelji kvantitativne analize u društvenim znanostima, Alinea, Zagreb, 1999., str.101.
[7] Milko Mejovšek, Uvod u metode znanstvenog istraživanja u društvenim i humanističkim znanostima, Naklada Slap, 2003., Zagreb, str.153.

Analiza započinje matricom početnih varijabli i njezin cilj je upravo analiza te matrice. Broj parova varijabli se može odrediti po formuli:

$$\frac{n(n-1)}{2} \qquad (2)$$

Gdje je:

n- ukupan broj varijabli

Na primjer, za izračunavanje 500 varijabli, pomoću gornje formule, broj korelacije koje bi se trebalo izračunati iznosi 124 750. Proučavanje tako velikog broja korelacija i donošenja određenih zaključaka na temelju toga je zadaća koju je nemoguće izvršiti. Upravo takva nerješiva zadaća može se rješiti faktorskom analizom.

U matričnoj algebri koja se koristi za opisivanje raznih matematičkih računskih operacija, opisuje se vektorima i matricama, te se ne samo u faktorskoj analizi koristi, nego i u svim multivarijatnim metodama. Matrica koja se obilježava slovom R kvadratnog je oblika i simetrična s obzirom na dijagonalu koja ide od gornjeg lijevog ugla matrice, pa sve do donjeg desnog. U dijagonali R matrice nalaze se jedinice gdje je varijabla u korelaciji sa samim sobom, a izvan dijagonale nalaze se korelacije svih mogućih parova početnih varijabli. Jedinica u dijagonali ne označuje samo povezanost varijable u korelaciji sa samim sobom, označuje i ukupnu varijancu varijable.

Faktorska analiza može biti:

1. Eksploracijska
2. Konformativna

Najprije se razvila eksploracijska analiza, a tek kasnije konformativna. Eksploracijska faktorska analiza je analiza kojom treba otkriti faktore u nekom području kada su broj i struktura faktora nepoznanica. Konformativana faktorska analiza se primjenjuje kada se npr. provjeravaju hipoteze o broju i strukturi faktora za koje postoje dokazi, tj. kada postoje podaci o broju i strukturi faktora u nekom području, te kada je to potrebno provjeriti.

Primarni je cilj identifikacija faktora i određivanje stupnja do kojeg su izvorne varijable objašnjene svakom dimenzijom - faktorom. Za razliku od PCA koja nije bazirana ni na kakvom statističkom modelu, faktorska analiza određena je specifičnim statističkim modelom. Zajednički faktor nevidljiva je, hipotetska varijabla koja pridonosi varijanci iz barem dvije izvorne varijable. Izraz faktor najčešće se odnosi na zajednički faktor. Jedinstveni ili specifični faktor, također je nevidljiva, hipotetska varijabla koja pridonosi varijanci u samo jednoj izvornoj varijabli.

U matričnom obliku ove jednadžbe mogu pisati:

$$Y = XB + E \qquad (3)$$

gdje je:
X- matrica faktorskih opterećenja
B - matrica zajedničkih faktora.

Faktorska opterećenja jednostavne su korelacije između bilo koje izvorne varijable i faktora i ključ su za razumijevanje prirode samog faktora. Kvadrat faktorskih opterećenja je komunalitet i predstavlja udio varijance određene izvorne varijable u ukupnoj varijanci (sumi varijanci svih varijabli u analizi) koji je objašnjen uvrštenim faktorom.

Ostatak koji nije objašnjen uvrštenim faktorom ili faktorima, dakle razlika ukupne varijance i komunalitete je dio varijance specifičan, jedinstven za svaku pojedinačnu varijablu.

Dakle:

(faktorsko opterećenje)^2 = komunalitet
ukupna varijanca - komunalitet = specifična varijanca

ili u standardiziranom obliku:

1 - komunalitet = specifična varijanca

Zadatak faktorske analize je procijeniti komunalitete za svaku varijablu. Ukoliko rezultate faktorske analize nije moguće interpretirati, moguće ih je pojasniti i učiniti manje subjektivnim metodama faktorske rotacije. Rotacija faktora se provodi primjenom linearne transformacije. Takvu rotiranu matricu, u kojoj svi koeficijenti, iznose 0 ili ± 1, lakše je interpretirati nego matricu punu intermedijarnih elemenata. Najviše metoda rotacije nastoje optimizirati funkcije matrice opterećenja koja mjeri koliko su bliski elementi 0 ili ±1. Rotacije mogu biti ortogonalne (orthogonal) ili kose (oblique). Za većinu problema, najbolja je ona rotacija koju je najlakše interpretirati. Ako dvije rotacije rezultiraju različitim interpretacijama ne znači da su one u konfliktu. One su dva različita načina gledanja na istu stvar, dva različita vidika u prostoru zajedničkih faktora. Zaključak koji ovisi o samo jednoj korektnoj rotaciji može biti neispravan. Dakle, faktorska analiza je postupak za istraživanje mogućnosti suzbijanja velikog broja varijabli na manji broj između kojih postoji povezanost.

3.3. Analiza glavnih komponenata

Karl Pearson je 1901. godine prvi opisao ovaj komponentni model predstavljen kao metoda glavnih komponenata koju je 1933. godine razvio Hotelling.

PCA ili analiza glavnih komeponenata je najjednostavnija metoda multivarijantne statistike. Možemo ju definirati kao način prikaza podataka u svrhu pronalaženja njihovih sličnosti i različitosti. Kako su te sličnosti i različitosti podataka teško uočljive ukoliko su podaci multidimenzionalni, a to znači da ih je nemoguće grafički prikazati, PCA je moćan alat za tu svrhu. Glavne komponente ovog koncepta su vrlo važne u kemometriji.

Jedan od glavnih razloga za korištenje analize glavnih komponenata je u njegovoj ogromnoj količini podataka izvedenoj iz modernih računala i tehnikama mjerenja. Ta metoda je bila samo teorijska metoda, no pojavom generacije poboljšanih računala, razvila se i u praktičnom smislu te postaje dominantna metoda faktorske analize. Najveći razlog je u njezinoj matematičkoj točnosti.

Metodom glavnih komponenata Harolda Hostellinga dobiva se onoliki broj glavnih komponenata koliki je broj početnih varijabli. Glavne komponente računaju se tako da se najprije izračuna prva, zatim druga i tako dalje redom. Prva glavna komponenta računa se na potpunoj matrici varijable R i ona nam objašnjava najveću količinu varijance varijabli, gdje svaka varijabla ima varijancu 1,0. Upravo taj broj 1,0 u dijagonali matrice R označuje i ukupnu varinacu varijabli. Kada se izračuna prva, daljnjim računjanjem računamo drugu glavnu komponentu, ali na matrici R koja ima znatno niže koeficijente u dijagonali i izvan nje. Razlog je taj jer je dio zajedničke varijance varijabli upotrebljen za forimiranje prve glavne komponente.

Nakon računanja prve i druge glave komponente, slijedi računanje treče glavne komponente, a zatim i ostalih. Kada se potroši zajednička varijanca varijable, počinje se trošiti specifična varijanca i varijanca pogreške varijabli, poznata kao unikna varijanca. Nakon izračunate posljedne komponente, matrica R postaje nul matrica. Nul matrica predstavlja matricu koja u dijagonali i izvan dijagonale nula.

Glavne komponente se računaju rješavanjem složenog sustava linearnih jednadžbi. Pomoću vektora izračunavaju se glavne komponente. Normalan položaj vektora je vertikalan, odnosno kao kolona brojeva. Množenje reda vektora s kolonom vektora predstavlja produkt vektora koji kao rezultat daje jedan broj. Elementi vektora koeficijenti su glavnih komponenata.

Sustav jednadžbi može se prikazati i riješiti kao produkt matrice i vektora kao. Glavni aspekti analize glavnih komponenata je analiza linearne povezanosti većeg broja multivarijatno, kvantitativnih, međusobno koreliranih varijabli u smislu njihove kondenzacije u manji broj komponenti, novih varijabli, međusobno nekoreliranih, sa minimalnim gubitkom informacija. Ulazni podaci za analizu glavnih komponenata čine p varijable i n opažaja (individua) i imaju oblik matrice p x n.

Cilj analize je kreiranje p linearnih kombinacija izvornih varijabli koje se nazivaju glavne komponente:

$$\xi_1 = w_{11}X_1 + w_{12}X_2 + \ldots + w_{1p}X_p \qquad (4)$$

$$\xi_2 = w_{21}X_1 + w_{22}X_2 + \ldots + w_{2p}X_p \qquad (5)$$

$$\xi_p = w_{p1}X_1 + w_{p2}X_2 + \ldots + w_{pp}X_p \qquad (6)$$

gdje su:

$\xi_1, \xi_2 \ldots \xi_p$- p glavnih komponenata

w_{ij}- koeficijenti tj. konstante koje čine koeficijente j-te varijable za i-tu glavnu komponentu

Konstante w_{ij} procijenjene su tako da je:

1. prva glavna komponenta, ξ_1, objašnjava maksimum varijance iz podataka
2. druga glavna komponenta, ξ_2, objašnjava maksimum varijance koja je ostala neobjašnjena prvom i tako dalje.

$w_{i1}^2 + w_{i2}^2 + ... + w_{ip}^2 = 1 \quad i = 1 ... p$ (7)

$w_{i1}w_{j2} + w_{i2}w_{j2} + ... + w_{ip}w_{jp} = 0 \quad$ za sve $i \neq j$ (8)

Uvjet da zbroj kvadrata konstanti iznosi 1, zadan je zbog fiksiranja skale novih varijabli. U protivnom, moguće bi bilo povećati varijancu linearne kombinacije jednostavnom promjenom skale.

Konstante w_{ip} nazivaju se svojstveni vektori ili latentni vektori (eigenvectors) i geometrijski su, u dvodimenzionalnoj strukturi, u stvari, sinusi i cosinusi kuteva novih osi tj. glavnih komponenata. Transformirane vrijednosti izvornih varijabli predstavljaju rezultate glavnih komponenata (principal component scores). Suma varijanci svih izvornih varijabli je ukupna varijanca. Dio te ukupne varijance objašnjen jednom glavnom komponentom naziva se svojstvena vrijednost ili latentni korijen (eigenvalue). Svojstvena vrijednost je, w_{ij}, najveća u prvoj glavnoj komponenti i u svakoj sljedećoj njena je vrijednost sve manja. Suma svih svojstvenih vrijednosti jednaka je ukupnoj varijanci.

Cilj je, iteracijskim postupkom, izdvojiti čim veći dio ukupne varijance u tek nekoliko prvih glavnih komponenata, što se uobičajeno izražava u kumulativnim postocima ukupne varijance, i time reducirati broj izvornih varijabli.

Svojstvena vrijednost je zapravo varijanca izračunata iz seta rezultata glavne komponente što se može prikazati setom jednadžbi:

$$w_{11}x_1 + w_{12}x_2 + ... + w_{1p}x_p = \lambda x_1 \quad (9)$$

$$w_{21}x_1 + w_{22}x_2 + ... + w_{2p}x_p = \lambda x_p \quad (10)$$

ili u obliku matrice:

$$Wx = \lambda x \quad (11)$$

ili $$(W - \lambda I)x = 0 \quad (12)$$

gdje je:

I- jedinična matrica p x p sa vrijednosti jedan u dijagonali

0- je p x 1 nul-vektor

λ- svojstvene su vrijednosti matrice W

Ako se za i-tu svojstvenu vrijednost λ_i, postavi $x_1 = 1$, tada se rezultirajući vektor sa x vrijednosti zove i- ti svojstveni vektor matrice A, kao što je prikazan na slici 4.

$$\mathbf{x_i} = \begin{bmatrix} 1 \\ x_{2i} \\ x_{3i} \\ \vdots \\ x_{ni} \end{bmatrix}$$

Slika 4. Prikaz vektora matrice A

Osnovu za interpretaciju glavnih komponenata čine svojstveni vektori. Njihove vrijednosti su u prvoj glavnoj komponenti, najčešće, relativno ravnomjerno raspoređene po svim izvornim varijablama. U drugoj glavnoj komponenti dolazi do njihove veće disproporcije, što omogućava izdvajanje izvorne varijable (ili tek nekoliko njih) i pomaže u objašnjavanju i sažimanju ukupne varijabilnosti.

Problem određivanja broja glavnih komponenata koje treba odbaciti, rješava se pomoću kriterija. No, glavni problem predstavlja što različiti kriteriji daju različita rješenja, a to znači da proglašavaju značajnima različit broj glavnih komponenata. Najpoznatiji kriteriji su Guttman-Kaiserov kriterij i Cattellov scree-test. Da bi glavna komponenta bila značajna, treba imati varijancu koja je približno jednaka varijanci pojedine početne varijable.

Cattellov scree-test je grafička metoda gdje je posljednja značajna glavna komponenta ona nakon koje dolazi do naglog smanjenja u opadanju vrijednosti karakterističnih korijenova (λ_j). Glavni aspekti analize glavnih komponenata je analiza linearne povezanosti većeg broja multivarijatno distribuiranih, kvantitativnih, međusobno koreliranih varijabli u smislu njihove kondenzacije u manji broj komponenti, novih varijabli, međusobno nekoreliranih, sa minimalnim gubitkom informacija.

3.4. Kanonička korelacijska analiza

Kanoničkom korelacijskom analizom utvrđuje se povezanost dvaju skupova varijabli te kada postoji više kriterijskih varijabli. Jedna varijabla sačinjava nezavisne ili prediktorske, a druga zavisne ili kriterijske varijable. Prema tome, kanonička korelacijska analiza omogućava statističku analizu u istraživačkim projektima u kojima se objekt mjeri na dva niza varijabli, a istraživač želi znati u kakvom su odnosu ta dva niza. Dakle, kanonička korelacijska analiza je mjerilo jačine povezanosti između dva seta varijabli.

Prva faza kanoničko korelacijske analize započinje supermatricom koja se sastoji od četiri matrice kao što je prikazano na slici 5.

$$X = \begin{bmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{bmatrix}$$

Slika 5. *Početna supermatrica*

X_{01} je matrica interkorelacija varijabli u prvom skupu, X_{11} je matrica interkorelacija varijabli u drugom skupu, a matrice X_{01} i X_{10} sadrže kroskorelacije varijabli dva skupa. Te dvije matrice sadrže jednake koeficijente, te matrice X_{00} i X_{11} sadrže jednake koeficijente u odnosu na glavnu dijagonalu. Kanonička korelacija je maksimalna korelacija između para linearnih funkcija, gdje su te linearne finkcije određene svaka u jednom skupu varijabli.

Analiza kanoničke korelacije može se prikazati:

$$Y_1 + Y_2 + \ldots + Y_n = X_1 + X_2 + \ldots + X_n \qquad (13)$$

Početni korak u analizi je formiranje kanoničkih jednadžbi za dobivanje para novih kanoničkih varijabli:

$$W_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \qquad (14)$$

$$V_1 = b_{11}Y_1 + b_{12}Y_2 + \ldots + b_{1q}Y_q \qquad (15)$$

U kanoničku korelaciju ulaze dva sete varijabli, Y i X. Analiza pronalazi novu varijablu, V_1, kao linearnu kombinaciju iz sete Y varijabli, s jedne strane i novu varijablu W_1, kao linearnu kombinaciju iz seta X varijabli, s druge strane. Ako uzmemo da je R_1, korelacija između kanoničkih varijabli V_1 i W_1, tada je cilj kanoničke korelacije procijeniti kanoničke koeficijente ili opterećenja a_{11}, a_{12}, ... a_{1p} i b_{11}, b_{12}, ... b_{1p}, tako da korelacija između prvog para kanoničkih varijabli R_1, bude maksimalna.

U kanoničkoj korelacijskoj analizi izračunavaju se parovi linearnih funkcija varijabli uz uvijet da njihova povezanost bude maksimalna:

$$R = \max \qquad (16)$$

Ova korelacija između dviju kanoničkih varijabli je prva kanonička korelacija. Sljedeći korak je formiranje drugog para kanoničkih varijabli V_2 i W_2, nekoreliranog sa prvim parom koji daje drugi najveći korelacijski koeficijent. Proces konstruiranja kanoničkih varijabli nastavlja se dok se ne izjednači broj parova kanoničkih varijabli i broja varijabli u manjem setu (X_p ili Y_q). Kanoničke koeficijente je uobičajeno standardizirati tako da svaka kanonička varijabla ima varijancu 1.

Dakle, svaka kanonička varijabla je nekorelirana sa bilo kojom drugom kanoničkom varijablom ili varijablom iz ulaznih setova, osim sa jednom korespondirajućom kanoničkom varijablom iz suprotnog seta. Kanonički koeficijenti općenito nisu ortogonalni, tako da kanoničke varijable ne predstavljaju zajednički okomiti pravac kroz prostor izvornih varijabli.
Prva kanonička varijabla najmanje je toliko velika koliko i multipla korelacija bilo koje varijable u suprotnom setu varijabli. Moguće je da je prva kanonička korelacija vrlo velika, dok su sve multiple korelacije za predikciju jedne od izvornih varijabli iz suprotnog seta male.

Osnovni parametri u interpretaciji kanoničke analize su matrice kanoničkih struktura ili kanoničkih opterećenja i matrice unakrsnih kanoničkih opterećenja kao mjerila jednostavne linearne korelacije između izvornih varijabli i novostvorenih u analizi. Rezultat čine četiri matrice korelacija u kombinacijama X vs.W, Y vs. V, X vs.V i Y vs. W.

Kanonička opterećenja odražavaju dio varijance koju izvorna varijabla dijeli sa novom kanoničkom varijablom, a mogu se interpretirati kao faktorska opterećenja u smislu relativnog učešća svake varijable u svakoj kanoničkoj funkciji.

Prema tome, traži se najpodudarniji mogući položaj ispitanika. Kanonička korelacijska analiza je primijena metode glavnih komponenata na problem povezanosti dvaju skupova varijabli.

3.5. Diskriminacijska analiza

Diskriminacijska analiza je metoda koja omogućava da se utvrdi koje varijable prave razliku između dviju ili više prirodno formiranih grupa (objekata). Cilj analize je da se definira manji broj novih varijabli, koje bi opisale razlike među grupama. Te se nove varijable nazivaju diskriminacijskim varijablama.

Diskriminacijske varijable dobivaju se kao linearne kombinacije orginalnih varijabli, uz uvjet da te varijable maksimalno razdvajaju grupe. Interpretacija diskriminacijskih varijabli temelji se na odnosu (korelaciji) orginalnih i diskriminacijskih varijabli, tj. matrici strukture.

Zadaća diskriminacijske analize je:

- određivanje varijabli na temelju kojih istraživač može izvršiti diskriminaciju između različitih (prirodno formiranih) grupa i
- klasificiranje entiteta (objekata) u različite grupe s većom točnošću nego što je slučajna (nasumice) klasifikacija.

Diskriminacijska analiza dijeli se na deskriptivnu diskriminativnu analizu i prediktivnu diskriminativnu analizu.

Deskriptivna diskriminativna analiza je statistička tehnika koja omogučava identificiranje onih varijabla ili atributa koji najbolje odvajaju ili diskriminiraju članove dviju ili više grupa prema nekim glavnim obilježljima. (1)[8]
Prediktivna analiza omogućuje predviđanje budućeg statusa subjekata u grupi čiji je sadašnji status nepoznat. (2)[9]

Na primjer, imamo slučaj vezan uz medicinu i zamišljamo da doktor treba odabrati uzorak pacijenata koji se podvrće operaciji srca. Taj uzorak se naziva pokusni uzorak. O svakom pacijentu postoje informacije koje se odnose na dob, spol, tjelesnu težinu, krvnu grupu, krvni pritisak, itd. Isto tako zna se broj preživljelih i umrlih u tijeku jedne godine tijekom takve operacije. Dakle, grupe se dijele na preživljele i umrle.

[8] Milko Mejovšek, Uvod u metode znanstvenog istraživanja u društvenim i humanističkim znanostima, Naklada Slap, 2003., Zagreb, str.153.
[9] Aleksandar Halmi, Multivarijatna analiza u društvenim znanostima, Alinea, Zagreb, 2003., str.138.

U ovom slučaju primjenjuje se deskriptivna analiza kako bi istraživač odredio one varijable koje najbolje odvajaju jednu od druge grupe. Prema tome istraživač, tj. u našem slučaju doktor primjenjuje one informacije poput krvnog pritiska, dobi, tjelesne težine. To su varijable koje mogu diskriminirati dvije navedene grupe pacijenata. Tako se na temelju poznatih informacija (varijabli) predviđa hoće li pacijent preživjeti operaciju.

Diskriminacijska analiza koristi entitete kao što je na primjer pacijenti, obiteljski sustav, korisnici i sl. kako bi se kategoriziralo u grupe koje predstavljaju različite dijagnostičke kategorije, razne slučajeve, eksperimentalne grupe i sl. Postoje dvije grupe i svaki taj entitet pripada jednoj grupi tako da se ostale grupe isključe. Entiteti se opisuju pomoću niza varijabli. Svaka grupa treba biti dobro definirana tako da grupiranje održava prave razlike između entiteta.

Deskriptivna diskriminacijska analiza opisuje razlike između grupa na temelju kvalitativnih obilježja entiteta, dok prediktivna se odnosi na pitanje kako označiti entitete pripadajućih grupa na temelju informacija koje su sadržane u varijablama.

Deskriptivna i prediktivna diskriminacijska analiza imaju različite ciljeve, no kombiniranjem s ostalim multivarijantnim tehnikama dolazi se do pouzdanih informacija. Mogućnosti su razne, ali cilj je uvijek isti: opis, te klasifikacija varijabli u različite grupe i kategorije.

3.6. Klaster analiza

Pojam klaster analiza ili grupiranje podataka obuhvača metode koje su primarno korisne za pronalaženje i stvaranje vidljivih struktura unutar promatranih i danih podataka. Klaster analiza je zajednički naziv za skup različitih klasifikacijskih postupaka, koji se ne temelje na nekim određenim statističkim testovima. Za razliku od drugih statističkih metoda, klaster analiza se najčešće koristi u slučajevima kada još nemamo unaprijed (tj. u eksplorativnom dijelu istraživanja) definirane tvrdnje koje želimo testirati. Statistički testovi se ne koriste u klaster analizi.

Kako se klaster analiza i diskriminacijska analiza odnose na problem klasifikacije objekata ili ispitanika u kategorije, diskriminacijska analiza zahtjeva poznavanje grupne pripadnosti za jedinice koje koristimo za utvrđivanje klasifikacijskih pravila. Npr., ako se nastoji razlikovati ispitanike koji spadaju u 3 dijagnostičke kategorije, mora se poznavati grupna pripadnost za svakog ispitanika. Na osnovu karakteristika tih ispitanika sa poznatom grupnom pripadnošću, diskriminacijska analiza omogućuje definiranje pravila za klasifikaciju ispitanika za koje ne znamo grupnu pripadnost.

U klaster analizi grupna pripadnost objekata nije poznata, kao ni konačni broj grupa. Cilj klaster analize jest utvrđivanje homogenih grupa ili klastera. Termin klaster dolazi od engl. riječi cluster (skupina "istovrsnih stvari", grozd, skupiti u hrpu). Prema tome, klaster analiza nastoji grupirati objekte prema nekim varijablama.

U tablici 1. objašnjava se primjena klaster analize s obzirom na sadržaj ulazne matrice koja može predstavljati slučaj kada su nam u kolonama varijable (varijable u tom slučaju predstavljaju objekte ili entitete) čije grupiranje nastojimo ispitati, dok su nam u redovima ispitanici koji su procjenjivali svaki od objekata prema nekoj karakteristici i koji nam služe kao izvor informacije o sličnosti ili različitosti varijabli.

	Varijable čije grupiranje želimo utvrditi			
Ispitanici	**VAR_1**	**VAR_2**	**VAR_3**	**VAR_k**
Ispitanik 1	X_{11}	X_{12}	X_{13}	X_{1k}
Ispitanik 2	X_{21}	X_{22}	X_{23}	X_{2k}
Ispiatnik 3	X_{31}	X_{32}	X_{33}	X_{3k}
...				
Ispitanik N	X_{N1}	X_{N2}	X_{N3}	X_{Nk}

Tablica 1. *Prikaz grupiranja varijabli koje predstavljaju objekte prema procjenama ispitanika*

Rezultat klaster analize uvijek predstavlja klasifikaciju objekata u neke grupe, što ovisno o korištenoj tehnici može dovesti do različitih rješenja.

3.7. Validacija

Metoda validacije u analitičkoj kemiji je posljednji korak u razvoju metode čiji postupak se provodi kako bi se osigurala kvaliteta metode. To je, dakle, bitan korak osiguravanja kvaliteta programa u laboratoriju. Osiguravanje kvaliteta općenito definirano kao: „ Sustav djelatnosti čija je svrha osigurati da proizvođač ili korisnik proizvoda ispunjava utvrđene standarde kvalitete s razinom povjerenja.“

Kemijska analiza također se može smatrati kao vrstom usluge, koja zadovoljava definirane standarde kvalitete, između ostaloga, analitičar koji treba definirati karakteristike te metode te dovodi do sljedeće definicije: "Postupak validacije sastoji se od dokumentiranja kvalitete analitičkih postupaka, uspostavljanjem odgovarajućih uvjeta za kriterije uspješnosti, kao što su točnost, preciznost, granice detekcije, itd., te mjerenjem vrijednosti tih kriterija."

Regulatorna tijela zahtijevaju da se detalji analitičkih postupaka koriste kao standardni operativni postupci, te zahtijevaju dokaz da se zaista provela validacija tih metoda koristeći validirane dokumente kao osiguranje kvalitete. Postoje tri vrlo važna pravila kod metode validacije:

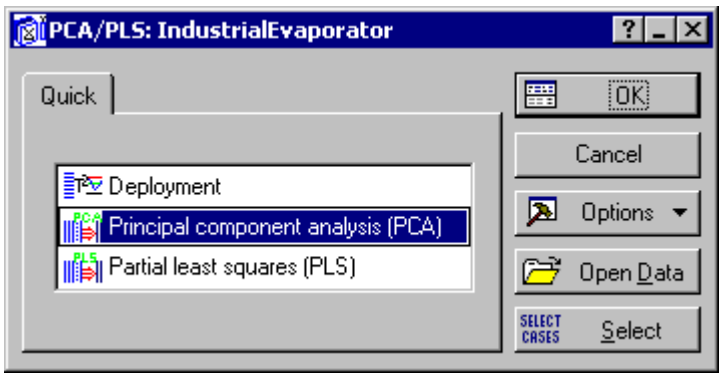
1. Validacija cijele metode
2. Validacija u cijelom rasponu koncentracija
3. Validacija preko cijele matrice

4. PRIMJENA

Primjer eksploracijsko-multivarijatne analize podataka baziran je na modelu PCA koji je preuzet iz programskog paketa Statistica.

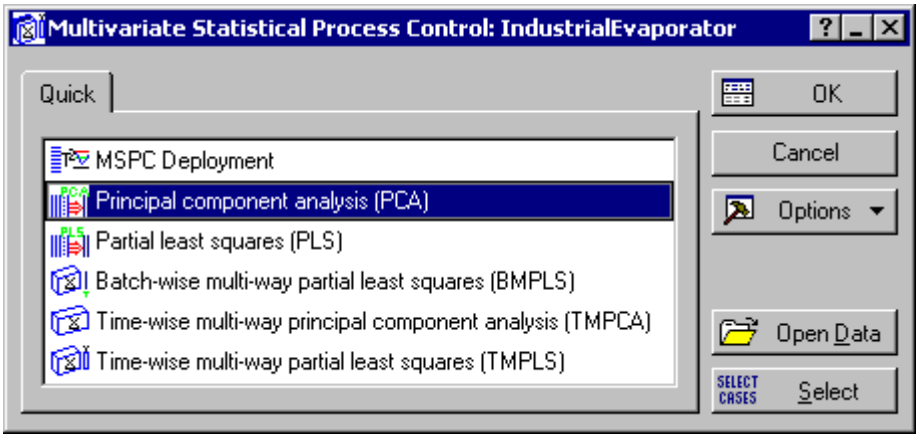
Podaci iz datoteke koriste se za demonstraciju primjene analize glavnih komponenata koji sadrže detalje o procesu sušenja broja mokrih produkata postavljenih u krevetu za isparavanje. Tijekom procesa isparavanja mjeri se 8 varijabla: temperatura kondenzacije, temperatura unosa, temperatura zraka u procesu, temperatura isparavanja, protok zraka, temperatura kreveta, tlak filtera i tlak kreveta. Varijable se mjere unutar jednakih vremenskih intervala radi nadziranja i kontrole kvalitete procesa isparavanja. Eksperiment se provodi radi otkrivanja abnormalnih uvjeta ako bi do njih došlo, te kako bi se osigurala kvaliteta krajnjeg proizvoda.

Iz izbornika potrebno je izabrati opciju *NIPALS Algorithm (PCA/PLS)* radi pristupa PCA modelu. Nakon odabira opcije, na ekranu će se pojaviti *PCA/PLS Startup Panel.*



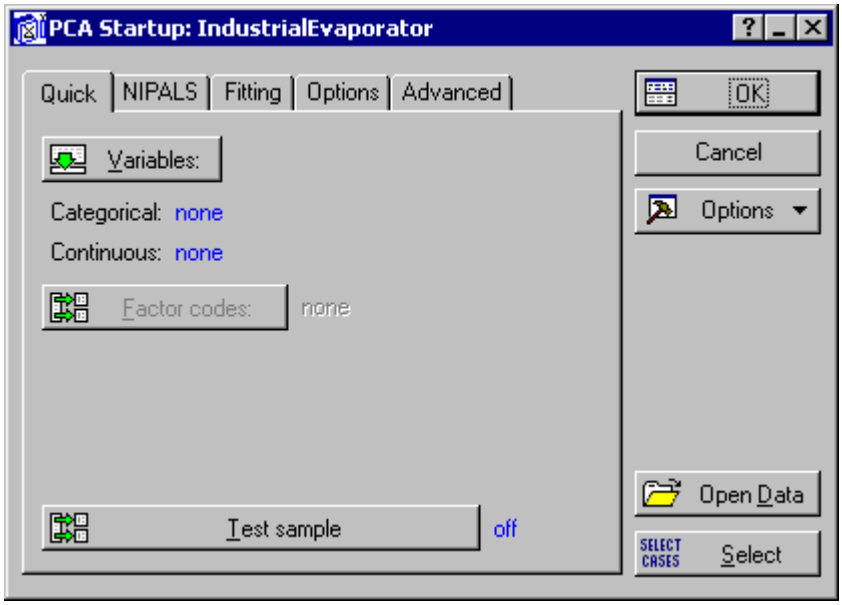
Slika 6. *PCA/PLS Startup Panel.*

Također je moguće izabrati i opciju *PLS, PCA Multivariate/Batch SPC* iz izbornika *Statistics* kako bi se prikazali multivarijatne analize podataka.



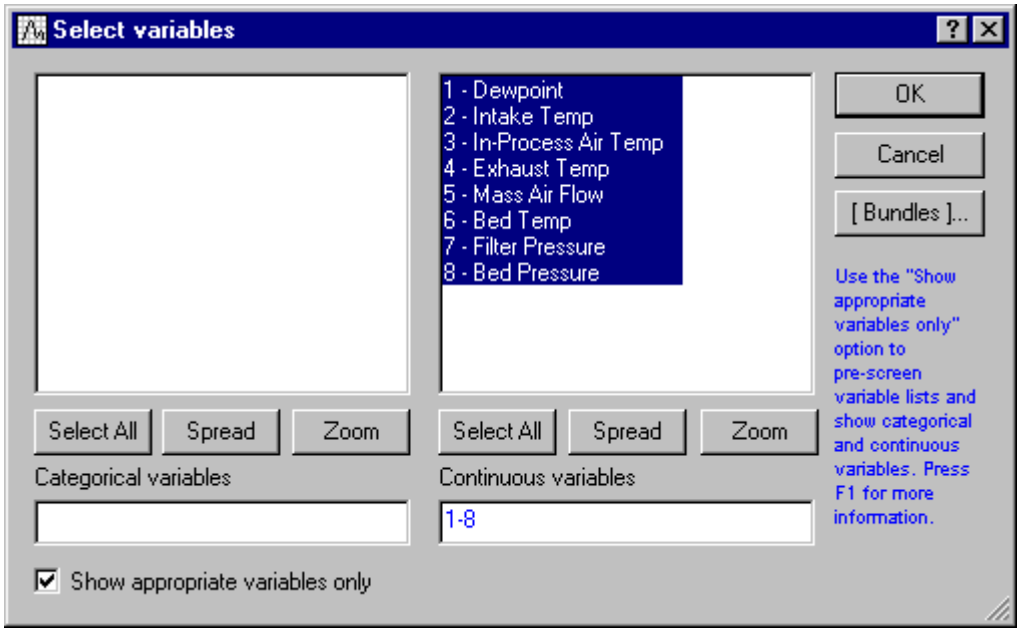
Slika 7. *Multivariate Statistical Process Control Startup Panel*

Unutar *Quick* tab-a bilo kojeg startup panel-a, potrebno je izabrati opciju *Principal component analysis (PCA)* i pritisnuti tipku *OK* kako bi se prikazao *PCA Startup* prozor.



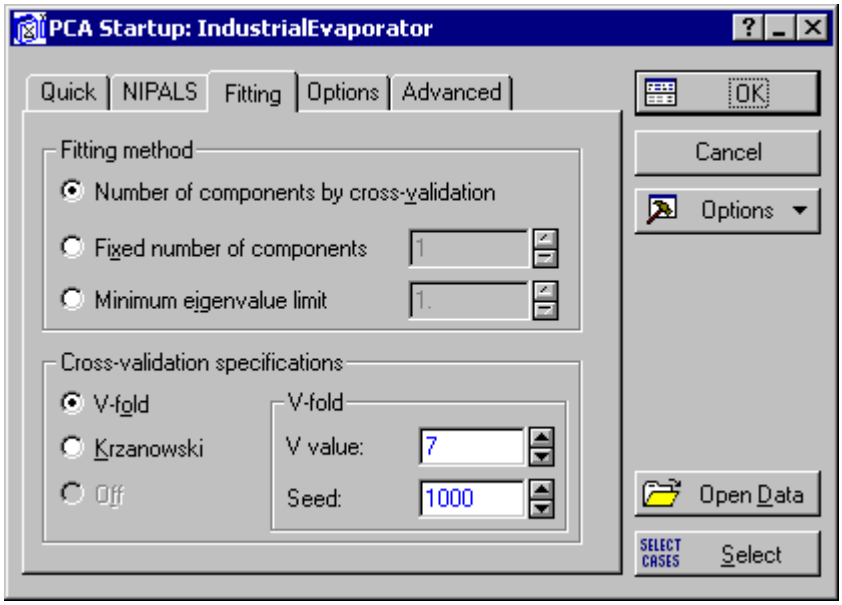
Slika 8. *PCA Startup*

Daljnim odabirom na tipku *Variables* potrebno je označiti svih osam varijabli kao trajne varijable za PCA analizu.



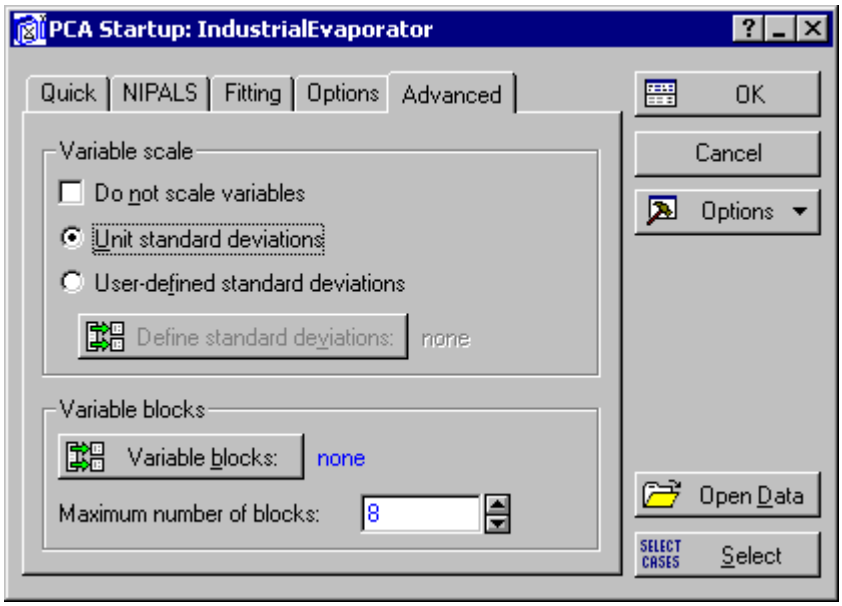
Slika 9. *Odabir varijabli*

Kako bi se zatvorio odabir varijabli, te izvršio povratak na PCA Startup prozor, potrebno je pritisnuti tipku *OK*. Unutar taba *Fitting*, moguće je odabrati metode za određivanje broja komponenti u PC modelu. Broj osnovnih komponenti određuje koliko će model biti kompleksan. Što više osnovnih komponenti model ima, bolje može uklopiti podatke za probu. Stoga, važno je pažljivo izabrati broj komponenti od PC analize. Moguće je ili prepustiti da metodom kružne validacije (eng. *cross-validation)* odredi taj faktor, ili je potrebno koristiti opciju *Fixed number of components* kako bi se ručno postavila kompleksnost modela.



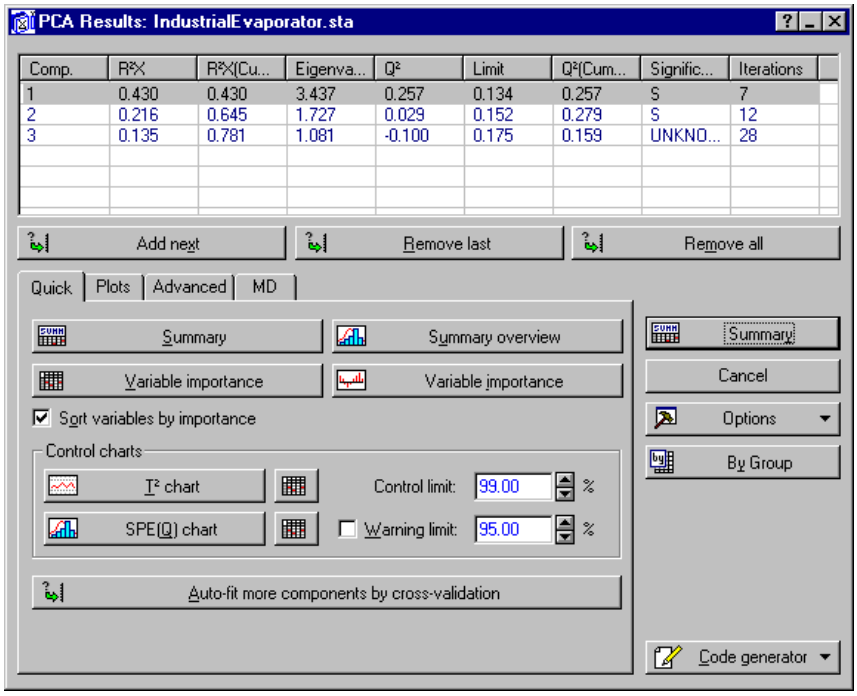
Slika 10. *PCA Fitting tab*

Jedna od bitnijih značajki STATISTICA PCA modela je njegova funkcionalnost obrade, koja omogućava skaliranje podataka radi bolje izrade modela. Podatke skaliramo kada ih želimo svesti na pozitivne vrijednosti (bez uključene 0). Zadana postavka je *Unit standard deviations* (ova opcija se nalazi unutar *Advanced* tab-a), što bi trebalo biti prikladno za većinu aplikacija. Međutim, ako to nije slučaj, moguće je postaviti vlastite fakore za skaliranje za individualne varijable odabirom opcije *User-defined standard deviations* i potom klikom na gumb *Define standard deviations* kako bi se prikazao *User-defined scale (standard deviation)* prozor gdje je moguće odrediti skalu. Na ovom primjeru će zadana opcija (*Unit standard deviations)* biti dovoljna.



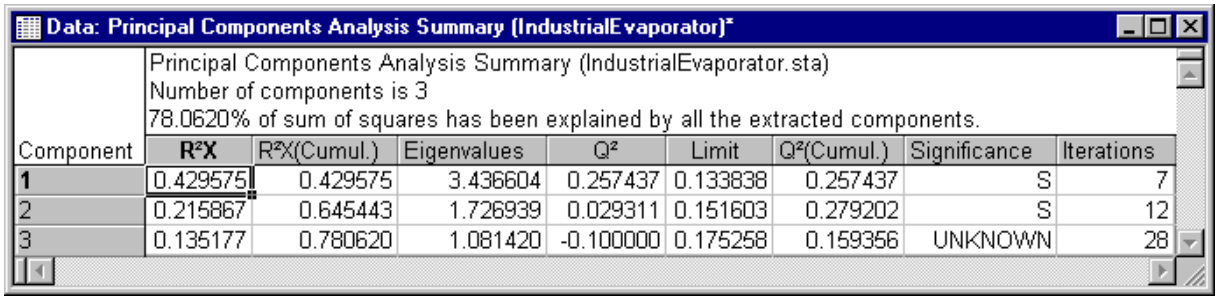
Slika 11. *PCA Advanced tab*

Spomenute postavke ne određuju samo rezultat, nego također i kvalitetu modela, tj. njegovu sposobnost predviđanja neviđenih primjera i uočavanja važnih svojstava koja mogu biti prisutna u podacima kao što su abnormalnosti. Uočavanje abnormalnosti jedan je od primarnih ciljeva procesa praćenja i kontrole kvalitete. Bitno je napomenuti da je svaka analiza jednistvena i potrebno je vrlo pažljivo konfigurirati njezine postavke. Po završetku ovog koraka, prikazati će se prozor *PCA Results.*



Slika 12. *PCA rezultati 1*

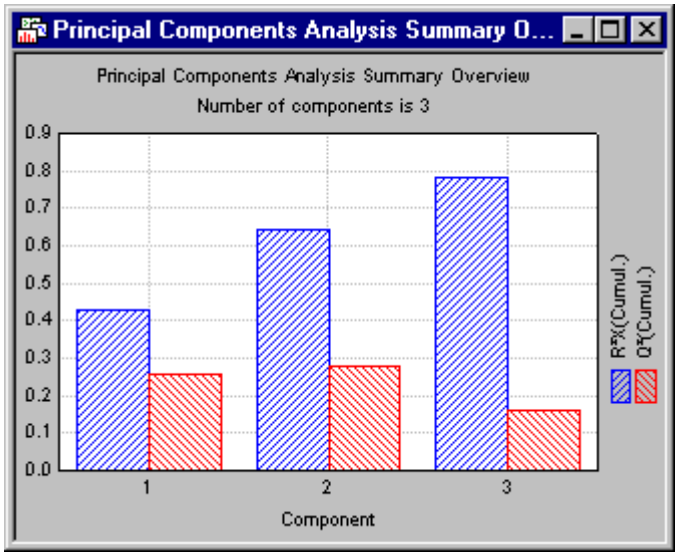
Okvir sa sažetkom nalazi se na vrhu prozora *Results* i sadrži informacije o PC modelu kao što su R^2X, svojstvene vrijednosti, Q^2 , limit, značaj i broj iteracija za svaku komponentu. Iste informacije se mogu prikazati i u proračunskoj tablici klikom na gumb *Summary* unutar *Quick* tab-a.



Component	R²X	R²X(Cumul.)	Eigenvalues	Q²	Limit	Q²(Cumul.)	Significance	Iterations
1	0.429575	0.429575	3.436604	0.257437	0.133838	0.257437	S	7
2	0.215867	0.645443	1.726939	0.029311	0.151603	0.279202	S	12
3	0.135177	0.780620	1.081420	-0.100000	0.175258	0.159356	UNKNOWN	28

Slika 13. *Proračunska tablica komponenti*

Također je moguće generirati i histogram grafičkih podataka od R^2X i Q^2 klikom na gumb *Summary overview* unutar *Quick* tab-a.



Slika 14. *Rezultati u histogramu*

Iz grafa se može vidjeti da se kumulativna R^2X poboljšava, tj. ima sklonost postati cjelinom, budući da je sve više i više komponenti dodavano u PC model. Treba istaknuti da je moguće dodavati i uklanjati jednu ili sve komponente iz PCA modela koristeći gumbe *Add next*, *Remove last* i *Remove all*. Malo dalje u ovom primjeru, koristit će se upravo te opcije kako bi se postepeno pregledavali PCA modeli sa drukčijim brojem komponenti za isti komplet podataka i u jednoj analizi koristeći samo opcije unutar prozora *Results*.

U ovom konkretnom primjeru korištena je kružna validacijska (eng. cross validation) metoda za određivanje optimalnog broja osnovnih komponenti (tj. kompleksnosti modela), koji je u ovom primjeru ispao 3. To znači da je, u ovom slučaju, kružno validacijski algoritam pronašao PC model sa 3 komponente koje će najbolje predstavljati komplet podataka.

Važnost varijable (eng. Variable Importance) je korisno svojstvo u PCA analizi. Ono mjeri koliko dobro je varijabla zastupljena od strane osnovnih komponenti. Također je poznato kao snaga čija količina može biti između 0 i 1. Za model sa dovoljnim brojem komponenata, varijable koje nisu dobro zastupljene (tj. imaju slabu vrijednost snage) će najvjerojatnije biti beznačajne. Kako bi se varijable prikazale poredane silazno u proračunskoj tablici, potrebno je izabrati check box *Sort variables by importance*. Klikom na gumb *Variable importance* generira se proračunska tablica s prikazom važnosti varijabli.

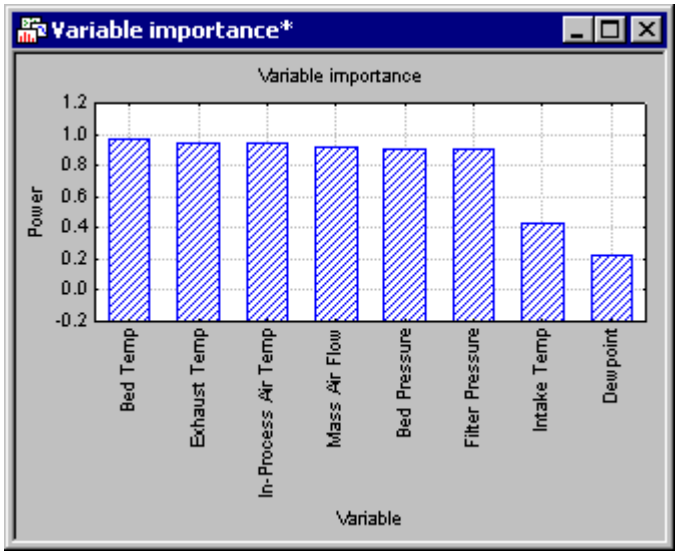
Data: Variable importance (IndustrialEvaporator)

Variable	Variable importance (IndustrialEvaporator) Number of components is 3		
	Variable number	Power	Importance
Bed Temp	6	0.971146	1
Exhaust Temp	4	0.946438	2
In-Process Air Temp	3	0.938109	3
Mass Air Flow	5	0.919734	4
Bed Pressure	8	0.910761	5
Filter Pressure	7	0.906128	6
Intake Temp	2	0.434358	7
Dewpoint	1	0.218288	8

Slika 15. *Proračunska tablica s prikazom važnosti varijabli*

Također, moguće je pregledavati snagu varijabli u histogram formatu klikom na gumb *Variable importance* s ikonom grafa.

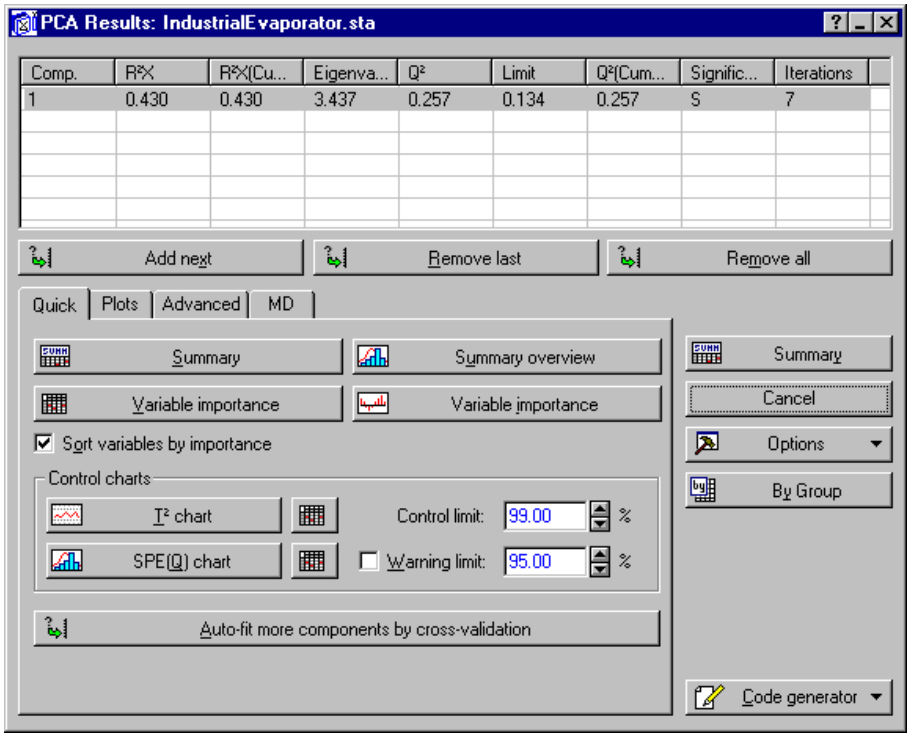
Slika 16. *Gumb Variable importance*



Slika 17. *Histogram važnosti varijable 1*

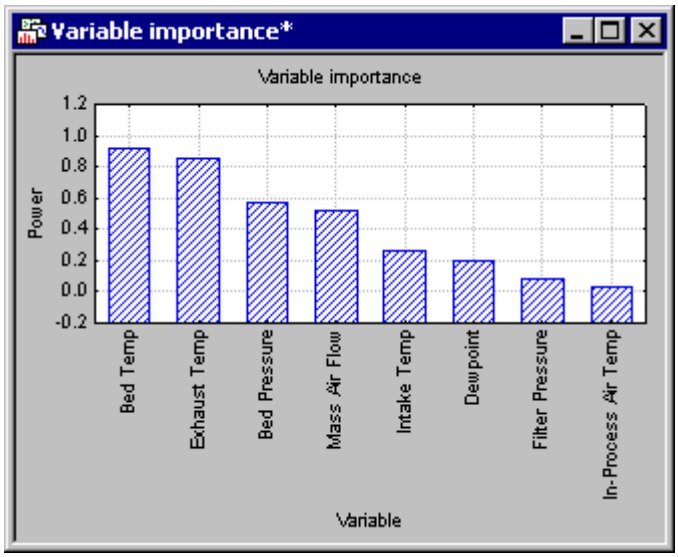
Moguće je ručno dodavati i uklanjati komponente iz PC modela koristeći gumbe *Add next*, *Remove last* i *Remove all* koji se nalaze unutar prozora *Results*. Ova funkcionalnost se može koristiti za praćenje promjena važnosti varijable koja ima povećan broj osnovnih komponenti.

Kako bi se to napravilo, potrebno je prvo kliknuti gumb *Remove all* kako bi se uklonile sve komponente, te potom na gumb *Add next* kako bi se dodala prva osnovna komponenta.



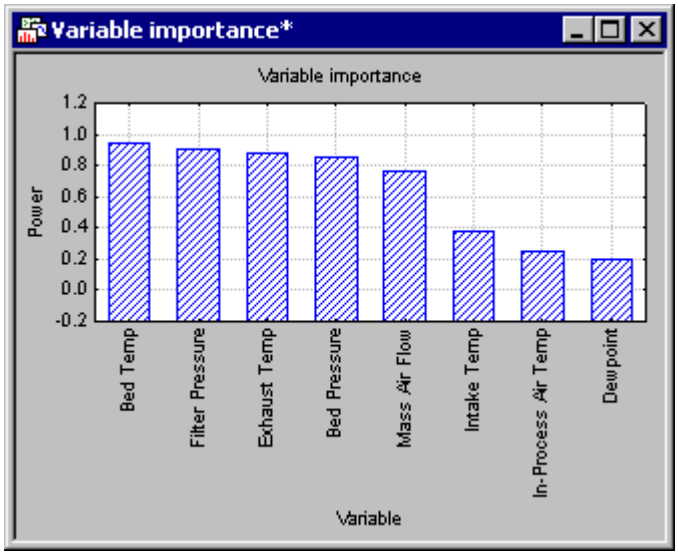
Slika 18. *PCA rezultati 2*

Klikom na gumb *Variable importance* generira se histogram važnosti varijable.



Slika 19. *Histogram važnosti varijable 2*

Graf pokazuje da za ovaj pojednostavljeni model tj., model koji nema dovoljan broj osnovnih komponenti, većina varijabli se čini beznačajnim. Razlog tomu je što model nema dovoljno komponenti kako bi efikasno modelirao varijable prema njihovom pravom značaju. Stoga je važno dodavati sve više komponenata modelu i izrađivati pripadajući histogram važnosti svaki put kada se modelu dodaje dimenzija.



Slika 20. *Histogram važnosti varijable 3*

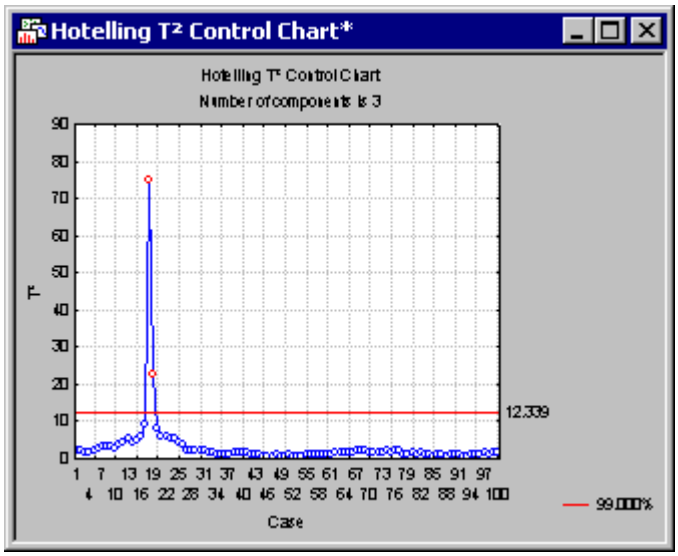
Sada je potrebno proučiti niz histograma koji su generirani. Prva stvar koju treba primjetiti je da što više komponenti model ima, to je veća snaga zasebne varijable. Posebno je bitno uočiti da varijable *Exhaust Temp* i *Bed Temp* su većinom modelirane od strane PC1, dok je varijabla *Filter Pressure* skoro isključivo modelirana od PC2. To sugerira da individualne komponente modeliraju različite individualne varijable (uz pretpostavku da su značajne).

Potrebno je ponovno ukloniti sve izvađene komponente i potom kliknuti na gumb *Auto-fit more components by cross-validation* unutar *Quick* tab-a. To će ponovno stvoriti početni PCA model koji je izrađen klikom na gumb *OK* unutar prozora *PCA Startup.* Drugim riječima, vrši se povratak na tu fazu analize prije nego što su ručno uklanjane i dodavane komponente iz modela i u model.

Dosad je analiziran PC model kako bi se proučile varijable. Drugim riječima, PC model je korišten za dijagnostiku varijabli pregledavanjem njihova značaja.

PC model također može pomoći i pri analizi podataka na temelju slučajeva generiranjem kontrolnih grafova koji se mogu koristiti za pregled i uočavanje abnormalnosti slučajeva. Ova značajka se može koristiti za uočavanje abnormalnosti.

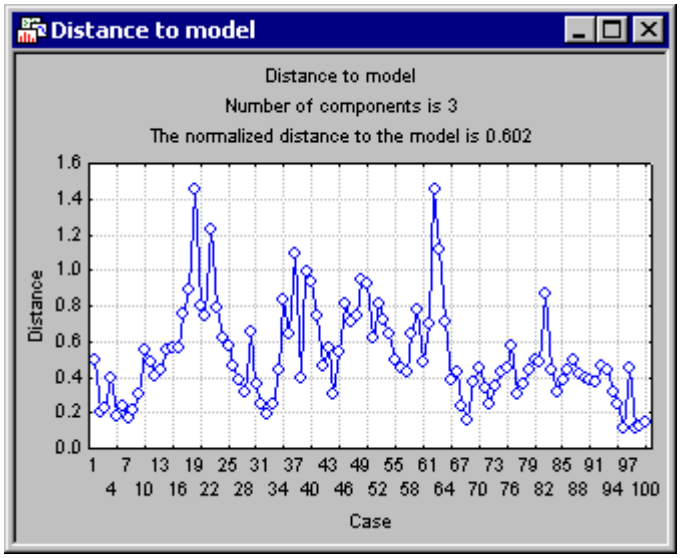
Za tu svrhu kontrole kvalitete, outlieri mogu biti naznaka abnormalnih radnih uvjeta koji mogu utjecati na kvalitetu završnog proizvoda i stoga bi trebali biti razlog za brigu.
Važan graf koji je potrebno pregledati je tzv. *Hotelling* T^2 koji se može koristiti za uočavanje umjerenih outliera. Graf se stvara klikom na gumb T^2 *chart* koji se nalazi unutar prozora *Results* pod tab-om *Quick.*



Slika 21. *Hotelling graf*

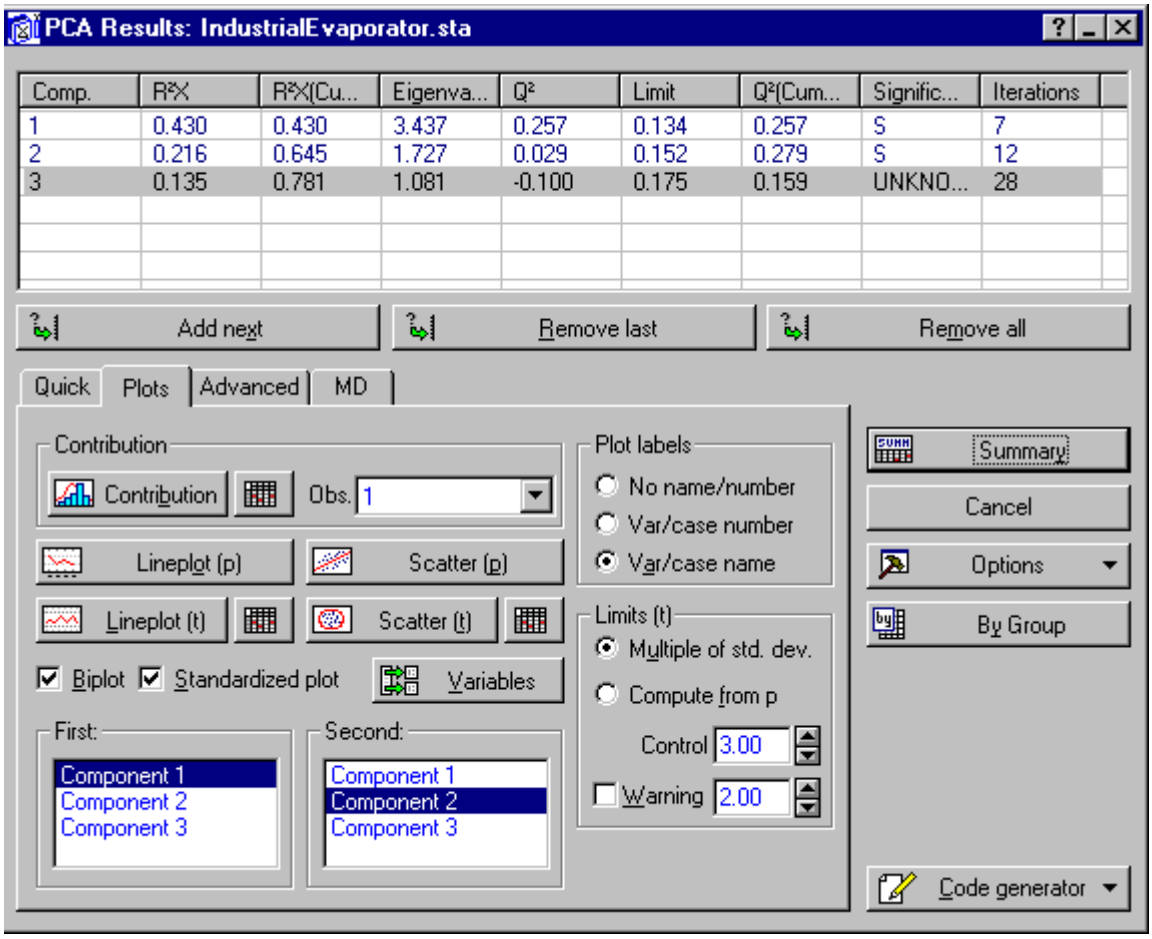
U slučaju ove analize, može se vidjeti da slučaj 18 sadrži posebno veliku vrijednost T^2 u usporedbi s ostalim zapažanjima. Slučaj 19 se također čini kao outlier, iako ne toliko ozbiljan. Stoga, možemo zaključiti da je proces isparavanja u vremenskim intervalima 18 i 19, izašao iz područja normalnosti. Međutim, proces se vratio u normalu nakon prolaženja kroz ova dva vremenska intervala, kao što vrijednosti T^2 pokazuju do kraja promatranja.

Sljedeći graf koji se može koristiti za uočavanje outliera je distance-to-model. Nalazi se unutar *Advanced* tab-a gdje je moguće kliknuti na bilo koji od tipki *D-To-Model* kako bi se generirao u obliku proračunske tablice, grafičkih podataka ili u histogram formatima (na slici su prikazani grafički podaci).



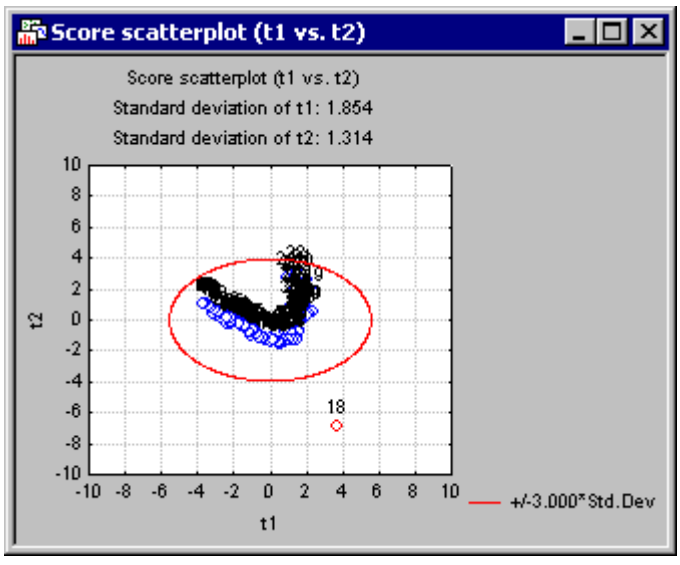
Slika 22. *Distance to model graf*

Daljnja dijagnostika slučajeva se može izvršiti koristeći scatterplot što znači raspoređivanje podataka od x-rezultata (x-scores). X-rezultati su promijenjene vrijednosti od X promatranja u sustavu osnovne komponente. X-rezultati koji imaju previsoku vrijednost (tj. onu koja previše odstupa od točke stvaranja) mogu ponovno biti smatrani outlierima ili abnormalnima. Kako bi generirali scatterplot od x-rezultata, treba odabrati tab *Plots*. U ovom primjeru su unutar komponentnih lista *First* i *Second*, odabrane opcije *Component 1* odnosno *Component 2*. U grupnom box-u *Plot labels*, potrebno je odabrati gumb *Var/case name* (kako bi se prikazala imena varijabli u scatterplot-u).



Slika 23. *PCA rezultati 2*

Nakon toga treba odznačiti check box *Biplot* i kliknuti na gumb *Scatter (t)* kako bi se kreirali raspršeni podaci od x-rezultata za *PC1* nasuprot x-rezultata od *PC2*.

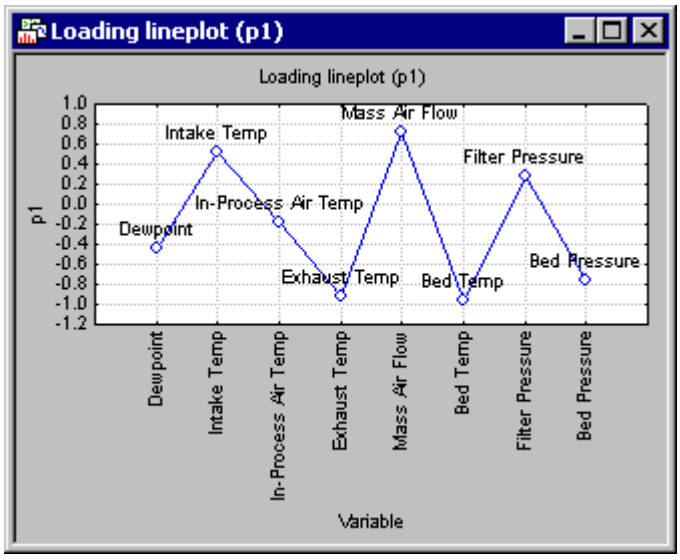


Slika 24. *Scatterplot graf 1*

Primjetno je da je slučaj 18 ponovno prikazan kao outlier budući da izlazi iz elipse normalnosti (koja je definirana u opcijama koje se nalaze u grupnom box-u *Limits*).

PCA također može pomoći pri analizi odnosa između orginalnih varijabli, načinu na koji imaju uzajmnu vezu i njihovom utjecaju pri određivanju novog koordinatnog sustava. U središtu takvih analiza su faktori x-opterećenja (eng. x-loading). X-opterećenja osnovne komponente sa posebnom točkom od varijable je kosinus kuta između smjera te komponente i osi od varijable. To podrazumijeva da što je varijabla značajnija u određivanju komponente, to više se varijablina os poravnava sa tom komponentom.

Za sljedeći korak, treba generirati grafičke podatke (eng. line plot) od x-opterećenja za *PC1*. Bitno je provjeriti da je opcija *Component 1* još uvijek odabrana unutar liste komponenata. Nakon provjere, treba kliknuti na gumb *Lineplot (p)* kako bi se stvorili grafički podaci (line plot) od varijabli nasuprot opterećenjima prve komponente.

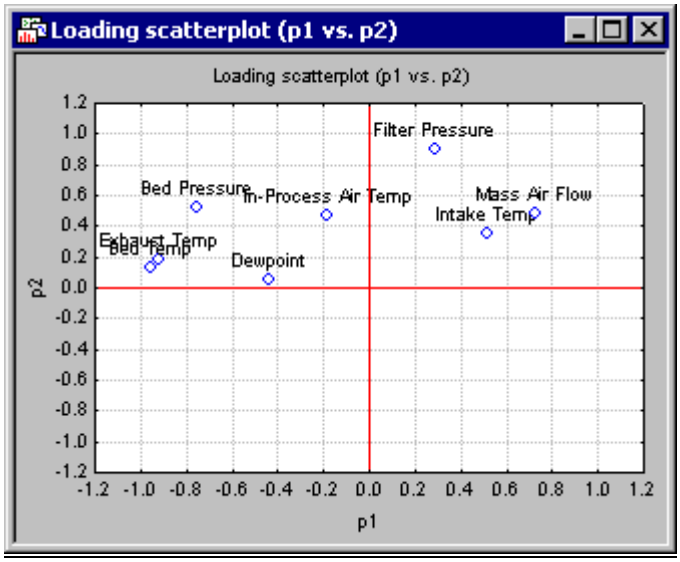


Slika 25. *Lineplot graf*

Proučavanjem grafičkih podataka možemo vidjeti da je varijabla *In-Process Air Temp* najmanje utjecajna u određivanju prve osnovne komponente, dok varijabla *Bed Temp* ima najznačajniju ulogu. Ovaj zaključak je potvrđen u proračunskoj tablici i histogramu varijable važnosti (gumb *Variable importance* unutar *Quick* tab-a) od PCA modela sa jednom osnovnom komponentom, što pokazuje da navedene varijable imaju modelirajuću snagu od 0,033526 i 0,921825.

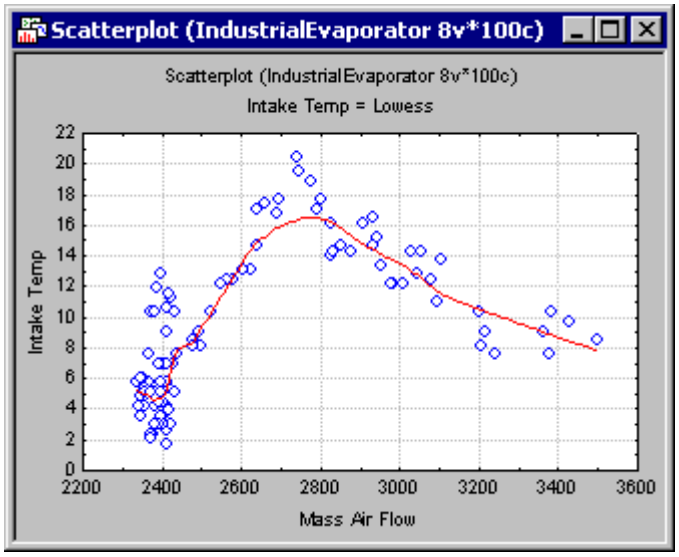
Moguće je napraviti isti graf za bilo koju osnovnu komponentu odabirom prikladne komponente unutar komponentne liste *First* koja se nalazi u *Plots* tab-u.

Nadalje, potrebno je koristiti scatterplot-ove od faktora opterećenja između raznih osnovnih komponenti kako bi se analizirao odnos između varijabli i identificirale najutjecajnije varijable u određivanju PCA modela. Nakon provjere jesi li opcije *Components 1* i *Components 2* odabrane u listama *First* i *Second* unutar taba *Plots*, potrebno je kliknuti na gumb *Scatter (p)* kako bi se stvorio scatterplot od faktora opterećenja.



Slika 26. *Scatterplot graf 2*

Graf pokazuje primjetnu količinu među varijablama. Varijable smještene u međusobnoj blizini utječu na PCA model na slične načine, što je također pokazatelj da su u uzajmnoj vezi. Varijable *Mass Air Flow* i *Intake Temp* su primjeri takvih varijabli sa znatnim stupnjem uzajamnosti. U stvari, scatterplot ovih dviju varijabli (koji se može stvoriti odabirom opcije *Scatterplots* iz izbornika *STATISTICA Graphs*) pokazuje nelinearni trend između njih.



Slika 27. *Scatterplot graf 3*

Druga korisna informacija u scatterplotu je odaljenost njegovih točaka od izvorišta. Što dalje se varijabla nalazi od izvorišta, to je više utjecajnija u određivanju PCA modela.

Kao što je navedeno u ovom primjeru, cilj PCA procedure je modelirati viševarijabilne strukture podataka uz pomoć novog koordinatnog sustava, poznatog kao osnovne komponente koji je dimenzijski manji od sustava s izvornim varijablama. To znači da se s određenim brojem osnovnih komponenti, mogu predvidjeti izvorne strukture podataka sa stupnjem točnosti koji se s vremenom povećava, barem u osnovi, dodavanjem sve više komponenata u model. Međutim, budući da je cilj ovdje modelirati izvornu strukturu podataka u manjim dimenzijama, uvijek postoji razlika (koja je vidljiva u ostacima) između izvornih promatranja i predviđanja PC modela. Moguće je napraviti proračunsku tablicu od ostataka klikom na gumb *Residuals* unutar *Advanced* tab-a.

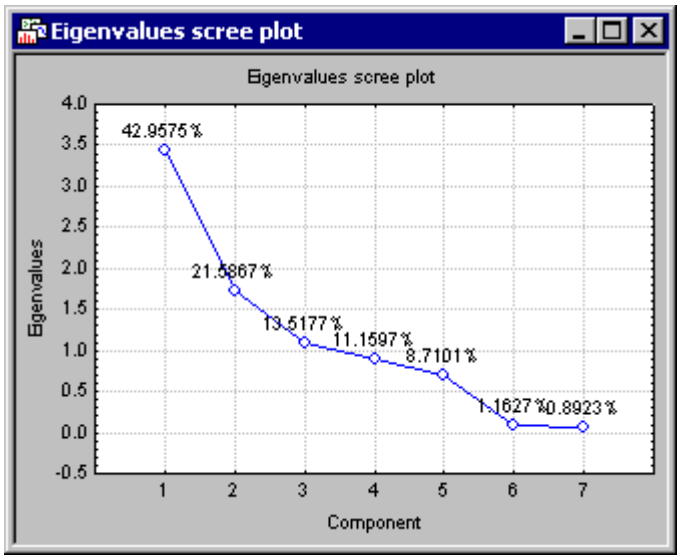
Data: Principal Component Analysis Residuals (IndustrialEvaporator)

Principal Component Analysis Residuals (IndustrialEvaporator)
Number of components is 3

Case	Dewpoint	Intake Temp	In-Process Air Temp	Exhaust Temp	Mass Air Flow	Bec
1	-0.89857	-0.32972	0.13691	-0.124849	-0.049114	-0
2	-0.29518	-0.27440	0.00614	-0.050168	0.019461	-0
3	0.01022	-0.29619	-0.03765	0.151685	0.157219	0
4	-0.34381	-0.35799	-0.00081	0.319325	0.121578	0
5	0.16354	-0.07906	-0.04839	0.144620	0.078592	0
6	0.38983	-0.09230	-0.08643	0.106194	0.124849	0
7	0.29826	0.13274	0.03611	0.075700	0.116920	0
8	0.27861	0.28341	0.04073	0.131010	0.032502	0
9	0.30362	0.49681	0.10081	0.125892	0.010387	0

Slika 28. *Proračunska tablica slučajeva*

Jedna od ključnih vrijednosti PCA procedure su svojstvene vrijednosti osnovnih komponenti, iz kojih skoro sva svojstva PCA modela mogu biti izvedena. Kako bi se generirali grafički podaci (line plot) od osnovnih svojstvenih vrijednosti, prvo je potrebno odabrati broj najutjecajnijih komponenti koje se žele prikazati u grafu prilagođavanjem vrijednosti opcije *Number of eigenvalues* unutar *Advanced* tab-a. Vrijednost je potrebno podesiti na 7 (maksimalan broj komponenti koje trenutačni model može imati „number of variables - 1“). Potom treba kliknuti na gumb *Scree plot.*



Slika 29. *Graf svojstvenih vrijednosti*

Prva osnovna svojstvena vrijednost bilježi 42,9575 % varijabilnosti unutar podataka. Međutim, ovaj trend se smanjuje dodavanjem još komponenata u model.

Kada je analiza osnovnih komponenti završena, često je potrebno pospremiti model kako bi se kasnije mogao koristiti za razvoj. Koristeći *STATISTICA PCA* moguće je spremiti PC modele u raznim formatima uključujući C/C++, STATISTICA Visual Basic i PMML (Predictive Markup Model Language). Kako bi se model spremio, potrebno je izabrati jedan od jezika iz izbornika *Code generator* koji se nalazi unutar prozora *PCA Results*. Za ovaj primjer izabran je PMML jezik. To će omogućiti modelov PMML kod u izvješću od STATISTICA-e. Iz izbornika *File* potom treba odabrati *Save As* kako bi se prikazao *Save As* prozor, koji se koristi za spremanje output-a na lokaciji po izboru korisnika s ekstenzijom XML. Sada je model spreman za razvoj.

5. ZAKLJUČAK

Znanstveno istraživanje je vrlo složen i dugotrajan proces koji se sastoji od prikupljanja i obrade podataka. Interpretiranje rezultata i izvođenje zaključaka na temelju postavljene hipoteze, teorijske su faze istraživačkog procesa. Svako znanstveno istraživanje započinje i završava teorijom. Svrha ovog rada bila je upoznati se s eksploracijskom analizom podataka, te primjenom na istraživačko multivarijatne modele. Istraživačka analiza podataka je proteklih godina znatno dobila na značaju. Upravo taj značaj možemo pripisati njezinim multivarijatnim modelima koji uz pomoć modernih računala, te još boljih i sofisticiranijih softvera mogu pomoći što lakšem i boljem odabiru bitnih varijabli.

6. LITERATURA

[1] J. W. Einax, H. W. Zwanziger, S. Geiß, Chemometrics in enviromental analysis, str.19.

[2] J. W. Einax, H. W. Zwanziger, S. Geiß, Chemometrics in enviromental analysis, str.20.

[3] Claus A. Andersson, M. Sc., Chew Eng, Exploratory Multivariate Data Analysis with Applications in Food Technology, str.5.

[4] Claus A. Andersson, M. Sc., Chew Eng, Exploratory Multivariate Data Analysis with Applications in Food Technology, str.5.

[5] J. W. Einax, H. W. Zwanziger, S. Geiß, Chemometrics in enviromental analysis, str.101.

[6] Aleksandar Halmi, Temelji kvantitativne analize u društvenim znanostima, Alinea, Zagreb, 1999., str.101.

[7] Milko Mejovšek, Uvod u metode znanstvenog istraživanja u društvenim i humanističkim znanostima, Naklada Slap, 2003., Zagreb, str.153.

[8] Milko Mejovšek, Uvod u metode znanstvenog istraživanja u društvenim i humanističkim znanostima, Naklada Slap, 2003., Zagreb, str.153.

[9] Aleksandar Halmi, Multivarijatna analiza u društvenim znanostima, Alinea, Zagreb, 2003., str.138.

[10] Elsevier, Handbook of Qualimetrics Part A

[11] Elsevier, Handbook of Qualimetrics Part B

[12] Kemometrija predavanje, Tomislav Bolanča

[13] http://www.agr.unizg.hr/multimedia/pdf/ds1905_metode_mva_osnove_2006.pdf, (pristup 7. kolovoza 2015.)

7. PRILOZI

- Korišten je programski paket Statistica pomoću kojega se izvršila istraživačko-multivarijatna analiza podataka na primjeru baziranome na PCA modelu koji je preuzet iz samog paketa (http://documentation.statsoft.com/STATISTICAHelp.aspx?path=MSPC%2FPCA%2FPCAExample)

8. ŽIVOTOPIS

Ime mi je Matea Dragoš. Rođena sam 09. listopada 1990. godine u Zagrebu. Godine 1997. započela sam osnovnoškolsko obrazovanje u Osnovnoj školi „Otona Ivekovića“ u Kustošiji, Zagreb. 2005. godine sam upisala Srednju školu „Škola za medicinske sestre vinogradska“ u Zagrebu, smjer medicinska sestra/tehničar. Srednjoškolsko obrazovanje sam završila 2009. godine, kada sam i maturirala sa odličnim uspjehom uz maturalni rad na temu Dijabetes. Upisala sam Fakultet kemijskog inženjerstva i tehnologije u Zagrebu, smjer Primijenjena kemija. Tijekom studiranja na preddiplomskom studiju odradila sam stručnu praksu na nastavnom Zavodu za javno zdravstvo „dr. Andrija Štampar“ uz mentorstvo dr.sc. Barbare Stjepanović.