ФОРМИРАНЕ НА КОМПЕТЕНТНОСТ ЗА ВИЗУАЛНО ПРЕДСТАВЯНЕ НА ИНФОРМАЦИЯТА В ДАННИТЕ ЧРЕЗ МЕТОДИ НА КОРЕСПОНДЕНТНИЯ АНАЛИЗ

Веска Нончева, Катерина Митева

Факултет по математика и информатика, ПУ "Паисий Хилендарски"
wesnon@uni-plovdiv.bg

***Резюме**: Тази статия се опитва да представи практика за изграждане на математическа компетентност за визуализиране на информацията, извлечена от данни.*

***Ключови думи**: кореспондентен анализ, пасивни точки, каноничен кореспондентен анализ*

1. Въведение

Съвременното образование е ориентирано към усвояване на компетентности чрез използване на нови научни методи и информационни ресурси [1].

Засилва се ролята на информацията в живота. Информацията и знанията се превръщат в основни продукти и определят характера на развитието на цивилизацията.

Хората възприемат информацията предимно визуално. Използването на визуални картини подобрява значително осмислянето на информацията. Когато информацията се представи с графики, хората изграждат стабилни асоциативни връзки, позволяващи дълготрайно съхраняване на получената информация.

Една от задачите на статистическия анализ е да получи информация от данните. Методите на кореспондентния анализ предоставят съвременни средства за визуализация на информацията носена от данните. Но един предварителен познавателен процес е необходим за да можем да виждаме информацията представена чрез графиките на кореспондентния анализ.

Целта на настоящата статия е да подпомогне формирането на компетентности за получаване на информация от данните като разкрие възможностите на някои методи на кореспондентния анализ. Притежаването на такива компетентности е добра основа за пълноценна професионална реализация.

2. Кореспондентен анализ

Математическият модел на кореспондентния анализ е следния: Имаме променлива, която разбива популацията на краен брой подпопулации. Такава променлива се нарича категорийна променлива. Стойностите, които може да приема една категорийна променлива, се наричат нива или категории. Данни, събрани от наблюдения над такава променлива се наричат категорийни данни. Тези данни могат да бъдат измерени в номиналната или наредената скала. Категорийните данни се представят обикновено в честотни таблици. Множеството от относителните честоти на една категорийна променлива се нарича профил.

Задачата на кореспондентния анализ е да визуализира категорийните данни в едномерното, двумерното или тримерното пространство [2]. Кореспондентният анализ (Correspondence Analysis, CA) работи с честотни таблици, като анализира разликите между относителните честоти. За да измери различието между категориите, CA използва *xu*-квадрат разстоянието. С χ2 статистиката е свързано понятието инертност. Инертността измерва разпръснатостта на категорийни данни. Инертност всъщност е частното (χ2 статистиката)/(обем на извадката). Тя е число от 0 до 2. Инертността е малка когато профилите лежат близо до средния профил.

Задача: Задачата на магазините е да удовлетворяват желанията и нуждите на своите клиенти. Изследването, проведено въз основа на удовлетвореността на хората, има за цел да изучи поведението на клиентите и подобри качеството на облужването им.

Нашата цел е да получим информация скрита в данните. За постигане на целта ще използваме методи на кореспондентния анализ. Като средство за анализа ще използваме функции на R и пакета *ca* в R ([4], [5]).

```
> library(ca)
```

Стойностите на променливата q20 са отговорите на въпроса Колко често посещавате сайтовете на търговците на дребно за да намерите отстъпки и промоции? Тази категорийна променлива има следните нива: Never, veryRarely, Rarely, Often, VeryOften. Променливата q28 е месечния доход на домакинството (представен в евро). Тази категорийна променлива има следните нива: LessThan668, 668-1330, 1330-2000, 2000-2670, 2670-3340, 3340-4000, MoreThan4000 и IdonotWantToTell.

```
> mytable <- with(mydata, table(q.20,q.28))    # create a 2 way table
> fit <- ca(mytable)
> print(fit)     # basic results
> summary (fit)
```

Principal inertias (eigenvalues):

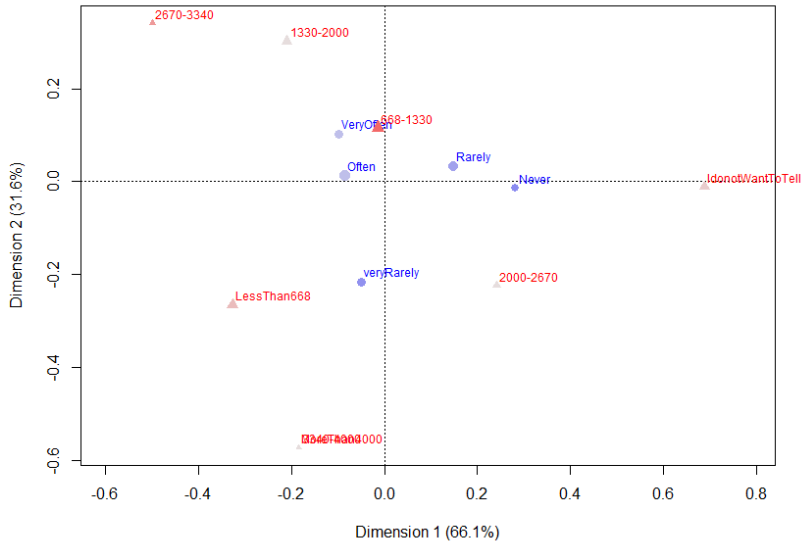
dim	value	%	cum%	scree plot
1	0.016635	66.1	66.1	**********
2	0.007949	31.6	97.7	********
3	0.000443	1.8	99.5	
4	0.000132	0.5	100.0	
		----------	----------	
Total:		0.025159	100.0	

Статистически извод: При проектиране на данните върху абцисната ос се запазва само 66.1% от информацията, а при проектиране върху двумерното пространство се запазва 97.7% от информацията. Инертността е 0.025.

Инертността може да бъде разположена като сума от положителни числа и това представлява числовия анализ на инертността, който е полезен при интерпретацията на резултатите.

Функцията plot() визуализира резултата от кореспондентния анализ:

> plot (fit, mass = TRUE, contrib = "absolute", map = "rowgreen")



Фигура 1. Графично визуализиране на резултатите от CA

Получените резултати се представят графично в Евклидовото пространство, което улеснява разбирането на данните и интерпретирането на резултатите. С най-наситен цвят (син и червен) са визуализирани тези точки, които имат най-голям принос при определяне на осите на координатната система.

Масите на точките отразяват очакваните средни относителни честоти. Тези маси са специфични тегла. СА подчертава (изтъква) нивата с по-големи тегла върху графиката по следния начин: Големината на кръгчетата и триъгълниците съответства на масата на точките. Инертността измерва колко близо са профилите до техния среден профил, който на графиката от СА съвпада с центъра на координатната система. Инертността може да бъде интерпретирана геометрично като претеглено средно χ2 разстояние между профилите и техния среден профил. Следователно, инертността може да бъде интерпретирана като степен на разпръснатост на профилите около техния среден профил, където точките са претеглени пропорционално на тяхната относителна честота.

Интерпретация на получените резултати: От графика виждаме, че хората имащи доход на член от семейството 668-1330 евра много често посещават сайтовете на търговците на дребно за да намерят повече отстъпки и промоции. Тези, които имат доход под LessThan668 евра много рядко се интересуват от отстъпки и промоции. А тези с доход 2000-2670 евра много рядко посещават сайтовете на търговците или никога.

3. Кореспондентен анализ с пасивни точки

Кореспондентният анализ визуализира данните в подходящо подпространство, като се опитва да запази разпръснатостта на данните. Когато имаме допълнителна информация за редовете или стълбовете на честотната таблица, можем да я визуализираме чрез допълнителни точки, наречени пасивни точки. Тези точки не влияят на решението на СА. Добавянето на пасивни точки не променя решението на СА, а само допринася за получаване на нова информация от данните[2].

Понякога може да искаме решението на СА да бъде пряко свързано с тези допълнителни точки. Тогава тези точки трябва да бъдат активни в кореспондентния анализ.

Задача: Променливата q25 представя броя на членовете от домакинството. Нивата на променливата q25 (т.е. възможните отговори на въпроса *От колко души се състои вашето домакинство*?) са One, Two, Three, Four, Five, MoreThanFive. Ще добавим тази променлива като пасивна точка в анализа. В този случай редовете на честотната таблица са нивата на

активната и пасивната променлива, подредени в азбучен ред. Векторът от пасивните точки трябва да бъде посочен във функцията ca().

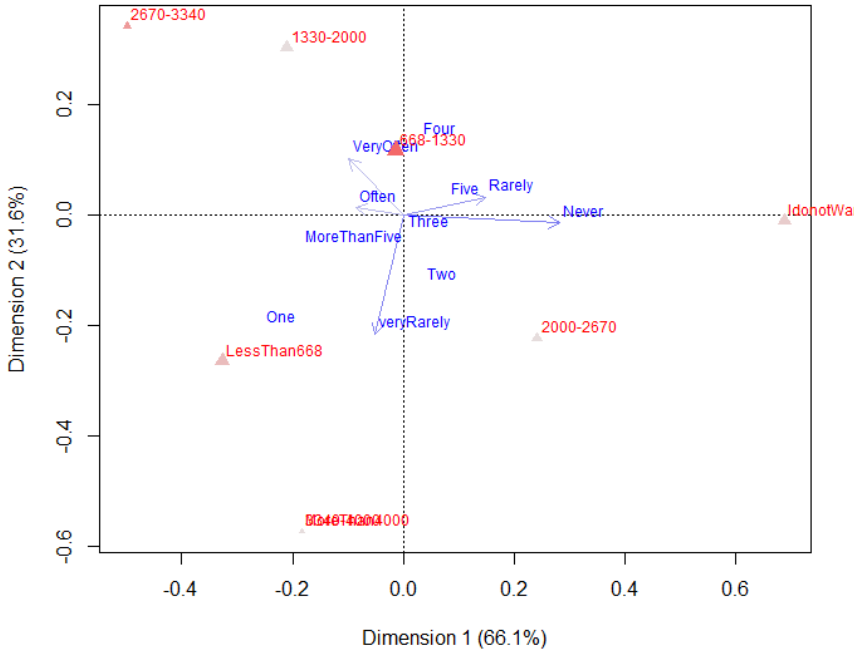
> fit <- ca(mytable,suprow=c(1,2,3,6,8,9))

> print(fit)

> summary(fit)

> plot.ca(fit, mass = c(TRUE,TRUE), contrib = "absolute", map = "rowgreen", arrows = c(TRUE,FALSE))

В графиката на Фигура 2 визуализираме активните точки със стрелки за да ги различим от пасивните.



Фигура 2. Графично визуализиране на резултатите от CA с пасивни точки

Интерпретация на получените резултати: От графика виждаме, че хората имащи доход на член от семейството 668-1330 евра много често посещават сайтовете на търговците на дребно за да намерят повече отстъпки

и промоции и това обикновено са четиричленни семейства. Тези, които имат доход LessThan668 евра много рядко се интересуват от отстъпки и промоции и това обикновено са едночленни семейства. А тези с доход 2000-2670 евра много рядко посещават сайтовете на търговците и това обикновено са двучленни семейства, или никога не посещават такива сайтове.

3. Каноничен кореспондентен анализ

Кореспондентният анализ намира основните (координатните) оси на пространство с по-ниска размерност като се стреми да запази разстоянието между точките. Каноничният кореспондентен анализ (Canonical Correspondence Analysis, CCA) е едно разширение на CA, което включва и външни обясняващи променливи, които често се наричат променливи на „околната среда". Тези допълнителни променливи могат да бъдат измерени в интервалната скала. Между осите на CA и външните променливи има линейна зависимост.

Каноничният кореспондентен анализ ограничава търсенето на оптималните основни оси в част от пространството, наречено канонично пространство. В каноничното пространство CA работи за да намери най-добрата размерност за да визуализира характерните особености на потреблението на населението.

Резултатът от CCA в каноничното пространство съдържа координатите на редовете и стълбовете от честотната таблица, както в CA. Графиката на CCA е същата като графиката на CA. Но към графиката на CCA могат да се добавят обясняващи променливи, които често се наричат „околната среда" на потребителя. Това са дадености, които ограничават потреблението на домакинството. Такива дадености могат да бъдат месечния доход, броят на членове в домакинството и други.

Графика на CCA се нарича триплот. Основната идея е, че в CCA осите са свързани с обясняващите потреблението променливи.

Задача: Ще изследваме процентния дял за храни и хранителни стоки от общите разходи на домакинствата през последния месец. Променливите Supermarket, SmallGrocery, Specialized и Online са количествени и показват какъв дял от тези разходи отива за пазаруване съответно в супермаркети, малки квартални магазини, специализирани магазини (напр. само млечни продукти, месни продукти и т.н.) и онлайн магазини. Променливите на „околната среда" са честотите на пазаруване в тези магазини и са запазени във вектора *data.env*.

Като средство за анализа ще използваме пакета *vegan* в R [3]. Функцията cca() реализира каноничния кореспондентен анализ.

> library (vegan)

> env.cca <- cca(store ~ Supermarket + SmallGrocery + Specialized + Online, data.env)

Пакетът vegan предоставя функции за отпечатване на резултатите от каноничния кореспондентен анализ print() и summary(). Ние предлагаме потребителска функция explanation.cca(), която обяснява статистическия резултат от този анализ.

> explanation.cca (env.cca)

[1] "The total variation in the data is 0.27"

[1] "The sum of all canonical eigenvalues is 0.26"

[1] "All explanatory variables explain 97 % of the total variation in the data."

[1] "The first two (canonical) eigenvalues are: "

CCA1 CCA2

0.13 0.08

[1] "So the first two canonical axes explain 78 of the variation that can be explained with all environmental variables."

[1] "But this is (the first two canonical axes explain) 76 % of the total variation in the data."

В нашия пример каноничният кореспондентен анализ е успешен: 97 % от общата разпръснатост на данните е „уловена" (отчетена) от каноничния кореспондентен анализ и 76 % от нея се запазва в равнината на първите две оси.

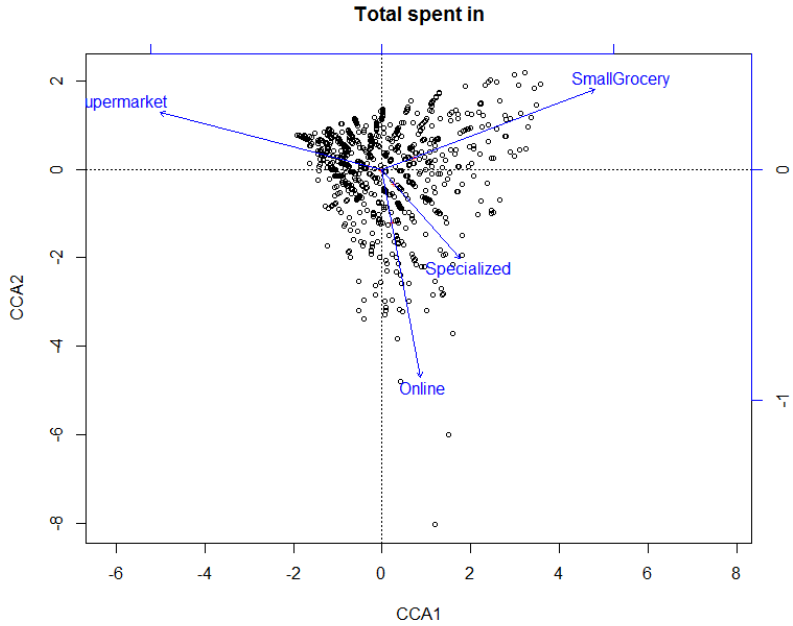
За да визуализираме резултата използваме функциите plot() и title() по следния начин:

> cca1.plot <- plot(env.cca) ; title ("Total spent in")

Интерпретация на получените резултати: В този анализ първата ос (абцисата) е свързана с намаляване на дяла на разходите за пазаруване в супермаркетите, а втората (ординатата) – с намаляване на дяла на разходите за пазаруване по Интернет. От графиката се вижда, че с намаляване на пазаруването в супермаркетите силно се увеличава пазаруването в малките квартални магазини.

Ще визуализираме домакинствата, за които дялът (в %) от общите разходи, похарчени в супермаркетите е по-малък от 60% .

> points (cca1.plot, what='sites', store$Supermarket < 60, col="green", bg="yellow", cex=0.7)



Фигура 3. Графично визуализиране на данните и резултатите от CCA

3.1. Добавяне на категорийни променливи в каноничния кореспондентен анализ

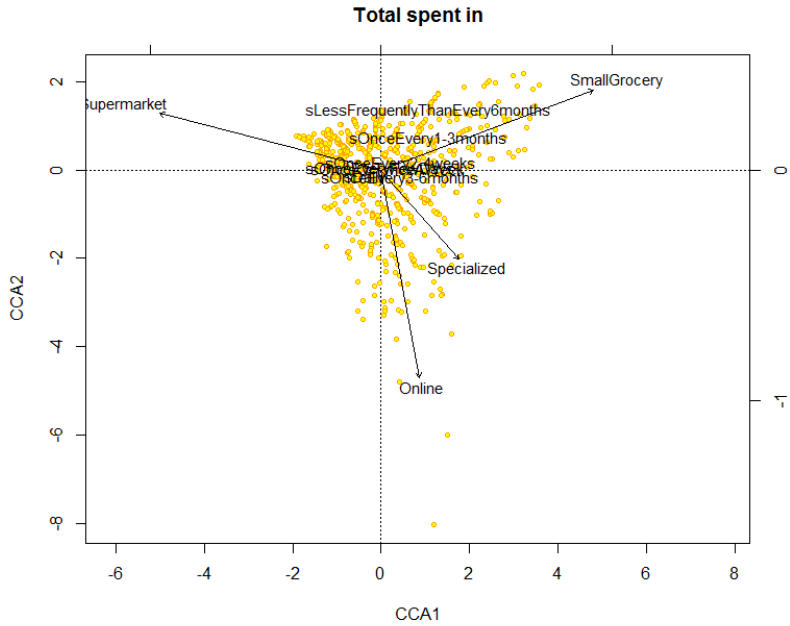
Всички променливи в каноничния кореспондентен анализ до този момент са количествени и са измерени в интервалната скала. Сега към анализа ще добавим и категорийна променлива s, която носи информация за честотата на пазаруване в супермаркетите.

```
> env.cca <- cca(store~Supermarket+SmallGrocery+Specialized+Online+s, data.env)
> plot(env.cca, type="n")
> points(env.cca, pch=21, col="orange", bg="yellow", cex=0.7)
> text(env.cca, dis="cn", cex=0.9)
> title ("Total spent in")
```

Новите категорийни променливи се включват като фиктивни променливи в CCA (dummy variables). В графиката на CCA към тези фиктивни променливи не сочат стрелки, както се визуализират непрекъснатите променливи. Те се

визуализират само с текст. Всяко ниво на категорийната променлива се представя в графиката на CCA.

Интерпретация на получените резултати: Тези, които пазаруват веднъж на шест месеца или по-рядко в супермаркетите, обикновено пазаруват в малките квартални магазини.



Фигура 4. Графично визуализиране на резултатите от CCA с добавяне на категорийната променлива честота на пазаруване в супермаркет

Заключение

Методите на кореспондентния анализ успешно се използват за извличане на информация от данните и за нейното визуализиране. .

Благодарности

Авторите считат за свой приятен дълг да отбележат благодарността си към Фонд "Научни изследвания" при ПУ "Паисий Хилендарски" за финансовата подкрепа при реализацията на проект СП15 ФМИИТ 015.

Литература

1. Кендеров П., Чехларова Т., Сендова Е. ЕВРОПЕЙСКИЯТ ПРОЕКТ KeyCoMath И ОРИЕНТИРАНОТО КЪМ УСВОЯВАНЕ НА КЛЮЧОВИТЕ КОМПЕТЕНТНОСТИ ОБРАЗОВАНИЕ ПО МАТЕМАТИКА, МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2015
2. Greenacre, Michael (2007). Correspondence Analysis in Practice, Second Edition. London: Chapman & Hall/CRC, 2010.
3. Jari Oksanen, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens and Helene Wagner. vegan: Community Ecology Package. R package version 2.3-2, 2015, https://CRAN.R-project.org/package=vegan
4. Nenadic, O., Greenacre, M. Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. Journal of Statistical Software 20(3):1-13, 2007
5. R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2016, https://www.R-project.org

MATHEMATICAL COMPETENCE FOR INFORMATION VISUALIZATION BY CORRESPONDENCE ANALYSIS METHODS

Veska Noncheva, Katerina Miteva

Abstract: *This paper attempts to outline teaching practice for building mathematical competency for visualization the information extracted from data.*