Korešpondenčná analýza

Soňa Coss[1]

Abstrakt
Korešpondenčná analýza je populárna metóda na vizualizáciu dát. Rozlišujeme dva druhy korešpondenčnej analýzy – jednoduchú a viacnásobnú. V oboch prípadoch sa výsledky zobrazujú do mapy pomocou bodov. Zakreslené body predstavujú relatívne početnosti kontingenčnej tabuľky. Pozícia bodov poukazuje na podobnosť medzi jednotlivými riadkovými kategóriami, medzi stĺpcovými kategóriami, ako aj na vzájomný vzťah medzi nimi. Pomocou tejto metódy sa snažíme zobraziť body znížením dimenzie priestoru, aby vynikli skryté vzťahy a asociácia medzi analyzovanými premennými.

Kľúčové slová
Korešpondenčná analýza, symetrická mapa

Abstract
Correspondence analysis is a popular visualization method. Two methods of analysis are commonly used, simple and multiple correspondence analysis. Both allow to vissualize the results to the correspondence map. Plotted points represent the relative frequencies of a contingency table. The position of the points shows the similarities within row categories, within column categories, as well as on the relationship between rows and columns. By this method, we try to display points by reducing the space dimension to capture the hidden relationships and association between the analyzed variables.

Key words
Correspondence Analysis, Symmetrical Map

JEL classification
C39

1 Úvod

Korešpondenčná analýza je štatistická metóda na analýzu vzťahov medzi kategóriami dvoch alebo viacerých premenných usporiadaných v kontingenčnej tabuľke. Umožňuje skúmať asociáciu kategoriálnych premenných a získať prehľadné grafické zobrazenie súvislostí v dvojrozmernom resp. viacrozmernom priestore. Cieľom je posúdiť vzájomný vzťah medzi premennými a vysvetliť štruktúru skúmanej závislosti.

Vstupnými premennými môžu byť akékoľvek kategoriálne premenné, ktoré sa dajú vyjadriť vo forme početností (absolútnych alebo relatívnych). Môžeme použiť nominálne, ordinálne alebo kvantitatívne diskrétne premenné. Ak chceme pracovať s kvantitatívnymi spojitými premennými, musíme ich hodnoty rozdeliť do kategórií.

Najdôležitejším výstupom analýzy je multidimenzionálna mapa, ktorú nazývame korešpondenčná mapa. V nej sú prehľadne zobrazené výsledky analýzy, zaznačením vzťahov medzi kategóriami v priestore v rovnakých dimenziách. Umožní nám posúdiť kategórie danej premennej, ich vzájomnú podobnosť a rozdiely medzi nimi, prípadne asociácie s kategóriami iných premenných. Táto metóda je obzvlášť vhodná pri analýze kontingenčných tabuliek

[1] Ekonomická univerzita, Fakulta hospodárskej informatiky, Katedra štatistiky, Dolnozemská cesta 1b, Bratislava, sona.coss@gmail.com

s veľkým počtom stĺpcov a riadkov, kde grafické zobrazenie môže byť omnoho prehľadnejšie ako tabuľkové výstupy.

Korešpondenčná analýza je analógiou metódy hlavných komponentov a faktorovej analýzy v prípade kategoriálnych premenných. Pomocou nej hľadáme latentné skryté faktory, ktoré predstavujú osy korešpondenčnej mapy. Aplikáciou metódy získame ordinačné osy (dimenzie) s klesajúcim stupňom dôležitosti. Snažíme sa nájsť také riešenie, v ktorom je možné zakresliť hlavnú informáciu z pôvodnej tabuľky do podpriestoru s nižším počtom dimenzií, pri čo najmenšej strate informácií. Najčastejšie využívame dvojrozmerný priestor.

Podľa Hebáka (2007) ide o redukciu mnohorozmerného priestoru vektorov riadkových a stĺpcových profilov pri zachovaní maximálnej informácie obsiahnutej v pôvodných dátach.

Korešpondenčná analýza je v oblasti marketingového výskumu stále viac využívanou metódou. Za popularitou metódy stojí jej základná výhoda - názorná a zrozumiteľná vizualizácia aj pomerne veľkých kontingenčných tabuliek. Rovnako aj fakt, že interpretácia korešpondenčnej mapy a hľadanie súvislostí je možné aj bez znalosti zložitého matematického aparátu na pozadí výpočtu metódy.

Jej nevýhodou je, že ide len o exploratívnu techniku, ktorou skúmame a popisujeme vzťahy medzi premennými. Nejde o verifikačnú metódu, ktorá by umožnila overenie vhodnosti dosiahnutého riešenia (Model Fit) a testovanie hypotéz.

Podľa počtu analyzovaných premenných rozlišujeme dva druhy korešpondenčnej analýzy. Jednoduchá korešpondenčná analýza sa používa pri analýze jednoduchej kontingenčnej tabuľky, teda pri skúmaní vzťahu dvoch premenných. V prípade, že chceme analyzovať viac ako dve kategoriálne premenné, využívame viacnásobnú korešpondenčnú analýzu.

V nasledujúcej časti popíšeme metodologické pozadie korešpondenčnej analýzy. Vstupnou maticou bude kontingenčná tabuľka z dvoch premenných, pričom riadky tvoria kategórie jednej premennej a stĺpce kategórie druhej premennej. V prípade viacnásobnej korešpondenčnej analýzy je potrebná ešte úprava vstupných premenných. Vstupnou maticou bude tabuľka, ktorú nazývame Burtova matica. Obsahuje všetky kombinácie kontingenčných tabuliek, ktoré sa dajú vytvoriť z daného počtu premenných. Na takto upravenú maticu môžeme následne aplikovať uvedený postup korešpondenčnej analýzy.

2 Metodologické pozadie korešpondenčnej analýzy

Pri jednoduchej korešpondenčnej analýze môžeme zadať vstupné premenné v rôznej podobe. Buď použijeme individuálne dáta, teda pôvodné hodnoty premenných u jednotlivých respondentov. Alebo môžeme priamo zadať absolútne početnosti z konkrétnej kontingenčnej tabuľky, (t.j. súhrnné počty prípadov), prípadne aj priemerné hodnoty za skupiny respondentov.

Vstupnou maticou analýzy je dvojrozmerná kontingenčná tabuľka matica **N** združených absolútnych početností n_{ij}. V jednotlivých políčkach tabuľky sú početnosti výskytu premennej X, ktorá nadobúda hodnoty x_i pre i=1,2,..r a premennej Y s hodnotami y_j pre j=1,2...s.

Z tabuľky môžeme vypočítať riadkové marginálne absolútne početnosti n_{i+} výskytu znaku X a stĺpcové marginálne absolútne početnosti znaku Y n_{+j} podľa vzťahov:

$$n_{i+} = \sum_{j}^{s} n_{ij} \qquad n_{+j} = \sum_{i}^{r} n_{ij} \tag{1}$$

Z nich sa následne vypočíta tzv. korešpondenčná matica, ktorú označíme **P**. Jej prvky tvoria relatívne početnosti p_{ij}, kde:

$$p_{ij} = \frac{n_{ij}}{n} \tag{2}$$

Okrem týchto početností napočítame aj marginálne relatívne početnosti, tzv. záťaže. Vznikajú vydelením marginálnych absolútnych početností celkovým počtom respondentov. Riadkové marginálne početnosti p_{i+} sa nazývajú riadkové záťaže. Stĺpcové marginálne početnosti p_{+j} nazývame stĺpcové záťaže. Pre riadkové a stĺpcové záťaže platia vzťahy:

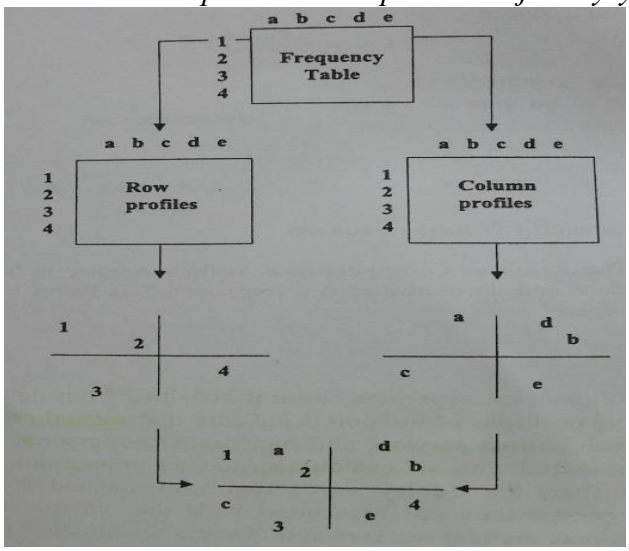
$$p_{i+} = \frac{n_{i+}}{n} \qquad p_{+j} = \frac{n_{+j}}{n} \tag{3}$$

Aby sa zabezpečila porovnateľnosť riadkových a stĺpcových kategórií, je v ďalšej analýze potrebné vypočítať profily. Riadkové profily $p_{j/i}$ sú podmienené relatívne početnosti, ktoré predstavujú štruktúru stĺpcovej premennej v prípade i-tej kategórie riadkovej premennej. Stĺpcové profily $p_{i/j}$ sú podmienené relatívne početnosti charakterizujúce štruktúru riadkovej premennej pri j-tej úrovni stĺpcovej premennej. Profily získame pomocou vzťahov:

$$p_{j/i} = \frac{n_{ij}}{n_{i+}} = \frac{p_{ij}}{p_{i+}} \qquad p_{i/j} = \frac{n_{ij}}{n_{+j}} = \frac{p_{ij}}{p_{+j}} \tag{4}$$

Zmenami v štruktúre riadkových a stĺpcových profilov sa prejavuje závislosť premenných. Jednotlivé riadkové a stĺpcové profily použijeme na výpočet súradníc bodov vo viacrozmernom priestore. Každému profilu priradíme váhu príslušného počtu pozorovaní, pričom váhou je práve záťaž. Tým prevedieme početnosti pôvodnej kontingenčnej tabuľky do porovnateľnej podoby.

Obr. 1: Schéma procesu korešpondenčnej analýzy



Zdroj: Nguyen Van Chuc (2011)

Korešpondenčná analýza rieši štyri základné otázky. A to: aké sú podobnosti a rozdiely medzi kategóriami riadkovej premennej z hľadiska rôznych úrovní stĺpcovej premennej. Aké sú podobnosti a rozdiely medzi kategóriami stĺpcovej premennej podľa rôznych obmien riadkovej premennej. Aký je vzájomný vzťah medzi riadkovými a stĺpcovými kategóriami navzájom. Poslednou je otázka, či môžeme graficky tento vzťah zobraziť v priestore s nižšou dimenziou.

Postup výpočtu korešpondenčnej metódy aj označenie symbolmi je podľa Řezánkovej (2007) nasledovný. Ak označíme maticu riadkových profilov symbolom **R** a maticu stĺpcových profilov ako **C**. Následne označíme r-členný vektor riadkových záťaží ako **r** a s-členný vektor stĺpcových záťaží ako **c**. Môžeme zapísať vzťahy medzi nimi:

$$\mathbf{R} = \mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P} = \begin{bmatrix} \mathbf{r}_1^{\mathrm{T}} \\ \mathbf{r}_2^{\mathrm{T}} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{r}_r^{\mathrm{T}} \end{bmatrix} \tag{5}$$

$$\mathbf{C} = \mathbf{D}_{\mathbf{c}}^{-1}\mathbf{P}^{\mathrm{T}} = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & . & . & . & \mathbf{c}_s \end{bmatrix} \tag{6}$$

, kde $\mathbf{D}_{\mathbf{r}}^{-1}$ je diagonálna matica s prvkami vektora **r** na diagonále a $\mathbf{D}_{\mathbf{c}}^{-1}$ je diagonálna matica s prvkami vektora **c** na diagonále. Korešpondenčnú maticu môžeme následne vyjadriť v tvare:

$$\begin{bmatrix} \mathbf{P} & \mathbf{r} \\ \mathbf{c}^{\mathrm{T}} & 1 \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1s} & r_1 \\ p_{21} & p_{22} & \cdots & p_{2s} & r_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{r1} & p_{r2} & \cdots & p_{rs} & r_r \\ c_1 & c_2 & \cdots & c_s & 1 \end{bmatrix} \tag{7}$$

Pričom pre vektor riadkových záťaží **r** a pre vektor stĺpcových záťaží **c** platia vzťahy:

$$\mathbf{r} = \sum_{j=1}^{s} p_{+j}\mathbf{c}_{\mathrm{j}} \qquad \mathbf{c} = \sum_{i=1}^{r} p_{i+}\mathbf{r}_{\mathrm{i}} \tag{8}$$

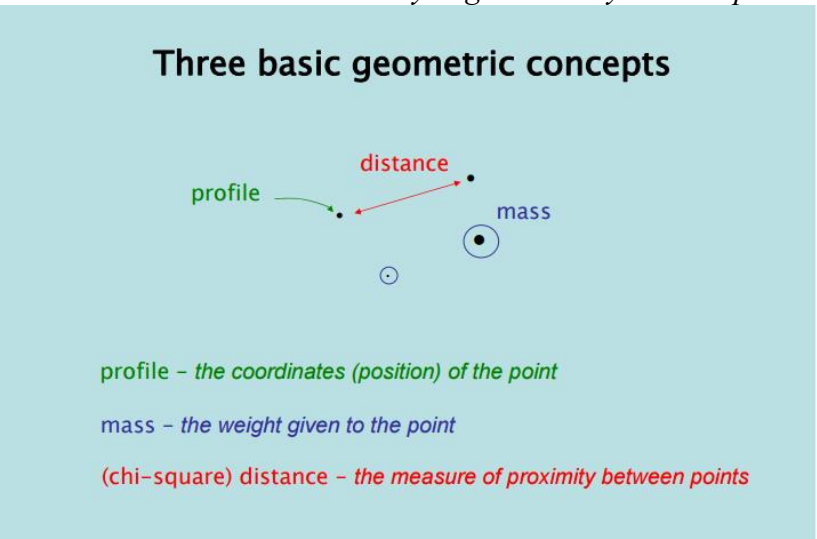
Pred samotným spustením procedúry korešpondenčnej analýzy je potrebné posúdiť vhodnosť vstupných premenných. Čo znamená, že pri analyzovaných premenných posúdime vzájomný vzťah, teda asociáciu. Na testovanie môžeme použiť Chí-kvadrát test nezávislosti, prípadne iné miery asociácie. Ak sa potvrdí štatisticky významná asociácia, sú dáta vhodné na aplikáciu metódy.

Následne zhodnotíme rozdiely, teda mieru nepodobnosti medzi kategóriami riadkovej a stĺpcovej premennej. Využijeme pritom výpočet Chí-kvadrát vzdialenosti definovaný:

$$D(i,i') = \sqrt{\sum_{j=1}^{s} \frac{(r_{ij} - r_{i'j})^2}{c_j}} \tag{9}$$

kde r_{ij} a $r_{i'j}$ sú prvky matice riadkových profilov **R** a c_j sú prvky vektora stĺpcových záťaží **c**. Vektor stĺpcových záťaží sa zároveň rovná priemernému stĺpcovému profilu, ktorý nazývame centroid (ťažisko) stĺpcových profilov. Analogicky vieme vypočítať mieru nepodobnosti, teda vzdialenosť medzi stĺpcovými kategóriami j a j'. Ako váhu použijeme r_i, teda prvky vektora riadkových záťaží **r**. Tento postup názorne zobrazuje schéma podľa Greenacra (2010).

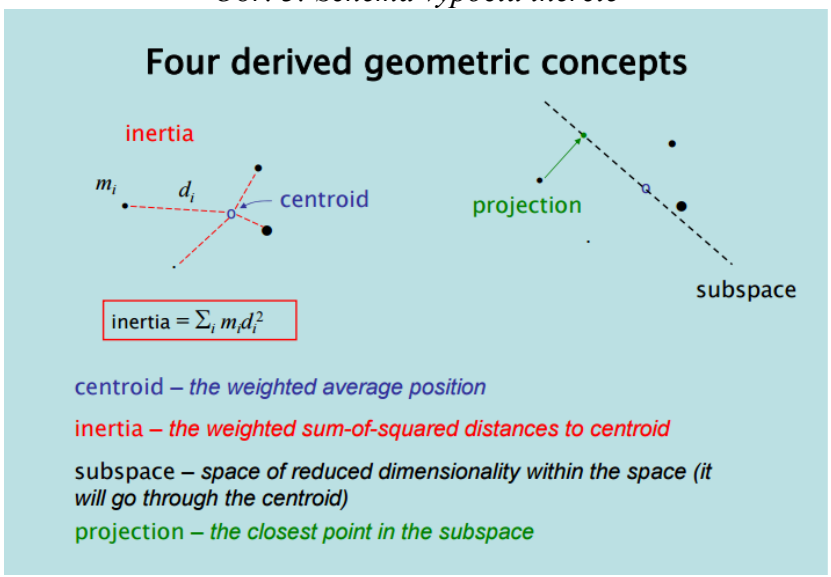
Obr. 2: Schéma troch základných geometrických konceptov



Zdroj: Greenacre (2010)

Podľa Holčíka a Komendu (2015) je algoritmus výpočtu korešpondenčnej analýzy obdobný ako pri analýze hlavných komponentov, teda pomocou vlastných čísiel. Avšak pri analýze hlavných komponentov, vypočítané vlastné čísla vyjadrujú vysvetlený rozptyl príslušným komponentom. V prípade korešpondenčnej analýzy predstavujú tzv. inerciu, teda vzťah medzi riadkovými a stĺpcovými kategóriami. Geometrická interpretácia inercie hovorí, že je mierou rozptýlenia profilov vo viacrozmernom priestore. Čím je väčšia jej hodnota, tým sú body v priestore viac rozptýlené. Inercia podľa Greenacra (2010) je definovaná ako vážený súčet štvorcov vzdialeností bodov od ich centroidu, čo naznačuje aj *Obr.3*.

Obr. 3: Schéma výpočtu inercie



Zdroj: Greenacre (2010)

Na výpočet vlastných čísiel použijeme rozklad na singulárne hodnoty. Najskôr vypočítame maticu štandardizovaných rezíduí **Z.** Je to z dôvodu, že pri zobrazovaní do grafu nehľadáme súradnice v pôvodných riadkových a stĺpcových profiloch. Robíme to na základe štandardizovaných rezíduí, ktoré predstavujú odchýlku riadkových a stĺpcových kategórií od nezávislosti. Prvky tejto matice nadobúdajú hodnoty podľa vzťahu:

$$z_{ij} = \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} \tag{10}$$

Maticu štandardizovaných rezíduí rozložíme na singulárne hodnoty podľa vzťahu:

$$\mathbf{Z} = \mathbf{U} \cdot \mathbf{\Gamma} \cdot \mathbf{V}^{\mathrm{T}} \tag{11}$$

, kde $\mathbf{\Gamma}$ je diagonálna matica, a platí, že $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = 1$, teda jednotková matica.

Pomocou tohto postupu získame vlastné čísla, ktoré sú usporiadané zostupne podľa veľkosti. Cieľom korešpondenčnej analýzy je redukovať priestor na nižší počet dimenzií, pri maximálnom zachovaní informácie z pôvodnej kontingenčnej tabuľky. Maximálny počet dimenzií, ako aj počet vlastných čísiel je minimum z počtu riadkov a počtu stĺpcov v pôvodnej vstupnej matici znížený o jednotku.

$$q = \min\{(r-1),(s-1)\} \tag{12}$$

Ďalším krokom v analýze je určiť spôsob, akým budú vypočítané súradnice bodov do korešpondenčnej mapy. Na výber máme niekoľko typov normalizácie. Ak nás zaujíma predovšetkým analýza riadkových kategórií, volíme ako normalizačnú metódu analýzu riadkových profilov. V prípade, že je cieľom skúmanie stĺpcových kategórií, rozhodneme sa pre analýzu stĺpcových profilov. Treťou možnosťou je symetrická normalizácia, ako kombinácia uvedených dvoch prístupov. Ide o postup, kde sa navzájom porovnávajú riadkové a stĺpcové kategórie. Normalizačná metóda nemá vplyv na výpočet singulárnych hodnôt, iba sa mení variabilita súradníc.

Postup pri analýze riadkových profilov podľa Řezánkovej (2007) je taký, že sa súradnice riadkových kategórií určia v stĺpcoch matice **F** a súradnice stĺpcových kategórií pomocou stĺpcov matice **Y**, podľa vzťahov:

$$\mathbf{F} = \mathbf{D}_{\mathbf{r}}^{-1/2}\mathbf{U}\mathbf{\Gamma} \qquad \mathbf{Y} = \mathbf{D}_{\mathbf{c}}^{-1/2}\mathbf{V} \tag{13}$$

Ak si zvolíme analýzu stĺpcových profilov, súradnice stĺpcových profilov nájdeme v stĺpcoch matice **G** a súradnice riadkových profilov pomocou stĺpcov matice **X.**

$$\mathbf{G} = \mathbf{D}_{\mathbf{c}}^{-1/2}\mathbf{V}\mathbf{\Gamma} \qquad \mathbf{X} = \mathbf{D}_{\mathbf{r}}^{-1/2}\mathbf{U} \tag{14}$$

Pri voľbe symetrickej normalizácie získame súradnice riadkových kategórií pomocou matice F a súradnice stĺpcových kategórií pomocou matice G.

$$\mathbf{F} = \mathbf{D}_{\mathbf{r}}^{-1/2}\mathbf{U}\mathbf{\Gamma} \qquad \mathbf{G} = \mathbf{D}_{\mathbf{c}}^{-1/2}\mathbf{V}\mathbf{\Gamma} \tag{15}$$

Po výpočte súradníc môžeme zobraziť kategórie v redukovanom súradnicovom systéme, v korešpondenčnej mape. Výhodou je, že do mapy môžeme zaznačiť kategórie obidvoch premenných súbežne. Tým získame predstavu jednak o vzájomnej podobnosti a rozdieloch kategórií v rámci každej premennej, o vzájomnom vzťahu oboch premenných, teda aj o štruktúre závislosti.

Podľa Holčíka a Komendu (2015) môžeme interpretovať vzdialenosti medzi riadkovými kategóriami, aj vzdialenosti medzi stĺpcovými kategóriami, nie však vzdialenosti riadkových a stĺpcových bodov. Čo je však možné, je posúdiť relatívnu pozíciu bodu z jednej sady, vzhľadom na všetky body druhej sady. Platí, že blízkosť dvoch riadkových (resp. stĺpcových) kategórií, poukazuje na podobnosti v profiloch týchto riadkov (stĺpcov). Ak sú od seba vzdialené, nemajú podobné profily. Blízkosť určitej kategórie riadka a určitej kategórie stĺpca naznačí, že táto kategória má dôležitú váhu v danom stĺpci. Ak sú tieto dve kategórie vzdialené, znamená to, že v danom stĺpci a riadku sa nevyskytujú takmer žiadne pozorovania. Ak sú nejaké body blízko stredu mapy (na pozícii v blízkosti nuly), nemajú výrazný profil, teda sú blízko ťažisku.

Okrem posúdenia vzájomnej pozície bodov v korešpondenčnej mape, je potrebné pokúsiť sa aj o vhodnú interpretáciu jednotlivých osí mapy. Ide nám o primerané pomenovanie osí, teda nájsť vhodný názov pre faktor na pozadí mapy. Niekedy môže byť výstižnejšie priradiť názov jednotlivým kvadrantom mapy.

Ako sme už spomenuli, korešpondenčná analýza nám neumožňuje otestovať štatistickú významnosť modelu ako celku. Ponúka nám však výpočet ukazovateľov, ktorých hodnoty naznačia, či je daný počet dimenzií postačuje na zobrazenie pôvodnej informácie z dát.

Najdôležitejším ukazovateľom je už spomínaná inercia. Keďže charakterizuje mieru rozptýlenia riadkových a stĺpcových kategórií, je v podstate analógiou rozptylu. S inerciou sa stretneme vo viacerých výstupoch korešpondenčnej analýzy.

Procedúra najskôr vypočíta hodnotu tzv. hlavných inercií (Principal Inertias), ktoré sú druhou mocninou vlastných čísiel matice štandardizovaných rezíduí. Vlastné čísla aj hlavné inercie sú napočítané pre každú dimenziu zvlášť. Sú usporiadané zostupne podľa veľkosti a charakterizujú dôležitosť poradia jednotlivých osí. Celková inercia (Total Inertia) je potom súčtom všetkých hlavných inercií.

Výstup analýzy nám ponúka aj vyjadrenie percentuálneho podielu inercií jednotlivých dimenzií na celkovej inercii, ako aj ich kumulatívne hodnoty. Relatívny podiel inercie danej dimenzie predstavuje hodnotu vysvetlenej inercie daným rozmerom, vyjadrenú v percentách. Kumulant relatívneho podielu informuje, koľko percent celkovej inercie je vysvetlené spoločne určitým počtom dimenzií. Tieto hodnoty slúžia na rozhodnutie analytika o znížení dimenzie, teda o vhodnom počte osí na vykreslenie grafu. Predpokladáme totiž, že podstatnú časť informácie vykreslíme pomocou dvoch, prípadne troch prvých osí. Postupujeme dvomi spôsobmi, buď zhodnotíme, či kumulatívne percento podielu inercií prvých dvoch osí je dostatočne vysoké. Alebo si zvolíme hraničnú hodnotu (napr. 80%) a zistíme, koľko prvých osí má kumulatívne percento väčšie, ako zvolená táto hodnota a tie použijeme v mape.

V ďalšom výstupe korešpondenčnej analýzy nájdeme ešte niekoľko ukazovateľov súvisiacich s inerciou. Nie sú už počítané z hľadiska jednotlivých dimenzií, ale samostatne z hľadiska jednotlivých riadkových, resp. stĺpcových kategórií.

Riadková inercia (resp. stĺpcová inercia), vyjadruje informáciu o variabilite, teda miere rozpýlenia jednotlivých riadkových (stĺpcových) kategórií.

Ďalším ukazovateľom sú príspevky riadkových bodov k inercii v príslušnej dimenzii (Contribution of Point to Inertia of Dimension). Vyjadrujú relatívnu mieru vplyvu danej kategórie na výslednú orientáciu jednotlivých osí. Získame z nich informáciu, ktoré riadkové (resp. stĺpcové) kategórie najviac prispievajú na orientáciu prvej osy, a ktoré kategórie majú najvyšší vplyv na orientáciu druhej osy.

Poslednou skupinou ukazovateľov sú príspevky osí k reprodukcii riadkových (stĺpcových) kategórií, tzv. Contribution of Dimension to Intertia of Point. Určia nám príspevok jednotlivých osí na vysvetlení príslušnej riadkovej (stĺpcovej) kategórie. Môžeme ich interpretovať ako koreláciu riadkových (stlpcových) profilov s jednotlivými osami. Ak následne sčítame príspevky hlavných osí v rámci danej kategórie, získame analógiu komunalít z faktorovej analýzy.

Pred samotným spustením korešpondenčnej analýzy pomocou softvéru SPSS musíme urobiť niekoľko rozhodnutí. Pri analýze nastavujeme preferovaný výpočet miery vzdialenosti. Buď použijeme štandardnú Chí-kvadrát vzdialenosť alebo euklidovskú mieru. Môžeme zvoliť štandardizáciu (odčítanie strednej hodnoty), teda centrovanie riadkov a stĺpcov. A v neposlednom rade aj normalizačnú metódu. Na výber máme riadkovú normalizáciu (ak preferujeme skúmanie kategórií riadkovej premennej), stĺpcovú normalizáciu (ak je hlavným cieľom skúmania podobnosť stĺpcových kategórií). Ďalšou možnosťou je symetrická normalizácia, ktorá v sebe kombinuje prístup oboch predošlých a umožňuje analýzu riadkových aj stĺpcových kategórií súbežne. Zároveň tiež môžeme nastaviť preferovaný počet dimenzií.

3 Aplikácia korešpondenčnej analýzy pri analýze frekvencie športovania v závislosti od veku respondentov

Pri aplikácii jednoduchej korešpondenčnej analýzy sme využili údaje z reprezentatívneho prieskumu, ktorý realizovala agentúra TNS Slovakia prostredníctvom online dopytovania v máji 2015. Výberová vzorka 1000 respondentov bola reprezentatívna pre populáciu Slovenska vo veku 18 a viac rokov z hľadiska pohlavia, veku, vzdelania, veľkosti miesta bydliska a kraja.

Zaujímalo nás, či a akým spôsobom závisí frekvencia športovania od vekovej štruktúry respondentov. Do analýzy sme zahrnuli dve kategoriálne premenné, teda sme použili jednoduchú korešpondenčnú analýzu. Výpočty boli realizované v prostredí IBM SPSS Statistics v.21, pomocou procedúry *Analyze/Dimension Reductions/Correspondence Analysis.*

Ako riadkovú premennú sme zvolili ordinálnu premennú vekové kategórie, ktorá nadobúda päť možných kategórií a to: 18-29 rokov, 30-39 rokov, 40-49 rokov, 50-59 rokov a 60 rokov a viac. Stĺpcovou premennou bola ordinálna premenná frekvencia športovania. Respondenti v nej odpovedali na otázku „Ako často športujete?“. Na výber mali jednu z piatich možností odpovedí: každý deň, viackrát do týždňa, jedenkrát týždenne, menej často alebo vôbec nešportujem.

Čo sa týka samotného nastavenia procedúry v IBM SPSS, zvolili sme výpočet Chi-kvadrát miery vzdialeností a štandardizáciu centrovania riadkov aj stĺpcov. Na vytvorenie mapy sme použili symetrickú normalizáciu, teda analýzu riadkových aj stĺpcových kategórií súbežne.

Tab. 1: Kontingenčná tabuľka absolútnych početností

	Ako často športujete?					
Vekové kategórie	Každý deň	Viackrát do týždňa	1 x týždenne	menej často	Vôbec nešportujem	Active Margin
18 - 29 rokov	21	61	49	78	17	226
30 - 39 rokov	11	50	41	56	46	204
40 - 49 rokov	3	27	28	59	50	167
50 - 59 rokov	6	25	33	50	62	176
60 rokov a viac	10	35	35	55	92	227
Active Margin	51	198	186	298	267	1000

Zdroj: Vlastné spracovanie

Prvým výstupom po spustení korešpondenčnej analýzy je dvojrozmerná kontingenčná tabuľka absolútnych početností (Tab.1). Obsahuje združené absolútne početnosti jednotlivých vekových kategórií podľa jednotlivých frekvencií športovania. Tabuľka tiež obsahuje marginálne absolútne početnosti 1. stupňa (Active Margin). Riadkové absolútne početnosti hovoria o rozdelení počtu respondentov v jednotlivých vekových skupinách. Stĺpcové marginálne početnosti informujú o počte prípadov pri jednotlivých úrovniach frekvencie športovania.

Následne boli vypočítané riadkové a stĺpcové profily (Tab.2 a Tab.3), ktoré umožnili previesť združené absolútne početnosti na porovnateľný základ, aby mohli byť zakreslené do jednej mapy.

Tab. 2: Riadkové profily (Row Profiles)

	Ako často športujete?					
Vekové kategórie	Každý deň	Viackrát do týždňa	1 x týždenne	menej často	Vôbec nešportujem	Active Margin
18 - 29 rokov	,093	,270	,217	,345	,075	1,000
30 - 39 rokov	,054	,245	,201	,275	,225	1,000
40 - 49 rokov	,018	,162	,168	,353	,299	1,000
50 - 59 rokov	,034	,142	,188	,284	,352	1,000
60 rokov a viac	,044	,154	,154	,242	,405	1,000
Mass	,051	,198	,186	,298	,267	

Zdroj: Vlastné spracovanie

Riadkové profily sú podmienené relatívne početnosti a predstavujú štruktúru súboru z hľadiska športovania v jednotlivých vekových kategóriách. Vo vekovej skupine 18-29 rokov športuje každý deň 9,3% mladých, avšak viackrát do týždňa je to až 27%. Keďže ide o tzv. riadkové percentá (Row %), suma jednotlivých riadkov je rovná jednej. V poslednom riadku Tab. 2 sa nachádzajú stĺpcové záťaže (Mass), teda stĺpcové marginálne relatívne početnosti. Vidíme, že až 26,7% všetkých respondentov deklaruje, že vôbec nešportujú.

Tab. 3: Stĺpcové profily (Column Profiles)

	Ako často športujete?					
Vekové kategórie	Každý deň	Viackrát do týždňa	1 x týždenne	menej často	Vôbec nešportujem	Mass
18 - 29 rokov	,412	,308	,263	,262	,064	,226
30 - 39 rokov	,216	,253	,220	,188	,172	,204
40 - 49 rokov	,059	,136	,151	,198	,187	,167
50 - 59 rokov	,118	,126	,177	,168	,232	,176
60 rokov a viac	,196	,177	,188	,185	,345	,227
Active Margin	1,000	1,000	1,000	1,000	1,000	

Zdroj: Vlastné spracovanie

Stĺpcové profily (Tab. 3) charakterizujú štruktúru vekovej skladby pri jednotlivých frekvenciách športovania. Najväčší podiel respondentov, ktorí športujú každý deň nájdeme v najmladšej vekovej skupine 18-29 ročných a predstavuje 41,2%. Keďže ide o vyjadrenie štruktúry, suma každého stĺpca je rovná jednej. V poslednom stĺpci tabuľky sú riadkové záťaže, ktoré hovoria o štruktúre vekových skupín v rámci celej vzorky respondentov.

Tab. 4: Súhrnná tabuľka (Summary)

					Proportion of Inertia		Value	
Dimension	Singular Value	Inertia	Chi Square	Sig.	Accounted for	Cumulative	Standard Deviation	Correlation 2
1	,284	,081			,894	,894	,026	-,002
2	,084	,007			,079	,973	,029	
3	,042	,002			,020	,993		
4	,026	,001			,007	1,000		
Total		,090	90,231	,000[a]	1,000	1,000		

a. 16 degrees of freedom

Zdroj: Vlastné spracovanie

Súhrnná tabuľka (Tab. 4) obsahuje množstvo dôležitých štatistík. Asi najdôležitejšou informáciou je výsledok Chí kvadrát testu nezávislosti premenných. Na základe neho posúdime, či medzi frekvenciou športovania a vekom je asociácia, teda či má zmysel pokračovať v interpretácii výsledkov korešpondenčnej analýzy. Hodnota testovacej charakteristiky je 90,231 a hodnota signifikancie (Sig.) je nižšia ako akákoľvek zvolená hladina významnosti. Preto zamietame hypotézu o nezávislosti a môžeme tvrdiť, že medzi skúmanými premennými je asociácia.

Z tabuľky ďalej vidíme, že riešenie je vypočítané pre 4 dimenzie. Je to z toho dôvodu, že obe premenné majú zhodne po 5 možných kategórií. Teda podľa vzorca (12), nám počet kategórií znížený o jednotku dáva riešenie práve v 4-rozmernom priestore.

V druhom stĺpci tabuľky máme singulárne hodnoty vypočítané pre jednotlivé dimenzie. Hlavné inercie v treťom stĺpci sú druhou mocninou príslušnej singulárnej hodnoty ($0{,}284^2=0{,}081$). Vyjadrujú zostupne mieru rozptýlenia bodov v danej dimenzii. Súčet všetkých hlavných inercií dáva celkovú inerciu, v našom prípade rovnú 0,09. Celkovú inerciu je možné vypočítať aj z Chí kvadrát štatistiky, a to: 90,231/1000=0,090.

Dva stĺpce, nazvané spoločne Proportion of Inertia, poskytujú informáciu o vhodnom počte dimenzií na zakreslenie do mapy. Stĺpec Accounted for predstavuje hodnotu inercie prepočítanú ako percento z celkovej inercie, teda relatívny podiel inercie danej dimenzie. Vyjadruje množstvo vysvetlenej inercie (variability) danou dimenziou. V našom prípade prvá dimenzia vysvetľuje 89,4% variability vzťahov v kontingenčnej tabuľke, druhá dimenzia už len 7,9% a pod. Stĺpec Cumulative obsahuje kumulatívne percento vysvetlenej variability. To znamená, že v prípade, ak sa rozhodneme pre zakreslenie hodnôt do dvojrozmernej mapy, podarí sa nám zachytiť 97,3% variability pôvodných bodov. Čo je pri výraznom znížení rozmeru zobrazenia len nepatrná strata informácie.

Tab. 5: Prehľad riadkových bodov (Overview Row Points)

Vekové kategórie	Mass	Score in Dimension		Inertia	Contribution				
					Dimension		Of Dimension to Inertia of Point		
		1	2		1	2	1	2	Total
18 - 29 rokov	,226	-,859	,019	,048	,587	,001	,993	,000	,993
30 - 39 rokov	,204	-,208	,176	,004	,031	,075	,589	,125	,714
40 - 49 rokov	,167	,232	-,577	,007	,032	,658	,347	,638	,985
50 - 59 rokov	,176	,402	-,078	,009	,100	,013	,930	,010	,940
60 rokov a viac	,227	,560	,307	,022	,250	,254	,907	,081	,988
Active Total	1,000			,090	1,000	1,000			

Zdroj: Vlastné spracovanie

Tab. 5 poskytuje prehľad o jednotlivých kategóriách riadkovej premennej. V stĺpci Mass sú zobrazené jednotlivé riadkové záťaže, čo je zároveň priemerný stĺpcový profil. Hovoria o relatívnom podiele výskytu jednotlivých kategórií v rámci danej vzorky respondentov. Ďalšie dva stĺpce, Score in Dimension, predstavujú súradnice jednotlivých riadkových kategórií, ktoré budú využité na zakreslenie do mapy. Stĺpec Inertia obsahuje riadkové inercie, teda mieru rozptýlenia jednotlivých kategórií.

Dva ďalšie stĺpce Contribution of Point to Inertia of Dimension, ponúkajú príspevky riadkových bodov k inercii danej osy. Vidíme z nich, že najväčší vplyv na orientáciu prvej osy má kategória 18-29 rokov (vplyv 58,7%) a kategória 60 a viac rokov (25%). Tieto dve kategórie spoločne určujú až 83,7% orientácie prvej osy. Na orientáciu druhej osy najviac vplývala kategória 40-49 ročných.

Posledné tri stĺpce predstavujú príspevky osí k reprodukcii riadkových kategórií. Označené sú ako Contribution of Dimension to Inertia of Point a vyjadrujú korelácie riadkových profilov s príslušnou osou. Posledný stĺpec Total dáva informáciu o kvalite zobrazenia danej kategórie pomocou dvoch osí. Vidíme, že hodnoty pri všetkých kategóriách, s výnimkou kategórie 30-39 rokov, prevyšujú 90%. To znamená, že kvalita zobrazenia jednotlivých kategórií riadkovej premennej pomocou týchto dvoch osí je veľmi dobrá.

Tab. 6: Prehľad stĺpcových bodov (Overview Column Points)

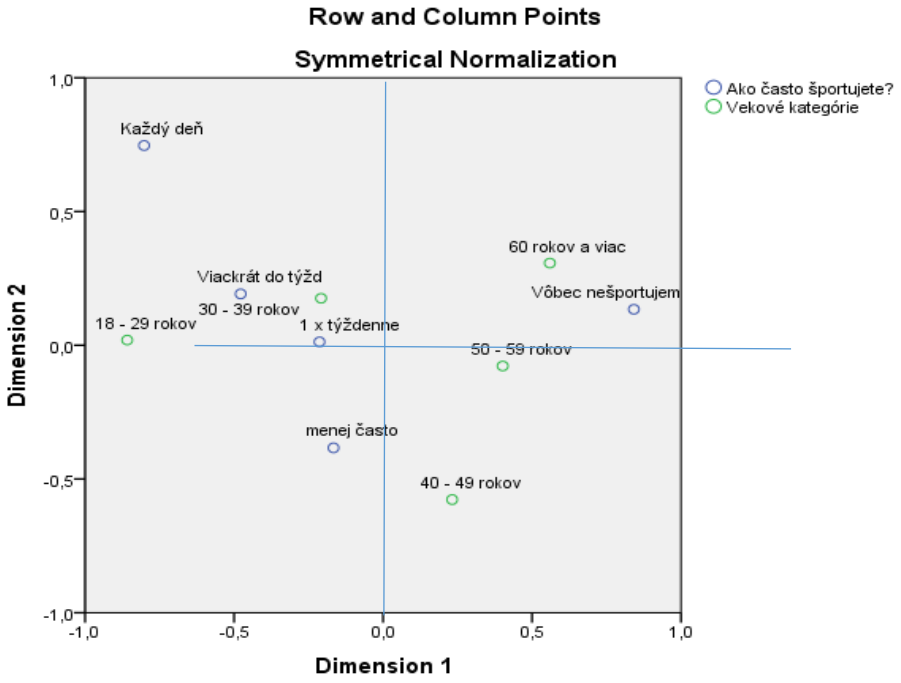
Ako často športujete?	Mass	Score in Dimension		Inertia	Contribution				
					Dimension		Of Dimension to Inertia of Point		
		1	2		1	2	1	2	Total
Každý deň	,051	-,802	,747	,013	,116	,337	,740	,191	,930
Viackrát do	,198	-,478	,192	,014	,159	,086	,900	,043	,943
1 x týždenne	,186	-,213	,013	,003	,030	,000	,817	,001	,818
menej často	,298	-,166	-,384	,006	,029	,519	,374	,593	,967
Vôbec	,267	,842	,134	,054	,666	,057	,992	,008	1,000
Active Total	1,000			,090	1,000	1,000			

Zdroj: Vlastné spracovanie

Následne sú napočítané tieto ukazovatele aj pre stĺpcové kategórie. Z Tab.6 okrem iného vidíme, že najväčší príspevok na inerciu prvej osy mala kategória vôbec nešportujem (66,6%). Orientáciu druhej osy najviac ovplyvnila kategória športujem menej často.

Čo sa týka kvality zobrazenia stĺpcovej premennej, vidíme, že kategória vôbec nešportujem bola týmito dvoma osami vysvetlená úplne (Total=100%). Ostatné kategórie je možné pomocou týchto dvoch osí, veľmi kvalitne zobraziť. Keďže vysvetlená variabilita daných kategórií predstavuje vo všetkých prípadoch viac ako 82 percent.

Obr. 2: Symetrická mapa riadkových a stĺpcových kategórií



Zdroj: Vlastné spracovanie

Výsledná korešpondenčná mapa nám názorne zobrazuje výsledky analýzy, čím umožňuje nájsť interpretáciu vzťahov a štruktúry závislosti v kontingenčnej tabuľke. Poloha riadkových a stĺpcových kategórií naznačuje, ktoré kategórie spolu súvisia, teda navzájom korešpondujú.

Respondenti vo veku 18-29 rokov a 30-39 rokov sú si z hľadiska frekvencie športovania veľmi podobní. Kategória 40-49 ročných respondentov má v tomto prípade špecifické postavenie. Z grafu tiež pozorujeme podobnosť profilov vekových kategórií 50-59 ročných a respondentov starších ako 60 rokov.

Čo sa týka jednotlivých frekvencií športovania, vidíme, že spolu korešpondujú kategórie 1x do týždňa a viackrát do týždňa, ktoré sa spolu s hraničnou kategóriou každý deň, nachádzajú v prvom kvadrante mapy. Kategórie menej často a vôbec nešportujem majú samostatné, izolované miesto na mape. Mladší respondenti inklinujú skôr k častejšej forme športovania.

Respondenti 40-49 roční sú na mape zachytení najbližšie ku kategórii menej často. Respondenti v staršom veku už sú zachytení najbližšie ku kategórii vôbec nešportujem.

Podľa vertikálnej osy je v grafe zachytené odlíšenie mladších vekových kategórií od starších. Táto os tiež zachytáva odlíšenie kategórií športujem pravidelne od respondentov, ktorí nešportujú vôbec. Názov tejto osy, ako skrytého faktora na pozadí mapy, by sme mohli označiť pojmom aktivita respondentov.

Podľa horizontálnej osy vidíme odlíšenie kategórií 40-49 ročných a 60 a viac ročných, kde pozorujeme najväčšie rozdiely. Z hľadiska frekvencie športovania je to odlíšenie kategórií menej často a ostatných foriem športovania. Druhá dimenzia zobrazenia tak ešte viac poukazuje na významný zlom v prístupe ku športu, ku ktorému dochádza v strednom veku respondentov.

4 Záver

Na základe uvedeného vyplýva, že korešpondenčná analýza umožňuje veľmi prehľadné a interpretovateľné výstupy, ktoré poskytujú hlbšie pochopenie vzťahov v kontingenčnej tabuľke. Dôležitým aspektom jej využitia je aj fakt, že táto metóda nemá limitujúce požiadavky v prípade predpokladov aplikácie a je možné ju využiť pri rôznorodej škále kategoriálnych premenných. To umožňuje jej široké použitie pri analýzach marketingových a sociologických dát.

Príspevok bol spracovaný v rámci riešenia grantovej úlohy VEGA 1/0092/15 *Moderné prístupy k navrhovaniu komplexných štatistických prieskumov*, ktorého vedúcim je prof. Ing. Milan Terek, PhD.

Literatúra

[1] Greenacre, M. (2010). *Correspondence Analysis and Related Methods*. Retrieved October 21, 2016, from http://statmath.wu.ac.at/courses/CAandRelMeth/CARME1.pdf

[2] Hebák, P. (2007). *Vícerozměrné statistické metody*. Praha: Informatorium.

[3] Holčík, J., & Komenda, M. (2015). *Matematická biologie: e-learningová učebnice* [online]. Retrieved October 21, 2016, from https://www.muni.cz/vyzkum/publikace/1334889

[4] Chuc, N. V. (2011). *Correspondence Analysis with XLSTAT*. Retrieved October 21, 2016, from http://www.slideshare.net/chucnv/correspondence-analysisstep-by-step

[5] Řezanková, H. (2011). *Analýza dat z dotazníkových šetření*. Praha: Professional Publishing.