Základy regionálnej geografie – cv. 8

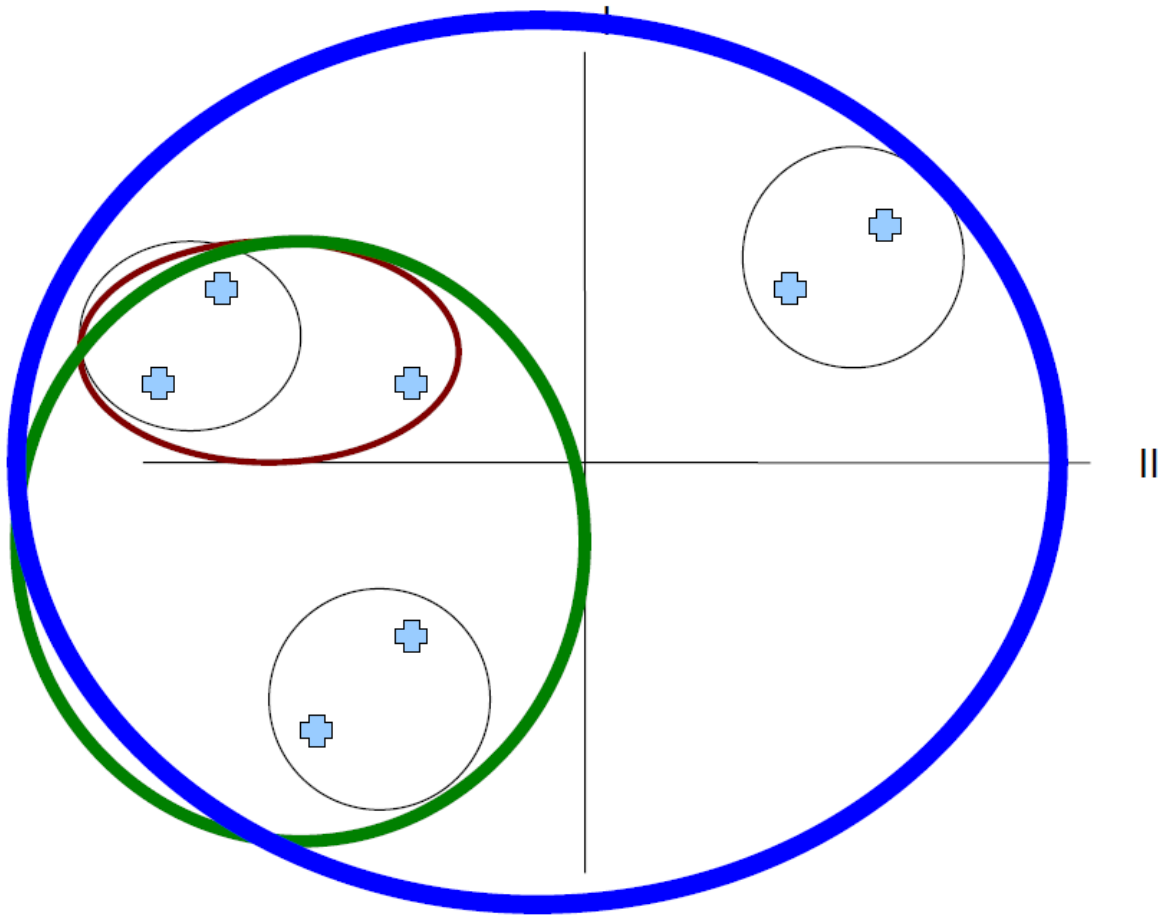
Jozef LOPUCH

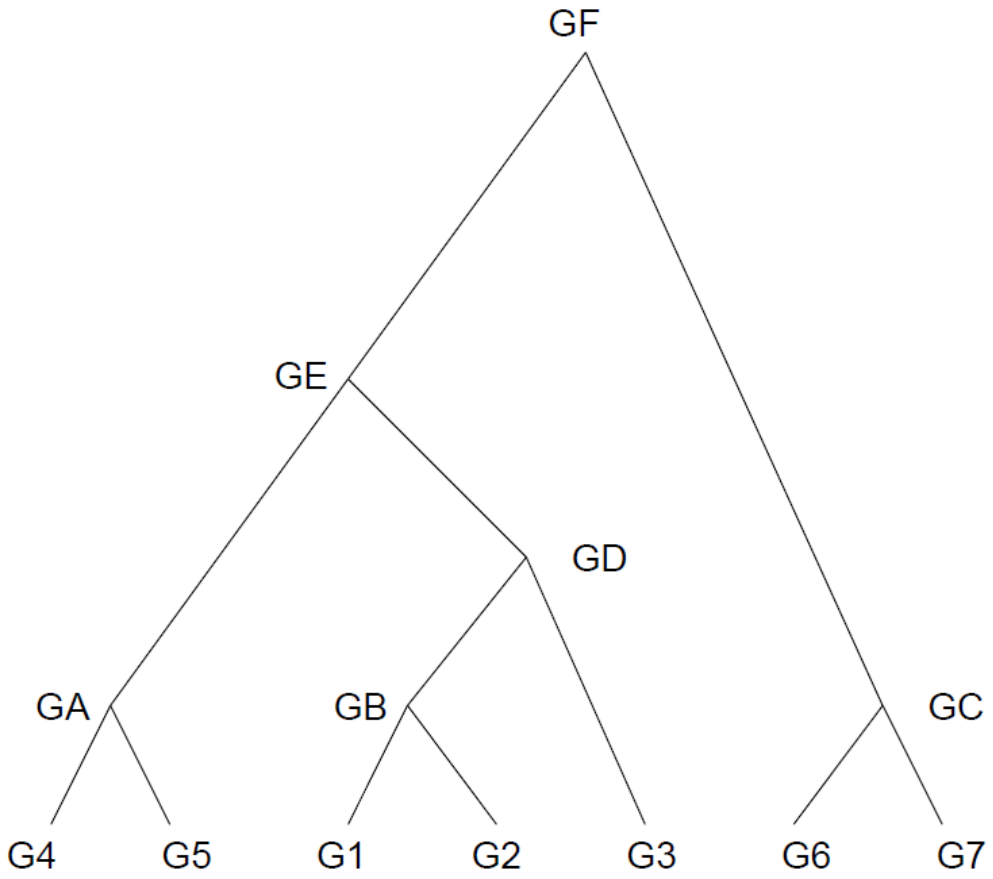
Zhluková analýza

- Cieľom je nájsť rozklad množiny objektov charakterizovanú skupinou premenných na podmnožiny – zhluky
- Objekty v zhluku by mali byť rovnaké, v odlišných zhlukoch by mali byť odlišné
- Počet zhlukov – menší ako počet objektov
- Používajú sa tu hierarchické aglomeratívne metódy – hierarchická postupnosť rozkladov pôvodnej množiny – postupné zlučovanie do väčších zhlukov
- Základ – podobnosť a odlišnosť objektov
- Viacero mier podobnosti a odlišnosti – napr. euklidovská vzdialenosť, jej štvorce a iné
- Objekty a pozorovania sa charakterizujú premennými – spravia sa z nich body na súradnicovej sústave – čí sú bližšie, tým sú podobnejšie

Zhluková analýza

- Postup:
- 1. krok – zoskupia sa 2 najpodobnejšie objekty, ktoré sú najpodobnejšie – tento zhluk sa stáva novým objektom
- Ďalej sa zhluky zhlukujú s inými objektami alebo zhlukmi
- Opakuje sa to, až kým nie je jeden zhluk
- Rôzne typy stratégie


I
II



Dendrogram, linkage tree

Zhluková analýza

- Neexistuje pravidlo na to, koľko je optimálny počet zhlukov
- Ale je možné spraviť analýzu nárastu vnútrozhlukovej variability, resp. sledujeme súčet vnútrozhlukových párových vzdialeností medzi objektami – jeho hodnota s klesajúcim počtom zhlukov vzrastá - najvhodnejšie je to pred najväčším rozdielom hodnôt

Zhluková analýza

- Nedostatky

- Najmä k povahe geografických dát
- Skôr na testovanie hypotézy ako na jej tvorbu
- Ťažkosti s nájdením optimálneho rozkladu vs. a priori určený počet zhlukov
- Vytvorené zhluky nie je možné v ďalších krokoch rozdeliť alebo spájať
- Finálny rozklad pôvodnej množiny – často menšie chyby, nepresnosti

Analýza hlavných komponentov

- Matica dát n x N (n premenných a N pozorovaní)
- PCA ju transformuje na inú maticu dát n x N, kde ale máme N pozorovaní a n nových premenných = komponentov– tie sú lineárnou kombináciou pôvodných premenných
- n ani N sa nezmení
- Komponenty:
  - reprezentujú pôvodný súbor premenných
  - sú navzájom nezávislé
- V súbore premenných začíname hľadaním niečoho čo by sa dalo nazvať priemernou premennou – je najbližšie k ostatným premenným
- Premenné majú jednotkový rozptyl – ten je nositeľom informácie
- Existuje viacero metód extrakcie komponentov – v súčasnosti programy najmä pomocou korelačnej matice

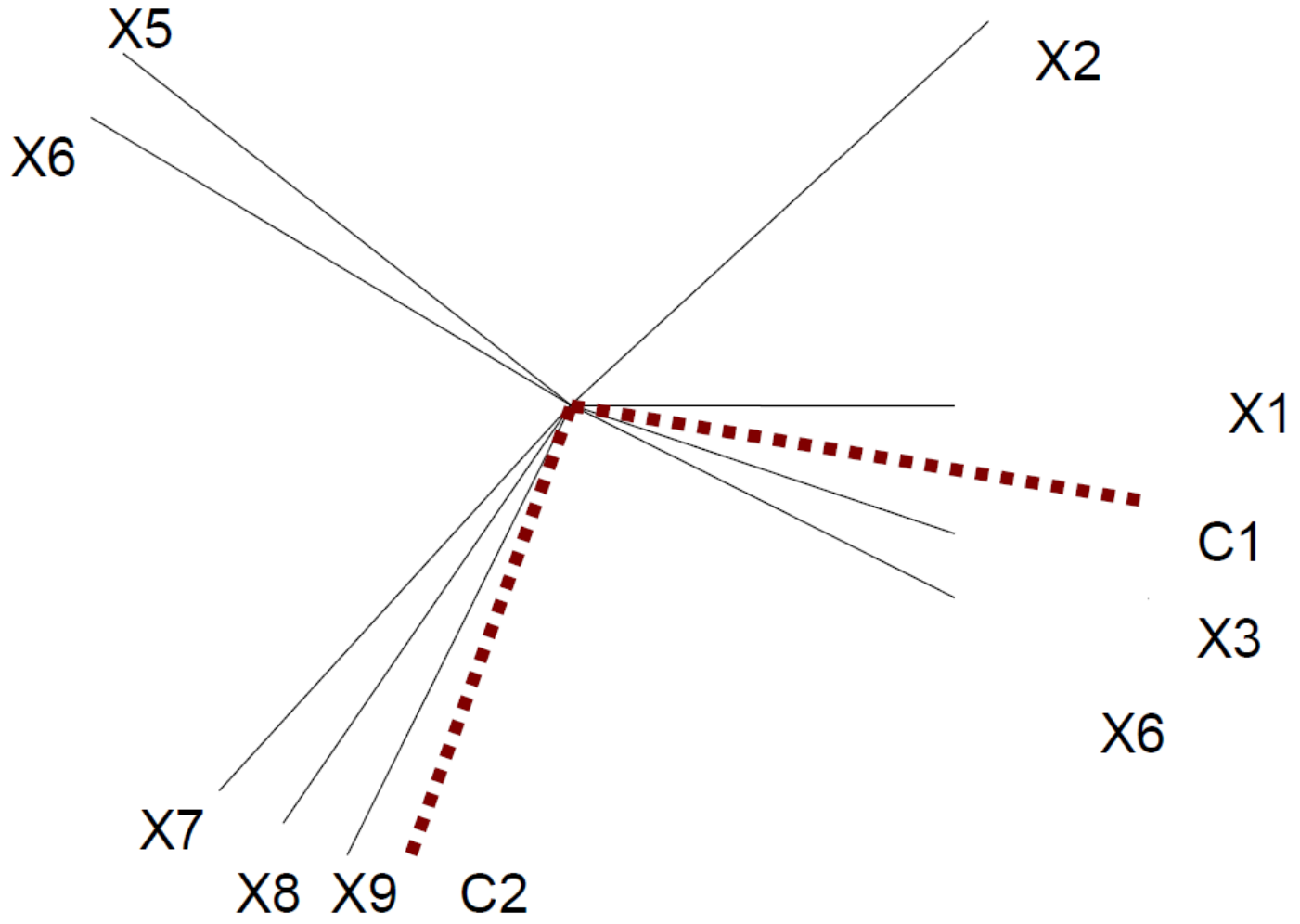
Analýza hlavných komponentov

- Pri vytvorení teda umiestnime novú premennú, ktorá je najbližšie k pôvodným premenným a ich vzťah je daný 3 ukazovateľmi:
  - uhol medzi komponentom a vektorom pôvodnej premennej
  - kosínus tohto uhla – tzv. korelačný koeficient
  - štvorec korelácie – určuje rozptyl tohto koeficientu
- Komponentná záťaž – korelácia medzi premennou a komponentom, existujú aj ich štvorce – tie zobrazujú rozptyl a vyjadrujú ako nový komponent nahrádza pôvodnú premennú
- Suma týchto štvorcov – vyjadruje celkový rozptyl vzťahujúci sa k danému komponentu a označuje sa ako „vlastná hodnota“ (lambda λ)
- Podiel celkového rozptylu vzťahujúceho sa ku komponentu vypočítame potom (λ/n)*100

Analýza hlavných komponentov

- Po vyextrahovaní prvého komponentu extrahujeme druhý – vychádza to už ale z matice, kde sa nenachádza rozptyl pripadajúci na prvý komponent
- Takisto hľadáme čosi ako priemernú hodnotu
- Čím väčší podiel rozptylu vyčerpá prvý komponent, tým menší bude vektor, premenné s nižšou komponentnou záťažou majú väčšie vektory
- Vektor druhého komponentu – bude lokalizovaný tak blízko ako je to možné k reziduálnemu rozptylu a zároveň bude umiestnený kolmo na vektor prvého (resp. pre predchádzajúceho) komponentu – komponenty sú teda navzájom ortogonálne
- Podiel na celkovom rozptyle bude vždy nižší ako u komponentu pred ním
- Pokračuje sa až kým sa nevyčerpá rozptyl a počet komponentov zodpovedá premenným

Analýza hlavných komponentov



Analýza hlavných komponentov

- Komunalita – suma štvorcov komponentných záťaží pre každú premenné a pretože mali jednotkový rozptyl, komunality premenných po PCA by mali byť 1 (prípadne s drobnou odchýlkou)
- Interpetácia sa vytvára hlavne z premenných s vysokými hodnotami komponentných záťaží, čiže hodnoty blízke 1 a -1
- Pri interpretácii komponentnej záťaže sa často zameriavame na premenné, ktoré ju majú nad 0,5 a pod -0,5 a interpretujeme ich ako korelačný koeficient
- Neexistuje pravidlo, koľko komponentov treba na interpretáciu

Analýza hlavných komponentov

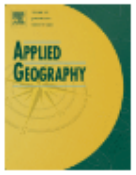
- Najčastejšie používame komponenty s vlastnou hodnotou (lambda) vyššou, rovnou 1
- Ďalšou možnosťou je analýza sutinového grafu (screen plot)
- Komponentné skóre je hodnotou komponentov pre dané pozorovanie a sú to vážené sumárne hodnoty premenných (zvlášť pre každé pozorovanie), váhou je komponentná záťaž
- V prípade pozorovania s vysokou hodnotou premennej s vysokou komponentnou záťažou vytvorí vysokú hodnotu komponentného skóre

Analýza hlavných komponentov

- PCA je vlastne ortogonálna transformácia pôvodného súboru premenných na súbor nových premenných – komponentov, ktoré sú navzájom nezávislé
- Umožňuje identifikovať skupiny súvisiacich premenných
- Umožňuje identifikovať premenné, ktoré sa najviac podieľajú na rozptyle
- Komponentné skóre umožňuje (priestorovú) interpretáciu základných vzorov
- Môže viesť k formulácii nových hypotéz (induktívna metóda)
- Nehodí sa na verifikáciu hypotéz (nerozlišuje spoločný a individuálny rozptyl)

Analýza hlavných komponentov

Applied Geography
Volume 102, January 2019, Pages 47-57



Mapping of climate vulnerability of the coastal region of Bangladesh using principal component analysis

Energy
Volume 160, 1 October 2018, Pages 1030-1046

Social acceptance of green energy determinants using principal component analysis

Case Studies on Transport Policy
Volume 7, Issue 1, March 2019, Pages 73-86

Principal component analysis of driver challenges in the shared taxi market in Ghana