VÁCLAV BEZVODA, TOMÁŠ KUČERA

K MOŽNOSTEM VYUŽITÍ DATAANALYTICKÝCH SYSTÉMŮ V GEOGRAFII

V. Bezvoda, T. Kučera: *On the possibilities of using data-analytic systems in geography.* — Sborník ČSGS, 91, 2, p. 133—139 (1986). — The transition from traditional to modern concepts in geography requires multidimensional data sets to be taken into account. Apart from other methods of corresponding data processing an important role is played by the factor analysis, possibilities and limitations of which are analysed in the article. Attention is also paid to the data-analytical systems (especially BMDP) which make the multivariant methods easy to use.

Pokrok v oblasti výpočetní techniky se v posledních dvou desetiletích promítl do mnoha oborů lidské činnosti; počítače se staly běžnou součástí našeho života. Jejich zpřístupnění širokému okruhu uživatelů poznamenalo vývoj vědních disciplín nejen technických, ale i většiny přírodních a společenských. K těmto disciplínám patří i geografie.

V geografii lze v širokém zavedení počítačů spatřovat (vedle přímého důsledku spojení vnějších technických podmínek s vnitřními potřebami oboru) podstatnou okolnost, která umožnila realizovat některé, do té doby neuskutečnitelné dílčí změny naznačované v průběhu procesu celkové transformace koncepce geografie. To však neznamená, že by počítače svoji úlohu v geografii dostatečnou měrou plnily nebo dokonce splnily. Stále ještě má jejich využívání převážně nejednotnou a extenzívní podobu a další intenzifikace a koncepčnost přístupu v sobě skrývají jednu z významných možností, ne-li přímo podmínku dalšího progresivního vývoje této disciplíny.

Přechod od tradičního k modernímu pojetí geografie představuje v prvé řadě přeorientování zájmu na pravidelnosti struktury a vývoje geografických systémů, což se vzhledem k jejich složitosti dané značnou kvalitativní pestrostí a rozsáhlými koexistenčními souvislostmi elementů odráží ve složitosti řešených problémů. Současná geografie je tak nucena pracovat se značným objemem informací, jejichž zvládnutí je bez počítačů téměř nemožné. Proto počítače nacházejí své uplatnění v celém procesu zpracování geografické informace, při vytváření a uchování datových souborů, jejich analýze, při modelování i při tvorbě kartografické dokumentace.

V této práci se nebudeme zabývat výčtem ani hodnocením použití počítačů v geografii (v této otázce můžeme odkázat na práci Š. Poláčika [4] a další v ní uváděnou literaturu); jde nám spíše o to, poukázat na dosud ne plně využívané možnosti, které skýtá programové vybavení řady dnešních počítačů.

O významu programového vybavení hovoří zejména to, že představuje větší část hodnoty současného standardně vybaveného počítače. Doplňování tohoto vybavení dalšími soubory programů se obvykle rozhodující měrou podílí na rozšiřování možností stávajícího zařízení a často také zvětšuje okruh uživatelů. Postupující standardizací se totiž zjednodušuje komunikace s počítačem, neboť ve stále více případech odpadá svízelné sestavování programů (mnohdy amatérské či poloprofesionální úrovně) a jejich následné odlaďování. Roste třída úloh řešitelných přímo s pomocí základního programového vybavení počítačů.

V geografii se v souvislosti s potřebou zachytit rozmanitost vztahů uvnitř geografických systémů setkáváme velmi často s *mnohorozměrnými soubory* dat. Každému prvku zpracovaného souboru je přiřazen jeden *objekt* — uspořádaná m-tice *veličin (proměnných, souřadnic, vlastností).* Matice, jejíž řádky jsou objekty a sloupce jednotlivé veličiny, je jedinou v této práci uvažovanou formou *matice pozorování.* K plnohodnotnému zpracování takových souborů, často čítajících desítky tisíc položek i více, existuje dnes řada různých metod vzniklých z potřeb mnoha vědních oborů, jako např. medicíny, biologie, ekonomie a sociologie a teoreticky rozpracovaných na pomezí statistiky a aplikované matematiky. Tyto matematicko-statistické metody, které jsou dnes označovány jako *mnohorozměrné metody,* jsou pak pro svoji početní složitost prakticky použitelné pouze za pomoci výkonné výpočetní techniky. Živelné zavádění takovéto techniky však v minulosti v geografii i v dalších oborech, kde znalost matematiky a výpočetní techniky není z pochopitelných důvodů všeobecně na potřebné úrovni, nevedlo k úspěšnému rozvinutí složitého formálního aparátu do aplikační roviny, neboť značné úsilí bylo nutno věnovat studiu teoretických základů jednotlivých metod a budování vlastního programového vybavení. Z tohoto hlediska je kvalitativním krokem vpřed zavedení *dataanalytických systémů,* které kromě toho, že umožňují přímou a širokou aplikaci i velmi složitých metod, jsou většinou budovány právě s ohledem na skutečnost, že uživatel není profesionálním programátorem. Mezi ně patří i u nás dostupné systémy jako SPSS — Statistical Package for Social Sciences, viz práci N. H. Nie (3) a BMDP — Biomedical Package, viz práci W. J. Dixon (1). I když se jedná o systém programů rozdílného zaměření, pokrývají oba prakticky celou problematiku zpracování mnohorozměrných souborů dat a jsou pro geografii stejně dobře použitelné.

Při této příležitosti je nutno upozornit na to, že ani jeden z uvedených systémů neumožňuje zpracování geografických dat v plné míře. Zejména nejsou zařazeny programy sloužící ke zpracování prostorové informace. Práce související s prostorovým rozmístěním zpracovávaných objektů je proto nutno provádět odděleně. Na druhé straně oba systémy se snaží respektovat to, že zpracovávaná data jsou často dosti vzdálená idealizacím, ze kterých metody zpracování většinou vycházejí. Přednost je např. dávána názornému vyjádření vlastností souborů před testováním (to se týká např. odhadu podobnosti rozdělení zpracovávaného souboru s rozdělením normálním). V posledních verzích systému BMDP se objevují aplikace robustních metod, tj. metod, které nejsou citlivé na přítomnost odlehlých hodnot. I tak je však nutno postupovat velmi opatrně. Zkoumané výběrové soubory nepředstavují např. často náhodný výběr ze souboru základního.

Poněkud podrobněji se budeme nyní věnovat dataanalytickému

systému BMDP. Ukážeme si na něm některé konkrétní možnosti využití dataanalytických systémů v geografickém výzkumu. Byl sestaven začátkem sedmdesátých let na Kalifornské univerzitě v Los Angeles týmem odborníků z oboru medicíny, biologie, informatiky a matematické statistiky. Je průběžně doplňován novinkami z oblasti zpracování dat a jeho jednotlivé programy lze nejrůznějším způsobem modifikovat a řetězit. Vzniká tak značný počet možných kombinací, k jejichž zvládnutí je třeba jisté praxe a dobré principiální znalosti jednotlivých metod. Předností systému však je, že respektuje i uživatele—začátečníky, neboť hodnoty většiny parametrů úloh není nutno zadávat; jak vyplyne z dále uvedeného příkladu, systém si je určí v takovém případě sám, a to v jistém smyslu optimálně.

Řídicí jazyk systému BMDP vychází z angličtiny a jazyka FORTRAN. Od uživatele se vedle nejzákladnějších znalostí programování v jazyce FORTRAN a schopnosti orientovat se v manuálech očekává, že má dobrou představu o metodách, které chce použít. Rozumí se tím znalosti podstaty a omezení jednotlivých metod tak, aby byla zajištěna jejich správná aplikace a interpretace získaných výsledků.*)

U řady dalších dataanalytických systémů (např. u zmíněného SPSS) je situace obdobná a příčina naší volby souvisí s tím, že BMDP nám byl snáze dostupný. Je třeba upozornit, že systém BMDP je určen především pro počítače řady IBM 360/370 a zařízení s nimi kompatibilní. Existující verze pro jiné počítače nejsou v ČSSR obecně dostupné. Některé jiné systémy tohoto typu lze sice beze změn implementovat na větší třídě počítačů, vždy však musí jít o dostatečně rychlé zařízení s pamětí řádově 250 KB vybavené výkonnou tiskárnou alespoň se 120 znaky na řádek.

V dalším budeme věnovat pozornost mnohorozměrným metodám, jejichž široká aplikace představuje kvalitativní změnu při zpracování vícerozměrných souborů, zejména *faktorové analýze*. Této více než 70 let staré metodě byla a je věnována stále značná pozornost. Z prací u nás vydaných jmenujeme alespoň práce K. Üblerly [6] a J. Kejkuly [2]. První z nich byla v minulosti často diskutována, mimo jiné také ve spojení s řešením úkolů geografického výzkumu.

Faktorová analýza je technika, která ve své častěji užívané R verzi umožňuje z množství pozorovaných, vzájemně závislých veličin izolovat hypotetické nezávislé (méně často definovaným způsobem závislé) veličiny nazývané *faktory*. Přitom hlavní snahou je reprodukovat s pomocí minimálního počtu faktorů (tedy co nejjednodušeji) informaci obsaženou v původním mnohorozměrném souboru dat při její minimální ztrátě. Toho se využívá při řešení řady obecných úloh, jejichž vzájemné odlišení není vždy zcela jednoznačné. O faktorové analýze se nejčastěji hovoří jako o metodě klasifikace, odhadu zásadních, leč přímo neměřitelných veličin, redukce datových souborů i jako o metodě tvorby hypotéz. Mimo tyto problémové oblasti je známa řada dalších, speciálních aplikací. Předností této metody je mj. i to, že faktory jsou seřazeny podle toho, jakou měrou se formálně podílejí na vysvětlení celkového rozptylu původních proměnných.

---

*) Informace o současném stavu a pravidlech používání BMDP lze získat u gestora systému pro ČSSR, kterým je Středisko biomatematiky Fyziologického ústavu ČSAV.

Z obecné charakteristiky samotné faktorové analýzy i šíře jejího aplikačního prostoru je zřejmé, že se jedná o metodu, s jejíž pomocí lze hledat odpovědi na nejednu z otázek zajímajících geografy. Obsáhlé výčty konkrétních aplikací faktorové analýzy v geografii je možné nalézt např. v pracích V. Toušek, M. Viturka (5) a Š. Poláčik (4).

Zpětný pohled na pronikání faktorové analýzy do geografie nám umožňuje rozlišit dva vývojové směry. Oba vycházely na jedné straně z potřeby řešit v úvodu vzpomenuté problémy a z nedostatečné metodologické saturace ve vztahu k těmto problémům na straně druhé. První směr byl poznamenán individuální snahou geografů o zvládnutí formální stránky metody, které by dovolilo provedení potřebných výpočtů. Složitost tohoto postupu však obyčejně vedla k jeho preferování a vynaložené úsilí zřídkakdy vyústilo v odpovídající aplikaci. Reálnější se proto jevila druhá cesta vývoje, formovaná na základě spolupráce s odborníky na příslušnou negeografickou problematiku. I v případě skupinové práce zůstala aplikace faktorové analýzy ve většině případů spornou záležitostí, neboť vedle postupného zvládání obtíží „technického" rázu nebyla obvykle věnována patřičná pozornost ostatním, věcným problémům. Dosažené výsledky promítnuté na pozadí původních, v mnoha směrech neadekvátních představ o možném přínosu často vedly k následnému odmítání této techniky a nejednou posloužily i k argumentaci proti snahám o kvantitativní přístup v geografii vůbec.

Jak jsme již naznačili, zavedením dataanalytických systémů se podstatně změnily technické podmínky aplikace faktorové analýzy. Přitom však přetrvává neuspokojivé řešení mnohých geografických problémů, v jejichž případě se použití metody zdá být přínosným. Domníváme se proto, že je nutné přezkoumat závěry týkající se použití faktorové analýzy a pokusit se o přístup k těmto otázkám na kvalitativně vyšší úrovni, odpovídající novým podmínkám. Zároveň je nezbytné si uvědomit, že využití programově sebelépe vybaveného počítače nás nezbavuje věcných problémů aplikace a že tedy k základním rysům nového přístupu musí patřit přenesení těžiště zájmu z formální na obsahovou stránku, resp. z výpočtové na přípravnou a interpretační část.

Interpretace numerických výsledků je konečným a zároveň ústředním momentem použití uvedených postupů. Vzhledem ke stavbě faktorové analýzy, kdy všechny kroky v jejím rámci jsou jednoznačně určeny, je právě zpětné nalezení či nenalezení cesty k realitě kritériem úspěšnosti té které konkrétní aplikace a jediným dokladem validity této metody v dané situaci. Přitom interpretace je ovlivněna nejen znalostí problematiky, na níž je metoda použita, ale i znalostí a respektováním příslušných metodologických principů. Práce geografa není tedy zatlačena do pozadí; spíše naopak je rozhojněna aspekty sloužícími ke komplexnějšímu poznání problému. Faktorová analýza vychází z tzv. *matice podobností* (jakožto kvalitativní vlastnosti všech dvojic veličin), zpravidla z některého typu *korelační matice.* Této zvyklosti se přidržíme i my. O úspěchu či neúspěchu aplikace však rozhoduje především vhodnost zpracovávaného souboru, tj. vhodnost výběru objektů a veličin. Kromě splnění samozřejmého požadavku souladu stanovených objektů a veličin s řešenou úlohou se většinou snažíme, aby objekty co nejrovnoměrněji pokrývaly celou zkoumanou oblast. Dále je třeba se vyvarovat takových skupin veličin, jejichž vzájemná korelace vyjádřená v absolutní hodnotě je blízká jedné a které jsou nositelem prakticky stejné informace. Stej-

ně jako v případě zahrnutí známé triviální či indukované závislosti se tak určuje společný výrazný faktor, což má často vliv na potlačení některých jemnějších, z hlediska interpretace důležitých vztahů. Obdobný efekt může mít též heterogenita souboru (jehož část je např. zatížena systematickou chybou), která vede ke vzniku „falešné" korelace mezi veličinami. Z existence těchto úskalí a z požadavku, aby korelační závislosti vyjádřené korelačními koeficienty představovaly skutečné (věcné) závislosti, je zřejmá nezbytnost důkladné logické analýzy jak vstupních dat, tak i všech mezivýsledků a výsledků.

Se snahou o dokonalé postižení skutečnosti souvisí výběr vhodné matice podobnosti. Jedná se o jeden z klíčových a často podceňovaných momentů faktorové analýzy, který může výrazně ovlivnit celkový výsledek a který souvisí s podstatou problémů. Jelikož není formálního charakteru, nemůže být jednoznačně řešen v matematicko-statistické oblasti a tedy ani algoritmem použitým v BMDP. Obvykle volíme mezi parametrickým lineárním Pearsonovým a neparametrickým Spearmanovým (pořadovým) koeficientem korelace. Nutným předpokladem použití Pearsonova koeficientu je lineární vazba mezi soubory dvou sledovaných veličin. Při výrazné nelinearitě se lze pokusit o přiblížení se ke splnění příslušného požadavku pomocí vhodných transformací souborů jednotlivých veličin. Je však třeba si uvědomit, že požadavek m-rozměrné linearity souboru dat, který souvisí s použitím Pearsonova koeficientu korelace při faktorové analýze, nelze v praxi nikdy splnit a jen obtížně lze testovat přiblížení našich dat tomuto ideálu. Přesto je tomuto koeficientu dávána v BMDP jednoznačná přednost. To souvisí s tím, že v případě dodržení alespoň přibližné linearity v rámci souborů jednotlivých proměnných dostaneme většinou vyhovující výsledky při přijatelné spotřebě strojového času i u velkých souborů. Naopak výpočty neparametrických koeficientů korelace jsou z hlediska spotřeby času velmi citlivé na růst počtu prvků. Jestliže se však nepodaří ani zhruba splnit uvedený požadavek pro výpočet parametrických korelačních koeficientů, je nutné použít neparametrický ukazatel zejména tehdy, jsou-li vstupní data zatížena trendem.

Vedle již zmíněných okolností může být výsledek faktorové analýzy ovlivněn také volbou jednotlivých metod řešení dílčích problémů, jako jsou problém použité varianty stanovení komunalit, počtu faktorů, rotace, ortogonality faktorů a problém faktorového skoru. Pro tyto rozhodovací kroky neexistují jednoznačná kritéria; dříve byly určující především technické možnosti, dnes se klade větší důraz na zkušenosti uživatele. Např. v rámci BMDP existují 4 principiálně odlišné metody faktorové analýzy a množství variant. Dostatečné zkušenosti v tomto ohledu však, alespoň v geografii, většinou doposud chybí. Tím víc je nutno cenit výhodu dataanalytických systémů, spočívající v možnosti snadno a rychle zopakovat výpočet za změněných podmínek. Uživatel tak může provést interpretaci na základě porovnání výsledků získaných buď z téhož datového souboru použitím různých dílčích metod, nebo stejným postupem z upraveného souboru dat (čehož lze využít také při hledání konečné podoby datového souboru). Dobrým vodítkem při odhadu správnosti postupu je *stabilita* získaných výsledků. Výsledek je stabilní, je-li málo citlivý na dílčí změny původních dat či zvoleného postupu. Velkou výhodou je pak to, že v případě, kdy nemůžeme zodpovědně zadat parametry určující kterýkoliv základní problém, určí jej systém sám. Vět-

šina klasických problémů faktorové analýzy tak přestává být ve skutečnosti problémem pro uživatele začátečníka i středně pokročilého.

Pro větší názornost uvedeme nyní základní informace o konkrétním použití BMDP při zpracování typické menší úlohy. Soubor 76 územních jednotek ČSR (okresů a hl. m. Prahy) byl posuzován z hlediska 9 číselných ukazatelů. V rámci jedné práce byly postupně aplikovány tři programy:

1. Program BMDP1D, který zajistil vytisknutí matice pozorování a navíc pro všechny veličiny stanovil základní statistické charakteristiky a vypsal největší a nejmenší hodnotu (jako takovou i standardizovanou). Poslední uvedené údaje slouží ke snadné identifikaci odlehlých (výjimečných či chybných) hodnot. Program se používá ke kontrole údajů, popřípadě k zařazení souboru do dataanalytického systému.
2. Program BMDP2D, který slouží ke stanovení statistických charakteristik souborů jednotlivých souřadnic. Pro každou souřadnici je v širší verzi programu vypočteno celkem 27 údajů. Zařazeny jsou i hodnoty zavedené v matematické statistice poměrně nedávno, jmenovitě robustní odhady centrálních hodnot. Navíc je vytisknuta kontingenční tabulka, pseudografickou technikou vykreslen jednoduchý histogram a na ose vyneseny důležité hodnoty příslušného souboru. Program slouží k základní orientaci o typech rozdělení v rámci jednotlivých veličin. To je důležité zejména u souborů menších, kde je často názorná představa cennější než statistický ukazatel.
3. Program BMDP4M realizující faktorovou analýzu. Tímto programem se budeme zabývat poněkud podrobněji.
   Jesliže předpokládáme, že data musíme zadat spolu s popisem úlohy, bude sestava štítků mít ve velmi jednoduchém případě např. následující tvar:

```
//NOKRESY JOB OAD-52,SILAR,CLASS=C,TIME=1
// EXEC BIMED,PROG=BMDP4M
//SYSIN DD *
/ PROBLEM
TITLE IS 'CHARAKTERISTIKA OKRESU CSR'.
/ INPUT
FORMAT IS '(A3, 9F 7.0)'.
VARIABLES ARE 10.
/ VARIABLE
  LABEL = 1.
  NAME=ZNAC, HUST, DETI, EKA, PRIM, SEC, MEST, BYT, FEM, DUCH.
  USE=3 TO 10.
/ END
  data
/*
//
```

Rolí řídících štítků se zabývat nebudeme. V řazení štítků řídícího jazyka BMDP se vstupními informacemi existuje v zásadě pevný řád a při změně vstupních dat je běžné měnit pouze příslušné číselné hodnoty, popřípadě název úlohy apod.

Význam jednotlivých *odstavců* (uvedených /) a k nim případně příslušných *příkazů* (ukončených tečkou) je celkem zřejmý. První odstavec slouží k pojmenování úlohy a příslušný příkaz není povinný. Druhý odstavec charakterizuje vstupní data. Povinný příkaz FORMAT se zapisuje tak, jak stanoví norma FORTRANU IV. Také příkaz definující počet proměnných je povinný. Následuje nepovinný odstavec VARIABLE,

jehož 3 nepovinné příkazy postupně zajišťují to, že první proměnná se chápe jako označení, že všechny proměnné mají svá jména a konečně, že faktorová analýza bude provedena na základě 3. až 10. veličiny. Veličina označená HUST má jako jediná rozdělení blízké lognormálnímu a byla v této verzi ze zpracování vyloučena.

Je zřejmé, že všechny parametry vlastního výpočtu faktorové analýzy jsou v tomto případě zadávány systémem. To konkrétně znamená, že maticí podobností je matice lineárních koeficientů korelace, zvolena je metoda hlavních komponent, přičemž jako faktory jsou brány ty komponenty, které odpovídají vlastním číslům větším než jedna; komunality pak odpovídají čtvercům množinových koeficientů korelace jednotlivých veličin ve vztahu k uvedeným faktorům. V rámci prostoru definovaného faktory je nakonec provedena ortogonální rotace (metodou Varimax).

Výpočet podle všech tří uvedených programů zadaný jako jediná práce trval celkem 35,2 s při využití 134 KB operační paměti, přičemž byla potištěna 31 strana tabulek, pseudografů a vysvětlujících textů. Připojením dalších programů a volbou parametrů jednotlivých úloh jakož i růstem matice pozorování množství informací dále narůstá. Také z toho přirozeně vyplývá, že vyhodnocení všech výstupních údajů zejména v případě, kdy uživatel není příliš zkušený, je mnohdy téměř neřešitelný problém. I na to však lze najít odpověď v manuálu k systému BMDP. Zní v překladu zhruba takto: „Programy BMDP se užívají snadno, jestliže ignorujeme to, co nepotřebujeme vědět."

Autoři považují za svou milou povinnost poděkovat dr. T. Havránkovi, CSc., pracovníku Střediska biomatematiky Fyziologického ústavu ČSAV, za pomoc, kterou poskytl na všech úrovních, které souvisely se vznikem tohoto článku.

Literatura:

1. DIXON, W. J.: Biomedical Computer Programs. Los Angeles, University of California 1972, 792 s.
2. KEJKULA, J.: Základy faktorové analýzy. In: Statistická revue 5, Praha, VÚSEI při FSÚ 1977, s. 107—132.
3. NIE, N. H., BRENT, D. H., HULL, C. H.: Statistical Package for the Social Sciences. 2. ed. New York, McGraw Hill 1975, 675 s.
4. POLÁČIK, Š.: Hlavné smery vo využívaní samočinných počítačov v geografii. Studia Geographica, 74, Brno, GGÚ ČSAV 1982, s. 1—209.
5. TOUŠEK, V., VITURKA, M.: Metoda faktorové analýzy a její aplikace ve výzkumu prostorových struktur. Zprávy Geografického ústavu ČSAV, 16, Brno, GGÚ ČSAV 1979, s. 132—148.
6. ÜBERLA, K.: Faktorová analýza. Bratislava, Alfa 1974, 334 s.

(Pracoviště autorů: V. Bezvoda — katedra aplikované matematiky přírodovědecké fakulty UK, Albertov 6, 128 43 Praha 2; T. Kučera — Geografický ústav ČSAV, Mendlovo nám. 1, 662 82 Brno.)
Došlo do redakce 29. 10. 1985.