GEOGRAFICKÝ ČASOPIS

ROČNÍK 44 1992 ČÍSLO 2

ANTON BEZÁK*

# KOMPARATÍVNA ANALÝZA NEHIERARCHICKÝCH METÓD REGIONÁLNEJ TAXONÓMIE (RASTOVÉ TYPY OBCÍ V GEMERI AKO PRÍKLAD)

Anton Bezák: Comparative Analysis of Non-Hierarchic Regional Taxonomic Methods (The Growth Types of Communes in the Gemer Region as an Example). Geogr. Čas., *44*, 1992, 2, 1 fig., 19 refs.

The main objective of this paper is a quantitative assessment of three non-hierarchic regional taxonomic procedures by applying them to the same data set consisting of 200 communes and eight growth attributes.

V poslednom období sa v regionálnej taxonómii začína prejavovať zvýšený záujem o nehierarchické metódy. Pozornosť, ktorú vzbudzuje táto kategória taxonomických procedúr, nie je vôbec náhodná. Nehierarchické metódy sú v prvom rade mimoriadne vhodné na riešenie rozsiahlych taxonomických problémov zahrňujúcich stovky alebo dokonca tisícky priestorových jednotiek. V porovnaní s hierarchickými metódami majú omnoho menšie požiadavky na výpočtový čas a kapacitu pamäte počítača. Ich numerická efektívnosť spočíva aj v tom, že pre vopred určený počet tried umožňujú nájsť triedenie, ktoré je v určitom zmysle optimálne.

Rastúca popularita nehierarchických metód vyvoláva, prirodzene, celý rad principiálnych otázok metodologickej povahy. Jednou z nich je aj otázka relatívnej kvality a účinnosti rôznych nehierarchických procedúr. Cieľom tejto štúdie je empiricky zhodnotiť tri nehierarchické metódy na základe dát opisujúcich populačný vývoj 200 gemerských obcí v období 1971-1990.

## NEHIERARCHICKÉ METÓDY REGIONÁLNEJ TAXONÓMIE

Prehľad nehierarchických metód začneme formuláciou špecifického problému regionálnej taxonómie, ktorý bude základom ich stručnej prezentácie i empirického zhodnotenia. Predpokladajme, že je daná množina pozostávajúca z $n$ priestorových jednotiek, ktoré sú charakterizované súborom $p$ atribútov. Hodnoty atribútov pozorované v jednotlivých priestorových jednotkách môžeme usporiadať do tzv. matice atribútov $\mathbf{X} = [x_{ij}]$ typu $n \times p$, kde $x_{ij}$ označuje hodnotu $j$-tého atribútu v $i$-tej priestorovej jednotke. V regionálnej taxonómii je zvykom zobrazovať priestorové jednotky ako body v $p$-rozmernom priestore atribútov tak, že súradnicami každého bodu sú hodnoty jednotlivých atribútov v korešpondujúcej priestorovej jednotke. Geometrická blízkosť dvoch alebo viacerých bodov v tomto priestore potom svedčí o podobnom charaktere zodpovedajúcich priestorových jednotiek. Ak sú všetky atribúty vyjadrené v rovnakých jednotkách

* Geografický ústav SAV, Štefánikova 49, 814 73 Bratislava.

merania a navyše sú vzájomne nekorelované, vhodnou mierou podobnosti (presnejšie nepodobnosti) dvoch priestorových jednotiek $i$ a $j$ je euklidovská vzdialenosť $d(X_i, X_j)$ dvoch korešpondujúcich bodov $X_i$ a $X_j$, definovaná formulou

$$d(X_i, X_j) = \left[ \sum_{m=1}^{p} \left( x_{im} - x_{jm} \right)^2 \right]^{1/2} . \tag{1}$$

Poznamenajme ešte, že v záujme jednoduchej prezentácie regionálno-taxonomických problémov sa priestorové jednotky a ich geometrické obrazy veľmi často stotožňujú.

Úlohou nehierarchickej regionálnej taxonómie je nájsť optimálny rozklad množiny $n$ priestorových jednotiek na $k$ disjunktných a neprázdnych tried ($k < n$) tak, aby vzhľadom na súbor $p$ atribútov priestorové jednotky patriace do tej istej triedy boli v určitom zmysle podobné a priestorové jednotky patriace do rôznych tried boli odlišné. Takto definované triedy priestorových jednotiek sa nazývajú regionálne triedy alebo regionálne taxóny. Požiadavka vnútrotriednej podobnosti je známa ako princíp internej homogenity a požiadavka medzitriednej odlišnosti ako princíp externej separácie (cf. [8], [10]).

Uvedená formulácia nehierarchického problému regionálnej taxonómie vyžaduje niekoľko doplňujúcich poznámok. V prvom rade si všimnime, že formulácia neobsahuje žiadnu podmienku susedstva priestorových jednotiek zaradených do jednej a tej istej triedy. Pripomíname, že priestorovo nesúvislé regionálne triedy sa nazývajú regionálne typy a proces ich identifikácie je známy ako regionálna typizácia. V našom prípade ide navyše o tzv. formálnu regionálnu typizáciu, pretože súbor atribútov vyjadruje vlastnosti priestorových jednotiek a nepostihuje vzťahy medzi nimi. Napokon ešte upozorníme na predpoklad disjunktných, t.j. vzájomne sa neprekrývajúcich regionálnych tried v tom zmysle, že žiadna priestorová jednotka nesmie patriť do dvoch alebo viacerých regionálnych tried. V súhrne sa náš problém dá označiť ako nehierarchický problém formálnej regionálnej typizácie s disjunktnými regionálnymi triedami.

Charakter rozkladu množiny priestorových jednotiek na $k$ regionálnych tried $R_1, R_2, \ldots, R_k$ sa v kontexte nehierarchickej taxonómie posudzuje na základe tzv. funkcionálov kvality rozkladu, definovaných na množine všetkých možných rozkladov danej množiny na určený počet tried. Takmer všetky funkcionály kvality, uvádzané v regionálno-taxonomickej literatúre (cf.[11, pp. 166-174]), vychádzajú buď z princípu internej homogenity alebo z princípu externej separácie. Najznámejší z nich je funkcionál súčtu vnútrotriednych rozptylov, definovaný vzťahom

$$Z = \sum_{i=1}^{k} \sum_{X_j \in R_i} d^2 \left( X_j, \overline{X}_i \right) , \tag{2}$$

kde $k$ označuje počet tried, $X_j$ bod v priestore atribútov reprezentujúci $j$-tu priestorovú jednotku, $\overline{X}_i$ ťažisko (priemer) triedy $R_i$ a $d^2$ štvorec euklidovskej vzdialenosti bodov uvedených v zátvorkách. Pod optimálnym rozkladom potom rozumieme rozklad, pri ktorom funkcionál $Z$ nadobúda minimálnu hodnotu.

Nemusíme azda ani zdôrazňovať, že žiadny problém nehierarchickej taxonómie nie je možné riešiť tak, že postupne zostrojíme a zhodnotíme všetky možné rozklady množiny $n$ priestorových jednotiek na $k$ regionálnych tried. Hlavným dôvodom konštrukcie rôznych nehierarchických procedúr je preto úsilie nájsť efektívne algoritmy umožňujúce zmenšiť počet rozkladov, ktoré je potrebné preskúmať. Väčšina nehierarchických procedúr začína stanovením alebo odvodením nejakého začiatočného rozkladu skúmanej množiny na určený počet tried. Začiatočný rozklad sa potom iteratívne zlepšuje premiestňovaním priestorových jednotiek z triedy do triedy podľa

určitého pravidla dovtedy, kým sa nezistí, že žiadne ďalšie zlepšenie sa už použitou metódou nedá dosiahnuť. Takto získaný rozklad sa nazýva finálnym rozkladom. Je zrejmé, že opísaný postup vo všeobecnosti nemusí viesť k absolútnemu, ale iba k lokálnemu extrému funkcionálu kvality.

Medzi najznámejšie nehierarchické procedúry, ktoré našli uplatnenie v regionálnej taxonómii, môžeme v zhode s M.M. Fischerom ([8], [11, pp. 176-183]) zaradiť metódu rozkladu s minimálnymi vzdialenosťami, metódu $k$ priemerov a tzv. horolezeckú metódu.

Najjednoduchšou nehierarchickou procedúrou je metóda rozkladu s minimálnymi vzdialenosťami (*minimum distance partition method*, v skratke metóda MDP). Jeden z variantov tejto metódy, známy pod názvom Forgyho procedúra, môžeme opísať ako postupnosť nasledujúcich štyroch krokov:

Krok 1. Zvolíme začiatočný rozklad množiny priestorových jednotiek na $k$ ( $k < n$ ) neprázdnych a disjunktných regionálnych tried $R_1, R_2, \ldots, R_k$.

Krok 2. Vypočítame ťažiská (priemery) $\overline{X}_i$ ($i = 1, 2, \ldots, k$) jednotlivých tried začiatočného rozkladu.

Krok 3. Postupne preskúmame každú priestorovú jednotku $X_m$ ($m = 1, 2, \ldots, n$). Vypočítame euklidovské vzdialenosti $d(X_m, \overline{X}_i)$ pre $i = 1, 2, \ldots, k$ a priradíme $X_m$ k triede $R_j$, pre ktorú platí

$$d(X_m, \overline{X}_j) = \min_i \left\{ d\left(X_m, \overline{X}_i\right) \right\}. \tag{3}$$

Postupným premiestňovaním priestorových jednotiek k najbližším ťažiskám vytvoríme nový rozklad množiny priestorových jednotiek, v ktorom určíme nové ťažiská $\overline{X}_i$.

Krok 4. Predchádzajúci krok opakujeme dovtedy, kým sa nedosiahne stabilná konfigurácia priestorových jednotiek, pri ktorej v priebehu jedného cyklu žiadna priestorová jednotka nezmení svoju triednu príslušnosť.

Metóda $k$ priemerov (*k-means method*, v skratke metóda KM) je tiež v podstate skupina procedúr, z ktorých si pozornosť zasluhuje tzv. konvergentný variant MacQueenovej procedúry. Od predchádzajúceho variantu metódy MDP sa líši iba v tom, že v kroku 3 sa nové súradnice ťažísk zväčšenej i zmenšenej triedy vypočítajú po každom premiestnení priestorovej jednotky. Iteračný cyklus sa v tomto kroku ukončí preskúmaním všetkých priestorových jednotiek. Dá sa ľahko dokázať (cf. [1]), že vzhľadom k známej vlastnosti priemeru (súčet štvorcov odchýlok od priemeru je minimálny) sa v rámci oboch spomenutých procedúr dosiahne minimálny celkový súčet štvorcov odchýlok od triednych priemerov. To však znamená, že v obidvoch procedúrach dochádza (implicitne) k minimalizácii funkcionálu kvality $Z$, definovaného rovnicou (2).

Pravdepodobne najefektívnejšou nehierarchickou procedúrou v regionálnej taxonómii je tzv. horolezecká metóda (*hill-climbing method*, v skratke metóda HC), ktorá v porovnaní s metódami MDP a KM explicitne minimalizuje funkcionál súčtu vnútrotriednych rozptylov. V rámci tejto metódy sa v priebehu jedného iteračného cyklu (krok 3) každá priestorová jednotka postupne premiestňuje z pôvodnej triedy do každej z ostatných tried. Pritom sa sleduje, či jej zaradenie do niektorej inej triedy nespôsobí pokles hodnoty funkcionálu $Z$. Definitívny presun priestorovej jednotky $X_m$ z triedy $R_i$ do triedy $R_j$ sa uskutoční vtedy, keď sú splnené nasledujúce dve podmienky. Po prvé, ak platí nerovnosť

$$\frac{n_j}{n_j + 1} d^2(X_m, \overline{X}_j) < \frac{n_i}{n_i - 1} d^2(X_m, \overline{X}_i), \tag{4}$$

kde $n_i$ (resp. $n_j$) je počet priestorových jednotiek v triede $R_i$ (resp. $R_j$) a po druhé, ak toto premiestnenie vedie k maximálnemu poklesu hodnoty funkcionálu $Z$ v rámci všetkých možných presunov jednotky $X_m$, ktoré vyhovujú nerovnosti (4). Po premiestnení jednotky $X_m$ sa vytvorí

nový rozklad množiny priestorových jednotiek a vypočítajú sa nové súradnice ťažísk v triedach $R_i$ a $R_j$. Ak sa pri presunoch jednotky $X_m$ nedosiahne zníženie hodnôt funkcionálu $Z$, pôvodné začlenenie tejto jednotky sa nezmení a do úvahy sa vezme ďalšia priestorová jednotka. Postupnosť iteračných cyklov končí dosiahnutím stabilnej konfigurácie priestorových jednotiek.

Na záver diskusie o metódach nehierarchickej taxonómie pripojíme ešte niekoľko odkazov na literatúru. Podrobné informácie o všetkých troch nehierarchických procedúrach poskytuje každá spoľahlivá príručka zhlukovej analýzy. Osobitne v tejto súvislosti upozorňujeme na monografie M.R. Anderberga [1], J.A. Hartigana [13] a H. Spätha [18]. Z českej literatúry si pozornosť zasluhuje práca A. Lukasovej a J. Šarmanovej [16], ktorá podáva dobrý prehľad metód MDP a KM. Veľmi stručne sa o obidvoch metódach zmieňujú aj P. Hebák a J. Hustopecký [14] vo svojej učebnici viacrozmerných štatistických metód. Ako príklady využitia nehierarchických metód v regionálno-taxonomickom kontexte môžeme spomenúť príspevky D. Barkera [4], J.H. Baumana, M.M. Fischera a U. Schuberta [5], H. Blumeho a R. Schwarza [7], M.M. Fischera a W. Drotha [12], B. Kostrubca [15] a W. Taubmanna [19]. Slovenskú geografickú literatúru zastupuje zatiaľ štúdia [6]. Napokon ešte uvedieme, že problémom komparatívneho hodnotenia nehierarchických procedúr sa veľmi podrobne a systematicky zaoberá M.M.Fischer v prácach [8], [10] a [12, pp. 186-204], ktoré nás inšpirovali aj k napísaniu tejto štúdie.

## EMPIRICKÉ POROVNANIE METÓD MDP, KM A HC

Komparatívna analýza troch nehierarchických procedúr sa v našej štúdii opiera o súbor dát opisujúci priestorové diferencie oboch zložiek populačného rastu na území Gemera v období 1971-1990. Množinu priestorových jednotiek tvorí 122 obcí okresu Rimavská Sobota a 78 obcí okresu Rožňava. Aby sa vylúčil vplyv územných zmien, ktoré znemožňujú porovnanie populačných dát v priebehu pozorovaného obdobia, všetkých 200 obcí sa uvažuje v administratívnych hraniciach ku dňu sčítania ľudu v roku 1980. Pozorované obdobie, pokrývajúce časový úsek od 1.1.1971 do 31.12.1990, je pre potreby našej analýzy rozdelené na štyri päťročné intervaly. V každom intervale sa intenzita populačného rastu meria pomocou dvojice charakteristík vyjadrujúcich prirodzený a migračný prírastok (resp. úbytok) vo forme priemerných ročných mier na 1000 obyvateľov stredného stavu. Dáta o oboch zložkách populačného rastu pochádzajú z priebežnej registrácie pohybu obyvateľstva podľa obcí za každý rok pozorovaného dvadsaťročia. Každú priestorovú jednotku (obec) teda charakterizuje osem rastových atribútov. V snahe vylúčiť vplyv rôznej variability jednotlivých atribútov, všetky stĺpcové vektory matice atribútov typu 200 x 8 sa normovali tak, aby mali jednotkovú normu.

Jedným z kľúčových problémov nehierarchickej taxonómie je určenie vhodného počtu regionálnych tried. Ak hodnota $k$ nie je stanovená a priori ako súčasť formulácie taxonomického problému, potom sa jej voľba môže opierať buď o znalosti vecnej povahy skúmaného problému, alebo o určité formálne kritériá. Keďže náš príspevok sleduje v prvom rade metodologické ciele, dali sme prednosť druhému spôsobu. Použili sme tzv. stopové kritérium [9, p. 60], založené na porovnaní stopy tr (**W**) matice vnútrotriednej variability a stopy tr (**B**) matice medzitriednej variability. Podľa tohto kritéria sa odporúča vybrať také $k$, pre ktoré je hodnota

$$M(k) = \frac{n-k}{k-1} \frac{\mathrm{tr}\left(\mathbf{B}\right)}{\mathrm{tr}\left(\mathbf{W}\right)} \tag{5}$$

maximálna.

V našej štúdii sme pri použití stopového kritéria zvolili nasledovný postup. Aplikáciou metódy HC na začiatočné rozklady generované pseudonáhodným spôsobom sme najprv získali niekoľko finálnych rozkladov skúmanej množiny na rôzny počet regionálnych tried. Pre každý finálny

rozklad sme potom vypočítali hodnotu $M(k)$ podľa rovnice (5). Na základe porovnania hodnôt stopového kritéria sme sa napokon rozhodli pre sedem regionálnych tried.

Druhou, nemenej závažnou otázkou nehierarchickej taxonómie je stanovenie začiatočných rozkladov skúmanej množiny priestorových jednotiek na určený počet tried. Analogicky ako v prípade výberu vhodnej hodnoty $k$, aj tu môžeme vychádzať buď z empirických poznatkov alebo z formálnych pravidiel. Využívajúc obidve možnosti, odvodili sme štyri odlišné začiatočné rozklady, označené symbolmi $P_1$, $P_2$, $P_3$ a $P_4$. Rozklad $P_1$ má charakter pseudonáhodného rozkladu. Je vytvorený na základe predpisu $j = \mathrm{mod}\ (i - 1, k) + 1$, kde $j$ značí index triedy, do ktorej je zaradená $i$-ta priestorová jednotka, $k$ je počet tried a mod ( $a,b$ ) označuje zvyšok pri delení čísla $a$ číslom $b$ [18]. Je zrejmé, že za predpokladu náhodného usporiadania priestorových jednotiek v rámci dátového súboru, rozklad $P_1$ evidentne preukazuje určité prvky náhodnosti. Poznamenávame, že sériu pseudonáhodných rozkladov uvedeného typu pre rôzne hodnoty $k$ sme použili aj pri hľadaní vhodného počtu tried pomocou stopového kritéria.

Na rozdiel od rozkladu $P_1$ ostatné tri druhy začiatočných rozkladov odzrkadľujú určité vysvetľujúce hypotézy týkajúce sa priestorovej diferenciácie rastového procesu v skúmanom území. Rozklad $P_2$ sa zhoduje s rozdelením gemerských obcí na sedem veľkostných kategórií podľa počtu ich obyvateľov k 1.11.1980. Hranice veľkostných intervalov pritom tvoria hodnoty 200, 500, 1000, 2000, 5000 a 10 000 osôb. Pri odvodení rozkladov $P_3$ a $P_4$ sme využili dve typizácie slovenských obcí, ktoré na základe údajov zo sčítania ľudu v roku 1970 spracoval J. Verešík pri tvorbe Atlasu SSR ([2, pp. 140 a 170], [3, pp. 84 a 94]). Rozklad $P_3$ vznikol úpravou typizácie obcí podľa podielu hlavných vekových skupín tak, že z najpočetnejšej kategórie gemerských obcí (obce s nadpriemerným podielom osôb vo veku 60 a viac rokov a s podpriemerným podielom detskej a strednej vekovej skupiny) sa vyčlenil osobitný typ obcí, v ktorých podiel najstaršej vekovej skupiny prekračuje hodnotu 22%. Odvodenie rozkladu $P_4$ sa opiera o veľmi podrobnú funkčnú typizáciu J. Verešíka, ktorá sa pre potreby našej analýzy zjednodušila v tom zmysle, že postupným spájaním príbuzných typov sa vymedzili tri triedy monofunkčných obcí (obce s dominantným zastúpením poľnohospodárstva a lesníctva, priemyslu a stavebníctva, dopravy a služieb), tri prechodné typy dvojfunkčných obcí a jeden zmiešaný typ polyfunkčných obcí.

Hlavné výsledky aplikácie troch nehierarchických procedúr na každý zo štyroch začiatočných rozkladov sú uvedené v tabuľke 1. Okrem hodnoty funkcionálu $Z$, definovaného rovnicou (2),

Tab. 1. Zhodnotenie kvality finálnych rozkladov

| Metóda | Začiatočný rozklad | Funkcionál $Z$ | Podiel vnútrotriedneho rozptylu |
|---|---|---|---|
| MDP | $P_1$ | 2,9560 | 52,05 |
| | $P_2$ | 2,9161 | 51,34 |
| | $P_3$ | 2,8980 | 51,02 |
| | $P_4$ | 2,9593 | 52,10 |
| KM | $P_1$ | 2,8971 | 51,01 |
| | $P_2$ | 2,9876 | 52,60 |
| | $P_3$ | 2,9153 | 51,33 |
| | $P_4$ | 2,9285 | 51,56 |
| HC | $P_1$ | 2,8781 | 50,67 |
| | $P_2$ | 2,9023 | 51,10 |
| | $P_3$ | 2,8853 | 50,80 |
| | $P_4$ | 2,8625 | 50,40 |

tabuľka udáva ešte percentuálny podiel vnútrotriedneho rozptylu na celkovom rozptyle v množine priestorových jednotiek, ktorý je pre všetkých 12 finálnych rozkladov konštatný. Ako sa dalo očakávať na základe teoretických úvah z predchádzajúceho odseku, metóda HC poskytuje pre všetky druhy začiatočného rozkladu jednoznačne najlepšie výsledky v zmysle minimalizácie funkcionálu $Z$. Prekvapuje však, že metódou MDP sa v prípade dvoch začiatočných rozkladov ($P_2$ a $P_3$) neočakávane dosiahli lepšie výsledky ako metódou KM. Ako vidieť, najlepšie riešenie sa získalo aplikáciou metódy HC na začiatočný rozklad $P_4$, naproti tomu najhoršie výsledky priniesla metóda KM pre začiatočný rozklad $P_2$.

Z údajov tabuľky 1 ďalej vyplýva, že vo všeobecnosti nie je možné jednoznačne určiť, aký druh začiatočného rozkladu vedie pri aplikácii rôznych metód k najlepším výsledkom. Pozoruhodné je okrem toho aj zistenie, že začiatočné rozklady $P_2$, $P_3$ a $P_4$, ktoré odzrkadľujú určité hypotézy týkajúce sa priestorovej diferenciácie populačného rastu v skúmanom území, neprinášajú v porovnaní s pseudonáhodným rozkladom $P_1$ žiadne výrazné zníženie hodnoty funkcionálu Z. Ako relatívne málo účinný sa prejavuje najmä začiatočný rozklad $P_2$ odrážajúci rozdelenie gemerských obcí podľa veľkosti. Na základe uvedených poznatkov môžeme preto konštatovať, že výber nehierarchickej procedúry má v našej štúdii závažnejšie dôsledky na riešenie taxonomického problému ako voľba začiatočného rozkladu.

Napriek relatívne malým rozdielom v celkovej účinnosti jednotlivých taxonomických riešení môžu existovať značné diferencie v ich priestorovej konfigurácii. Z tohto dôvodu je vhodné vzájomne porovnať všetkých 12 finálnych rozkladov pomocou jednoduchého koeficientu konkordancie $c$, ktorý navrhol W.M. Rand [17]. Konkordancia $c\,(P', P'')$ dvoch rozkladov $P'$ a $P''$ je definovaná formulou

$$c\,(P', P'') = \left\{ \binom{n}{2} - \left[ \frac{1}{2} \left( \sum_{i=1}^{k} \left( \sum_{j=1}^{k} n_{ij} \right)^2 + \sum_{j=1}^{k} \left( \sum_{i=1}^{k} n_{ij} \right)^2 \right) - \sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij}^2 \right] \right\} \Big/ \binom{n}{2}, \qquad (6)$$

kde $P'$ a $P''$ sú dva finálne rozklady množiny $n$ priestorových jednotiek na $k$ disjunktných a neprázdnych tried a $n_{ij}$ je počet priestorových jednotiek v $i$-tej triede rozkladu $P'$ a v $j$-tej triede rozkladu $P''$($i,j = 1, 2, \ldots, n$). Je zrejmé, že pre hodnoty koeficientu $c$ platí nerovnosť $0 < c \leq 1$, pričom $c = 1$ pri úplnej zhode oboch rozkladov.

Pri posudzovaní stupňa zhody dvoch rozkladov pomocou koeficientu konkordancie je však nevyhnutné vedieť, ako sa vypočítaná hodnota $c$ odlišuje od korešpondujúcej minimálnej hodnoty. M.M.Fischer [10] totiž upozornil, že ak hodnota $c_{\min}$ nie je známa, potom prakticky každú vypočítanú hodnotu $c$ môžeme interpretovať ako určitý prejav zhody dvoch rozkladov, a to aj v prípade, keď $c = c_{\min}$. V súvislosti s tým odporúča transformovať hodnoty $c$ do intervalu $\langle 0,1 \rangle$. Transformované hodnoty koeficientu konkordancie $\hat{c}$ sú potom definované vzťahom

$$\hat{c} = \frac{c - c_{\min}}{1 - c_{\min}}. \qquad (7)$$

Hodnotu $c_{\min}$, vyjadrujúcu maximálny stupeň nezhody dvoch finálnych rozkladov, určíme tak, že v rovnici (6) pozorované hodnoty $n_{ij}$ nahradíme korešpondujúcimi teoretickými hodnotami $\hat{n}_{ij}$, ktoré sa vypočítajú ako súčin príslušných marginálnych súčtov matice $\mathbf{N} = [n_{ij}]$ delený hodnotou $n$.

Hodnoty $\hat{c}$ pre všetky možné dvojice finálnych rozkladov sú uvedené v tabuľke 2. Ako vidieť, stupeň zhody medzi rôznymi riešeniami taxonomického problému nie je vo všeobecnosti vysoký, pretože koeficient $\hat{c}$ veľmi citlivo reaguje aj na malé diferencie medzi finálnymi rozkladmi. Je zaujímavé, že najvyšším stupňom vzájomnej zhody sa vyznačuje trojica riešení, ktoré sú odvodené

Tab. 3. Priemerné hodnoty prirodzeného a migračného prírastku (v ‰) v pozorovaných päťročných intervaloch pre sedem regionálnych typov a celú množinu obcí Gemera

| Regionálny typ | Prirodzený prírastok | | | | Migračný prírastok | | | |
|---|---|---|---|---|---|---|---|---|
| | 1971 – 75 | 1976 – 80 | 1981 – 85 | 1986 – 90 | 1971 – 75 | 1976 – 80 | 1981 – 85 | 1986 – 90 |
| $R_1$ | 2,5 | 3,6 | – 0,1 | – 1,3 | – 8,1 | – 12,3 | – 12,9 | – 4,6 |
| $R_2$ | 1,7 | – 0,3 | 2,5 | 5,4 | – 18,8 | – 18,9 | – 28,2 | – 40,3 |
| $R_3$ | 1,8 | 0,6 | – 0,8 | – 1,0 | – 61,4 | – 61,9 | – 50,1 | – 30,7 |
| $R_4$ | – 4,2 | – 9,3 | – 10,5 | – 16,1 | – 11,9 | – 10,6 | – 21,8 | – 17,3 |
| $R_5$ | 6,8 | 8,3 | 8,9 | 8,5 | – 0,4 | 1,3 | – 0,6 | – 2,1 |
| $R_6$ | 2,6 | 0,1 | – 3,6 | – 8,5 | – 19,0 | – 26,9 | – 24,1 | – 24,8 |
| $R_7$ | 9,7 | 9,5 | 7,3 | 3,5 | – 15,5 | – 20,8 | – 17,7 | – 16,7 |
| Gemer | 4,1 | 3,5 | 1,5 | – 0,8 | – 13,6 | – 16,8 | – 17,4 | – 15,2 |

zo pseudonáhodného rozkladu $P_1$. Relatívne dobrú zhodu prejavujú aj dve riešenia získané metódou KM pre začiatočné rozklady $P_3$ a $P_4$ a okrem nich ešte dvojice riešení KM/$P_2$ a HC/$P_2$, MDP/$P_2$ a HC/$P_3$ a napokon aj MDP/$P_3$ a KM/$P_2$. Všimnime si však, že väčšina z uvedených riešení patrí medzi menej úspešné riešenia v zmysle minimalizácie funkcionálu Z. Za povšimnutie stojí aj zistenie, že najnižšie hodnoty koeficientu konkordancie sa spravidla viažu na druhé najhoršie riešenie MDP/$P_4$. Naproti tomu najlepší finálny rozklad HC/$P_4$ preukazuje iba mierny stupeň zhody s ostatnými finálnymi rozkladmi. Získané hodnoty $\hat{c}$ vo všeobecnosti potvrdzujú známy poznatok, že voľba začiatočného rozkladu má podstatný vplyv na *štruktúru* finálneho rozkladu.

Jeden z mnohých spôsobov, ako zlepšiť niektoré riešenia nehierarchického problému spočíva v tom, že finálne rozklady získané jednou nehierarchickou metódou sa použijú ako začiatočné rozklady pre inú nehierarchickú metódu. Keďže metódy MDP a KM nemôžu zlepšiť žiadne riešenie dosiahnuté metódou HC (pretože táto metóda explicitne minimalizuje funkcionál Z), aplikovali sme iba metódu HC na všetkých osem finálnych rozkladov získaných metódami MDP a KM. Hoci metóda HC vo všetkých prípadoch spôsobila ďalšie zníženie hodnôt funkcionálu Z, ani v jednom prípade nedošlo k poklesu pod hodnotu 2,8625. Z tohto dôvodu za optimálne riešenie nášho problému musíme považovať riešenie HC/$P_4$, aj keď, pravda, nie je možné prehlásiť, že toto riešenie vedie k absolútnemu minimu funkcionálu Z.

## RASTOVÉ TYPY OBCÍ V GEMERI

V tomto odseku podáme stručnú charakteristiku siedmich regionálnych typov získaných aplikáciou metódy HC na začiatočný rozklad $P_4$. Využijeme pritom údaje z tabuľky 3, ktorá udáva priemerné hodnoty jednotlivých atribútov v každej regionálnej triede a na porovnanie aj v celej množine gemerských obcí. Na uľahčenie identifikácie diferencií medzi regionálnymi typmi sa vnútrotriedne i celkové priemery vypočítali z pôvodných, t.j. nenormovaných hodnôt každého atribútu. Priestorové rozloženie jednotlivých regionálnych typov je znázornené na obr. 1.

*Regionálny typ* $R_1$ (počet priestorových jednotiek: $n_1 = 53$; vnútrotriedny rozptyl: $v_1 = 0{,}0086$)

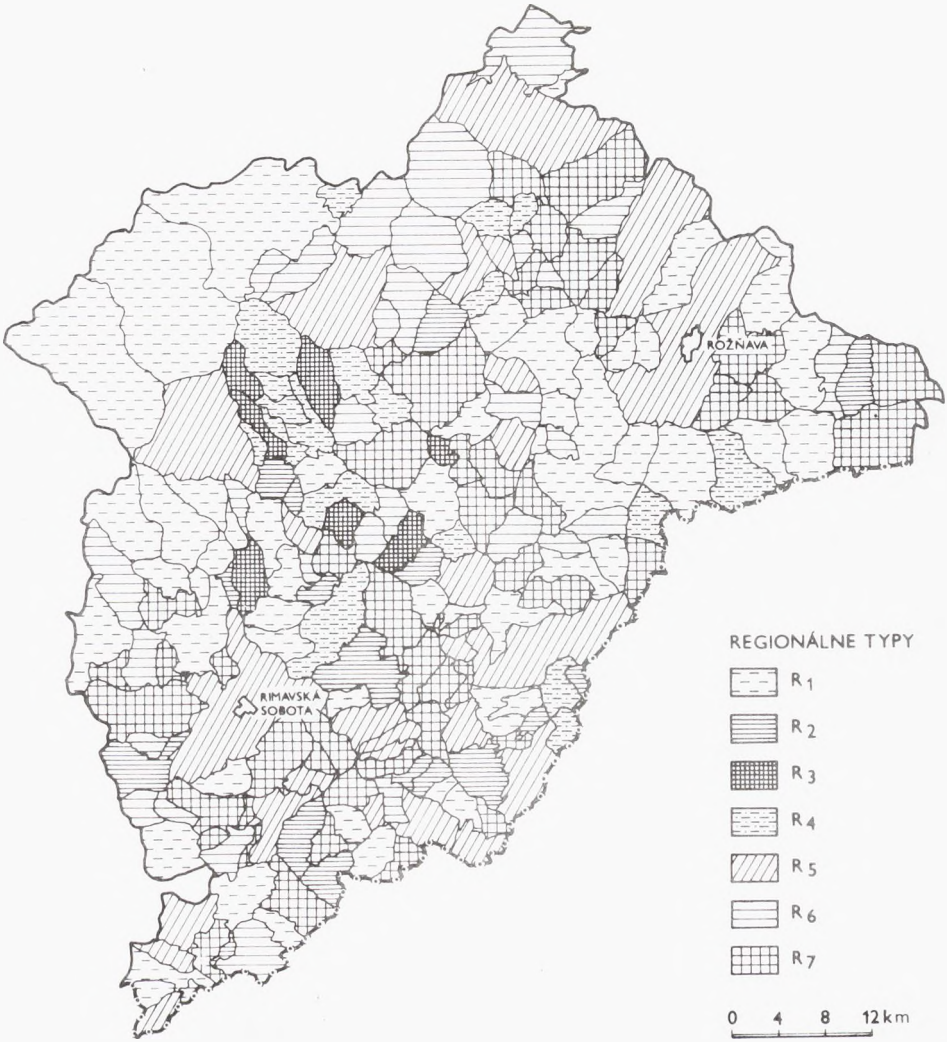
Tento najpočetnejší a súčasne najhomogénnejší regionálny typ zahrňuje 50 vidieckych obcí, dve mestá (Plešivec a Tisovec) a jednu obec prechodného charakteru (Štítnik). Vyznačuje sa nízkymi hodnotami prirodzeného prírastku (v druhom desaťročí pozorujeme dokonca malý prirodzený úbytok) a mierne vysokými hodnotami migračného úbytku. Hoci obce patriace k tomuto typu sú rozptýlené po celom území Gemera, väčšina z nich leží v severnej časti okresu Rimavská Sobota a vo východnej polovici okresu Rožňava. Sú to obce všetkých veľkostných kategórií až po hranicu 5 tisíc obyvateľov.

*Regionálny typ* $R_2$ ($n_2 = 16$; $v_2 = 0{,}0204$)

Od predchádzajúceho typu sa regionálny typ $R_2$ odlišuje kladnými hodnotami prirodzeného prírastku v druhej polovici pozorovaného obdobia a podstatne vyššími hodnotami migračného úbytku počas celého dvadsaťročia. Patrí k nemu 16 vidieckych obcí, z ktorých iba dve ležia na území okresu Rožňava. Takmer 70% obcí tohto typu malo v roku 1980 menej ako 500 obyvateľov.

*Regionálny typ* $R_3$ ($n_3 = 7$; $v_3 = 0{,}0247$)

K najmenšiemu typu $R_3$ zaraďujeme sedem vidieckych obcí ležiacich v strednej časti Revúckej vrchoviny po oboch stranách okresnej hranice. S výnimkou jednej obce (Sirk) majú všetky ostatné




Obr. 1. Priestorové rozloženie regionálnych typov obcí v Gemeri.

obce menej ako 350 obyvateľov. Charakteristickou črtou tohto typu sú takmer nulové prirodzené prírastky a extrémne vysoké migračné úbytky v priebehu celého pozorovaného obdobia.

*Regionálny typ* $R_4$ ($n_4 = 19$; $v_4 = 0{,}0336$)

Vnútorne najheterogénnejší regionálny typ $R_4$ pozostáva z 19 vidieckych obcí, z ktorých takmer polovica má menej ako 200 a 85% menej ako 500 obyvateľov. Vyznačuje sa jednoznačne negatívnym populačným vývojom počas celého sledovaného dvadsaťročia. Svedčia o tom nielen záporné hodnoty migračného prírastku, ale predovšetkým extrémne vysoké a postupne sa zväčšujúce prirodzené úbytky.

Tab. 2. Koeficienty konkordancie medzi finálnymi rozkladmi

| | | Metóda MDP | | | | Metóda KM | | | | Metóda HC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_1$ | $P_2$ | $P_3$ | $P_4$ |
| | $P_1$ | 1,000 | 0,535 | 0,356 | 0,305 | 0,671 | 0,349 | 0,394 | 0,387 | 0,639 | 0,396 | 0,546 | 0,418 |
| MDP | $P_2$ | 0,535 | 1,000 | 0,482 | 0,311 | 0,565 | 0,436 | 0,417 | 0,428 | 0,548 | 0,439 | 0,621 | 0,480 |
| | $P_3$ | 0,356 | 0,482 | 1,000 | 0,521 | 0,359 | 0,603 | 0,414 | 0,374 | 0,386 | 0,554 | 0,510 | 0,344 |
| | $P_4$ | 0,305 | 0,311 | 0,521 | 1,000 | 0,288 | 0,471 | 0,514 | 0,505 | 0,287 | 0,411 | 0,317 | 0,459 |
| | $P_1$ | 0,671 | 0,565 | 0,359 | 0,288 | 1,000 | 0,385 | 0,365 | 0,370 | 0,857 | 0,440 | 0,514 | 0,385 |
| KM | $P_2$ | 0,349 | 0,436 | 0,603 | 0,471 | 0,385 | 1,000 | 0,481 | 0,383 | 0,375 | 0,621 | 0,496 | 0,395 |
| | $P_3$ | 0,394 | 0,417 | 0,414 | 0,514 | 0,365 | 0,481 | 1,000 | 0,664 | 0,347 | 0,420 | 0,487 | 0,554 |
| | $P_4$ | 0,387 | 0,428 | 0,374 | 0,505 | 0,370 | 0,383 | 0,664 | 1,000 | 0,365 | 0,415 | 0,501 | 0,573 |
| | $P_1$ | 0,639 | 0,548 | 0,386 | 0,287 | 0,857 | 0,375 | 0,347 | 0,365 | 1,000 | 0,465 | 0,495 | 0,365 |
| HC | $P_2$ | 0,396 | 0,439 | 0,554 | 0,411 | 0,440 | 0,621 | 0,420 | 0,415 | 0,465 | 1,000 | 0,406 | 0,442 |
| | $P_3$ | 0,546 | 0,621 | 0,510 | 0,317 | 0,514 | 0,496 | 0,487 | 0,501 | 0,495 | 0,406 | 1,000 | 0,484 |
| | $P_4$ | 0,418 | 0,480 | 0,344 | 0,459 | 0,385 | 0,395 | 0,554 | 0,573 | 0,365 | 0,442 | 0,484 | 1,000 |

Je zaujímavé, že všetky obce tohto typu ležia v horskej časti Gemera. Najviac z nich sa nachádza na území Revúckej vrchoviny, hlavne v hornej časti povodia Východného Turca.

*Regionálny typ* $R_5$ ($n_5 = 27$; $v_5 = 0{,}0145$)

K regionálnemu typu $R_5$ patrí 19 vidieckych obcí, šesť miest (Dobšiná, Hnúšťa, Revúca, Rimavská Sobota, Rožňava a Tornaľa) a dve obce prechodného charakteru (Jesenské a Slavošovce). Dôsledkom relatívne početného zastúpenia obcí mestského a prechodného charakteru sú aj vysoké a vyrovnané hodnoty prirodzeného prírastku a výnimočne nízke hodnoty migračného úbytku. Ako prezrádza obr. 1, obce tohto typu sa vyskytujú najmä v juhovýchodnej časti okresu Rimavská Sobota a v severozápadnej časti okresu Rožňava. Zastúpené sú všetky veľkostné kategórie, pričom zhruba polovica obcí má viac ako 1000 obyvateľov.

*Regionálny typ* $R_6$ ($n_6 = 30$; $v_6 = 0{,}0128$)

Typ $R_6$ pozostáva z 30 vidieckych obcí, z ktorých viac ako 75% patrí do dvoch najnižších veľkostných kategórií. Príznačné pre tento typ sú pomerne značné prirodzené úbytky v druhej polovici pozorovaného obdobia a vysoké migračné straty počas celého dvadsaťročia. Pozoruhodné je, že väčšina obcí vytvárajúcich typ $R_6$ leží mimo hlavných cestných komunikácií v západných častiach oboch okresov.

*Regionálny typ* $R_7$ ($n_7 = 48$; $v_7 = 0{,}0102$)

Posledný regionálny typ tvorí 45 vidieckych obcí, jedno mesto (Jelšava) a dve obce prechodného charakteru (Lubeník a Nižná Slaná). Výraznou črtou tohto početného a relatívne homogénneho typu sú značne vysoké hodnoty prirodzeného prírastku a stredne vysoké hodnoty migračného úbytku. Obce zaradené k typu $R_7$ sa rozkladajú takmer na celom území Gemera s výnimkou severného výbežku okresu Rimavská Sobota. Najviac z nich leží v Rimavskej a Rožňavskej kotline, zvyšok nachádzame v dolinách Muráňa a hornej Slanej. Nadpriemerne početné zastúpenie majú kategórie stredne veľkých obcí, úplne chýbajú obce s menej ako 200 obyvateľmi.

## ZÁVEREČNÉ POZNÁMKY

Táto štúdia si kládla za cieľ poskytnúť stručný prehľad troch najvýznamnejších metód nehierarchickej taxonómie a na základe dát opisujúcich priestorové diferencie populačného rastu v Gemeri pokúsiť sa o ich komparatívne zhodnotenie. Ak ponecháme stranou detaily, môžeme výsledky tohto hodnotenia zhrnúť nasledovným spôsobom.

Pravdepodobne najdôležitejším rezultátom našej analýzy je poznatok, že metóda HC poskytuje pre všetky uvažované druhy začiatočného rozkladu jednoznačne najlepšie riešenie nehierarchického problému. Prioritný charakter metódy HC sa pritom neprejavuje iba vo vzťahu k minimalizácii funkcionálu kvality, ale aj vzhľadom na výpočtový čas, ktorý je potrebný na dosiahnutie stabilného finálneho rozkladu. Iným zaujímavým aspektom nášho hodnotenia je zistenie, že voľba začiatočného rozkladu množiny priestorových jednotiek má podstatný vplyv na štruktúru jej výsledného triedenia. V tejto súvislosti stojí za zmienku, že relatívne vysokým stupňom zhody finálnych rozkladov sa vyznačujú riešenia odvodené všetkými tromi nehierarchickými metódami zo pseudonáhodného začiatočného rozkladu.

Všetky uvedené výsledky našej štúdie sa zhodujú s poznatkami, ku ktorým dospeli M.M. Fischer a jeho spolupracovníci v analogicky koncipovanom príspevku [8]. Na druhej strane existuje niekoľko sporných miest, v ktorých sa výsledky oboch štúdií značne rozchádzajú. V prvom rade musíme upozorniť, že v našej štúdii sa nepotvrdila lepšia účinnosť metódy KM v porovnaní

s metódou MDP. Okrem toho výsledky vzájomného porovnania finálnych rozkladov nepodporujú tvrdenie, že metódy HC a KM vedú k podobným riešeniam, ak sa aplikujú na rovnaké začiatočné rozklady. Navyše sa zdá, že obdobné tvrdenie neplatí ani pre inú dvojicu nehierarchických procedúr. Pozoruhodné v tomto kontexte je aj zistenie, že voľba nehierarchickej procedúry mala v našej štúdii omnoho závažnejšie dôsledky na celkovú úspešnosť taxonomického riešenia ako výber začiatočného rozkladu. Napokon môžeme ešte konštatovať, že začiatočné rozklady odvodené na pozadí určitých vysvetľujúcich hypotéz, týkajúcich sa veecnej stránky skúmaného problému, nepriniesli v porovnaní so pseudonáhodným rozkladom výraznejší pokles hodnoty funkcionálu kvality. Na základe tohto poznatku môžeme vysloviť domnienku, že vysvetlenie priestorovej diferenciácie populačného rastu na území Gemera nebude práve jednoduchou úlohou.

## LITERATÚRA

1. ANDERBERG, M.R.: Cluster analysis for aplications. New York 1973. - 2. Atlas Slovenskej socialistickej republiky. Bratislava 1980. - 3. Atlas SSR: textová časť. Bratislava 1982. - 4. BARKER, D.: Hierarchic and non-hierarchic grouping methods: an empirical comparison of two techniques. Geografiska Annaler, *58* B, 1976, pp. 42-58. - 5. BAUMANN, J.H., FISCHER, M.M., SCHUBERT, U.: A multiregional labour supply model for Austria: the effects of different regionalisations in multiregional labour market modelling. Papers of the Regional Science Association, *52*, 1983, pp. 53-83. - 6. BEZÁK, A.: Migračné toky a regionálna štruktúra Slovenska: nehierarchická regionalizácia. Geografický časopis, *43*, 1991, pp. 265-275. - 7. BLUME, H., SCHWARZ, R.: Zur Regionalisierung der USA. Geographische Zeitschrift, *64*, 1976, pp. 262-295. - 8. ERTL, H., FISCHER, M.M, WOHLSCHLÄGL, H.: A methodological approach for large regional taxonomic problems: spatial patterns of population development in Austria. Papers of the Regional Science Association, *44*, 1980, pp. 119-135. - 9. EVERITT, B.: Cluster analysis. London 1974. - 10. FISCHER, M.M.: Regional taxonomy: a comparison of some hierarchic and non-hierarchic strategies. Regional Science and Urban Economics, *10*, 1980, pp. 503-537.

11. FISCHER, M.M.: Eine Methodologie der Regionaltaxonomie: Probleme und Verfahren der Klassifikation und Regionalisierung in der Geographie und Regionalforschung. Bremer Beiträge zur Geographie und Raumplanung, *3*, 1982. - 12. FISCHER, M.M., DROTH, W.: Factorial ecology versus cluster analysis: the intra-urban residential structure of the city of Hamburg. Bremer Beiträge zur Geographie und Raumplanung, *8*, 1986, pp. 99-121. - 13. HARTIGAN, J.A.: Clustering algorithms. New York 1975. - 14. HEBÁK, P., HUSTOPECKÝ, J.: Vícerozměrné statistické metody s aplikacemi. Praha 1987. - 15. KOSTRUBIEC, B.: Taksonomia numeryczna w badaniach geograficznych. Acta Universitatis Wratislaviensis, Studia Geograficzne, *38*, 1982. - 16. LUKASOVÁ, A., ŠARMANOVÁ, J.: Metody shlukové analýzy. Praha 1985. - 17. RAND, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, *66*, 1971, pp. 846-850. - 18. SPÄTH, H.: Cluster analysis algorithms for data reduction and classification of objects. Chichester 1980. - 19. TAUBMANN, W.: Zur Abgrenzung von sogenannten Problemgebieten: Bemerkungen zu Indikatoren und Verfahrensweisen. Karlsruher Manuskripte zur Mathematischen und Theoretischen Wirtschafts- und Sozialgeographie, *32*, 1979.

Anton Bezák

## COMPARATIVE ANALYSIS OF NON-HIERARCHIC REGIONAL TAXONOMIC METHODS (THE GROWTH TYPES OF COMMUNES IN THE GEMER REGION AS AN EXAMPLE)

The aim of this paper has been to review and evaluate three non-hierarchic procedures for solving regional taxonomic problems by applying them to the same data set. As basic spatial units of the study the 200 communes of the Gemer Region (one of the historical counties of Slovakia) were taken. In order to characterize the population development of these communes during the last twenty years, eight attributes

were chosen in the following way. The 1971–1990 period was first divided into four five–year intervals. Then two mean annual rates were defined to measure both the natural and migration components of population growth in each interval. Finally, all attribute values were reduced to the same scale by means of the $L_2$–normalization procedure.

Three iterative non–hierarchic procedures were selected and compared in this study. They were: the minimum distance partition method (MDP), the k–means method (KM), and the hill–climbing method (HC). In order to examine the influence of different initial partitions upon the result, the following four partitions of the set of communes into seven regional taxa were used:

$P_1$ – (pseudo–) random partition in which commune $i$ is assigned to regional taxon $j$ according to the rule: $j = \mathrm{mod}(i - 1, k) + 1$ for $i = 1, 2, \ldots, n$ ($k$ is number of regional taxa)

$P_2$ – partition in which the communes are assigned to regional taxa in accordance to their population size in 1980

$P_3$ – partition derived from the typology of Slovak communes with respect to their population age composition (see [2, p. 140])

$P_4$ – partition completed on the basis of the functional classification of Slovak communes (see[2, p. 170]).

The number of regional taxa ($k = 7$) was determined according to the trace criterion [9, p. 60] in such manner that the HC method was applied to random initial partition in a series of runs with varying values of $k$. Finally, it should be noted that no contiguity constraint prohibiting disconnected regional taxa was imposed.

The regional taxonomic results for seven regional taxa obtained by the MDP, KM and HC methods were compared and evaluated by the within–taxa variance criterion $Z$ (see formula (2)). The corresponding criterion values for the 12 final partitions are shown in Table 1, whereas concordance coefficients between each two final partitions (defined by formula (6)) are summarized in Table 2. From these tables, the following main conclusions can be drawn.

As expected, the HC method provides the best results for all initial partitions considered. Contrary to expectations, however, no superiority of the KM method to the MDP method was confirmed. On the other side it is impossible to determine what kind of initial partition yields, in general, better results than the other ones. Another interesting point in this context is that the initial partitions $P_2$, $P_3$ and $P_4$, reflecting some a priori hypotheses concerning the population growth in the study area, do not lead to substantial decrease of the $Z$ value in comparison with the random partition $P_1$. These findings suggest that in our study the choice of the non–hierarchic procedure had a more important effect upon the solution of the taxonomic problem than the choice between the four different initial partitions.

As regards the concordance between final partitions, it is worth mentioning that the highest degree of similarity can be found between final partitions derived by all the three non–hierarchic procedures from the random initial partition. In general, the values from Table 2 confirm the well–known result that the choice of the initial partition has a substantial influence upon the *structure* of the final partition.

The last section of the paper is devoted to empirical interpretation of the best final partition obtained by the hill–climbing method combined with initial partition $P_4$. The spatial distribution of the seven regional taxa is shown in Figure 1. A special attention is given to their distinctions in terms of the mean attribute values (Tab. 3).

Tab. 1. Evaluation of the final partitions
(non–hierarchic method, type of initial partition, variance criterion value, the within–taxa variance in percent of the total variance).
Tab. 2. Coefficients of concordance between the final partitions.
Tab. 3. Mean annual rates of natural and migration growth (in o/oo) in four five–year intervals for seven regional taxa and the whole set of communes in the Gemer Region.
Fig. 1. Spatial distribution of seven regional types of communes in the Gemer Region.

English by the author