Některé z možných aplikací fuzzy množin na oblast databázových systémů

Nejvíce v relačních SŘBD, protože jsou v praxi nejrozšířenější, i když i další datové modely přicházejí do úvahy.

Použití fuzzy množin v R-SŘBD převážně spadá do dvou kategorií:

1. Reprezentace vágně určených konceptů pomocí fuzzy relací.
2. Nekompletní informace uvnitř záznamů.

Ad 1) Každý záznam (n-tice, tuple) t relace r má stupeň příslušnosti, který udává, do jaké míry t náleží do r (tj. do fuzzy relace r) – říkáme, že r obsahuje vážené záznamy.

Takovéto relace lze ovšem chápat různě, v závislosti na zamýšleném významu vah. Několik možností:

- stupeň, do nějž je splněn fuzzy koncept reprezentovaný danou relací. Např. jestliže relace EMP-MARE(#emp, name, date-of-birth, date-of-employment, ...) reprezentuje „zaměstnance, kteří jsou středního věku a nedávno přijatí", pak váha přidělená konkrétnímu záznamu udává, do jaké míry je zaměstnanec „středního věku a nedávno přijatý".

Ad 2) Fuzzy R-SŘBD založené na pojmu „možnost".

Zde se vychází ze skutečnosti, že informace o tom, jaké hodnoty může nabýt nějaký atribut v záznamu, je reprezentována distribucí možnosti:

$\pi_{A(t)} \in D \cup \{e\}$, kde A je skalární atribut, t je záznam, D je doména atributu A, a e je přídavný element pro případy, kdy atribut A není aplikovatelný na záznam t.

Distribuce možnosti $\pi_{A(t)}$ je chápána jako fuzzy restrikce možných hodnot A(t) a definuje mapování $D \cup \{e\} \to [0, 1]$.

Např. informace Pepa má dostatečnou zkušenost bude reprezentována jako

$$\pi_{zkušenost(Pepa)}(e) = 0, \quad \pi_{zkušenost(Pepa)}(d) = \mu_{dostatečný}(d)$$

kde $d \in D$. Zde $\mu_{dostatečný}$ je funkce příslušnosti reprezentující vágní predikát dostatečný pro daný kontext (jako např. počet let zaměstnání, vzdělání atp.).

Posibilistický přístup umožňuje jednotným způsobem vyjadřovat přesné hodnoty (singletony), prázdné hodnoty (NULL), a přibližné (fuzzy množiny) či nepřesné (klasické množiny).

Takový typ relace může vzniknout z obyčejné (ne-fuzzy) relace, na níž byla použita nějaká „fuzzy" podmínka. Případně jsou aplikovány „fuzzy" hodnoty vyjádřené lingvistickými výrazy (vůbec ne, přiměřeně, trochu, velice ...). Např. relace MÁ-RÁD(osoba, film) – váhy zde odpovídají lingvistickým hodnotám (zde ovšem, na rozdíl od předchozího příkladu se zaměstnancem, nelze váhy určit z hodnot atributů) v relaci.

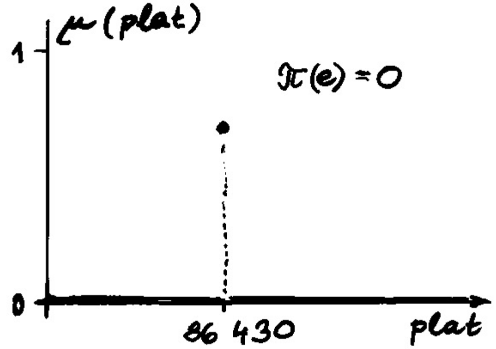
- míra jistoty (jak je informace zatížena nejistotou) informace uložené v záznamech. Informace s sebou nese kvalifikátor (jistota chápaná ve formě váhy záznamu):

  0.8/(Pepa, Hvězdné války) může vyjadřovat, že je na 80% jisté, že Pepovi se líbí film Hvězdné války.

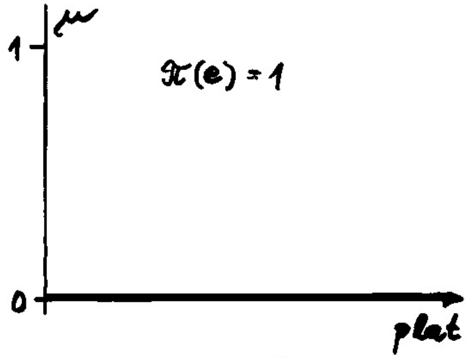
- míra možnosti (jak možná je informace uložená v záznamu): POSNÍDÁ-VAJÍČEK(osoba, počet)

  1.0/(Pepa, 1) 1.0/(Pepa, 2), 0.85/(Pepa, 3), 0.4/(Franta, 2) ...

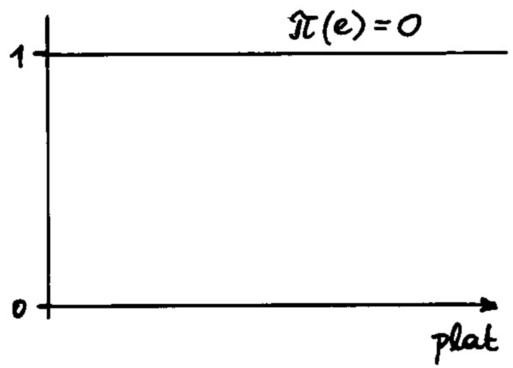
Distribuce možnosti pro obvyklé situace:



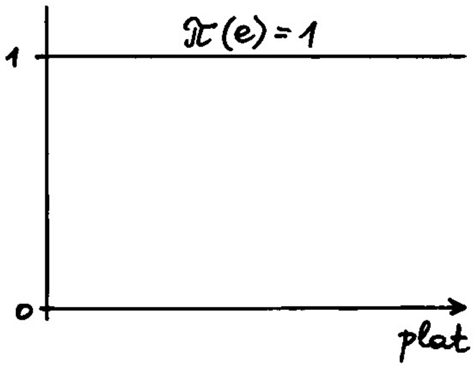
(přesná hodnota)



(neaplikovatelná hodnota – hodnota neexistuje)



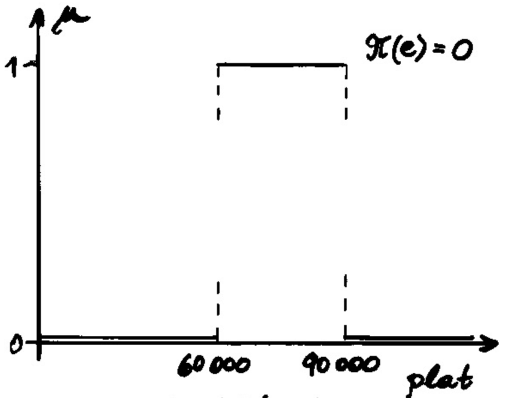
(neznámá, avšak existující hodnota)



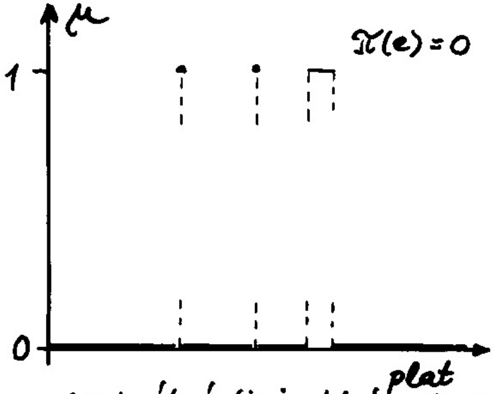
(úplná neznalost – cokoliv je zcela možné)

může i nemusí existovat

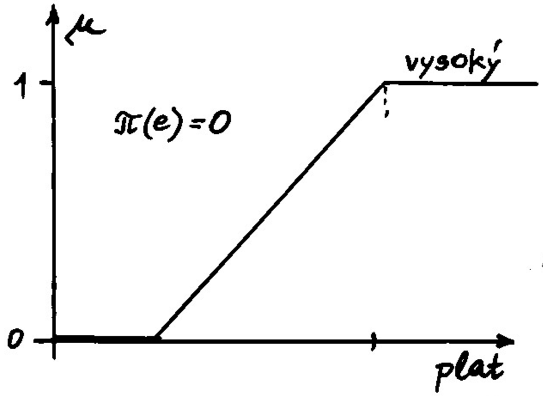
Distribuce možnosti pro nepřesně/přibližně/neúplně/... známé hodnoty atributu:



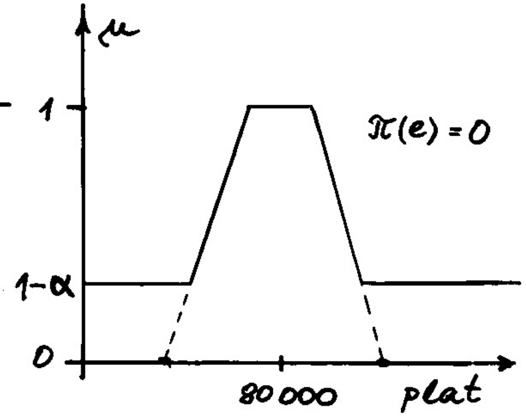
(obvyklý interval)



(diskrétní disjunktní infor.)



(distribuce odpovídající restrikci ~~prom~~ hodnot proměnné plat fuzzy množině vysoký)



(fakt, že Pepův plat je okolo 80 000 je tzv. α-jistý)

Pozn.: Skutečnost, že fuzzy hodnota je množina, nenarušuje první normální formu relace (1NF), neboť atomickou hodnototou z tohoto hlediska je ta množina jako celek (na kódování nezáleží).

Dotazovací jazyky – flexibilní dotazy

Vyhledávání a výběr dat v databázi často tvoří převažující aktivitu uživatelů. Dotazovací funkčnost databázových jazyků byla v některých případech rozšířena na tzv. flexibilní dotazování, kdy výsledkem nemusí být jediná odpověď, nýbrž soubor odpovědí navzájem odlišených.

Fuzzy predikáty

Např. dotaz směřuje na vyhledání (Internet...) nepříliš drahých lokalit blízko dopravě. Aplikace klasického množinového výběru může často vrátit prázdnou odpověď např. důsledkem přílišných požadavků, zatímco flexibilní systém může poskytnout v takovém případě řadu odpovědí seřazených od nejlépe vyhovujících uživatelovu požadavku. Přitom lze vyřadit mnoho „podprahových" odpovědí. To může zabránit mnohonásobnému přeformulovávání ~~odp~~ dotazů. Lingvisticky zadané přibližné hodnoty lze modelovat pomocí fuzzy množin a dále aplikovat i lingvistické operátory umožňující modelovat mnoho, více-méně, spíše, velmi ... (koncentrace, dilatace).

Složené podmínky výběru dat, mající formu logických výrazů, jsou reprezentovány pomocí operátorů nad fuzzy množinami (možnosti jsou bohatší než s Booleovskou algebrou).

Konjunkci/disjunkci predikátů lze zpracovat buď pomocí operátorů min/max nebo také pomocí kompenzačních funkcí. (resp. t-norem)

Je-li zapotřebí vyjádřit, že některé elementární podmínky jsou méně důležité než jiné, použiji se tzv. <u>váhy významnosti</u>:

konjunktivní agregace

$$\min_i \max\left(\mu_{P_i}(A_i(x)),\ 1-w_i\right)$$

disjunktivní agregace

$$\max_i \min\left(\mu_{P_i}(A_i(x)),\ w_i\right)$$

kde P_i je podmínka aplikovaná na $A_i(x)$, w_i je váha významnosti ($\max_i w_i = 1$).
(tj. jsou-li všechny podmínky stejně významné, $w_i = 1,\ \forall i$), pak se obě operace redukují na standardní min a max).

<u>Příklad:</u> Hledá se byt jenž je levný a dostatečně velký, přičemž druhá ($w<1$) podmínka (velikost) je méně významná než první ($w=1$) (láce). Odpovídající formule pak je:

$$\min\left(\max\left(\mu_{levný}(cena),\ 1-w_{levný}\right),\ \max\left(\mu_{dost\text{-}velký}(plocha),\ 1-w_{dost\text{-}velký}\right)\right) =$$

$$= \min\left(\mu_{levný}(cena),\ \max\left(\mu_{dost\text{-}velký}(plocha),\ 1-w_{dost\text{-}velký}\right)\right)$$

Pro vyhodnocování výrazů uvedeného typu se používají dvě veličiny: možnost a jistota, že podmínka je splněna.

<u>Možnost:</u> možnost; fuzzy množina; distribuce možnosti

$$\mu_{\Pi P}(t) = \underbrace{\Pi}_{possibility}(P;\ A(t)) = \sup_{d\in D} \min\left(\mu_P(d),\ \pi_{A(t)}(d)\right)$$

<u>Jistota:</u> necessity (nutnost)

$$\mu_{NP}(t) = N(P;\ A(t)) = 1 - \Pi(\bar{P};\ A(t)) =$$

$$= 1 - \sup_{d\in D\cup\{e\}} \min\left(\mu_{\bar{P}}(d),\ \pi_{A(t)}(d)\right) =$$

$$= \inf_{d\in D\cup\{e\}} \max\left(\mu_P(d),\ 1-\pi_{A(t)}(d)\right)$$

<u>$\Pi(P;\ A(t))$</u> udává odhad, do jaké míry <u>nejméně jedna</u> <u>hodnota</u> $A(t)$ omezená $\pi_{A(t)}$ je kompatibilní s P (tj. s definovanou hodnotou, jako je STŘEDNÍ-VĚK).

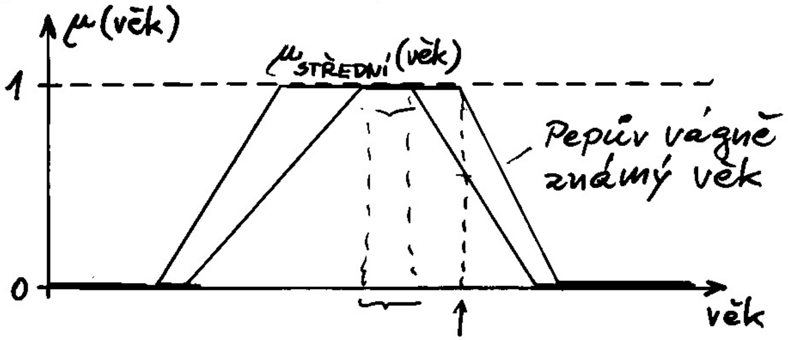
<u>$N(P;\ A(t))$</u> udává, do jaké míry <u>všechny hodnoty</u>, které více či méně připadají pro $A(t)$ do úvahy, jsou zahrnuty v P.

Pozn.: bylo ukázáno, že ΠP a NP (fuzzy množiny) vždy splňují inkluzivní relaci $\Pi P \supseteq NP$ (tj., $\forall t$, $\mu_{NP}(t) \le \mu_{\Pi P}(t)$), za předpokl., že $\pi_{A(t)}$ je normální.

Fuzzy dotazování nad relačními SŘBD

Implementace (praktické) často nepoužívají striktně teoretický rámec, jsou spíše založeny na intuitivním očekávání a představě, co by měl systém dělat.

Aplikace fuzzy hodnot zapříčiňuje, že odpověď není ve formě jediné hodnoty. Nelze totiž říci, zda hodnota splňuje či nesplňuje danou podmínku (ta může být splněna mnoha záznamy v různém stupni !).



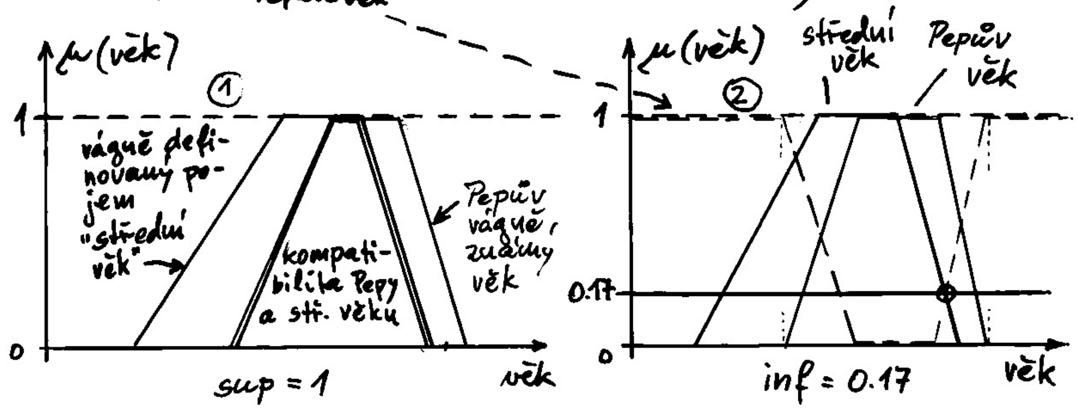
Otázka zní: <u>Je Pepa středního věku</u> (a má tedy být z databáze vybrán) ?

Odpověď nemusí být snadná, zejména je-li hodnota STŘEDNÍ i hodnota Pepova věku vyjádřena přibližně – odpověď pak je samozřejmě také přibližná a interpretace musí být učiněna dodatečně.

<u>Příklad</u>: Je-li Pepův věk a pojem „střední věk' vymezen dle předchozího obrázku, pak vyhodnocení podmínky věk(Pepa) = "střední věk" se vypočte následovně:

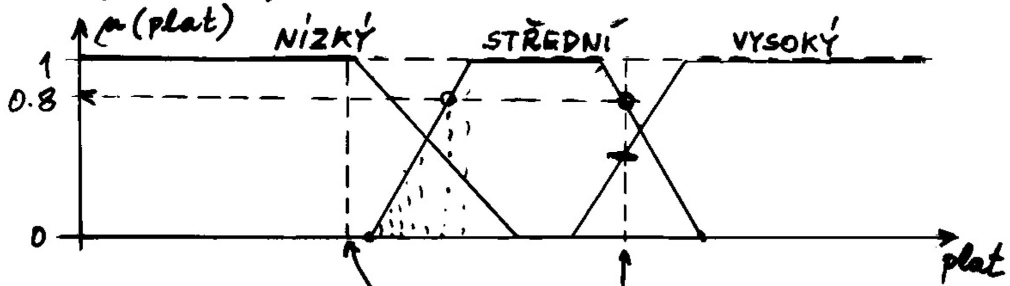
① $\min(\pi_{Pepův\text{-}věk}(u),\ \mu_{střední\text{-}věk}(u))$ a

② $\max(1-\pi_{Pepův\text{-}věk}(u),\ \mu_{střední\text{-}věk}(u))$



Protože supremum = 1, tak je <u>zcela</u> možné, že Pepa má střední věk. <u>Jistota tvrzení</u>, že Pepův věk je střední, je dána stupněm jistoty = 0.17 (např. <u>interpretace může být</u>: Ano, je 100% <u>možné</u>, že Pepa je středního věku, a jsme si tím <u>jisti</u> na 17%).

Fuzzy dotazy nad konvenční databází



ZAMĚSTNANEC

OS. ČÍSLO	JMÉNO	FUNKCE	PLAT
043672	DONALD K.	03	85 500
122053	MOUSE M.	12	36 000

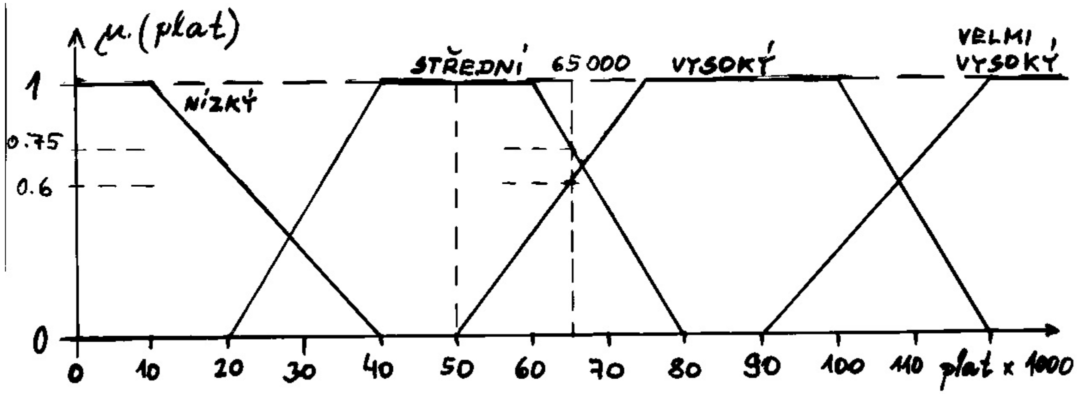
SELECT (OS.ČÍSLO, JMÉNO)
WHERE PLAT = "STŘEDNÍ"

⋮

↓

| 043672 | DONALD K. | 0.8

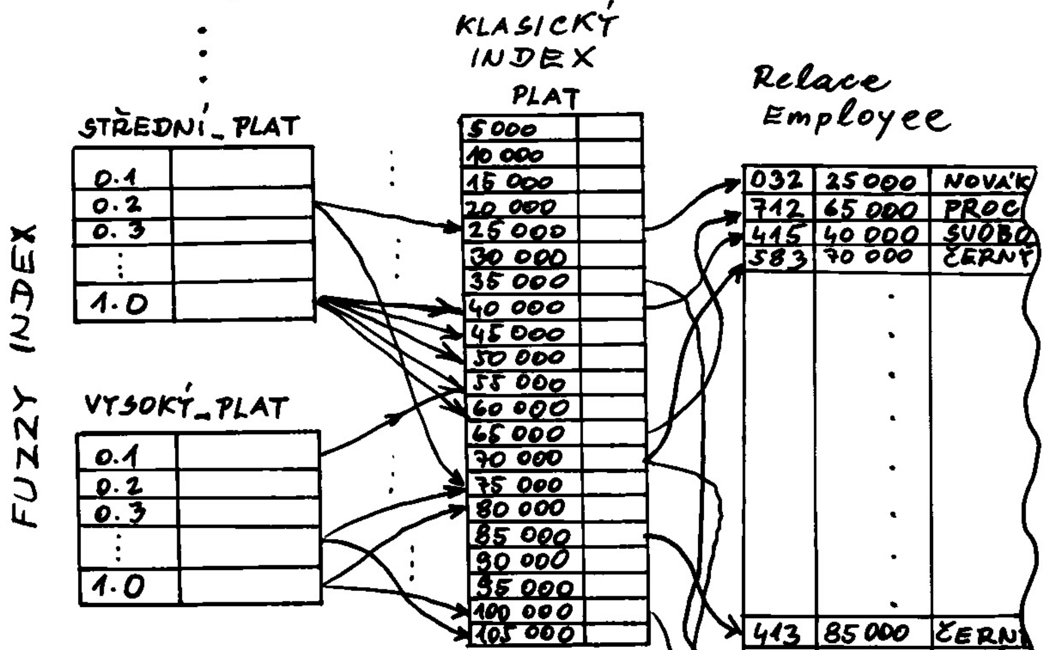
Vybraný záznam nesplňuje danou podmínku dokonale



SELECT (Emp#) FROM Employee
WHERE plat = STŘEDNÍ

SELECT Emp# FROM Employee
WHERE plat = VYSOKÝ

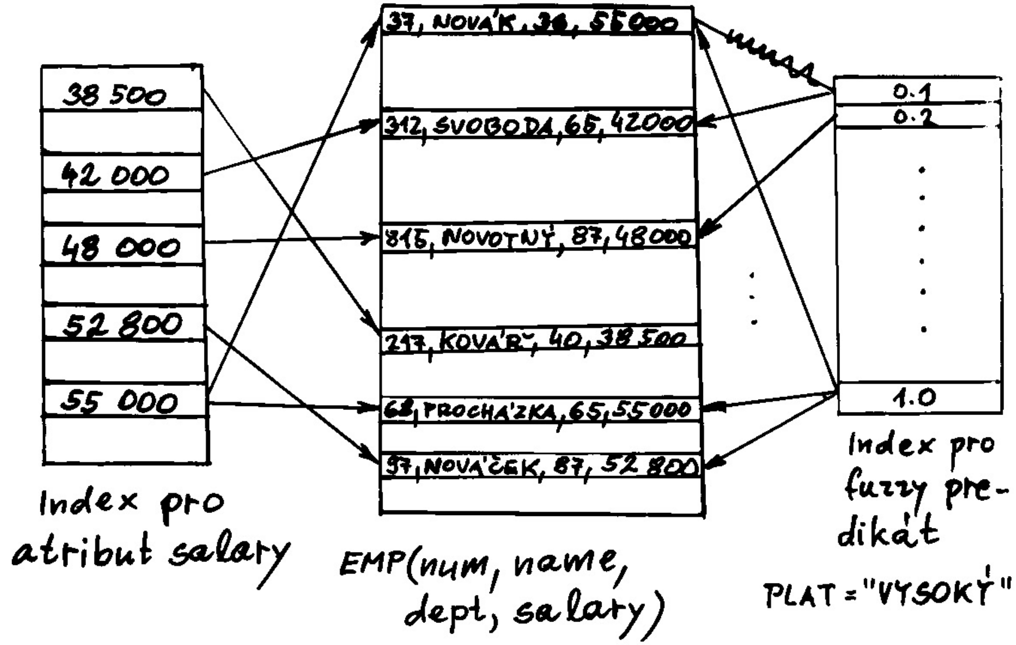
SELECT (Emp#, Emp_Name) FROM Employee
WHERE plat = VELMI_VYSOKÝ AND věk = STŘEDNÍ ...



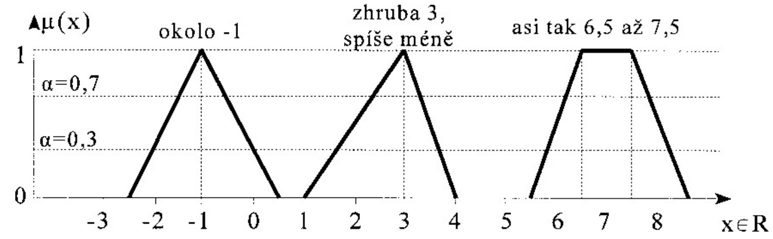
032	25 000	NOVÁK
712	65 000	PROC
415	40 000	SVOBO
383	70 000	ČERNÝ
⋮		
413	85 000	ČERN
205	35 000	BÍLÝ J
131	100 000	NOVOTN
613	70 000	PROCH

konkrétní fuzzy index lze vytvářet podle potřeby

Implementační poznámka



ZÁKLADNÍ ARITMETIKA PRO SPOJITÁ FUZZY ČÍSLA POMOCÍ α–ŘEZŮ



$$M_\alpha = \{x \in X \mid \mu_M(x) \geq \alpha\} \qquad M = \bigcup_{\alpha=0}^{1} \alpha M_\alpha = \bigvee_{\alpha=0}^{1} \alpha M_\alpha$$

Nechť M a N jsou spojitá fuzzy čísla, $M_\alpha=[m_1, m_2]$ je α-řez fuzzy čísla M pro nějakou hodnotu α, a podobně $N_\alpha=[n_1, n_2]$. Nejjednodušší možností je využití intervalové aritmetiky:

1. **Sečítání**: $M_\alpha+N_\alpha = [m_1+n_1, m_2+n_2]$
   Vlastnosti:
   a) $M+N = N+M$
   b) $(M+N)+K = M+(N+K)$
   c) $M+0 = 0+M = M$
   d) *monotónnost, konvexnost, normálnost* jsou zachovány (důkaz uveden v Kaufmann A., Gupta M.M.: *Introduction to Fuzzy Arithmetics*, Van Nostrand, N.Y., 1985)

2. **Odčítání**: $M_\alpha-N_\alpha = [m_1-n_2, m_2-n_1]$
   M-N je ekvivalentní sečtení $M+(-N)$, kde $-N$ je tzv. *obraz N* definovaný jako $\mu_{-N}(x) = \mu_N(-x)$, $\forall x$.
   Protože jako výsledek odčítání může vzniknout záporné číslo, komutativnost a asociativnost *nejsou* zachovány.

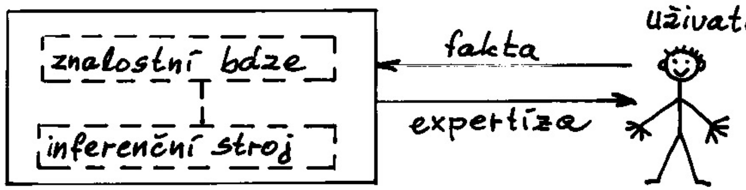
3. **Násobení**: $M_\alpha \cdot N_\alpha = [m_1 \cdot n_1, m_2 \cdot n_2]$
   Vlastnosti:
   a) $M \cdot N = N \cdot M$ d) $M \cdot 1 = 1 \cdot M = M$ g) $(-M) \cdot N = -(M \cdot N)$
   b) $(M \cdot N) \cdot K = M \cdot (N \cdot K)$ e) $M_\alpha^{-1} = [1/m_2, 1/m_1]$ výjimka: nulové fuzzy číslo !
   c) $M+0 = 0+M = M$ f) $M \cdot M^{-1} \neq 1$

4. **Dělení**: $M_\alpha/N_\alpha = M_\alpha \cdot N_\alpha^{-1} = [m_1/n_2, m_2/n_1]$
   Dělení je převedeno na ekvivalent násobení inverzní hodnotou.

(Fuzzy) Expertní systém

Expertní systém je počítačový systém, který (v ideálním případě) emuluje schopnosti člověka-experta provádět rozhodování v příslušné doméně (oblasti problematiky). Expertní systémy (ES) fungují velmi dobře ve svých omezených doménách (obchod, medicína, věda a výzkum).

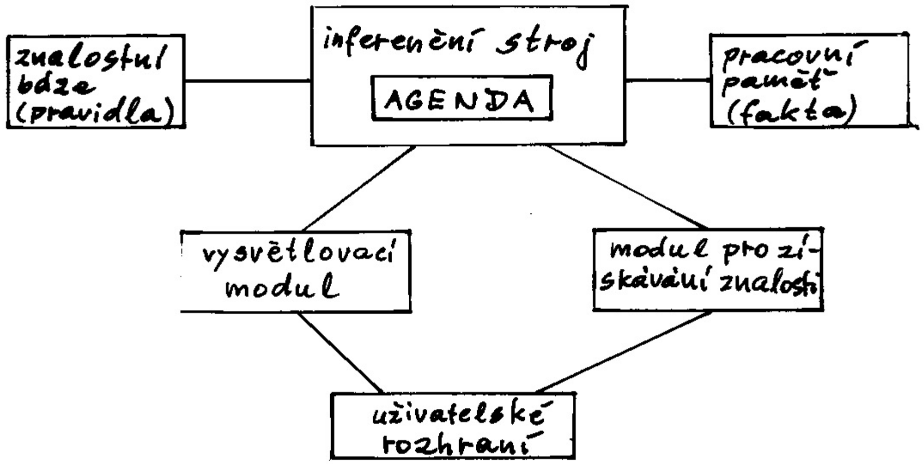
Základní koncept ES jako systému založeného na zpracování znalostí:



ES mají řadu výhod:

- zvýšená dostupnost expertízy (je poskytována hromadně vhodným HW+SW);
- redukovaná cena na uživatele;
- redukovaná rizika (lze užít v prostředí nebezpečném pro člověka);
- permanence (člověk-expert může změnit činnost, odejít do penze, zemřít...);
- vícenásobná expertíza (znalost více expertů může být kdykoliv kdekoliv použita zároveň a může převýšit znalost jediného experta);
- zvýšená spolehlivost expertízy (ES jako doplněk jiného experta, různého od tvůrce báze znalostí);
- vysvětlení (jak ES došel k poskytnutému závěru) je ihned k dispozici;
- rychlá doba odezvy (ES může být rychlejší či být ihned k dispozici);
- trvalá, bezemoční a úplná odpověď kdykoliv;

Struktura ES řízeného pravidly:



Fuzzy ES

Fuzzy ES zpracovávají přibližnou znalost prostřednictvím fuzzy (přibližné) logiky. Přibližné kvantity jsou modelovány pomocí fuzzy množin.

Přibližná znalost je reprezentována formou fuzzy pravidel IF A THEN B zachycujících vztahy mezi přibližnými ~~p~~ hodnotami.

Fuzzy rozhodování

Rozhodování tvoří každodenní součást života. Podstatnou vlastností reálného rozhodování je, že téměř všechny rozhodovací problémy mívají mnohonásobná, často protichůdná, kritéria.

V principu existují dvě hlavní kategorie:

- víceatributové rozhodování MADM (multiple attribute decision making)
- vícecílové rozhodování MODM (multiple objective decision making)

Z praktického hlediska je MADM spojeno s problémy, kde počet alternativ řešení je předem stanoven.

Rozhodovací mechanismus má za cíl vybrat/dát prioritu/seřadit konečný počet akcí.

MODM nemá předem stanovené alternativy. Cílem je navrhnout „nejslibnější" možnost vzhledem k omezeným zdrojům.

Při rozhodování ve světě lidí prakticky vždy existuje nejistota, proto se pro řešení rozhodovacích problémů využívá teorie pravděpodobnosti (pro stochastické problémy) a pro rozhodování nemající stochastický charakter se využívají např. fuzzy množiny.

MADM se zabývá rozhodováním mezi existujícími směry potenciálních akcí za předpokladu přítomnosti mnoha atributů, často konfliktních (např. laciný AND kvalitní).

Typický MADM problém:

Volba zaměstnání – výběr z několika možností (plat, místo, možnosti kariéry, kolegové ...).

Koupě automobilu – výběr z několika značek a typů (cena, bezpečnost, komfort, spotřeba ...).

Plánování vodárny (cena, možnost nedostatku vody, energie, ochrana čistoty vody, kvalita vody ...).

Výběr vysokoškolského učitele (výzkumné schopnosti, pedagogické schopnosti, komunikativnost, vyspělost ...).

Problém MADM lze vyjádřit formou matice:

$$D = \begin{array}{c c} & \begin{array}{cccc} X_1 & X_2 & \cdots & X_n \end{array} \\ \begin{array}{c} A_1 \\ A_2 \\ \vdots \\ A_m \end{array} & \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \end{array}$$

kde A_i je alternativa (možná akce), X_j jsou atributy, jimiž je měřena účinnost alternativy, x_{ij} je účinnost alternativy A_i vzhledem k atributu X_j.

Není neobvyklé, že x_{ij} je známo pouze přibližně.

Nepřesnost či přibližnost má několik zdrojů:

- nekvantifikovatelná informace: cenu auta lze snadno kvantifikovat, zatímco míru komfortu nikoliv. Bezpečnost a komfort se obvykle vyjadřují lingvistickými termíny (dobrý, přijatelný, špatný ...);
- nekompletní informace: např. rychlost rychle se pohybujícího objektu vyjádřená jako „okolo 120 km/h" a nikoliv jako „přesně 120 km/h".
- nepřístupná informace: někdy lze přesná data získat, avšak za příliš vysokou cenu, takže rozhodnutí je nutno provést pomocí „aproximace" přesných dat. Rovněž pro tzv. senzitivní data (státní tajemství, velikost bankovního účtu jedince, věk ženy) se používá pouze aproximace či lingvistický popis;
- částečná ignorance (neznalost): je známa pouze část faktů nebo fakta jsou známa nedokonale.

Klasické MADM metody nejsou schopny s uvedenými typy přibližnosti pracovat. Teorie fuzzy množin je schopna přibližné hodnoty modelovat jako fuzzy množiny, resp. fuzzy čísla.

Numerický příklad pro MADM:

Jedna země se rozhodla zakoupit stíhačky od USA. Odborníci z Pentagonu nabídli charakteristiky čtyř modelů na prodej: maximální rychlost (X_1), dolet (X_2), maximální hodnota nákladu (X_3), nákupní cena (X_4), spolehlivost (X_5), manévrovací schopnosti (X_6), tj. 6 atributů. Jednotky pro jednotlivé atributy jsou Mach, míle, libry, mil. dolarů, škála vysoký-nízký, a škála vysoký-nízký. Rozhodovací matice pro výběr stíhačky:

$$D = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \end{array} \begin{array}{c} \begin{array}{cccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \end{array} \\ \left[\begin{array}{cccccc} 2.0 & 1500 & 20\,000 & 5.5 & \text{průměrná} & \text{velmi vysoké} \\ 2.5 & 2700 & 18\,000 & 6.5 & \text{nízká} & \text{střední} \\ 1.8 & 2000 & 21\,000 & 4.5 & \text{vysoká} & \text{vysoké} \\ 2.2 & 1800 & 20\,000 & 5.0 & \text{průměrná} & \text{střední} \end{array} \right] \end{array}$$

Každý atribut X_j může být ještě váhován (tj. stanovena jeho relativní významnost):

$$\vec{w} = (w_1, w_2, \dots, w_n)$$

Rozhodovací mechanismus vypočítá (hodnotu) tzv. účelové funkce $U_i(x_1, x_2, \dots, x_n)$ pro A_i. Hodnota účelové funkce je použita pro konečné rozhodnutí (maximální nebo minimální hodnota).

Obecně je problém seřazení (tj. porovnání) hodnot U_i, $\forall i$, velmi obtížný.

$U(x_1, \dots, x_n)$ je definována rozhodovacím mechanismem.

Klasické MADM metody předpokládají, že všechny hodnoty x_{ij}, w_j jsou tzv. ostrá čísla.

Pro každou alternativu A_i agreguje účelová funkce U koeficienty efektivity x_{ij}, $\forall j$, a výsledkem je nějaká konečná hodnota U_i. Alternativy s vyšší U_i jsou preferovány před nižšími.

Existuje mnoho metod MADM (cca 18 hlavních).

Fuzzy jednoduchá aditivní váhová metoda

Tato metoda je v klasické (non-fuzzy) formě definována jako

$$U_i = \sum_{j=1}^{m} w_j r_{ij} \Big/ \sum_{j=1}^{m} w_j$$

kde r_{ij} je ohodnocení (rating) i-té alternativy A_i pro j-tý atribut na numerické srovnávací škále.

Nejlepší alternativa A^* se vybere takto:

$$A^* = \left\{ A_i \;\middle|\; \max_i U_i \right\}$$

Fuzzy varianta předpokládá, že jak w_j tak r_{ij} jsou fuzzy množiny definované jako:

$$w_j = \left\{ \left(y_j, \mu_{w_j}(y_j) \right) \right\}, \ \forall j \qquad x_{ij} \in R \quad y_j \in R$$

$$\mu_{w_j} \in [0.0,\ 1.0]$$

$$r_{ij} = \left\{ \left(x_{ij}, \mu_{r_{ij}}(x_{ij}) \right) \right\}, \ \forall i, \forall j \qquad \mu_{r_{ij}} \in [0.0,\ 1.0]$$

Zbývá spočítat hodnoty účelové funkce U_i pro jednotlivé alternativy A_i:

$$U_i = \left\{ \left(u_i, \mu_{U_i}(u_i) \right) \right\}, \ \forall i$$

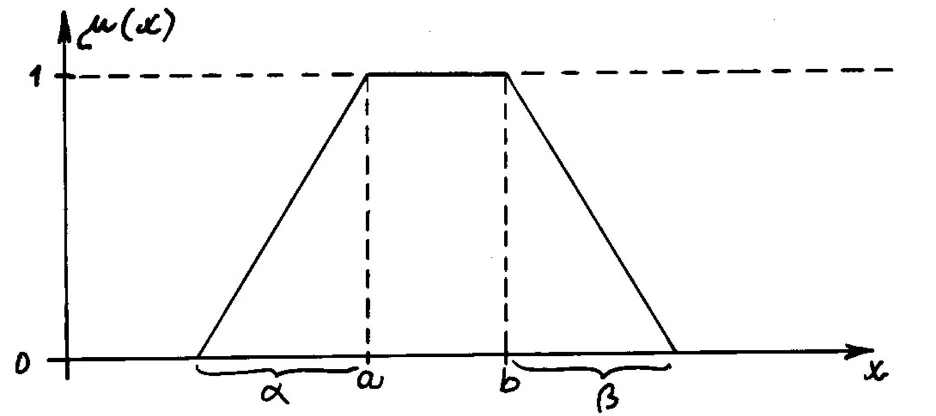
$$u_i = \sum_{j=1}^{m} y_j x_{ij} \Big/ \sum_{j=1}^{m} y_j, \ \forall i, \forall j$$

Je třeba stanovit hodnoty funkcí příslušnosti:

$$\mu_{U_i}(u_i) = \sup_{v} \left\{ \left[\bigwedge_{j=1}^{m} \mu_{w_j}(y_j) \right] \wedge \left[\bigwedge_{j=1}^{m} \mu_{r_{ij}}(x_{ij}) \right] \right\}$$

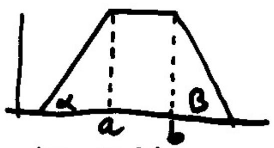
kde $v = (y_1, y_2, \dots, y_m, x_{i1}, x_{i2}, \dots, x_{im})$.

Aritmetické výpočty lze provádět různými metodami. Jednoduchou a účinnou je metoda, kterou navrhl Bonissone. Vychází z předpokladu, že ostrá a fuzzy informace v rozhodovacích problémech může být aproximována pomocí tzv. L-R trapezoidálních (lichoběžníkových) fuzzy čísel (a, b, α, β):



Bonissonova fuzzy aritmetika:

Nechť $M=(a,b,\alpha,\beta)$ a $N=(c,d,\gamma,\delta)$ jsou L-R fuzzy čísla, $M>0$, $N>0$. Aritmetické operace jsou definovány následovně:



$$M+N=(a+c,\ b+d,\ \alpha+\gamma,\ \beta+\delta)$$
$$M-N=(a-d,\ b-c,\ \alpha+\delta,\ \beta+\gamma)$$
$$M\cdot N=(a\cdot c,\ b\cdot d,\ a\gamma+c\alpha-\alpha\gamma,\ b\delta+d\beta+\beta\delta)$$
$$M/N=\left(\frac{a}{d},\ \frac{b}{c},\ \frac{a\delta+d\alpha}{d(d+\delta)},\ \frac{b\gamma+c\beta}{c(c-\gamma)}\right)$$

Pomocí těchto operací lze spočítat hodnoty účelové funkce U_i pro alternativy A_i:

$$U_i=\sum_{j=1}^{n} w_j r_{ij}$$

kde w_j a r_{ij} mohou být ostrá nebo fuzzy čísla representovaná pomocí L-R trapezoidální formy.

Numerický příklad: (dle Bonissona)

Mají se vyhodnotit 3 alternativy investic: komoditní trh, trh s akciemi, a trh s nemovitostmi vzhledem ke 4 atributům: X_1 riziko ztráty kapitálu, X_2 vliv inflace, X_3 úrokový zisk, X_4 realizovatelnost peněžního kapitálu.

Rozhodovací matice:

$$\mathcal{D}=\begin{array}{c}\\ A_1\\ A_2\\ A_3\end{array}\begin{array}{cccc} X_1 & X_2 & X_3 & X_4\\ \end{array}$$

$$\mathcal{D}=\begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4\\ \hline A_1 & \text{vysoký} & \text{v-m vysoký} & \text{velmi vysoký} & \text{přijatelný}\\ A_2 & \text{přijatelný} & \text{přijatelný} & \text{přijatelný} & \text{v-m dobrý}\\ A_3 & \text{nízký} & \text{velmi nízký} & \text{v-m vysoký} & \text{špatný}\end{array}$$

Vektor vah:

vice-méně

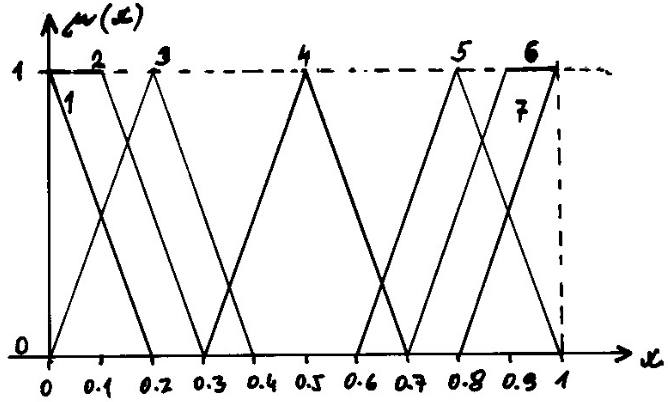
$\vec{w}$ = [v-m důležitý, v-m důležitý, velmi důležitý, v-m nedůlež.]

Význam lingvistických termů a trapez. čísel:

tvar	fuzzy č.	X_1	X_2	X_3	X_4	váhy
1	(0, 0, 0, 0.2)	vv	vv	vn	vš	vned
2	(0, 0.1, 0, 0.2)	v	v	n	š	ned
3	(0.2, 0.2, 0.2, 0.2)	vmv	vmv	vmn	vmš	vmned
4	(0.5, 0.5, 0.2, 0.2)	p	p	p	p	nez
5	(0.8, 0.8, 0.2, 0.2)	vmn	vmn	vmv	vmd	vmdůl
6	(0.9, 1, 0.2, 0)	n	n	v	d	důl
7	(1, 1, 0.2, 0)	vn	vn	vv	vd	vdůl

v ... vysoký
vv ... velmi vysoký
vn ... velmi nízký
vš ... velmi špatný
š ... špatný
n ... nízký
vmv ... více-méně vysoký
vmš ... více-méně špatný
vmn ... více-méně nízký
p ... přijatelný
d ... dobrý
vd ... velmi dobrý

vmd ... více-méně dobrý
vned ... velmi nedůležitý
ned ... nedůležitý
vmned ... více-méně nedůležitý
nez ... nezáleží
vmdůl ... více-méně důležitý
důl ... důležitý
vd ... velmi důležitý



Fuzzy reprezentace lingvistických termů

S použitím definované aritmetiky dostaneme např.:

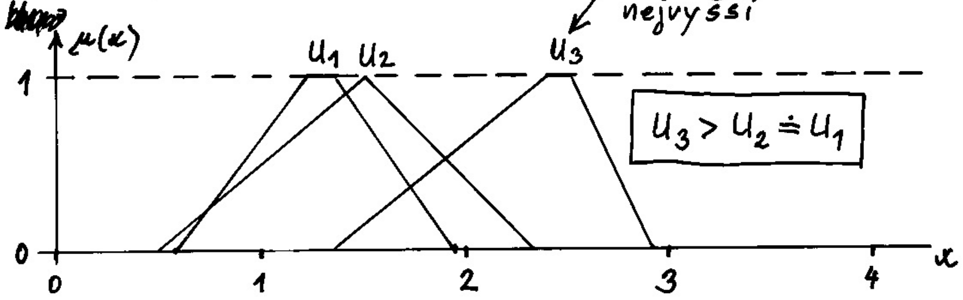
$$U_1 = \sum_{j=1}^{4} w_j\, x_{ij} = w_1(\text{vysoký}) + w_2(\text{v-m vysoký}) + w_3(\text{velmi vysoký}) + w_4(\text{přijatelný}) =$$

$$= (1.26,\ 1.34,\ 0.62,\ 0.64)$$

Ostatní (U_2, U_3, $\cancel{U_4}$) se vypočtou obdobně:

$U_1 = (1.26,\ 1.34,\ 0.62,\ 0.64)$
- $U_2 = (1.46,\ 1.46,\ 0.86,\ 0.80)$
- $U_3 = (2.32,\ 2.42,\ 0.94,\ 0.52)$

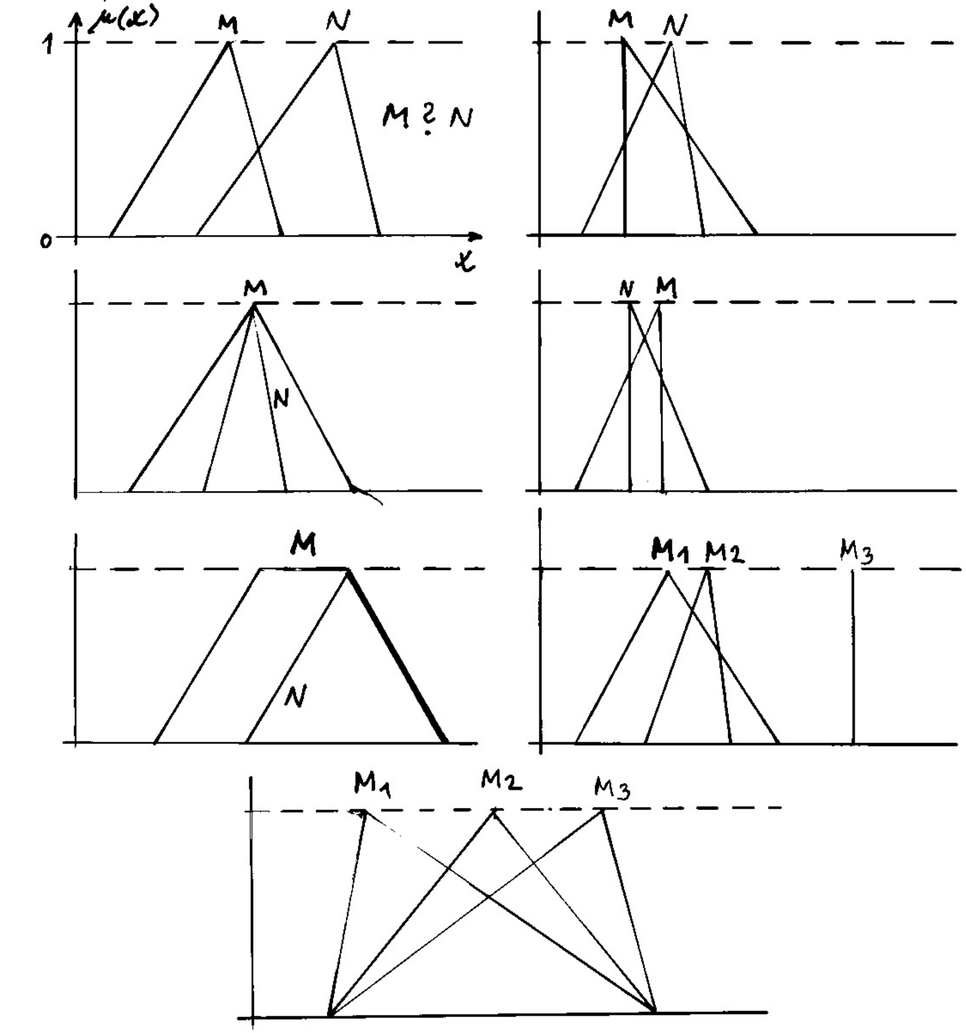
U_3 pro A_3 je nejvyšší



$$\boxed{U_3 > U_2 \doteq U_1}$$

Porovnávání fuzzy čísel

Při řešení problémů MADM je nutno po vyhodnocení účelové funkce U_i pro jednotlivé alternativy A_i výsledky seřadit. Obecně je seřazování fuzzy čísel obtížný problém, protože se pracuje s množinami čísel, takže je zapotřebí vyřešit srovnání množin obsahujících prvky s různým stupněm příslušnosti:



Fuzzy množiny a rozpoznávání vzorů (algoritmus FCM).

K technikám používaným pro rozeznávání vzorů patří tzv. řízené a neřízené rozeznávání (supervised a unsupervised pattern recognition). U řízeného je poskytnuta ke vzorku dat příslušná klasifikace, u neřízeného klasifikace (zařazení do třídy) není.

K nejznámějším algoritmům neřízené klasifikace patří tzv. fuzzy c-means (FCM) algoritmus.

Neřízené rozeznávání (klástrování - clustering)

Cílem neřízeného klástrování (shlukování podle určitých charakteristik) je hledání zajímavých vzorů či skupin v daném souboru dat. Např. tržní analýzy o zákaznících apod.

V oblasti rozpoznávání vzorů je neřízené shlukování používáno často k tzv. segmentaci obrazů (dělení pixelů v obraze do různých oblastí).

Konvenční (ne-fuzzy) metody nalézají tzv. ostré („tvrdé") hranice mezi oblastmi – pixel může náležet pouze do jediného shluku (klástru) v oblasti.

Definice: Nechť X je soubor dat (množina) a $x_i \in X$. Oblast $P = \{C_1, C_2, \dots, C_\ell\}$ na X je tzv. ostrá, pokud (a jen pokud) platí:

a) $\forall x_i \in X \;\; \exists C_j \in P$ takové, že $x_i \in C_j$

b) $\forall x_i \in X \;\; x_i \in C_j \Rightarrow x_i \notin C_k$ kde $k \neq j$, C_k a $C_j \in P$ ■

Ve mnoha problémech shlukování v reálném světě však některé datové body částečně mohou náležet do více shluků, např. pixel z obrazu magnetické rezonance může odpovídat dvěma druhům tkáně.

Uvedená skutečnost motivovala vznik tzv. shlukaž s neostrými hranicemi („měkkých" shluků) a algoritmů pro jejich vyhledávání.

Definice: Nechť X je datová množina a $x_i \in X$. Oblast $P = \{C_1, C_2, \dots, C_\ell\}$ množiny X je tzv. neostrá, pokud (a jen pokud) platí:

a) $\forall x_i \in X \;\; \forall C_j \in P \;\; 0 \le \mu_{C_j}(x_i) \le 1$

b) $\forall x_i \in X \;\; \exists C_j \in P$ takový, že $\mu_{C_j}(x_i) > 0$

kde $\mu_{C_j}(x_i)$ označuje stupeň náležení x_i do C_j. ■

Je zřejmé, že koncept neostré oblasti je podobný konceptu fuzzy množiny.

Nejčastěji se používá neostrých shluků, které splňují podmínku:

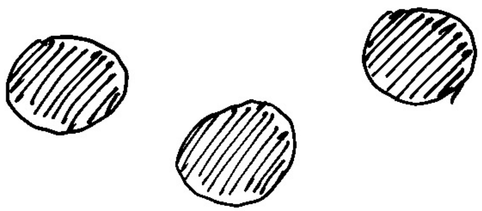
$$\sum_j \mu_{C_j}(x_i) = 1 \qquad \forall x_i \in X$$

Taková neostrá oblast se nazývá (neostrá) oblast s omezením.

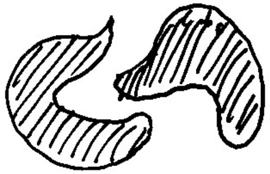
Nejužívanější algoritmus, c-means algoritmus, vytváří právě takovéto neostré oblasti.

Jsou rozeznávány obecně 3 typy shluků:

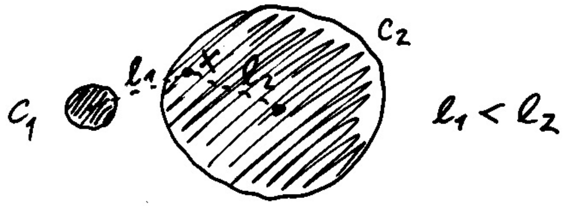
a) Kompaktní, dobře separované shluky:



b) Nekompaktní a nedobře separované shluky:



c) Kompaktní a nedobře separované shluky:



(zde je bod x blíže nějakému bodu v C_1 než v C_2).

Definice: Oblast $P = \{C_1, C_2, \dots, C_k\}$ datové množiny X má kompaktní separované shluky tehdy (a jen tehdy), když libovolné 2 body ve shluku si jsou blíže než je vzdálenost mezi 2 body v různých shlucích: $\forall x, y \in C_p \quad d(x,y) < d(z,w)$ kde $z \in C_q$, $w \in C_r$, $q \neq r$ a d značí míru vzdálenosti.

Za předpokladu, že datová množina obsahuje c kompaktních dobře separovatelných shluků, cílem ostrého algoritmu c-means je:

1) najít středy těchto shluků
2) určit shluky (tj. názvy, přiřazení) každému bodu z datové množiny.

Známe-li středy shluků, pak určitému bodu $x_i \in X$ přiřadíme shluk s nejbližším středem:

$x_i \in C_j$ pokud $\|x_i - v_j\| < \|x_i - v_k\|$, $k = 1, 2, \dots, c$, $k \neq j$

kde v_j označuje střed C_j.

K nalezení středů shluků je zapotřebí znát kritérium pro jejich vyhledání. Jedním takovým kritériem je součet vzdáleností mezi body v každém shluku a jejich středy:

$$J(P,V) = \sum_{j=1}^{c} \sum_{x_i \in C_j} \|x_i - v_j\|^2$$

kde V je vektor středů shluků, který má být nalezen. Jedná se o minimalizaci J, které je ovšem i funkcí oblasti P určených středy V v souladu s rovnicí. Proto se jedná o iterativní hledání:

1) Výpočet aktuální oblasti založený na aktuálních shlucích.
2) Modifikace aktuálních středů shluků s použitím gradientního sestupu k minimalizaci funkce J.

Iterace končí, když vzdálenost mezi středy ve dvou iteračních krocích po sobě je menší než předem stanovený práh.

Algoritmus pro fuzzy c-means

středy shluků (cluster means)

FCM (fuzzy c-means) zobecňuje „ostrý" c-means. Bod zde může náležet do více shluků.

Cílová funkce je rozšířena dvěma způsoby:

1) je přidán parametr m jako váha,
2) J obsahuje stupně příslušnosti:

$$J_m(P,V) = \sum_{i=1}^{k} \sum_{x_k \in X} \left(\mu_{c_i}(x_k)\right)^m \| x_k - v_i \|^2$$

kde P je fuzzy oblast v X tvořená shluky $c_1, \dots, c_k$. Parametr m je váha určující stupeň, do něhož částeční členové shluku ovlivňují výsledek shlukování.

Cílem je rovněž minimalizovat J_m nalezením dobré oblasti pomocí v_i. Kromě toho je nutno nalézt také μ_{c_i} takové, aby J_m byla minimalizována.

Věta: Fuzzy oblast s omezením $\{c_1, c_2, \dots, c_k\}$ může být lokálním minimem cílové funkce J_m pouze tehdy, jsou-li splněny tyto 2 podmínky:

1)

$$\mu_{c_i}(x) = \frac{1}{\sum_{j=1}^{k} \left(\frac{\| x - v_i \|^2}{\| x - v_j \|^2} \right)^{\frac{1}{m-1}}}$$

pro $1 \le i \le k$, $x \in X$

2)

$$v_i = \frac{\sum_{x \in X} \left(\mu_{c_i}(x)\right)^m \times x}{\sum_{x \in X}^{n} \left(\mu_{c_i}(x)\right)^m}$$

pro $1 \le i \le k$

FCM tedy aktualizuje v_i a μ_{c_i} iterativně pomocí obou výše uvedených rovnic, dokud není dosaženo konvergenční kritérium.

Iterační algoritmus FCM

$FCM(X, c, m, \varepsilon)$

X: množina dat
c: počet shluků, které se mají vytvořit
m: parametr cílové funkce
ε: práh pro konvergenční kritérium

Inicializuj $V = \{v_1, v_2, \dots, v_c\}$

Opakuj (REPEAT)

$V^{\text{Předchozí}} \leftarrow V$
Spočítej μ_{c_i}
Spočítej v_i a aktualizuj V

Dokud (UNTIL) $\sum_{i=1}^{c} \| v_i^{\text{Předchozí}} - v_i \| \leq \varepsilon$

Příklad: Nechť je dána datová množina šesti bodů, z nichž každý má 2 vlastnosti: F_1 a F_2. Předpokládejme, že chceme body rozdělit do 2 shluků ($c = 2$). Nechť $m = 2$ a počáteční hodnoty $v_1 = (5, 5)$ a $v_2 = (10, 10)$.

	f_1	f_2
x_1	2	12
x_2	4	9
x_3	7	13
x_4	11	5
x_5	12	7
x_6	14	4

Počáteční μ_c jsou spočteny pomocí rovnice:

$$\mu_{c_1}(x_1) = \frac{1}{\sum_{j=1}^{2} \left(\frac{\|x_1 - v_1\|}{\|x_1 - v_j\|} \right)^2}$$

$$\|x_1 - v_1\|^2 = 3^2 + 7^2 = 9 + 49 = 58$$
$$\|x_1 - v_2\|^2 = 8^2 + 2^2 = 64 + 4 = 68$$

$$\mu_{c_1}(x_1) = \frac{1}{\frac{58}{58} + \frac{58}{68}} = \frac{1}{1 + 0.853} = \underline{0.5397}$$

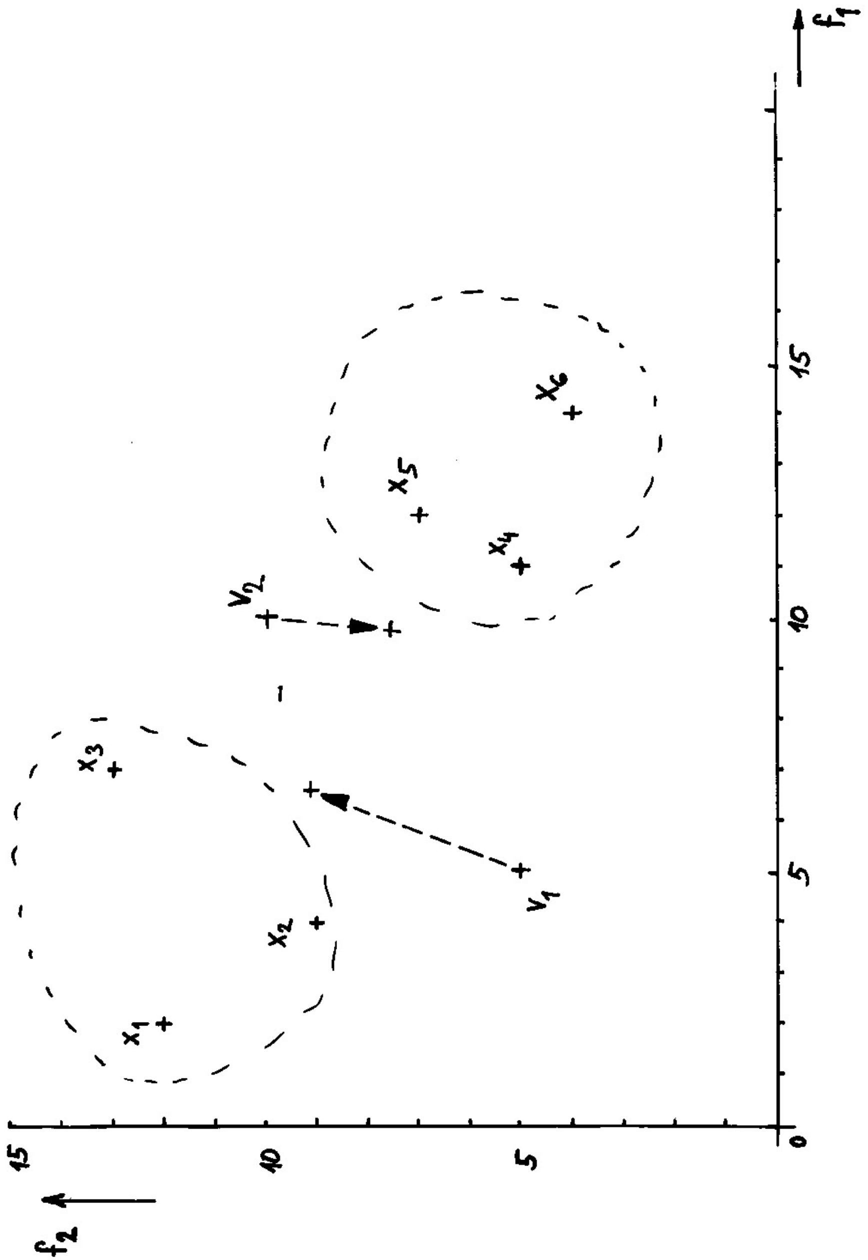
c_1

Podobně dostaneme:

$$\mu_{c_2}(x_1) = \frac{1}{\frac{68}{58} + \frac{68}{68}} = \underline{0.4603}$$

$$\mu_{c_1}(x_2) = \frac{1}{\frac{17}{17} + \frac{17}{37}} = 0.6852$$

$$\mu_{c_2}(x_2) = \frac{1}{\frac{37}{17} + \frac{37}{37}} = 0.3148 \quad \text{atd.} \dots$$



Z toho vyplyne, že x_1 a x_2 jsou spíše v C_1, zatímco ostatní body spíše v C_2.

Aktualizace v_1 a v_2:

$$v_1 = \frac{\sum_{k=1}^{6} \left(\mu_{c_1}(x_k)\right)^2 \times x_k}{\sum_{k=1}^{6} \left(\mu_{c_1}(x_k)\right)^2} =$$

$$= \frac{0.5397^2 \times (2,12) + 0.6852^2 \times (4,9) + 0.2093^2 \times (7,13) + \cdots}{0.5397^2 + 0.6852^2 + 0.2093^2 + \cdots}$$

$$\frac{\cdots + 0.4194^2 \times (11,5) + 0.197^2 \times (12,7) + 0.3881^2 \times (14,4)}{\cdots + 0.4194^2 + 0.197^2 + 0.38812} =$$

$$(5 \quad , \quad 5)$$

$$= \left(\frac{7.2761}{1.0979}, \frac{10.044}{1.0979}\right) = (6.6273, 9.1484)$$

Podobně pro v_2: $(9.7374, 8.4887)$

Aktualizovaný v_1 se posune blíže k centru formovanému pomocí x_1 a x_2, x_3 zatímco v_2 k druhému centru formovanému x_4, x_5, x_6.

FCM zaručuje konvergenci pro $m>1$.
FCM najde lokální minimum (či sedlový bod) funkce J_m.
Výsledek závisí nejen na volbě <u>c</u> a <u>m</u>, ale i na $V^{počáteční}$.

Fuzzy geografické informační systémy

Geografické informační systémy (GIS) označují obecně skupinu metod pro řízení a zpracování kartografické informace. Obvykle se pod pojmem GIS rozumí organizovaný soubor SW systémů a zeměpisných dat schopných representovat, ukládat a přistupovat veškerou formu informace související s geografií. Srdcem GIS je prostorová databáze, v níž jsou informace popisující umístění a tvar zeměpisných charakteristik v termínech bodů, čar a ploch.

Celkově je systém tvořen složitou kombinací kartografických údajů, majících rozličné formy (mřížka, barva...) kombinované s obrazy a náčrtky.

Je rovněž zapotřebí účinný dotazovací systém, který umožní zkombinovat a zpřístupnit veškeré rozličné formy informace konsistentním a účinným způsobem.

Přesnost prostorové databáze

Problém přesnosti byl vždy vnímán jako kritický pro úspěšnou implementaci a dlouhodobou aplikaci technologie GIS. Hodnota GIS např. jako nástroje pro rozhodování závisí na možnosti a schopnosti vyhodnocení informace, na níž je rozhodnutí založeno.

Uživatelé GIS musí tedy být schopni odhadnout a určit původ a stupeň chyby v prostorové databázi, sledovat průchod této chyby operacemi GIS a odhadnout přesnost tabulární i grafické výstupní informace.

Vyhodnocování přesnosti prostorové databáze v rámci GIS zahrnuje množství konceptů, metod a modelů.

Situace je dále komplikovaná tím, že význam různých dimenzí přesnosti je funkcí datového typu, aplikace a zdrojů chyb.

Existuje množství různých kategorizací chyb a nepřesnosti, např.:

a) 1. měření chyb v prostorových databázích, 2. přesnost kartometrických odhadů, 3. chyby zavlečené během kompilací dat, 4. šíření chyb skrze operace GIS, 5. obecné problémy přesnosti prostorové databáze.

b) 1. chyby v souborech dat, 2. chyby v klasifikaci dat, 3. chyby v analýze dat, 4. chyby ve vizualizaci obsahu map, 5. chyby při representaci reality.

Problémy s nejistotou

Existuje množství zdrojů nepřesností v GIS, vyjádřitelných pomocí různých druhů nejistoty:

1. vliv proměnnosti chyb
2. vliv vágnosti
3. vliv neúplnosti (např. kvůli nesprávné vzorkovací frekvenci)

Interpretační nejistota a inherentní nejednoznačnost může být ilustrována pomocí označování dat získaných např. z obrazů satelitu LANDSAT. Obrazy jsou napřed zpracovány pomocí neřízené klasifikace (klasifikace bez dohledu), přičemž jsou obrazy tříděny do tříd.

Poté jsou výsledkům přiřazeny (subjektivní) interpretace člověkem (např. zda snímek představuje zemi nebo moře ...). Zde se jedná o subjektivní spojování objektivně získaných tříd obrazů s lingvistickými koncepty, existujícími v mysli interpretujícího. To může ve skutečnosti vést k různé interpretaci téhož různými interprety. Vzniká problém při ukládání do databáze, protože takovéto inherentně nepřesné koncepty vyžadují zvláštní interpretaci.

V aplikacích pracujících s informací získanou vzdálenými senzory a navíc z více zdrojů použitých k formulaci geografických dat je problém nepřesnosti a nejistoty zesílen.

Na data prostorové databáze se aplikuje množství operací, které jsou korektní za předpokladu, že atributy a jejich vztahy byly stanoveny a priori přesně.

Obecně není tento předpoklad oprávněn, protože nepřesnost, vágnost a přibližnost jsou součástí prostorových dat.

Byly navrženy různé modely nejistoty pro informaci v GIS; tyto modely zahrnují myšlenky z oblasti zpracování přirozeného jazyka, non-monotonické logiky, fuzzy množin, pravděpodobnostní teorie aj.

Každý takový model je adekvátní pro jiný typ nejistoty. Zdroje nejistoty/nepřesnosti/vágnosti jsou v principu tři:

① Náhodnost: může se vyskytnout v případech, kdy pozorování může předpokládat interval hodnot.

② Vágnost: může vzniknout jako výsledek nepřesnosti v taxonomických definicích.

③ Neúplnost podkladů: může se vyskytovat v případech, kdy je např. použita určitá vzorkovací frekvence při pozorování, čímž „vznikají" chybějící hodnoty, nebo se použily náhradní hodnoty, atp.

Aplikace fuzzy databáze na GIS

Zdrojem pro většinu informace uložené v databázi GIS je výstup klasifikací a analýz aplikovaných na mnohospektrální data získaná pomocí vzdálených senzorů (např. kamera v družici). Mnohé z těchto klasifikátorů jsou založeny na fuzzy-množinovém přístupu. Následkem toho se objevuje potřeba ukládat tyto výsledky do GIS způsobem, který umožní reprezentaci fuzzy výsledků.

Tato representace závisí na vztahu mezi nejistotou a nepřesností v modelování a ve skutečných datech, která mají být ukládána.

Vágní model a přesná data

Tento přístup je zaměřen na fuzzy representaci, i když je možné získat data přesně. Takový postup je vhodný tehdy, když dvě (či více) domény mající vzájemný vztah, nemohou být z dat rekonstruovány perfektně. Uvažme příklad, kde je v záznamu uložena velikost mokřiny či suché země:

místo	Mokřina	Suchá země
2. 10	29	60
2. 11	25	62

V tomto případě lze snadno připustit, že mokřina i suchá země nemusejí být navzájem výlučné třídy informace (může existovat určité překrytí mezi oběma doménami). I když tato situace není neobvyklá, dosud neexistuje mnoho pokusů o její fuzzy representaci.

Přesný model a nepřesná data

V tomto případě je sémantika datového modelu vyjádřena přesnými logickými omezeními, avšak vlastní informaci je obtížné či (z podstaty) nemožné zachytit přesně. Toto je nejobvyklejší model používaný pro representaci fuzzy dat a pro její správu a zpracování.

Existují v zásadě dva přístupy k řešení problému:

a) Klasifikátor povrchu poskytuje stupeň příslušnosti oblasti vymezené nějakou souřadnicovou mřížkou (buňkou) do příslušné domény:

buňka	země	voda
0101	1.00	0.14
0102	1.00	0.09
0103	1.00	0.85
0104	0.47	1.00

b) Stupeň příslušnosti representuje sílu závislosti mezi klíčem a atributem. V následujícím příkladu je <u>klíčem</u> buňka a <u>atributem</u> povrch:

buňka	povrch	μ
0101	země	1.00
0101	voda	1.00
0102	země	1.00
0102	voda	1.00

Vícenásobné hodnoty nejsou neobvyklé. Často bývá určitý pixel klasifikován různě. Např. použije-li se jak digitální klasifikace tak interpretace fotografie k určení třídy povrchu, vznikne typická podmnožina pixelů.