Posudek na diplomovou práci ing. Davida Fišera
„Využití data miningových metod při zpracování dat
z demografických šetření“

Předkládaná diplomová práce se skládá ze 102 stran textové části a dalších 4 stran citované literatury. Cílem práce bylo popsat a následně demonstrovat na modelové úloze principy dolování dat z databází (viz abstrakt). Práce je rozdělena do pěti dále strukturovaných kapitol: *1. Úvod*, *2. Data minig, 3. Statistické šetření, 4. Data miningová úloha,* následované *5. Závěrem.* Již v úvodu je třeba se pozastavit nad používáním termínu data mining (DM), protože se v češtině již delší dobu používá standardní termín dolování dat, stejně jako dobývání (lepší než dolování, protože je obecnější) znalostí z databází. Problémy s terminologií se objevují i v dalších částech práce.

V úvodu autor vymezuje předmět problematiky, včetně vazby na demografii (s.13). Obávám se však, že zde začíná určité nepochopení toho, k čemu by se mohly metody DM v demografii využívat. Ta má svůj předmět zájmu (zkoumání procesu reprodukce lidských populací) a k tomu používá demografickou analýzu, demografickou metodologii a další instrumenty. Právě v oblasti demografické metodologie, např. pro dosažení lepší kvality demografických dat (dopočty), by se mohly metody a instrumenty využít, nikoliv však *ad hoc* zkoumáním zda by nástroje DM nemohly objevit v datech nějaké (dosud neznámé) „zajímavosti“ (s.13). To souvisí s tím, že DM byl původně vyvíjen pro business, nikoliv pro vědecké bádání.

Druhá kapitola začíná vymezením pojmů, to že autoři používají termíny v odlišném vymezení, vyplývá dle mého názoru z relativní mladosti oboru. Je dobře, že autor práce cituje různé koncepty, bohužel se však nevyvaroval chyb: typ regrese (s. 18) je přeci matematicko-statistickou metodou a obecně do ní patří jak kvantitativní, tak i kvalitativní, jak spojité, tak i nespojité veličiny. Pouze se liší pro jednotlivé typy dat různé metody regrese, např. logistická regrese pro kvalitativní veličiny, ta se dá ale nahradit zobecněným regresním modelem (GLM) – podle charakteru jednotlivých statistických úloh. V tab. 2 (s. 19) jsou uvedeny praktické příklady, které bohužel jsou demografii na hony vzdáleny, chtělo by to reálnější příklady, kde se dají metody DM použít. Zbývající část kapitoly docela pěkně a podrobně vysvětluje jednotlivé metody, až na ty nešťastné příklady (např. s. 28: párky, hořčice, rohlíky) a poněkud nejasnou práci se statistickými pojmy, např. (s. 35): „Velkou výhodou neuronových sítí je zejména to, že jsou koncipovány na práci s numerickými atributy, na rozdíl od rozhodovacích stromů a pravidel, které jsou určeny primárně na práci s kategoriálními daty“. Oponent netvrdí, že rozumí neuronovým sítím, ale domnívá se, že rozumí statistice – stavět numerické atributy proti kategoriálním datům, nějak nedává smysl. Přehled výpočetních (alias softwarových) nástrojů v části 2.5. je však docela pěkný.

Třetí kapitola je věnována statistickým šetřením a jsou zde popsány všechny typy statistických šetření tak, jak je popisuje základní literatura (Čermák, Valentová). I zde se nachází určité nejasnosti, viz např. tvrzení o statistické efektivitě (s. 50). Základním kritériem všech výběrových šetření je jejich reprezentativnost (brána samozřejmě z různých hledisek), co se míní efektivitou, není z textu patrné. Mezi metodami sběru dat není uvedena ta, která je v mnoha statistikách nyní aktuální: přebírání informací z administrativních zdrojů dat (zde se dá očekávat opět možnost využití v DM). Chybné je tvrzení o využití regresní analýzy jen pro numerické (autor měl asi na mysli kvantitativní proměnné): logistická regrese se využívá právě pro kvalitativní proměnné. Čtvrtá kapitola je věnována konkrétní aplikaci, tj. data miningové úloze. Zcela chybné je tvrzení o nedostupnosti primárních dat ze sčítání. To, že ČSÚ nenabízí data ze sčítání na svých webových stránkách je snad z hlediska jejich ochrany jasné, je však docela jednoduché se datům dostat. V současné době se data ze sčítání nachází i na Přírodovědecké fakultě UK, v době vzniku této práce tomu tak zřejmě nebylo.

Pokud se autor práce rozhodl pracovat s daty z projektu IPUMS, je to jeho volba, zde se nachází data 17 evropských populačních censů, American Community Survey však není náhražkou sčítání (s. 57). Co se týče výběru proměnných pro DM, bylo by vhodnější pracovat s proměnnými ekonomické aktivity, ta se ve sčítáních standardně sleduje, non-response je zde rovněž vysoká. Rovněž není jasná vazba zkoumání výše platu (s. 58) a hledání nuggetů, kde jsou definované analytické otázky na úplně jiná témata (až na první dvě). Obecné charakteristiky (např. věk, rodinný stav) nijak nevážou na předchozí text – logika této části kapitoly není příliš jasná.

Poslední kapitola shrnuje výsledky především předchozí kapitoly. Vysvětlení je hledáno především z hlediska kvantifikace DM metod. Pokud autor dospěl pomocí DM nástroje (AA kvantifikátoru) k závěru (s.102) o vyšším podílu obyvatel určité rasy na určitém území, tak není jasné, zda je to obecné tvrzení, které by potvrdil jiný kvantifikátor ale hlavně by byla třeba demografická interpretace a nikoliv skončit u potenciální zajímavosti této skutečnosti.

Jedná se o práci nesporně zajímavou, v rámci práce je zpracováno množství informací z oblasti, která by mohla být pro demografii přínosem, aplikační část práce je však diskutabilní. S touto (nikoliv formální výhradou) doporučuji práci k obhajobě.

Ing. Jaroslav Kraus, Ph.D.
6. září 2015