

Mgr. Miroslav Otáhal, Ph.D.

Oddělení koordinace šetření u domácností

Český statistický úřad

Na padesátém 81, 100 82, Praha 10

Posudek oponenta diplomové práce

Bc. Markéta Nárožná: Imputace chybějících hodnot v rozsáhlých datových souborech

Diplomová práce studentky druhého ročníku oboru Aplikace matematiky v ekonomii má 64 stran, 19 stran příloh, uvádí 27 literárních pramenů (z nichž pouze 4 jsou v češtině, zbytek je v jazyce anglickém) a je rozčleněna do šesti kapitol a tří příloh.

První z nich (str. 6-7) stručně popisuje šetření Životní podmínky, které je každoročně organizované Českým statistickým úřadem, se kterým diplomantka při psaní práce spolupracovala.

Druhá kapitola (str. 8-13) se věnuje mechanismům vzniku chybějících hodnot. Systematicky rekapituluje tři základní mechanismy a zavádí jednoduchý matematický aparát k jejich popisu.

Třetí kapitola (str. 14-18) je zaměřena na vizualizaci nekompletních dat. Je v ní uveden rozsáhlý přehled statistických grafů, ale kromě jednoho zde nejsou uvedeny žádné příklady, čímž se vytrácí přehlednost textu.

Ve čtvrté kapitole (str. 19-32) je hezky zpracován přehled metod imputací chybějících hodnot. Je zde bohužel několik drobných chyb v uvedených matematických vzorcích - na straně 25 v sumách kvadrátů pod odmocninou má být místo y_{ij} člen y_{ik} (u obou definovaných vzdáleností) a u třetí chybí transpozice vektoru $(y_i - y_j)$, který zleva násobí matici Σ^{-1} . Trochu těžkopádné je zavedení zobecněného lineárního modelu (na jehož podrobnější a přesnější zavedení ale v práci určitě nebylo místo).

Pátá kapitola (str. 33-35) přibližuje čtenáři šetření Životních podmínek prováděné na ČSÚ. Text by byl pochopitelnější při lepším odlišení jednotlivých typů chybějících hodnot, což se dá omluvit chybějící praxí studentky. V tabulce 2 na straně 34 řádek 2 (Příspěvky do soukromého penzijního pojištění - pravděpodobně výdaje) nezapadá do kontextu ostatních řádků (které jsou čistě o příjmech).

Praktická část práce je v kapitole 6 (str. 36-58). Je zde ukázáno jak použití dříve zavedených statistických grafů tak i použití rekapitulovaných metod imputace, obojí v softwaru R.

V přílohách jsou na 19-ti stranách vytištěné formuláře, pomocí nichž ČSÚ provádí šetření Životních podmínek v ČR.

Předkládaná diplomová práce tak, jak je pojatá, není po matematické stránce nijak zvlášť náročná. Cenný je na ní zejména komplexní, ucelený a srozumitelný přehled o používaných metodách imputace chybějících hodnot rozsáhlých datových souborů a samostatně vypracované příklady na reálných datech. Tento postup psaní diplomové práce se v aplikované matematice vyskytuje běžně.

Diplomantka musela prostudovat velké množství anglicky psané literatury a vyzkoušet si teoretické znalosti na praktických příkladech. Oponovaná práce je logicky velmi dobře členěná na přiměřeně velké samostatné a na sebe navazující celky. Je jasně patrný záměr práce a rozsah věnovaný jednotlivým kapitolám je adekvátní. Narazit ale lze i na velmi těžkopádné věty

a příliš dlouhá souvětí. Je to pravděpodobně důsledkem vyhýbání se používání matematické symboliky, přepisem z anglicky psané literatury nebo nedostatkem praktických zkušeností při řešení reálných problémů. Práce je sice bez použití vzorců, definic, vět a důkazů dobře čitelná, ale k tomu, aby byla jednoduchou příručkou pro výzkumníky v oblasti výběrových šetření, kteří se nechťejí matematikou hlouběji zabývat, by mohla být ještě propracovanější.

Celková úprava práce je na vysoké úrovni a množství chyb a překlepů je minimální. Proložení vlastního textu příklady příkazů softwaru R považuji za užitečné a demonstrativní.

Diplomantka prokázala schopnost samostatné orientace v komplexním problému, který je doposud zpracován pouze v anglicky psané literatuře. Patrná je též její schopnost práce v softwaru R. Matematický aparát na úrovni základních kurzů statistiky (s výjimkou zmínky o zobecněných lineárních modelech) je redukován na uvedení několika základních vzorců, z čehož se nadá zvládnutí vyšší matematiky diplomantkou posoudit. V práci trochu chybí větší samostatná polemika nad podstatou problému nebo nad závěrečnými výsledky. Z textu naopak číší zájem studentky o dané téma a pozitivně hodnotím dobrou strukturovanost a provázanost jednotlivých kapitol. Za předpokladu úspěšné obhajoby navrhuji vzhledem k výše uvedenému hodnocení B.

V Praze, dne 30. dubna 2013