

Oponentský posudek na disertační práci  
*Ing. Lukáš Sobišek: Shluková a regresní analýza mikropanelových dat*

Disertační práce ing. Sobiška je velmi obsáhlá (162 stran) a profesionálně napsána. Již na první pohled je z ní patrné nemalé úsilí, pečlivost a vytrvalost autora při shromažďování znalostí, literárních údajů i vlastních výsledků. Dobrou ilustrací rozsahu materiálu prostudovaného při přípravě práce je reprezentativní, 232 položkový seznam použité literatury. Práce je napsána svěžím jazykem a srozumitelně. Anglický abstrakt i členění do kapitol i přehled materiálu stručně uvedený na začátku každé z nich ukazuje autorovu záběhlost v prezentaci odborného textu (což dobře koresponduje s přiloženým Přehledem publikační činnosti). Způsob výkladu prozrazuje nezanedbatelné pedagogické kvality ing. Sobiška. Z textu jsou ovšem patrné i rozsáhlé praktické zkušenosti ing. Sobiška se statistickou analýzou reálných dat získané v průběhu nesporně úctyhodné řady datově-analyticky orientovaných pracovních pozic které již během své dosavadní kariéry zastával. Z výčtu dosavadních zaměstnání i z textu samotné práce je patrné že autor má široký a nikoli jednostranný/specializovaný záběr. To je ostatně explicitně patrné jak z výčtu jeho odborných zajjmů, tak z toho že v současnosti pracuje jak na VŠE tak na 1. lékařské fakultě. Tyto zkušenosti i znalosti konkrétních zajímavých problémů dokáže autor v textu dobře využít pro motivaci různých přístupů které v disertaci rozpracovává. L. Sobišek v textu dobře propojuje své ekonomické a statistické znalosti. To je patrné i na kvalifikované volbě příkladů z ekonomického modelování. Kromě toho má ale dobrý přehled i o aplikacích metod pro panelová/longitudinální data i v mnoha jiných oblastech. Nejde přitom ale ani zdaleka jen o načtené znalosti – výčet problémů a reálných dat na nichž autor pracuje (na str. 24-25 i v samotném Závěru) je impozantní.

Jako přípravu pro vývoj a popis vlastních metod práce obsahuje rozsáhlý a dobře napsaný úvod do lineárních smíšených (LME), zobecněných lineárních modelů se smíšenými efekty (GLMM) a kratší diskusi marginálních modelů a technik pro jejich odhad (GEE). Velmi krátce jsou zmíněny i některé modernější partie regresních modelů se smíšenými efekty (GAMM) či Bayesovský přístup k inferenci pro korelovaná data (tato pasáž je ale opravdu rudimentární a ve srovnání s ostatními kapitolami působí poněkud nevyzrále a „nuceně“ – práce je obsažná a klidně by se obešla i bez ní). Dobře je zpracován i přehled shlukovací metodologie. Velká pozornost je přirozeně věnována shlukování longitudinálních trajektorií (jako příprava pro vývoj autorem navržené metodologie). Probírané koncepty jsou dobře ilustrovány nejen standardními vzorečky ale i informativními a dobře provedenými obrázky.

V rámci přípravy pro vlastní práci autor nastudoval nejen velké množství literatury ale i prozkoumal řadu různých dostupných softwarových implementací metod pro longitudinální analýzy. Své zkušenosti podrobně a přehledně shrnul do poměrně obsáhlé kapitoly 4 (str. 51-64). Vedlejším produktem této přípravné aktivity je tak materiál potenciálně užitečný pro začínající zájemce o analýzu longitudinálních dat (např. pro studenty jako doplnková četba či výzkumníky z různých oborů).

Poněkud méně šťastná je autorova úporná snaha o český překlad některých zavedených anglických termínů (ta ovšem není jen autorovým specifikem - v Čechách účast v soutěži o vlastní, co nejoriginálnější překlad který nikdo jiný nezná a tedy ani nepoužívá, velmi oblíbená). Některé z takto vytvořených „českých názvů“ jsou vyloženě úsměvné (např. konstanta kontrolující zmatenosť, spojovací funkce, kontinuální poměr, zastavený poměr, přilehlé rozdelení apod.), jiné zcela zbytečné (průsečík jako překlad anglického intercept namísto zaužívaného absolutního člena, sklon jako překlad slope apod.). V případě slova „prostorový“ je nově zavedená terminologie až téměř matoucí – jde zcela proti ve statistice hluboce vžité konotaci související buď s klasickou geostatistikou (a jejími modernějšími na modelech

založenými alternativami) nebo s abstraktnějším pojetím prostoru (jako v případě Markovských náhodných polí, MRF a podobných struktur) – nic z toho ovšem text nemá na mysli.

V práci se (podobně jako v jiných pracech tohoto rozsahu) vyskytují různé drobné nekonzistence ve značení (viz např.  $n_i$  versus  $T$  v popisu scénářů, drobný zmatek v dimenzi matic  $X_i, Z_i$  a následného značení pod vztahem (2.12), typografická chyba ve vzorci (2.15) apod.), ale není jich mnoho – i jinak je typografie velmi pečlivá. Na str. 90 (a dále) se popisuje simulační schéma ve kterém jsou parametry  $\beta_0, \beta_1$  generovány jako náhodné. Jsou (uvnitř daného shluku) tedy individuálně-specifické? Pokud ano, měly by jejich indexy obsahovat  $i$  ( $\beta_{0i}, \beta_{1i}$ )?

Občas se v textu také vyskytují ne definované pojmy typu připomínající až žargon srozumitelný odborníkovi ale nikoli běžnému čtenáři („hodnoty v dlouhém formátu a shlukovací proměnné v širokém formátu“).

Na některých (naštěstí nemnoha) místech se vyskytují nepřenosná tvrzení která mohou být leckdy až matoucí. Např. opakované tvrzení o tom že s regresními modely upravenými pro mikropanelová data je spjato úskalí v nejednoznačnosti odhadu jejich parametrů - nejde, jak by se z výroku mohlo zdát, o nějaký problém s neidentifikovatelností některé z probíraných tříd modelů ale o problém nahlížený buď čistě z numerického/výpočetního pohledu, nebo z pohledu praktického – v souvislosti se specifikací modelu vhodného pro danou strukturu a rozsah dat. Podobně ambivalentní výroky se vyskytují v souvislosti s několikrát opakovaným domnělým problém s „definicí pevných a náhodných efektů“ (v definici na úrovni konkrétního pravděpodobnostního modelu samozřejmě žádny problém není – ten nastává až s volbou vhodného modelu pro danou praktickou situaci).

Smíšené modely a shlukování tvoří jádro původního příspěvku této práce. Je třeba vyzdvihnout originalitu celého přístupu – autor jde svou vlastní cestou. Nutno zdůraznit že kompletně samostatně od počáteční formulace problému, přes návrh původních metod, softwarové implementaci (ta je k dispozici ve formě R kódu na přiloženém CD) až k výpočetnímu ověření na simulovaných datech. Nápad shlukovat individuální trajektorie a využít příslušnosti ke shlukům jako informace v následném longitudinálním modelování, stejně jako (poněkud nešťastně nazvaný) „regresní model prostorových shluků“ jsou dobře prakticky motivovány – zejména viz např. Stachová, Sobišek (2015), Král, Stachová, Sobišek (2014). Navržený přístup je orientován spíše „algoritmicky“ (tedy spíše jako Metoda než jako skutečný statistický Model jež důsledkem by pak byla metodologie přísně odvozená z vlastnosti modelu). Důsledkem jsou pak velmi komplikované statistické vlastnosti odhadů – jejich teoretická analýza (byť asymptotická) je myslitelná jen ztěží. Jediným realisticky použitelným nástrojem k průzkumu vlastnosti navržených postupů jsou pak simulační studie. To je i postup který autor v práci používá – ověřuje výkonnost svých metod datech simulovaných za různých scénářů stupňující se složitosti. Kromě toho výsledky svých metod autor ilustruje na reálných datech.

Drobná terminologická poznámka – zatímco autor považuje obě navrhované metody (i] shlukování trajektorií následované použitím indikátoru shlukové příslušnosti jako vysvětlující proměnné, ii] blokový prostorový model) za dvoukrokové, my bychom ii] považovali spíše za metodu tříkrokovou (charakteristiky, shlukování, analýza „prostorových“ bloků). Poznámka věcná, I: přístup i] nezohledňuje nejistotu spojenou s odhadem shlukové příslušnosti (natožpak odhadem počtu shluků), proto odhady středních chyb (a tím i hodnoty) získané přímočarým použitím shlukových indikátorů ve standardním SW pro smíšený model (jako např. v tab. 8.4) nejsou tak úplně korektní. Poznámka věcná, II: v přístupu ii] použité charakteristiky mohou být korelované (přinejmenším nejsou záměrně konstruovány jako ortogonální). Proto se pro následné shlukování zdá být velmi přirozenou

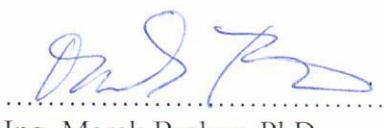
metrikou Mahalanobisova vzdálenost. Dále: je třeba dát pozor na to, že  $\hat{\beta}_{ct}$  odhadu budou velmi často i dosti výrazně korelovány přes  $t$  uvnitř  $c$  – to bude komplikovat různá srovnání či vyhledávání extrémních hodnot zmiňovaná na str. 80 (formalizovaný výpočet toho „jak velké je velké“ bude na rozdíl např. od standardního LME velmi komplikovaný). Poznámka věcná III: verze ICC diskutovaná na str. 80 (založená na „odhadu variability trajektorií“  $S_{y,B}$ ) není tak úplně korektní (jde o analog přímočarého odhadu náhodného efektu založeného na rozptylu individuálních odhadů pořízených jako efekty pevné – o něm lze s elementární teorií dokázat že je systematicky vychýlen).

Celkově je přístup této práce cenný i po pedagogické stránce. Snaží se novým způsobem těžit ze souvztažnosti mezi heterogenitou a korelací v datech. Ing. Sobíšek na něm rozhodně demonstroval penzum svých znalostí. Problém tradičně řešený pomocí mikropopisu heterogenity založeného na statistické formulaci korelačních (kovariančních) vlastností obchází pohlcením co možná největší části heterogenity do variability mezi shluky a standardním modelováním s jednodušší korelační strukturou (v extrému až s nezávislostí) uvnitř (víceméně) homogenních shluků. Přístup je to zajisté originální. Jak je vidět z provedených simulací, jsou situace kde může být relativně úspěšný a hledat si tak doménu svého použití. Častější ale asi bude nadále přístup založený na strukturovaném statistickém modelování korelační/kovarianční struktury v datech. Navržený „regresní model prostorových bloků“ by mohl být užitečný zejména jako neformální nástroj explorativní analýzy dat, pro předběžnou kontrolu předpokládaných vlastností dat před formalizovaným modelováním tradičnějšího statistického zaměření (tedy např. pro inspekci (ne)homogeneity, vyhledávání outlierů v prostoru trajektorií, různého chování sledovaného procesu v různých časových bodech apod.). Zajímavý je v tomto kontextu také návrh indexu datové kvality (DQI).

V práci popsaný dvoukrokový přístup k odhadu směsného modelu v longitudinálních datech není ani zdaleka jediný možný. Lze zformulovat plně formalizovaný statistický model např. se směsí v náhodných efektech (nebo jiných latentních proměnných) který povede ke shlukování v principu podobnému tomu které navrhuje autor. Odhadu z takového modelu jsou zajisté výpočetně složitější ale umožňují kompletně a korektně zohlednit nejistotu v odhadech jakýchkoli parametrů. To je při dvoukrokovém přístupu obtížné (například je obtížné zohlednit nejistotu v odhadu shluků, včetně nejistoty odhadu počtu shluků a exaktní vliv této nejistoty na kovarianční matici odhadů regresního modelu). Určitě by nebylo od věci se seznámit s R balíčkem (package) mixAK doc. Komárka (a také s jeho prací Komárek, A., Hansen, B. E., Kuiper, E. M. M., van Buuren, H. R., and Lesaffre, E. (2010). Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. Statistics in Medicine, 29(30), 3267–3283).

Připomínky a podněty (záměrně podrobně rozvedené) jsou zamýšleny hlavně jako náměty pro budoucí práci. Předloženou práci ing. Sobíška rozhodně doporučuji k obhajobě před příslušnou komisí. Její obsah, stejně tak jako další materiály předložené spolu s ní hodnotím velmi kladně - jsou dle mého názoru více než dostačující pro úspěšné zakončení disertace na VŠE.

V Praze dne ..... 20.12.2016



Ing. Marek Brabec, PhD

# Hodnocení disertační práce v oboru Statistika (VŠE v Praze)

(vyplňuje školitel a oponenti - příloha posudku oponenta)

Jméno doktoranda: Ing. Lukáš Šobášek

Název práce: Shlašová a rozhovorní analýza  
měřopanologických dat

Jméno hodnotitele: Ing. David Brabec, PhD

## Kritéria hodnocení

1. Odpovídá název práce jeho obsahu (zcela, částečně)? ANO, ZČBLA

2. Je vymezení cílů vyhovující, je uvedeno v úvodu, abstraktu? ANO

3. Jak jsou vytčené cíle splněny, je shrnuto v závěru, abstraktu? ZČBLA SPLEŇENY, SHRUVTÍ OPOMÍJÍCÍ

4. Jsou zřetelně odlišeny metody převzaté z literatury a vlastní přístupy? ANO  
DOBÉZ A ZRBTBLWE OPLISBY

5. Co je vědeckým přínosem (např. z hlediska teorie statistiky či metod analýzy dat)? VÝUZ VLASTNÍ METODY

6. Úroveň rešerše poznatků a použité literatury ve zkoumané oblasti: VBLNc DOBRA

7. Úroveň aplikací, hodnocení a porovnání metod (zařazení tabulek a grafů): VZLICB KVALITN.

8. Způsob vyjadřování, jazyková úroveň: BBZ UYHRA D

9. Jsou řádně definovány používané pojmy, zkratky a symboly? ANO

10. Formální úroveň (matematických výrazů, tabulek, grafů): V POŽADKU

Datum: 19.12.2016



podpis