Posudek školitele na doktorskou disertační práci

Doktorand: Ing. Štěpán Chalupa

Studijní program: Systémové inženýrství a informatika

Studijní obor: Informační a znalostní management

Školicí pracoviště: Univerzita Hradec Králové, Fakulta informatiky a managementu

Školitel: Doc. RNDr. Zdena Lustigová, CSc.

Název práce: Využití nástrojů data miningu v hotelnictví a cestovním ruchu

Práce Ing. Štěpána Chalupy nepochybně představuje přínos pro obor hotelnictví a cestovní ruch, ve kterém se pokročilé statistické metody používají spíše výjimečně a pokud ano, tak spíše formou obecných matematických modelů.
Doktorand, který je prakticky zaměřen, chtěl zpracovat reálná data a své případné modely pak chtěl postavit přímo na nich.

Práce samotná, včetně cílů, se vyvíjela a souvisela s úrovní aktuálního poznání doktoranda, i aktuální dostupností dat. Samozřejmě se v ní odráží i slepé cestičky, kterými doktorand nutně musel projít, aby k určitému poznání dospěl.
Tento dlouhodobý vývoj pak patrně vedl k určité nejednotnosti ve formulaci cílů, která je patrná nejen v abstraktu (zejména anglický je poněkud nesrozumitelný), ale i v práci samotné.
Doktorandovi se bohužel také nepodařilo úplně přehledně popsat tzv. „předchozí paradigmata“, o kterých zmiňuje, že je na nich postaven předchozí výzkum a podniková praxe. Bohužel ani neidentifikoval jejich nedostatky.
Podobnými problémy trpí také formulace hypotéz, které nevyplývají z nutnosti opravit zmiňovaná nevhodná paradigmata, ale spíše až z dat samotných. Jejich opodstatnění není úplně zřejmé.

Patrně z výše uvedených důvodů působí kapitoly 2, 3, a 4 jako deskriptivní úvod, nikoliv jako jasná analýza předchozích přístupů, s identifikací případných rozporů a z nich vyplývajících autorských hypotéz. Svůj podíl na tom mohl mít i fakt, že v dostupné literatuře v oblasti hotelnictví není nadbytek prací s jasně identifikovanou metodologií, dobře popsaným vzorkem respondentů či jednotek (například hotelů), o smysluplném samplingu či dokonce jakkoliv zaslepených studiích ani nemluvě. Prostě autoři těchto prací jsou vděční za jakákoliv data, která mají k dispozici. To je patrné i

z přehledu studií (str. 29-41), kde doktorand nejprve zamýšlel provést meta-analýzu, ale pro naprostou nesourodost datovou i metodologickou toto nebylo možné. Ačkoliv statistické zpracování těchto dat bylo patrně na profesionální úrovni, o samotných datech a metodice jejich sběru lze pochybovat, často však není ani uvedena. To by mělo vést (nejen doktoranda) k obecnější úvaze o dlouhodobé krizi opakovatelnosti (replikability) v současné vědě, a to nejen sociální či ekonomické.

Na druhou stranu je třeba zdůraznit, že z hlediska rozsahu datového vzorku, který je opravdu velmi rozsáhlý, a způsobu jeho zpracování, je předkládaná práce tohoto zaměření (obor hotelnictví a cestovní ruch) výjimečná. Odhad parametrů empirického regresního modelu na základě výběrové kovariance a výběrového rozptylu považuji za originální (viz zmiňované spoluautorství v práci Petříček a kol z roku 2020). I zde se však autor dopouští drobných nedostatků. Například není jasné, proč pro zjištění/odhad cenové elasticity (autor používá termín „měření“, který nepovažuji za vhodný), byla použita log log regresní analýza.

V práci jsou z hlediska statistického některé formální (i neformální nedostatky), které by před jakoukoliv případnou publikací měly být opraveny. Uvádím namátkově pouze některé:
1/Ocenila bych například v tabulkách 7 a 8 nejen údaje o mediánu, ale i průměru, případně Q1 a Q3. To by umožnilo rychlý náhled na typ rozdělení a zejména jeho šikmost. Rovněž zde (a nejen zde) postrádám grafické zobrazení, například alespoň formou box plotu.
2/Grafy 1, 2, 3 například mají pouze ilustrativní charakter, v podstatě z nich nelze téměř nic vyčíst. Graf 3 například uvádí OCC Corporate a OCC Leisure, ale nejsou rozlišitelné.
3/V naprosté většině tabulek doktorand používá k popisu průměru a směrodatné odchylky řecká písmena mí a sigma, ta ale dle zvyku slouží k popisu populace, nikoliv výběru.
4/Výše zmíněný průměr a směrodatná odchylka (v autorově podání písmena mí a sigma) by měla být použita v případě normálního Gaussova rozdělení, a i když je lze dobře předpokládat, autor neprovedl jeho ověření a ani nezmínil tento předpoklad.
5/Autor zmiňuje nutnost normálního rozdělení pro provedení Studentova T-testu, (Cituji: „*t-test při hladině významnosti α=0,05 a jeho oboustranné rozdělení. Samotné testování předpokládá normální rozdělení, a proto je nutné pracovat s daty očištěnými o extrémní abnormální případy (tabulka 10, Corporate Epd (očištěné).“), ale* způsob očištění neuvádí a to může být značný metodologický problém. Pro případnou publikaci doporučuji ověřit normalitu všech rozdělení, u kterých ji autor předpokládá, a neprovádět „očištění“ dat jak se mi hodí…

Rovněž formální úprava práce, anglicko české nejednotné popisky i způsob vyjadřování, nejednotnost ve statistických parametrech a jejich popisu, jsou jistými, byť pochopitelnými, nedostatky, kterých by se měl autor napříště vyvarovat.

Celkově lze konstatovat, že předkládaná práce prezentuje metodologicky originální a netriviální postupy, aplikované na velké objemy reálných dat. Nahrazení modelu RFM modelem varianty EFM (Elasticity, Frequency, Monetary) může být a jistě i bude

předmětem diskuze. Netradičnost, originalita a aplikabilita předkládaných postupů však zcela jistě odpovídá nárokům disertačních prací.
Práci doporučuji k obhajobě.

V Praze 2.1. 2022

doc. RNDr. Zdena Lustigová, CSc.

školitel