Jak nepoužívat víceúrovňové modely? Na okraj textu o výzkumné efektivitě slovenských vysokých škol

Petr Soukup
Katedra sociologie FSV UK, Praha

Tato stať je reflexí na článek Kanovského (2018), který se meritorně věnuje vědecké efektivitě slovenských vysokých škol a užívá mj. víceúrovňové modelování. V tomto příspěvku se zaměřuji zejména na provedené analýzy, meritum článku hodnotím pouze okrajově, Na počátku musím přiznat, že můj vztah k víceúrovňovému modelování je poměrně osobní. Byl jsem v česko-slovenském prostoru první, kdo o těchto modelech detailněji referoval (Soukup 2006). Nadto jsem modely opakovaně využíval ve výuce (jak studentů, tak akademiků) i publikovaných článcích, nyní dopisuji knihu, kde se tématu dále věnuji. Cítím tedy jistou zvýšenou míru odpovědnosti za používání těchto modelů a myslím, si že vzhledem k jejich komplikovanosti je více než dobré dodržovat standardy v této oblasti. Právě prismatem této zvýšené odpovědnosti se chci zaměřit na publikovaný článek (Kanovský 2018).

Věcná reflexe textu

Před rozborem statistických postupů užitých v citovaném článku se chci krátce vyjádřit i k věcnému zaměření a výsledkům. Text si klade za cíl (s. 430) prověřit hypotézu o tom, že je malý podíl vysokých škol, které mají zásadní podíl na vědeckém výzkumu na Slovensku[1]. Zkoumání této hypotézy probíhá velice diskutabilním způsobem. Prvním problémem je dle mne operacionalizace, kdy podíl na vědeckém výzkumu (správně by asi měl být vědecký výkon, či produkce) je měřen skrze objem prostředků, které vysoká škola získá od státu „na vědu“ v příslušném roce. Pro srovnatelnost je tento objem vztažen k počtu tvůrčích zaměstnanců. Zde se nabízí dvě otázky k autorovi článku. Proč jsou jako tvůrčí zaměstnanci bráni jen ti, co mají celý úvazek (srov. s. 430)? Proč je index efektivity konstruován jako podíl mezi počtem zaměstnanců a finančními prostředky a nízká hodnota (pod 1 znamená efektivitu), vysoká (nad 1, znamená neefektivitu)? Jako velký problém (i ve světle výsledků) lze vnímat změny v rozdělování prostředků za vědu v SR, kdy váha kritérií se meziročně měnila. Autor sám na str. 443 píše, že zjištěné změny (v zásadě hlavní náplň článku) lze z 89 % vysvětlit změnami metodiky. Jinými slovy, žádné velké změny ve vědeckém výkonu se nejspíš neudály a jejich modelování je tak zbytné, resp. problematické. Analyticky je totiž text v zásadě analýzou

[1] Dodejme, že přes deklaraci této hypotézy se téměř celý text věnuje desetiletému vývoji, k hypotéze o podílech se vrací autor až na konci textu.

časových řad. Ty musí pro modelování splňovat dva základní požadavky, stejné časové i věcné vymezení. S časem problém není, ale věcné vymezení napříč roky variuje a v důsledku toho pozorujeme (umělé) změny. Ještě jedním problémem dle mého názoru trpí analyzované řady. Neznám detailně slovenský systém financování vědy, ale předpokládám (s ohledem na ostatní systémy v Evropě), že mnohé ukazatele pro financování vědy jsou brány za více než jeden rok, často nadto s několikaletým zpožděním. To věcně znamená, že výsledky nevypovídají o letech 2008 – 2017, jak uvádí název článku. Analyticky je ovšem podstatnější provázanost pozorování v čase, resp. nereflektování tohoto v analýze samotné (srov. dále).

Statistická reflexe textu

Věcné problémy článku nejsou ovšem jediné, samotná analýza vybízí k mnoha otázkám a skýtá mnoho problémů. Nejdříve uvedu několik obecnějších postřehů, poté se budu věnovat rozboru použitých analytických technik a interpretací výsledků.

Obecně lze u dat za deset let a 20 škol (v zásadě cenzu všech slovenských veřejných VŠ za posledních 10 let) vyslovit pochybnost, zda se má uplatňovat statistické testování a zejména zda se mají užívat složitější postupy typu víceúrovňových modelů. Připomeňme, že statistické testy byly vyvinuty pro randomizované experimenty a náhodné výběry, o jejich užití pro celopopulační data se vede mezi statistiky diskuse (srov. Soukup 2019 a lit. tam uvedená). I pokud bychom připustili spíše okrajový názor o superpopulacích a užívání statistických testů pro celopopulační data (autor recenzního příspěvku jej nesdílí a preferuje pro tyto situace popisnou statistiku, srov. Soukup – Rabušic 2007), nabízí se další problém. V případě všech testů v článku je základ velice malý (10 pozorování v čase). Je namístě se ptát, jaká byla síla použitých statistických testů (osobně odhaduji max. 20-30 %). Další obecný problém, který spatřuji v textu je záměna statistické a věcné významnosti. Autor nesprávně užívá slovo „signifikantní“ nebo „výrazné“ (s. 433) pro vyhodnocení výsledků statistické významnosti. I když jde o častý problém (srov. Soukup 2019), jde o chybu, nadto zavádějící interpretaci. Nic na tom nemění skutečnost, že díky malému počtu pozorování jsou statisticky významné jen velké hodnoty koeficientů. Správně měly být interpretovány právě tyto koeficienty, a nikoliv P hodnoty.

Vše výše popsané se plně vztahuje zejména na testy monotonie trendu v počátku analytické části, netřeba tedy více argumentovat. U testu zlomů (change points) je situace ještě složitější. Autor vybral poměrně neobvyklý test (to by samo o sobě nebylo špatné). Nicméně použití testů je zcela v rozporu s tím, co autor o testu deklaruje v obecné části (s 433). Dle autora je test vhodný pro normálně rozdělenou veličinu. Dodejme, že již toto tvrzení je

chybné, test předpokládá normální rozdělení reziduí (dodejme, že normálně rozdělená časová řada je v zásadě nemyslitelná). Test pro body zlomu, ale meritorně slouží pro jiné typy dat, než užívá autor (dlouhé meteorologické řady) a snaží se nalézt jeden bod, kdy se řada z určité průměrné úrovně dostane na jinou. Předpokládá se, že pro tento posun je nějaký důvod (v meteorologii změna přístroje či stanoviště). Autor tohoto vůbec nedbá, zlomů v jeho řadách je typicky více (většinou jsou způsobené změnou metodiky hodnocení vědy v SR), a proto je test nepoužitelný. Autor pak interpretačně zcela chybně a neudržitelně popisuje rozdíl mezi (chybnými) výsledky testu a správnými úsudky o zlomech, pro které stačí pohled na graf a selský rozum. Závěr o tom, že vizualizace ne vždy umožňuje detekci zlomů by mohl být úsměvný, pokud by nebyl v článku v časopise s impakt faktorem, nadto s odkazem na texty statistiků. Takto jde o jasně chybné tvrzení založené na chybné aplikaci testu, jehož výsledek není v souladu se skutečností. Zde by bylo namístě dodání errat k článku, kde by došlo k odstranění těchto chyb. Obecně zde opět zde platí klasické doporučení, že prostý popis by byl mnohem lepší než užívání statistických testů. K testu zlomů je nutno ještě dodat, že autor sice udává vzorce (neudává odkud je převzal, z původního textu nejsou, tam je symbolika jiná), ale zcela absentuje vysvětlení symbolů ve vzorci, což je nepřijatelné. Zarážející je, že výsledky o zlomech (byť chybné) nejsou v dalších interpretacích nijak využity.

Samostatnou kapitolou je pak použití víceúrovňového modelu. Jednoduše řečeno je tato část kompletně problematická a zcela neúplná. Není cílem recenzní stati vše detailně popsat, upozorním na základní problémy. Při užití víceúrovňových modelů bývá zvykem, že autor buď detailně slovně model popíše, nebo uvede rovnici. Článek sice naoko nabízí obojí, ale nikoli v uspokojivé podobě. Náznak rovnice je na straně 434, ale nejde o rovnici konkrétního analyzovaného modelu, ale o zcela obecnou rovnici, nadto operující s obecnými pojmy (subjekt), maticemi a vektory, které jsou pro běžného čtenáře zcela nesrozumitelné. Pokud by bylo cílem replikovat provedenou analýzu dle uvedené rovnice to nelze. Nepomůže ani přidaný popis (opakovaně na s. 431 a 440), že hodnoty za roky jsou fixní efekty a hodnoty za školy jsou efekty náhodné. Jde totiž zcela jasně o chybný popis. Chápu, že diskuse o fixních a náhodných efektech je komplikovaná, ale model pro vývoj užívá na první úrovni jednotlivé časové okamžiky a na druhé úrovni kontext (zde školy)[2]. Ve dvouúrovňovém modelování vždy platí, že pro koeficienty proměnných první úrovně lze volit mezi fixními a pevnými efekty, na druhé úrovni to nelze. Je tedy vyloučeno, aby efekty pro školy byly náhodné (samozřejmě není vyloučeno, že koeficient pro vývoj v letech bude fixní).

[2] Snahou je postihnout základní vývoj (pokles, stagnace, růst) a poté individuální odlišnosti trajektorií. Stručný popis je uveden v závěru článku Soukup (2006).

Z mého předchozího popisu navíc plyne, že náhodné či fixní nejsou proměnné (jak se uvádí v článku), ale koeficienty jim náležející. Tedy i slovně je popis chybný. Již výše byl zmíněn problém provázanosti měření v čase, i toto musí model reflektovat a musí být uvedeno, jaký model pro korelovanost reziduí byl zvolen, volba by měla být též zdůvodněna. Shrnu-li stručně, specifikace modelu je zcela nedostačující, místy chybná.

Zvláštní je též prezentace modelu, interpretaci nelze precizně zhodnotit právě s ohledem na chybějící části prezentace modelu. I když se autor v předchozích analýzách zarytě drží výsledků statistické významnosti (P hodnot), u víceúrovňového modelu je čtenář nenajde. Nenajde v zásadě žádný výsledek sloužící ke zobecnění (standardní chyby odhadů, hodnoty testových kritérií či intervaly spolehlivosti). Ve dvojím provedení (proč?) nalezne čtenář jen hodnoty koeficientů (označených jako random efekty, tabulka č. 5, obrázek č. 6). Dále chybí hodnoty klasických ukazatelů jako je ICC či R^2. Výsledek logaritmu LR testu (s. 439) má hodnotu, která vypadá velice zvláštně a vůbec není jasné, jak z ní autor usoudil na vhodnost modelu (běžně se užívá záporný dvojnásobek této hodnoty). Lze jen odhadovat, že je uvedeno něco jiného než hodnota testového kritéria. I pokud by se jednalo o tuto hodnotu (s nepřesným označením), nelze ji minimálně bez počtu stupňů volnosti statisticky vyhodnotit. Popis hodnot koeficientů na s. 440 je značně tendenční a odhlíží od relativnosti výsledků. Autor zde rozděluje školy na výzkumně efektivní a neefektivní, ale zapomíná na to, že toto dělení uměle vytvořil tím, že výkony srovnává mezi sebou a relativizuje. Při takovém dělení bude vždy část (cca polovina) nad průměrem a polovina pod ním. Možná jsou všechny slovenské školy efektivní, či všechny neefektivní, ale liší se jen v míře efektivity. Vše je totiž v analýze relativní a to musí reflektovat interpretace, což se ovšem neděje.

Tímto se již dostáváme k závěru článku. Zde už čtenář autora v zásadě nezajímá a autor prezentuje, co se mu hodí. Celá strana 442 je v zásadě neověřitelná na výsledcích publikovaných v článku. Daleko horší je ovšem samotný závěr (s. 444). Zde už autor rezignuje na své relativní pojetí efektivity (umožňující srovnání, byť s výše uvedenými výhradami). Natvrdo a zcela absolutně se zde uvádí výkon pěti vysokých škol (aniž by se zohlednila jejich velikost) a autor konstatuje, že tyto školy zajišťují cca 70-80 % podíl vědecké činnosti VŠ na Slovensku. Zvláštní je, že závěr (i v dalších částech) zcela odhlíží od výsledků analýz (byť místy chybných či problematických, viz výše) a vypadá jako když si autor vyřizuje své účty s tvůrci slovenské vzdělávací politiky. Namísto vědecké argumentace však nastupují zcela normativní a nepodložená tvrzení, což je pro článek v odborném časopise zcela nepřijatelné.

Závěrem jen mohu dodat, že mne jako člověka dlouho působícího v oboru sociálněvědní statistiky zaujal i medailonek autora, kde píše, že se kromě antropologie věnuje využívání robustních a bayesovských metod v sociálních

vědách. Při náhledu na publikační a výukový profil autora se mi nic podobného nepodařilo nalézt. Na základě publikovaného článku soudím, že tvrzení v medailonku je výrazně nadsazené.

Obecně mohu svůj recenzní příspěvek uzavřít několika doporučeními, snad na ně jako dlouholetý učitel a autor textů z oblasti sociálněvědní analýzy dat mám nárok. Prvním doporučením je, aby autoři používali ideálně co nejjednodušší přístupy, kterými lze problém řešit. Druhé doporučení se týká složitějších technik. Zde by měl autor nejdříve techniku základně ovládnout a ideálně se o konkrétní aplikaci poradit s někým, kdo ji rutinně užívá (lépe pak rozvíjí či reflektuje v odborných textech). Při publikaci výsledků složitějších technik je dobré dodržovat publikační standardy (typicky se dají dovodit z knih těmto technikám věnovaným, případně z četby článků). V neposlední řadě, s ohledem na současné požadavky na tzv. reproducibility research či open science, lze doporučit, aby autoři v článcích popsali všechny nezbytné kroky, které by umožnily, výzkumnou studii kompletně replikovat. Ostatně patří dnes již k dobrému zvyku, že autor společně s článkem zveřejní i svou datovou matici, případně (u složitějších technik) příkazy, jimiž dospěl k výsledkům. Ostatně autora recenzovaného článku bych k tomuto kroku vyzval bylo by možné přesněji posoudit jednotlivé analytické kroky. Ostatně poté se může ukázat, že jsem se v některých popisech lehce mýlil, resp. že můj odhad byl místy nepřesný.

LITERATURA

KANOVSKÝ, M., 2018: Výskumná efektivita slovenských vysokých škôl: kvantitatívna analýza trendov 2008 – 2017. Sociológia – Slovak Sociological Review 50(4): 429-447.
DOI: https://doi.org/10.31577/sociologia.2018.50.4.17.

SOUKUP, P., 2006: Proč užívat hierarchické lineární modely? Sociologický časopis / Czech Sociological Review 42 (5): 987-1012

SOUKUP, P., 2019: P a d (Používání statistické a věcné významnosti v českých sociálních vědách).“ Sociologický časopis / Czech Sociological Review 55 (2): 215-254

SOUKUP, P. – RABUŠIC, L., 2007: Několik poznámek k jedné obsesi českých sociálních věd – statistické významnosti. Sociologický časopis/Czech Sociological Review 43 (2): 379-395.