Implementácia metodiky prideľovania dotácii zo štátneho rozpočtu verejným vysokým školám

Jozef Číž, Viktória Pravdová

10. júna 2023

1 Úvod

Témou projektu bola implementácia metodológie prideľovania dotácii zo štátneho rozpočtu verejným vysokým školám na rok 2023. Metodológia sa zameriava na spravodlivé rozdelenie dotácií vo výške 18 658 146.30€ zo štátneho rozpočtu medzi verejné univerzity na Slovensku. Naším cieľom bolo zreprodukovať systém rozdeľovania, ktorý zohľadňuje faktory ako publikačná excelentnosť, schopnosti hanobiť peniaze a populáciu fakulty. Závery poskytujú hodnotenie účinnosti metodológie a jej dôsledky pre verejné univerzity, prispievajú k transparentnosti a potenciálnym vylepšeniam pri budúcich prideľovaniach dotácií.

2 Proces

2.1 Zdroje dát

Dáta sme čerpali z príloh na stránkach ministerstva školstva, konkrétne https://www.minedu.sk/33286-sk/rozpis-dotacii-zo-statneho-rozpoctu-verejnym-vysokym-skolam-na-rok-2023/ a https://www.minedu.sk/32743-sk/navrh-predlozeny-na-vyjadrenie-reprezentaciam-vysokych-skol/. Na týchto stránkach sú vo viacerých prílohách poskytnuté Excelovské tabuľky obsahujúce do istej miery predspracované dáta, ako zoznamy počtu zamestnancov pracovísk, relevantných odborných publikácii, monografií, grantov a iné. Rovnako tu vieme nájsť aj dokument obsahujúci prílohu 10, popisujúcu metodiku analýzy daných údajov.

V rámci metodiky boli spomínané aj iné, nepriložené prílohy. Oslovili sme pána Kanovského z ministerstva o doloženie niektorých údajov, avšak nedostali sme to po čom sme prahli. Niektoré vybrané údaje ohľadom publikácii a monografií sme kontrolovali s podkladným zdrojom publikácii na CREPČ.

2.2 Jazyky, knižnice

Na spracovanie dát sme používali Python a knižnicu Pandas v odporúčanej forme Jupyter zápisníku. Na štatistické modelovanie sme použili jazyk R s modifikovaným referenčným kódom uvedeným v metodike. Našu implementáciu spolu s relevantnými dátami je možné nájsť na GitHube.

2.3 Postup

Postup reprodukovania výsledku metodiky je principiálne jednoduchý. Skladá sa zo štyroch častí:

1. Načítanie a spracovanie všetkých dát z jednotlivých hárkov príloh do vhodného formátu
2. Vytvorenie regresných modelov pre štatistické distribúcie
3. Vytvorenie pomocných metrík a výslednej alokácie
4. Zobrazenie výsledkov

Načítavanie a spracovanie všetkých dát robím pomocou knižnice Pandas. Cieľom spracovania dát je pre každú odbornú oblasť určiť pracoviská, ktoré majú aspoň 5 zamestnancov a pre ne prislúchajúce publikácie, monografie a granty z danej oblasti.

Po spracovaní dáta pošleme R programu, ktorý natrénuje dva regresné modely pre štatistické distribúcie Sichelova (pre publikácie) a Tweedie (pre granty), pre každú oblasť zvlášť. Tieto distribúcie boli v referenčnej implementácii zvolené na základe štatistickej literatúry a dobrej zhody s modelovanými dátami. Výstupom sú predikcie určujúce očakávaný priemerný publikačný výkon alebo grantový zisk na jedného priemerného zamestnanca pre pracovisko s daným množstvom zamestnancov.

Z týchto predikcií vypočítame pre jednotlivé pracoviská z-skóre na základe rozdielov od ich predikovaného výkonu. Tieto skóre pre každé pracovisko priamo určujú jeho relatívnu excelentnosť, teda pracoviská s vysokým z-skóre považujeme za dobré.

Zostáva nám určiť vrchný kvartil pracovísk a rozdeliť alokované dotácie. Postupujeme množstvom sčítavaní, váhovaní, priemerovaní a normalizovaní, pokým sa nedopracujeme k, podľa tohoto postupu, férovému rozdeleniu dotácii.

3 Problémy

3.1 Tabuľky

Prílohy obsahujú kvantum stroho pomenovaných tabuliek. Nie je jasné, ktoré hárky s dátami sú reálne použité ako podkladné dáta pre tento proces, nieto ešte či sú kompletné alebo správne. Pôvod týchto podkladných údajov a postup ako ich zreprodukovať nie je uvedený.

Rôzne tabuľky popisujúce podobné dáta sú tvorené rôznymi stĺpcami.

Veľký zlomok údajov v tabuľkách chýba (políčko je prázdne alebo obsahuje `NULL`). Toto je problém ak je daný údaj povinný (napríklad názov fakulty prislúchajúci určitej publikácii), alebo ide o údaj ktorý by bolo vhodné vedieť. Napríklad publikácie nemajú vyplnený ISBN a monografie neobsahujú žiaden identifikátor. Toto celkom spôsobuje problémy pri spracovaní dát. V metodike je zdôraznené aby jednotlivé publikácie neboli zarátané viackrát. V niektorých hárkoch však sú publikácie zjavne uvedené duplicitne (rovnaký riadok pod sebou niekoľko krát). Nemôžeme teda úplne veriť miere predspracovanosti dát a pri iných hárkoch, pri ktorých nie je možné overiť unikátnosť položiek, nemáme istotu správnosti výsledkov.

Ako bonus chceme uviesť príklad vzorcov ktoré sme našli v referenčných prílohách. Bunky okrem invalidných referencií a referencovania prázdnych buniek obsahovali aj vzorce ako

`=AVERAGE(CR22:CR26,CR3:CR14,CR15,CR16:CR17,CR18:CR21)` namiesto jednoduchšieho `=AVERAGE(CR3:C26)`.

3.2 Granty

V rámci hárkov je prevažne používaná určitá množina pomenovaní pracovísk. V tabuľke grantov je však použité celkom iné pomenovanie. Museli sme implementovať heuristiku na určenie relácie týchto pomenovaní. Okrem toho sa niektorým položkám nedá určiť príslušnosť pracovisku, keďže nemajú uvedenú fakultu alebo je ako fakulta uvedený rektorát danej vysokej školy. Nie je jasné, ako s týmito dátami narába referenčná implementácia.

Zoznam grantov obsahuje viacero stĺpcov ktoré by mohli byť použité na určenie príslušnosti grantu do odbornej oblasti. Rôzne stĺpce vedia grant priradiť do inej odbornej oblasti. Znova, z metodiky nie je jasné, ktorý stĺpec bol použitý v referenčnej implementácii. Domnievame sa, že bol dokonca použitý nevhodný stĺpec, čo má za následok nesprávne priradenie niektorých grantov.

3.3 Odlišné výsledky

V metodike je popísaný spôsob, ako váhovaným súčtom agregovať všetky publikácie a monografie pracoviska do jednej metriky. Opísané sú rôzne kategórie vydavateľstiev, na základe ktorých sa určuje váha publikácie. Ďalej sa zohľadňuje aj percentuálny prínos pracoviska pri jej tvorbe. Súčasťou tohoto procesu je nejaké násobenie dvomi a zaokrúhľovanie. Nie je nám však jasné, ako to má celé naozaj fungovať, keďže sme sa nedokázali priblížiť k referenčným výsledkom.

Iné údaje, ktoré sa v našej analýze nezhodujú, sú v niektorých prípadoch aj počty zamestnancov pracovísk. Či už v rámci porovnania jednoduchého súčtu hodnôt v jednej tabuľke a jeho výsledku v druhej. Alebo v zmysle, že tieto údaje sa nezhodujú s reálnymi aspoň v prípade našej fakulty.

3.4 Štatistika

Často sme mali problém úspešne natrénovať regresný model. Problémy sme mali rôzne, avšak nie sme si istý, čo ich spôsobovalo. Iba zdvojnásobenie hodnôt dokázalo spôsobiť, že Sichel model sa nepodarilo natrénovať.

Je otázne, do akej miery je naozaj potrebné používať tieto pokročilé modely. Z prvotného pohľadu sa nám zdá, že jednoduchá lineárna regresia by mala podobný efekt.

4 Záver

Zistili sme, že popísaná metodika nie je dostatočne presná a dáta nie sú dostatočne transparenté. Zverejnené tabuľky ministerstva obsahujú chybné vzorce ale hlavne množstvo pevne zakódovaných hodnôt, ktorých správnosť sa nedá overiť a pôvod nesprávnosti nedá určiť.

V niektorých oblastiach, ako počet zamestnancov alebo agregácia grantov, sme sa dostali pomerne blízko k referenčným výsledkom. Avšak aj v tomto prípade sme mali na mnohých miestach odlišné výsledky. Naopak v prípade publikačnej excelentnosti sme sa žiadnym spôsobom nedokázali priblížiť k referenčným hodnotám.

Kvôli veľkému množstvo uvedených problémov sa naša snaha zreprodukovať referenčné výsledky nevydarila. Zatiaľ nie je úplne jasné, či a do akej miery sú referenčné výsledky nesprávne.