Komentarji, poudarki in dodatne informacije v zvezi s primeri »big data«

Cene

Numbeo ima filtre za odstanjevanje napak. Gre za postopek urejevanja podatkov, ki je poznan tudi v uradni statistiki, npr. pri podatkih podjetij. A v uradni statistiki se avtomatsko izvaja le pri majhnih enotah, ki nimajo bistvenega vpliva na končne rezultate, medtem ko se pri večjih enotah vedno poslužijo osebnega kontakta z respondenti in/ali domenskega znanja o obravnavanem pojavu (npr. če gre za administrativne podatke, kjer nimajo stika z respondenti).

Potrebno je vedeti, da je v ozadju Numbea ena oseba in uporabniki (user-generated content), pri katerih pa težko ugotavljamo in določamo točnost, zato je veljavnost take metode težko zagovarjati. Načeloma pričakujemo, da veljavnost boljša pri skupini znanstvenikov na MIT (the Billion Prices project) ali pa v uradni statistiki, ki ima načeloma boljši dostop do podatkov in veliko znanja o tem (če, seveda, ne potvarja).

Za primerjavo podatki za Slovenijo:

- Numbeo: These data are based on 4448 entries in the past 18 months from 438 different contributors.
- SURS: 11.000 cen mesečno za neživilske proizvode, 70.000 cen mesečno za živilske proizvode (iz baz največjih trgovcev), pri čemer se te cene nanašajo na skupno 749 proizvodov/storitev.

Inflacija ima to lepo lastnost, da zanjo potrebujemo izračun indeksa (indeks cen življenjskih potrebščin = consumer price index), kar pomeni, da nas ne zanima raven cen (koliko evrov nekaj stane), ampak gibanje ravni (koliko % več moramo za nekaj plačati). Zato so morebitne pristranskosti v spletnih cenah irelevantne, če so ves čas enake.

Pomislek glede reprezentativnosti ene same spletne trgovine je na mestu, zato so v znanstvenem članku raziskovalci pretestirali več hipotez glede morebitne nereprezentativnosti in v bistvu eno po eno ovrgli, s tem pa tudi potrdili predvidevanja nekaterih argentinskih in tujih strokovnjakov, da mora vlada potvarjati kazalce.

O prednosti: takojšnja dostopnost. Inflacijo merimo na mesečni ravni (v Sloveniji boste zanjo izvedeli vsak zadnji delovni dan v mesecu). Večinoma to zadostuje. Če pa bi prišlo do kakih nihanj, v časih ekonomske krize itd., pa bi hitrejša objava prišla prav. The Billion Prices project je bil sicer postavljen zaradi potvarjanja ključnih ekonomskih indikatorjev v Argentini, s čimer so tudi pred mednarodnimi institucijami hoteli skriti ekonomske težave. Hitrost objave je bila zgolj dobrodošel stranski produkt.

Google Flu Trends

Vir, ki ni namenjen za pripravo statističnih indikatorjev, lahko doživlja spremembe z (nezaznanim) vplivom na te indikatorje. Podoben primer imamo v uradni statistiki, ko administrativni viri nadomeščajo neposredno zbiranje podatkov z anketami. Če bi se npr. bistveno spremenila davčna zakonodaja (zavezanci, davčne stopnje, kazenske sankcije itd.), bi samo to dejstvo lahko vplivalo na (boljše ali slabše) poročanje davčnih podatkov ob nespremenjeni realnosti. Torej bi se ustvaril vtis, kot da se nekaj dogaja, pa se ne. Tak primer smo doživeli, ko je bil napovedan davek na nepremičnine z obdavčitvijo vikendov, pa so se ljudje začeli "seliti" na naslove svojih vikendov.

To je tudi primer, da brez domenskega znanja lahko hitro zaidemo v napačne interpretacije (kakršnihkoli podatkov, ne samo grafikonov). Za 45 poizvedb niso uporabili nobenega teoretičnega znanja, samo korelacije, ki pa so lahko tudi umetne/lažne (spurious correlation) ali pa naključne. Data-driven approach je zelo privlačen, vendar tudi tvegan.

Razkritje metodologije je ključno za presojo v znanstvenoraziskovalni dejavnosti in tudi v uradni statistiki, zato vedno ohranimo skeptičnost do rezultatov, pri kateri ne poznamo metodologije.

Netflix

Good enough is better than the best. Ekipe so dokaj hitro dosegle nekaj odstotkov izboljšanja, najtežje pa je bilo doseči zadnji odstotek (za določen input torej dobiš vedno manj outputa, kar bi lahko označili kot padajoči mejni donos).

Podatki in njihova analiza ne moreta nadomestiti kreativnosti in prevzemanja tveganj.

50.000 kandidatom so omogočili dostop do dveh velikih baz podatkov, kar je že samo po sebi lahko problematično. Do (nedvoumne) identifikacije pa je prišlo, ko so te "anonimizirane" podatke uparili z javno dostopnimi komentarji na IMDB.

Pristranskost zaradi kvantifikacije

Noben posamični vir podatkov ne more povedati vsega. Zato tudi Facebook ob vseh podatkih, ki jih avtomatsko zajema od uporabnikov, uporablja npr. tudi intervjuje z uporabniki kot tipično metodo zbiranja kvalitativnih podatkov. Kvalitativni podatki npr. pomagajo pri razumevanju, zakaj opazimo določen pojav v kvantitativnih podatkih (npr. funkcija »hide« uporabljena kot »prebrano«, namesto kot »nočem«). Pri tem se je seveda potrebno zavedati, da kvalitativni podatki tipično prihajajo od nekaj deset enot in zato ne omogočajo statističnega posploševanja. Vendar pa kvalitativni podatki osvetljujejo mehanizme delovanja, zato je potrebno od primera do primera presoditi, v kolikšni meri jih lahko posplošimo.

Na področju informacijske podpore za pametna mesta obstaja samo nekaj velikih korporacij, ki obvladujejo trg in preko tovrstnega poslovanja prihajajo do ogromne količine podatkov, kar odpira vprašanja moči in pravic.

Promet in vreme

Nizozemski primer obcestnih senzorjev nam lepo kaže, da imamo opravka z veliko količino podatkov, a tudi z veliko šuma in relativno majhno informativno vsebino, zato je potrebnega precej čiščenja, da se do te informacije prebijemo. V poduk nam je tudi, da naprave niso nujno boljši vir podatkov od osebe, saj tudi tu zaznavamo celo serijo problemov (v tem primeru predvsem manjkajoči podatki).

Seveda pa je človek lahko tudi vir napak: ko so temperaturo odčitavali opazovalci, je lahko prišlo do precejšnjih odstopanj, tudi če so zjutraj zamudili samo pol ure pri odčitku, ker se takrat temperatura hitro spreminja. Kdaj je pa človek samo posredno vplival na napake, npr. postavitev stavbe, ki meče senco na napravo ipd.

Pametne stavbe

Pri Microsoftovem primeru se je treba zavedati, da gre za njihov »showcase«. Čeprav verjamem, da so bili prihranki ogromni, je vedno tudi vprašanje, v kakem stanju je bil campus pred vpeljavo pametnih sistemov.

Stavba Edge je bila res vrhunsko domišljena. Izpostaviti pa velja eno od osnovnih idej o organizaciji dela v tej poslovni stavbi – ker je veliko osebja na terenu, bi bil eden od ključnih prihrankov, da bi dodeljevali delovni prostor na dnevni ravni glede na vrsto obveznosti (samoten kotiček za individualno delo; diskusija za dva v prostoru, ki omogoča pogovor brez motenj drugih; sestanek ekipe v sejni sobi itd.). To se zdi logično. Pri tem pa so čisto pozabili na človeške vidike, saj tak način dela sociološko in psihološko ne zdrži: ljudje se radi vračajo na svoje mesto in tako mesto tudi sebi prilagodijo (družinska slika itd.). Je pa fino, da te pri avtomatu pričaka tvoja vrsta kave/pijače ☺.

Odvisnost od zunanjega ponudnika pri pametnih stavbah, še bolj pa pri pametnih mestih.

Tako kot velja za lokacijo, velja tudi za temperaturo, da je načeloma merjenje dobro, vendar lahko pride do napak, zato ne gre napravam nikoli slepo zaupati. Samo pomislite na primer iz šempetrske bolnišnice, kjer so na urgenci pacienti dobivali smejalni plin namesto kisika.

Target

Eden prvih negativnih primerov. Govori o ameriški verigi veleblagovnic Target, kjer naj bi analitik na osnovi transakcijskih in drugih podatkov ugotovil, da z veliko verjetnostjo (0.87) lahko identificira noseče ženske, še preden te organizirajo zabavo »baby shower«, za katere trgovine vodijo liste želja v ZDA. Ker je nosečnost eden od redkih večjih dogodkov v življenju, ko je oseba pripravljena temeljito spremeniti nakupne navade, je za trgovce izjemno pomembno, da začnejo ciljati te osebe prvi. Med proizvodi je npr. analitik identificiral specifične proizvode (npr. kreme brez parfuma in vitaminske dodatke). Problem pa je nastal, ko naj bi neki oče prišel v Target in nadrl managerja, da pošiljajo kupone za nosečnice njegovi najstniški hčerki; kasneje pa naj bi se opravičil, češ da ni vedel za hčerkino nosečnost. Primer, ki ga je sicer lansiral eden od medijev, bi sicer lahko bil tudi izmišljen ali prirejen, vendar pa ne nemogoč. Etično sporno je to, da lahko trgovec tako posega v zasebnost posameznika in informacije razkriva preko svojih aktivnosti. Če naj bi bili tiskani oglasi verige Target že takrat personalizirani, potem so se po tej zgodbi zagotovo zatekli k drugačnim pristopom – informacije so bolje zamaskirali, npr. verjetno noseči ženski poleg njej in naraščaju namenjenih izdelkov ponudili še čisto nepovezane artikle, da se osebe ne bi prestrašile vdora v zasebnost.