Metodika zberu dát

VÝPOČET SOCIAL PROGRESS INDEX PRE TVORBU LEPŠÍCH VEREJNÝCH POLITÍK V TRNAVSKOM SAMOSPRÁVNOM KRAJI

Európska únia
Európsky sociálny fond

Operačný program
Efektívna
verejná správa

APOS
Asociácia pre občiansku spoločnosť

Tento projekt je podporený z Európskeho sociálneho fondu.

Zhromažďovanie údajov predstavuje podstatnú súčasť výpočtu indexu. Je dôležité udržiavať čo najviac štandardizované procesy a záznamy, pretože počas celého procesu sa údaje zhromažďujú v rôznych fázach a rôznymi ľuďmi. Cieľom tohto metodického dokumentu je poskytnúť usmernenie a navrhnúť najlepšie postupy, hoci každý projekt si môže vyžadovať iný prístup. Je však nevyhnutné, aby bolo všetko dobre zdokumentované.

Tieto jednoduché pokyny pomáhajú orientovať sa v zložitom bludisku viac ako 40 ukazovateľov, ktoré zvyčajne tvoria index sociálneho pokroku.

1) Štruktúrovanie údajov

Rámec SPI ponúka užitočnú metódu štruktúrovania priečinkov, je veľmi intuitívny a ľahko sa v ňom naviguje. V ideálnom prípade by sa údaje mali ukladať podľa dimenzií a komponentov. Použili sa tieto skratky:

- BHN
  - NBMC
  - WS
  - S
  - PS
- FoW
  - ABK
  - AIC
  - HW
  - EQ
- Opp
  - PR
  - PFC
  - Incl
  - AAE

V niektorých prípadoch, keď jeden súbor údajov pokrýva viac ako jednu zložku, je možné pridať buď „všeobecný“ priečinok alebo súbor, ktorý nemusí byť uložený do priečinka. Platí to najmä pre zdroje, ako sú sčítanie ľudu, prieskum demografie a zdravia (Demographic and Health Survey) atď.

2) Ukladanie a manipulácia s údajmi

A) UKLADANIE SÚHRNNÝCH INFORMÁCIÍ O UKAZOVATEĽOCH NA JEDNOM MIESTE

Od samého začiatku je užitočné viesť súhrnnú tabuľku s kľúčovými informáciami o zvažovaných ukazovateľoch bez ohľadu na to, či sa dostanú do konečného výberu. Tabuľka by mala byť štruktúrovaná podľa rámca SPI a mala by obsahovať nasledujúce informácie: názov ukazovateľa, definíciu ukazovateľa, zdroj ukazovateľa, časové obdobie, prepojenie zdroja ukazovateľa, zahrnutie do konečného indexu (áno/ nie), dôvody zamietnutia, transformáciu (modifikáciu).

B) ZACHOVANIE ORIGINÁLNYCH ÚDAJOV

Výpočet SPI zahŕňa množstvo krokov a manipulácií s údajmi. Je dôležité tento proces dobre zdokumentovať a zachovať originálny súbor bez zmien, aby bolo jasné, aké boli originálne údaje, a aby sa počas vývoja indexu bolo možné k originálnym súborom údajov stále vracať. Ak sa originálny súbor zmení, niekedy nie je možné originálne údaje obnoviť. Platí to napríklad

aj pre údaje, ktoré boli skopírované manuálne z webového zdroja. Originálny súbor by mal obsahovať hlavičku alebo textové okno s informáciami o tom ako bol zhromaždený, presný zdroj a dátum. Zdroj musí byť jednoznačný. Napríklad všeobecná webová stránka neumožní lokalizovať údaje. Súbor obsahujúci originálne údaje by mal byť označený ako „originál“. Je tiež dôležité, aby sa originálny zdroj údajov používal v najväčšej možnej miere namiesto zdrojov, ktoré originálne dáta iba preberajú z iných zdrojov. Napríklad ukazovatele rozvoja Svetovej banky obsahujú veľa údajov o vzdelávaní, ale mnohé z nich pochádzajú z UNESCO - vždy je potrebné skontrolovať pôvodný zdroj údajov a prostredníctvom tohto zdroja ich aj sťahovať a odkazovať na ne vo všetkých metodických dokumentoch.

C) CHÝBAJÚCE HODNOTY

(Viď. Príloha 1: Chýbajúce hodnoty a ich doplňovanie)

Pri hodnotení globálneho indexu sa používa regresia na odvodenie chýbajúcich hodnôt. V prípade subnárodných indexov sú často vhodnejšie iné prístupy. Medzi ne môže patriť použitie historických hodnôt, priemerovanie hodnôt pre väčšinu „rovnakých“ jednotiek, použitie hodnoty pre vyššiu úroveň geografie atď. Takéto manipulácie sa ľahšie uskutočňujú v Exceli, kde jednotlivé chýbajúce dátové bunky môžu byť prepísané vhodnou hodnotou, avšak spôsob vloženia chýbajúcich údajov musí byť jasný a uvedený v dokumente. Ak je napríklad chýbajúca hodnota doplnená priemerovaním väčšiny rovnakých jednotiek, výpočet priemeru by sa mal v dokumente zaznamenať a mal by sa vytvoriť nový dokument na doplňovanie chýbajúcich hodnôt.

D) PRESUN A TRANSFORMÁCIA UKAZOVATEĽOV

Nespracované údaje sa veľmi často nejakým spôsobom transformujú alebo sa presúvajú ešte predtým, ako sa dostanú do fázy konečného výpočtu. Niekedy je ľahšie robiť takéto presuny a transformácie v Exceli, v iných prípadoch môže byť vhodnejšia STATA. Akýkoľvek presun dát musí byť dobre zdokumentovaný. Ak realizujete úpravy v Exceli, vytvorte nový súbor na základe pôvodnej sady údajov. Zachovajte všetky kroky výpočtu, nielen konečné čísla. Napríklad pri výpočte podielu obyvateľstva s vysokoškolským vzdelaním pomocou dvoch ukazovateľov - celkový počet obyvateľov a celkový počet obyvateľov s vysokoškolským vzdelaním - sa zachovajú oba ukazovatele, ako aj všetky vzorce a výpočty použité na dosiahnutie konečnej hodnoty obyvateľstva s vysokoškolským vzdelaním. Ak používate Excel, vždy zabezpečte, aby sa výpočtové vzorce uložili do dokumentu, aby bolo ľahké spätne zistiť, ako sa dosiahol. Pre všetky výpočty v STATA uložte “do-files”.

E) ODDELENIE JEDNOTLIVÝCH ÚPRAV UKAZOVATEĽOV

Pokiaľ viacero indikátorov nepochádza z rovnakého zdroja, je lepšie realizovať úpravy indikátorov oddelene. Ak pôvodný zdrojový súbor obsahuje viac ako jeden indikátor, úpravy indikátorov by sa mali vykonávať na samostatných kartách, ktoré sú príslušne označené.

3) Zjednocovanie a označovanie ukazovateľov

Po dokončení úprav údajov v Exceli môžeme začať vytvárať kompletný súbor údajov, ktorý sa importuje do STATA (alebo R) a použije sa na výpočet indexu. Tento súbor údajov by mal byť samostatnou tabuľkou a bude obsahovať všetky zvažované ukazovatele. Aby sa výpočty v STATA uľahčili, štítky indikátorov by mali obsahovať predponu podľa zložky, do ktorej patrí, napríklad: nbmc_indicatorname, ws_indicatorname, sh_indicatorname, ps_, k_, i_, hw_, eq_, pr_, pf_, in_, ae_.

NA ČO SI DAŤ POZOR:

Rôzna taxonómia pre pozorovacie jednotky:
Zdroje veľmi často používajú rozdielne taxonómie, pravopis alebo notáciu pre jednotky pozorovania. Tieto musia byť prepojené, aby bolo možné konsolidovať všetky údajové body do jedného súboru údajov. Toto by sa malo vykonať v rovnakej

tabuľke ako iné úpravy údajov. Namiesto prepísania pôvodného názvu je lepšie vytvoriť nový stĺpec, ktorý bude obsahovať názov použitý pre index. Je tiež užitočné viesť súhrnnú tabuľku celej taxonómie z rôznych zdrojov.

Výpočet Social Progress Index pre tvorbu lepších verejných politík v Trnavskom samosprávnom kraji: Príloha 1: Chýbajúce hodnoty a ich doplňovanie

CHÝBAJÚCE A CHYBNÉ HODNOTY

Mnohé údajové body môžu chýbať alebo môžu byť chybné, napríklad keď je chýbajúca hodnota nesprávne označená ako 0.
Je preto dôležité identifikovať ich hneď na začiatku a zvážiť vhodný prístup. Nasleduje niekoľko návrhov na riešenie tohto problému:

1) Identifikujte chýbajúce hodnoty a nuly

Vo všetkých indikátoroch a jednotkách pozorovaní sa identifikujú prázdne bunky a bunky s nulami. To sa dá urobiť ľahko v Exceli pomocou podmieneného formátovania. Môžete tiež použiť nasledujúce funkcie na súčet medzi ukazovateľmi a jednotkami. COUNTBLANK (rozsah), COUNTIF (rozsah, 0)

2) Identifikujte extrémne hodnoty (tzv. outliers)

Pre každý ukazovateľ uveďte extrémne hodnoty. To sa dá znova ľahko urobiť v Exceli pomocou podmieneného formátovania a zvýraznenia niekoľkých (v závislosti od veľkosti vzorky 5-10) najvyšších a najnižších hodnôt.

HĽADAJTE VZORCE – „PATERNY“

Po identifikácii hodnôt je pravdepodobné, že sa ukážu vzorce. Vo viacerých sledovaných jednotkách bude chýbať viac ako jedna hodnota vo všetkých ukazovateľoch alebo v niekoľkých ukazovateľoch bude chýbať niekoľko hodnôt vo všetkých sledovaných jednotkách. Je dôležité poznať takéto hodnotya rozhodnúť sa, či je možné tieto ukazovatele nahradiť inými, ktoré majú lepšie pokrytie, alebo či sa musí zvážiť vhodná metóda imputácie (doplňovanie chýbajúcich hodnôt).

SÚ NULY SKUTOČNE NULY?

Chyby sa stávajú každému, preto je dôležité starostlivo preskúmať hodnoty, ktoré sa zdajú nezvyčajné alebo neočakávané - napríklad extrémy alebo nuly. Majú všetky pozorované jednotky nenulovú hodnotu okrem jednej alebo dvoch? Existuje extrémna hodnota, ktorá takmer vyzerá, že desatinná čiarka nie je na svojom mieste? Všetky tieto prípady môžu byť chybné, a preto sa musia dôkladne skontrolovať. Hneď ako sú tieto hodnoty identifikované, musíme zistiť, či je hodnota správna alebo či ide skutočne o chybu. Preto je užitočné nahliadnuť do alternatívnych zdrojov, ktoré by mohli chybu potvrdiť alebo vyvrátiť alebo sa obrátiť na inštitúciu zodpovednú za súbor údajov a priamo to s nimi skontrolovať. Ak sa potvrdí, že hodnota je chybná a nemôžeme vykonať opravu, musíme s ňou zaobchádzať tak, akoby chýbala.

ČO S CHÝBAJÚCIMI HODNOTAMI?

Ak hodnoty chýbajú, index sociálneho pokroku sa nedá vypočítať v plnom rozsahu. Všeobecne platí, že v jednej zložke by nemala chýbať viac ako jedna hodnota ukazovateľa pre konkrétnu jednotku pozorovania. Existujú však výnimky. Na vyplnenie chýbajúcich hodnôt sa môžu použiť rôzne metódy. Každý ukazovateľ sa musí posudzovať osobitne, aby sa určila vhodná metóda. (Nasleduje neúplný zoznam, ktorý možno použiť ako usmernenie.)

1) Skontrolujte, prečo hodnota chýba

Je tiež dôležité hlbšie preskúmať, prečo hodnota chýba. Dôvodom môže byť skutočnosť, že obec alebo región nepredložil správu, opatrenie sa nesleduje vo všetkých správnych jednotkách, ale tiež preto, že opatrenie nie je pre tento región relevantné

alebo preto, že jeho uverejnenie by mohlo narušiť súkromie osôb. Pri posudzovaní chýbajúcich hodnôt a pri rozhodovaní o najvhodnejšom prístupe k ich riešeniu je potrebné zohľadniť všetky tieto aspekty.

Bezvýznamnosť opatrenia

Jedným z kľúčových kritérií na zahrnutie ukazovateľa do rámca SPI je jeho význam pre všetky sledované jednotky. Ukazovateľ je však niekedy koncepčne veľmi dôležitý a relevantný pre takmer všetky jednotky pozorovania. Je to typické pre vidiecke a mestské členené ukazovatele, ale možno aj iné. Napríklad globálny SPI obsahuje ukazovateľ verejnej defekácie obyvateľstva na vidieku. Singapur však nemá vidiecke oblasti. Bolo by nevhodné hodnotiť Singapur hodnotami 0. Skôr, iba na účely výpočtu, priradíme Singapuru teoreticky najlepšie hodnoty, ale pri prezentácii údajov konštatujeme, že hodnota chýba. Upozorňujeme, že tento prístup by sa mal uplatňovať iba v zriedkavých prípadoch a iba pre veľmi obmedzený počet pozorovaní.

Potlačené hodnoty

Niekedy, zvyčajne z dôvodov ochrany súkromia, môžu byť hodnoty ukazovateľov, ako je úmrtnosť matiek alebo dojčiat alebo iné zdravotné ukazovatele, potlačené. Čo znamená, že sú zhromažďované a k dispozícii, avšak nie sú verejne dostupné. V takom prípade existuje niekoľko možných prístupov. Môžeme sa obrátiť na zodpovednú inštitúciu a opýtať sa, či by boli ochotní poskytnúť nám údaje na účely výpočtu, pričom tvrdíme, že hodnota by nezostala potlačená a v našich publikáciách sa neuvádzala ako nedostupná. Je nepravdepodobné, že by to bola inštitúcia ochotná urobiť, ale možno sa to oplatí vyskúšať. Druhou možnosťou je vypočítať chýbajúcu hodnotu na základe poskytnutých informácií. Inštitúcia niekedy poskytuje podrobnejšie informácie o potlačených hodnotách - napríklad počet prípadov bol menší ako 5, 10, 20. Na základe týchto informácií by sme mohli vypočítať aproximáciu chýbajúcej hodnoty.
Ak nie je uvedený konkrétny počet prípadov, imputovaná chýbajúca hodnota by mali byť vždy nižšia ako najnižšia zaznamenaná hodnota. Dávajte pozor, aby ste nezamieňali počet prípadov s pomerom.

2) Ak hodnota jednoducho chýba

Existuje niekoľko ďalších prístupov, ktoré môžeme použiť na vyplnenie medzier, ak hodnoty „jednoducho“ chýbajú.
Pri globálnom SPI sa používa kombinácia metód vrátane regresie na odvodenie chýbajúcich hodnôt. Pri subnárodných indexoch však nemusí byť regresia tou najlepšou metódou a iné prístupy by mohli byť vhodnejšie. Medzi ne môže patriť použitie historických hodnôt, priemerovanie hodnôt pre väčšinu „rovnakých“ jednotiek, použitie hodnoty pre geografiu vyššej úrovne atď. Mali by sme sa vždy usilovať o:

- Posúdenie všetkých chýbajúcich hodnôt od prípadu k prípadu, namiesto toho, aby sa použil jednotný prístup vyhovujúci všetkým.
- Rransparentnosť metódy imputácie a dôvod, prečo sme ju vybrali.
- Neinterpretovať a priamo porovnávať imputované hodnoty so zaznamenanými hodnotami.

Subnárodné indexy sa často vytvárajú a používajú na informovanie o politike a rozhodovaní, čo je potrebné zohľadniť pri rozhodovaní o najlepšej metóde imputácie. Niekedy to tiež môže znamenať, že ukazovateľ sa musí vylúčiť, ak existujú nejaké chýbajúce hodnoty, pretože akýkoľvek typ imputácie by nebol pre tvorcov politiky prijateľný. Z týchto dôvodov sú napríklad imputácie, ktoré sú založené na použití skutočných kódovaných hodnôt, ako sú hodnoty uvedené nižšie, vhodnejšie ako regresia:

Historické alebo novšie hodnoty

Na vyplnenie medzery môžu byť k dispozícii vzdialenejšie historické hodnoty. V prípade indexov, ktoré pokrývajú viac ako jeden rok, je tiež možné spätne použiť najnovšiu dostupnú hodnotu, ak nie je k dispozícii iná alternatíva.
To sa však musí robiť veľmi sporadicky, t. j. pre niekoľko ukazovateľov z rôznych komponentov.

Priemerovanie všetkých alebo susedných jednotiek

Použitie vhodného priemeru môže byť tiež vhodným spôsobom na vyplnenie medzier v údajoch. Napríklad je možné použiť priemer všetkých zaznamenaných hodnôt v ukazovateli alebo vybrať iba niekoľko regiónov / okresov a použiť ich priemer. Napríklad v prípade indických okresov bolo vhodné použiť skôr okresy v rovnakom štáte ako všetky indické okresy.

Vyššia úroveň geografie

Pre indexy produkované na podrobnejších úrovniach - ako sú obce, okresy, oddelenia, je tiež možné použiť hodnotu vyššej administratívnej jednotky - napríklad región, štvrť, štát, krajina. To by mohlo byť užitočné najmä pre mierne vyššie úrovne geografie, ako sú štáty, kde by sa iné metódy priemerovania nepovažovali za vhodnejšie.

Regresia

Chýbajúce hodnoty môžu byť vyplnené aj predpokladanými hodnotami získanými z regresie. Závislou premennou regresie je individuálny ukazovateľ s chýbajúcou hodnotou a regresor je individuálny ukazovateľ, ktorý vykazuje silný vzťah so závislou premennou, t.j. zvyčajne vysoký stupeň korelácie. Globálny index sociálneho pokroku používa na regresiu predpovedanie regresie hodnoty v spojení s niektorými z vyššie uvedených metód.

3) Zabezpečenie presnosti

Každý imputovaný údajový bod by sa mal vyhodnotiť, aby sa zabezpečila presnosť. V prípade, že imputovaná hodnota nespĺňa očakávania, je potrebné zvážiť a otestovať alternatívne metódy imputácie. Niektoré manipulácie sa ľahšie vykonávajú v Exceli, kde jednotlivé chýbajúce dátové bunky môžu byť prepísané vhodnou hodnotou, iné sú priamočiarejšie v štatistickom softvéri, ako napríklad STATA alebo R. V každom prípade musí byť metóda pripisovania chýbajúcich údajov jasná a uvedená v konečnom metodickom dokumente. Nikdy neprepisujte pôvodný dokument ani nezamieňajte imputované hodnoty so skutočnými kódovanými hodnotami.