A5: Projecte de visualització de dades (PAC2)

Daniel López Ramírez

M2.959 – Visualització de dades

Índex

1. Títol de la visualització

Cronologia dels acords de pau intraestatals segons els grups de persones, la governança i la seguretat.

2. Descripció del document

Aquest document descriu els resultats obtinguts després d'analitzar l'exportació de dades realitzada sobre les dades obertes de la base de dades Peace Agreements de la Universitat d'Edimburg, que conté informació sobre tots els acords de pau signats des del 1990 fins el 2019.

En aquesta exportació, s'ha seleccionat els acords de pau intraestatals (1273 acords) per a poder avaluar segons la cronologia en el temps les mencions a grups de persones, la governança i la seguretat. L'anàlisi realitzat pretén mostrar que en aquesta selecció d'acords de pau, hi ha grups de persones més mencionats que d'altres i que fins i tot poden estar més o menys relacionats com són els grups racials i els grups religiosos. D'altra banda, també es volen tenir en compte com afecten les mesures de governança i seguretat a aquests acords de pau intraestatals.

Per arribar a la visualització de les dades proposada, primer, farem una descripció de les dades extretes de la base de dades de Peace Agreements, revisant les variables, el seu tipus i els seus possibles valors. A continuació, explorarem les dades seleccionades, cercant possibles relacions entre variables, tendències,...Seguidament, parlarem dels mètodes i eines utilitzats per al tractament de les dades i la seva preparació. I finalment presentarem les visualitzacions realitzades amb el conjunt de dades.

3. Descripció de les dades

Tal com s'ha comentat anteriorment, he seleccionat un subconjunt de les dades de PAX, seleccionant a la seva web, només els registres que com a *"Agreement/conflict level"* continguin el valor *"Intrastate/intrastate conflict"*. Per tant, el conjunt de dades seleccionat (pax_corpus_1273_agreements_14-04-20.csv i pax_data_1273_agreements_14-04-20) conté 1273 registres. He descarregat tant les dades amb les metadades com el corpus, amb la intenció de fusionar aquests dos fitxers en un de sol que contingui tota la informació dels acords de pau. En el cas del corpus, cada registre conté 12 variables i el fitxer de dades, conté 265 variables. La fusió dels dos documents la podem realitzar utilitzant la variable *"AgreementId"* ("AgtId" en el document d'agreements), que és comú a tots dos documents i és un valor únic. Revisant les variables addicionals que conté el fitxer de corpus, veiem que estàn duplicades també a l'altre fitxer, i per tant podem eliminar-les després de la fusió dels dos fitxers. En aquest cas, hem eliminat les variables *Name, Region, Country, Peace.Process, Peace.Process.Name, Signed.Date, Agreement.Conflict.Level,Agtp,Agreement.Status,Conflict.Nature,Stage.y.*

Amb aquest preprocessament inicial, finalment tenim 1273 registres amb 265 variables, incloent també el text dels acords, és a dir, les dades del fitxer de corpus. D'aquestes variables, una petita part (17 variables) d'aquestes són les característiques generals dels acords, i la resta són metadades sobre els acords agrupades en diferents categories.

Les dades generals contenen majoritàriament dades categòriques, a part de la data de l'acord, que podríem dir que és una dada continua que ens permetrà generar la cronologia. En quant a les metadades de les categories contenen dades categòriques i dades binàries.

4. Exploració de les dades

Tal com he comentat a la descripció de les dades, he fusionat les dades dels acords amb el corpus per a tenir totes les dades disponibles.

Amb aquest conjunt de dades s'ha revisat el seu tipus, veient que la majoria de dades són de tipus enter o factor. Les dades de les diverses categories s'hauran de convertir en categoria per a poder tractar-les correctament, ja que la càrrega de les dades les ha tractat com enters, igual que els valors binaris de moltes de les variables que també estan tractades com enters.

Un cop revisades les variables, he avaluat si les dades contenien valors nuls. En aquest cas, s'ha trobat que 6 variables contenen valors nuls: *Loc1GWNO, Loc2GWNO, UcdpCon, UcdpAgr, PamAgr i CowWar*. Avaluant el seu contingut, s'ha observat que les dues primeres estan relacionades amb el país del conflicte però amb un codi diferent. Les tres següents estan relacionades amb un nou programa per a identificar els conflictes, que està en funcionament des del 2017 o bé pertanyen a un altre programa, i per tant, els conflictes anteriors no tenen aquesta dada i la última relaciona el conflicte amb una guerra, però pot ser que la guerra no compleixi amb el criteri necessari de morts, i per tant no estigui informat el camp. Per tant, s'ha decidit eliminar del conjunt de dades aquestes 6 variables.

Un cop extrets els valors nuls, he començat a revisar els valors de les diverses variables de les categories, per veure com es distribuïen entre els seus possibles valors. Mitjançant una primera revisió visual comptabilitzant els diferents casos de les metadades, he observat que a la mostra hi podia haver correlació entre les variables *"GRa"* i *"GRe"*, ja que els seus valors coincidien força. Per revisar aquesta correlació, s'ha creat una nova variable que pot tenir 3 valors: 0 si no coincideixen, 1 si una o altre té un valor superior a 0 i 2 si el valor de les dues coincideixen.

Amb la inspecció visual també he trobat diverses variables que poden tenir un pes important dintre de la mostra com: *"GRef", "GeWom", "StDef", "StGen", "Cons", "Ele", "PolPar", "SsrGua", "CE", "SsrPol", "SsrArm", "SsrDdr", "SsrPsf", "ImE".*

En aquesta primera revisió de les dades, i observant que la informació que ens poden proporcionar les metadades més genèriques ens permeten obtenir molta informació sobre els diversos tractats de pau, obviarem les dades proporcionades per les subcategories en la part d'agrupacions de persones i amb el repartiment de poder; per exemple de *"GRef"*, ens quedarem només amb *"GRef"* i descartarem *"GRefRhet", "GRefSubs" i "GRefOth"*.

I per finalitzar amb l'exploració de les dades, he realitzat una matriu de correlació entre les metadades, per veure quines variables poden tenir una relació entre elles, seleccionant les que tenen un valor més alt, per a poder revisar-les posteriorment.

5. Eines i procediments utilitzats sobre les dades

Per a poder analitzar i tractar les dades, finalment he utilitzat R, ja que amb els fitxers csv descarregats de la web, hi havia problemes a l'hora de tractar-los tant amb excel com amb libreoffice. Un cop carregats els dos fitxers descarregats a R, s'ha realitzat la fusió dels dos ftxers en un de sol, tal com s'ha comentat anteriorment, amb la funció "merge" de R. Al dataframe resultant, s'han eliminat les variables duplicades del procés de fusió.

Amb aquestes dades, hem revisat els tipus de dades de la mostra, veient que totes les variables o bé son factors o són enters; també s'han buscat si hi havia valors nuls, eliminant les variables que contenien aquests valors, ja que aquestes variables no les utilitzarem pel nostre estudi.

Amb aquesta mostra, hem realitzat una inspecció visual de les dades de la mostra, i mitjançant la funció *"count"* de la llibreria *"plyr"*, s'ha extret informació relativa dels valors de cada variable. D'aquesta manera, hem pogut veure les variables que podien tenir més inferència en la mostra de dades. Aquesta visualització també ha permès veure una possible relació entre dues variables de la mostra. Per veure aquesta possible

relació, s'ha creat una nova variable *"RelGRaGRe"* que depenent dels valors de *"GRa"* i *"GRe"* conté un 0 si no coincideixen, un 1 si una o altre té un valor superior a 0 i un 2 si el valor de les dues coincideixen.

Veient aquesta possible relació entre variables, he decidit fer una matriu de correlació entre totes les metadades de la mostra mitjançant dues funcions de correlació. La primera, *"cor"*, per extreure totes les dades en brut, i la segona *"correlate"* de la llibreria *"corrr"* per a poder filtrar les correlacions que tinguessin un valor més alt de 0.6.

Revisant les dades de les variables relacionades amb el país (*Con, Loc1ISO i Loc2ISO*), s'ha observat que, per una banda, *"Con"* conté també països relacionats amb els conflictes (com per exemple si el país pertanyia a l'antiga Iugoslàvia) i "*Loc1ISO" i "Loc2ISO"* contenen els codis ISO dels països, però no el nom. Per tant, utilitzant la llibreria *"countrycode"* de R, afegirem dues variables més amb el nom del país segons les variables abans esmentades.

Per finalitzar amb el tractament de les dades, hem convertit les metadades a factor, i hem generat el fitxer final amb totes les dades ("pax_alldata_1273.csv").

S'adjunta amb la pràctica el fitxer en R amb les passes realitzades per al tractament de les dades.

6. Visualització de les dades

Amb les dades tractades tal com s'ha indicat als punts anteriors, les he carregat a Tableau per a realitzar alguna visualització. He realitzat dues visualitzacions:

- Acords de pau per pais: és una visualització amb tres fulles amb les que es pot interactuar seleccionant països mitjançant accions.
    - Mapa de països dels acords de pau intraestatals: és un mapa per països on es conten el número d'acords de cada país.
    - Acords de pau intraestatals: diagrama de barres de cada país amb el número d'acords de cadascun.
    - Tree Map dels acords de pau intraestatals per pais i any: és un tree map on s'agrupen els tractats primer per pais i després per any. Ens permet veure com s'han anat desenvolupant els diversos tractacs a cada pais durant els anys.
    - URL: https://public.tableau.com/profile/daniel.lopez6830#!/vizhome/Acordsdepauintraestatalsperpais/AcordsdePauperPais
- Metadades dels acords: en aquesta visualització es vol comparar quines metadades són les més nombrades en els acords de pau. Igual que l'anterior, es pot interactuar amb la visualització.
    - Metadades segons país: mapa que conté com es distribueixen les dades per pais.
    - Ocurrències de les metadades: diagrama de barres amb el nombre d'ocurrències de les diverses metadades.
    - Heat Map per pais i any de les metadades: mapa de calor amb la distribució de les diveres metadades per país i any.
    - URL: https://public.tableau.com/profile/daniel.lopez6830#!/vizhome/PAC2_alldata/MetadadesdelsAcords

Aquestes visualitzacions donen una visió generalista de les dades, i ens permeten veure quines metadades ens poden donar més informació sobre els acords de pau.