

README

Szymon Talaga

10.12.2014

[English follows]

Repozytorium zawiera dane, skrypty i własne funkcje, które są niezbędne do zreplikowania wyników badań, które przeprowadziłem w ramach swojej pracy magisterskiej pisanej na Wydziale Psychologii Uniwersytetu Warszawskiego pod opieką profesor Marii Lewickiej. Dodatkowo, znajdują się w nim również raporty (w formatach .pdf/.html i .Rmd), które w sposób syntetyczny, ale opisowy i wyczerpujący przedstawiają wszystkie najważniejsze etapy skalowania i redukcji danych oraz właściwych analiz służących weryfikacji postawionych w pracy hipotez.

Wszystkie dane są udostępnione zarówno w formie surowej jak i w postaci przygotowanej do poszczególnych etapów analizy. Jedyną zmianą w danych surowych to ich pełna anonimizacja. Ponadto niektóre etapy transformacji i rekodowania danych nie są w pełni udokumentowane. Dotyczy to przede wszystkim klasyfikacji i grupowania surowych danych tekstowych dotyczących odwiedzanych przez respondentów miejsc. Wynika to z faktu, że ilość tych danych była bardzo duża (ponad 2 tys. wskazań na miejsca w danych surowych) i były one zupełnie nieustrukturyzowane przez co musiały być rekodowane w dużym stopniu ręcznie i przy użyciu setek wyrażeń regularnych, co oczywiście sprawiało, że pełne udokumentowanie wszystkich transformacji było zadaniem ponad siły jednej osoby. Więcej informacji na temat rekodowania surowych danych ze wskazaniem na miejsca w opisie danych dotyczących miejsc.

Mapa repozytorium

- **Katalog główny (MA)** - zawiera plik z kwestionariuszem użytym do zebrania danych (Kwestionariusz.pdf) oraz plik projektu Rstudio o nazwie MA.Rproj. Ponadto zawiera w sobie katalogi z poszczególnymi elementami:
 - **rawData** - tutaj znaleźć można wszystkie pliki z danymi surowymi oraz pliki pomocnicze użyte przy rekodowaniu i klasyfikowaniu odpowiedzi z pytań otwartych. Znajduje się tutaj również skrypt ("RelabellingAndBasiDataTrans.R"), który został użyty do przekształcenia surowego zbioru "raw244.csv" w zbiór "DatInd231.csv", który znaleźć można w katalogu "MainData". Skrypt zawiera również opis kolejnych czynności i transformacji.
 - **MainData** - tutaj znajdują się wszystkie zbiory danych po kolejnych etapach rekodowania i transformacji. Wszystkie z nich zapisane są jako obiekty R (format .RData), ponieważ ten format zapewnia utrzymanie poprawnej kolejności poziomów zmiennych jakościowych. Niemniej jednak najważniejsze zbiory danych zapisane są również w formacie .csv w celu zapewnienia możliwości replikacji wyników również przy użyciu innych niż język R narzędzi. Dokładny opis kolejności powstawania zbiorów danych i relacji między nimi w sekcji "Generowanie Danych".
 - **HelperFunctionsMisc** - ten folder zawiera skrypty z różnymi dodatkowymi funkcjami, które napisałem, żeby ułatwić sobie pracę oraz zwiększyć czytelność kodu.
 - **Imputation** - tutaj znaleźć można skrypty służące podstawaniu brakujących odpowiedzi. Znajduje się tu również skrypt z funkcjami pomocniczymi, których używałem przy podstawianiu a także obiekt .RData zawierający całą przestrzeń zmiennych związanych z finałowym etapem podstawiania brakujących danych przy pomocy algorytmu MICE. Służy to temu, żeby nie trzeba było za każdym razem od nowa dokonywać wszystkich obliczeń, ponieważ algorytm MICE jest dosyć kosztowny obliczeniowo, przez co cały proces może zająć nawet do kilkunastu minut.
 - **Scaling** - tutaj znajdują się skrypty i funkcje służące opracowaniu skal aktywności obywatelskiej, kapitału społecznego oraz przywiązania do miejsca.

- **DataReduction** - ten folder zawiera skrypty i funkcje, których przeprowadzają redukcję danych w oparciu o Analizę Głównych Składowych, Wielokrotną Analizę Korespondencji oraz Analizę Skupień (k-średnich).
- **Raporty** - ten folder zawiera raporty (napisane po polsku) dokumentujące w przystępny ale wyczerpujący sposób najważniejsze etapy przeprowadzonych analiz.

Generowanie Danych

UWAGA: poprawne wykonanie wszelkich skryptów zawartych w tym repozytorium jest zależne od struktury środowiska całego projektu. Innymi słowy struktura katalogów oraz lokalizacja poszczególnych plików musi pozostać niezmieniona.

Dane na temat respondentów

Aby wygenerować ostateczny zbiór danych z poziomu danych surowych należy wykonać kolejno następujące kroki:

1. Wykonaj skrypt "rawData/RelabellingAndBasiDataTrans.R" : raw244.csv -> DatInd231.RData
2. Wykonaj skrypt "Imputation/BasiImpute.R" : DatInd231.RData -> DatIndBIM.RData
3. Wykonaj skrypt "Imputation/LimitData.R" : DatIndBIM.RData -> DatIndBIM2.RData
4. Wykonaj skrypt "Imputation/ImputeGenMICE.R" : DatIndBIM2.RData -> MainData1.RData
5. Wykonaj skrypt "Scaling/Scaling-CivicActivity.R" : MainData1.RData -> MainData2.RData
6. Wykonaj skrypt "Scaling/ResourceGenerator.R" : MainData2.RData -> MainData3.RData
7. Wykonaj skrypt "Scaling/Scaling-Attachment.R" : MainData3.RData -> MainData4.RData
8. Wykonaj skrypt "DataReduction/DatRed-CultPred.R" : MainData4.RData -> MainData5.RData : MainData4.RData -> MainData6.RData
9. Wykonaj skrypt "DataReduction/PrepClustMCA.R" : MainData6.RData -> MainData7.RData
10. Wykonaj skrypt "DataReduction/MCA.R" : MainData7.RData -> MainData8.RData
11. Wykonaj skrypt "DataReduction/Kmeans.R" : MainData8.RData -> MainData9.RData
12. Wykonaj skrypt "Heterogeneity/HeterogeneityCoefs.R" : MainData9.Rdata, Places.RData, Places-Data.RData -> MainData10.RData
13. Wykonaj skrypt "Scaling/ResGen-Revisited.R" : MainData10.RData -> MainData11.RData

Wskazania na miejsca

Tak jak zostało wspomniane, dane ze wskazaniami na miejsca zostały opracowane w dużym stopniu ręcznie i w oparciu o pracę z wyrażeniami regularnymi w konsoli R. Ze względu na rozmiar tego przedsięwzięcia jego pełne udokumentowanie nie było możliwe. W katalogu "Places" znajdują się skrypty, które dokumentują część zabiegów i w oparciu o nie można sobie wyrobić ogólne pojęcie na temat tego w jaki sposób dane na temat miejsc zostały wyprodukowane. Obiekt "Places.RData" zawiera finalną wersję zrekodowanego zbioru danych ze wskazaniami na miejsca. Wiersze ramki danych to respondenci a kolumny to dziwności wskazań na miejsca + dwudziesta kolumna z identyfikatorami respondentów. Na podstawie tego zbioru danych (oraz zbioru respondentów) tworzony jest zbiór danych dla miejsc, który zawiera profile miejsc będące uśrednionymi profilami osób, które je odwiedzają. Dodatkowo posłużył on do wygenerowania wskaźników różnorodności przestrzeni społecznej, w której zanurzeni są badani. Wskaźniki te są oparte na entropii rozkładów osób odwiedzających dane miejsce, gdzie respondenci są poklasyfikowani zgodnie z wynikami analizy skupień.

Profile miejsc

1. Wykonaj skrypt "Places/PlaceProfiles.R" : MainData9.RData, IncidenceMatrix.RData, Places.RData -> PlaceData.RData

Jedno- i dwudzielne macierze sąsiedztwa

1. Wykonaj skrypt “Networks/BipartiteNetwork.R” : Places.RData i MainData9.RData \rightarrow IncidenceMatrix