FACTOQGIS: GUI ALAT ZASNOVAN NA R SKRIPTI ZA IZVOĐENJE GEOMETRIJSKE ANALIZE PODATAKA U SLOBODNOM I OTVORENOM GIS-u

Florent Demoraes, Univ Rennes, CNRS, ESO - UMR 6590, F-35000 Rennes, France
florent.demoraes@univ-rennes2.fr

Marc Souris, UMR Unité des Virus Emergents (UVE : Aix-Marseille Univ – IRD 190 – Inserm 1207 – IHU Méditerranée Infection), Marseille, France
marc.souris@ird.fr

FactoQGIS je algoritam koji omogućava implementaciju geometrijske analize višedimenzionalnih podataka u QGIS. Konkretnije, ovaj alat je dizajniran tako da lako izvrši tipološku analizu kvantitativnih podataka agregiranih u prostornim jedinicama. Ova metoda se široko koristi u geografiji, ali se do sada izvršavala izvan GIS okruženja, u specifičnom statističkom softveru. FactoQGIS je alat koji precizno popunjava ovu prazninu među GIS funkcionalnostima. Prvo izvodi PCA (Analiza glavnih komponenti) i drugo HAC (Hijerarhijska rastuća klasifikacija) na prve faktore. FactoQGIS se bazira na R skripti koja uglavnom koristi FactoMineR paket koji je razvio François Husson et al. (Agrocampus Ouest, Rennes, Francuska). Rezultati (tabele i grafikoni) se izvoze u Excel i png format, a zatim se ubacuju u html datoteku koja se automatski pojavljuje u web pregledniku na kraju procesa. Algoritam također kreira novi sloj sa kolonom koja označava klaster u kojem svaka prostorna jedinica pripada, tako da olakšava mapiranje tipologije. FactoQGIS je dostupan iz grafičkog korisničkog interfejsa direktno u QGIS okruženju. To će biti od posebnog interesa za geografe i sve korisnike koji žele da jednostavno izgrađuju i mapiraju višedimenzionalnu tipologiju bez poznavanja R jezika. Da bismo ilustrirali kako funkcionira FactoQGIS, napravili smo kao primjer, tipološku analizu socio-demografskih podataka koji su agregirani u "arrondissements" i "communes" u Parizu.

Ključne riječi: *Geometrijska analiza podataka;Tipološka analiza;Zbirni podaci u prostornim jedinicama; R skripta; Slobodni i otvoreni GIS*

FACTOQGIS: A GUI TOOL BASED ON AN R SCRIPT TO PERFORM GEOMETRIC DATA ANALYSIS IN A FREE AND OPEN SOURCE GIS

Florent Demoraes, Univ Rennes, CNRS, ESO - UMR 6590, F-35000 Rennes, France
florent.demoraes@univ-rennes2.fr

Marc Souris, UMR Unité des Virus Emergents (UVE : Aix-Marseille Univ – IRD 190 – Inserm 1207 – IHU Méditerranée Infection), Marseille, France
marc.souris@ird.fr

FactoQGIS is an algorithm that allows the implementation of a geometric analysis of multidimensional data in QGIS. More specifically, this tool was designed to easily perform a typological analysis on quantitative data aggregated in spatial units. This method is broadly used in geography but it was up to now executed out of GIS environments, in

specific statistical software. FactoQGIS is a tool which precisely fills this gap among GIS functionalities. It first performs a PCA (Principal Component Analysis) and second a HAC (Hierarchical Ascending Classification) on the first factors. FactoQGIS is based on an R script that mainly uses the FactoMineR package developed by François Husson et al. (Agrocampus Ouest, Rennes, France). The results (tables and plots) are exported respectively in Excel and png format and then inserted into an html file that automatically pops up in a web browser at the end of the process. The algorithm also creates a new layer with a column indicating the cluster each spatial unit belongs to, so as to make it easy to map the typology. FactoQGIS is accessible from a graphical user interface directly in the QGIS environment. It will be of particular interest to geographers and to any users who wish to simply build and map a multidimensional typology without knowing the R language. To illustrate how FactoQGIS works, we performed as an example, a typological analysis on socio-demographic data that are aggregated by "arrondissements" and "communes" in Paris.

Keywords: Geometric Data Analysis; Typological Analysis; Aggregated data in spatial units; R script; Free and Open Source GIS

UVOD
INTRODUCTION

Danas postoje sve bogatiji i raznovrsniji izvori podataka, od kojih se veliki dio agregira u geografske divizije (administrativne jedinice, popisni traktati, slivovi itd.). Ova masa informacija zahtijeva posebne analitičke metode za generiranje znanja za podršku javnim politikama, definiranje orijentacija upravljanja okolišem, usmjeravanje razvoja projekata i usmjeravanje strategija za geomarketing kompanija. Potreba za analitičkim metodama stoga se tiče širokog spektra javnih aktera (lokalne vlasti, državne uprave), polu-javnih organizacija (poljoprivredni uredi, agencije za urbanističko planiranje) i privatnih aktera (uredi za projektiranje, vizualizacije podataka i konsultantske kompanije za komunikacije koje rade za biračke institucije, izdavačke kuće za novine i časopise). MeĐu analitičkim metodama, definicija sintetičkih profila ili glavnih tipova meĐu prostornim jedinicama, često nastoji istaknuti, u odreĐenom vremenu, dnevne obrasce mobilnosti (Demoraes et al., 2013), ruralne strukture (Walsh, 2000), sastav urbanih područja (Metzger, 2001), izborno ponašanje (Rivière, 2012), ili razlikovanje socijalno-ekoloških jedinica (Hanspach et al., 2016), itd. TakoĐer je korisno pratiti putanju prostornih jedinica tokom vremena, kako bi se pratile urbane transformacije (Piron, et al., 2004), ili dinamika tokova domaćinstva (Robson, et. Al., 2009). U ovoj perspektivi, tipološka analiza, koja je dio područja analize geometrijskih podataka (Benzécri J.-P., 1973; Le Roux B. i Rouanet H., 2004; Greenacre M. i Blasius J., 2006), je od primarnog interesa. Prvo dozvoljava da se trendovi identifikuju u skupovima podataka. Drugo, omogućava stvaranje klastera koji povezuju slične prostorne jedinice. Preciznije, tipološka analiza kombinira analizu faktora (kao što je analiza glavnih komponenti, analiza višestruke korespondencije, analiza faktora miješanih podataka) i hijerarhijska rastuća klasifikacija (koja se naziva i hijerarhijska aglomerativna klasterizacija) zasnovana na prvim faktorskim dimenzijama (vidi Lebart L. et al., 2006; Husson et al., 2009).

Još jedna primjedba se odnosi na softversku ponudu, koja se može shematski podijeliti u dvije glavne grupe:

- softverska rješenja posvećena analizi statističkih podataka (R, SPAD, SPSS, Stata, itd.) koji omogućavaju primjenu tehnika analize geometrijskih podataka,
- GIS softver (QGIS, OpenJump, gvSIG, SavGIS, ArcGIS, Mapinfo, itd.) koji omogućava obradu podataka sa prostornom komponentom, kao što su agregirani podaci u prostornim jedinicama, i za izradu kartografskih prikaza, uključujući tipološke karte.

Ova dva glavna alata su generalno veoma interoperabilna i zahtevaju konverziju formata i čitav niz uvoza i izvoza kako bi prešli iz jednog u drugi. MeĎutim, ova dihotomija mora biti kvalifikovana. Softverske kompanije ili razvojne zajednice razvile su relativno integrisana softverska rješenja. Što se tiče softvera za analizu statističkih podataka, R softver nudi mogućnost implementacije toka integracije upravljanja prostornim podacima (rgdal / sp / rgeos i sf paketi), tipološke analize (FactoMineR, ade4 i paketi klastera) i alata za mapiranje (ggplot2, rCarto, maptools, ggmap i kartografski paketi). MeĎutim, ovaj radni proces zahtijeva dobro poznavanje R jezika.

Što se tiče GIS softvera, meĎu vlasničkim rješenjima možemo spomenuti implementaciju PCA i multivarijatne klasterizacije u ArcGIS-u. MeĎutim, upotreba ove dvije metode ograničena je na analizu višestrukih satelitskih snimaka. U SavGIS-u, PCA se može primijeniti na vektorske i rasterske slojeve, ali HAC nije dostupan. U QGIS-u postoje ekvivalentni alati koji su takoĎer posvećeni analizi slike kroz GRASS funkcionalnosti. QGIS takoĎe omogućava izvršavanje R skripti, od kojih su neke posvećene faktorskoj analizi (PCA, MCA, FAMD, zasnovane na paketima ade4 i FactoMineR) i klasteriranju (HAC sa paketom klastera). Ove skripte, kao i sve druge funkcije koje su dostupne u QGIS toolboxu, mogu biti dodane tijeku rada koji se može dizajnirati u grafičkom modeleru. MeĎutim, analiza je ekstremno razbijena: potrebno je izvršiti prvi skript kako bi se dobio krug korelacije, druga skripta da bi se dobila faktorska karta sa varijablama, treća skripta da se dobije faktorska karta sa pojedincima i četvrta, da bi dobili doprinos. Pored toga, sa ovim skriptama, PCA se može primijeniti samo na 4 varijable, što uvelike ograničava njihov interes. Što se tiče HAC skripte, ona se mora primijeniti na 5 varijabli i stoga se ne može koristiti za kreiranje tipologije na prva dva ili tri faktora dobivena iz prethodne PCA.

Ova zapažanja su nas navela da razvijemo alat zvan FactoQGIS koji zadovoljava sljedeće kriterije:

- Alat koji se lako može izvršiti iz grafičkog interfejsa, bez skriptovanja,
- Alat integriran u besplatni i otvoreni GIS softver koji se široko koristi,
- Alat koji proizvodi izlaze koji se lako interpretiraju i koji se mogu direktno koristiti u QGIS-u,
- Alat sa detaljnom kontekstualnom pomoći i podrazumevanim postavkama, najčešće korišćenim

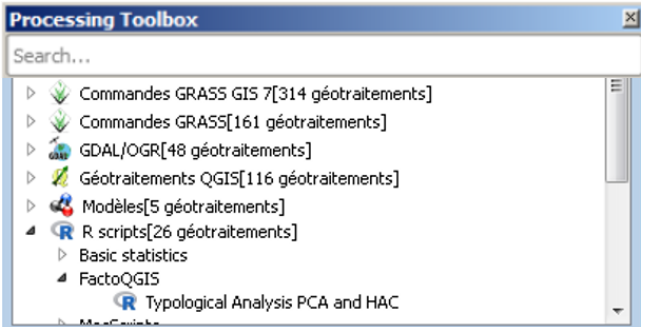
FactoQGIS će stoga biti od posebnog interesa za akademske svrhe, posebno za geografe, urbane planere i studente GIS analitičare.

RADNI PRINCIPI FACTOQGIS
OPERATING PRINCIPLES OF FACTOQGIS

Licenca, jezici, verzije softvera i sadržaj skripte
License, languages, software versions and script content

FactoQGIS je razvijen pod GNU General Public licencom v2.0 i radi sa QGIS 2.18 i R 3.5.1 verzijom ili novijom. Razvoj je u toku da bi mogli nastaviti koristiti R skripte u novijim izdanjima QGIS-a (izdanje 3.0 i novije) zahvaljujući dodatku Processing R Provider. Da biste pokrenuli R skripte u QGIS-u, R mora naravno biti instaliran na računalu. FactoQGIS se uglavnom zasniva na FactoMineR paketu koji je razvio François Husson et al. (2009). On takoĎer koristi sekundarnu upotrebu factoextra, stringr, openxlsx, R2HTML i corrplot paketa. Ovi paketi moraju biti prethodno instalirani u R softveru (ili preko R Studio). Za izvršavanje R skripti, QGIS koristi modul Obrada (Graser & Olaya, 2015), koji se sam temelji na Python modu potprocesa. FactoQGIS alat se sastoji od dva fajla koja su dostupna na GitHubu. Prva datoteka "Typological_Analysis_PCA_PCA_and_HAC.rsx" sadrži skriptu. Druga datoteka "Typological_Analysis_PCA_PCA_and_HAC.rsx.help" sadrži pomoć. Ove datoteke moraju biti pohranjene u folderu:
C:\Users\...\...\qgis2\processing\rscripts.

Zaglavlje skripte sadrži python parametre povezane s argumentima koje korisnik popuni u dijaloškom okviru. Ispod zaglavlja počinje R skripta, koja je sama po sebi podijeljena na 7 dijelova kao što je prikazano ispod.

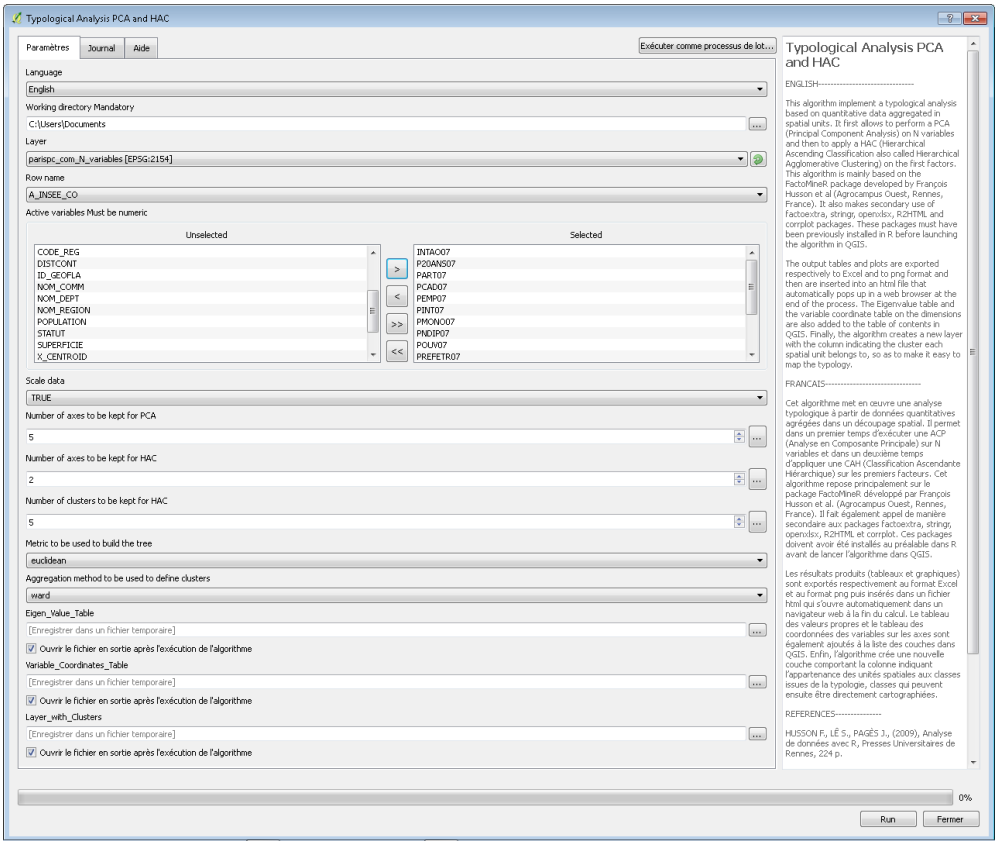


Sl. 1. FactoQGIS u QGIS alatu
Fig. 1. FactoQGIS in the QGIS toolbox

1 - Učitava pakete potrebne za izvršavanje skripte.
2 - Vraća u R objekte parametre koje je korisnik unio u dijalog box i pretvara ih u vrijednosti argumenta za R funkcije.
3 - Uvozi skup podataka (tabelu atributa sloja) i kreira skup podataka koji odgovara samo aktivnim kvantitativnim varijablama.
4 - Pokreće PCA, izračunava rezultate (tabele, parcele) u različitim formatima.
5 - Pokreće HAC, izračunava rezultate (tabele, parcele) u različitim formatima.
6 - Dodaje rezultate u html datoteku koja se automatski pojavljuje na kraju procesa.

7 - Kreira sloj koji u svojim atributima sadrži kolonu sa klasterima kojoj pripadaju prostorne jedinice, koje proizlaze iz tipologije.

U dijaloškom okviru (slika 2), korisnik mora unijeti 14 parametara. Prvih 11 su ulazni parametri i obavezni su. Neki imaju default vrijednosti. Posljednja tri su izlazni parametri i nisu obavezni. Ako korisnik ne odredi bilo koje ime za izlazne datoteke, one će biti spremljene u privremenu mapu. Parametri su detaljno opisani u prilozima na kraju članka.



Sl. 2. Okvir za dijalog FactoQGIS
Fig. 2. The FactoQGIS tool dialog box

Izlazni formati i izlazne datoteke
Outputs format and output files

Tablica 1 prikazuje za svaki od rezultata koje je proizveo FactoQGIS alat njegov format, da li je datoteka kreirana u radnom prostoru, da li je dodana u html datoteku i da li je dodana u tablicu sadržaja u QGIS-u. Većina rezultata (tabela i grafika) se ubacuje u html datoteku koja se automatski pojavljuje u web pregledniku na kraju procesa.

Tabela 1. Sažetak rezultata koje je kreirao FactoQGIS
Table 1. Summary of the outputs created by FactoQGIS

Izlaz	Format	Izlazne datoteke pohranjene u radnom direktoriju	Dodato u html datoteku	Dodato sadržaju u QGIS-u
Tabela Eigen vrijednosti	xlsx, csv	x	x	x
Scree plot (dobitak inercije)	png	x	x	
Prva faktorska mapa koja pokazuje varijable (osi 1 i 2)	png, pdf	x	x	
Tabela varijabilnih koordinata	xlsx, csv	x		x
Kvalitet prikaza varijabli (Cos2)	png	x	x	
Prva faktorska mapa koja prikazuje koordinate pojedinaca (dimenzije 1 i 2)	png, pdf	x	x	
Hijerarhijsko stablo klastera	png	x	x	
Hijerarhijsko stablo klastera na prvoj faktorskoj mapi	png	x	x	
Bar-ploče koje prikazuju varijable koje najbolje opisuju klastere *	png	x	x	
Tabele koje daju opis klastera po varijablama			x	
Sloj s atributom koji označava klaster kojem pripada svaka prostorna jedinica	shp	x (samo ako je korisnik dao ime)		x

** Samo varijable sa v-testom> | 1.96 | su iscrtane.*

PRIMJENA FACTOQGISA ZA ANALIZU SOCIO-DEMOGRAFSKOG PROSTORNOG UZORAKA PARIZA
APPLICATION OF FACTOQGIS TO ANALYZE THE SOCIO-DEMOGRAPHIC SPATIAL PATTERN OF PARIS

Da bismo ilustrirali kako funkcionira FactoQGIS, napravili smo primjer, tipološku analizu podataka koji se dostavljaju uz knjigu koju su napisali Commenges et al. (2014) i koji odgovaraju uzorku podataka iz popisa iz Francuske za 2007. godinu. Ovi podaci su agregirani u "arrondissements" i "communes" u Parizu i predstavljaju njegov prvi vanjski prsten. Skup podataka uključuje 143 prostorne jedinice raštrkane na 4 administrativna "odjela". S obzirom na niz od 14 socio-demografskih pokazatelja (Tabela 2), autori su nastojali razumjeti kako su organizirani Pariz i njegov prvi vanjski prsten te ukazati na sličnosti i različitosti prostornih jedinica.

U ovom članku nije cilj detaljno prikazati analizu pariškog socio-demografskog obrasca koji je već uradio Commenges et al., već samo da bi ilustrirali kako FactoQGIS funkcionira i kako je primijenjen na te podatke. U tom smislu, sljedeći snimci ekrana

pokazuju rezultate koji se pojavljuju u izlaznom html fajlu i svaki rezultat je povezan sa kratkim komentarom.

Tabela 2. Spisak 14 socio-demografskih indikatora koji se koriste u tipološkoj analizi
Table 2. List of the 14 socio-demographic indicators used in the typological analysis

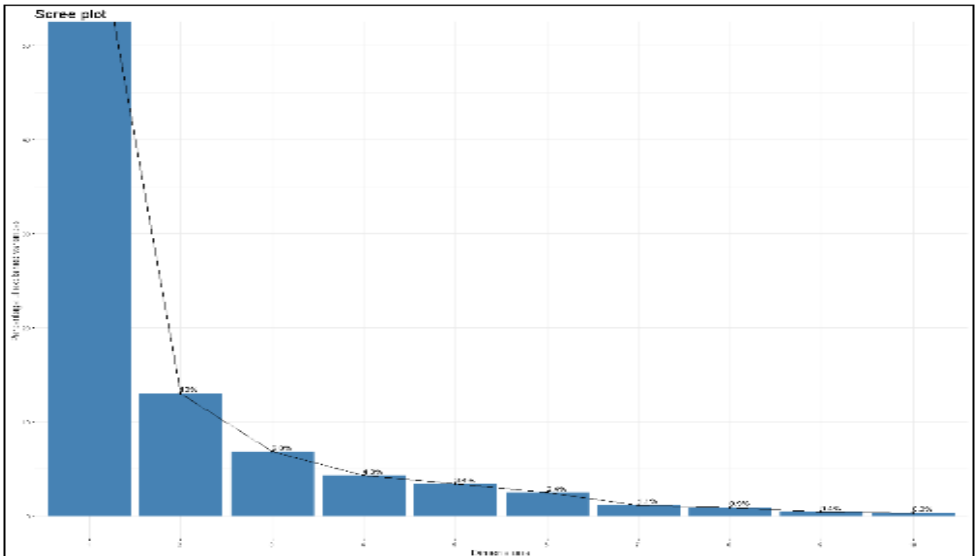
Oznaka	Opis
INTAO07	Udio privremenih radnika (zaposlena radna snaga)
P20ANS07	Udio osoba mlađih od 20 godina (ukupna populacija)
PART07	Udio obrtnika (zaposlena radna snaga)
PCAD07	Udeo rukovodilaca (zaposlena radna snaga)
PEMP07	Udio zaposlenih (zaposlena radna snaga)
PINT07	Udeo posredničkih zanimanja (zaposlena radna snaga)
PMONO07	Udio jednoroditeljskih obitelji (obitelji)
PNDIP07	Procenat onih koji nisu diplomirali (populacija mlađa od 15 godina)
POUV07	Udio radnika (zaposlena radna snaga)
PREFETR07	Udeo domaćinstava sa glavom koji je strano lice (domaćinstva)
PRET07	Udio umirovljenika (zaposlena radna snaga)
REFEROUI07	Procenat glasova za DA na evropskom referendumu (2006)
RFUCQ07	Srednji prihod (tekući euro)
TXCHOM07	Udio nezaposlenih (radna snaga)

Izvor: INSEE (2007), presented in Commenges et al. (2014)

PCA dozvoljava sažimanje opažanja koja su definirana u prostoru od p varijabli te do prostora p glavnih komponenti. Prednost ove metode je smanjenje dimenzija i eliminisanje kolinearnosti između varijabli. Komponente (koje se nazivaju i faktori, dimenzije ili osi) odgovaraju linearnim kombinacijama svih analiziranih indikatora. Glavne komponente su sintetičke varijable koje identificiraju glavne faktore diferencijacije unutar početne tablice. Nekoliko metrika se može koristiti za karakterizaciju ovih komponenti, počevši od svojstvenih vrijednosti. Potonji odražavaju inerciju oblaka tačaka koju objašnjava svaki faktor. Zbir ovih sopstvenih vrijednosti daje ukupnu varijansu (koja se naziva i inercija). Da bi dobili ideju o tome kako su strukturirane varijable, potrebno je ispitati relativni dio varijanse za svaku komponentu, kao i njihov kumulativni dio. U našem primjeru, ove indikacije su sadržane u tabeli 3 i također su prikazane u grafičkom obliku na scree parceli (Slika 3).

Tabela 3. Eigen vrijednosti za svaku dimenziju
Table 3. Eigen values for each dimension

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	9.4e+00	6.7e+01	6.7e+01
Dim.2	1.8e+00	1.3e+01	8.0e+01
Dim.3	9.5e-01	6.8e+00	8.7e+01
Dim.4	6.0e-01	4.3e+00	9.1e+01
Dim.5	4.7e-01	3.4e+00	9.4e+01
Dim.6	3.4e-01	2.4e+00	9.7e+01
Dim.7	1.6e-01	1.1e+00	9.8e+01
Dim.8	1.2e-01	8.8e-01	9.9e+01
Dim.9	5.5e-02	3.9e-01	9.9e+01
Dim.10	4.0e-02	2.8e-01	1.0e+02
Dim.11	2.9e-02	2.1e-01	1.0e+02
Dim.12	2.2e-02	1.6e-01	1.0e+02
Dim.13	1.4e-02	1.0e-01	1.0e+02
Dim.14	1.3e-03	9.6e-03	1.0e+02



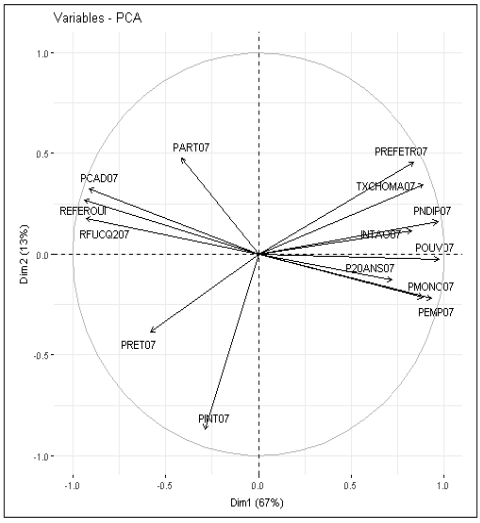
Sl. 3. Scree plot (dobitak inercije)
Fig. 3. Scree plot (gain of inertia)

Čini se da je 67% informacija sadržanih u 14 početnih atributa sažeto iz prvog faktora (vidi procenat varijance u tabeli 3). Diskriminacijska snaga sljedećih osi je relativno niska (Slika 3). Dvije prve dimenzije sažimaju 80% ukupne inercije i prve pete dimenzije gotovo 95% (vidi kumulativni procenat varijance u tabeli 3). Ove informacije potvrĐuju da su zadane vrijednosti za 7. i 8. ulazne parametre u dijaloškom okviru (Slika 2) dobre i da ih u ovom slučaju ne treba mijenjati.

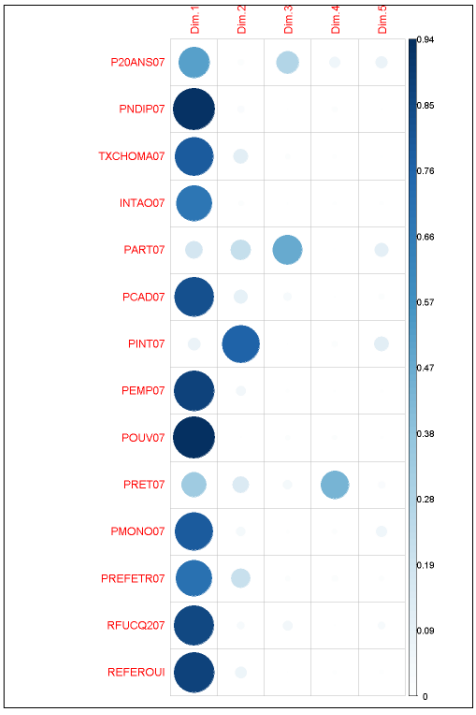
Prva faktorska karta definirana je u prve dvije dimenzije (slika 4). Prva osa (dim. 1) jasno razlikuje na lijevom dijelu parcele prostorne jedinice s visokim udjelom rukovoditelja (PCAD07) i visokim prihodima (RFUCQ07), a na desnom dijelu parcele prostorne jedinice karakterizirane varijablama koje ukazuju na veću društveno-ekonomsku štetu (kao što je visoka stopa nezaposlenosti, TXCHOM07).

Kvalitet predstavljanja varijable na dimenzijama je još jedna važna metrika za karakterizaciju komponenti. Kvalitet se mjeri kvadratnim kosinusom (cos^2) kuta izmeĐu varijable i njegove projekcije na osi. Ovaj kvadratni kosinus se izračunava za svaku dimenziju. Slika 5 ilustruje ovaj kvalitet. Što je krug veći i tamniji, to je kvalitet bolji. Strogo govoreći, treba tumačiti samo dobro projektovane elemente.

Na primjer, varijabla POUV07 ima visoku kvalitetu na prvoj dimenziji. To znači da je cos^2 veoma blizu 1, a ugao je vrlo blizu 0 (POUV07 vektor je gotovo usklaĐen sa dimenzijom 1 na slici 4). Sa svoje strane, varijabla PINT07 ima nisku kvalitetu na prvoj dimenziji. Ova varijabla je mnogo manje projektovana na dimenziju 1 (vidi njen visoki ugao sa dimenzijom 1 na slici 4).



Sl. 4. Prva faktorska karta koja pokazuje varijable (dimenzije 1 i 2)
Fig. 4. First factorial map showing the variables (dimensions 1 and 2)

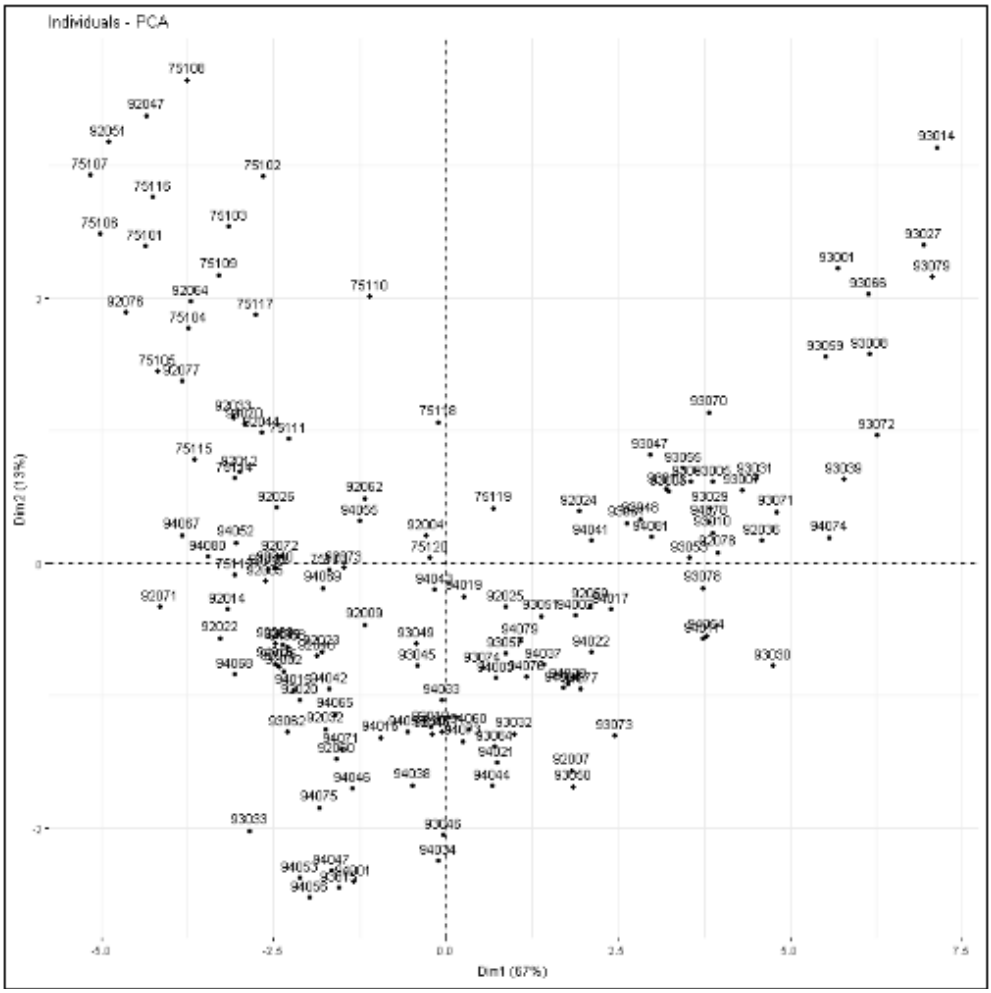


Sl. 5. Kvalitet prikaza varijabli na dimenzijama (Cos^2)
Fig. 5. Quality of the representation of the variables on the dimensions (Cos^2)

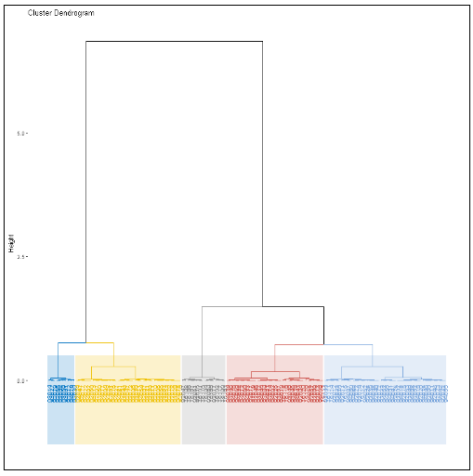
Slika 6 je prva faktorska mapa koja prikazuje koordinate pojedinaca i može se preklapati s prvom faktorijalnom mapom koja pokazuje varijable (slika 4). Izgleda da se pojavljuju profili: na prvoj osi možemo razlikovati na lijevoj strani prostorne jedinice s ID-om počevši od 75 (unutarnji Pariz) koje imaju visok udio rukovoditelja i visokih prihoda (vidi sliku 4) i na desnoj, prostornoj jedinice s ID-om počevši od 93 (Seine Saint-Denis), karakterizirane varijablama koje ukazuju na veću društveno-ekonomsku štetu (vidi sliku 4).

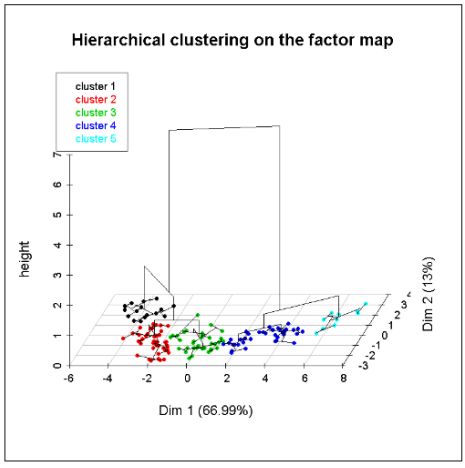
Hijerarhijska stabla klastera omogućavaju dobijanje mnogo preciznijih profila. Klastering algoritmi imaju za cilj definiranje grupa pojedinaca (prostornih jedinica, u našem slučaju) koje su homogene u smislu njihovih statističkih atributa. Grupe su homogene ako su statistički pojedinci u svakoj grupi što je moguće sličniji unutar svake grupe. Algoritam u primjeru vraća 5 optimalnih klastera (Slika 7).

Sljedeći grafički prikazi (slika 8) pokazuju za svaki klaster vrijednosti koje su uzete od strane varijabli u odnosu na ukupnu srednju vrijednost (tablica 4). Ove vrijednosti su korisne za kvalifikaciju svakog klastera. Samo varijable povezane s v-testom> | 1.96 | su značajne i stoga su iscrtane.

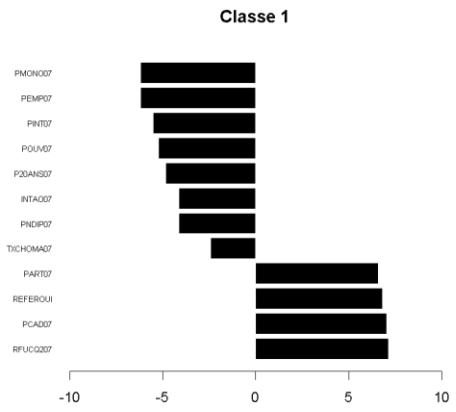


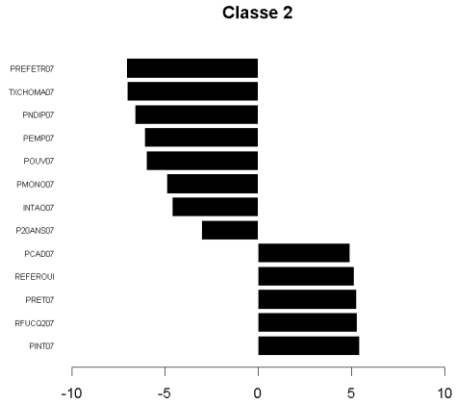
Sl. 6. Prva faktorska mapa koja prikazuje koordinate pojedinaca (dimenzije 1 i 2)
Fig. 6. First factorial map showing the coordinates of the individuals (dimensions 1 and 2)

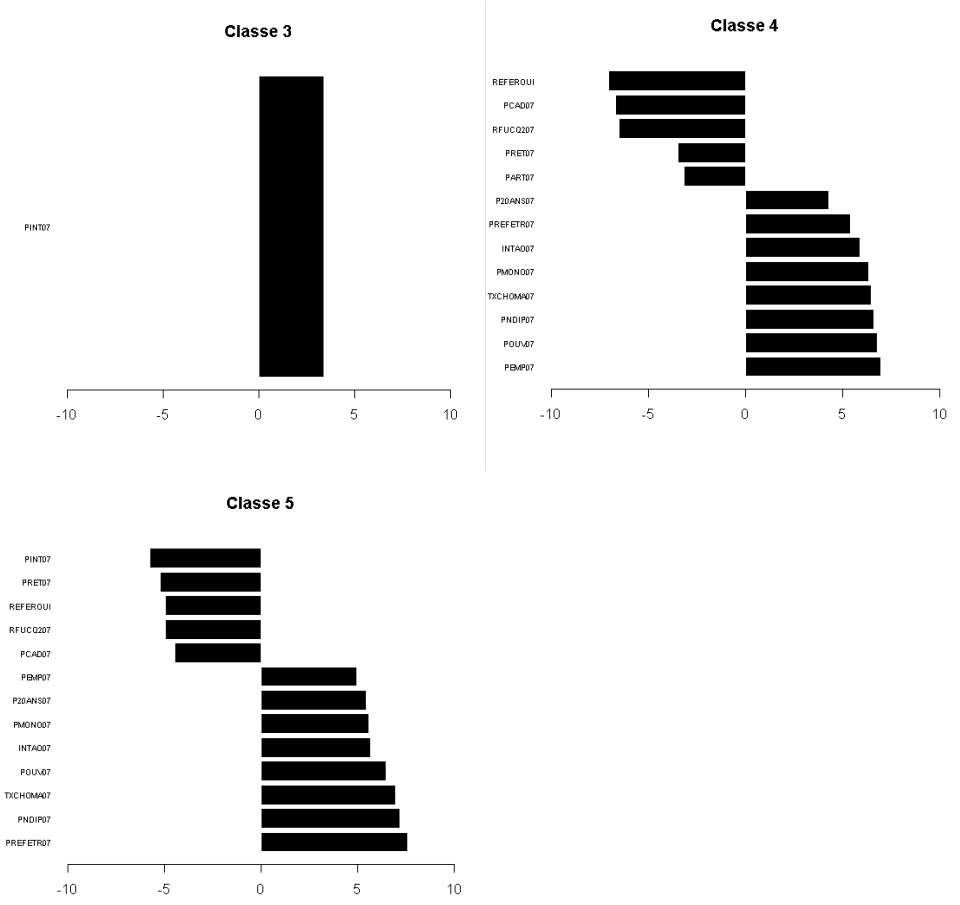




Sl. 7. Stabla hijerarhijskog klastera
Fig. 7. Hierarchical cluster trees







Sl. 8. Grafikoni koji prikazuju varijable koje najbolje opisuju klastere (grafički su prikazane samo varijable s v-testom> | 1.96 |)
Fig. 8. Bar plots showing the variables which best describe the clusters (only the variables with a v-test >|1.96| are plotted)

- Klaster 1 odgovara bogatim opštinama (Srednji prihod - RFUCQ visok; Udio rukovodilaca - PCAD visok; Procenat glasova za DA na evropskom referendumu 2006 - REFERUI visok; Procenat zanatlija - Visok udio).
- Klaster 2 odgovara opštinama gornje srednje klase sa visokim procentom srednjih zanimanja (PINT07), prilično visokim srednjim prihodima (RFUCQ07), pretjeranom zastupljenošću penzionera (PRET07), procentom glasova za DA na evropskom referendumu 2006. - REFEROUI prilično visoka. S druge strane, udio domaćinstava sa glavom koja je strano lice (PREFETR), udio onih koji ne diplomiraju (PNDIP) i udio nezaposlenih (TXCHOM) je nizak.

- Klaster 3, u središtu grafikona (slika7), ima samo preveliku zastupljenost srednjih zanimanja kao posebnu karakteristiku.
- Klaster 4 odgovara općinama niže srednje klase sa visokim udjelom zaposlenih (PEMP07) i radnika (POUV07), kao i visokim udjelom nekvalificiranih (PNDIP07) i nezaposlenih osoba (TXCHOM).
- Klasa 5 uključuje posebno ugrožene opštine.

1

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
RFUCQ207	7.136	33451.31	21686.64	5872.162	6973.155	0
PCAD07	7.0243	49.5	25.951	2.2638	14.1798	0
REFEROUI	6.827	76.5062	52.3035	4.2268	14.9946	0
PART07	6.5859	7.25	4.8811	1.5207	1.5214	0
TXCHOMA07	-2.4288	9.125	11.4615	1.3636	4.0689	0.0151
PNDIP07	-4.1155	10	18.8182	1.7678	9.0628	0
INTAO07	-4.1354	0.6875	1.3427	0.4635	0.6701	0
DISTCONT	-4.6416	4.7828	10.4849	4.4925	5.196	0
P20ANS07	-4.8515	21.0625	25.6434	4.6296	3.9937	0
POUV07	-5.2392	5.125	15.1329	1.1659	8.0794	0
PINT07	-5.505	20.1875	25.4196	2.0377	4.0199	0
PEMP07	-6.179	17.8125	28.6224	1.8445	7.3995	0
PMONO07	-6.1847	6.3125	10.4545	0.9164	2.8327	0

2

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
PINT07	5.4139	28.1591	25.4196	3.4241	4.0199	0
RFUCQ207	5.3079	26345.68	21686.64	2413.373	6973.155	0
PRET07	5.2822	16.8409	15.1538	2.3543	2.5373	0
REFEROUI	5.1529	62.0295	52.3035	6.8878	14.9946	0
PCAD07	4.9169	34.7273	25.951	8.529	14.1798	0
P20ANS07	-3.0429	24.1136	25.6434	2.9559	3.9937	0.0023
INTAO07	-4.6013	0.9545	1.3427	0.2083	0.6701	0
PMONO07	-4.9078	8.7045	10.4545	1.1981	2.8327	0
POUV07	-5.9856	9.0455	15.1329	2.763	8.0794	0
PEMP07	-6.0851	22.9545	28.6224	3.1691	7.3995	0
PNDIP07	-6.6142	11.2727	18.8182	2.1676	9.0628	0
TXCHOMA07	-7.0246	7.8636	11.4615	1.3072	4.0689	0
PREFETR07	-7.0488	9.0909	16.3497	2.4292	8.1809	0

5

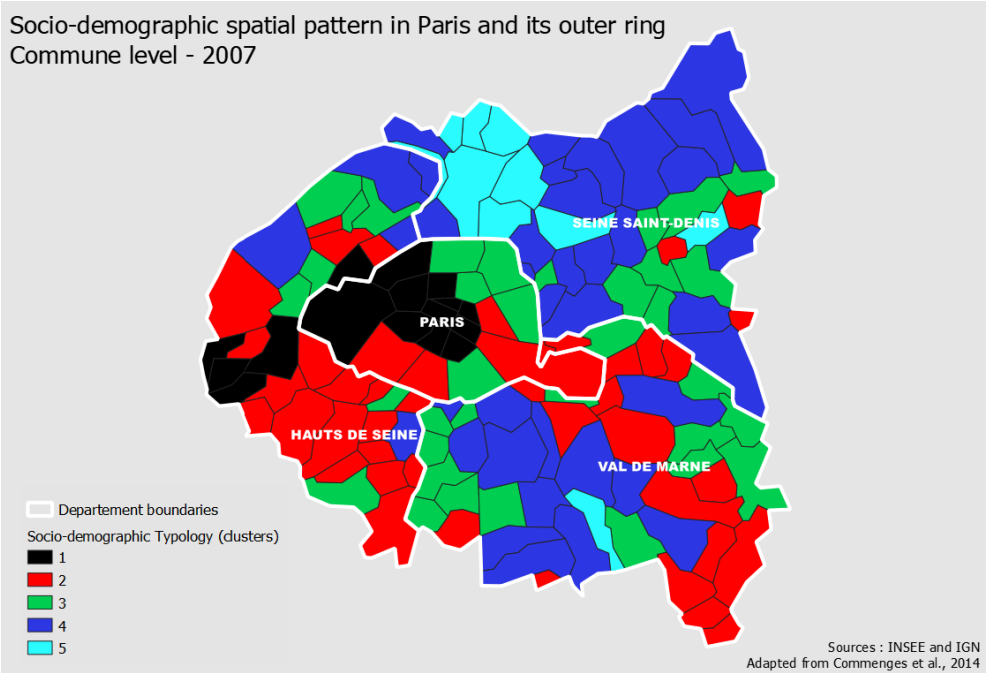
	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
PREFETR07	7.569	35.3	16.3497	5.2924	8.1809	0
PNDIP07	7.1682	38.7	18.8182	4.2673	9.0628	0
TXCHOMA07	6.937	20.1	11.4615	1.578	4.0689	0
POUV07	6.4576	31.1	15.1329	2.6627	8.0794	0
INTAO07	5.6435	2.5	1.3427	0.5	0.6701	0
PMONO07	5.5892	15.3	10.4545	1.6155	2.8327	0
P20ANS07	5.4462	32.3	25.6434	2.3685	3.9937	0
PEMP07	4.9359	39.8	28.6224	2.0396	7.3995	0
PCAD07	-4.4822	6.5	25.951	1.9105	14.1798	0
RFUCQ207	-4.9436	11136.7	21686.64	1094.014	6973.155	0
REFEROUI	-4.9626	29.53	52.3035	3.4954	14.9946	0
PRET07	-5.2205	11.1	15.1538	1.6401	2.5373	0
PINT07	-5.787	18.3	25.4196	1.4177	4.0199	0

4

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
PEMP07	6.9691	35.8158	28.6224	2.6343	7.3995	0
POUV07	6.7936	22.7895	15.1329	3.7216	8.0794	0
PNDIP07	6.5967	27.1579	18.8182	4.1646	9.0628	0
TXCHOMA07	6.4659	15.1316	11.4615	2.2144	4.0689	0
PMONO07	6.3086	12.9474	10.4545	1.5381	2.8327	0
INTAO07	5.9062	1.8947	1.3427	0.552	0.6701	0
PREFETR07	5.4125	22.5263	16.3497	4.5116	8.1809	0
P20ANS07	4.2774	28.0263	25.6434	2.2418	3.9937	0
PART07	-3.1599	4.2105	4.8811	0.8322	1.5214	0.0016
PRET07	-3.4831	13.9211	15.1538	1.5455	2.5373	5e-04
RFUCQ207	-6.5239	15340.74	21686.64	1642.102	6973.155	0
PCAD07	-6.7071	12.6842	25.951	4.1174	14.1798	0
REFEROUI	-7.0472	37.5632	52.3035	5.1588	14.9946	0

Tabela 4. Tabele koje daju opis klastera po varijablama (navedene su samo varijable s v-testom> 1.96)
Table 4. Tables giving the description of the clusters by the variables (only the variables with a v-test >|1.96| are listed)

Zajednička analiza karte (slika 9) i prethodnih paragrafa i tabela pokazuje skup dobrostojećih opština na zapadu Pariza i Hauts-de-Seine (klaster 1 u crnoj i klaster 2 u crvenom). S druge strane, sjeverni dio istraživanog područja, koji odgovara većini općina Seine-Saint-Denis, koncentrira više društvene nepovoljnosti (klaster 4 u tamnoplavoj i klaster 5 u svijetloplavoj boji). Na svojoj strani Val-de-Marne ima veoma kontrastnu situaciju sa 4 klastera od 5.



Sl. 9. Izlazna mapa koja prikazuje tipologiju u QGIS-u
Fig. 9. Output map portraying the typology in QGIS

DISKUSIJA I ZAKLJUČAK

FactoQGIS je algoritam koji omogućava besprijekorno izvoĐenje geometrijskih analiza podataka u QGIS okruženju. Konkretnije, ovaj alat je dizajniran da izvrši tipološku analizu kvantitativnih podataka agregiranih u prostorne jedinice. Ovaj alat se oslanja na R pakete i može se lako izvršiti iz GUI. To je od posebnog interesa za geografe i sve korisnike koji žele da jednostavno izgraĐuju i mapiraju u slobodnom i otvorenom GIS softveru, višedimenzionalnu tipologiju bez poznavanja R jezika. Alat predstavljen u ovom članku je prvo izdanje. Razmatra se nekoliko načina za poboljšanje. Na primjer, planiramo dodati mogućnost odabira dodatnih kvantitativnih i kvalitativnih varijabli. Planirana je i mogućnost primjene metode particioniranja na osnovu k-sredstava prije izvoĐenja HAC-a. Ovo će omogućiti da se FactoQGIS koristi na velikim skupovima podataka (nekoliko hiljada prostornih jedinica). Konačno, dodatni razvoj će omogućiti integraciju drugih tipova faktorske analize, posebno MCA i FAMD, kako bi se lako proizvele tipološke mape bez obzira na vrstu ulaznih podataka. Skripte svih ovih predstojećih razvoja će biti dostavljene zajednici pod GNU GPL licencom u GitHubu.

ZAHVALE

Studenti druge godine SIGAT Master studija 2018-2019 (Université Rennes 2, Francuska) i Mégane Bouquet (UMR ESO 6590 CNRS, Rennes, Francuska)

Reference
References

1. Benzécri, J.P., (1973) *L'Analyse des données*, Dunod, 619 p. ISBN 2-04-007225-X
2. Commenges, H. (dir.), (2014) *R et espace - Traitement de l'information géographique*. Groupe ElementR- Framabook. Available online: https://framabook.org/r-et-espace/ (accessed on May 13, 2019)
3. Demoraes, F.; Gouëset, V.; Piron, M.; Figueroa, O.; Zioni, S. (2013) Desigualdades socioterritoriais e mobilidades cotidianas nas metrópoles de América Latina: uma comparação entre Bogotá, Santiago de Chile e São Paulo. *Revista dos Transportes Públicos - ANTP*, Planejamento urbano, 35 (134), 9-30. Available online: https://halshs.archives-ouvertes.fr/halshs-01110019 (accessed on May 13, 2019)
4. Graser, A.; Olaya, V. (2015) Processing: A Python Framework for the Seamless Integration of Geoprocessing Tools in QGIS. Vol. 4, *ISPRS Int. J. Geo-Information*, 2219-2245. Available online: https://doi.org/10.3390/ijgi4042219 (accessed on May 13, 2019)
5. Greenacre M. J.; Blasius J. (2006). Multiple Correspondence Analysis and Related Methods. CRC press. ISBN 978-1-58488-628-0.
6. Hanspach, J.; Loos, J.; Dorresteijn, I.; Abson, D.; Fischer, J. (2016) Characterizing social–ecological units to inform biodiversity conservation in cultural landscapes, *Diversity and Distributions*, 1-12, Available online: https://doi.org/10.1111/ddi.12449 (accessed on May 13, 2019)
7. Husson, F.; Lê, S.; Pagès, J., (2009) *Analyse de données avec R*, Presses Universitaires de Rennes. 224 p. ISBN 978-2753509382
8. Le Roux, B.; Rouanet, H. (2004) *Geometric Data Analysis - From Correspondence Analysis to Structured Data Analysis*, Springer Netherlands, 475 p. ISBN 978-1-4020-2236-4
9. Lebart, L.; Piron, M.; Morineau, A. (2006) *Statistique exploratoire multidimensionnelle : visualisation et inférence en fouille de données*, Dunod. 464 p. ISBN 978-2100496167
10. Metzger, P. (2001) *Perfiles ambientales de Quito*. Quito: MDMQ; IRD, 117 p. ISBN 9978-41-682-X, Available online: http://www.documentation.ird.fr/hor/fdi:010026340 (accessed on May 13, 2019)
11. Piron, M.; Dureau, F.; Mullon, C. (2004) Dynamique du parc de logements à Bogota : Analyse par typologies multi-dates. *Cybergeo*, 1-23, Available online: https://journals.openedition.org/cybergeo/2925 (accessed on May 13, 2019)
12. Rivière, J. (2012) Mapping votes and social inequalities: the case of Paris and its inner suburbs, *Metropolitics*, 1-8, Available online: https://www.metropolitiques.eu/Mapping-votes-and-social.html (accessed on May 13, 2019)
13. Robson, B.; Lymperopoulou, K.; Rae, (2009) A. A typology of the functional roles of deprived neighbourhoods, Centre for Urban Policy Studies, Manchester University, Department for Communities and Local Government, 63 p. ISBN: 978-1-4098-1017-9, Available online: https://webarchive.nationalarchives.gov.uk/20120919132719/http:/www.communities.gov.uk/documents/communities/pdf/1152966.pdf (accessed on May 13, 2019)
14. Walsh, J. (2000) *Irish Rural Structure and Gaeltacht Areas*. National Spatial Strategy Report. Maynooth and Brady Shipman Martin. Available online: http://www.irishspatialstrategy.ie/docs/report10.pdf (accessed on May 13, 2019)

ONLINE IZVORI :

1. Blog on how to execute R scripts in QGIS 3.0 and later. Available online: https://github.com/north-road/qgis-processing-r/releases/tag/v0.0.2 (accessed on May 13, 2019)
2. List of the R scripts that can be executed from the QGIS Toolbox. Available online: https://github.com/qgis/QGIS-Processing/tree/master/rscripts (accessed on May 13, 2019)

3. Documentation of the FactoMineR package used in FactoQGIS. Available online: https://www.rdocumentation.org/packages/FactoMineR/versions/1.41 (accessed on May 13, 2019)
4. Documentation of the factoextra package used in FactoQGIS. Available online: https://www.rdocumentation.org/packages/factoextra/versions/1.0.5 (accessed on May 13, 2019)
5. Documentation of the stringr package used in FactoQGIS. Available online: https://www.rdocumentation.org/packages/stringr/versions/1.3.1 (accessed on May 13, 2019)
6. Documentation of the openxlsx package used in FactoQGIS. Available online: https://www.rdocumentation.org/packages/openxlsx/versions/4.1.0 (accessed on May 13, 2019)
7. Documentation of the R2HTML package used in FactoQGIS. Available online: https://www.rdocumentation.org/packages/R2HTML/versions/2.3.2 (accessed on May 13, 2019)
8. Documentation of the corrplot package used in FactoQGIS. Available online: https://www.rdocumentation.org/packages/corrplot/versions/0.84 (accessed on May 13, 2019)
9. Data set used in this article as an example to illustrate how FactoQGIS works. Available online: http://framabook.org/docs/Respace/RetEspace_Donnees.zip (accessed on May 13, 2019)

PRILOZI

FactoQGIS ulazni parametric

1 - Jezik

Francuski ili engleski. Ovaj parametar će definirati jezik koji će se primijeniti na opise tablica i grafika u izlaznoj html datoteci.

2 - Radni direktorij

Ovo polje je obavezno. Putanja do radnog direktorija mora biti kratka i ne smije sadržavati posebne znakove ili razmake. U njemu će biti pohranjene sve izlazne tablice i parcele.

3 - Sloj

Sloj na kojem treba primijeniti PCA i HAC. Atributna tablica ovog sloja mora sadržavati kvantitativne varijable. Ovaj sloj se mora učitati u QGIS.

4 - Naziv reda

Polje koje sadrži identifikator prostornih jedinica. Taj ID će se tada pojaviti na faktorskim mapama i takoĐer je potreban za spajanje podataka na kraju algoritma.

5 - Aktivne varijable

Aktivne varijable na kojima će se izvršiti PCA. Mora biti numerički. Aktivne varijable koje se pojavljuju na slici 2 detaljno su prikazane u tabeli 2.

6 - Skaliranje podataka

Mogućnost skaliranja i centriranja podataka. Trebalo bi se primijeniti u velikoj većini slučajeva, posebno kada je varijacija jedinice veoma različita između varijabli.

7 - Broj osi koje treba čuvati za PCA

Broj osi koje treba čuvati za PCA. 5 je podrazumjevana vrijednost. Generalno, zadržavamo N prvih faktora koji objašnjavaju najmanje 95% inercije. Preporučuje se da prvo podesite podrazumjevanu vrijednost i da provjerite tabelu vrijednosti Eigen i scree. Ako je potrebno, možete promijeniti zadanu vrijednost i izvesti drugi put PCA.

8 - Broj osi koje treba čuvati za HAC

Broj osi koje treba čuvati za HAC. 2 je podrazumevana vrijednost. Uopšteno, zadržavamo N prvih faktora koji objašnjavaju najmanje 80% inercije kako bi dobili stabilnije grupiranje. Preporučuje se da prvo podesite podrazumjevanu vrijednost i da provjerite tabelu vrijednosti Eigen i scree. Ako je potrebno, možete promijeniti zadanu vrijednost i izvesti drugi put HAC.

9 - Broj klastera za HAC

Broj klastera koji će se čuvati za HAC. 5 je podrazumjevana vrednost. Preporučuje se da prvo prepustite podrazumjevanu vrijednost i da provjerite hijerarhijsko stablo. Ako je potrebno, možete promijeniti zadanu vrijednost i izvesti drugi put HAC.

10 - Metrika za korištenje izgradnje stabla
Metrika se koristi za izračunavanje razlika između pojedinaca. Trenutno dostupne opcije su "euklidski" i "manhattan". Euklidske udaljenosti su korijeni kvadrata razlika, a manhattanove udaljenosti su zbroj apsolutnih razlika. Podrazumjevana vrijednost je "euklidska".
11 - Metoda agregacije koja će se koristiti za definiranje klastera
Metoda grupisanja. Četiri primenjene metode su "prosječna" (metoda aritmetičkih sredina neparnih parova grupa), "jednostruka" (jednostruka veza), "potpuna" (potpuna veza) i "odjeljenje" (Wardova metoda). Wardova metoda je najčešće korišćena i podrazumjevana je vrijednost.

FactoQGIS izlazi
12 - Tabela vrijednosti Eigen
Tabela Eigen vrijednosti koja za svaku varijablu daje svoj dio globalnoj inerciji. Ova tablica se automatski dodaje sadržaju u QGIS i također se izvozi u tablični list programa Excel.
13 - Tabela varijabilnih koordinata
Tabela koja daje koordinate svake varijable na osi. Ova tablica se automatski dodaje u tablicu sadržaja u QGIS-u i također se izvozi u tablični list programa Excel.
14 - Sloj sa klasterima
Izlazni sloj vektora sa kolonom koja označava klaster kojem pripada svaka prostorna jedinica. Ovaj sloj se automatski dodaje sadržaju u QGIS-u kako bi se olakšala mapiranje tipologije.

Postavke i aktiviranje R skripti u QGIS
Postupak je dostupan online:
https://docs.qgis.org/2.18/en/docs/training_manual/processing/r_intro.html (Pristupljeno 13. maja 2019)

Pisanje novih algoritama za obradu poput python skripti
Postupak je dostupan online:
https://docs.qgis.org/2.8/en/docs/user_manual/processing/scripts.html (Pristupljeno 13. maja 2019)

SUMMARY

FACTOQGIS: A GUI TOOL BASED ON AN R SCRIPT TO PERFORM GEOMETRIC DATA ANALYSIS IN A FREE AND OPEN SOURCE GIS

Florent Demoraes, Univ Rennes, CNRS, ESO - UMR 6590, F-35000 Rennes, France
florent.demoraes@univ-rennes2.fr

Marc Souris, UMR Unité des Virus Emergents (UVE : Aix-Marseille Univ – IRD 190 – Inserm 1207 – IHU Méditerranée Infection), Marseille, France
marc.souris@ird.fr

FactoQGIS is an algorithm that allows a seamless execution of geometric data analysis in the QGIS environment. More specifically, this tool was designed to perform a typological analysis on quantitative data aggregated in spatial units. This tool relies on R packages and can easily be executed from a GUI. It is of particular interest to geographers and to any users who wish to simply build and map in a free and open source GIS software, a

multidimensional typology without knowing the R language. The tool presented in this article is a first release. Several avenues for improvement are being considered. For example, we plan to add the possibility of choosing supplementary quantitative and qualitative variables. The option of applying a partitioning method based on the k-means before performing the HAC is also planned. This will allow FactoQGIS to be used on large datasets (several thousand spatial units). Finally, additional developments will make it possible to integrate other types of factor analysis, in particular MCA and FAMD, in order to easily produce typology maps whatever the kind of input data. The scripts of all these forthcoming developments will be provided to the community under GNU GPL license in GitHub.

Authors

Florent Demoraes, Full professor at the Université Rennes 2 (areas: Latin American metropolises, teaching activities, teaching topics, GIS theory, basics of spatial analysis, etc.), Director of ESO-Rennes Research Unit - UMR CNRS 6590 Spaces and Societies. He is a member of the Scientific Council of the American Institute in Rennes. Head of ESO collection in HAL (Open Archive).
Scientific area of research includes geomatics, spatial analysis, spatial statistics, spatial mobility, residential segregation and social inequalities (spatial dimension).

Dr. Marc Souris, is senior research director at the Institut de Recherche pour le Développement (IRD), in the UMR 190 "Unité des virus émergents". He holds a PhD in Computer Science.
Mathematician and computer scientist, his work ainly concerns information sciences applied to geography and epidemiology. Since 1983, he has been developing research, innovation, software, and teaching in geomatics: geographic information systems, spatial analysis, statistics, modelling. He is the main author of the SavGIS GIS software package (www.savgis.org).