Klasterovanje “the OpenFlights Airports Database extended” skupa podataka

Seminarski rad u okviru kursa Istraživanje podataka 1
Matematički fakultet

David Popov
Avgust 2019.

Sadržaj

1 Opis skupa podataka

The OpenFlights Airports Database extended je skup koj sadrži podatke o areodromima, lukama i železničkim stanicama, tako da svaki slog sadrži geografsku širinu i dužinu, nadmorsku visinu, vremensku zonu, ime, grad i državu gde se nalazi. Podaci se dobijaju u CSV formatu.

Detaljan opis skupa podataka:

Ime atributa	Tip podataka	Opis
Airport id	Integer	Jedinstven indetifikacion broj.
Name	String	Naziv areodroma, železničke stanice ili luke.
City	String	Ime grada gde se nalazi areodrom, železničke stanica ili luka.
Country	String	Ime države gde se nalazi areodrom, železničke stanica ili luka.
IATA	String	Kod koj se sastoji od tri slova. Ukoliko nema vrednost postavljeno je na null.
ICAO	String	Kod koj se sastoji od četri slova. Ukoliko nema vrednost postavljeno je na null.
Latitude	Float	Geografska širina. Negativna vrednost predstavlja jug, a pozitivna sever.
Longitude	Float	Geografska dužina. Negativna vrednost predstavlja zapad, a pozitivna istok.
Altitude	Float	Nadmorska visina u fitima.
Timezone	Float	Vremenska zona koja je prikazana kao razlika sati u odnosu na vreme u Griniču.
DST (Daylight Savings Time)	Char	Predstavlja da li se koristi zimski ili letnji način računanja vremena. Vrednosti: E (Evropa), A (US/Kanada), S (Južna Amerika), O (Australia), Z (Novi Zeland), N (Nije obrađeno) ili U (Nepoznato).
Tz database	String	Prikazivanje vremenske zone u „tz“ formatu.
Type	String	Kog je tipa. Vrednost „airport“ je za areodrome, „station“ za železničke stanice, „port“ za luke i „unknown“ ukoliko je nije poznato.
Source	String	Odakle potiču podaci

2 Korišćeni alati

Za predprocesiranje je korišćen jezik Python sa njegovim bibliotekama Pandas, Timezonefinder, Pytz, Datetime. Za obradu podataka (klasterovanje) i vizuelizaciju je korišten IBM SPSS.

3 Preprocesiranje i vizuelizacija podataka

U ovom delu ćemo se upoznati detaljnije sa podacima i koracima koje sam preduzeo u preprocesiranju radi dobijanja što boljih rezultata.

Kako bih se što bolje upoznao sa podacima morao sam da izvršim analizu. Kako neki podaci sadrže nepostojeće vrednosti, ti podaci nisu korišćeni u daljoj analizi tj. izbačeni su tokom preprocesiranja.

Kolone koje izbacujem iz analize su one koje ili imaju jedinstvenu vrednost za svaki slog ili ima tolko različitih vrednosti da je približno jednako jedinstvenoj vrednosti za svaku kolonu.

Airport id se prvo izbacuje jer je jedinstven za svaki slog.

Polje City se izbacuje jer ima previse razlicitih gradova, tj. jedan grad moze da ima svega nekoliko areodroma, stanica i luka, pa mi taj udeo u 12000 slogova ništa ne znači.

Što se tiče polja Country kako i tu ima previše različitih vrednosti, a nisam hteo da skroz izbacim to polje, napravio sam novo polje Country New Values koje ima vrednosti 19 zemalja sa najvecim brojem areodroma, stanica i luka, dok sam svim ostalim slogovima dodelio vrednost Others.

Iz polja DST sam uklonio nepostojeće vrednosti.

Polja IATA i ICAO se izbacuju jer ima previše raziličitih vrednosti za slogove pa mi u daljoj analizi ne znače.

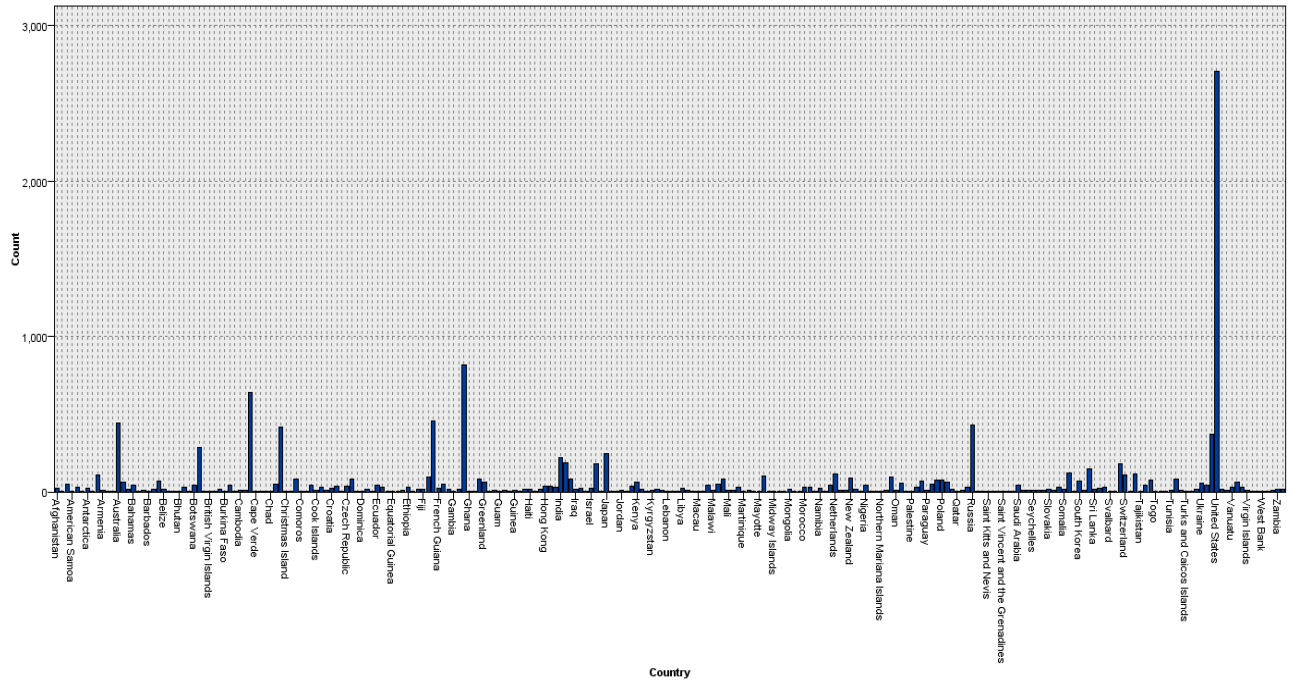
Polje Name se izbacuje jer svaki areodrom, luka i železnička stanica ima različito ime pa mi u daljoj analizi ne znače.

Iz polja Source sam uklonio nepostojeće vrednosti.

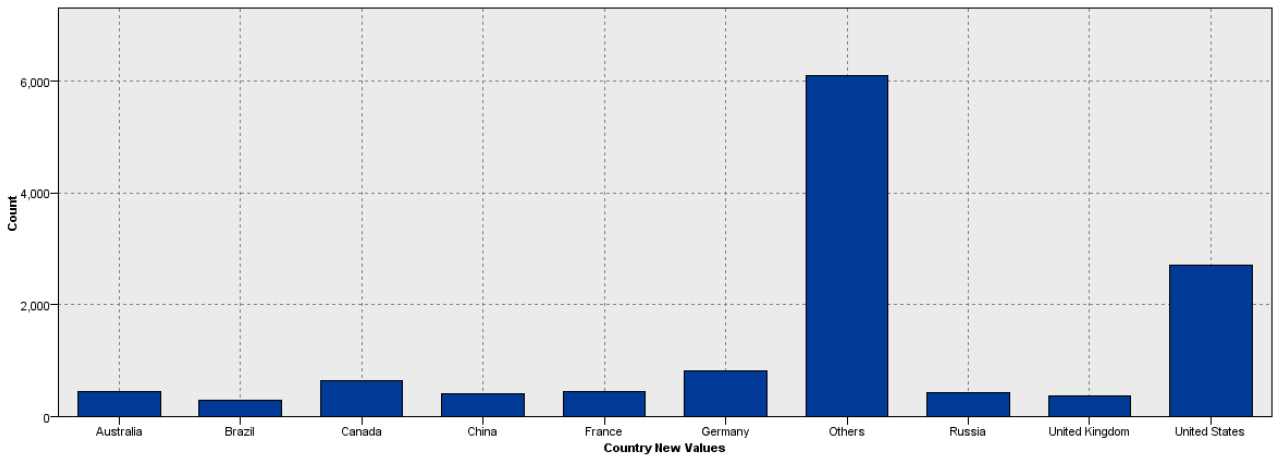
Kako je polje Timezone sadržalo nepostojeće vrednosti, nisam želeo i da ih izbacim već sam u Python-u iskoristio bibiloteku timezonefinder gde sam pomoću geografske širine i dužine dobio koja su vremenska zona i samo zamenio sa nepostojećim vrednostima.

Iz polja Type sam uklonio nepostojeće vrednosti.

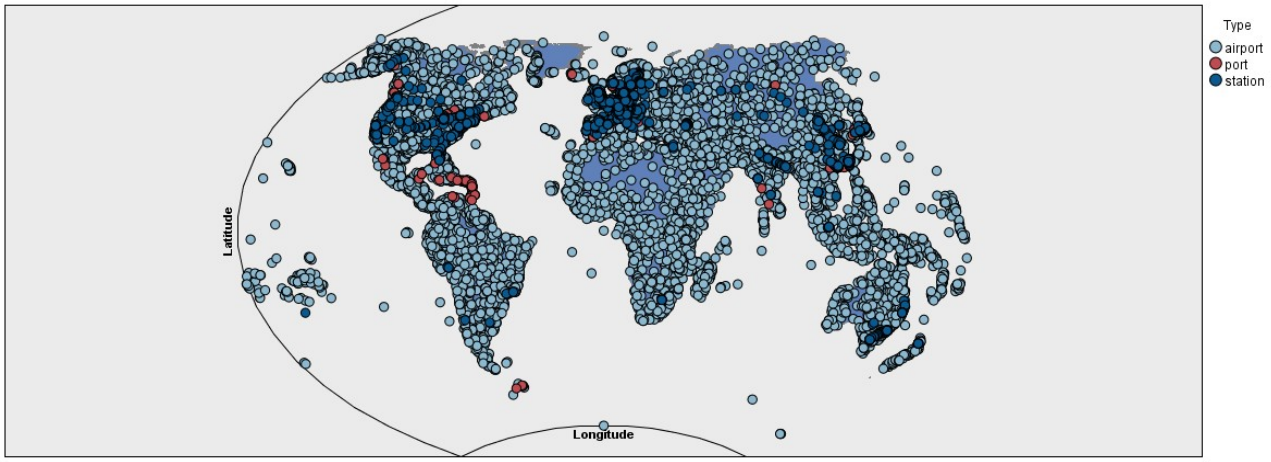
Polje Tz database sam takođe izbacio jer ima previše različitih vrednosti pa mi onda nista ne zanči u daljoj analizi.



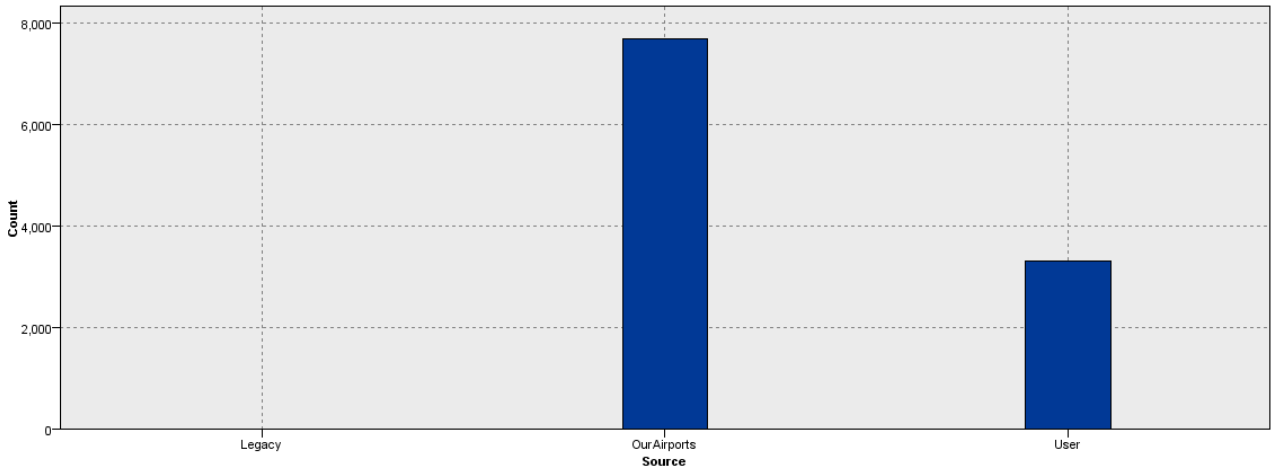
Slika 3.1: *Polje Country pre izmene*



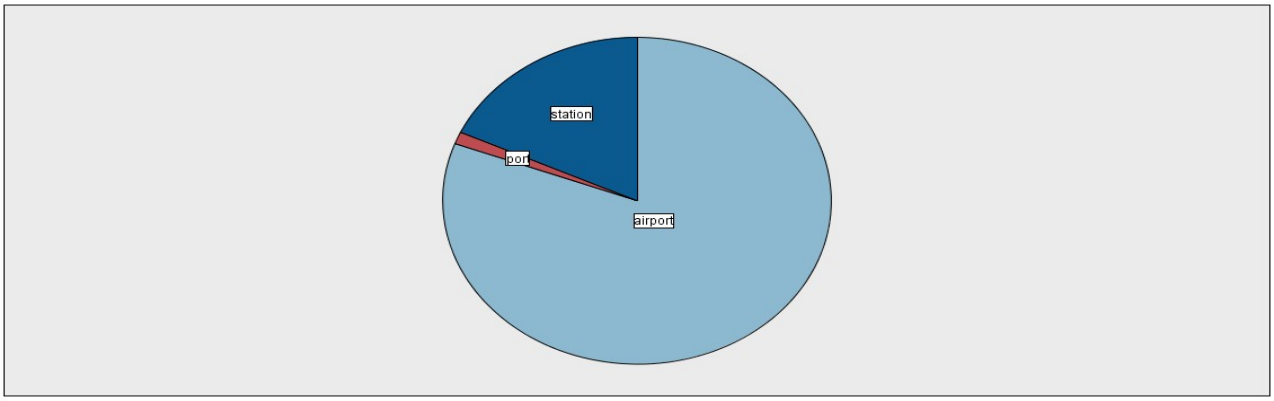
Slika 3.2: *Country New Values (Izmenjeno polje Country)*



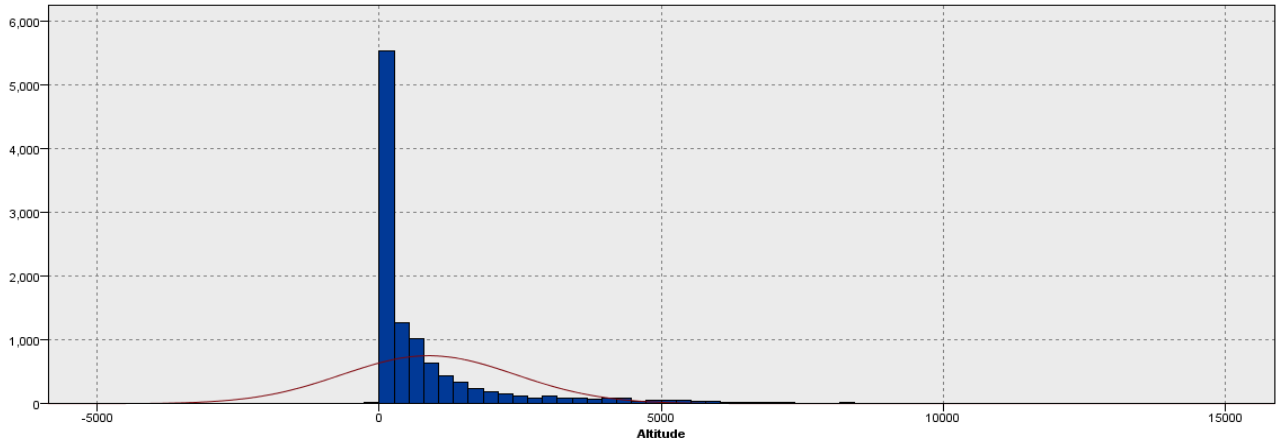
Slika 3.3: *Mapa rasporeda areodroma, luka i železničkih stanica*



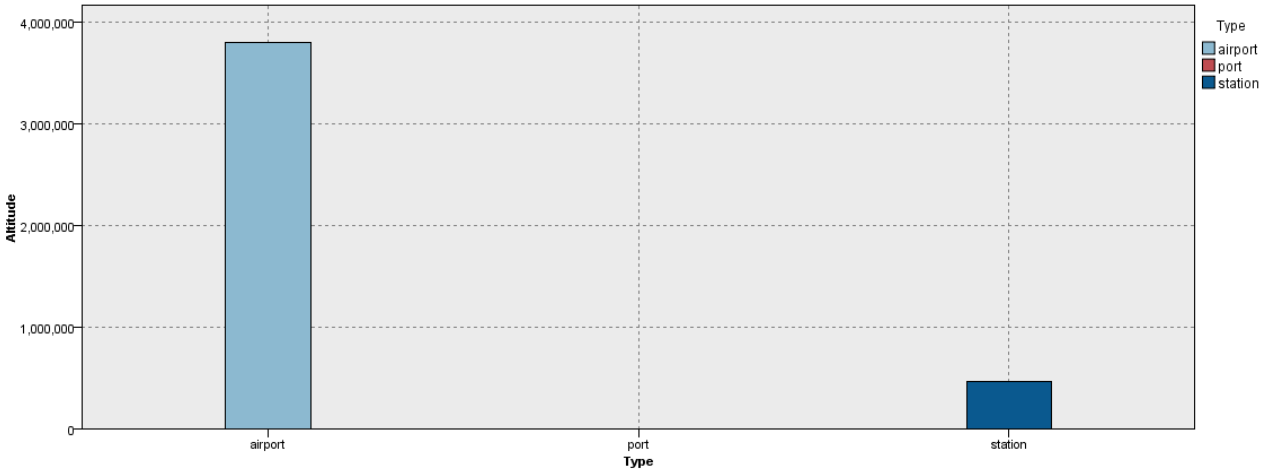
Slika 3.4: *Izvor podataka u odnosu na broj slogova*



Slika 3.5: *Zastupljenost areodroma, luka i železničkih stanica*



Slika 3.6: *Histogram nadmorske visine*



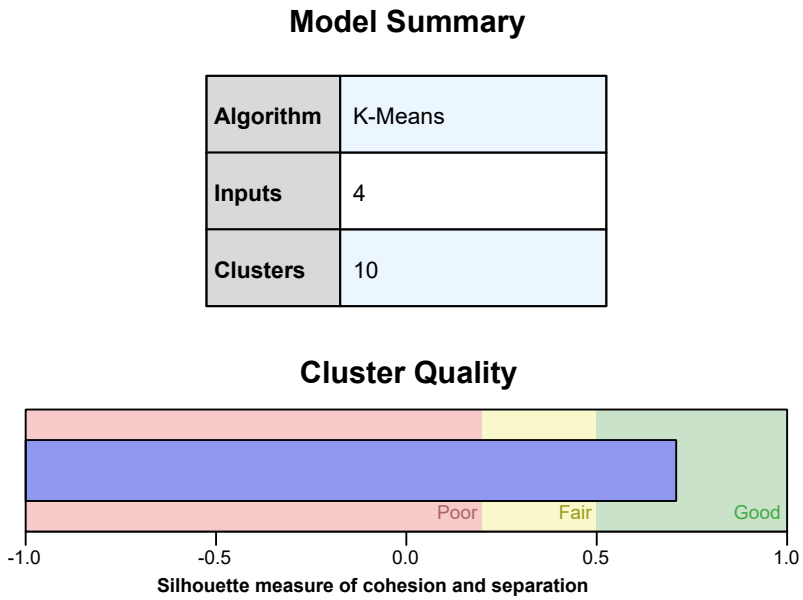
Slika 3.7: *Odnos nadmorske visine i tipa (areodrom, luka, železnička stanica)*

4 Klasterovanje

Ideja klasterovanja jeste da podelimo naše podatke u grupe (klastere) prema određenim zavisnostima. Prilikom analize koristili su se K-means i Kohonen algoritam

4.1 K-means

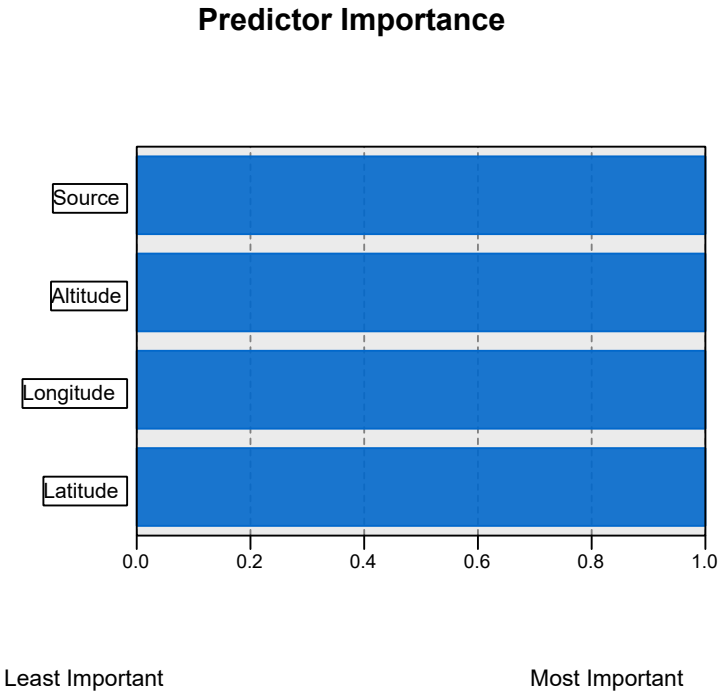
Prva ideja je bila da se vidi koj je maksimalan broj atributa koji može da se koristi tako da klasterovanje prikazuje dobre rezultate.



Slika 4.1.1: *Koeficijent seneke za prvu ideju*

Dobilo se da je najbolje kad je broj klastera 10 i kad se se kao ulazni podaci koriste atributi: Altitude, Latitude, Longitude i Source. Koeficijent senek koj se dobije je 0.7.

Dalje možemo videti da sva 4 atributa maksimalno utiču na rezulate klasterovanja.



Slika 4.1.2: *Zavisnost atributa od važnosti za predviđanje*

Clusters

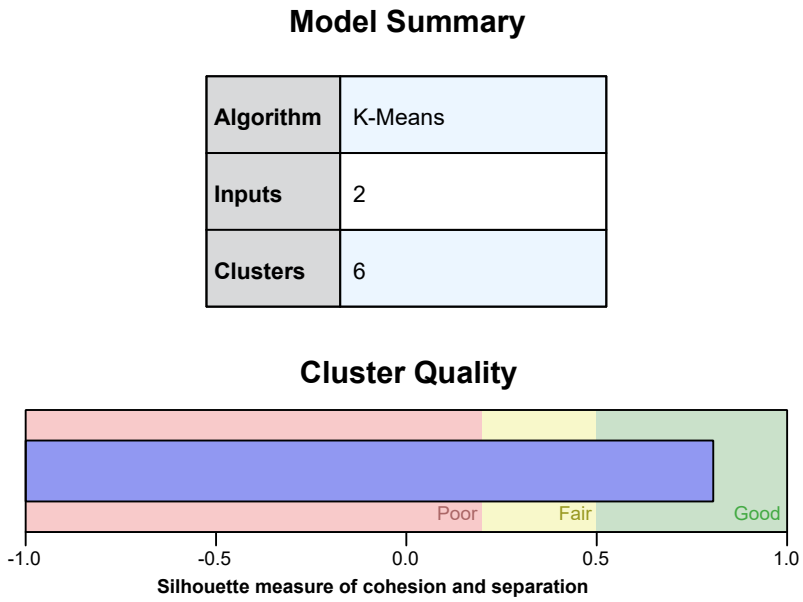
Input (Predictor) Importance

1.0 0.8 0.6 0.4 0.2 0.0

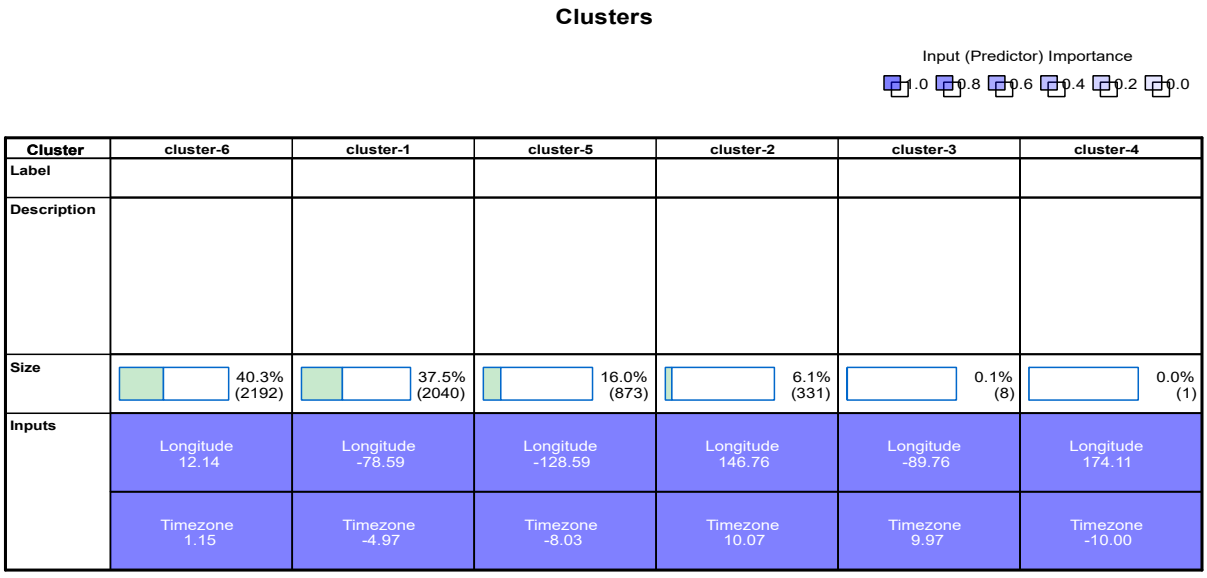
Cluster	cluster-1	cluster-9	cluster-6	cluster-4	cluster-7	cluster-5	cluster-10	cluster-2	cluster-8	cluster-3
Label										
Description										
Size	30.6% (1666)	29.1% (1584)	12.9% (705)	11.0% (597)	6.2% (336)	5.2% (281)	3.9% (214)	0.9% (47)	0.3% (14)	0.0% (1)
Inputs	Altitude 589.27	Altitude 693.30	Altitude 506.26	Altitude 446.32	Altitude 1,119.10	Altitude 453.92	Altitude 4,792.61	Altitude 371.91	Altitude 1,755.07	Altitude 5,355.00
	Latitude 43.23	Latitude 48.24	Latitude 42.44	Latitude 50.86	Latitude -18.35	Latitude -29.85	Latitude 36.95	Latitude -34.11	Latitude -31.02	Latitude 35.04
	Longitude -98.28	Longitude 13.60	Longitude -96.60	Longitude 9.02	Longitude -51.72	Longitude 146.95	Longitude -107.71	Longitude 149.64	Longitude -58.74	Longitude -106.61
	Source OurAirports (100.0%)	Source OurAirports (100.0%)	Source User (100.0%)	Source User (100.0%)	Source OurAirports (100.0%)	Source OurAirports (100.0%)	Source OurAirports (100.0%)	Source User (100.0%)	Source User (100.0%)	Source Legacy (100.0%)

Slika 4.1.3: *Rezultati klasterovanja prve ideje*

Druga ideja je bila da se izvrši klaster analiza samo za atribute Timezone i Longitude. Najbolji rezultati se dobijaju kad imam 6 klastera, dok se za koeficijent senke dobija vrednost 0.8.

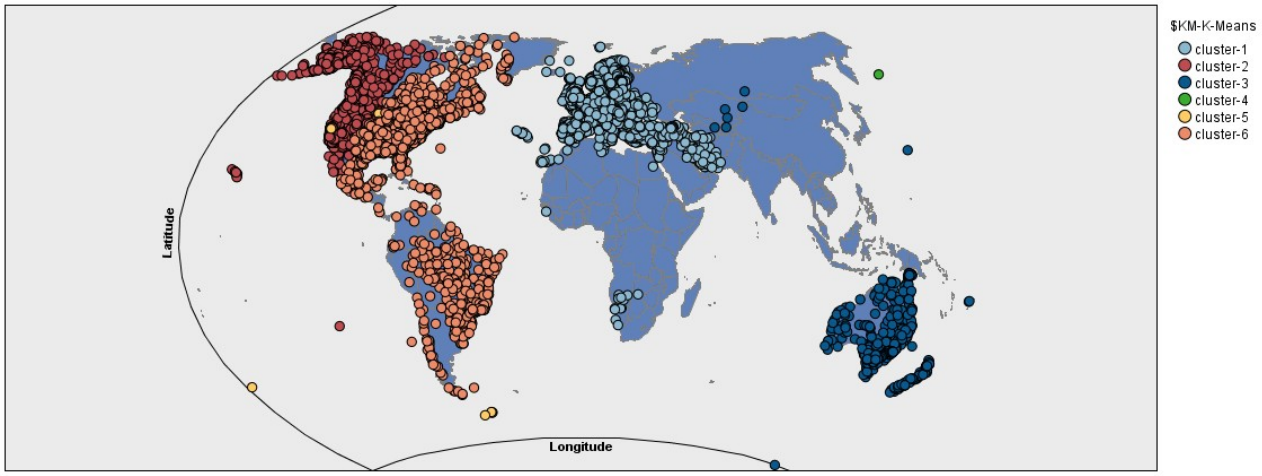


Slika 4.1.4: *Koeficijent senke za drugu ideju*



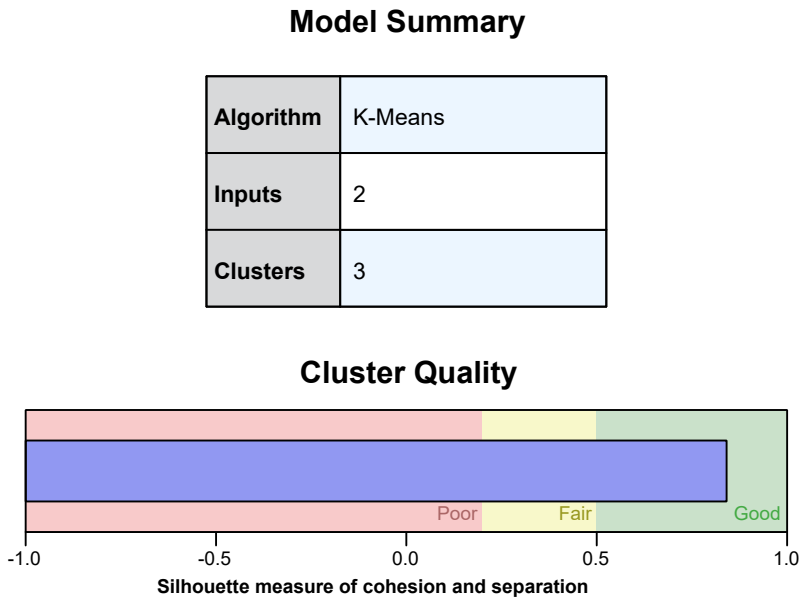
Slika 4.1.5: *Rezultati klasterovanja za drugu ideju*

Ono sto možemo da zaključimo jeste da su se klasteri grupisali po meridijanima i vremenskim zonama, što nam i jeste ocekivan rezultat.
Prvi klaster se prostire kroz Evropu i Afriku.
Drugi klaster je zauzeo predeo oko Aljaske i zapadnog dela Severne Amerike.
Treći klaster se prostire kroz Australiju, Okeaniju i Novi Zeland.
Četvrti klaster je zauzeo mali deo kod severne Azije.
Peti klaster je zauzeo delove u tihom okeanu.
Šesti klaster se protire kroz istočni deo Severne Amerike i kroz Južnu Ameriku.



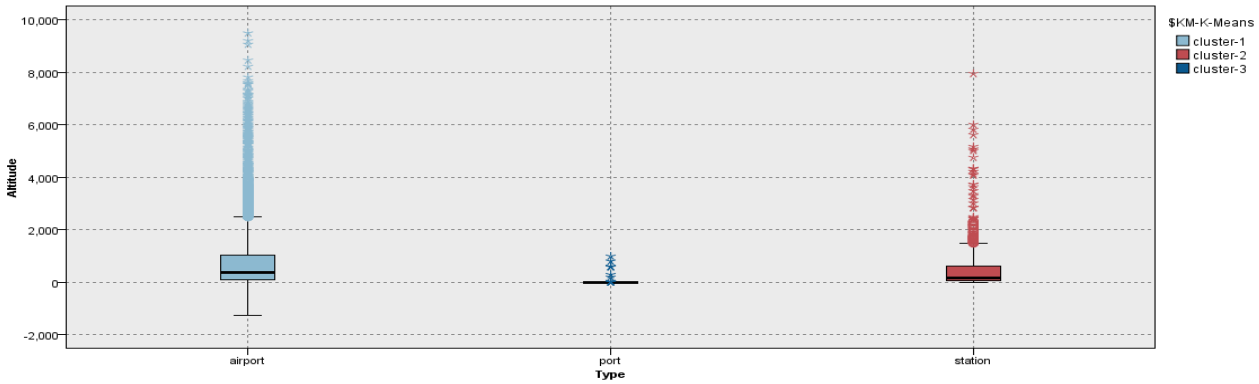
Slika 4.1.6: *Mapa klastera za drugu ideju*

Treća ideja je bila da se izvrši klaster analiza samo za atribute Type i Altitude. Najbolji rezultati se dobijaju kada se koriti 3 klastera, a za koeficijent senke dobijamo rezulata 0.8.



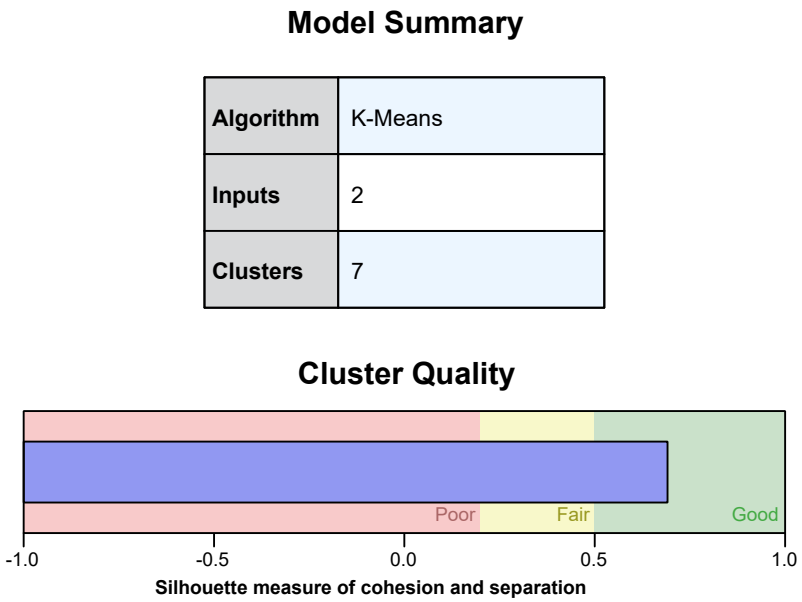
Slika 4.1.7: *Koeficijent seneke za treću ideju*

Rezultat koj se dobija je da sva tri klastera sadrze različite tipove.
Prvi sadrži samo areodrome, gde je prosečna visina 866.54 ft. Drugi sadrži samo železničke stanice, gde je prosečna visina 474.71 ft. Treći sadrži samo luke, gde je prosečna visina 71.30 ft. Rezultati koji se dobijaju su krajnje logični, luke će uvek biti na nadmorskoj visina koja je bliža nadmorskoj visina vode (0 ft), dok areodromi i železničke stanice mogu da se grade i na većim nadmorskim visinama.



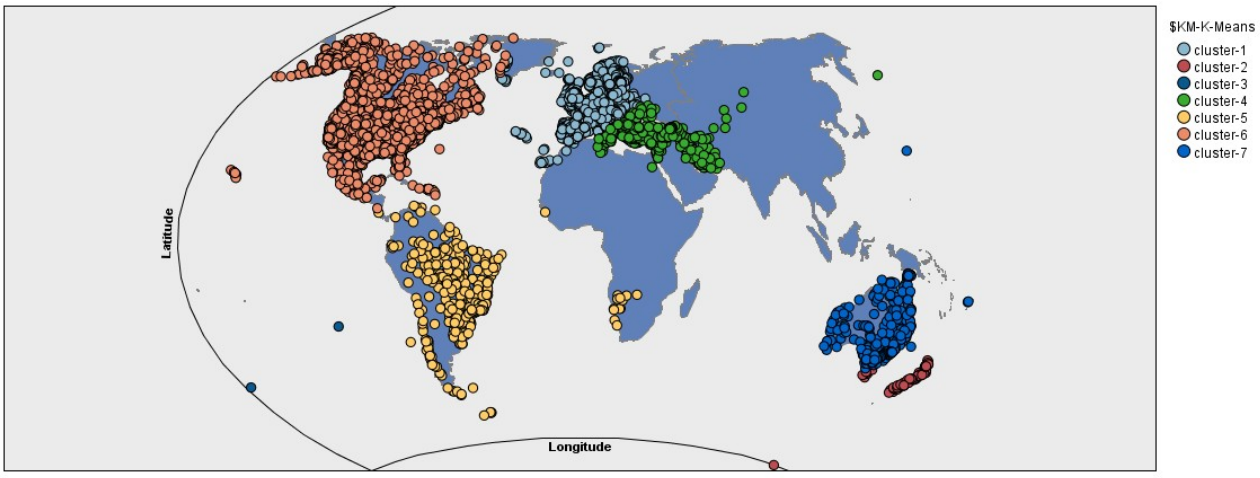
Slika 4.1.8: *Boxplot klasterovanja za treću ideju*

Četvrta ideja je bila da se izvrši klaster analiza samo za atribute Longitude i Latitude. Najbolji rezultati se dobijaju kada se koriti 7 klastera, a za koeficijent senke dobijamo rezulata 0.7.



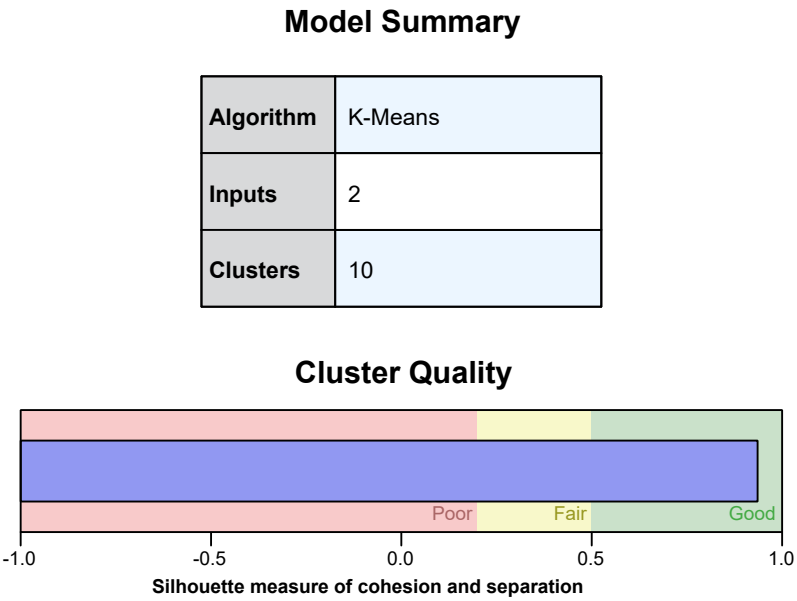
Slika 4.1.9: *Koeficijent seneke za četvrtu ideju*

Rezultati koj se dobiju su da svaki klaster zauzima po neki deo kontinenta.
Prvi zauzima područije Evrope. Drugi zauzima prostor Novog Zelanda. Treći zauzima prostor Austrlije. Četvrti zauzima deo prostora Azije, tj. Malu Aziju i bliski istok. Peti deo Afrike i Južnu Ameriku. Šesti zauzima prostor Severne Amerike. Sedmi obuhvata ostrva u Tihom okeanu.

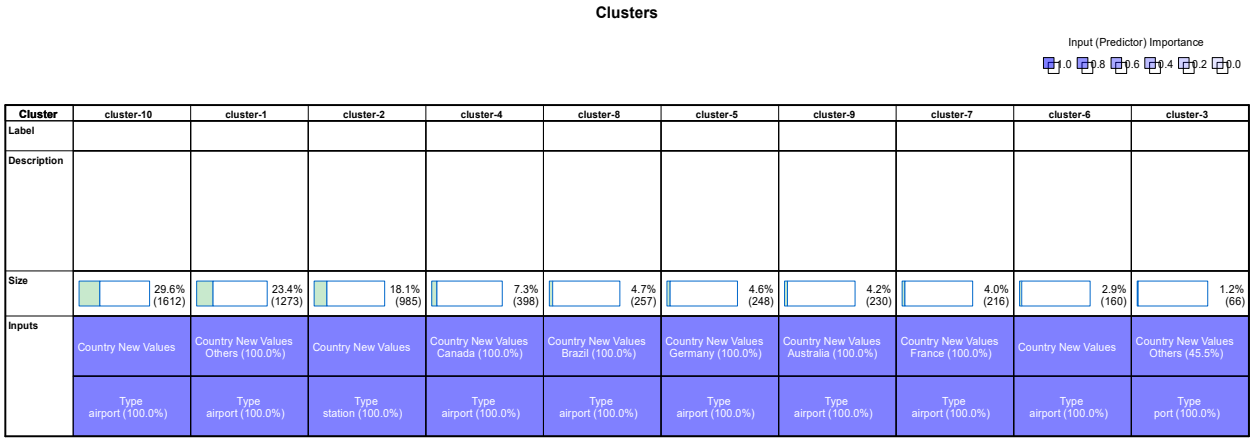


Slika 4.1.10: *Mapa klastera za četvrtu ideju*
Rezultati su ocekivani s obzirom da smo klasterovali po geografskoj širini i dužini.

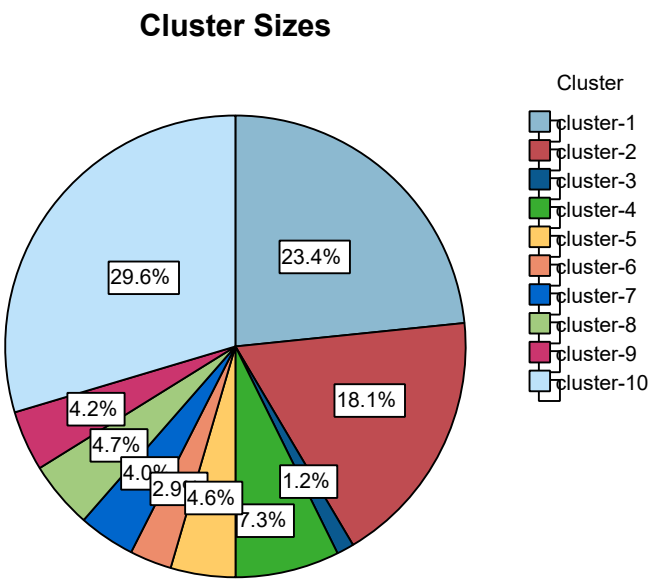
Peta ideja je bila da se izvrši klaster analiza samo za atribute Country New Values i Type. Najbolji rezultati se dobijaju kad imam 10 klastera, dok se za koeficijent senke dobija vrednost 0.9.



Slika 4.1.11: *Koeficijent seneke za petu ideju*



Slika 4.1.12: *Rezultati klasterovanja pete ideje*

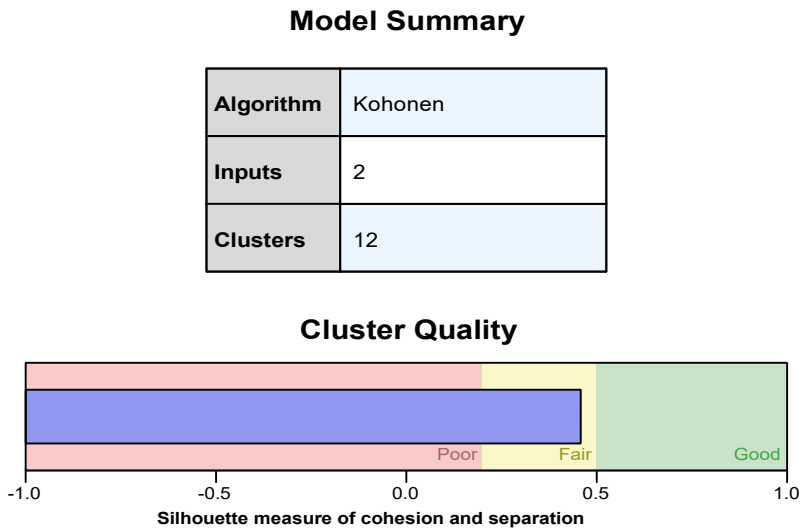


Size of Smallest Cluster	66 (1.2%)
Size of Largest Cluster	1612 (29.6%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	24.42

Slika 4.1.13: *Klasteri pete ideje*

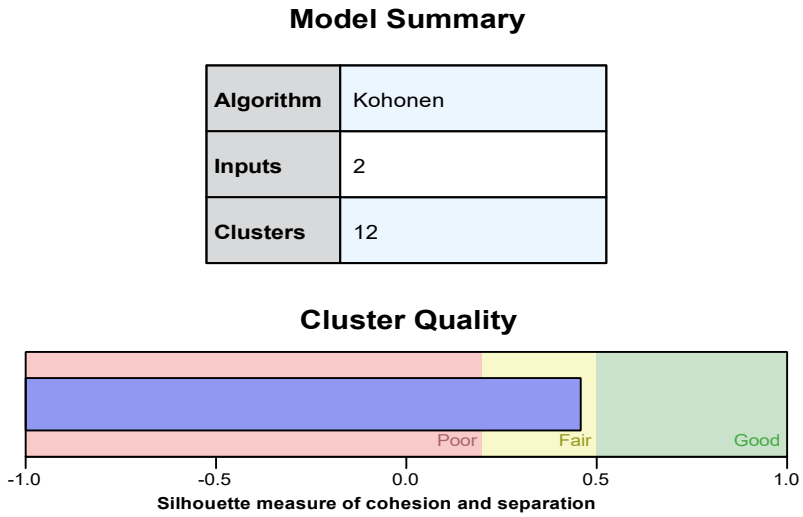
4.2 Kohonen

Prva ideja je bila da se izvrši kohonen algoritam po podrazumevanim vrednostima, ali je klasterovanje dalo lose rezultate, tj. koeficijent senke je imao vrednost 0.4.



Slika 4.2.1: *Koeficijent seneke za prvu ideju*

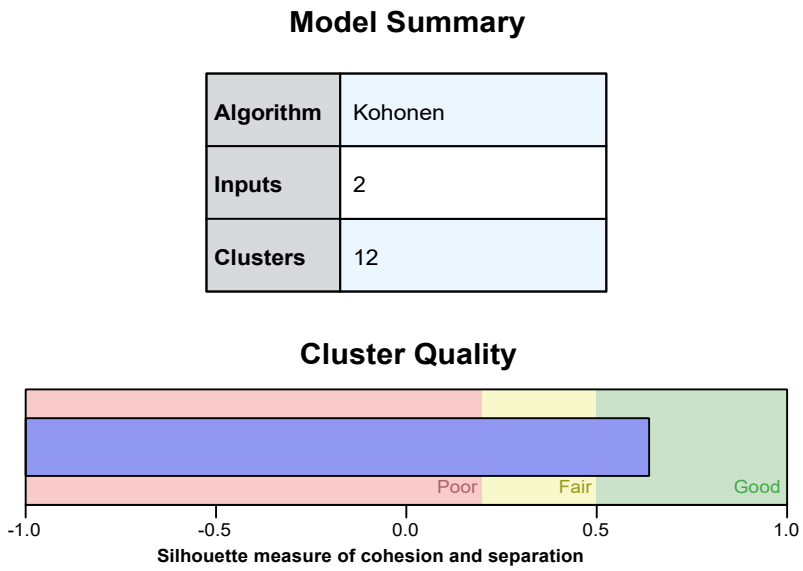
Druga ideja je bila da se izvrši kohonen algoritam nad atributima Longitude i Latitude. Ali i on je dao lose rezultate, vrednost koeficijenta senke je bila nesto manje od 0.4.



Slika 4.2.2: *Koeficijent seneke za drugu ideju*

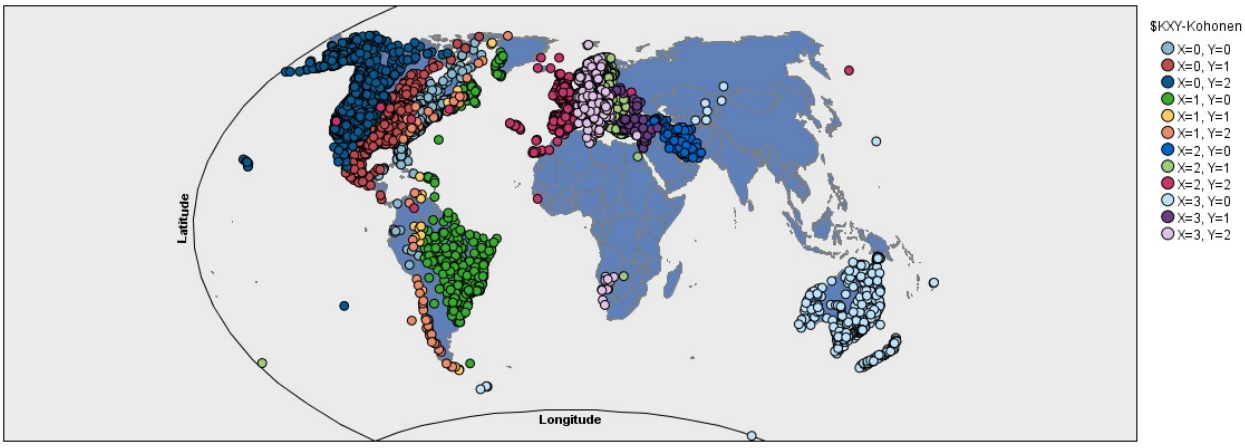
Prethodna klasterovanja nećemo dodatno objašnjavati jer su dala loše rezultate.

Treća ideja je bila da se izvrši kohonen algoritam nad atributima Longitude i Time. Ovaj put klasterovanje je dalo nešto bolje rezultate i koeficijent senke je imao vrednost 0.6.



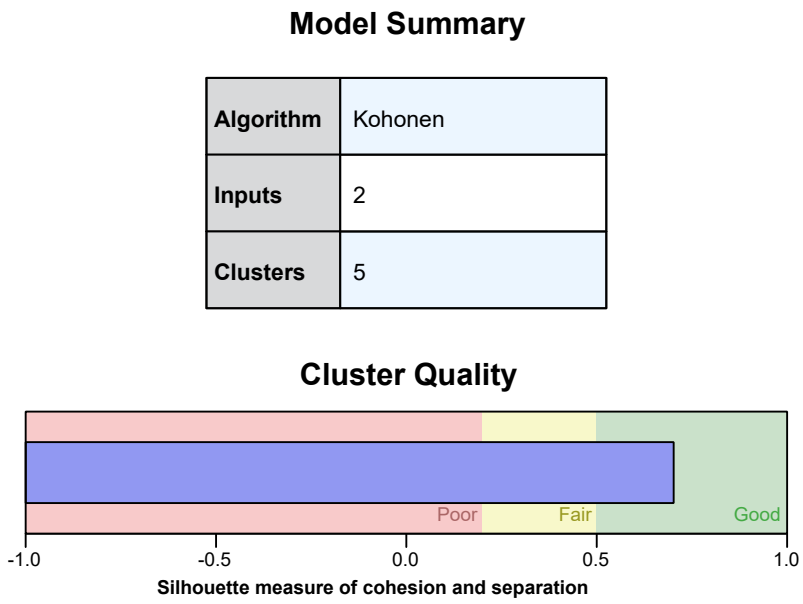
Slika 4.2.3: *Koeficijent seneke za treću ideju*

Rezultata koj se dobija sličan je ko i kada se koristi k-means algoritam, s tim da ima više klastera pa su samim tim gušće raspoređeni. Klasteri se rasporedjuju po medijanama, što je i za očekivati.



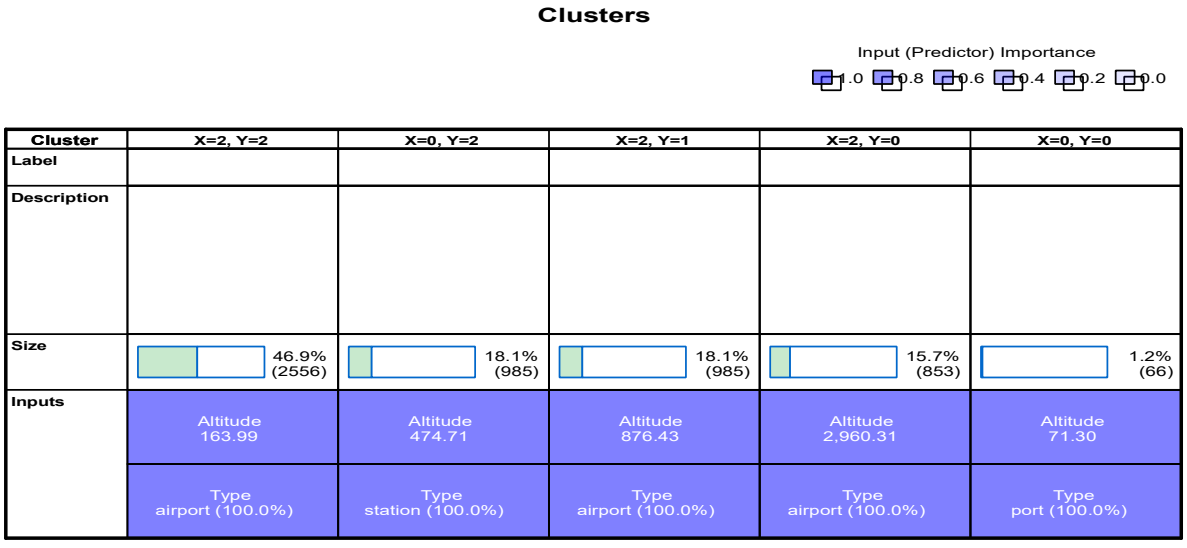
Slika 4.2.4*: Mapa klastera za drugu ideju*

Četvrta ideja je bila da se izvrši kohonen algoritam nad atributima Altitude i Type. Dobili smo 5 klastera, a vrednost koeficijenta senek je 0.7.



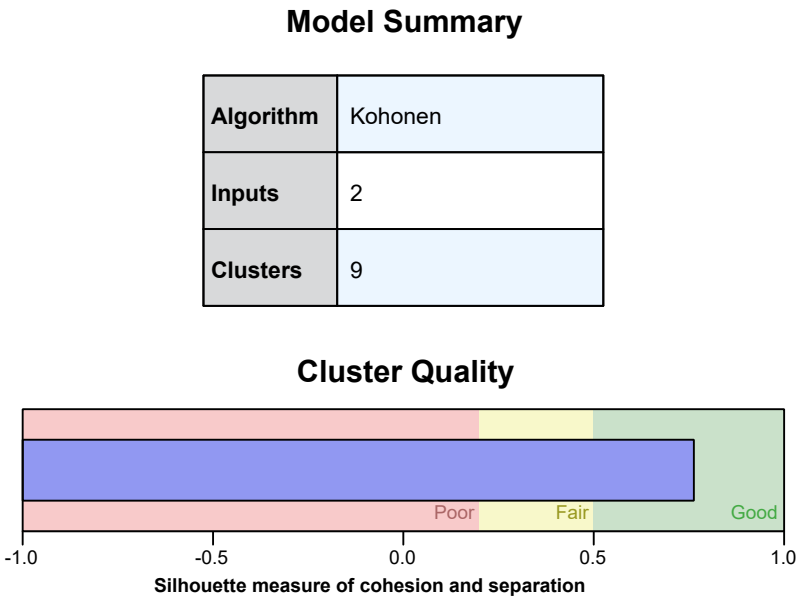
Slika 4.2.5: *Koeficijent seneke za četvrtu ideju*

Rezultata koj se dobija sličan je ko i kada se koristi k-means algoritam, s tim da su klaster koj sadrže luke i železničke stanice ostaju isit, dok klaster koj sadrži areodrome je razbijen na tri, dok u kad smo koristili k-means imali smo samo jedan klaster.

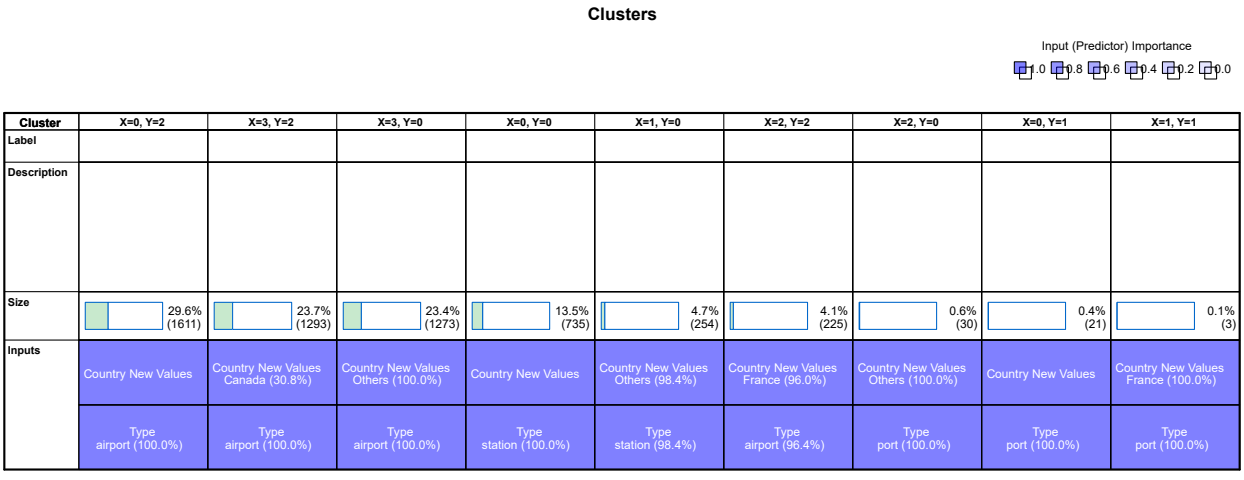


Slika 4.2.6: *Rezultati klasterovanja četvrte ideje*

Peta ideja je bila da se izvrši kohonen algoritam nad atributima Country New Values i Type. Dobili smo 9 klastera, a vrednost koeficijenta senek je 0.8. Ovde dobijamo najbolje klasterovanje kohonenovim algoritmom. U odnosu na k-means imamo jedan klaster manje. Ostali klasteri su dosta slični sa klasterima dobijenim k-means algoritmom.



Slika 4.2.7: *Koeficijent seneke za petu ideju*



Slika 4.2.8: *Rezultati klasterovanja pete ideje*