

---

---

# Klaster analiza “the OpenFlights Airports Database extended” skupa podataka

Seminarski rad u okviru kursa Istraživanje podataka 1

Matematički fakultet

Avgust 2019.

David Popov 102/16.

---

# Opis skupa podataka

The OpenFlights Airports Database extended je skup koji sadrži podatke o aerodromima, 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.

Ime atributa	Tip podataka	Opis
Airport id	Integer	Jedinstven identifikacion 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 četiri 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

# Alati korišćeni za rad

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





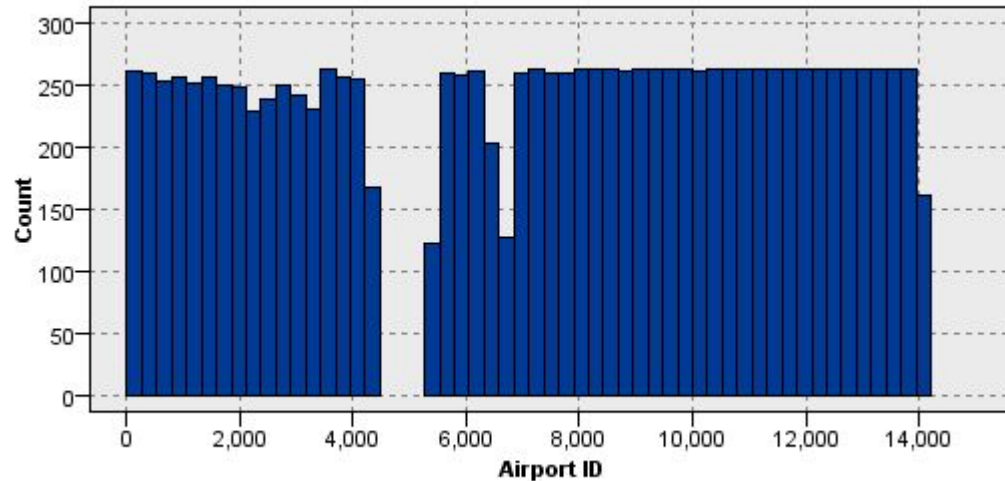
# Preprocesiranje

Prilikom analize podatke, izbacuju se oni podaci i kolone koji su ispunili sledeće uslove:

- Kolona koje imaju jedinstvenu vrednost za svaki slog
- Kolone koje su uzimale vrednosti iz velikog skupa podataka
- Slogove koji sadrže nepostojeće vrednosti

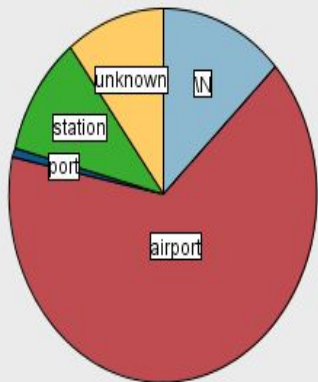
—

Kolona Airport Id se izbacuje jer je sadržala jedinstvenu vrednost za svaki slog. Dok kolone City, IATA, ICAO, Name i Tz database imaju preveliki skup podataka odakle uzimaju vrednost, pa se izbacuju jer bi se u daljoj analizi bi se ponašale slično kolonama koje imaju jedinstvenu vrednost za svaku kolonu.

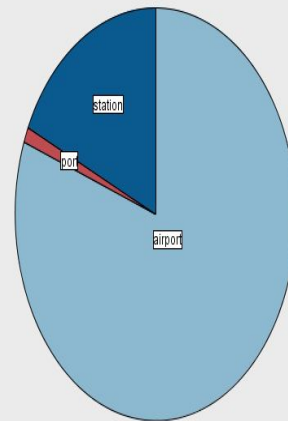


Iz kolona DST, Source, i Type se uklanjaju samo nepostojeće vrednosti.

*Zastupljenost areodroma, luka i železničkih stanica,  
pre preprocesiranja*



*Zastupljenost areodroma, luka i železničkih stanica,  
posle preprocesiranja*

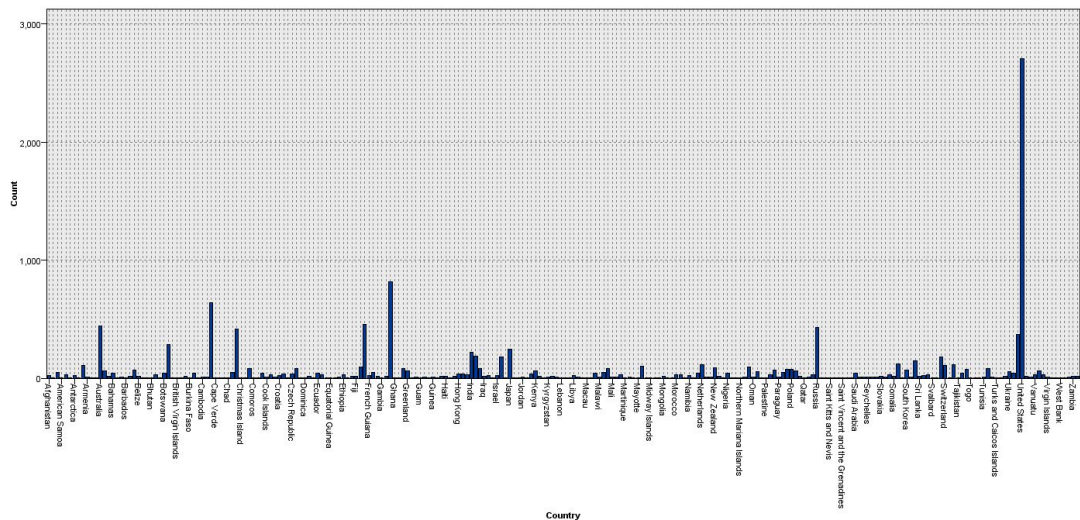


Kolone Country ima previše različitih vrednosti, napravljeno je polje koje ima vrednosti zemalja sa najvećim brojem areodroma, stanica , luka i vrednost Others.

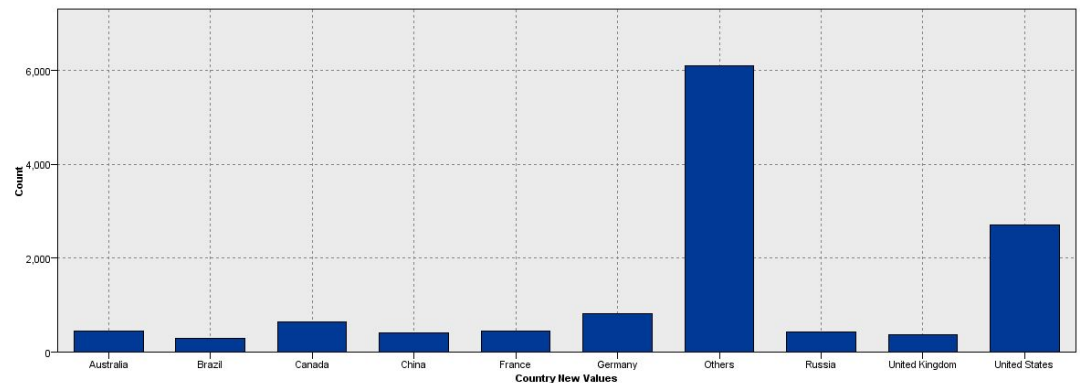




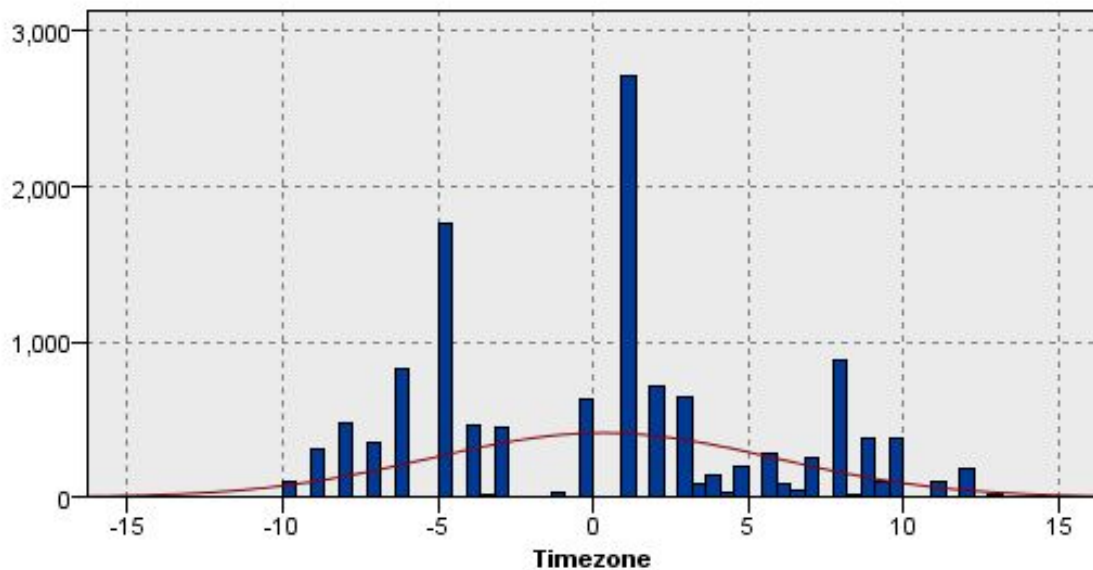
Kolona Country  
pre  
preprocesiranja



Kolona Country  
posle  
preprocesiranja  
(Country New  
Values)



—  
Kako je polje Timezone sadržalo nepostojeće vrednosti, na osnovu širine i dužine te vrednosti su izračunate.





# Klasterovanje

Ideja klasterovanja jeste da podelimo naše podatke u grupe (klaster) prema određenim zavisnostima. Prilikom analize podataka koristili su se algoritmi:

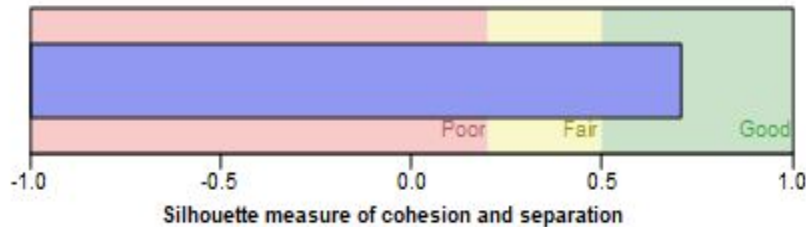
- **K-sredina**
- **Kohonenov**

# K-sredina

Model Summary

Algorithm	K-Means
Inputs	4
Clusters	10

Cluster Quality














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

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

## Clusters

Input (Predictor) Importance  
 1.0 0.8 0.6 0.4 0.2 0.0

Cluster	X=0, Y=0	X=3, Y=2	X=1, Y=2	X=0, Y=2	X=2, Y=0	X=3, Y=0	X=3, Y=1	X=2, Y=1	X=2, Y=2	X=1, Y=0	X=0, Y=1
Label											
Description											
Size	 24.2% (1317)	 23.7% (1291)	 19.4% (1057)	 14.1% (769)	 5.2% (281)	 4.5% (243)	 4.4% (238)	 2.4% (128)	 1.4% (74)	 0.8% (46)	 0.0% (1)
Inputs	Latitude 45.47	Latitude 51.09	Latitude 38.07	Latitude 47.23	Latitude -29.85	Latitude -23.76	Latitude 35.35	Latitude 8.53	Latitude 53.22	Latitude -35.47	Latitude 35.04
	Longitude -48.13	Longitude 10.02	Longitude -84.52	Longitude -123.12	Longitude 146.95	Longitude -48.95	Longitude 38.98	Longitude -41.33	Longitude -49.12	Longitude 149.56	Longitude -106.61
	Source User (100.0%)	Source OurAirports (100.0%)	Source OurAirports (100.0%)	Source OurAirports (100.0%)	Source OurAirports (100.0%)	Source OurAirports (100.0%)	Source OurAirports (100.0%)	Source OurAirports (100.0%)	Source OurAirports (100.0%)	Source User (100.0%)	Source Legacy (100.0%)
	Altitude 491.99	Altitude 528.26	Altitude 571.68	Altitude 1,807.78	Altitude 453.92	Altitude 1,398.53	Altitude 1,693.45	Altitude 402.49	Altitude 144.77	Altitude 379.78	Altitude 5,355.00

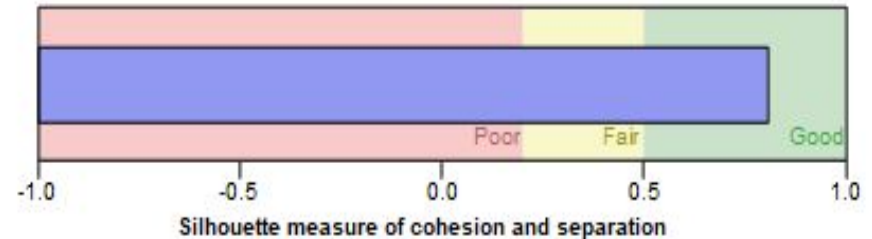
*Druga ideja* je bila da se izvrši klaster analiza samo za attribute Timezone i Longitude.

Najbolji rezultati se dobijaju kad imam 6 klastera, dok se za koeficijent senke dobija vrednost 0.8.

**Model Summary**

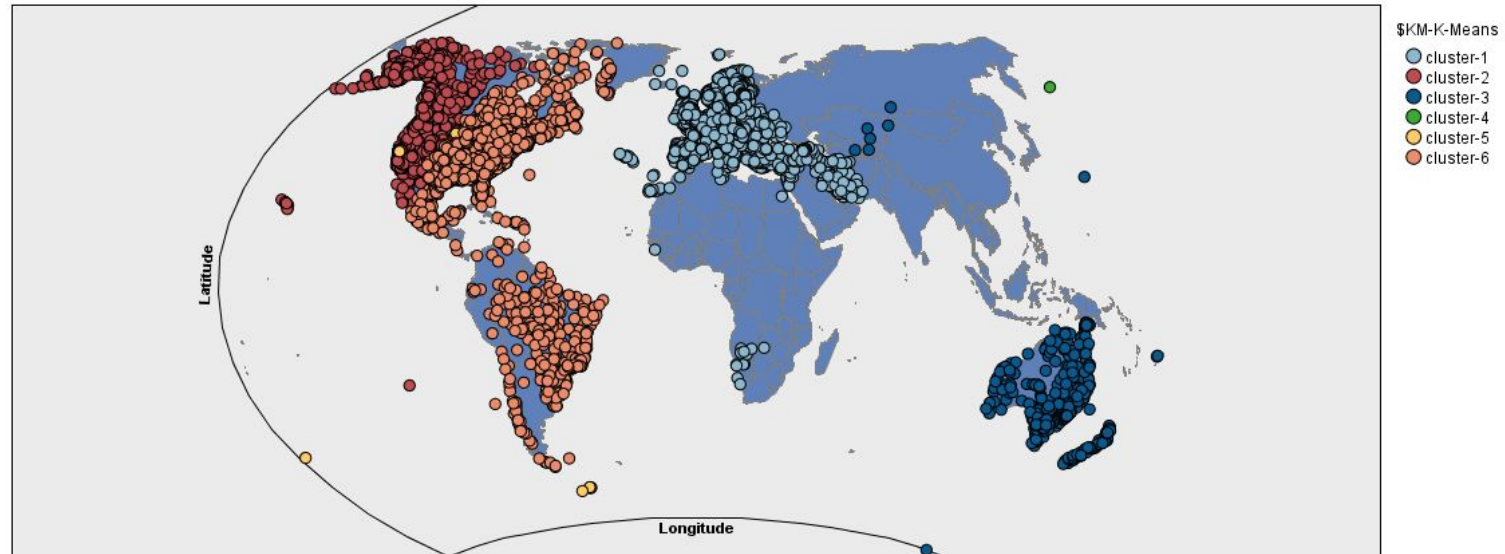
Algorithm	K-Means
Inputs	2
Clusters	6

**Cluster Quality**



—

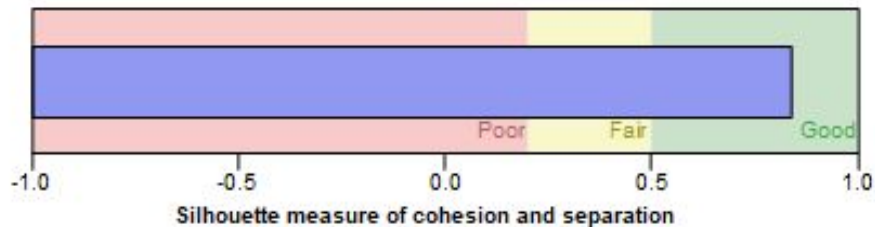
Prvi klaster se prostire kroz Evropu i Afriku. Drugi klaster je zauzeo predeo oko Aljaske i zapadnog dela Severne Amerike. Treći klaster se prostire kroz Australiju, Okeaniju i Novi Zeland. Četvrti klaster je zauzeo mali deo kod severne Azije. Peti klaster je zauzeo delove u tihom okeanu. Šesti klaster se protire kroz istočni deo Severne Amerike i kroz Južnu Ameriku.



### Model Summary

Algorithm	K-Means
Inputs	2
Clusters	3

### Cluster Quality



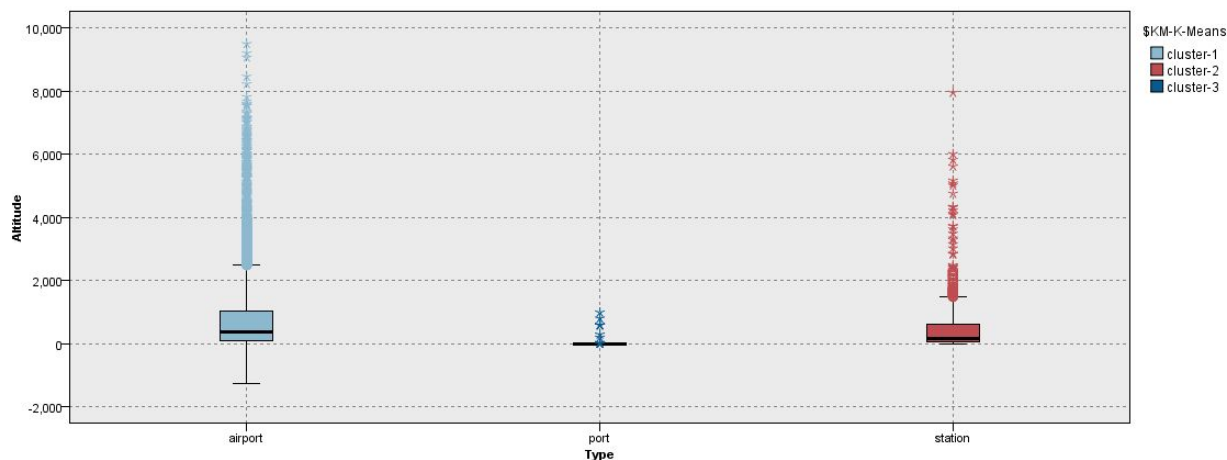
*Treća ideja* je bila da se izvrši klaster analiza samo za attribute Type i Altitude.

Najbolji rezultati se dobijaju kada se koristi 3 klastera, a za koeficijent senke dobijamo rezultata 0.8.



Rezultat koji se dobija je da sva tri klastera sadrže 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.



Četvrta ideja je bila da se izvrši klaster analiza samo za attribute Longitude i Latitude.

Najbolji rezultati se dobijaju kada se koriti 7 klastera, a za koeficijent senke dobijamo rezultata 0.7.

### Model Summary

Algorithm	K-Means
Inputs	2
Clusters	7

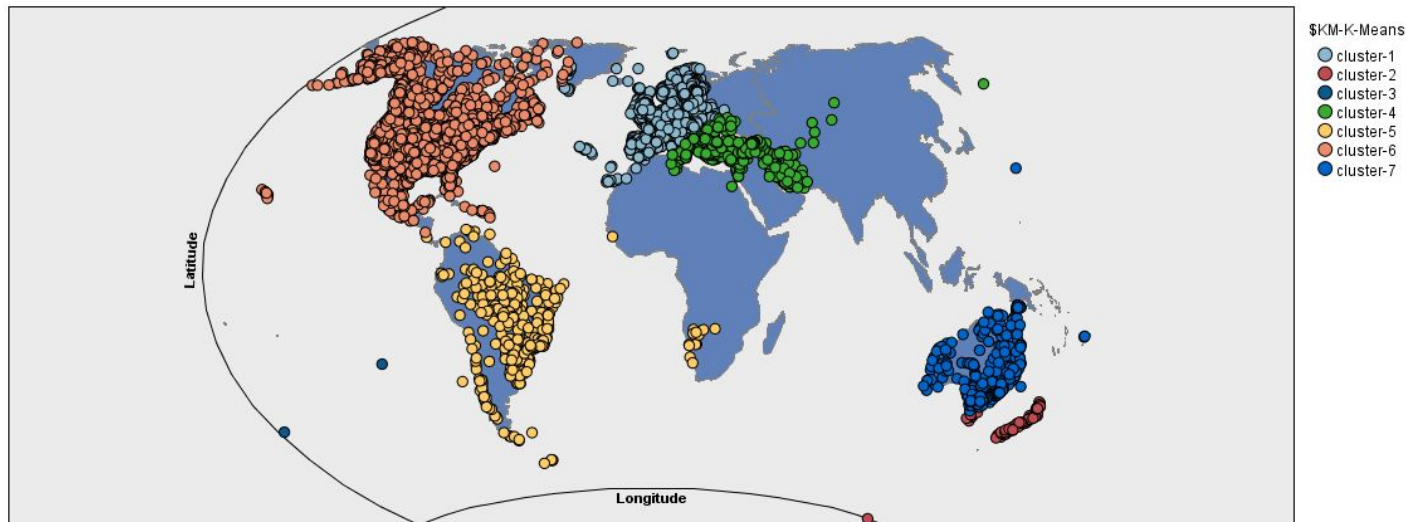
### Cluster Quality



—

Rezultati koji se dobiju su da svaki klaster zauzima po neki deo kontinenta.

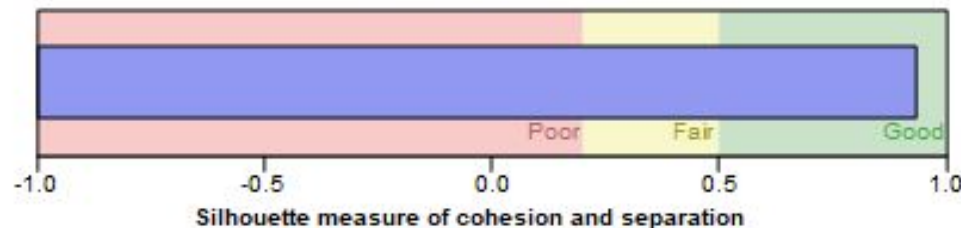
Prvi zauzima područje Evrope. Drugi zauzima prostor Novog Zelanda. Treći zauzima prostor Austrije. Četvrti zauzima deo prostora Azije, tj. Malu Aziju i bliski istok. Peti deo Afrike i Južnu Ameriku. Šesti zauzima prostor Severne Amerike. Sedmi obuhvata ostrva u Tihom okeanu.



### Model Summary

Algorithm	K-Means
Inputs	2
Clusters	10

### Cluster Quality

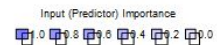


*Peta ideja* je bila da se izvrši klaster analiza samo za attribute Country New Values i Type.

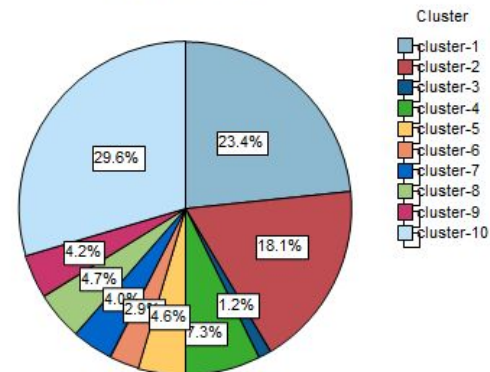
Najbolji rezultati se dobijaju kad imam 10 klastera, dok se za koeficijent senke dobija vrednost 0.9.

Clusters

Cluster	cluster-10	cluster-1	cluster-2	cluster-4	cluster-3	cluster-5	cluster-9	cluster-7	cluster-6	cluster-8
Label										
Description										
Size	<div><div></div><div>29.6% (1612)</div></div>	<div><div></div><div>23.4% (1273)</div></div>	<div><div></div><div>18.1% (985)</div></div>	<div><div></div><div>7.3% (398)</div></div>	<div><div></div><div>4.7% (257)</div></div>	<div><div></div><div>4.6% (248)</div></div>	<div><div></div><div>4.2% (230)</div></div>	<div><div></div><div>4.0% (216)</div></div>	<div><div></div><div>2.9% (160)</div></div>	<div><div></div><div>1.2% (66)</div></div>
Inputs	Country New Values United states (100.0%)	Country New Values Others (100.0%)	Country New Values United states (36.6%)	Country New Values Canada (100.0%)	Country New Values Brazil (100.0%)	Country New Values Germany (100.0%)	Country New Values Australia (100.0%)	Country New Values France (100.0%)	Country New Values United Kingdom (100.0%)	Country New Values Others (45.5%)
	Type airport (100.0%)	Type airport (100.0%)	Type station (100.0%)	Type airport (100.0%)	Type airport (100.0%)	Type airport (100.0%)	Type airport (100.0%)	Type airport (100.0%)	Type airport (100.0%)	Type port (100.0%)



Cluster Sizes



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

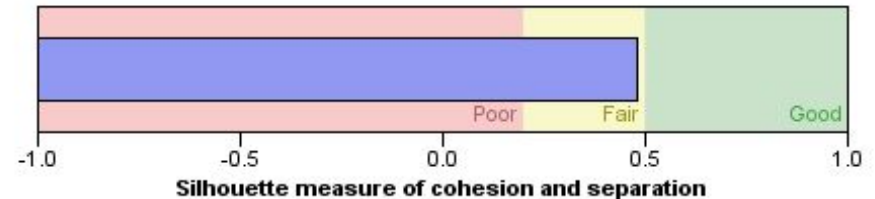
# Kohonen

*Prva ideja* je bila da se izvrši kohonen algoritam po podrazumevanim vrednostima, ali je klasterovanje dalo loše rezultate, tj. koeficijent senke je imao vrednost 0.5.

Model Summary

Algorithm	Kohonen
Inputs	4
Clusters	11

Cluster Quality



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

Prethodna klasterovanja nećemo dodatno objašnjavati jer su dala loše rezultate.

### Model Summary

Algorithm	Kohonen
Inputs	2
Clusters	12

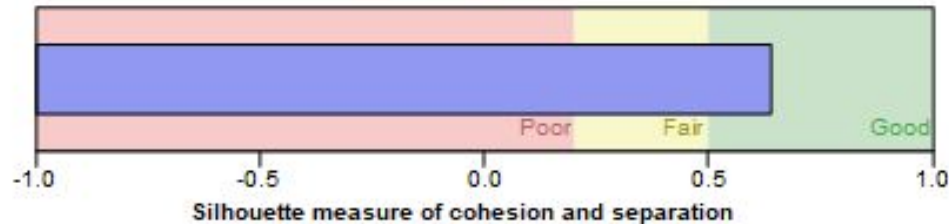
### Cluster Quality



### Model Summary

Algorithm	Kohonen
Inputs	2
Clusters	12

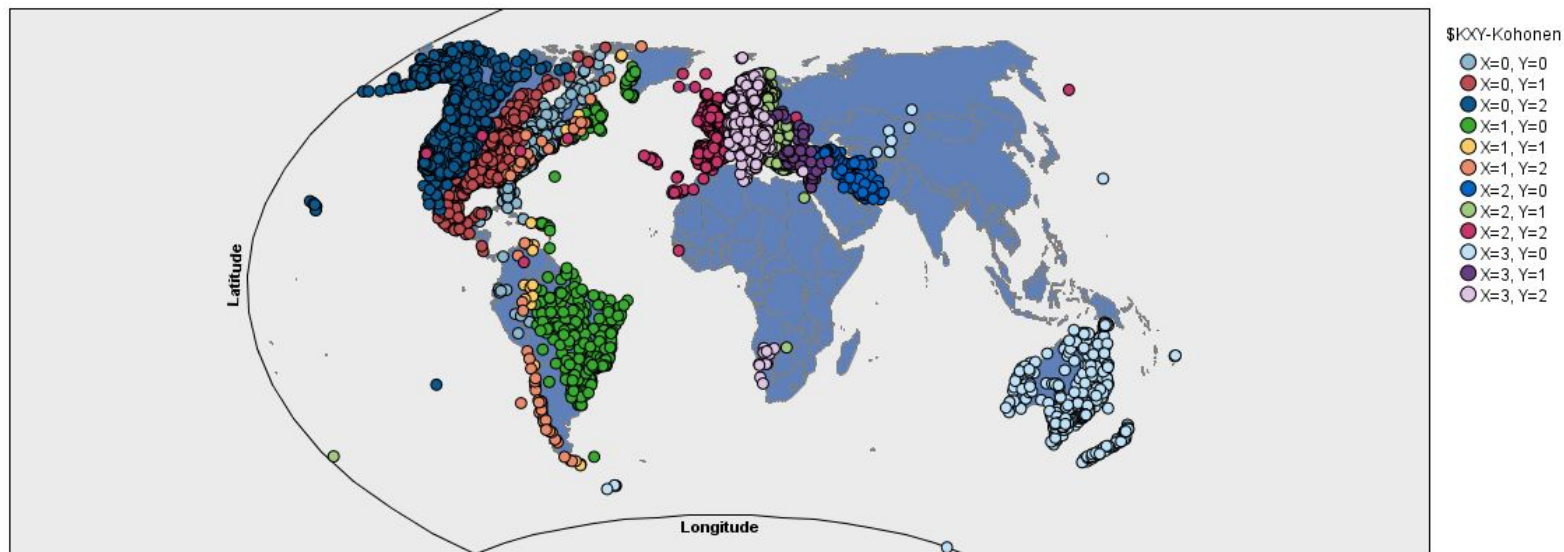
### Cluster Quality



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



Rezultat koji se dobija sličan je ko i kada se koristi k-sredina algoritam, s tim da ima više klastera pa su samim tim gušće raspoređeni. Klasteri se raspoređuju po medijanama, što je i za očekivati.



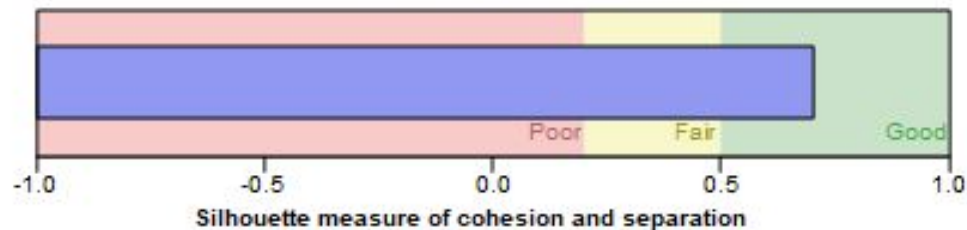
—

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

### Model Summary

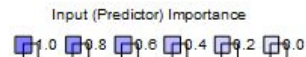
Algorithm	Kohonen
Inputs	2
Clusters	5

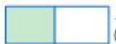




### Cluster Quality



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

### Clusters



Cluster	X=2, Y=2	X=0, Y=2	X=2, Y=1	X=2, Y=0	X=0, Y=0
Label					
Description					
Size	 46.9% (2556)	 18.1% (985)	 18.1% (985)	 15.7% (853)	 1.2% (66)
Inputs	Altitude 163.99	Altitude 474.71	Altitude 876.43	Altitude 2,960.31	Altitude 71.30
	Type airport (100.0%)	Type station (100.0%)	Type airport (100.0%)	Type airport (100.0%)	Type port (100.0%)

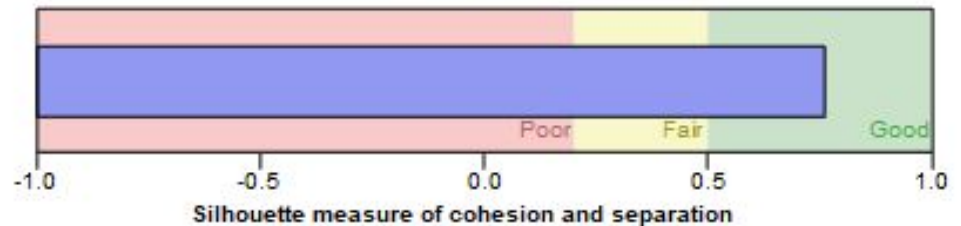
—

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

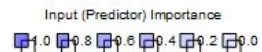
### Model Summary










Algorithm	Kohonen
Inputs	2
Clusters	9

### Cluster Quality



## Clusters



Cluster	X=0, Y=2	X=3, Y=2	X=3, Y=0	X=0, Y=0	X=1, Y=0	X=2, Y=2	X=2, Y=0	X=0, Y=1	X=1, Y=1
Label									
Description									
Size	 29.6% (1611)	 23.7% (1293)	 23.4% (1273)	 13.5% (735)	 4.7% (254)	 4.1% (225)	 0.6% (30)	 0.4% (21)	 0.1% (3)
Inputs	Country New Values United states (100.0%)	Country New Values Canada (30.8%)	Country New Values Others (100.0%)	Country New Values United states (49.1%)	Country New Values Others (98.4%)	Country New Values France (96.0%)	Country New Values Others (100.0%)	Country New Values United states (100.0%)	Country New Values France (100.0%)
	Type airport (100.0%)	Type airport (100.0%)	Type airport (100.0%)	Type station (100.0%)	Type station (98.4%)	Type airport (96.4%)	Type port (100.0%)	Type port (100.0%)	Type port (100.0%)