Wstęp

Celem projektu jest hierarchizacja oraz pogrupowanie krajów europy.

Dane:

- nation - kraj,

- region - region świata,

- fertility - współczynnik dzietności, liczba dzieci przypadających na kobietę,

- ppgdp - produkt krajowy brutto per capita w dolarach amerykańskich,

- lifeExpF - oczekiwana długość życia kobiet,

- pctUrban - procent ludności miejskiej,

- infantMortality - zgony niemowląt w wieku 1 roku na 1000 żywych urodzeń,

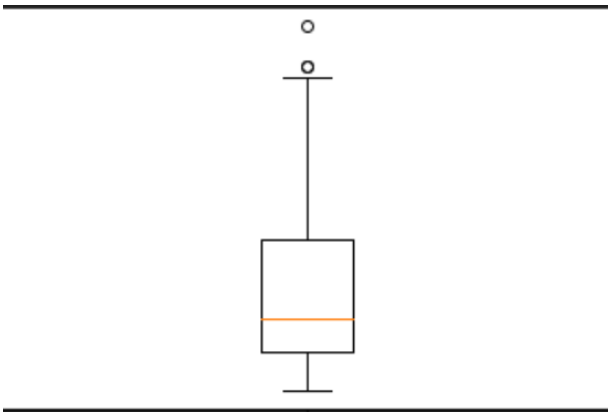
- income - zmienna binarna (1 - gospodarki o wysokich i średnio-wysokich dochodach, 0 - gospodarki o niskich i średnio-niskich dochodach).

Statystyki opisowe zmiennych

	fertility	ppgdp	lifeExpF	pctUrban	infantMortality	income
count	199.00	199.00	199.00	199.00	199.00	199.00
mean	2.7613	13011.9	72.2931	57.929648	31.2333	0.582915
std	1.3395	18412.4	10.1237	23.429565	29.0930	0.494321
min	1.1340	114.800	48.1100	11.000000	1.91600	0.00
25%	1.7535	1282.95	65.6600	39.000000	7.64800	0.00
50%	2.2620	4684.50	75.8900	59.000000	19.8480	1.00
75%	3.5445	15520.5	79.58500	75.000000	47.7010	1.00
max	6.9250	105095.4	87.12000	100.000000	124.5350	1.00
coefficient of variation cv	0.4851	1.415041	0.140038	0.404449	0.931474	0.84

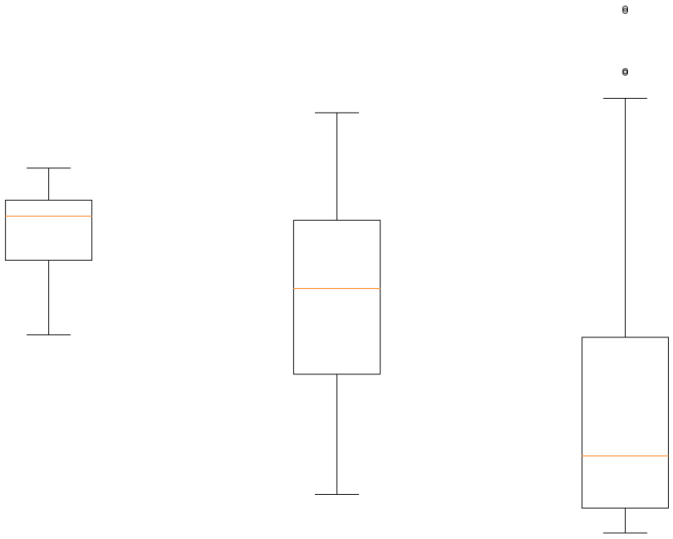
Zmienne ppgdp oraz infantMortality charakteryzują się dużą zmiennością co może zaburzyć wyniki poniższych metod.

Fertility



Zmienna fertility charakteryzuje się skośnością prawostronną I potencjalnie posiada outliery.

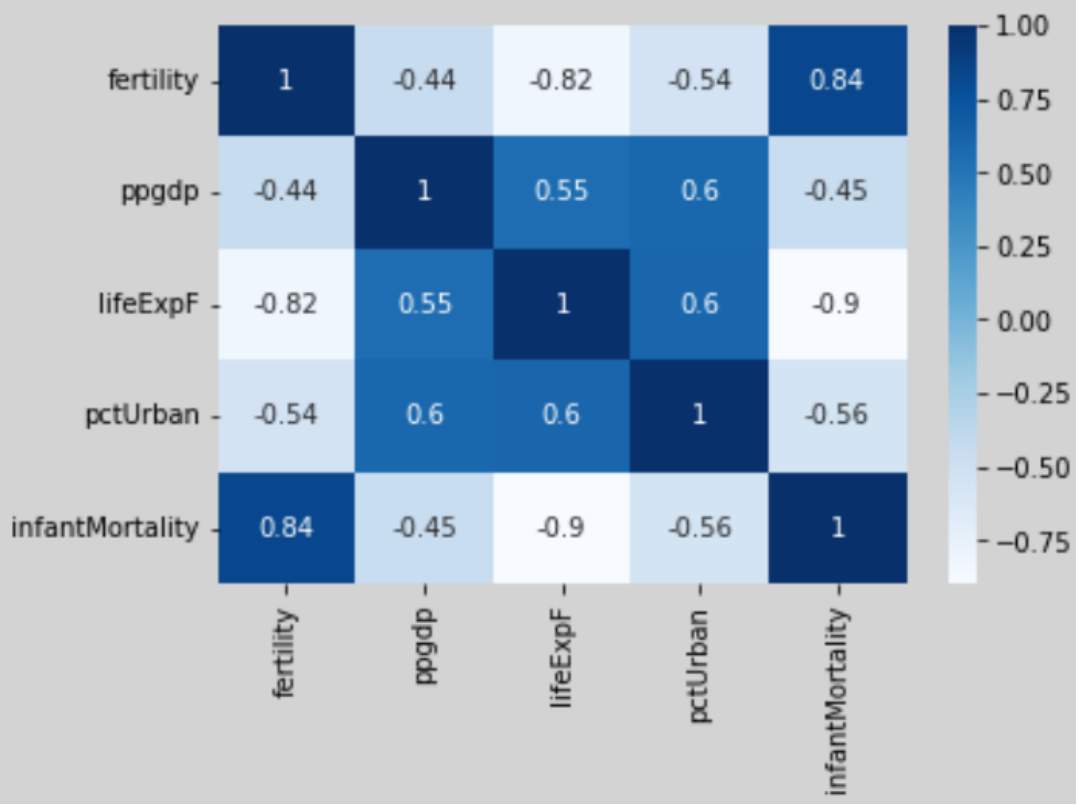
Zmienne: lifeExp, pctUrban, infantMort.



Zmienna lifeExp jest skośna lewostronnie a zmienna pctUrban jest relatywnie symetryczny co może sugerować rozkład normalny.

Zmienna dotycząca śmiertelności niemowląt jest skośna prawostronnie I może posiadać outliery.

Macierz korelacji



Najmocniej ze sobą są skorelowane zmienne: oczekiwana długość życia I dzietność oraz śmiertelność niemowląt oraz dzietność.

Porządkowanie liniowe

Pierwszą metodą jest porządkowanie liniowe tj. metoda hellwiga. Dane zostały zamienione na nominanty (przyjęto iż optymalną ilością urodzonych dzieci są 3) , stymulanty oraz destymulanty (infant mortality). Wyznaczono wzorzec jako wartości maksymalne każdego wymiaru oraz obliczono odległości euklidesowe które pozwoliły na utworzenie zmiennej syntetycznej.

Utworzony Ranking

Country	Zmienna Syntetyczna	
0	Norway	0.765283
1	Luxembourg	0.687974
2	Sweden	0.650148
3	Denmark	0.649164
4	Iceland	0.644582

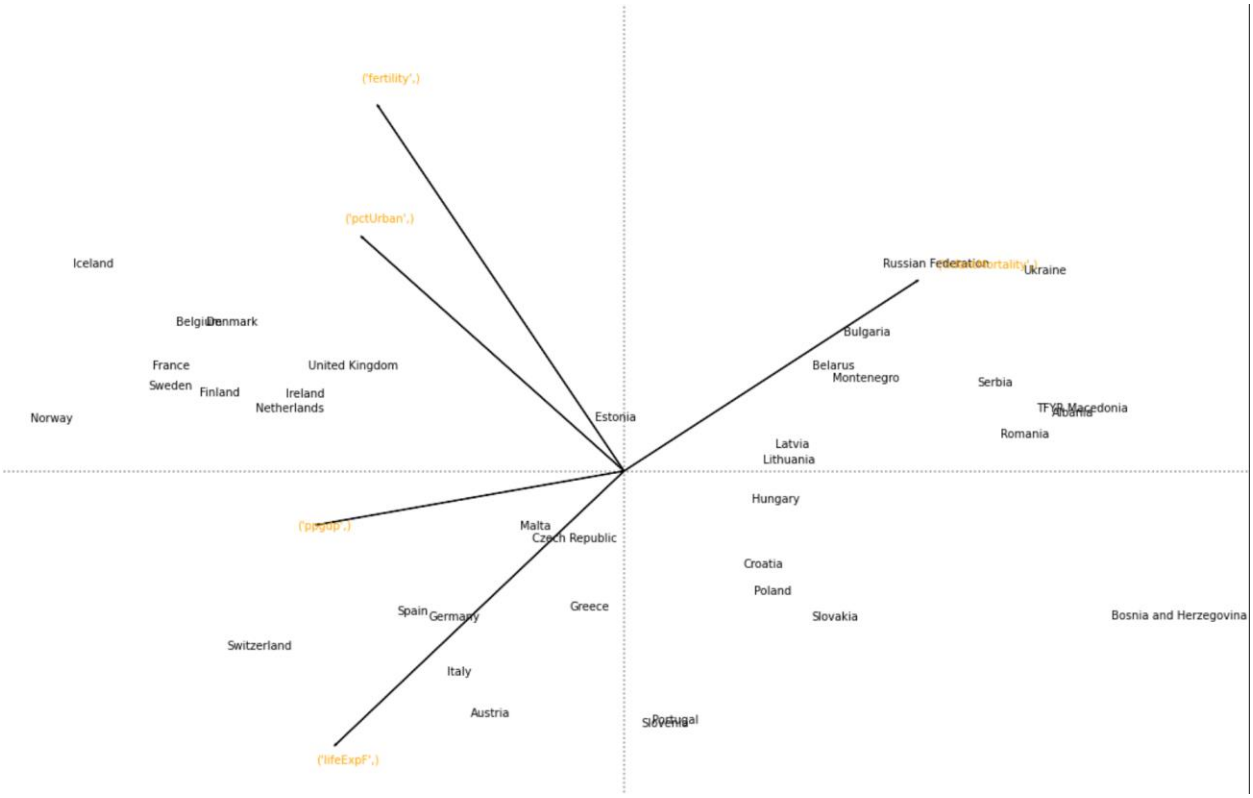
Country	Zmienna Syntetyczna	
5	France	0.624662
6	Finland	0.614804
7	Belgium	0.61405
8	Netherlands	0.591412
9	United Kingdom	0.553185
10	Switzerland	0.542134
11	Ireland	0.532364
12	Spain	0.429356
13	Germany	0.428499
14	Italy	0.398958
15	Austria	0.388905
16	Estonia	0.356036
17	Greece	0.351113
18	Czech Republic	0.350963
19	Malta	0.343139
20	Portugal	0.271972
21	Latvia	0.269634
22	Lithuania	0.261058
23	Croatia	0.259342
24	Hungary	0.259119
25	Bulgaria	0.254375
26	Slovenia	0.253343
27	Poland	0.249783
28	Montenegro	0.23551
29	Belarus	0.232145
30	Russian Federation	0.227335
31	Slovakia	0.212941
32	Albania	0.203408
33	Romania	0.191311
34	Serbia	0.188452
35	Ukraine	0.170723

Country	Zmienna Syntetyczna	
36	TFYR Macedonia	0.16636
37	Bosnia and Herzegovina	0.082776
38	Moldova	0.051873

Na podstawie danych najlepsze kraje to Norwegia, Luxemburg oraz Szwecja.

PCA

Biplot dla danych:



Z użyciem metody głównych składowych możliwe jest zobrazowanie odległości krajów od siebie względem kryteriów.

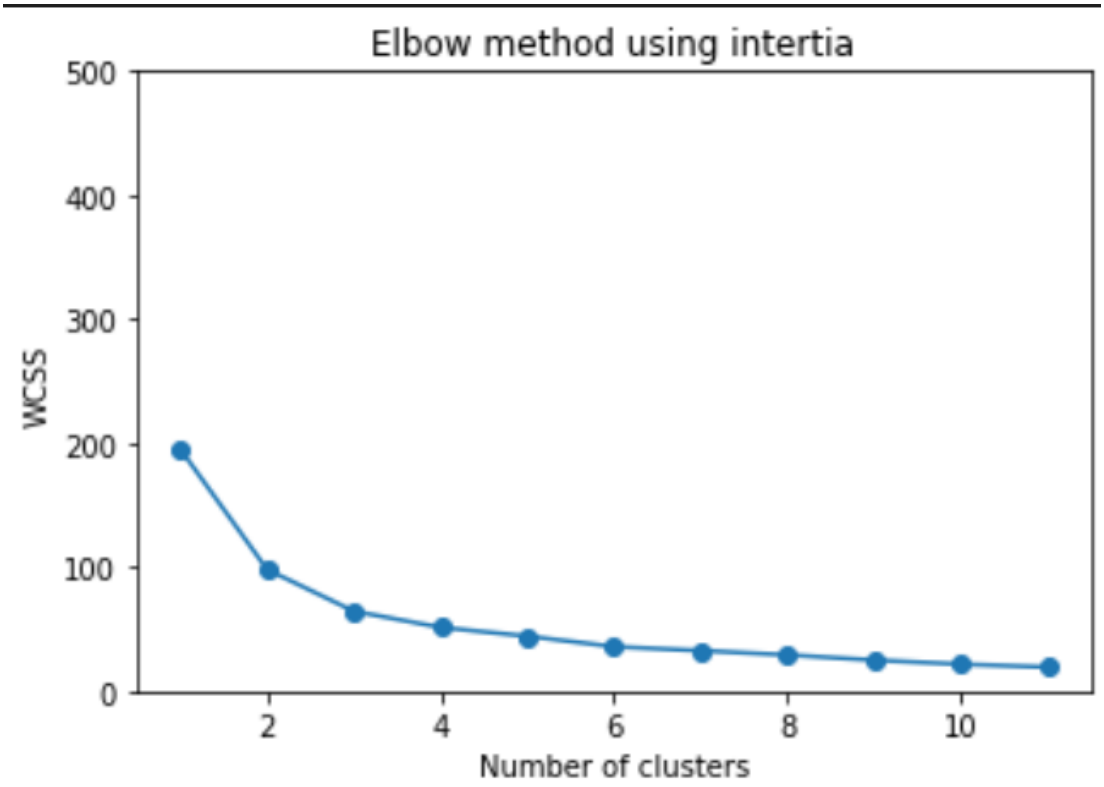
Można zauważyć, że zmienne lifeExpF oraz ppgdp trochę ze sobą korelują.

Kraje takie jak Islandia, Norwegia mają dużą populację żyjącą w miastach oraz duży współczynnik dzietności.

Rosja oraz Ukraina wykazują wysoką śmiertelność noworodków.

Klasteryzacja metodą k-średnich

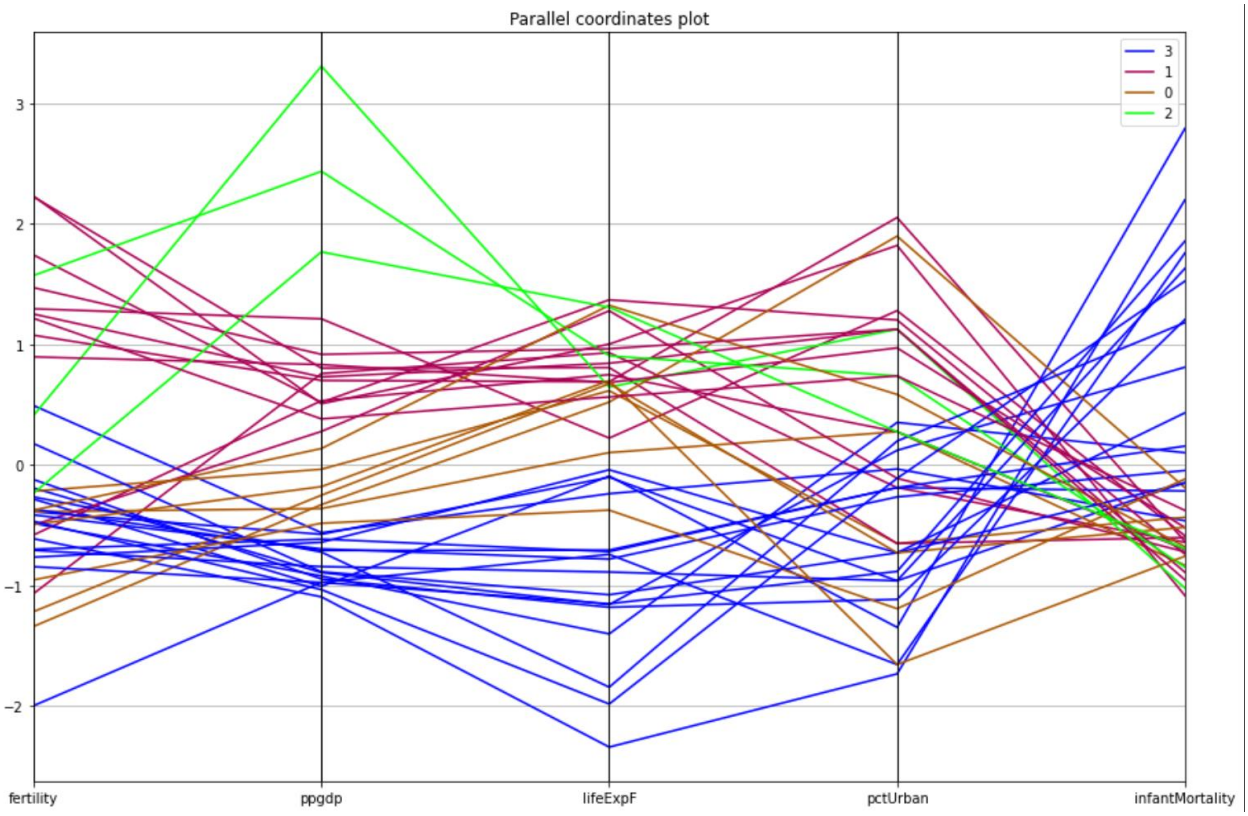
Metododą k-średnich wyznaczono 4 grupy krajów europy o podobnym łącznym wyniku. Do Znalezienia liczby klastrów wykorzystano “elbow method”. Na poziomie właśnie 4 klastrów suma kwadratów odległości obserwacji od wyznaczonych im centroid zaczyna liniowo maleć.



Wyniki Grupowania podziałowego

Grupa 1 (0)	Grupa 2 (1)	Grupa 3 (2)	Grupa 4 (3)
Lithuania	Czech Republic	Sweden	Ukraine
Latvia	Spain	Norway	Romania
Poland	Slovenia	Netherlands	Russian Federation
Hungary	Malta	United Kingdom	Serbia

Estonia	Germany	Luxembourg	Bosnia and Herzegovina
Croatia	Greece	Ireland	TFYR Macedonia
Slovakia	Switzerland	Iceland	Moldova
Belarus	Portugal	France	Albania
Bulgaria	Italy	Finland	
Montenegro	Austria	Denmark	
		Belgium	



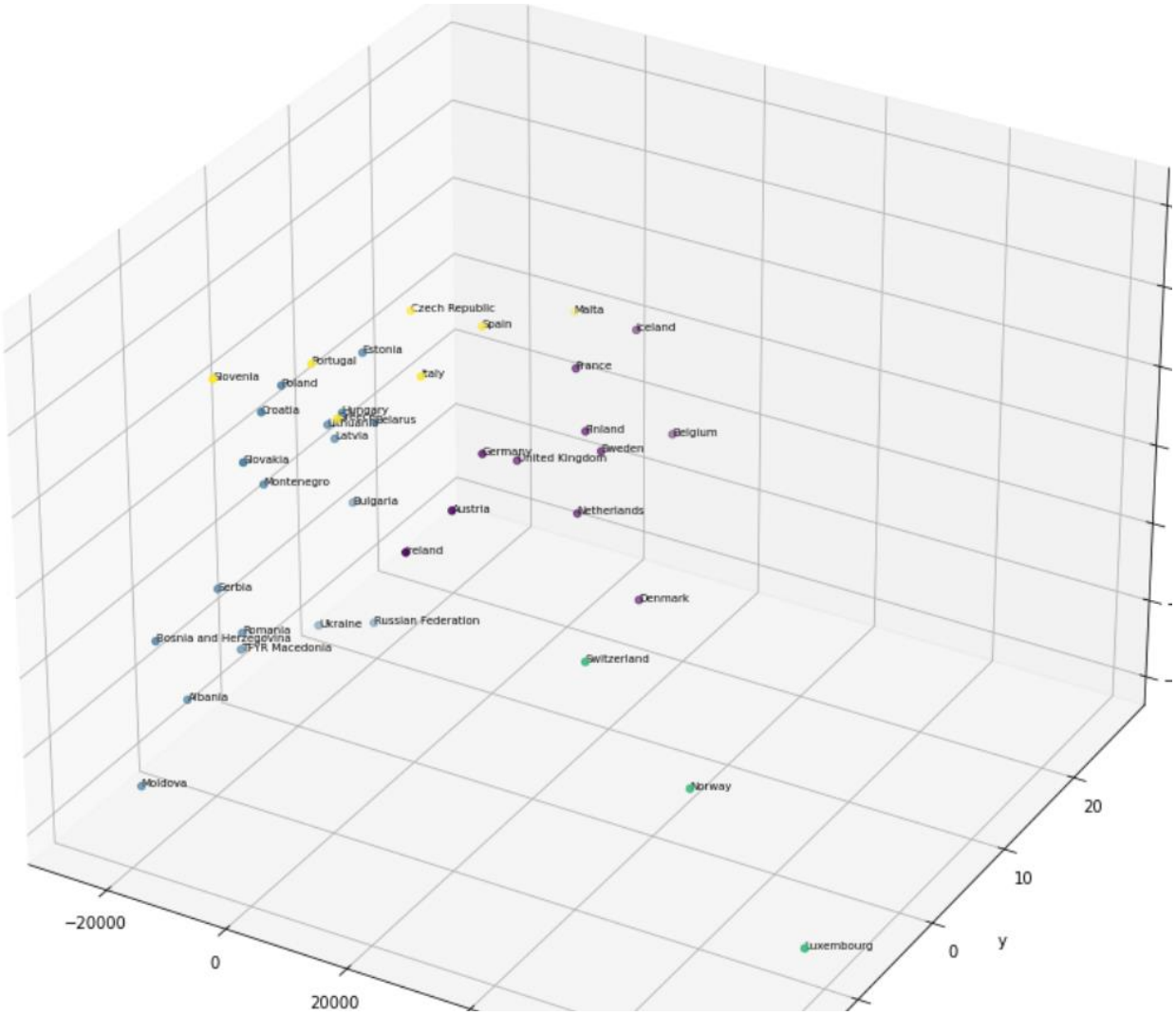
Za pomocą pararell coordinates plot można zinterpretować cechy charakterystyczne dla danych grup.

Można zauważyć, że w grupie pierwszej oraz 4 znajdują się głównie Kraje będące wcześniej częścią ZSRR. Grupa 1 (kolor brązowy) oraz 4 (kolor niebieski) charakteryzują się większymi niż reszta grup śmiertelnością noworodków oraz niskim procentem ludności miejskiej.

W grupie 3 (kolor zielony) znalazły się kraje o najwyższym pkb zaś w grupie 2 (kolor malinowy) te o dużej dzietności I procencie ludności żyjącej w miastach.

Kraje w tych samych grupach są gospodarczo do siebie podobnie I prawdopodobnie poziom życia w nich jest porównywalny.

Wizualizacja w 3 wymiarach


Czech Republic
Malta
Spain
Iceland
Estonia
Portugal
France
Slovenia
Italy
Poland
Croatia
Hungary
Belarus
Finland
Belgium
Latvia
Germany
Sweden
United Kingdom
Slovakia
Montenegro
Bulgaria
Austria
Netherlands
Ireland
Serbia
Denmark
Ukraine
Russian Federation
Romania
Bosnia and Herzegovina
FYR Macedonia
Switzerland
Albania
Moldova
Norway
Luxembourg
−20000
0
20000
0
10
20
y