LABORATOIRE HUBERT CURIEN

UMR • CNRS • 5516 • SAINT-ETIENNE

UNIVERSITÉ DE LYON

From Statistics to Data Mining

Master 1
COlour in Science and Industry (COSI)
Cyber-Physical Social System (CPS2)
Saint-Étienne, France

Fabrice MUHLENBACH
https://perso.univ-st-etienne.fr/muhlfabr/
e-mail: fabrice.muhlenbach@univ-st-etienne.fr

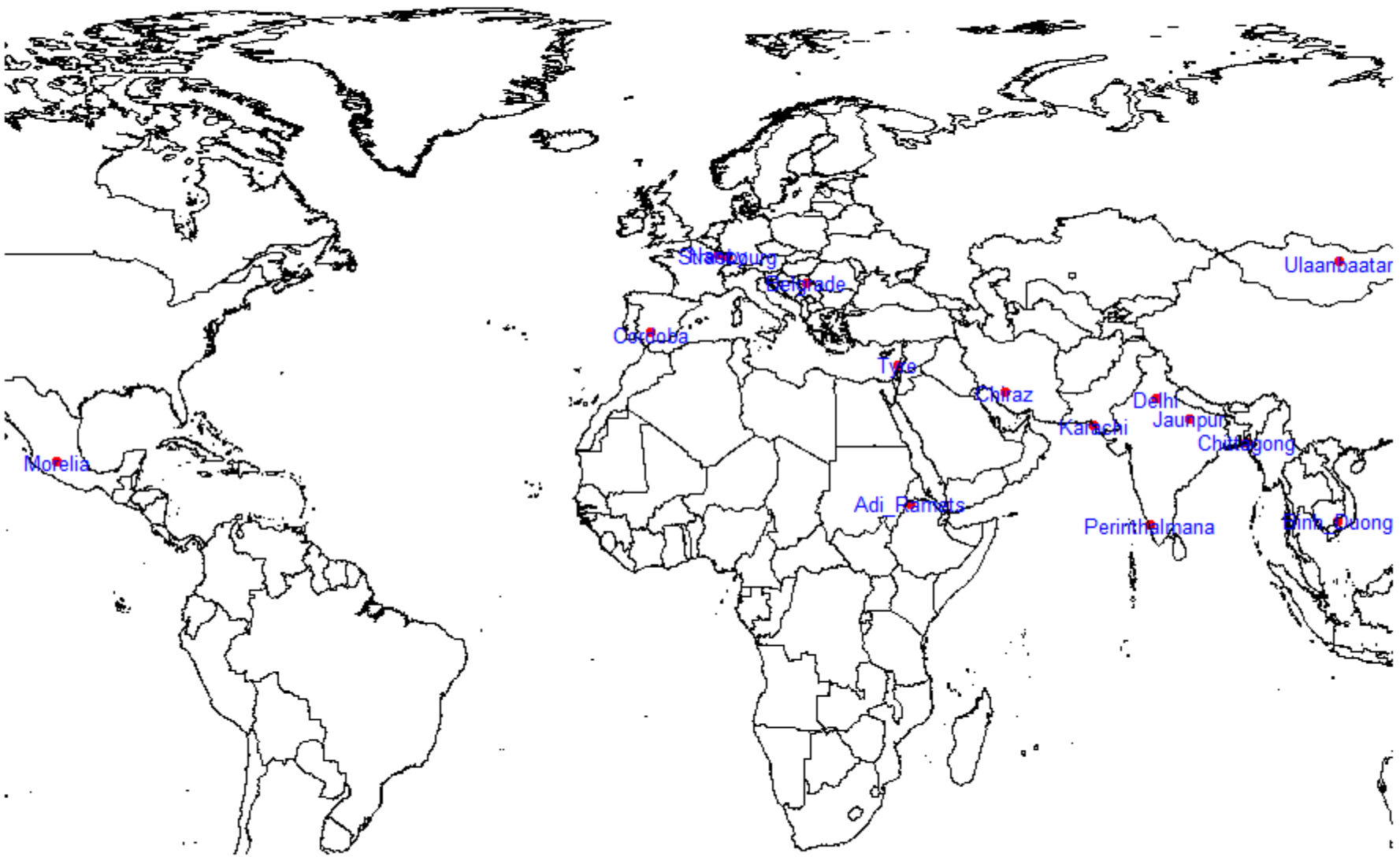
Tutorial

- Clustering

➢ consider the following cities with their spatial coordinates:

City	Latitude	Longitude
Adi_Ramets	13,75	37,32
Belgrade	44,79	20,46
Binh_Duong	11,33	106,48
Chiraz	29,59	52,58
Chittagong	22,34	91,83
Córdoba	37,89	-4,77
Delhi	28,70	77,10
Jaunpur	25,73	82,68
Karachi	24,86	67,01
Perinthalmana	10,98	76,23
Morelia	19,77	-101,19
Nancy	48,69	6,18
Strasbourg	48,58	7,75
Tyre	33,27	35,20
Ulaanbaatar	47,89	106,91

➢ the spatial coordinates correspond to points not on a plane but on a sphere


Ulaanbaatar
Belgrade
Cordoba
Tyre
Chiraz
Delhi
Jaunpur
Karachi
Chittagong
Morelia
Adi Ramats
Perinthalmana


Strasbourg
Belgrade
Ulaanbaatar
Cordoba
Tyre
Chiraz
Delhi
Jaunpur
Karachi
Chittagong
Morelia
Adi_Ramets
Perinthalmana
Binh_Duong

Tutorial

- Clustering

➢ import the cities coordinates in a spreadsheet

➢ compute the distance matrix between the cities

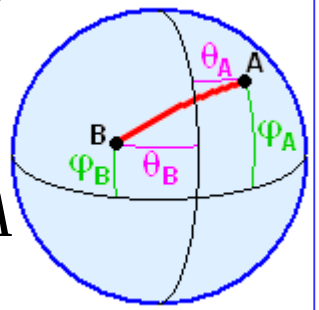
City	Adi_Ram	Belgrade	Binh_Du	Chiraz	Chittago	Córdoba	Delhi	Jaunpur	Karachi	Perintha	Morelia	Nancy	Strasbou	Tyre	Ulaanba
Adi_Ramets	0														
Belgrade		0													
Binh_Duong			0												
Chiraz				0											
Chittagong					0										
Córdoba						0									
Delhi							0								
Jaunpur								0							
Karachi									0						
Perinthalmana										0					
Morelia											0				
Nancy												0			
Strasbourg													0		
Tyre														0	
Ulaanbaatar															0

Tutorial

- Clustering

➢ What is the appropriate distance for this kind of problem?

→ the great-circle distance (or “orthodromic distance”)



If x_1 is the latitude and x_2 is the longitude

Let $\varphi_A = x_1(A)$ be the geographical latitude of point A

Let $\theta_A = x_2(A)$ be the geographical longitude of point A

Let $\varphi_B = x_1(B)$ be the geographical latitude of point B

Let $\theta_B = x_2(B)$ be the geographical longitude of point B

The shortest distance $d(A, B)$ between any two points on the surface of a sphere (the Earth) is called the orthodromic distance and is given (in kilometers) by:

$$d = 6378 \times \arccos\left(\sin(\varphi_A) \times \sin(\varphi_B) + \cos(\varphi_A) \times \cos(\varphi_B) \times \cos(\theta_A - \theta_B)\right)$$

where 6378 (km) is the radius for spherical Earth

Tutorial

- Clustering

➢ **City distance matrix:**

City	Adi_Ram	Belgrade	Binh_Du	Chiraz	Chittago	Córdoba	Delhi	Jaunpur	Karachi	Perintha	Morelia	Nancy	Strasbou	Tyre	Ulaanba
Adi_Ramets	0	3806	7494	2362	5819	4939	4425	4910	3346	4238	14157	4824	4732	2184	7370
Belgrade	3806	0	8821	3281	6844	2233	5244	5884	4715	6486	10735	1170	1057	1802	6288
Binh_Duong	7494	8821	0	5919	1981	11042	3613	2971	4420	3302	15435	9815	9707	7595	4070
Chiraz	2362	3281	5919	0	3991	5306	2381	2990	1521	3210	13864	4445	4330	1699	5015
Chittagong	5819	6844	1981	3991	0	9062	1639	1004	2544	2087	15141	7861	7751	5640	3145
Córdoba	4939	2233	11042	5306	9062	0	7444	8087	6810	8485	9221	1491	1560	3627	8340
Delhi	4425	5244	3613	2381	1639	7444	0	644	1090	1976	14638	6304	6191	4005	3331
Jaunpur	4910	5884	2971	2990	1004	8087	644	0	1579	1777	14955	6934	6822	4637	3250
Karachi	3346	4715	4420	1521	2544	6810	1090	1579	0	1826	14907	5852	5737	3220	4326
Perinthalmana	4238	6486	3302	3210	2087	8485	1976	1777	1826	0	16603	7641	7526	4859	5006
Morelia	14157	10735	15435	13864	15141	9221	14638	14955	14907	16603	0	9581	9689	12533	12001
Nancy	4824	1170	9815	4445	7861	1491	6304	6934	5852	7641	9581	0	115	2953	6863
Strasbourg	4732	1057	9707	4330	7751	1560	6191	6822	5737	7526	9689	115	0	2844	6784
Tyre	2184	1802	7595	1699	5640	3627	4005	4637	3220	4859	12533	2953	2844	0	6049
Ulaanbaatar	7370	6288	4070	5015	3145	8340	3331	3250	4326	5006	12001	6863	6784	6049	0

Tutorial

- Clustering

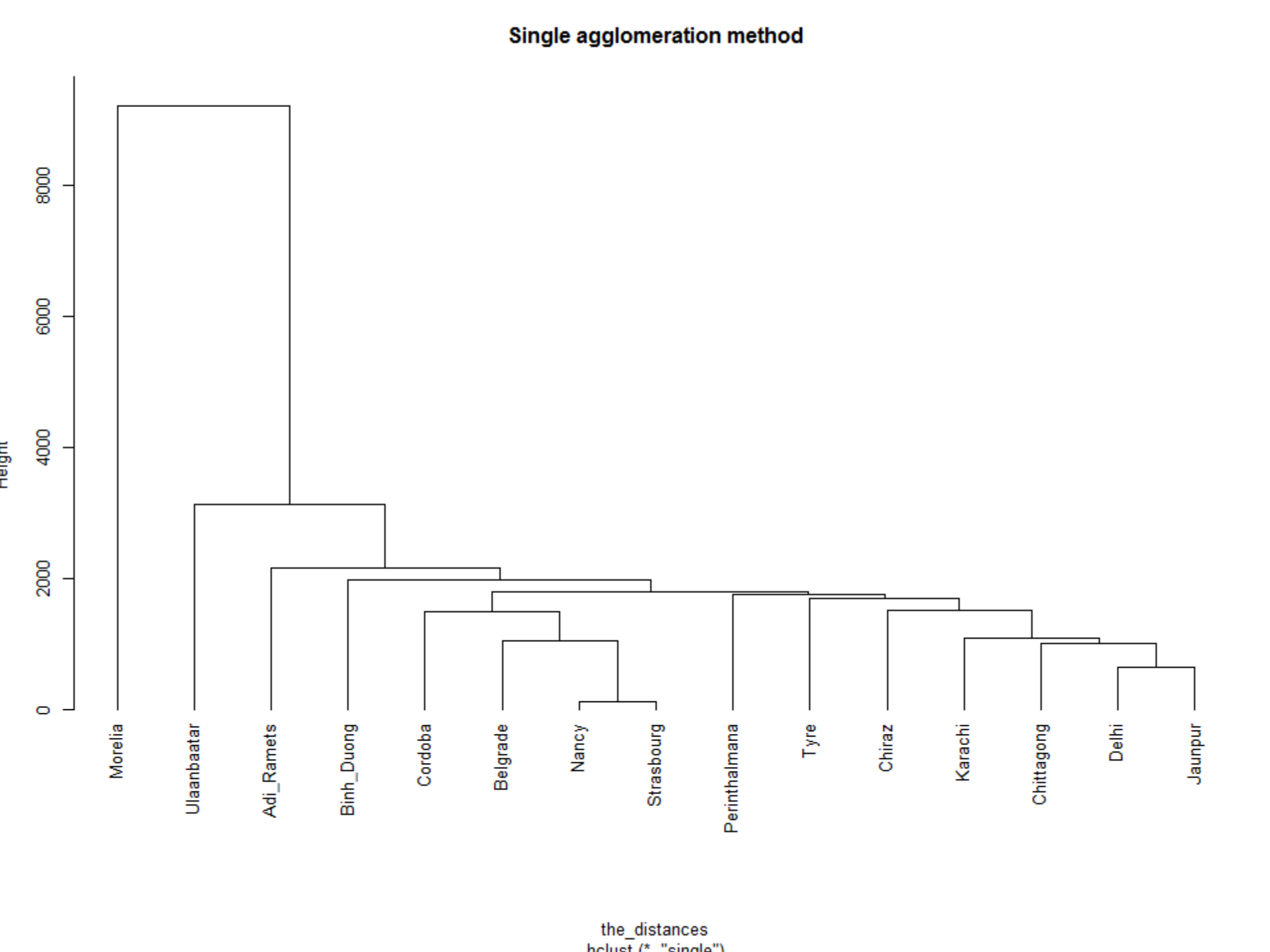
- **Questions:**
- run a hierarchical clustering algorithm (agglomerative)
- compute the distance matrix between the cities with the “single” agglomeration method
- plot the dendrogram associated with this clustering
- do the same with the “complete” agglomeration method

Tutorial

- Clustering

➢ **City distance matrix:**

City	Adi_Ram	Belgrade	Binh_Du	Chiraz	Chittago	Córdoba	Delhi	Jaunpur	Karachi	Perintha	Morelia	Nancy	Strasbou	Tyre	Ulaanba
Adi_Ramets	0	3806	7494	2362	5819	4939	4425	4910	3346	4238	14157	4824	4732	2184	7370
Belgrade	3806	0	8821	3281	6844	2233	5244	5884	4715	6486	10735	1170	1057	1802	6288
Binh_Duong	7494	8821	0	5919	1981	11042	3613	2971	4420	3302	15435	9815	9707	7595	4070
Chiraz	2362	3281	5919	0	3991	5306	2381	2990	1521	3210	13864	4445	4330	1699	5015
Chittagong	5819	6844	1981	3991	0	9062	1639	1004	2544	2087	15141	7861	7751	5640	3145
Córdoba	4939	2233	11042	5306	9062	0	7444	8087	6810	8485	9221	1491	1560	3627	8340
Delhi	4425	5244	3613	2381	1639	7444	0	644	1090	1976	14638	6304	6191	4005	3331
Jaunpur	4910	5884	2971	2990	1004	8087	644	0	1579	1777	14955	6934	6822	4637	3250
Karachi	3346	4715	4420	1521	2544	6810	1090	1579	0	1826	14907	5852	5737	3220	4326
Perinthalmana	4238	6486	3302	3210	2087	8485	1976	1777	1826	0	16603	7641	7526	4859	5006
Morelia	14157	10735	15435	13864	15141	9221	14638	14955	14907	16603	0	9581	9689	12533	12001
Nancy	4824	1170	9815	4445	7861	1491	6304	6934	5852	7641	9581	0	115	2953	6863
Strasbourg	4732	1057	9707	4330	7751	1560	6191	6822	5737	7526	9689	115	0	2844	6784
Tyre	2184	1802	7595	1699	5640	3627	4005	4637	3220	4859	12533	2953	2844	0	6049
Ulaanbaatar	7370	6288	4070	5015	3145	8340	3331	3250	4326	5006	12001	6863	6784	6049	0

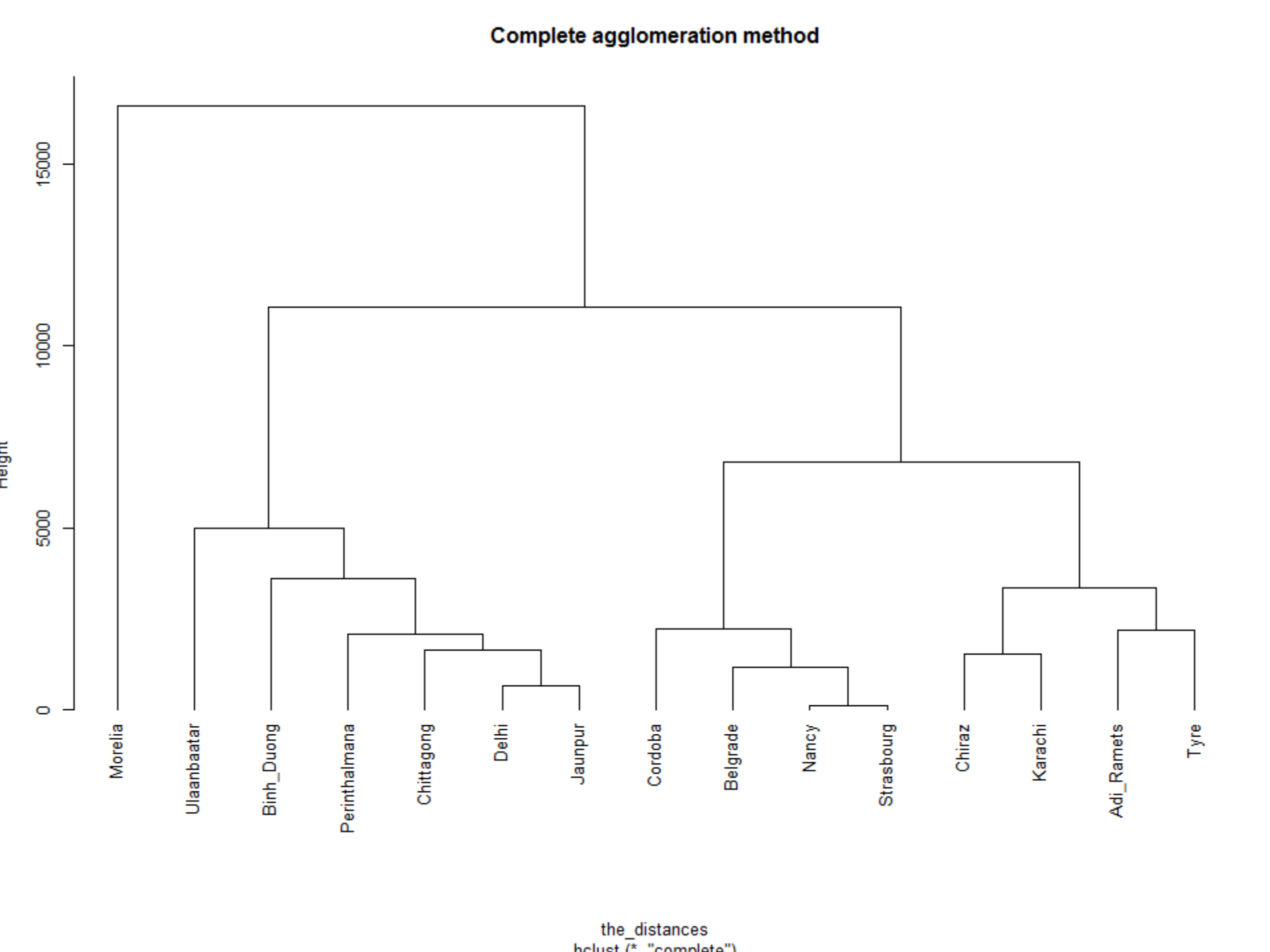

Single agglomeration method
Height
0
2000
4000
6000
8000
Morelia
Ulaanbaatar
Adi_Ramets
Binh_Duong
Cordoba
Belgrade
Nancy
Strasbourg
Perinthalmana
Tyre
Chiraz
Karachi
Chittagong
Delhi
Jaunpur
the_distances
hclust (*, "single")

Tutorial

- Clustering

➢ **City distance matrix:**

City	Adi_Ram	Belgrade	Binh_Du	Chiraz	Chittago	Córdoba	Delhi	Jaunpur	Karachi	Perintha	Morelia	Nancy	Strasbou	Tyre	Ulaanba
Adi_Ramets	0	3806	7494	2362	5819	4939	4425	4910	3346	4238	14157	4824	4732	2184	7370
Belgrade	3806	0	8821	3281	6844	2233	5244	5884	4715	6486	10735	1170	1057	1802	6288
Binh_Duong	7494	8821	0	5919	1981	11042	3613	2971	4420	3302	15435	9815	9707	7595	4070
Chiraz	2362	3281	5919	0	3991	5306	2381	2990	1521	3210	13864	4445	4330	1699	5015
Chittagong	5819	6844	1981	3991	0	9062	1639	1004	2544	2087	15141	7861	7751	5640	3145
Córdoba	4939	2233	11042	5306	9062	0	7444	8087	6810	8485	9221	1491	1560	3627	8340
Delhi	4425	5244	3613	2381	1639	7444	0	644	1090	1976	14638	6304	6191	4005	3331
Jaunpur	4910	5884	2971	2990	1004	8087	644	0	1579	1777	14955	6934	6822	4637	3250
Karachi	3346	4715	4420	1521	2544	6810	1090	1579	0	1826	14907	5852	5737	3220	4326
Perinthalmana	4238	6486	3302	3210	2087	8485	1976	1777	1826	0	16603	7641	7526	4859	5006
Morelia	14157	10735	15435	13864	15141	9221	14638	14955	14907	16603	0	9581	9689	12533	12001
Nancy	4824	1170	9815	4445	7861	1491	6304	6934	5852	7641	9581	0	115	2953	6863
Strasbourg	4732	1057	9707	4330	7751	1560	6191	6822	5737	7526	9689	115	0	2844	6784
Tyre	2184	1802	7595	1699	5640	3627	4005	4637	3220	4859	12533	2953	2844	0	6049
Ulaanbaatar	7370	6288	4070	5015	3145	8340	3331	3250	4326	5006	12001	6863	6784	6049	0


Complete agglomeration method
Height
0
5000
10000
15000
Morelia
Ulaanbaatar
Binh_Duong
Perinthalmana
Chittagong
Delhi
Jaunpur
Cordoba
Belgrade
Nancy
Strasbourg
Chiraz
Karachi
Adi_Ramets
Tyre
the_distances
hclust (*, "complete")