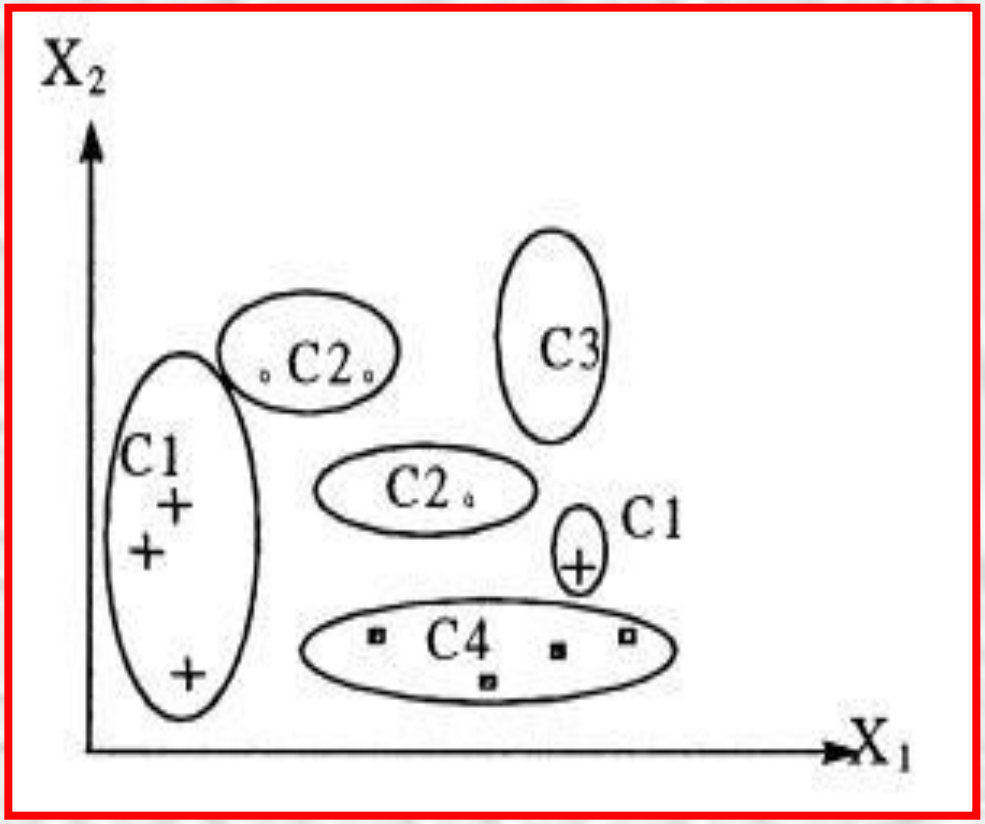
Starptautiskā zinātniska konference
VIDE. TEHNOLOĢIJA. RESURSI
Rēzeknes Augstskola, Latvija
2011.gada 20.-22. jūnijs

PĒTERIS GRABUSTS

THE CHOICE OF METRICS FOR CLUSTERING ALGORITHMS

Klasiskās klasterizācijas metodes



Algoritmi -

- K-Means Clustering -

FOREL u.c.

- **Mērķis -**
  Ieejas vektorus sadalīt klasēs (klasteros) un noteikt to centrus

- **Objektu grupas -**
  klasteri, klases, taksoni

K-Means Clustering algorithm

1 solis. Inicializē klasteru centrus w_j (j – nepieciešamo klasteru skaits uzdevuma risināšanai).

2.solis. Grupē visus apmācības izlases punktus ap tuvākā klastera centru t.i. katru punktu x_i saista ar klasteru j^*, kuram

$$\| x_i - w_{j^*} \| = \min_j \| x_i - w_j \|$$

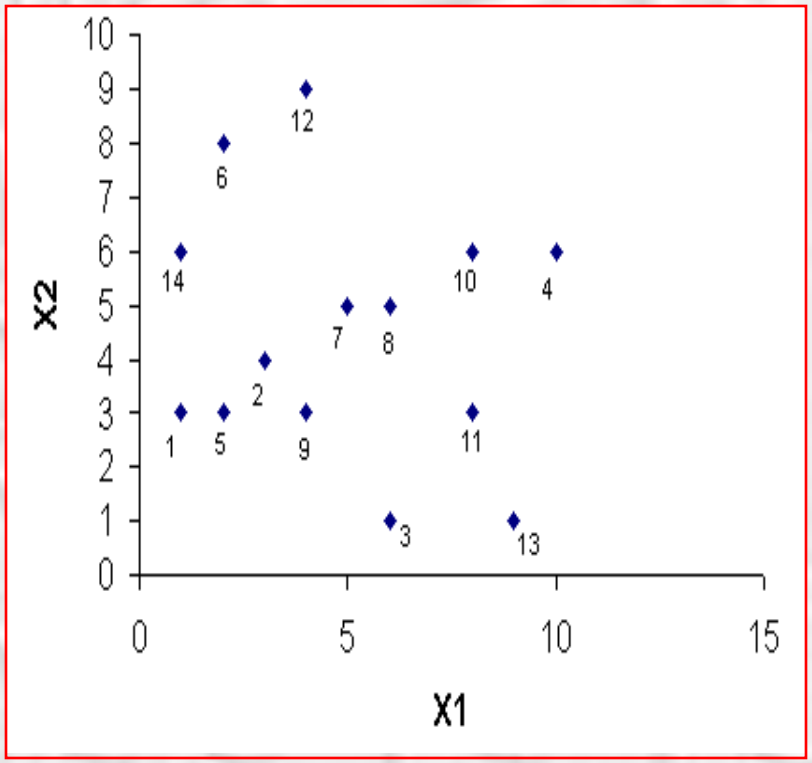
3.solis. Izskaitļo jaunos klasteru centrus, t.i. visiem w_j izskaitļo :

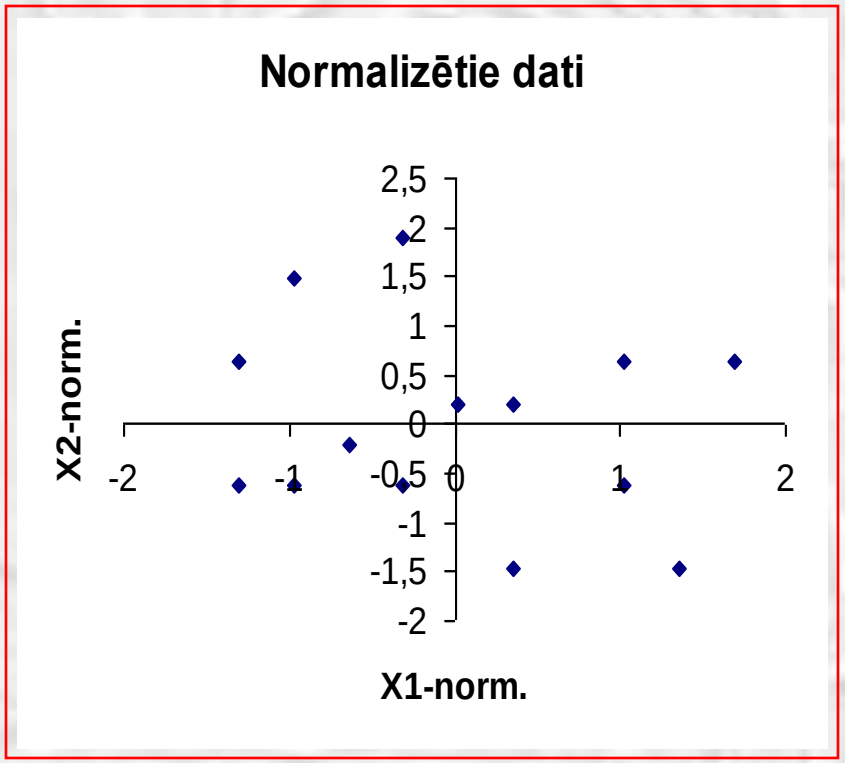
$$w_j = \frac{1}{m_j} \sum_{x_i \in klasteram\ j} x_i \quad ,$$

kur m_j – klasteram j piederošo punktu skaits.

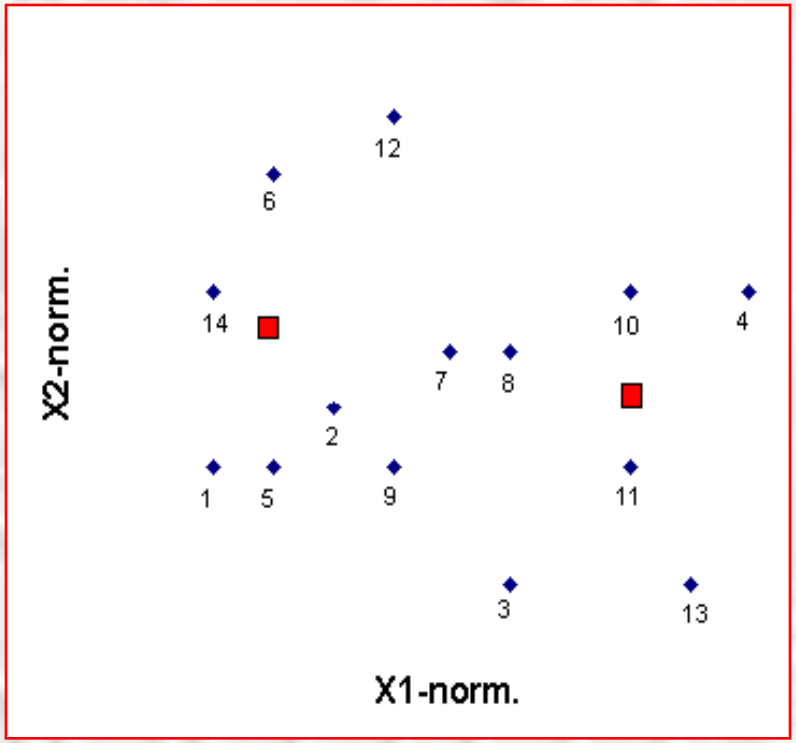
4 solis. Atkārtot soli 2 tik ilgi, kamēr iterāciju laikā nemainās klasteru centru vērtības.

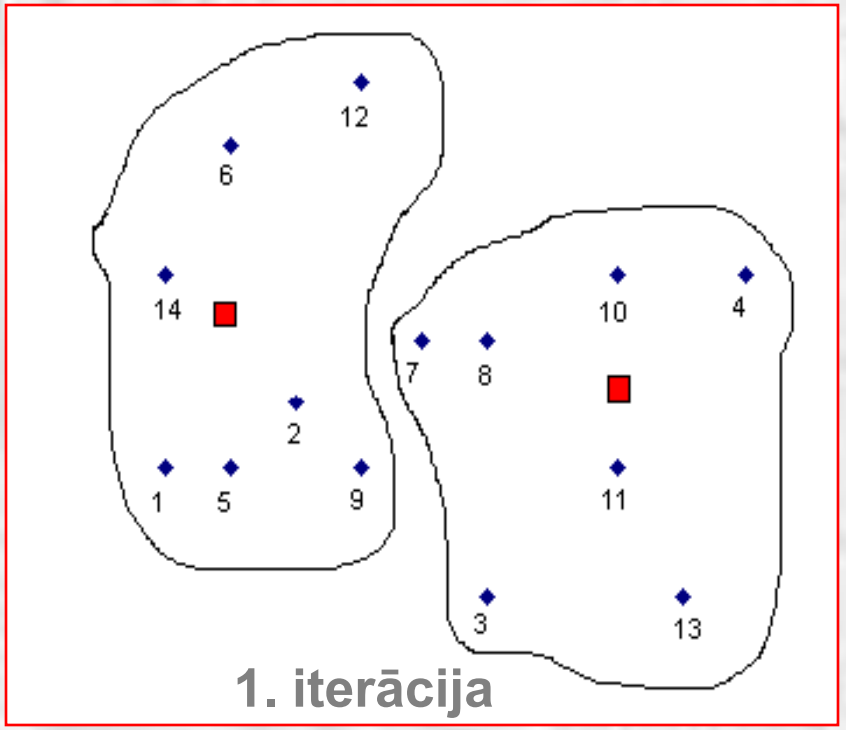
Klasterizācijas piemērs



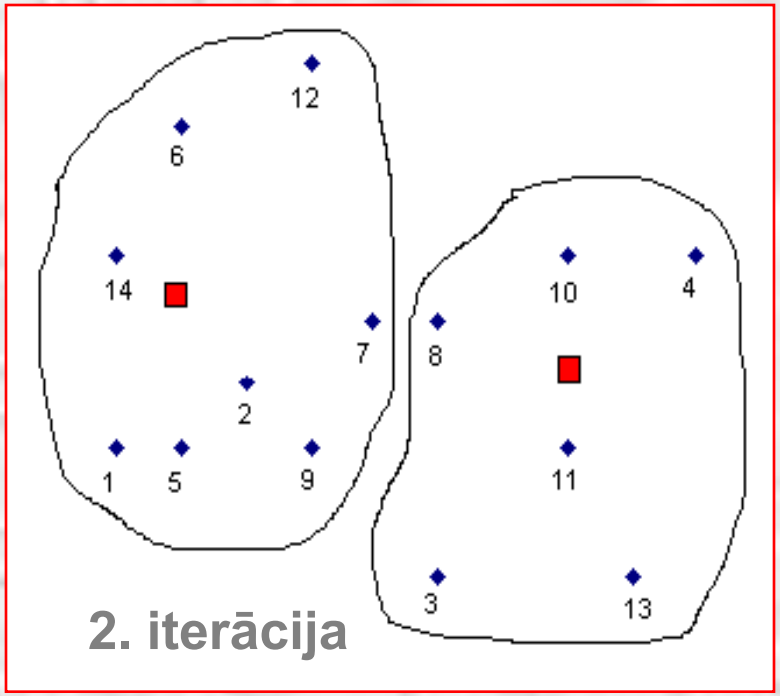


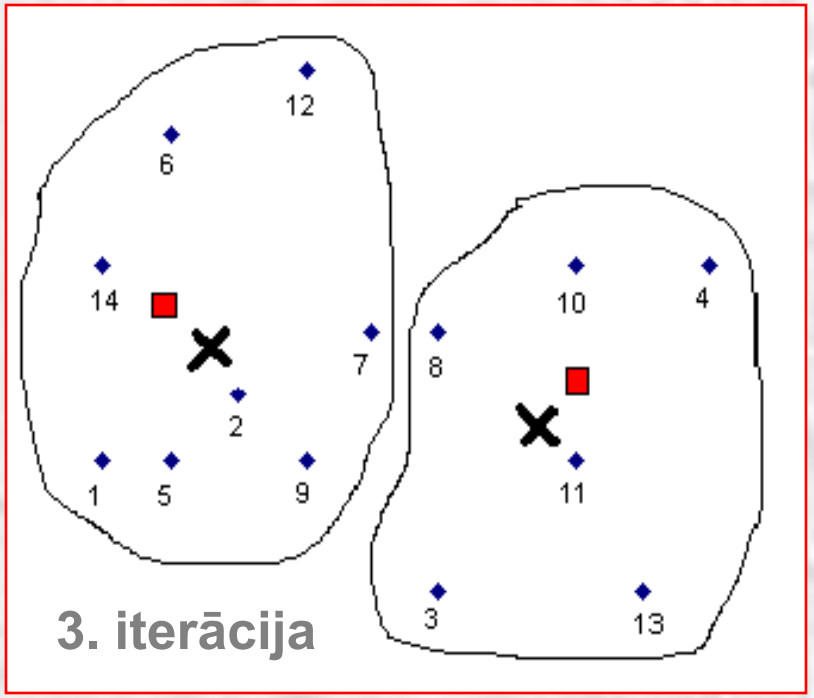
Klasterizācijas piemērs - (turp. 1)





Klasterizācijas piemērs (turp.2)





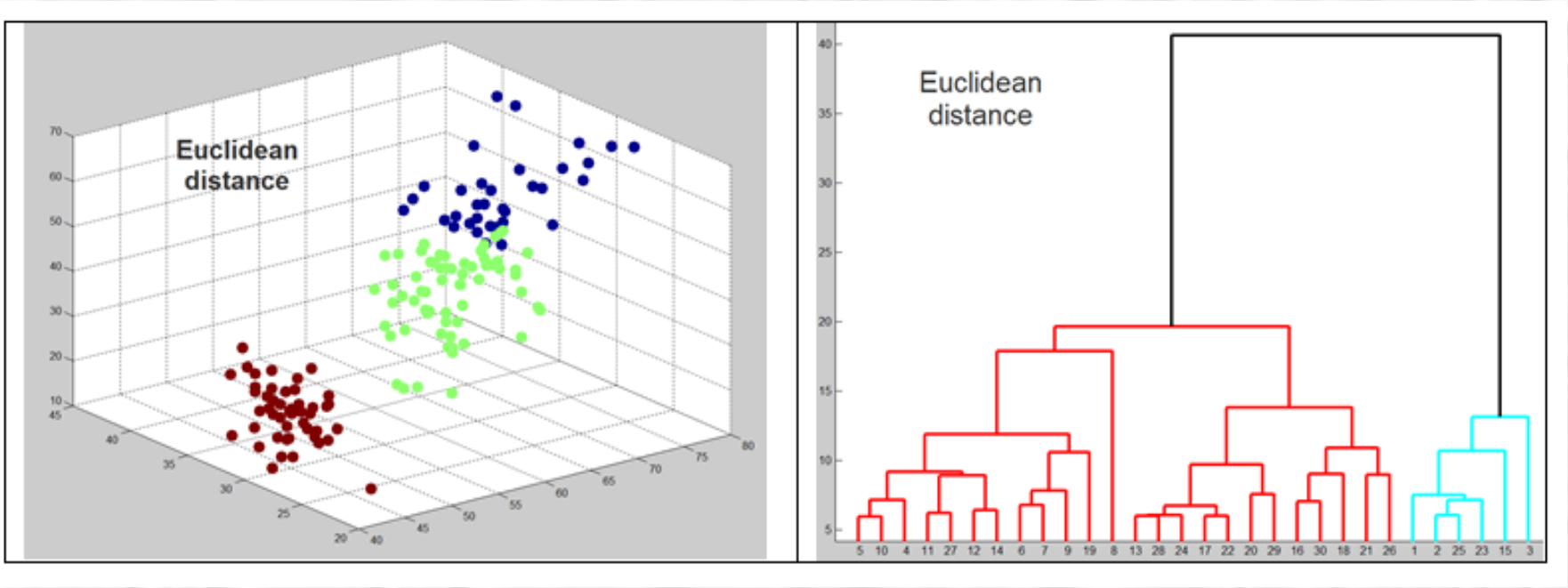
Distances metrika

Measure	Examples and applications
Euclidean distance	K-means with its variants
Manhattan distance	Fuzzy ART, clustering algorithms
Chebyshev distance	Fuzzy C-means clustering
Minkowski distance	Fuzzy C-means clustering
Pearson correlation	Widely used as the measure for microarray gene expression data analysis

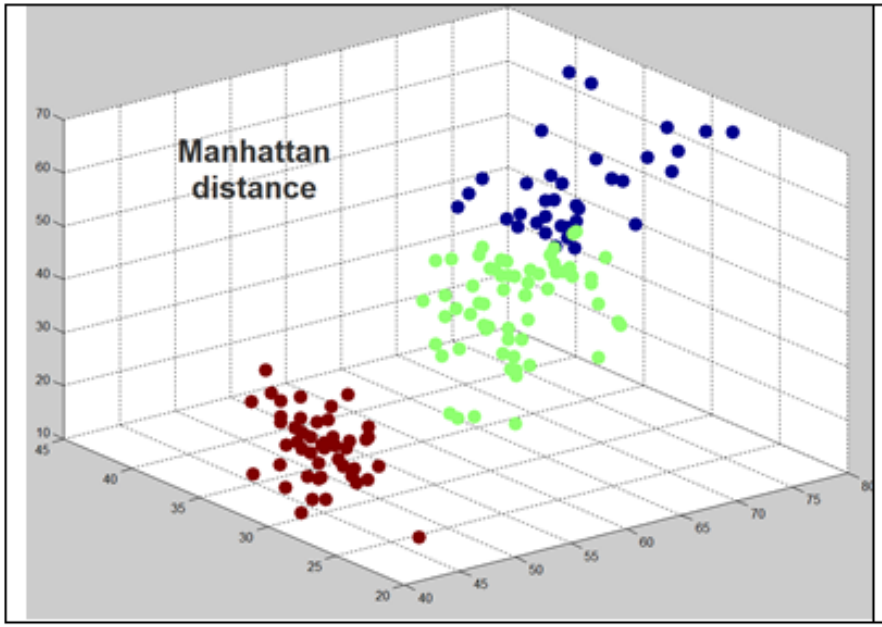
Eksperiments

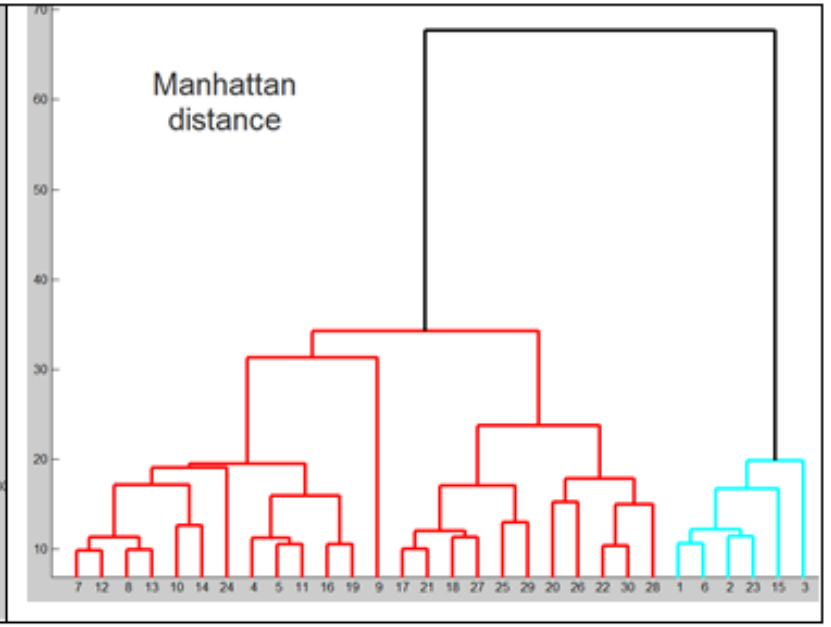
Distance	Euclidean	Manhattan	Correlation
Cluster centres	50.06 34.28 14.62 2.46 68.50 30.74 57.42 20.71 59.02 27.48 43.94 14.34	50 34 15 2 57 27 42 13 65 30 54 19	0.68 0.24 -0.29 -0.63 0.62 -0.35 0.34 -0.61 0.69 -0.23 0.20 -0.66
Cluster1 contains:	Records from cluster 1 – **50** Records from cluster 2 – 0 Records from cluster 3 – 0	Records from cluster 1 – **50** Records from cluster 2 – 0 Records from cluster 3 – 0	Records from cluster 1 – **50** Records from cluster 2 – 0 Records from cluster 3 – 0
Cluster2 contains:	Records from cluster 1 – 0 Records from cluster 2 – **48** Records from cluster 3 – **2**	Records from cluster 1 – 0 Records from cluster 2 – **39** Records from cluster 3 – **11**	Records from cluster 1 – 0 Records from cluster 2 – **47** Records from cluster 3 – **3**
Cluster3 contains:	Records from cluster 1 – 0 Records from cluster 2 – **14** Records from cluster 3 – **36**	Records from cluster 1 – 0 Records from cluster 2 – **4** Records from cluster 3 – **46**	Records from cluster 1 – 0 Records from cluster 2 – **3** Records from cluster 3 – **47**
Correct-ness:	For cluster 1 – **100 %** For cluster 2 - **96 %** For cluster 3 - **72 %**	For cluster 1 – **100 %** For cluster 2 - **78 %** For cluster 3 - **92 %**	For cluster 1 - **100 %** For cluster 2 - **94 %** For cluster 3 - **94 %**

Eiklīda distance

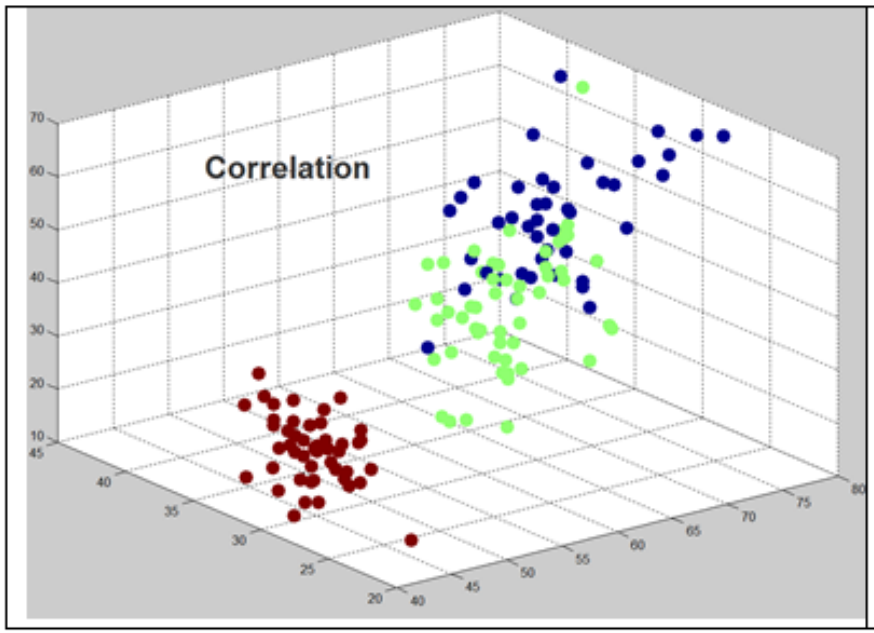


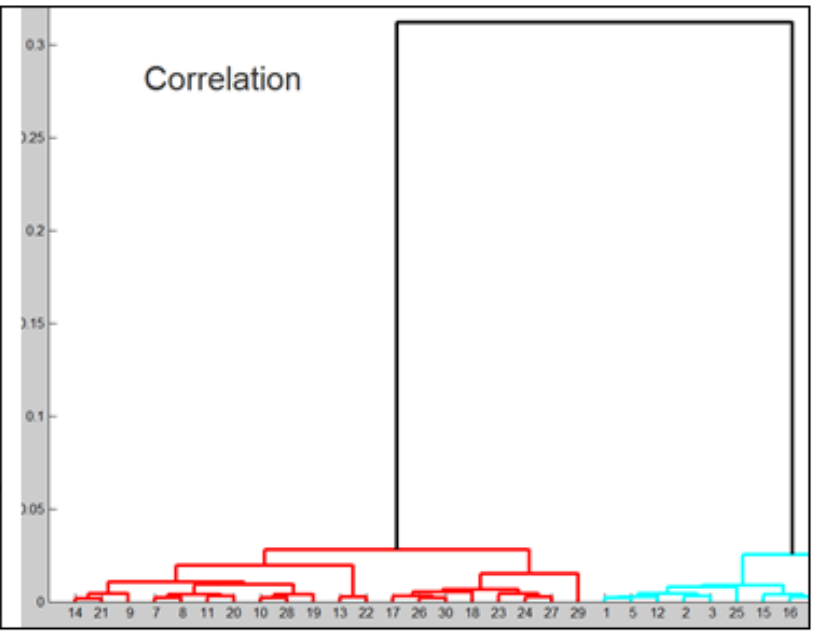
Manhetenas distance



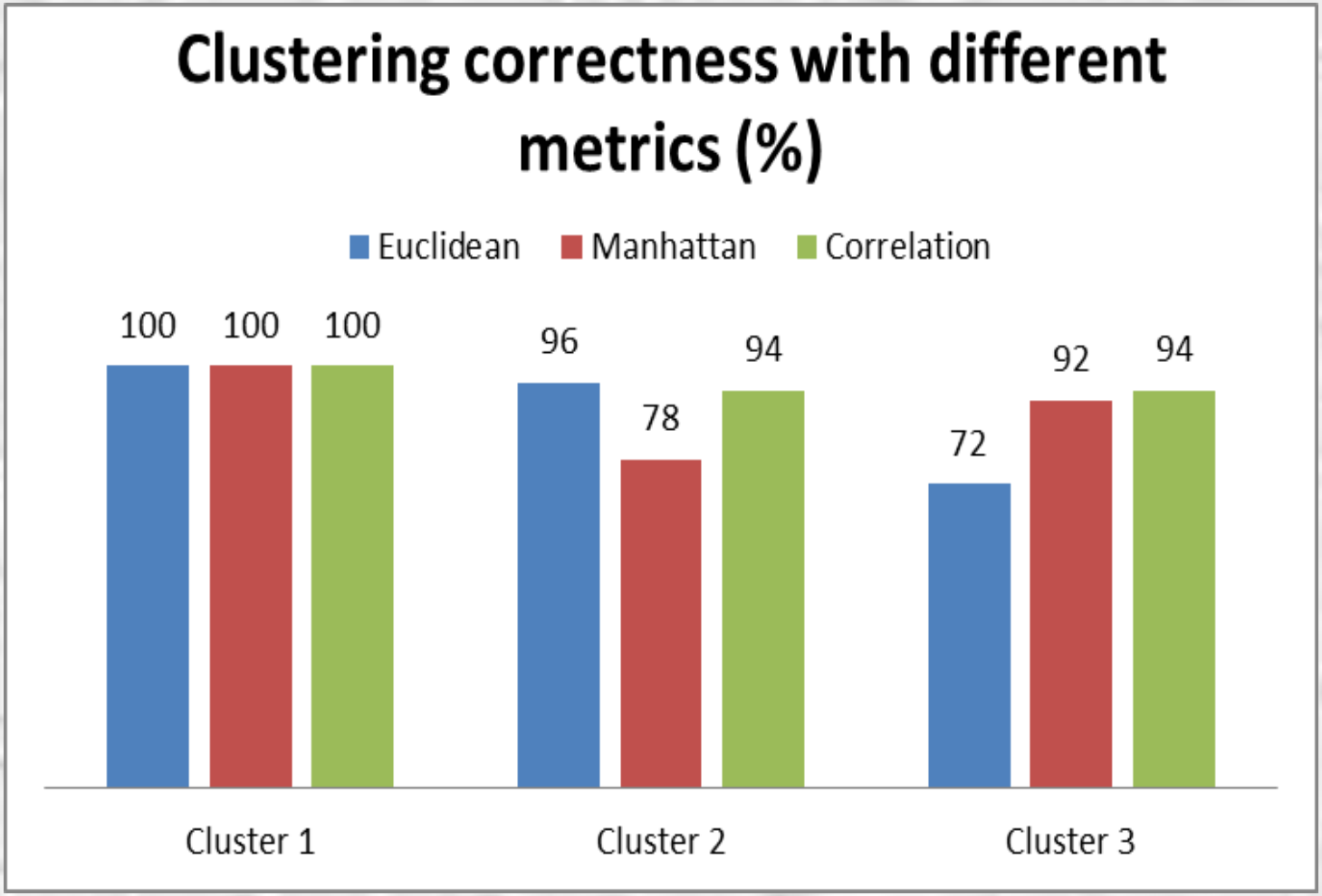


Korelācijas koeficients





Eksperiments



Secinājumi

- **Klasterizācijas algoritmu popularitāte, efektivitāte, procedūras vienkāršība**
- **Problēmas ar klasteru centru sākumvērtību noteikšanu**

- **Klasterizācijas skaitliskie rezultāti var būt diskutējami un parasti tie rūpīgi jāanalizē.**

PALDIES PAR UZMANĪBU !