2019 esri User Conference

July 8-12, 2019 | San Diego | California

Zonation of groundwater quality using spatially constrained multivariate clustering

Fahad Alahmadi

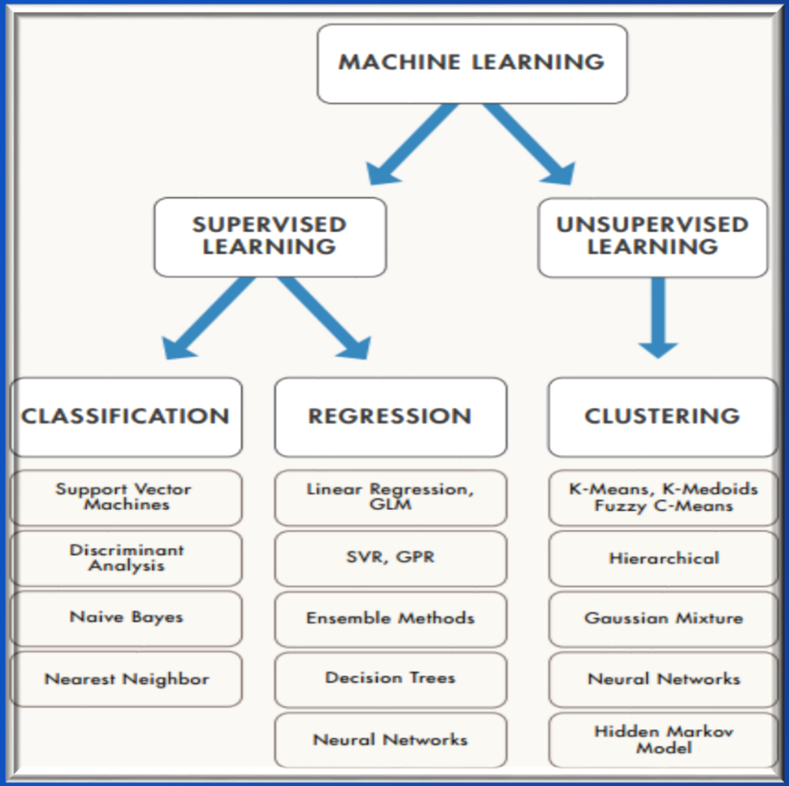
Presentation Outline

- Introduction
- Study area description
- Data Set
  - General description, descriptive statistics, histogram, boxplot
- Methodology used
  - Scenarios formulation
- Results & Discussions
- Conclusion

Introduction

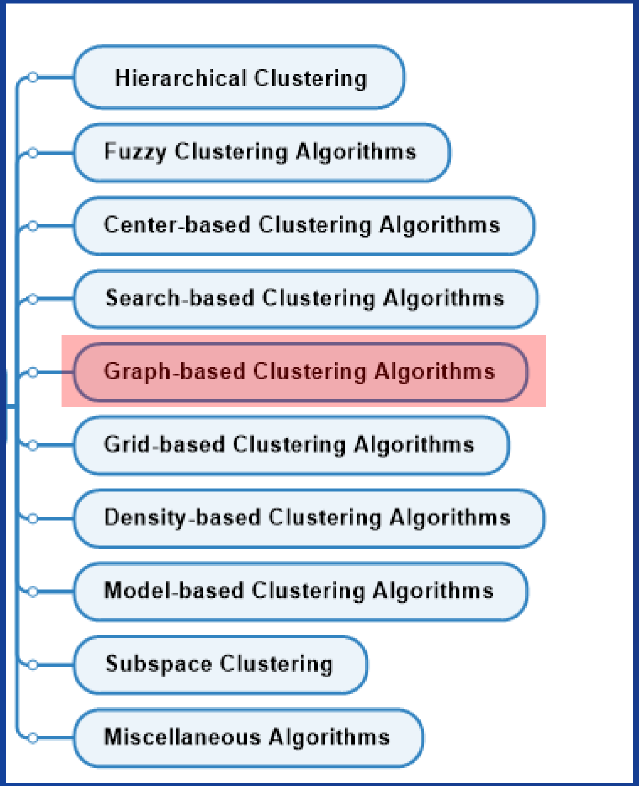
- **Urban groundwater (GW) is usually vulnerable to pollution.**
- **The main sources of GW quality degradation are:**
  - **Anthropogenic activities**
  - **Natural processes**
  - **Atmospheric input**
- **Subdividing the region into zones based on GW quality is usually undertaken.**
- **Recently, several methods are used, such as Genetic Algorithm, Model-Based Approach, Bayesian Approach, cluster analysis … etc.**
- **In this study, spatially constrained multivariate clustering method (SCMC) is used to subdivide Madina city (West KSA) into several zones based on six GW chemicals.**

Introduction (cluster analysis overview)



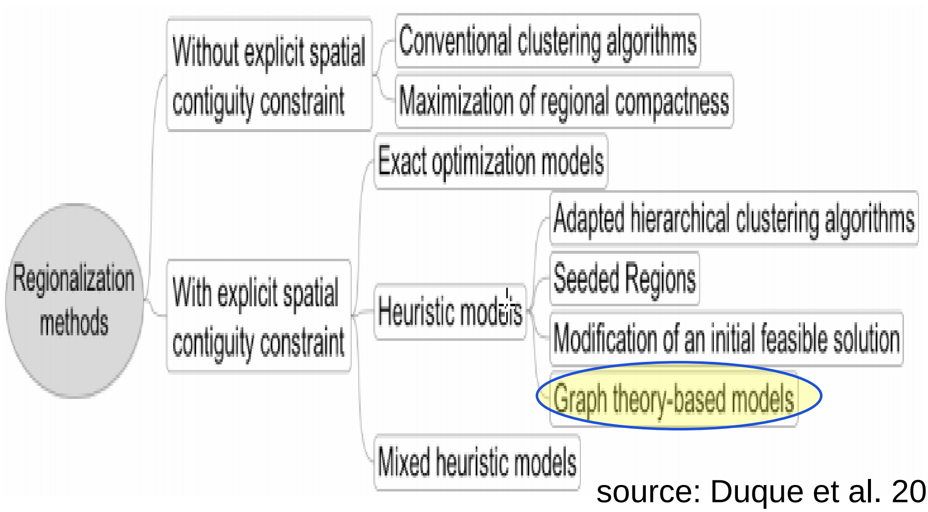
MACHINE LEARNING

- SUPERVISED LEARNING
  - CLASSIFICATION
    - Support Vector Machines
    - Discriminant Analysis
    - Naive Bayes
    - Nearest Neighbor
  - REGRESSION
    - Linear Regression, GLM
    - SVR, GPR
    - Ensemble Methods
    - Decision Trees
    - Neural Networks
- UNSUPERVISED LEARNING
  - CLUSTERING
    - K-Means, K-Medoids Fuzzy C-Means
    - Hierarchical
    - Gaussian Mixture
    - Neural Networks
    - Hidden Markov Model



- Hierarchical Clustering
- Fuzzy Clustering Algorithms
- Center-based Clustering Algorithms
- Search-based Clustering Algorithms
- Graph-based Clustering Algorithms
- Grid-based Clustering Algorithms
- Density-based Clustering Algorithms
- Model-based Clustering Algorithms
- Subspace Clustering
- Miscellaneous Algorithms

Figure 1. Taxonomy of methods for solving regionalization problems

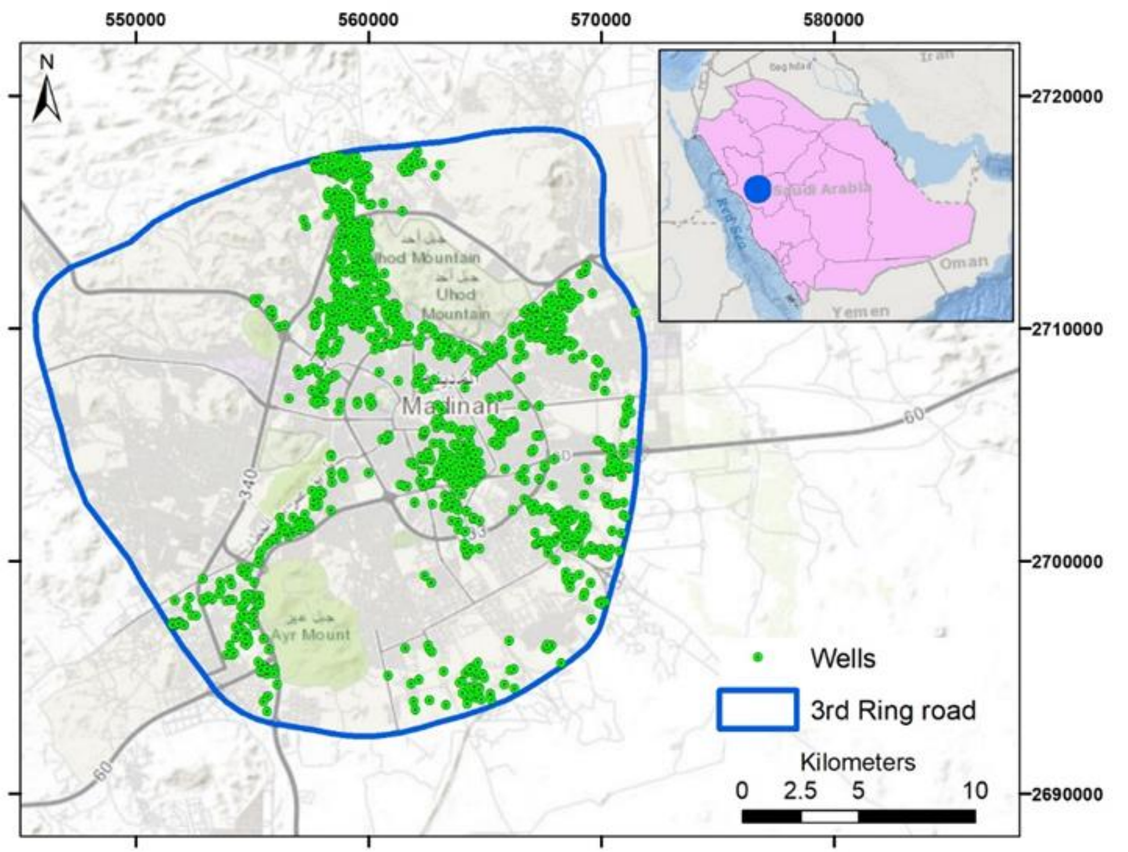


- Regionalization methods
  - Without explicit spatial contiguity constraint
    - Conventional clustering algorithms
    - Maximization of regional compactness
  - With explicit spatial contiguity constraint
    - Exact optimization models
    - Heuristic models
      - Adapted hierarchical clustering algorithms
      - Seeded Regions
      - Modification of an initial feasible solution
      - Graph theory-based models
    - Mixed heuristic models

source: Duque et al. 2007

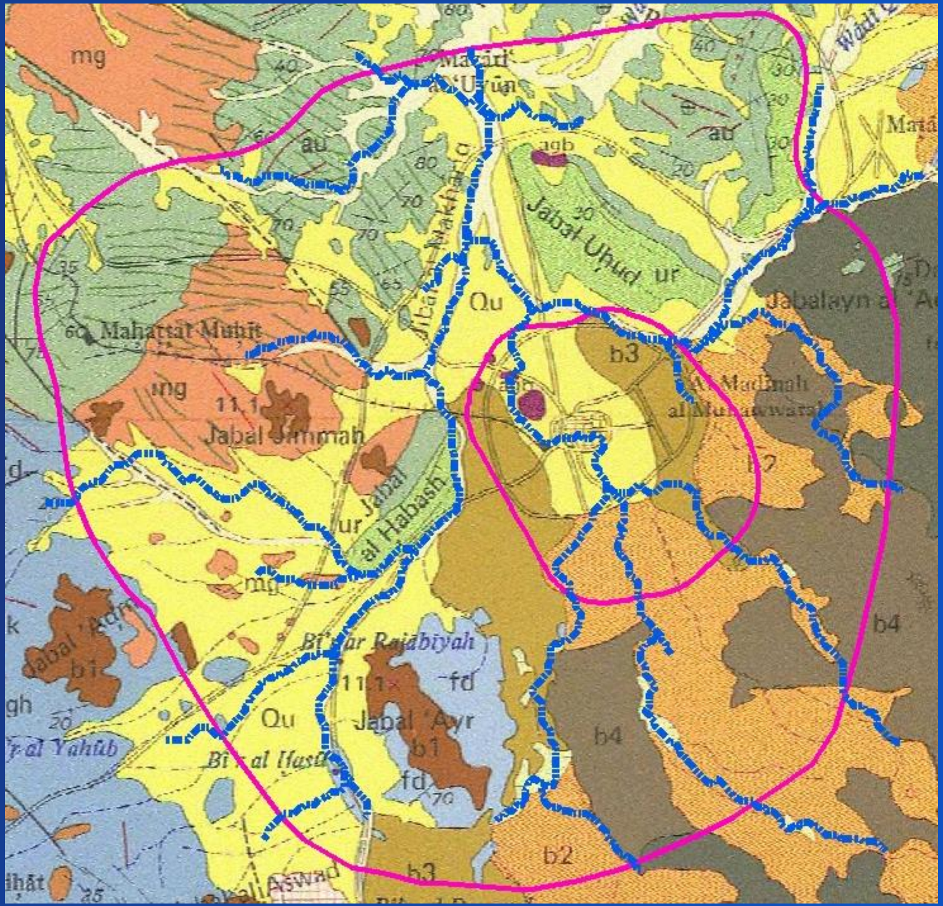
Study area description (General Location)

- Madinah city (West KSA) is selected.
- The population about 1.25 million + 10 million annual visitors.
- covering an area of 522 km^2.
- In this study, The wells inside the 3rd ring road is selected for analysis.
- These wells are located in private farms.



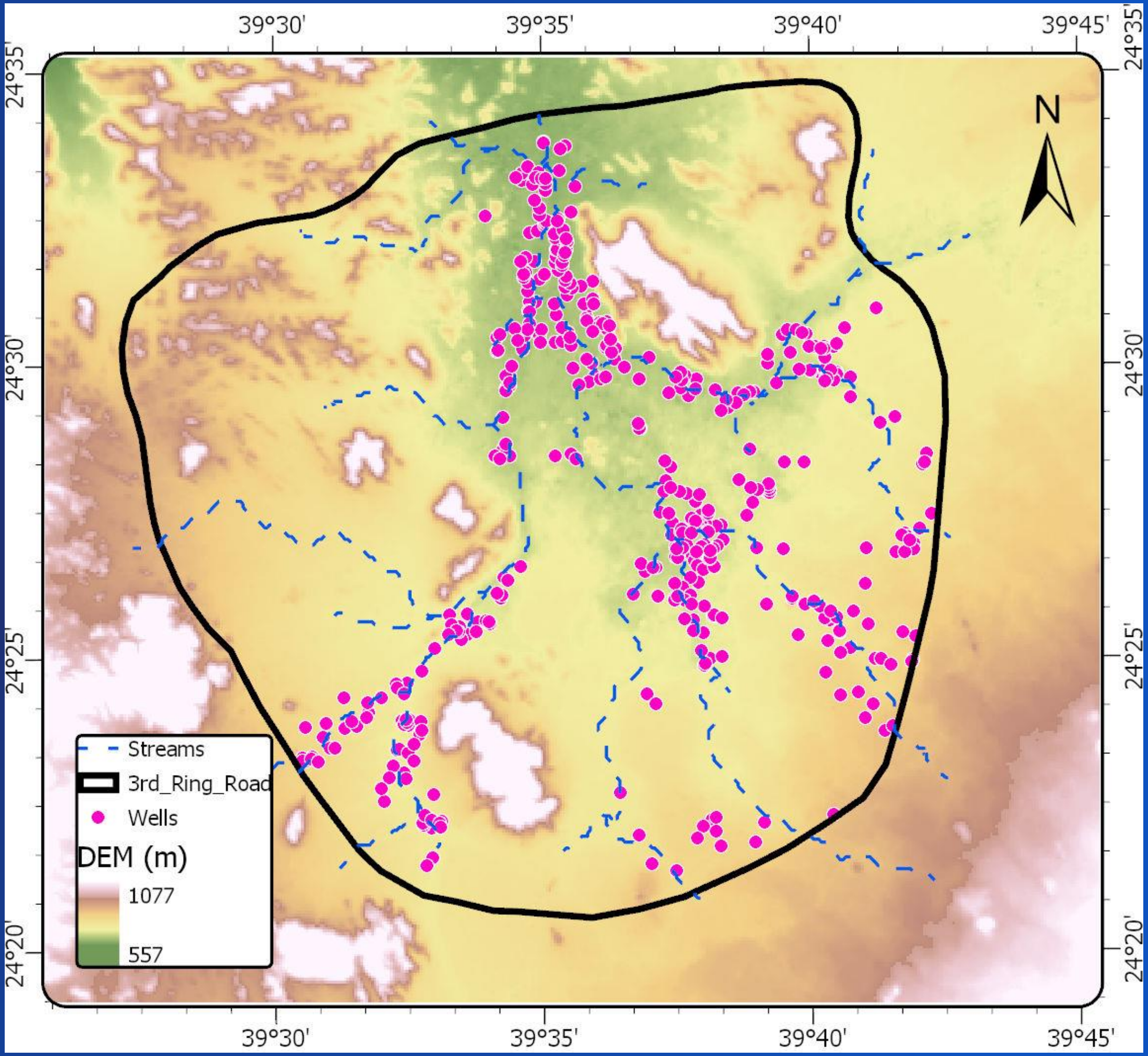
Study area description (Geology)

- The geology of the study area consists mainly of three parts;
  - lava plateaus (volcanic basalt flows)
  - alluvial deposits
  - rock outcrops (pre-cambrian rock).
- The first two parts are the places of shallow groundwater aquifers.
- 50% of the study area is covered by volcanic basalt rocks

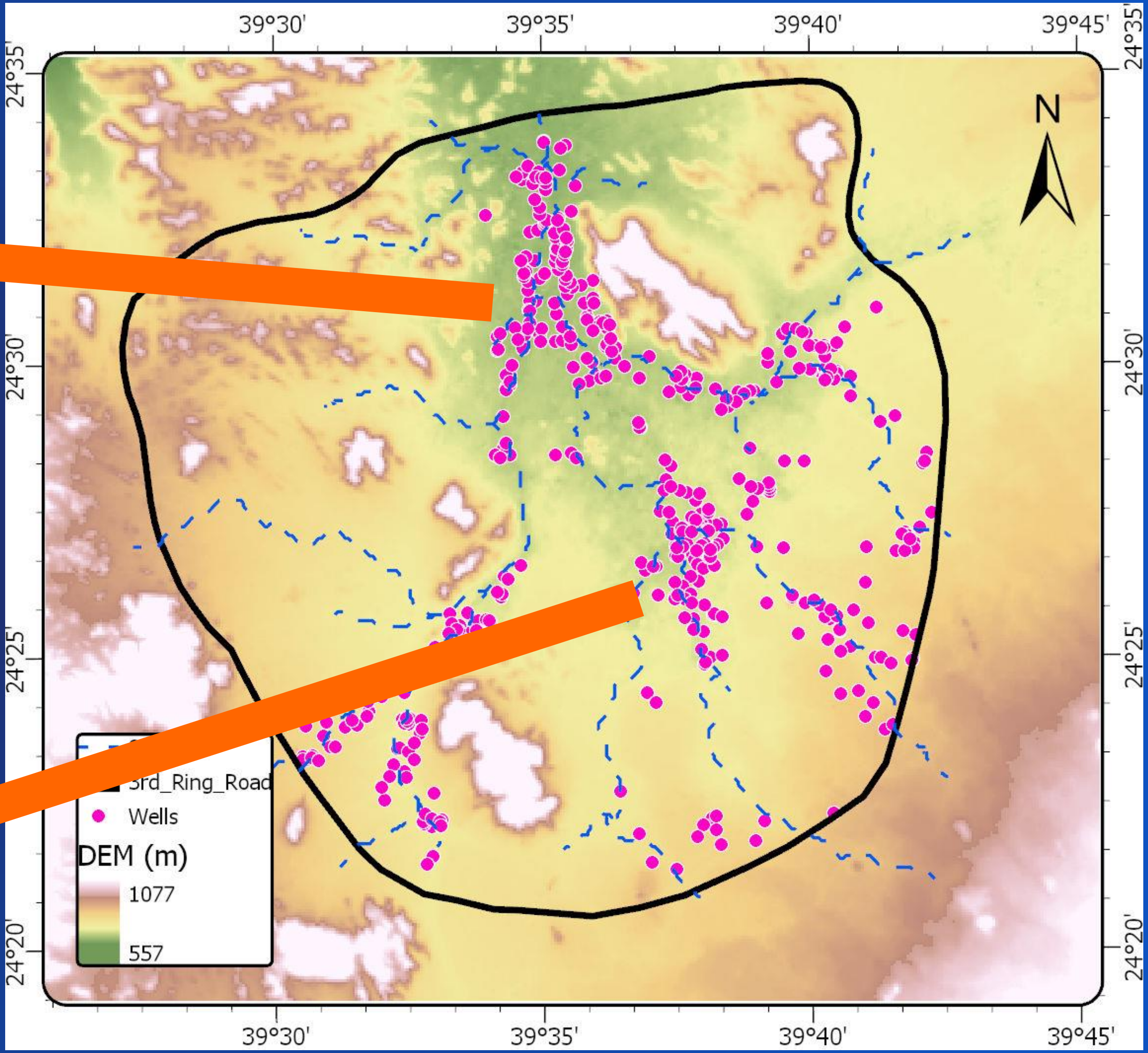


Study area description (Topography)

- The elevation ranges from 570 m (a.m.s.l.) up to 1,100 m.
- Strong relationship between wells location and watercourses (ephemeral streams)



Study area description (Wells location)

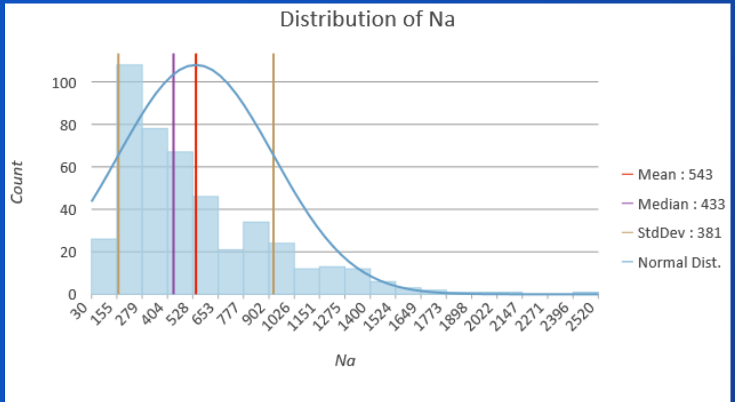


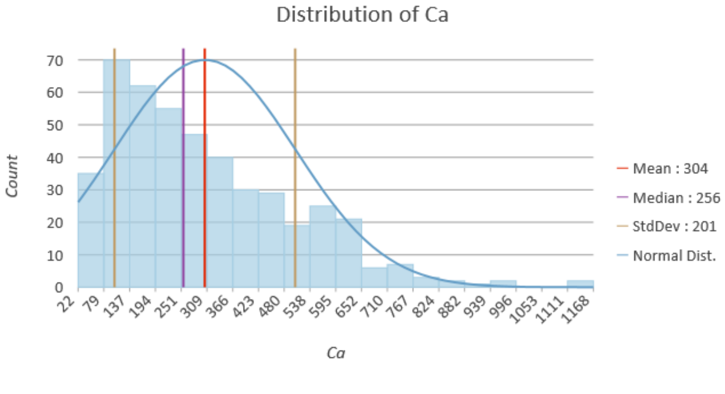
Data set

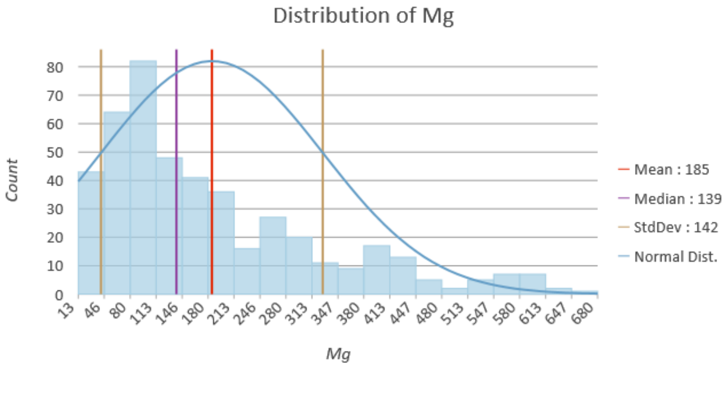
- **456 private farms inside the 3rd ring road are visited.**
- **From each farm, one well location is registered using GPS.**
- **Water samples from the wells are collected and taken to the laboratory for analysis (pH, TDS, EC, hardness, turbidity, alkalinity, color, ions (*cations and anions*).**
- **Three cations (Na, Ca, Mg) and three anions (Cl, HCO_3, SO_4) are selected for cluster analysis.**

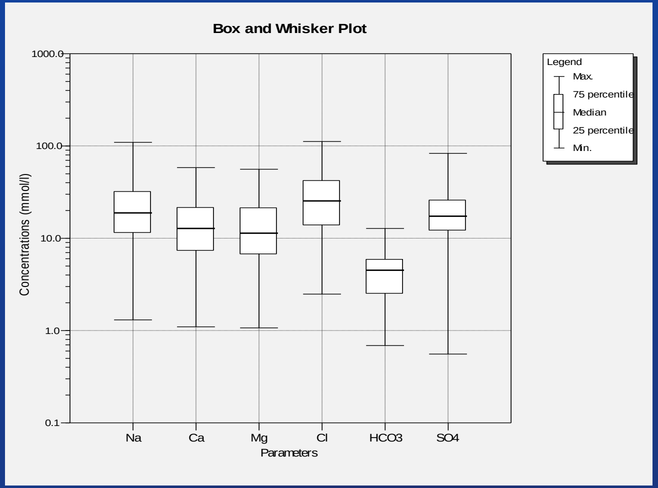
	Na	Ca	Mg	Cl	HCO_3	SO_4
Min.	30	22	13	88	42	26
Max.	2520	1168	680	3970	780	3996
Avg.	543	304	185	1080	271	982
Media	433	256	139	901	276	836
STD	381	201	142	740	131	611

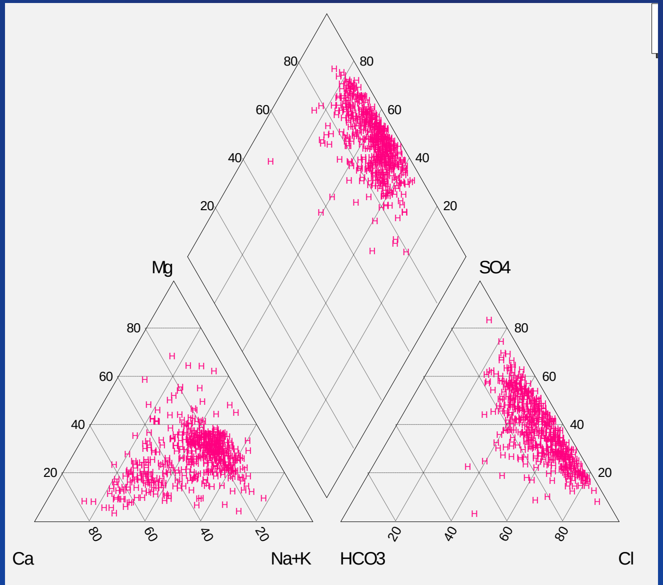
Data set



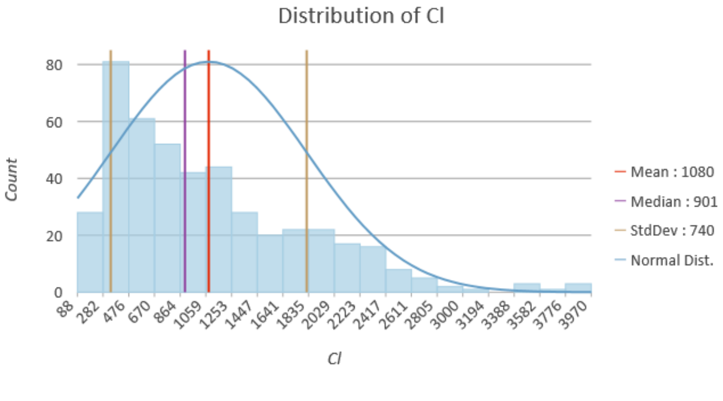


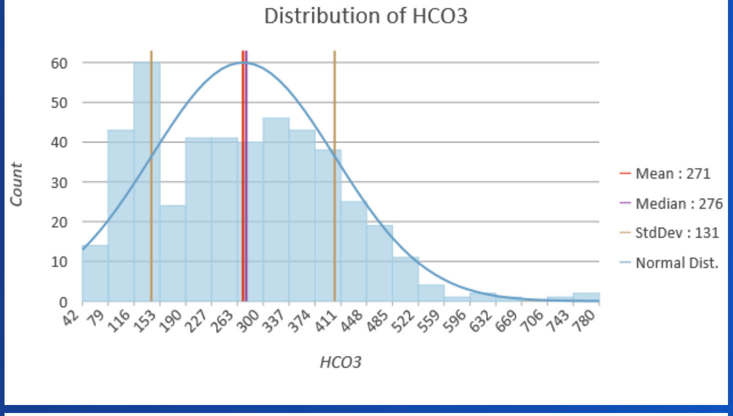


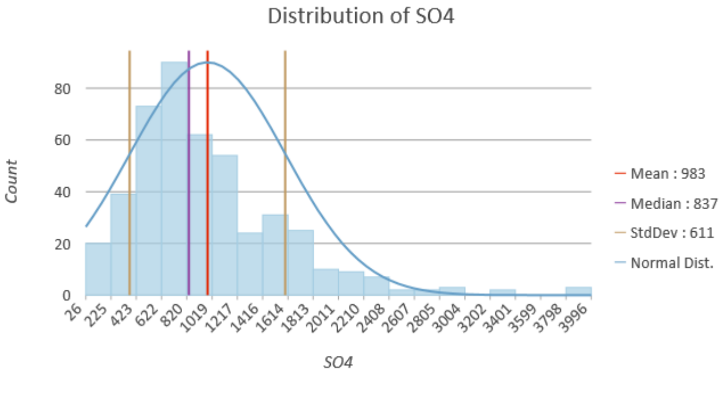




Piper plots (Trilinear diagram)

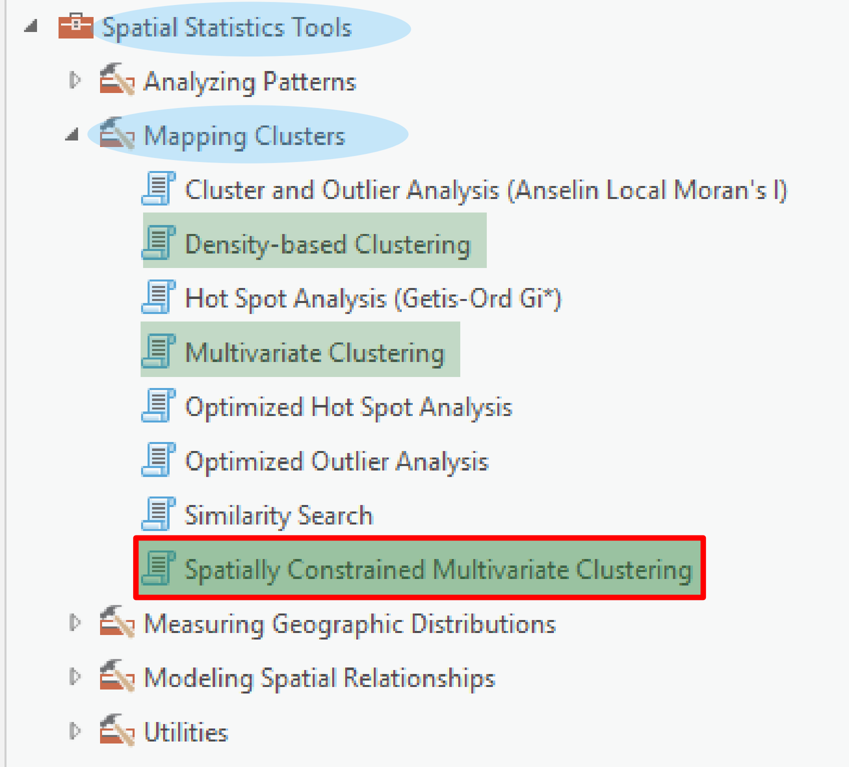






Methodology (clustering in ArcGIS Pro)

- **The Three new cluster methods in ArcGIS Pro are:**
  - **density-based clustering,**
  - **multivariate clustering and**
  - **spatially constrained multivariate clustering (SCMC).**



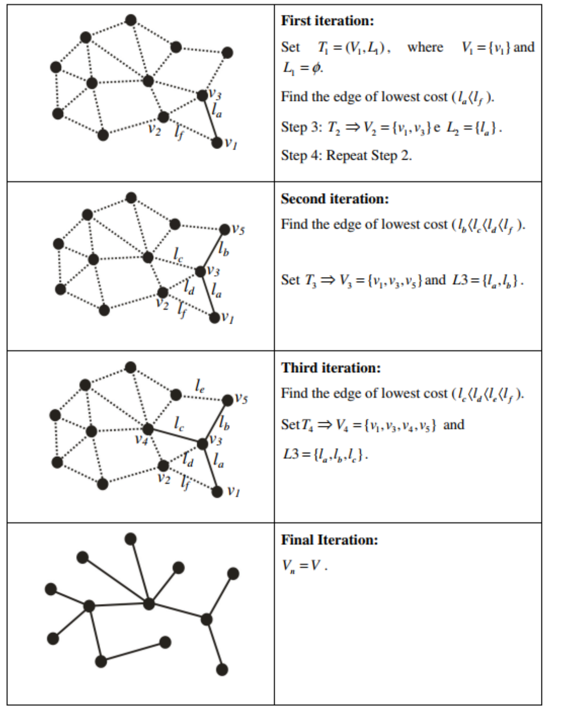
Spatial Statistics Tools
Analyzing Patterns
Mapping Clusters
Cluster and Outlier Analysis (Anselin Local Moran's I)
Density-based Clustering
Hot Spot Analysis (Getis-Ord Gi*)
Multivariate Clustering
Optimized Hot Spot Analysis
Optimized Outlier Analysis
Similarity Search
Spatially Constrained Multivariate Clustering
Measuring Geographic Distributions
Modeling Spatial Relationships
Utilities

SCMS is the process of grouping of the observations based on the attributes similarity and location similarity using multiple objective optimization.

maximizing within-group similarity

Methodology (theoretical background)

- **In ArcGIS Pro, spatially constrained multivariate clustering tool uses Spatial K'luster Analysis by Tree Edge Removal (SKATER) algorithm which is based on minimum spanning tree (MST) method**



First iteration:

Set $T_1 = (V_1, L_1)$, where $V_1 = \{v_1\}$ and $L_1 = \phi$.

Find the edge of lowest cost ($l_a \langle l_f$).

Step 3: $T_2 \Rightarrow V_2 = \{v_1, v_3\}$ e $L_2 = \{l_a\}$.

Step 4: Repeat Step 2.

Second iteration:

Find the edge of lowest cost ($l_b \langle l_c \langle l_d \langle l_f$).

Set $T_3 \Rightarrow V_3 = \{v_1, v_3, v_5\}$ and $L3 = \{l_a, l_b\}$.

Third iteration:

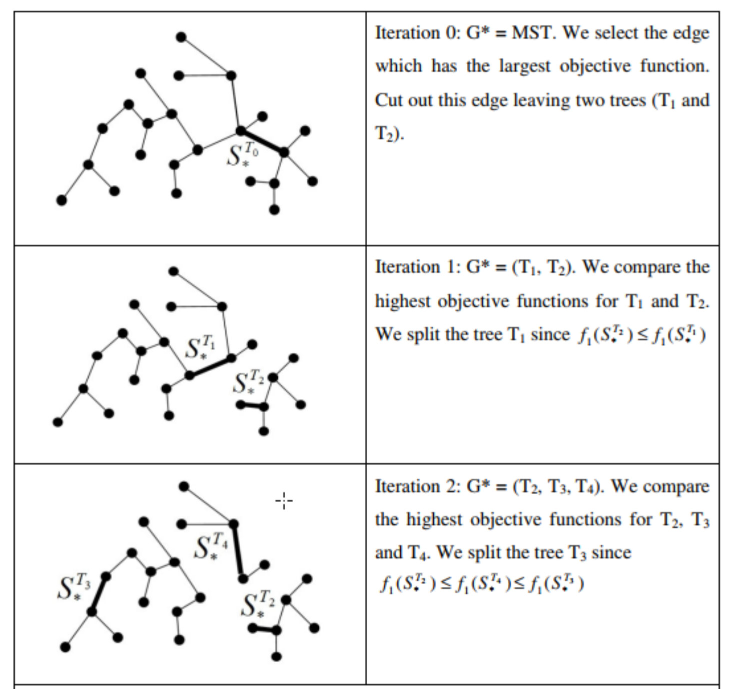
Find the edge of lowest cost ($l_c \langle l_d \langle l_e \langle l_f$).

Set $T_4 \Rightarrow V_4 = \{v_1, v_3, v_4, v_5\}$ and $L3 = \{l_a, l_b, l_c\}$.

Final Iteration:

$V_n = V$.

Figure 2 - Construction of the minimum spanning tree.

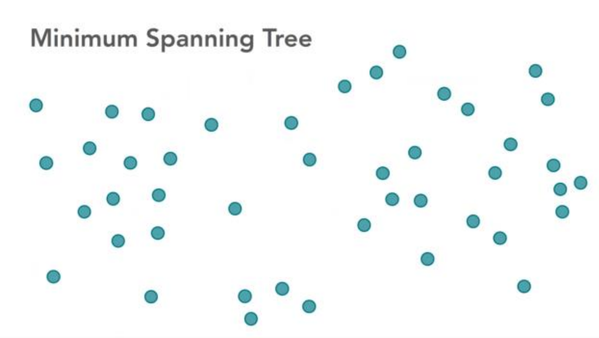


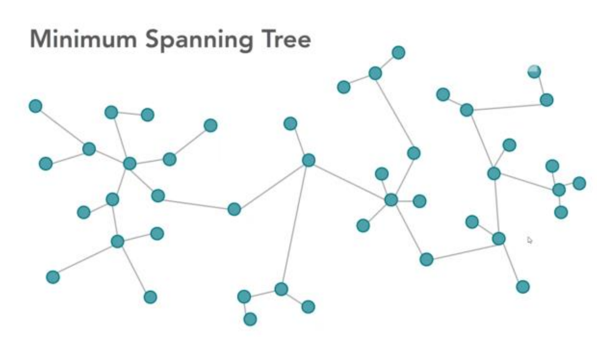
Iteration 0: G* = MST. We select the edge which has the largest objective function. Cut out this edge leaving two trees (T_1 and T_2).

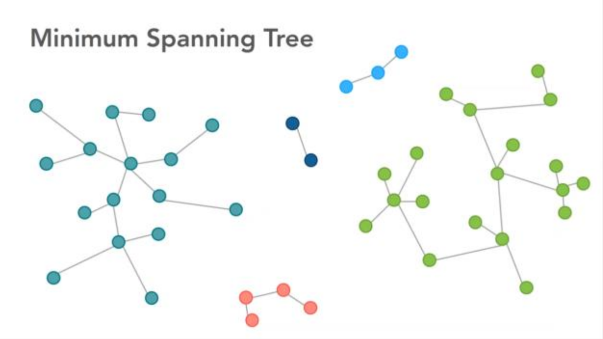
Iteration 1: G* = (T_1, T_2). We compare the highest objective functions for T_1 and T_2. We split the tree T_1 since $f_1(S_*^{T_2}) \leq f_1(S_*^{T_1})$

Iteration 2: G* = (T_2, T_3, T_4). We compare the highest objective functions for T_2, T_3 and T_4. We split the tree T_3 since $f_1(S_*^{T_2}) \leq f_1(S_*^{T_4}) \leq f_1(S_*^{T_3})$

Figure 3 – Partitioning of the MST







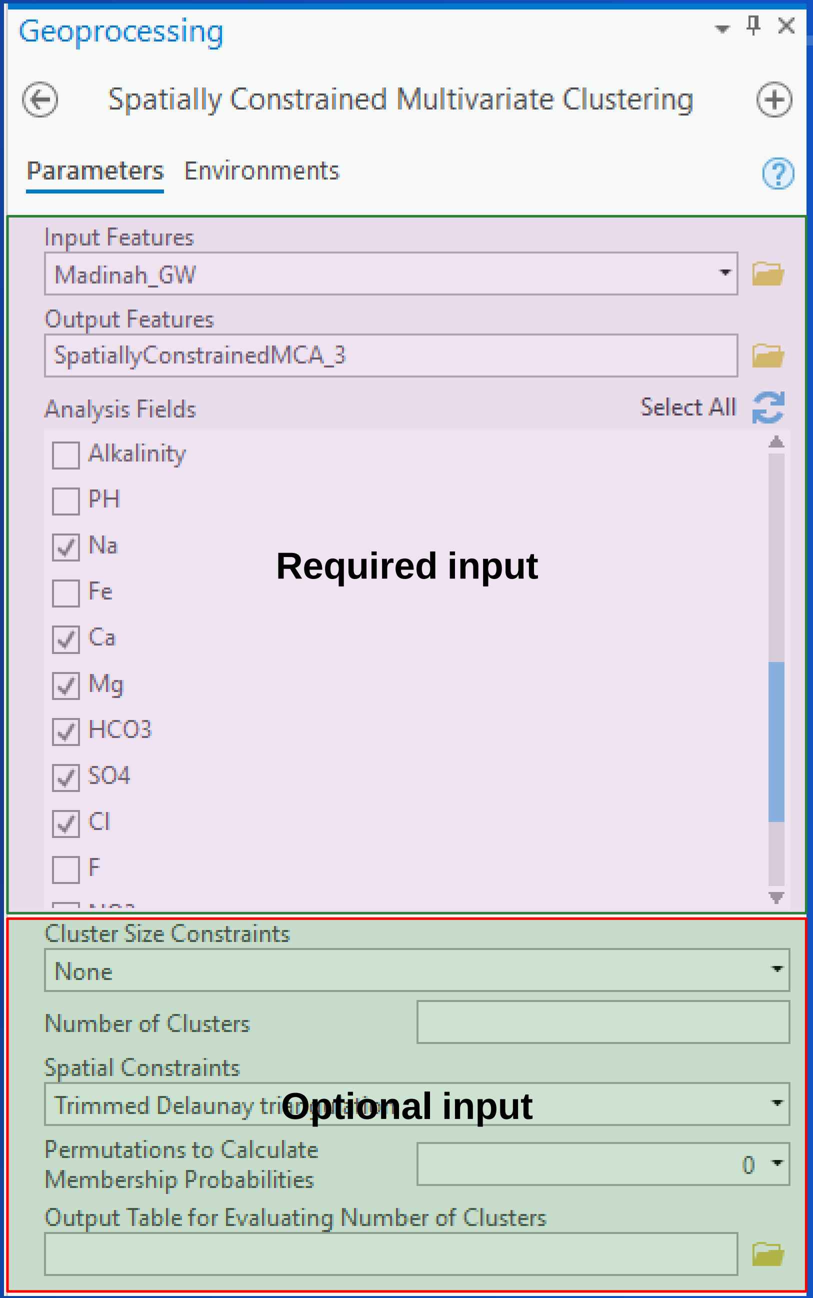
construct the network graph by connecting the contiguous nodes with lowest cost.

find the shortest path that minimizes the sum of dissimilarity (or maximizing the sum of similarities) – minimum spanning tree

Source: Assuncao et al (2006)

Methodology

- Spatially constrained Multivariate Clustering (SCMC) methods in ArcGIS Pro has two main groups of input parameters:
- **Three required input:**
  - Input layer
  - The name of output layer
  - Selected attributes for analysis
- **Five optional input:**
  - Cluster size constraints (None, No. of features, Attribute value)
  - Number of clusters
  - Spatial constraints
  - Permutations Membership probabilities
  - Output table for evaluating number of clusters



Input Features: Madinah_GW

Output Features: SpatiallyConstrainedMCA_3

Analysis Fields (Select All)

□ Alkalinity
□ PH
☑ Na
□ Fe
☑ Ca
☑ Mg
☑ HCO3
☑ SO4
☑ Cl
□ F

Cluster Size Constraints: None

Number of Clusters:

Spatial Constraints: Trimmed Delaunay triangulation

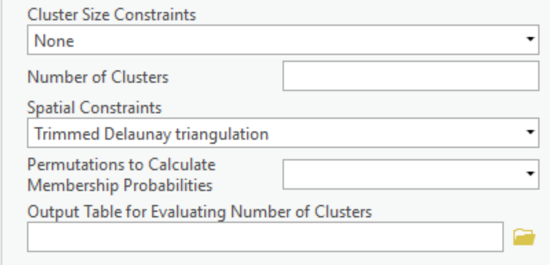
Permutations to Calculate Membership Probabilities: 0

Output Table for Evaluating Number of Clusters:

Methodology (Scenarios formulation)

- Four groups of scenarios are developed based on the optional input parameters:
  - **Group (A) scenario** : no optional input and the optimum number of clusters is computed automatically.
  - **Group (B) scenarios** : three optimum No. of clusters are specified (2, 3 and 7).
  - **Group (C) scenarios**: min. No. of features per cluster is specified (20, 40), with fixed number of cluster (= 7 clusters).
  - **Group (D) scenarios**: min. no. and max. no. of features per clusters are specified (“25, 150” and “50, 100”), the optimum number of clusters is computed automatically.

Group (A) scenario



Cluster Size Constraints: None

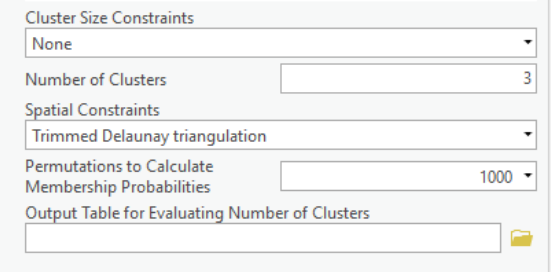
Number of Clusters:

Spatial Constraints: Trimmed Delaunay triangulation

Permutations to Calculate Membership Probabilities:

Output Table for Evaluating Number of Clusters:

Group (B) scenarios



Cluster Size Constraints: None

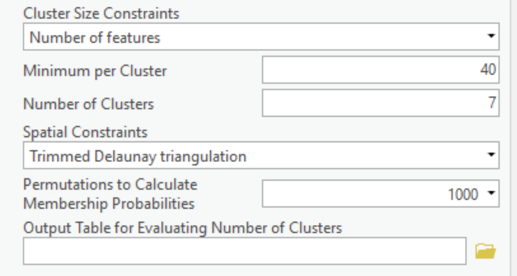
Number of Clusters: 3

Spatial Constraints: Trimmed Delaunay triangulation

Permutations to Calculate Membership Probabilities: 1000

Output Table for Evaluating Number of Clusters:

Group (C) scenarios



Cluster Size Constraints: Number of features

Minimum per Cluster: 40

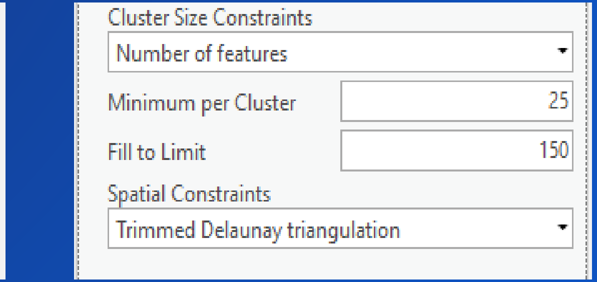
Number of Clusters: 7

Spatial Constraints: Trimmed Delaunay triangulation

Permutations to Calculate Membership Probabilities: 1000

Output Table for Evaluating Number of Clusters:

Group (D) scenarios



Cluster Size Constraints: Number of features

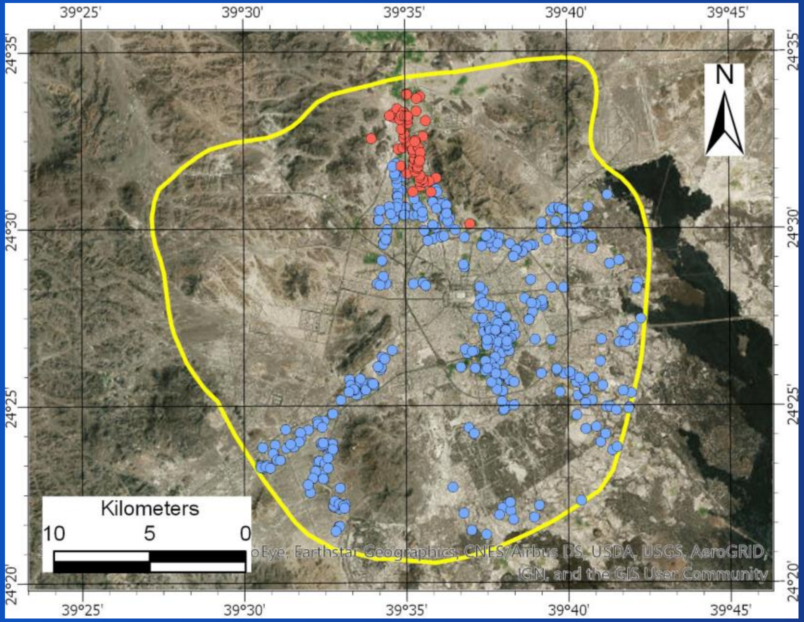
Minimum per Cluster: 25

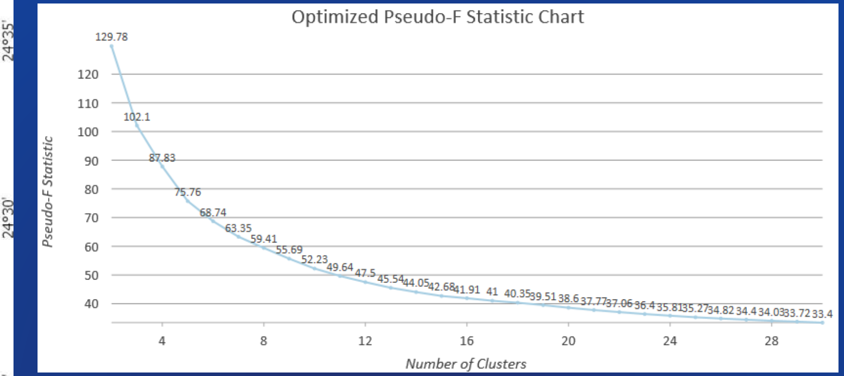
Fill to Limit: 150

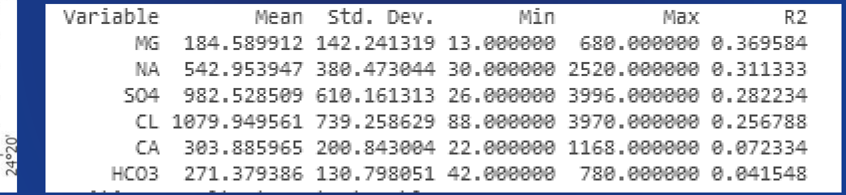
Spatial Constraints: Trimmed Delaunay triangulation

Results & Discussions (Group (A) scenario)

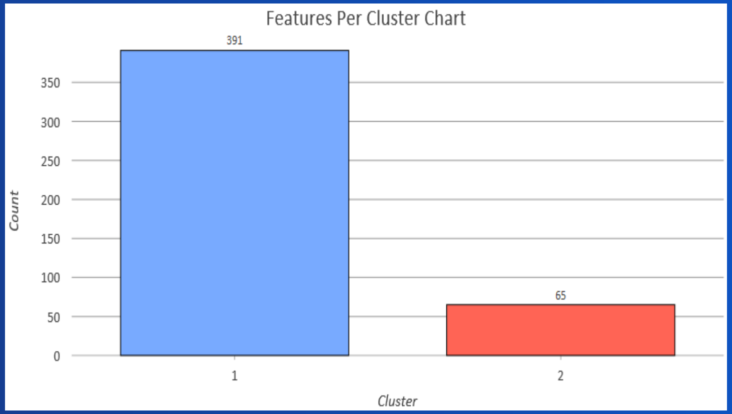
No optional input and the optimum number of clusters is computed automatically

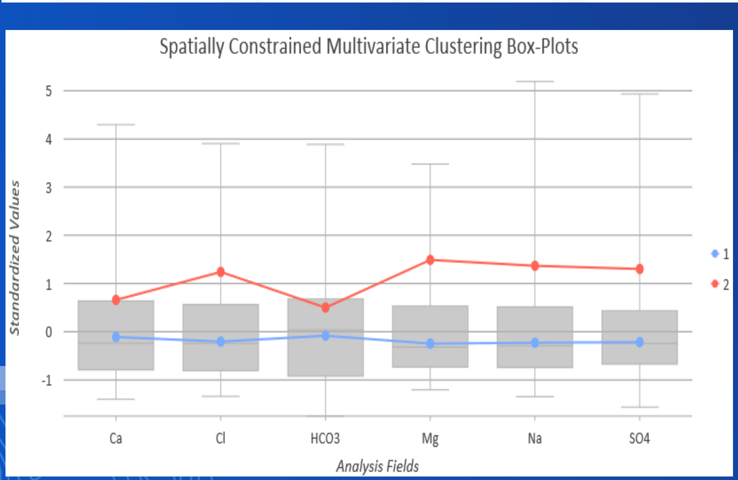

39°25'
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
N
Kilometers
10
5
0
oEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

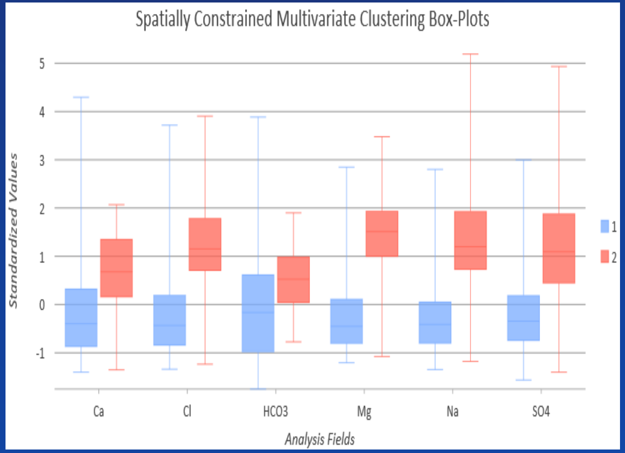


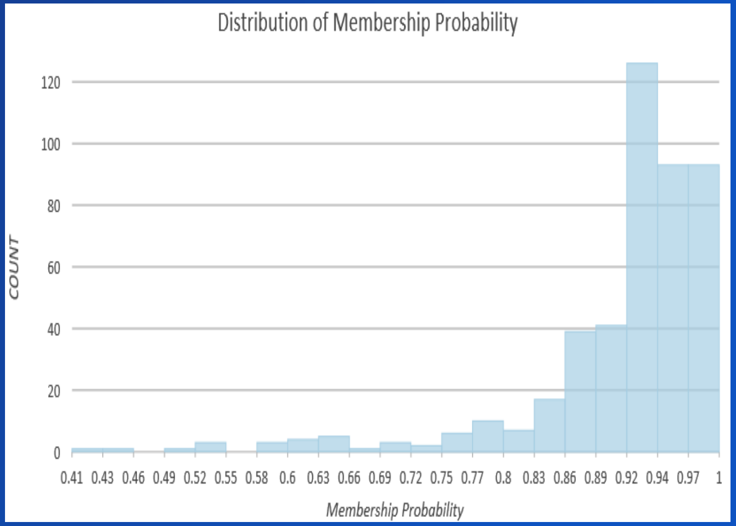


```
Variable          Mean   Std. Dev.        Min          Max       R2
      MG   184.589912  142.241319  13.000000   680.000000 0.369584
      NA   542.953947  380.473044  30.000000  2520.000000 0.311333
     SO4   982.528509  610.161313  26.000000  3996.000000 0.282234
      CL  1079.949561  739.258629  88.000000  3970.000000 0.256788
      CA   303.885965  200.843004  22.000000  1168.000000 0.072334
    HCO3   271.379386  130.798051  42.000000   780.000000 0.041548
```



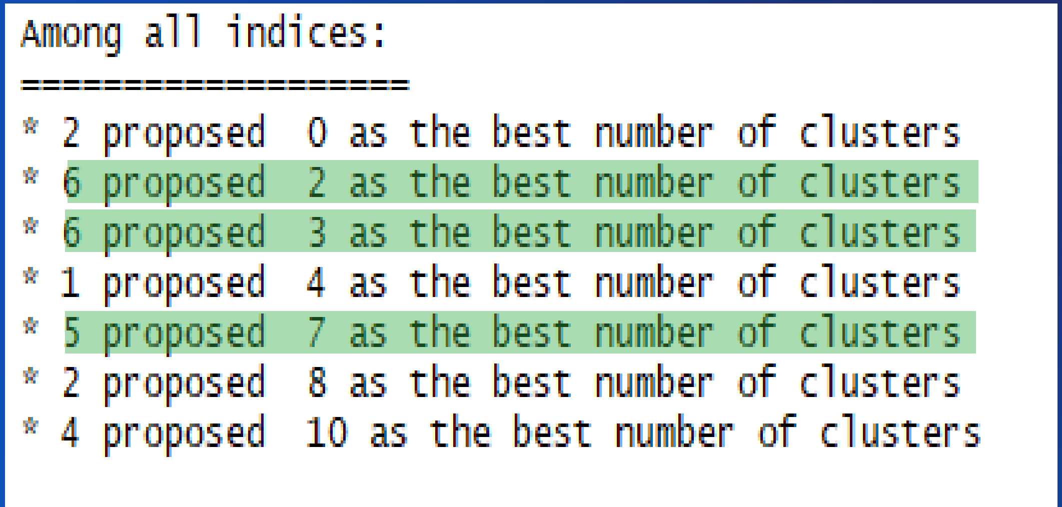




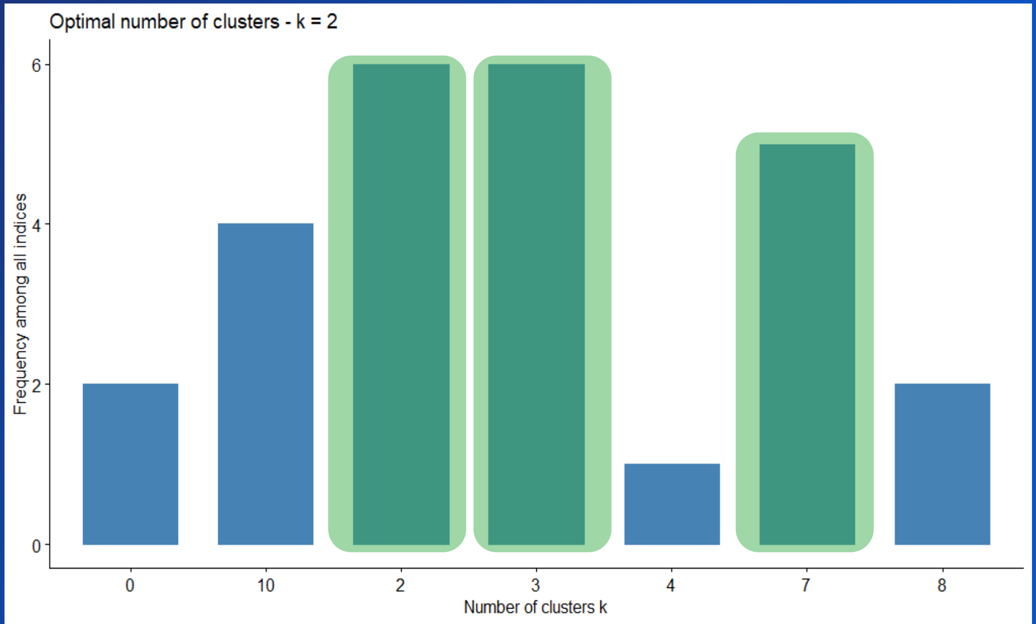


Results & Discussions (Group (B) scenarios)

- Most of the clusters methods need from the user to specify the optimum number of clusters
- Unfortunately, this is still unsolved problem and there is no definitive answer to this question.
- Determining the optimal number of clusters is somehow subjective.
- In this study, the optimum number of clusters is determined by evaluating 30 indices using R programming language (NbClust R package).
- Selection of the optimum No. of cluster is based on the “majority rule”.



```
Among all indices:
===================
* 2 proposed  0 as the best number of clusters
* 6 proposed  2 as the best number of clusters
* 6 proposed  3 as the best number of clusters
* 1 proposed  4 as the best number of clusters
* 5 proposed  7 as the best number of clusters
* 2 proposed  8 as the best number of clusters
* 4 proposed  10 as the best number of clusters
```

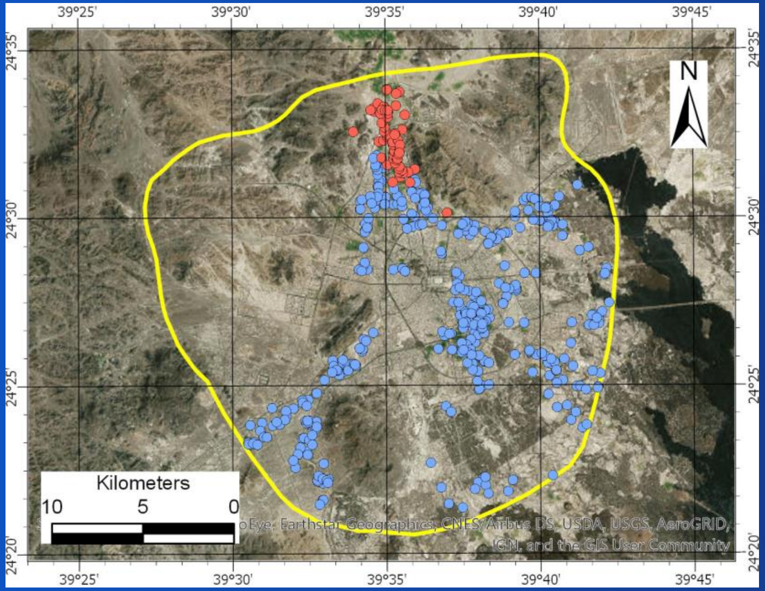


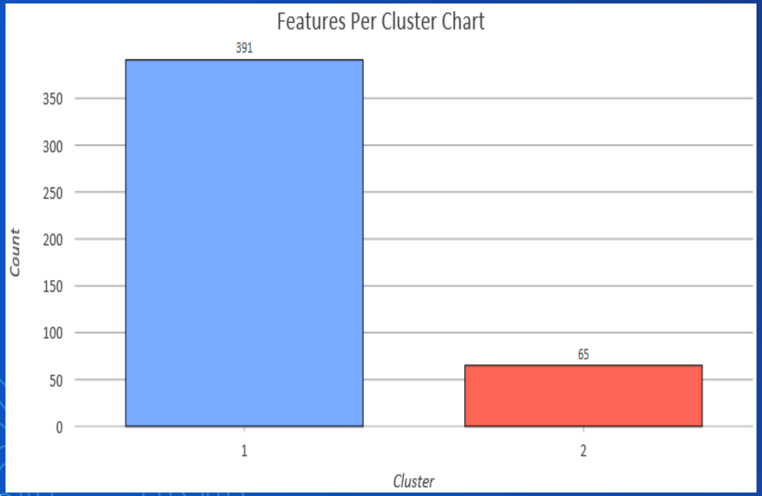
- 2, 3 and 7 are selected as the optimum number of clusters

Results & Discussions (Group (B) scenarios)

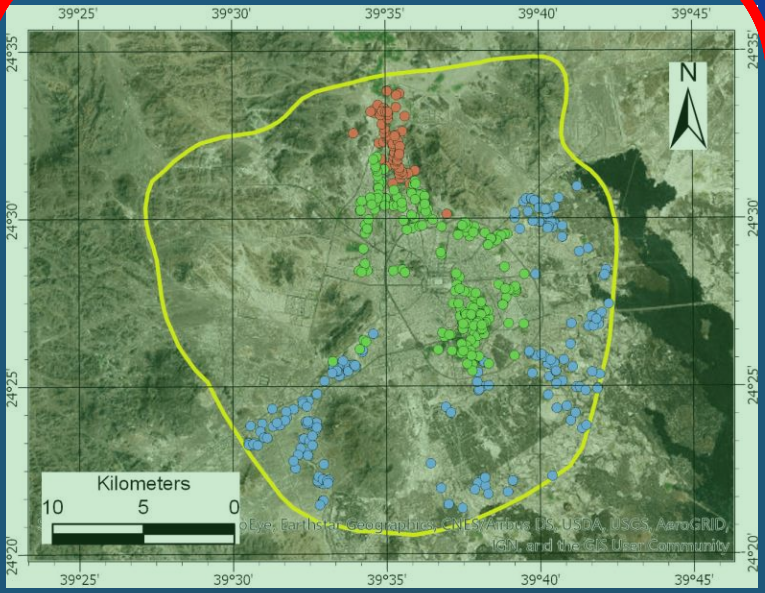
Three optimum No. of clusters are specified (2, 3 and 7)

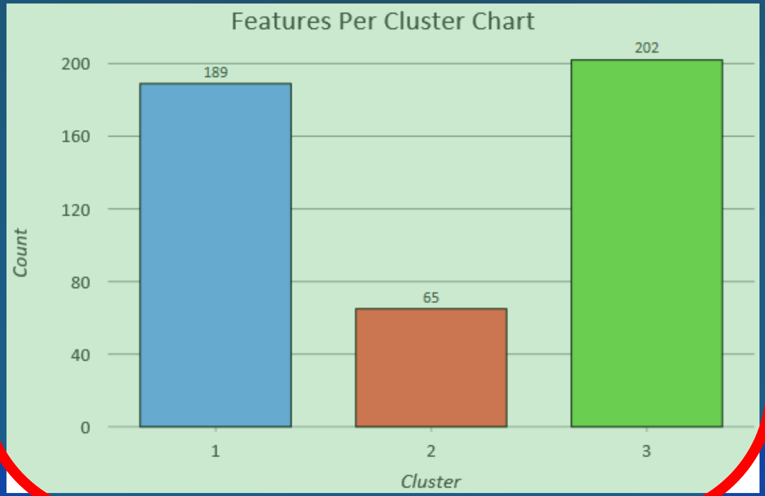
No. of clusters = 2


39°25'
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
N
Kilometers
10
5
0
oEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

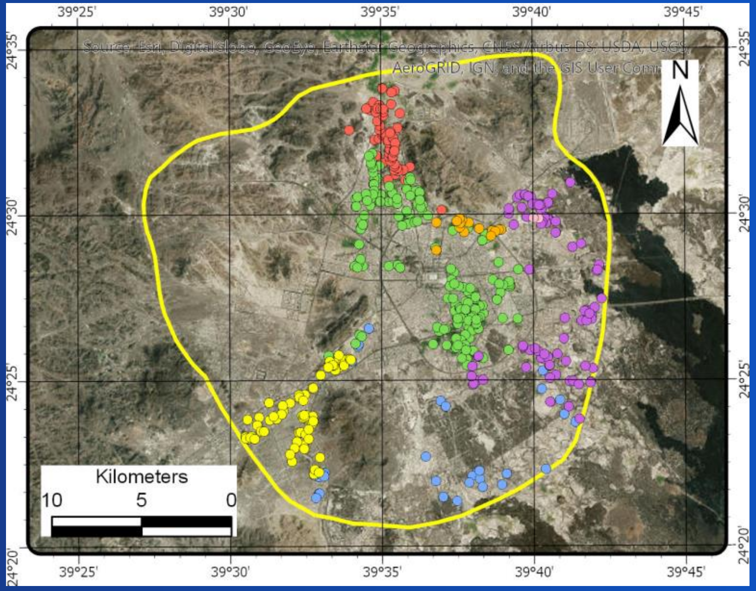


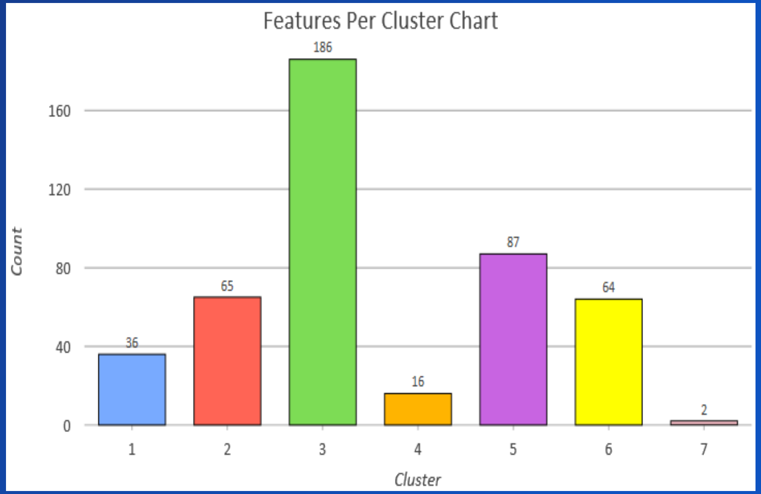
No. of clusters = 3


39°25'
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
N
Kilometers
10
5
0
oEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community



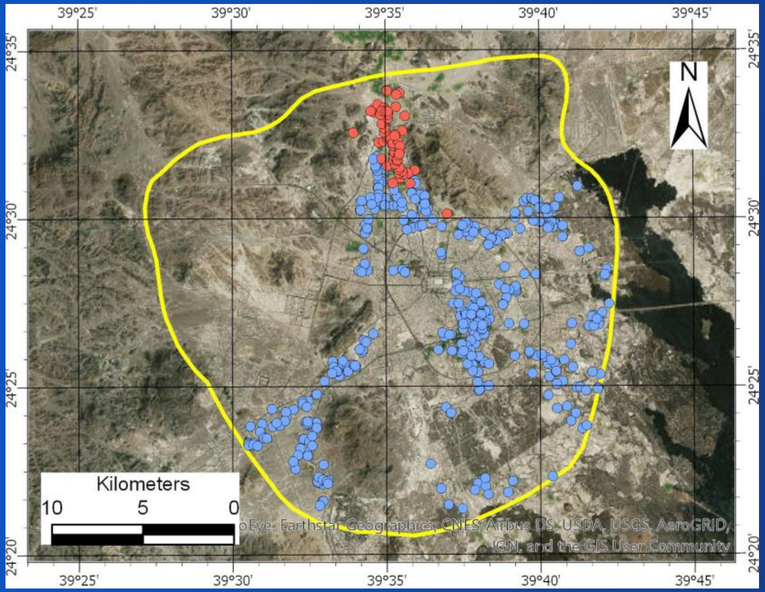
No. of clusters = 7


39°25'
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
N
Kilometers
10
5
0
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Comm

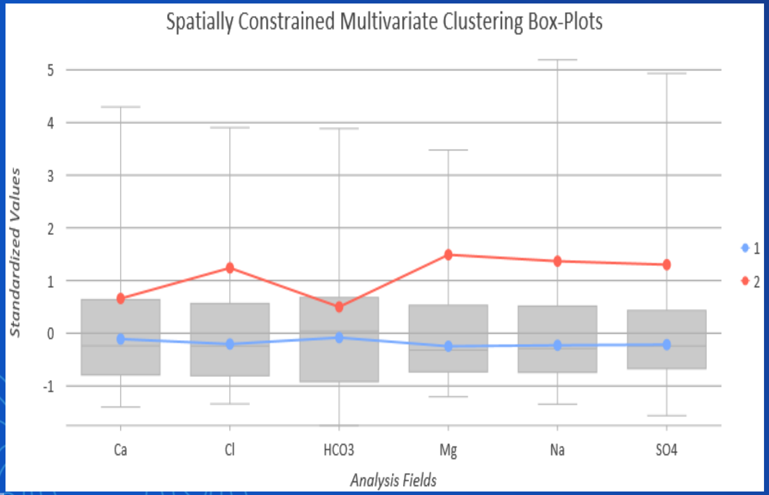


Results & Discussions (Group (B) scenarios)

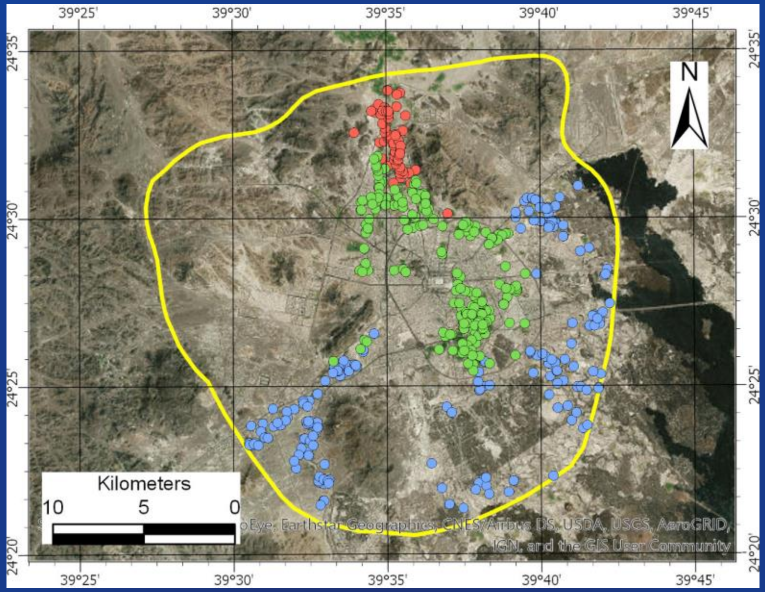
No. of clusters = 2



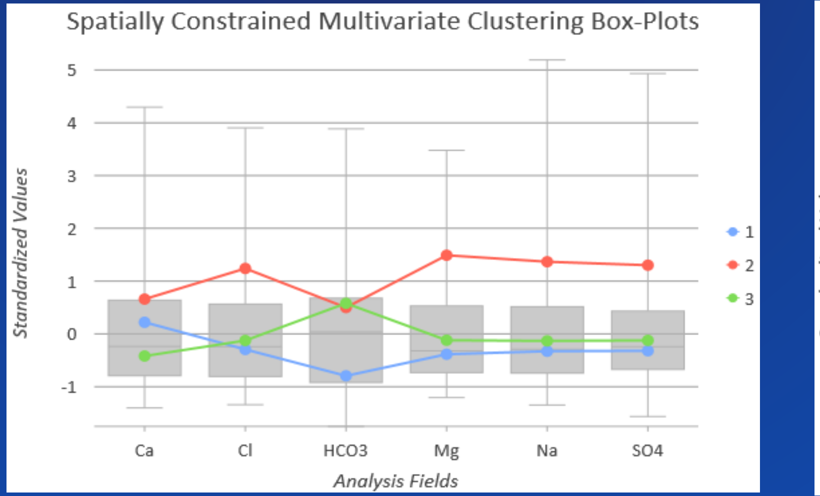
39°25'
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
N
Kilometers
10
5
0
oEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community



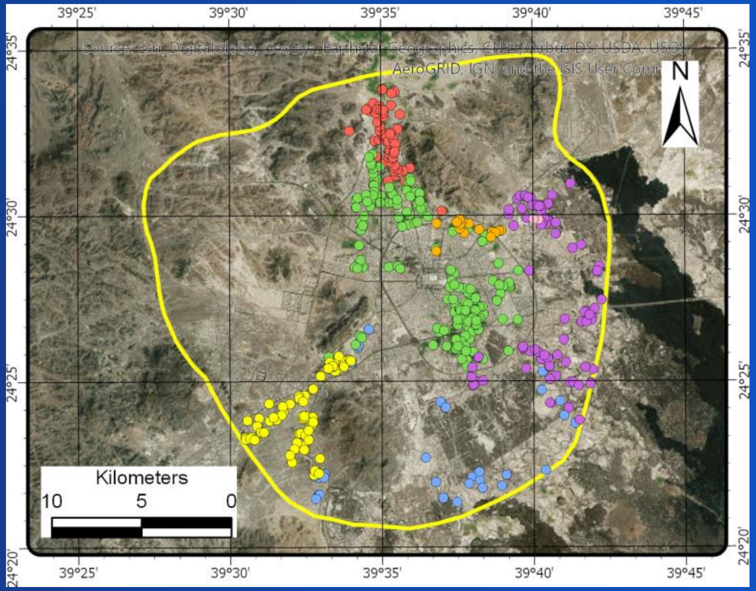
No. of clusters = 3



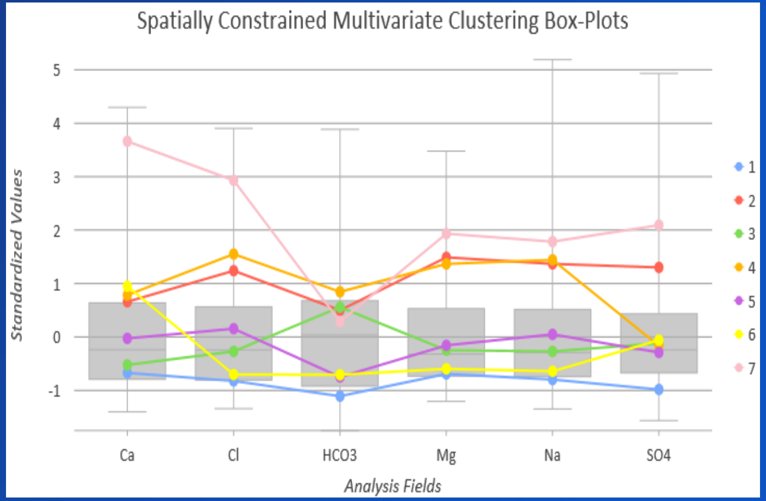
39°25'
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
N
Kilometers
10
5
0
oEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community



No. of clusters = 7

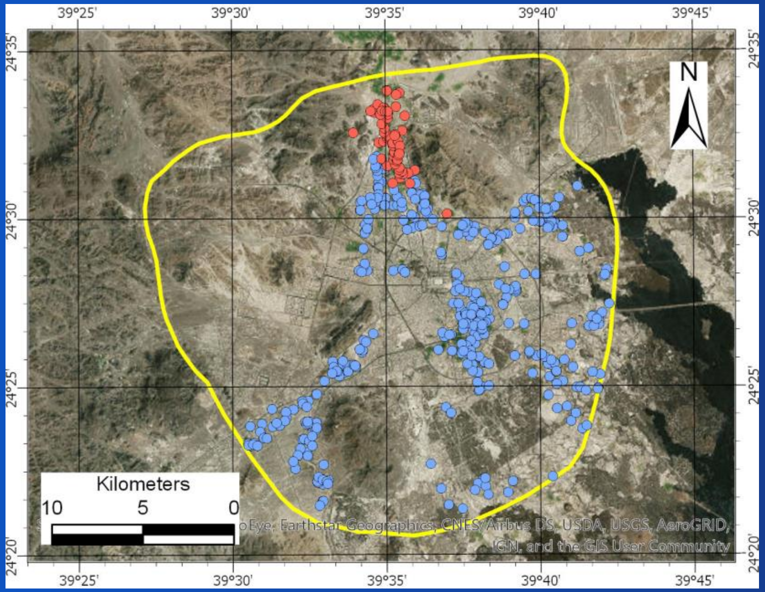


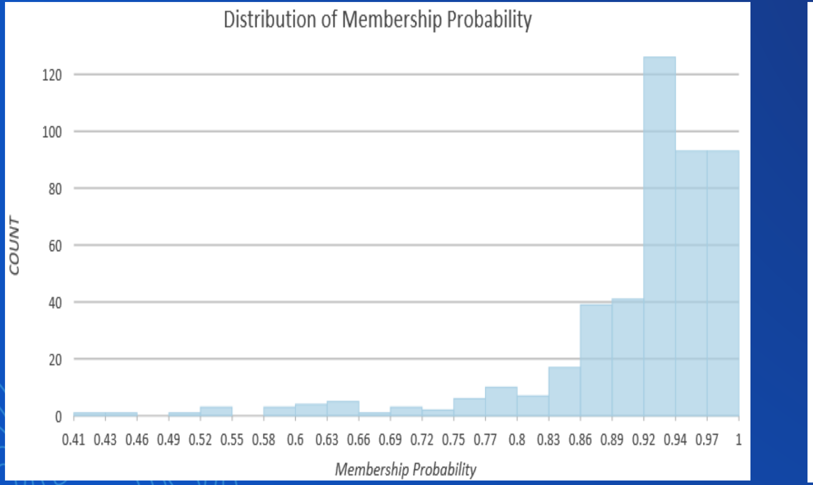
39°25'
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
Sources: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Comm
N
Kilometers
10
5
0



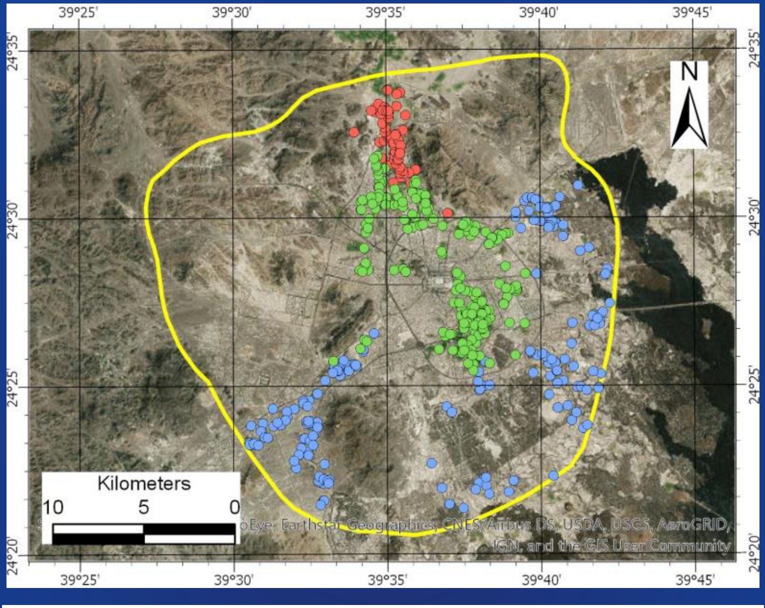
Results & Discussions (Group (B) scenarios)

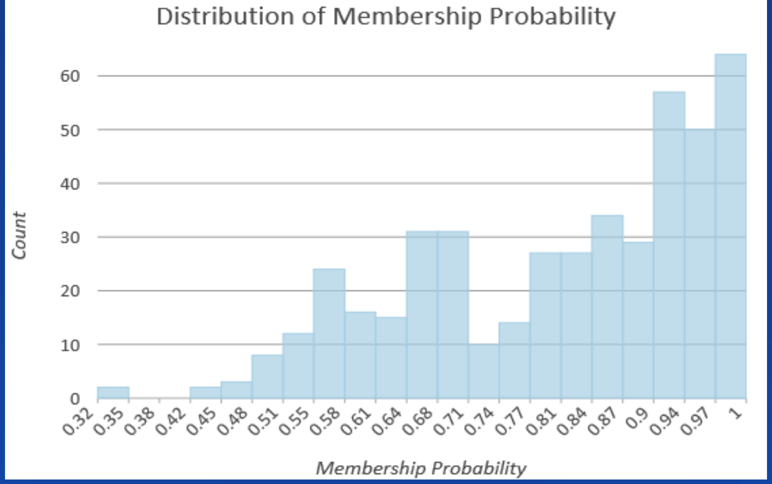
No. of clusters = 2


39°25'
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
N
Kilometers
10
5
0
oEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID,
IGN, and the GIS User Community

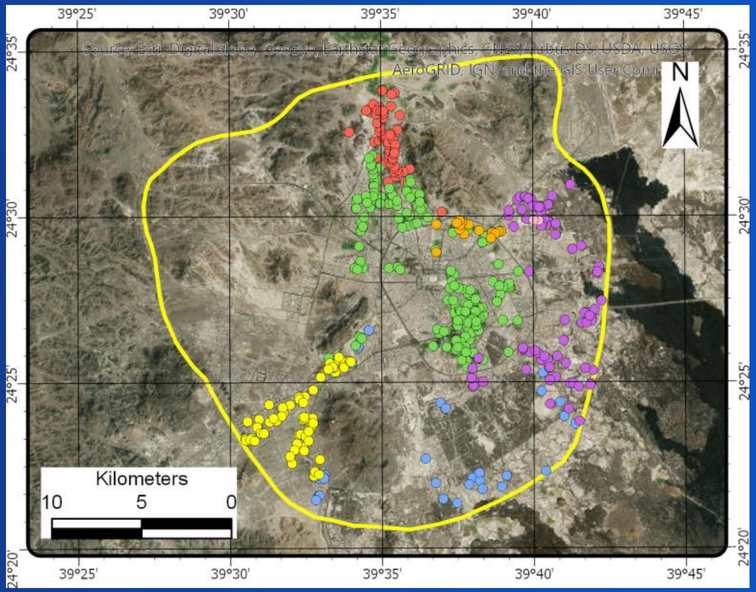


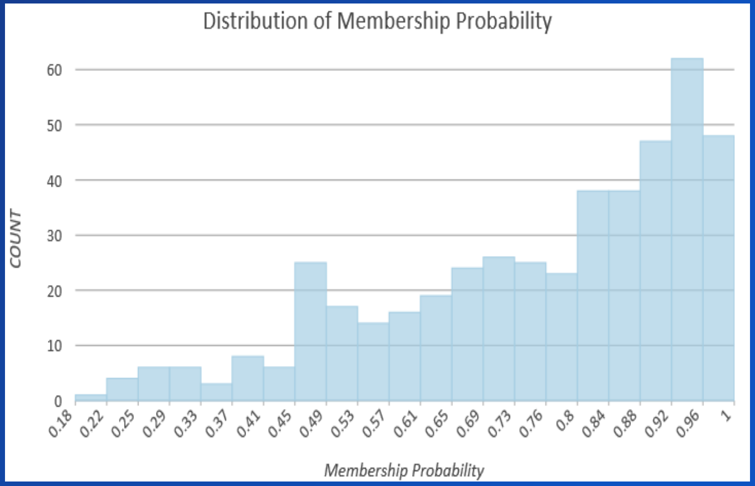
No. of clusters = 3


39°25'
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
N
Kilometers
10
5
0
oEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID,
IGN, and the GIS User Community



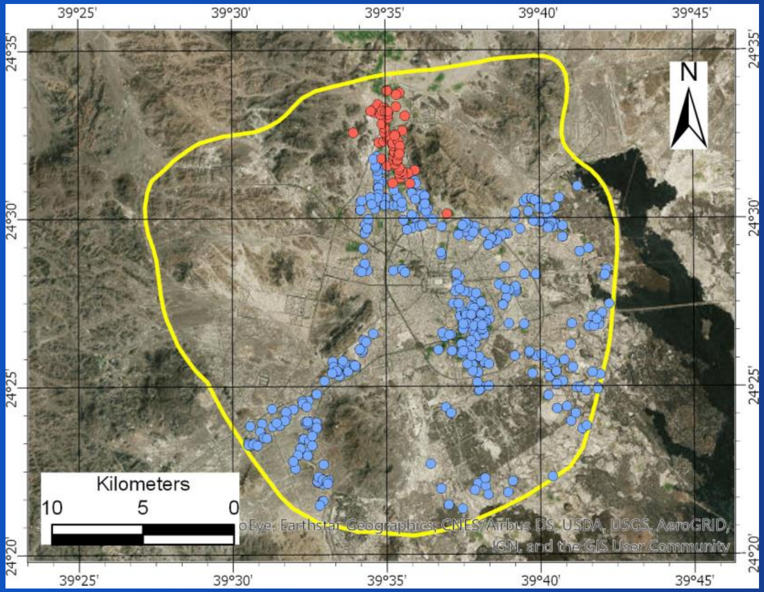
No. of clusters = 7

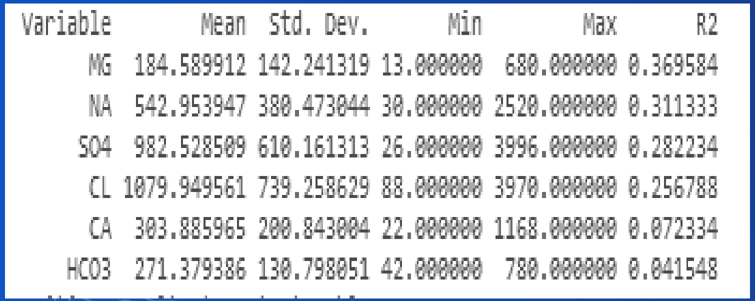

39°25'
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS,
AeroGRID, IGN, and the GIS User Community
N
Kilometers
10
5
0



Results & Discussions (Group (B) scenarios)

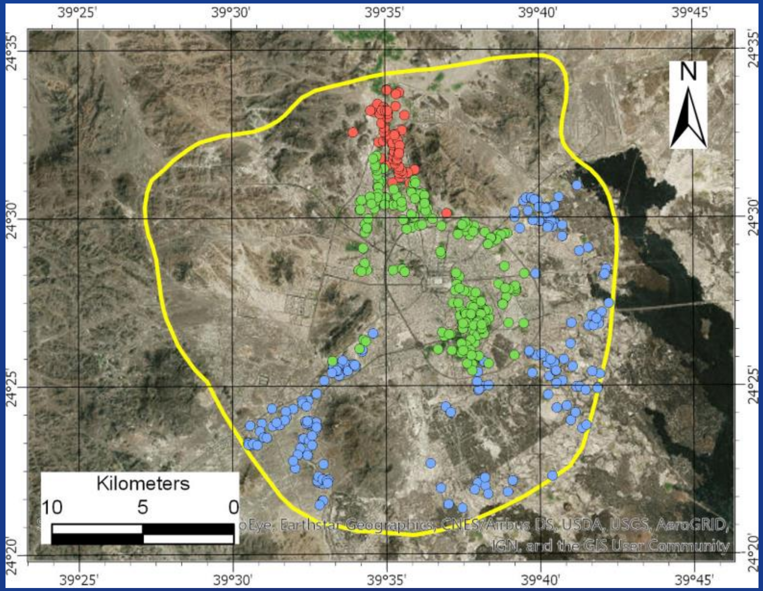
No. of clusters = 2

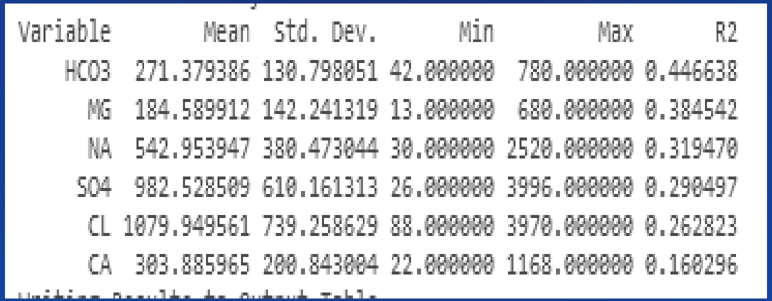

39°25'
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
N
Kilometers
10
5
0
oEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community



```
Variable        Mean   Std. Dev.        Min          Max       R2
      MG  184.589912 142.241319 13.000000  680.000000 0.369584
      NA  542.953947 380.473044 30.000000 2520.000000 0.311333
     SO4  982.528509 610.161313 26.000000 3996.000000 0.282234
      CL 1079.949561 739.258629 88.000000 3970.000000 0.256788
      CA  303.885965 200.843004 22.000000 1168.000000 0.072334
    HCO3  271.379386 130.798051 42.000000  780.000000 0.041548
```

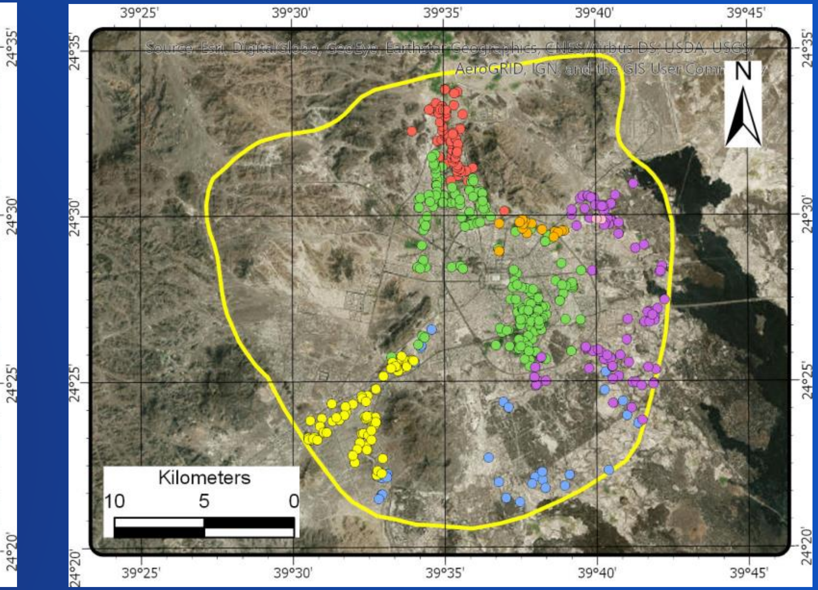
No. of clusters = 3

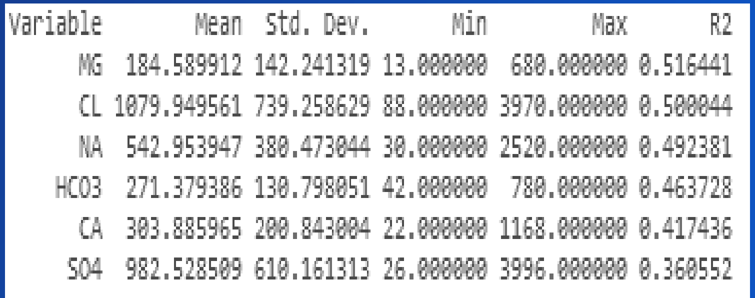

39°25'
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
N
Kilometers
10
5
0
oEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community



```
Variable        Mean   Std. Dev.        Min          Max       R2
    HCO3  271.379386 130.798051 42.000000  780.000000 0.446638
      MG  184.589912 142.241319 13.000000  680.000000 0.384542
      NA  542.953947 380.473044 30.000000 2520.000000 0.319470
     SO4  982.528509 610.161313 26.000000 3996.000000 0.290497
      CL 1079.949561 739.258629 88.000000 3970.000000 0.262823
      CA  303.885965 200.843004 22.000000 1168.000000 0.160296
Writing Results to Output Table
```

No. of clusters = 7


39°25'
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
N
Kilometers
10
5
0

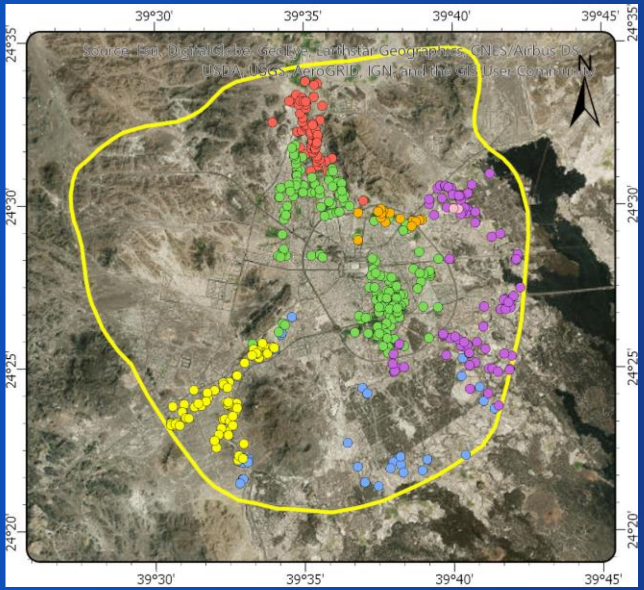


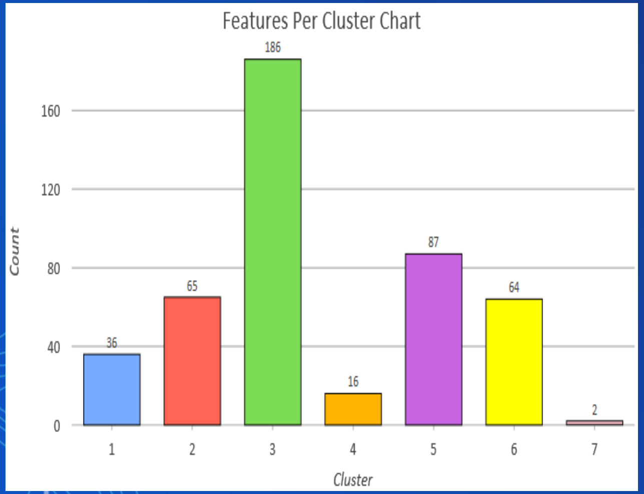
```
Variable        Mean   Std. Dev.        Min          Max       R2
      MG  184.589912 142.241319 13.000000  680.000000 0.516441
      CL 1079.949561 739.258629 88.000000 3970.000000 0.500044
      NA  542.953947 380.473044 30.000000 2520.000000 0.492381
    HCO3  271.379386 130.798051 42.000000  780.000000 0.463728
      CA  303.885965 200.843004 22.000000 1168.000000 0.417436
     SO4  982.528509 610.161313 26.000000 3996.000000 0.360552
```

Results & Discussions (Group (C) scenarios)

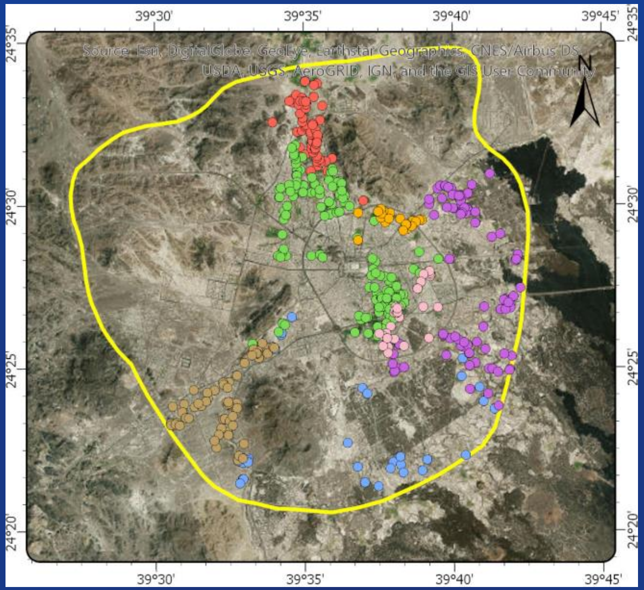
Assuming the optimum No. of cluster = 7

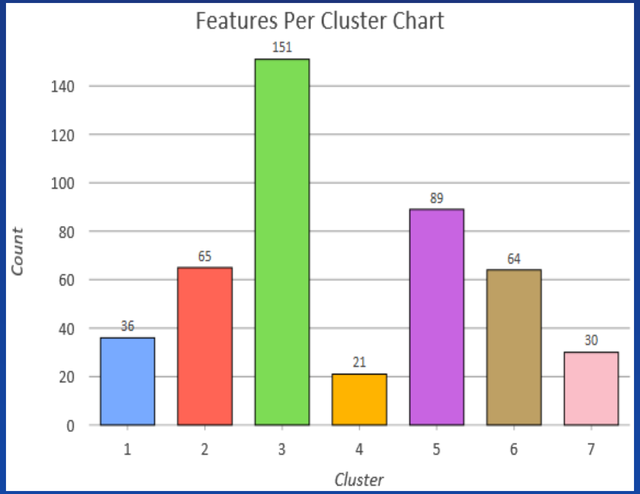
7 clusters, no min. No. of features


39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
N

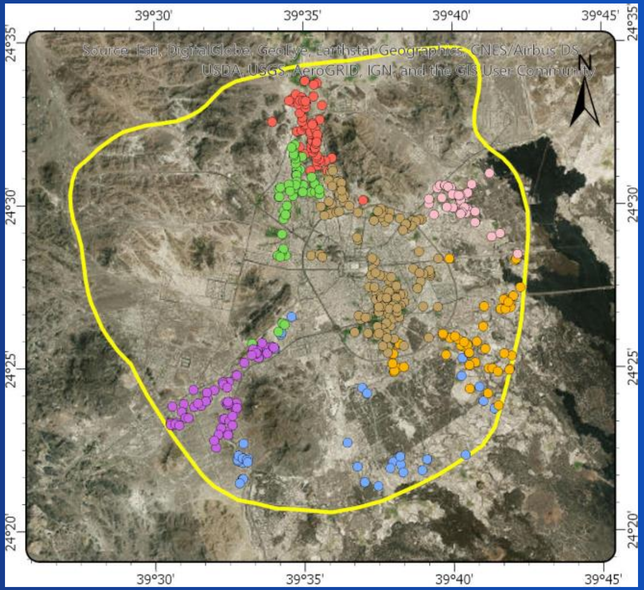


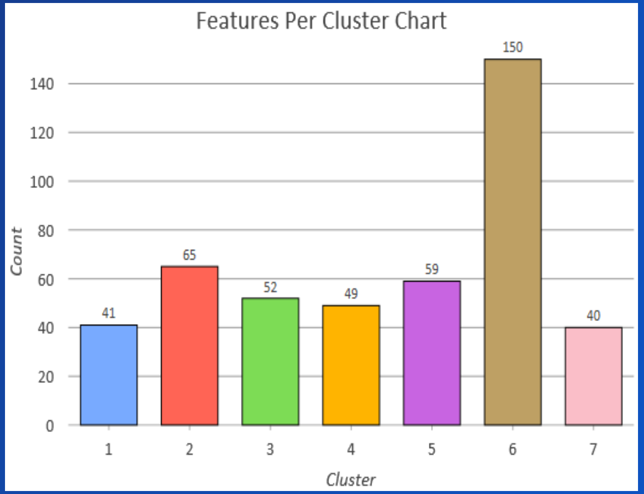
7 clusters, min. No. of features = 20


39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
N



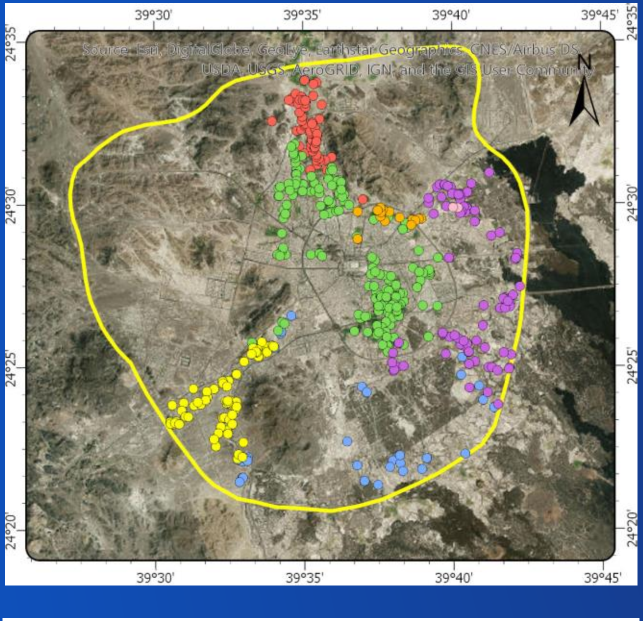
7 clusters, min. No. of features = 40


39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
N

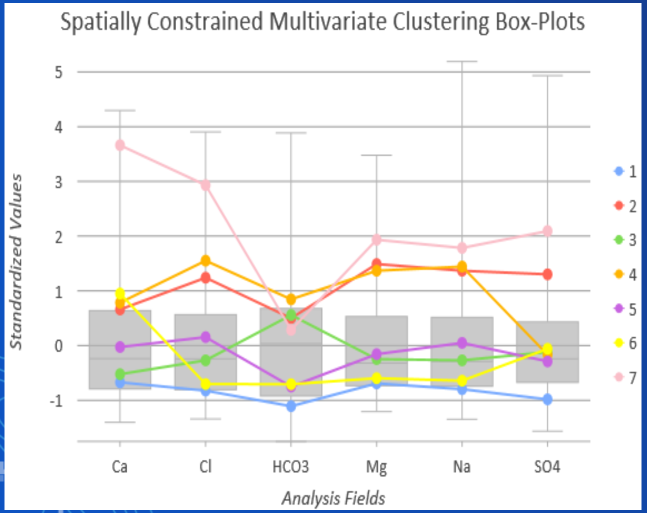


Results & Discussions (Group (C) scenarios)

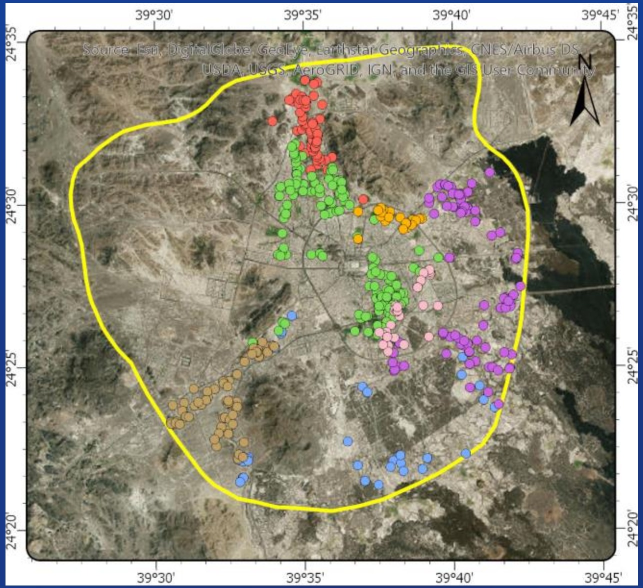
7 clusters, no min. No. of features



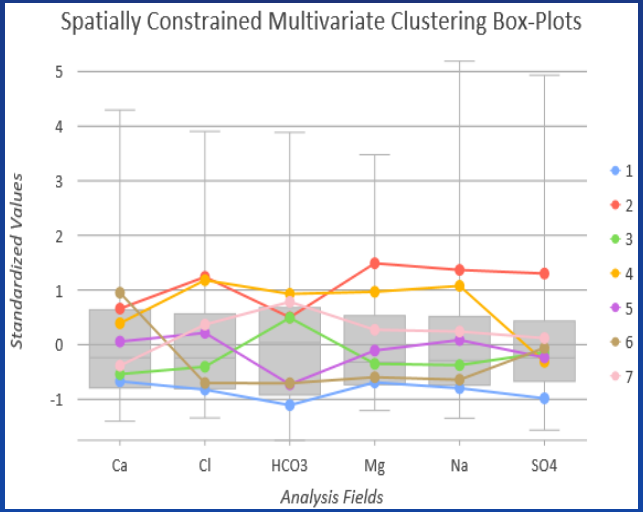
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
N



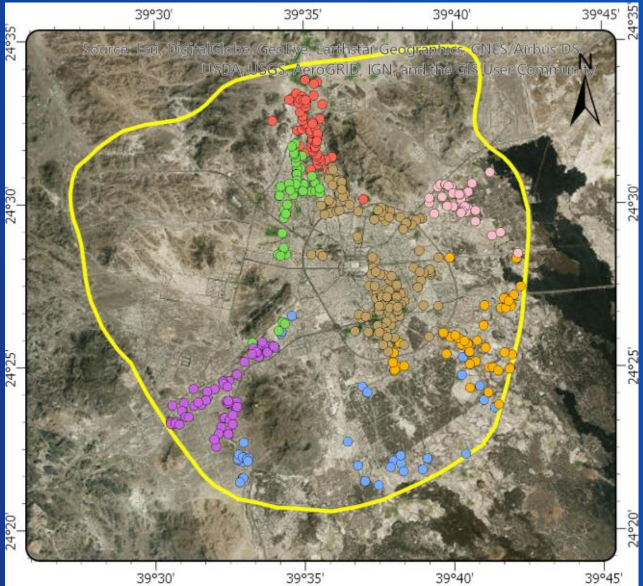
7 clusters, min. No. of features = 20



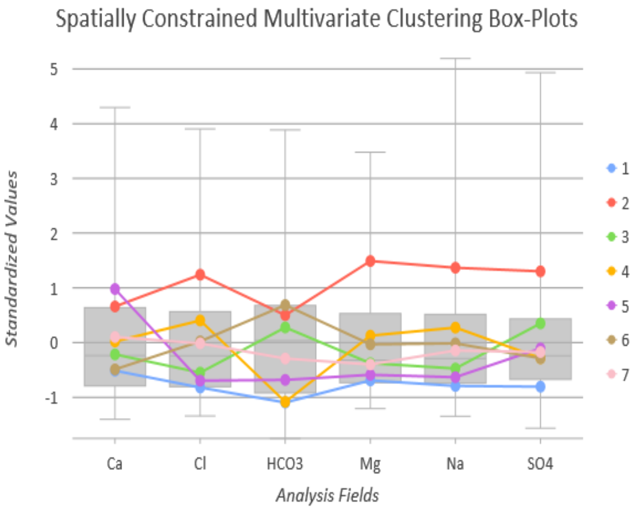
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
N



7 clusters, min. No. of features = 40

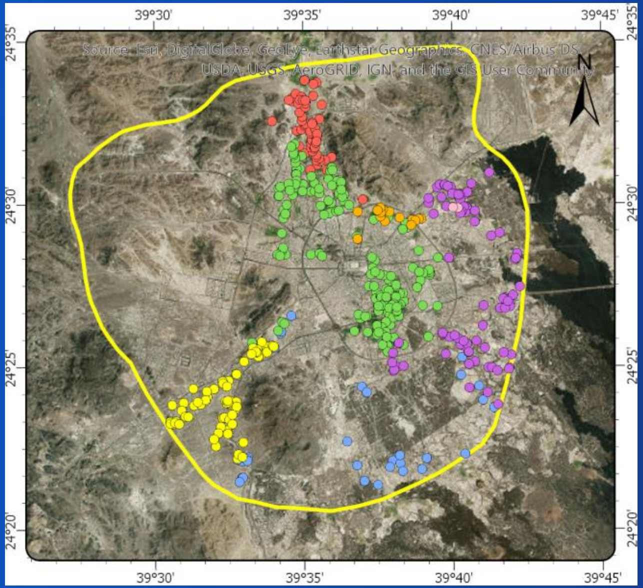


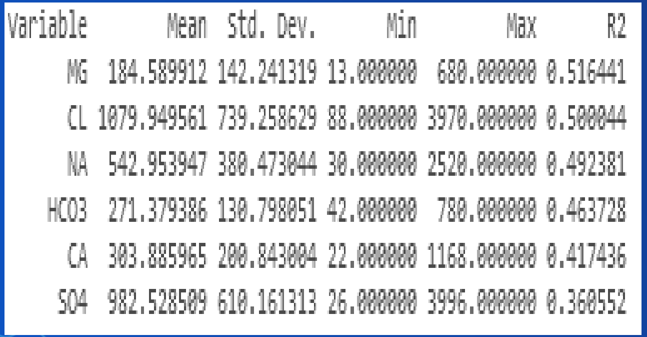
39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
N



Results & Discussions (Group (C) scenarios)

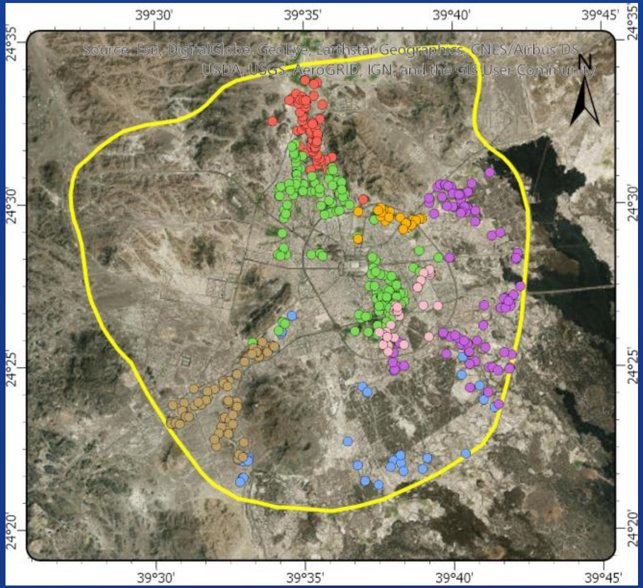
7 clusters, no min. No. of features

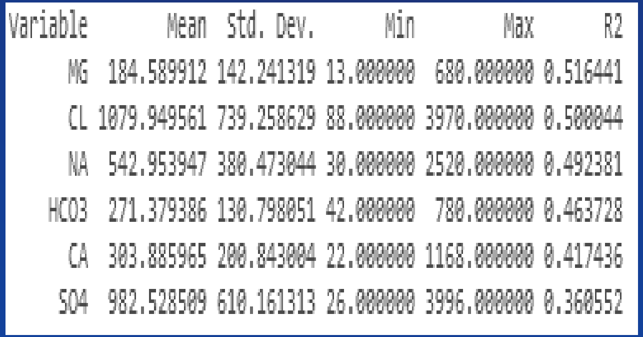

39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
N



Variable	Mean	Std. Dev.	Min	Max	R2
MG	184.589912	142.241319	13.000000	680.000000	0.516441
CL	1079.949561	739.258629	88.000000	3970.000000	0.500044
NA	542.953947	380.473044	30.000000	2520.000000	0.492381
HCO3	271.379386	130.798051	42.000000	780.000000	0.463728
CA	303.885965	200.843004	22.000000	1168.000000	0.417436
SO4	982.528509	610.161313	26.000000	3996.000000	0.360552

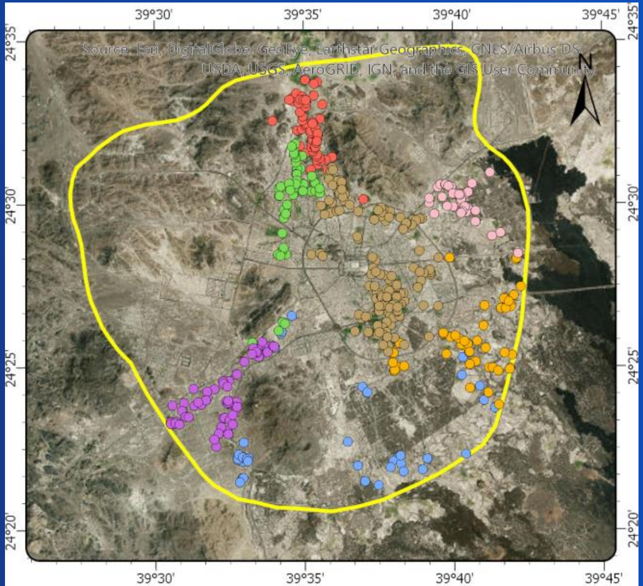
7 clusters, min. No. of features = 20

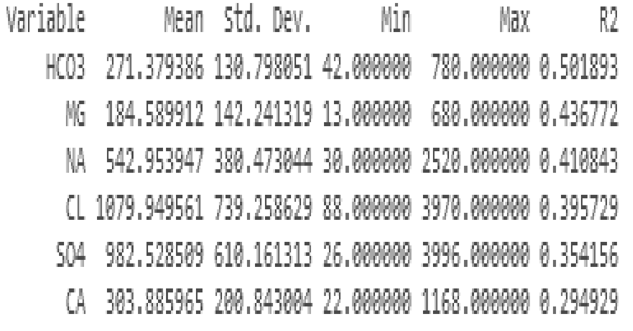

39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
N



Variable	Mean	Std. Dev.	Min	Max	R2
MG	184.589912	142.241319	13.000000	680.000000	0.516441
CL	1079.949561	739.258629	88.000000	3970.000000	0.500044
NA	542.953947	380.473044	30.000000	2520.000000	0.492381
HCO3	271.379386	130.798051	42.000000	780.000000	0.463728
CA	303.885965	200.843004	22.000000	1168.000000	0.417436
SO4	982.528509	610.161313	26.000000	3996.000000	0.360552

7 clusters, min. No. of features = 40


39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
N

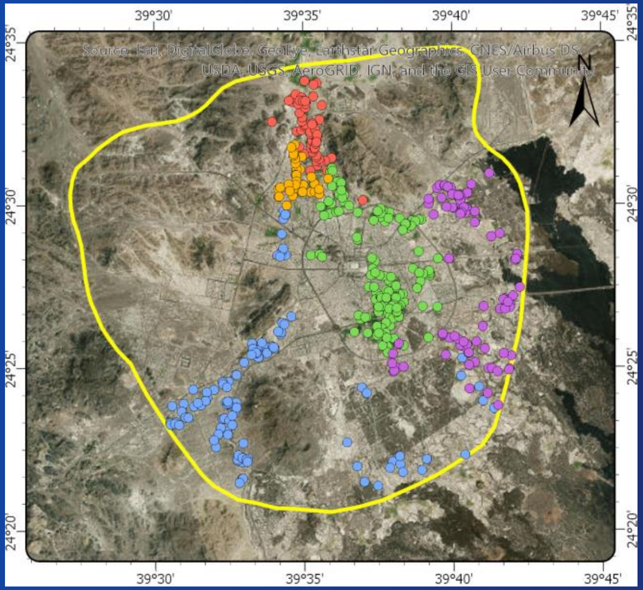


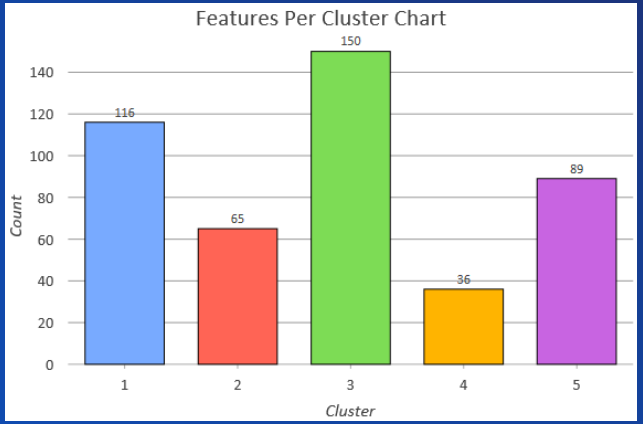
Variable	Mean	Std. Dev.	Min	Max	R2
HCO3	271.379386	130.798051	42.000000	780.000000	0.501893
MG	184.589912	142.241319	13.000000	680.000000	0.436772
NA	542.953947	380.473044	30.000000	2520.000000	0.410843
CL	1079.949561	739.258629	88.000000	3970.000000	0.395729
SO4	982.528509	610.161313	26.000000	3996.000000	0.354156
CA	303.885965	200.843004	22.000000	1168.000000	0.294929

Results & Discussions (Group (D) scenarios)

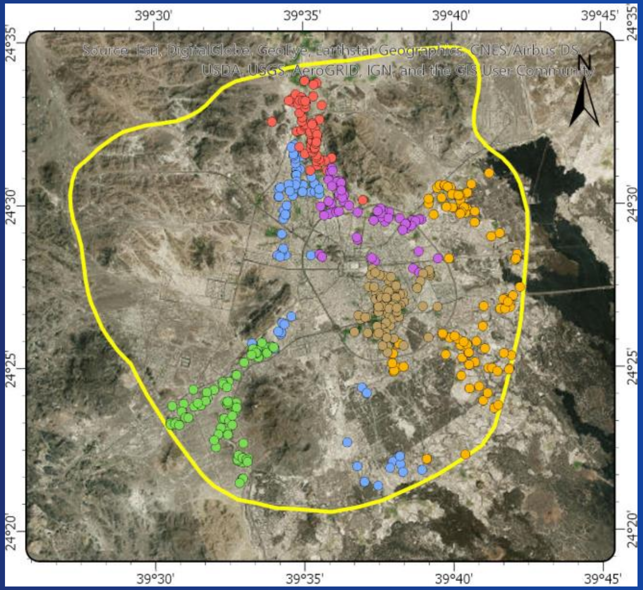
Specifying the margins (boundaries) of the No. of features per cluster

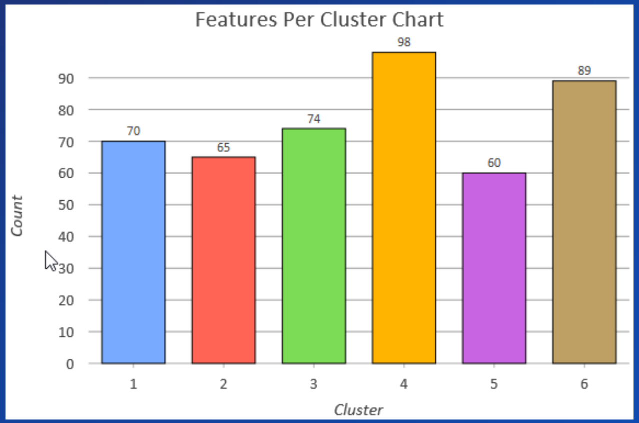
Min. No. of features/cluster = 25
Max. No. of features/cluster = 150


39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
N
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community



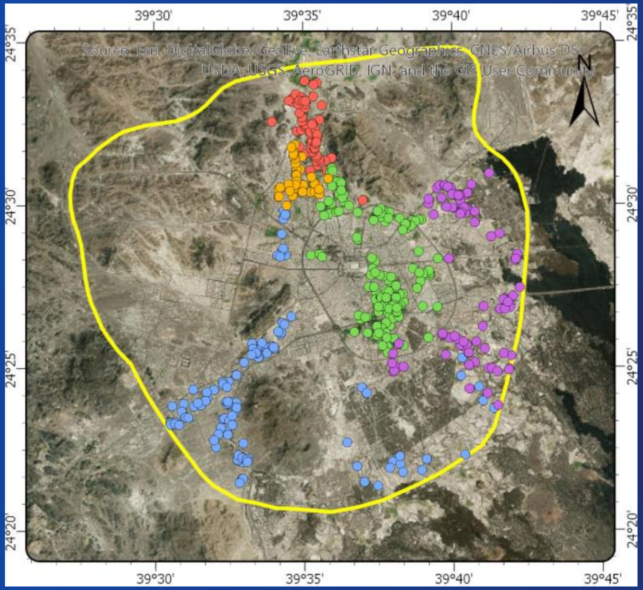
Min. No. of features/cluster = 50
Max. No. of features/cluster = 100

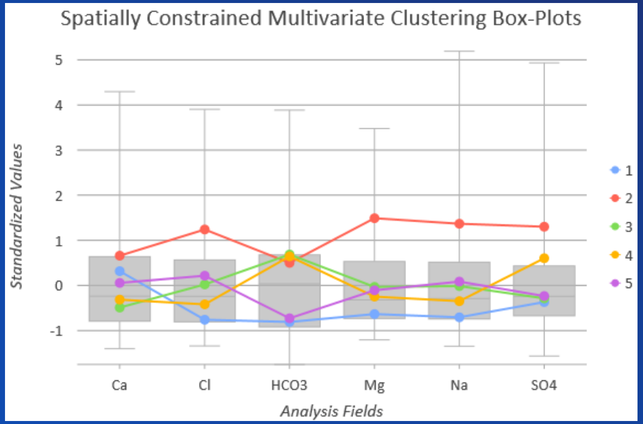

39°30'
39°35'
39°40'
39°45'
24°35'
24°30'
24°25'
24°20'
N
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community



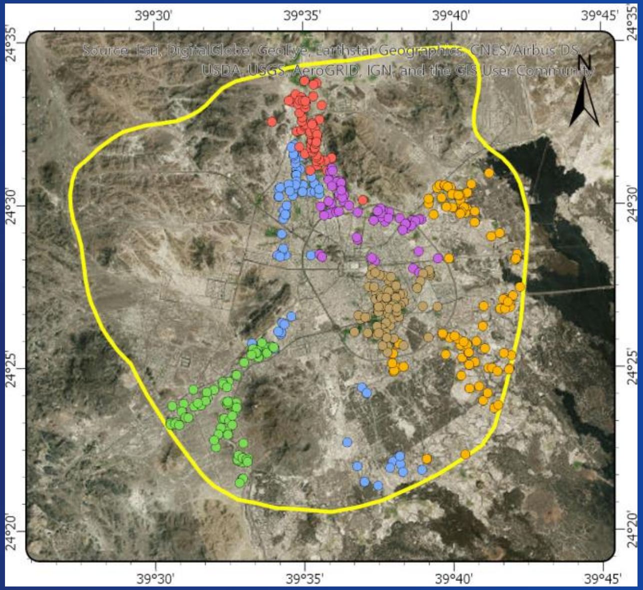
Results & Discussions (Group (D) scenarios)

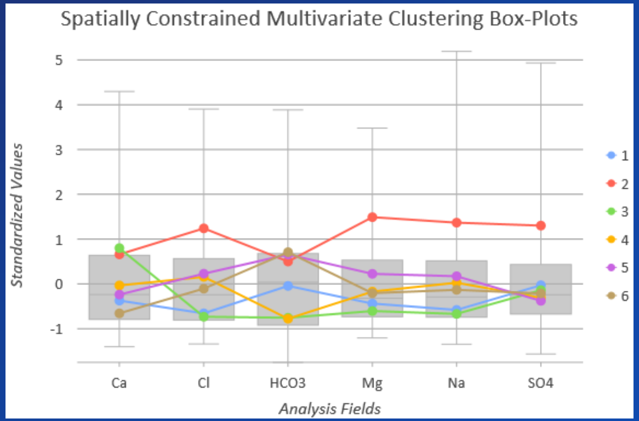
Min. No. of features/cluster = 25
Max. No. of features/cluster = 150


Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community


Spatially Constrained Multivariate Clustering Box-Plots

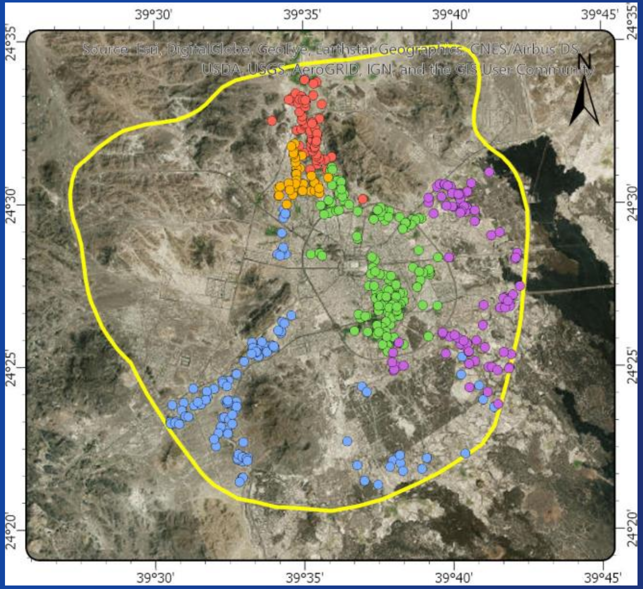
Min. No. of features/cluster = 50
Max. No. of features/cluster = 100


Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

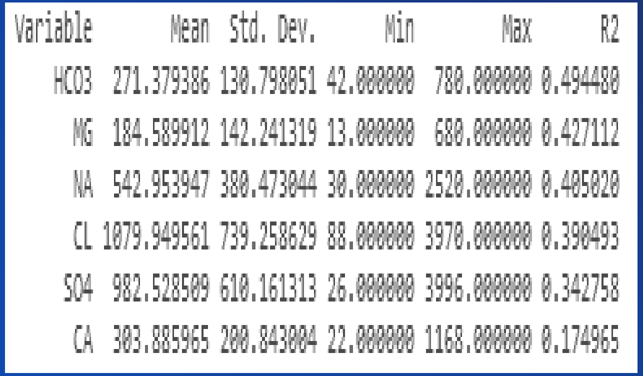

Spatially Constrained Multivariate Clustering Box-Plots

Results & Discussions (Group (D) scenarios)

Min. No. of features/cluster = 25
Max. No. of features/cluster = 150

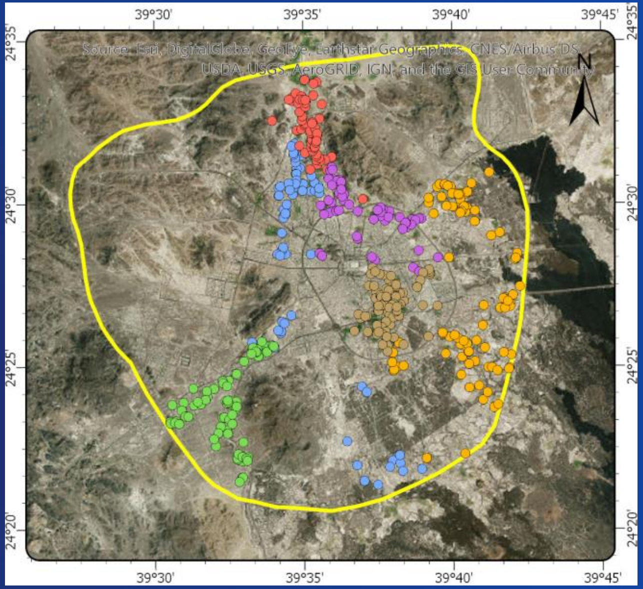


Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

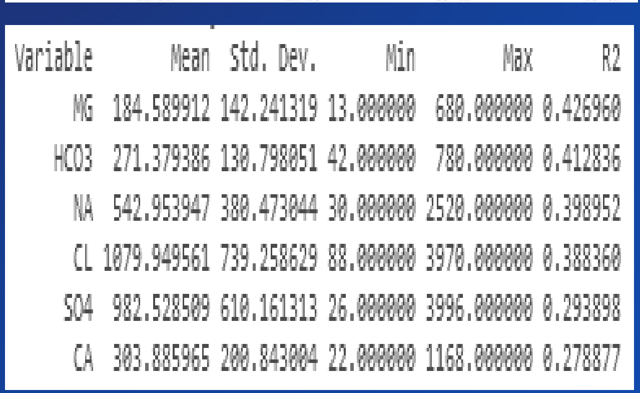


Variable	Mean	Std. Dev.	Min	Max	R2
HCO3	271.379386	130.798051	42.000000	780.000000	0.494480
MG	184.589912	142.241319	13.000000	680.000000	0.427112
NA	542.953947	380.473044	30.000000	2520.000000	0.405020
CL	1079.949561	739.258629	88.000000	3970.000000	0.390493
SO4	982.528509	610.161313	26.000000	3996.000000	0.342758
CA	303.885965	200.843004	22.000000	1168.000000	0.174965

Min. No. of features/cluster = 50
Max. No. of features/cluster = 100

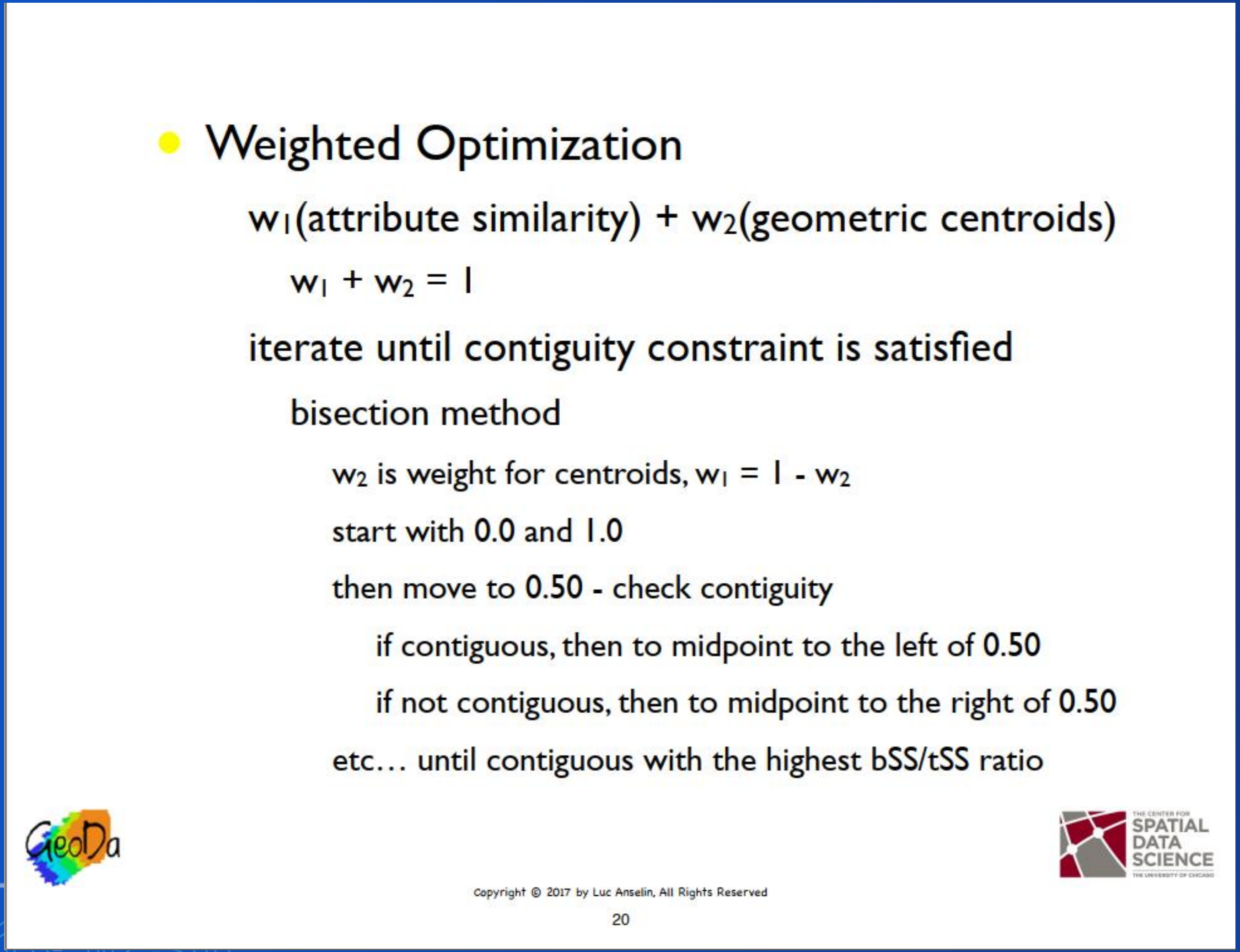


Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community



Variable	Mean	Std. Dev.	Min	Max	R2
MG	184.589912	142.241319	13.000000	680.000000	0.426960
HCO3	271.379386	130.798051	42.000000	780.000000	0.412836
NA	542.953947	380.473044	30.000000	2520.000000	0.398952
CL	1079.949561	739.258629	88.000000	3970.000000	0.388360
SO4	982.528509	610.161313	26.000000	3996.000000	0.293898
CA	303.885965	200.843004	22.000000	1168.000000	0.278877

Results & Discussions (more spatial control using spatial weights)



- Weighted Optimization

  w_1(attribute similarity) + w_2(geometric centroids)

  $w_1 + w_2 = 1$

  iterate until contiguity constraint is satisfied

  bisection method

  w_2 is weight for centroids, $w_1 = 1 - w_2$

  start with 0.0 and 1.0

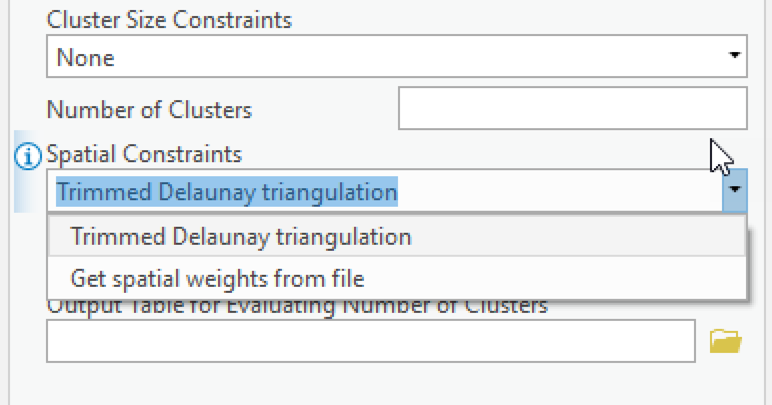
  then move to 0.50 - check contiguity

  if contiguous, then to midpoint to the left of 0.50

  if not contiguous, then to midpoint to the right of 0.50

  etc… until contiguous with the highest bSS/tSS ratio

GeoDa
THE CENTER FOR SPATIAL DATA SCIENCE THE UNIVERSITY OF CHICAGO



Cluster Size Constraints: None

Number of Clusters:

Spatial Constraints: Trimmed Delaunay triangulation

Trimmed Delaunay triangulation
Get spatial weights from file
Output Table for Evaluating Number of Clusters

Results & Discussions (Cluster results evaluation)

- Higher Goodness of fit index is better
- B = between-cluster sum of square error (SSE) – need to be maximized
- W = within-cluster sum of square error (SSE)
- K = the number of clusters
- N = the number of features (observation)

$$Goodness\ of\ fit\ idex = \frac{B/(k-1)}{W/(n-k)}$$

Conclusion

- SCMC method in ArcGIS Pro found to be powerful tool for spatial clustering with many options and functionalities.
- SCMC method as many Clustering methods needs full understanding of the data used.
- One of the best scenarios is sub-dividing the city based on GW quality into three zones which are; Upper zone with good quality, city center zone with moderate quality, and lower (downstream) zone with low quality.
- The results of this study will be beneficial not only for the farmers but also for the local government, environmental agencies and investors in agriculture.

References and Resources

- **Spatially Constrained Multivariate Clustering**
  - https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/spatially-constrained-multivariate-clustering.htm
- **How Spatially Constrained Multivariate Clustering works**
  - https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/how-spatially-constrained-multivariate-clustering-works.htm
- **Determining Number of Clusters in One Picture**
  - https://www.datasciencecentral.com/profiles/blogs/determining-number-of-clusters-in-one-picture
- **Assunção, R. M., Neves, M. C., Câmara, G., & da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. International Journal of Geographical Information Science, 20(7), 797-811.**
- **Duque, J. C., Ramos, R., & Suriñach, J. (2007). Supervised regionalization methods: A survey. *International Regional Science Review*, *30*(3), 195-220.**
- ***Luc Anselin, (2017),* Cluster Analysis (3) *Spatially Constrained Clustering Methods,***
  - https://geodacenter.github.io/workbook/8_spatial_clusters/lab8.html
- **Kassambara, A. (2017). *Practical guide to cluster analysis in R: unsupervised machine learning* (Vol. 1). STHDA.**
- **Fischer, M. M., & Getis, A. (Eds.). (2009). *Handbook of applied spatial analysis: software tools, methods and applications*. Springer Science & Business Media.**

Thank you