UNIVERSITY OF NORTH DAKOTA

CLUSTERING ALGORITHMS FOR STREAMING OPENPDC DATA SETS

ANUPAM MUKHERJEE & RANGANATH VALLAKATI

DEPARTMENT OF ELECTRICAL ENGINEERING

UNIVERSITY OF NORTH DAKOTA

ADVISOR: DR. PRAKASH RANGANATHAN

2015 GPA USER'S FORUM AND TUTORIAL, AUGUST 4 & 5, 2015

GRID PROTECTION ALLIANCE

This Research acknowledges ND EPSCoR (UND0014140), the Office of RD&C (21418-4010-02294), and the UND Graduate School for the grant support.

OUTLINE OF THE PRESENTATION

- Introduction : Need for Situational Awareness of Smart-grid
- Proposed Situational Awareness Framework
- Development of User Interface for openPDC
- Data Visualization
- Data Clustering
  - DBSCAN Clustering
  - k-means Clustering
  - Multi-Tier k-means Clustering
- Results and Discussions
- Conclusion

NEED FOR SITUATIONAL AWARENESS OF SMART GRID

Blackout Events	Affected Areas	Cause
August 14, 2003 – Northeast Blackout.	Northeastern and Mid-western United States and Canadian province of Ontario. **People affected** – 55 million.	Software bug in the alarm system.
July 31, 2012 – Blackout in India.	22 states and union territories. **People affected** – 600 million.	Collapse of Northern and Eastern grids.
December 22, 2013 – Major ice-storm caused power failure.	Ontario to the maritime province in the far east and Michigan **People affected** – 1.1 million.	Ice storm
March 31, 2015 – Black-out, caused by technical failure, affected about 90% of Turkey.	90% of Turkey. **People affected** – 70 million.	Probable cyber attack.

INTEGRATED SOFTWARE SUITE (ISS)

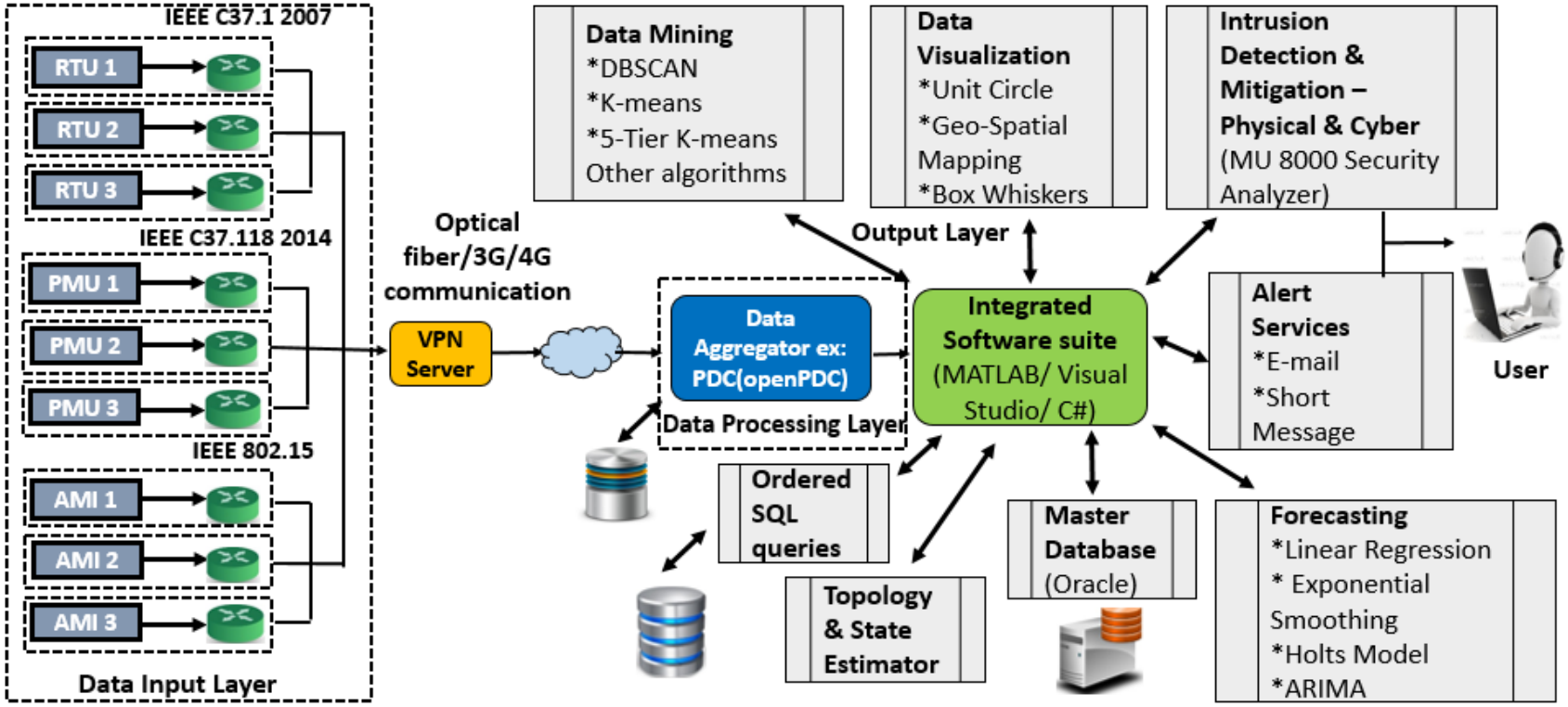


Figure 1: Integrated Software Suite

UNIVERSITY OF NORTH DAKOTA

DEVELOPMENT OF USER INTERFACE

OpenPDC functions by receiving data broadcasted by a PMU and concentrating it, enabling archiving, rebroadcasting, and analysis of the phasor data. It provides around 30 samples per second.

Functionalities:

- E-mail Alarm
- Short Message Service alarm
- Location based monitoring

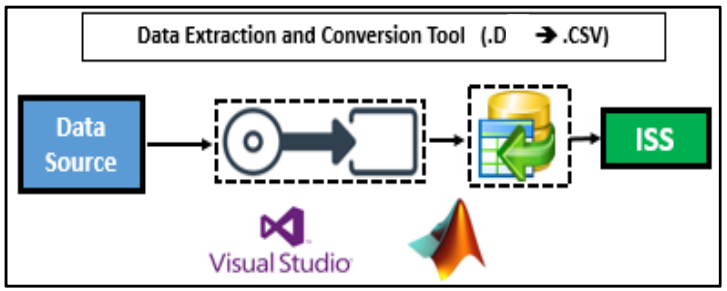


Figure 2: Data Processing Layer

Methodologies

- C# used for all coding
- Visual Studio 2012 IDE used for development
- External libraries utilized:
  - Grid Solutions Framework
  - Google Static Maps API
  - .NET Framework 4.5

UNIVERSITY OF NORTH DAKOTA

ALERT SYSTEMS DEVELOPED FOR OPENPDC

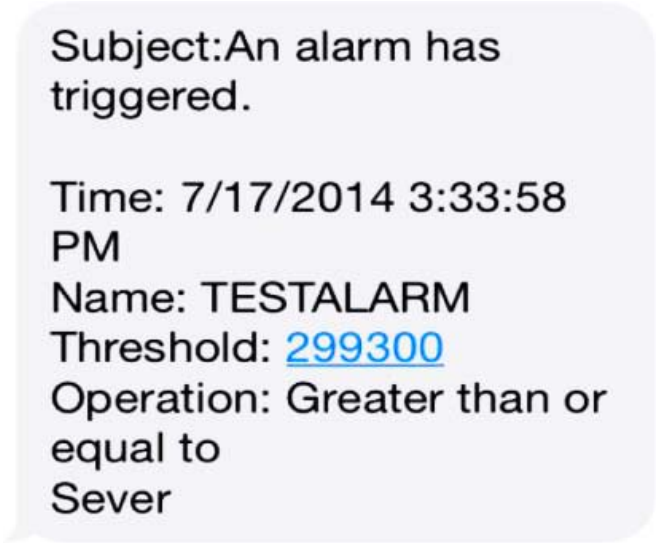


Figure 3: Short Message Service Alarm

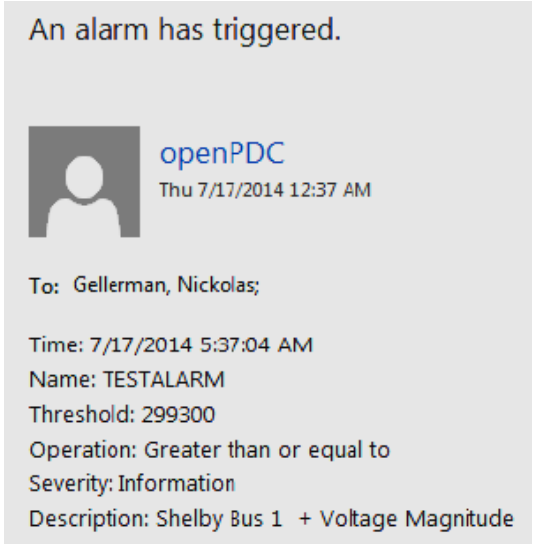


Figure 4: E-mail Alarm

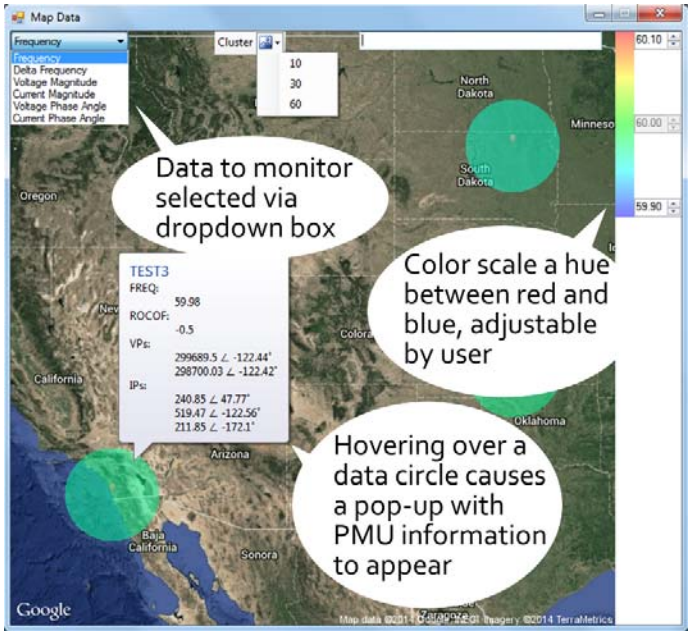


Figure 5: Location Based Monitoring System

UNIVERSITY OF NORTH DAKOTA

DBSCAN CLUSTERING SCHEME

- DBSCAN is a density-based clustering algorithm that divides large regions with sufficiently high density into multiple clusters.
- DBSCAN considers two parameters as input excluding the data. They are ε (Eps) and $MinPts$. Minpts are the minimum number of points that are required to form a core, and eps is the distance threshold from center of the cluster to its circumference of the cluster

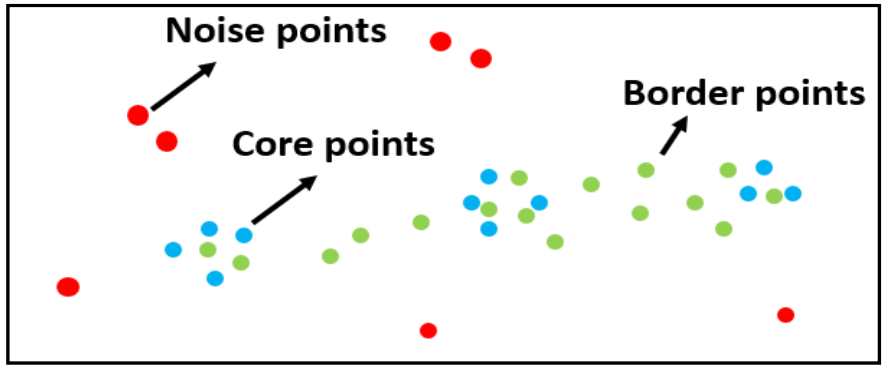


Figure 6: DBSCAN Cluster Formation

K-MEANS CLUSTERING SCHEME

- The k-means technique is a well-known and popular algorithm which was first proposed by Lloyd.
- Here, each cluster is represented by an adaptively changing centroid (also called a cluster center), starting from some initial values

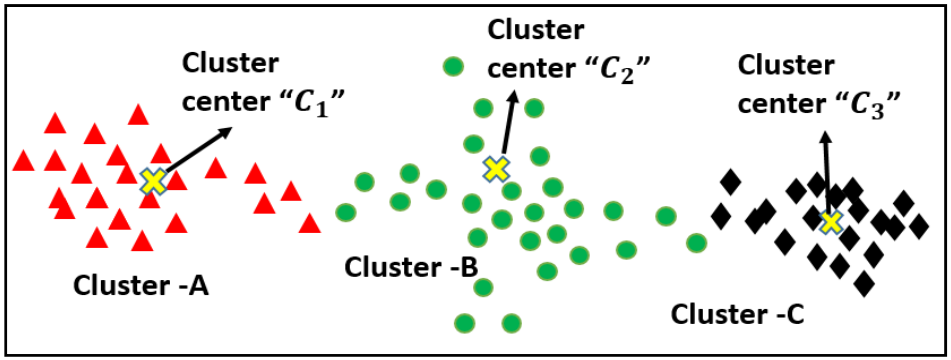


Figure 7: k-means Clustering

MULTI-TIER K-MEANS CLUSTERING SCHEME

- This paper presents a different version of k-means which we refer as multi-tier k-means clustering tailored for power system data sets.
- The proposed approach dynamically forms clusters from 1 to 5 clusters depending on the data thresholds and fault type. They are : High Noise, High Border, Good Data, Low Border, and Low Noise points
- Capable of clearly distinguish the good, bad and the noisy data with the threshold inputs from the operator.

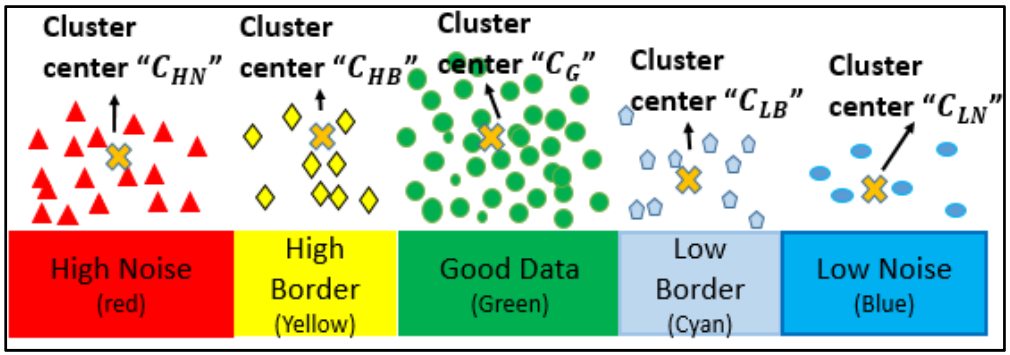


Figure 8: Multi-tier k-means Cluster Formation

UNIVERSITY OF NORTH DAKOTA

DATA CLUSTERING SCHEME

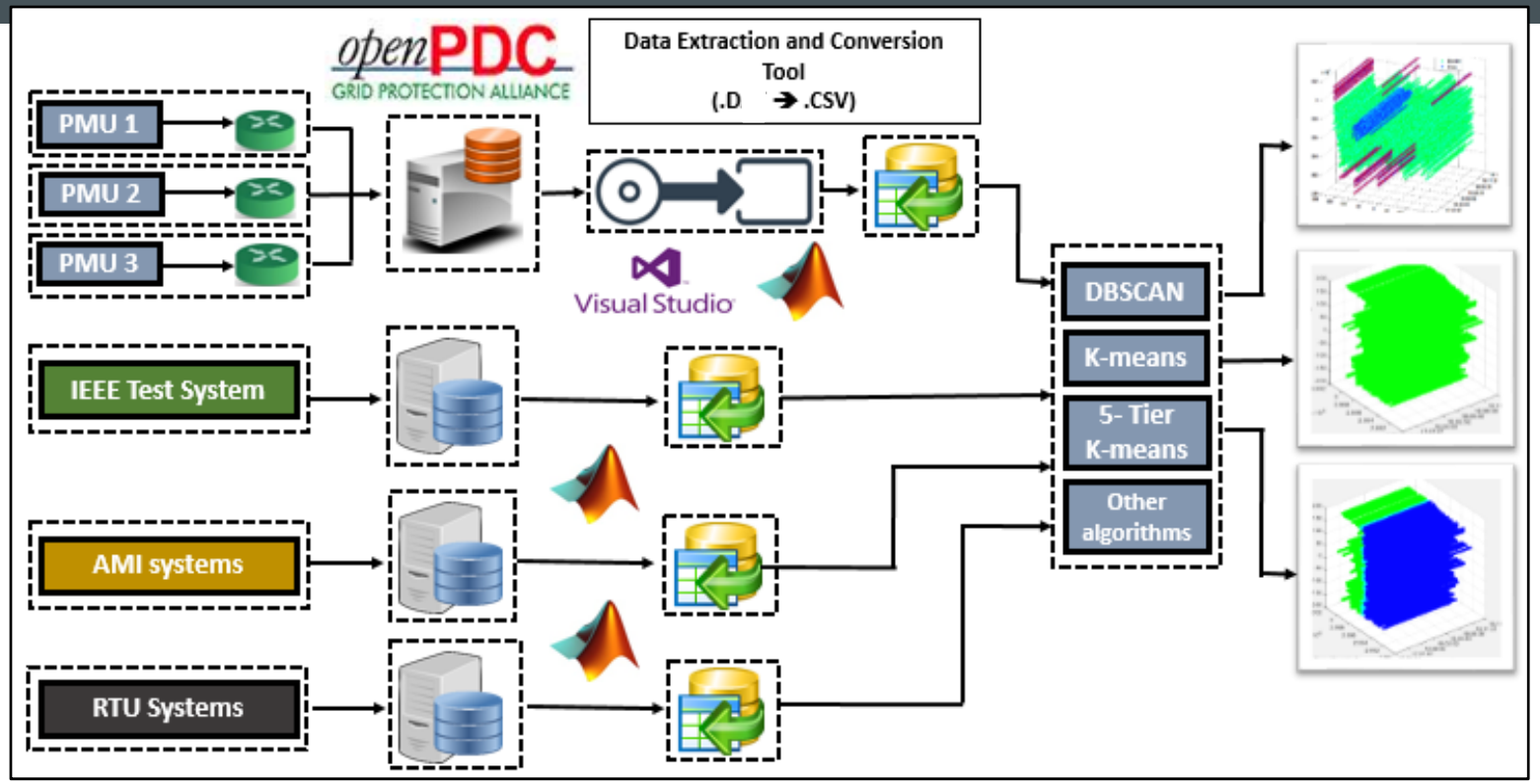


Figure 9: Smart Grid Data Management Framework (SGDMF)

UNIVERSITY OF NORTH DAKOTA

RESULTS AND DISCUSSIONS

- Data Visualization
  - Box Plot
  - Circle Representation
- Data Clustering
  - DBSCAN Clustering
  - k-means Clustering
  - Multi-Tier k-means Clustering

DATA VISUALIZATION

- As phase angle varies between -π to +π (0 to 360 degrees) and the magnitudes are above 0 for the electric signals, unit circle representation is ideal smart-grid data
- The "Box Whiskers" is a statistical tool that allows observing a time-series data with minimum and maximum values in the series, standard deviations, mean and median values.

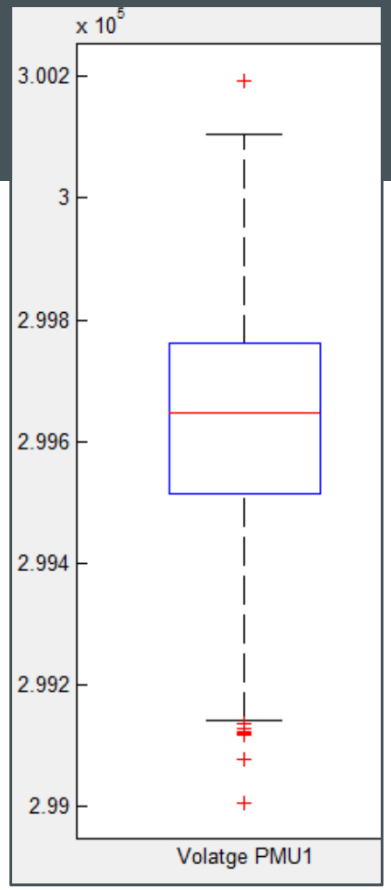


Figure 10: Box Whisker Representation of openPDC Voltage Data

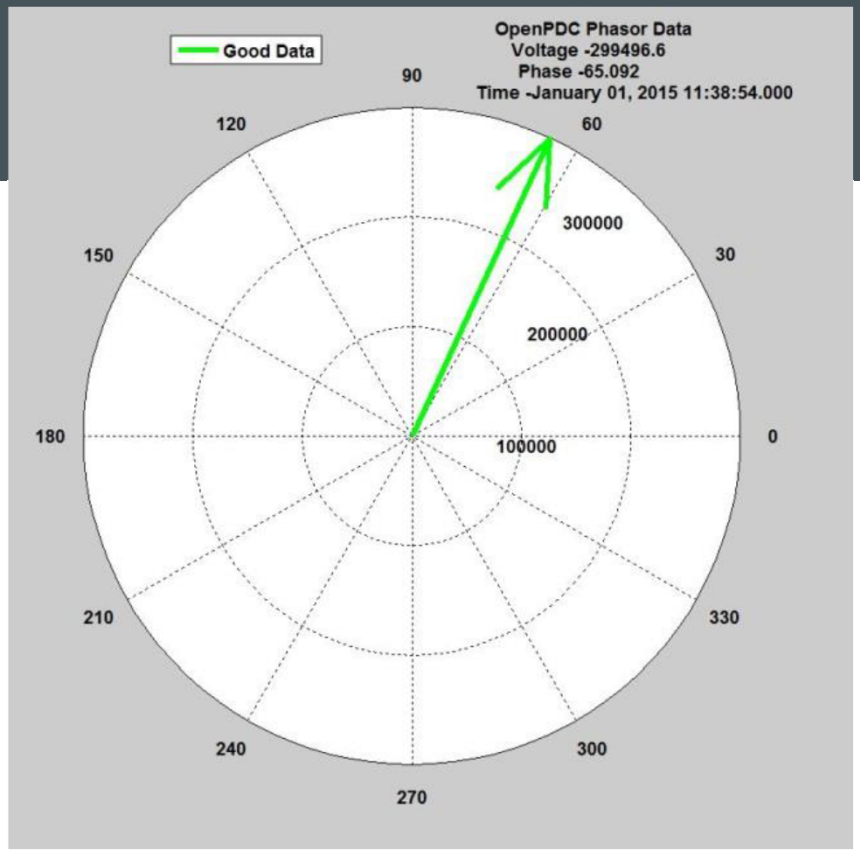
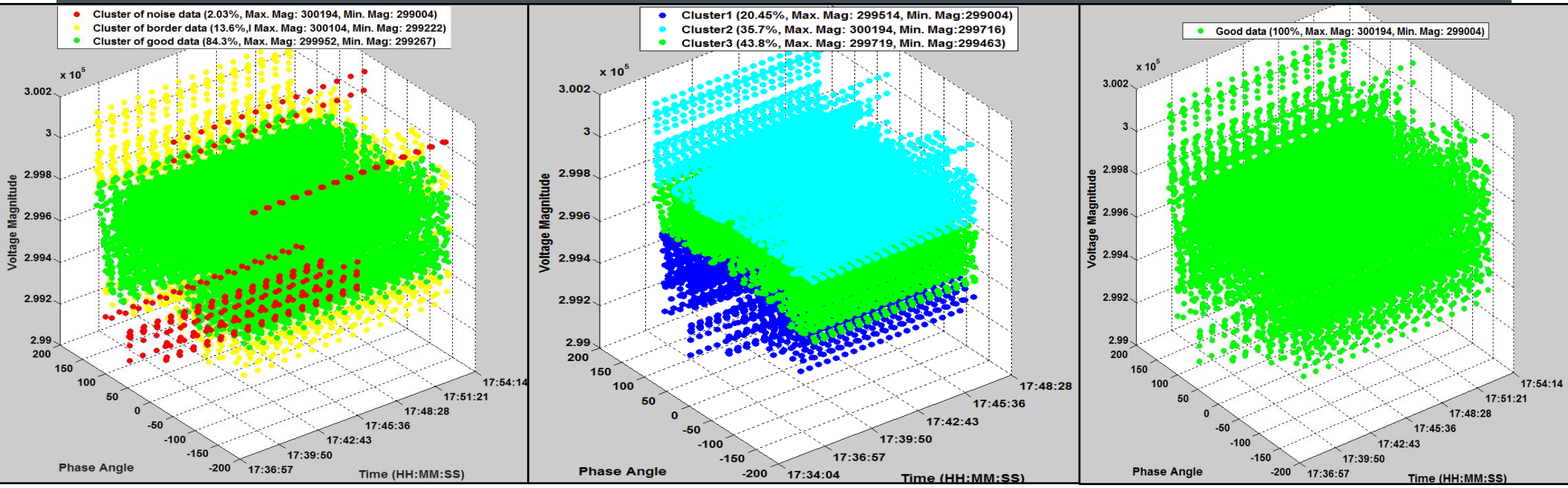


Figure 11: Circle Representation of openPDC Voltage Data

UNIVERSITY OF NORTH DAKOTA

TEST SCENARIO: STEADY-STATE CONDITION



(a) (b) (c)

Figure 12: Clustering Schemes Applied on openPDC data under steady state condition (a) DBSCAN, (b) k-means, (c) Multi-Tier

UNIVERSITY OF NORTH DAKOTA

TEST SCENARIO: HEAVY LOAD (HIGH DEMAND) CONDITION

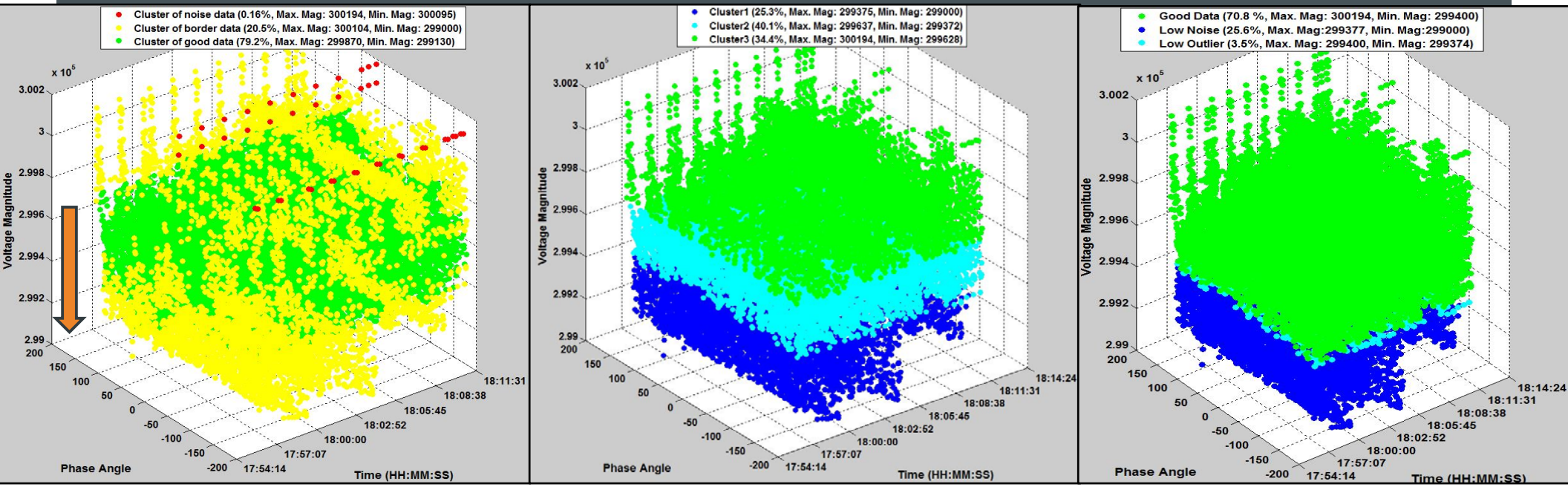


Figure 13: Clustering Schemes Applied on openPDC data under Heavy Load Conditions (a) DBSCAN, (b) k-means, (c) Multi-Tier

TEST SCENARIO: LIGHT LOAD (LOW DEMAND) CONDITION

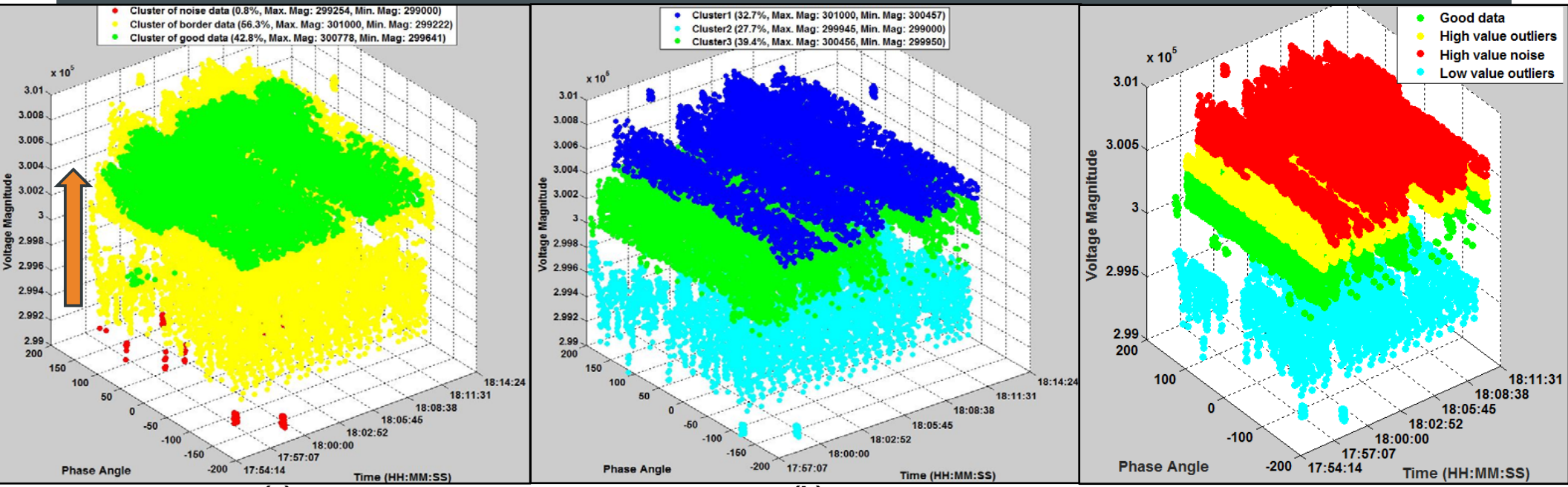
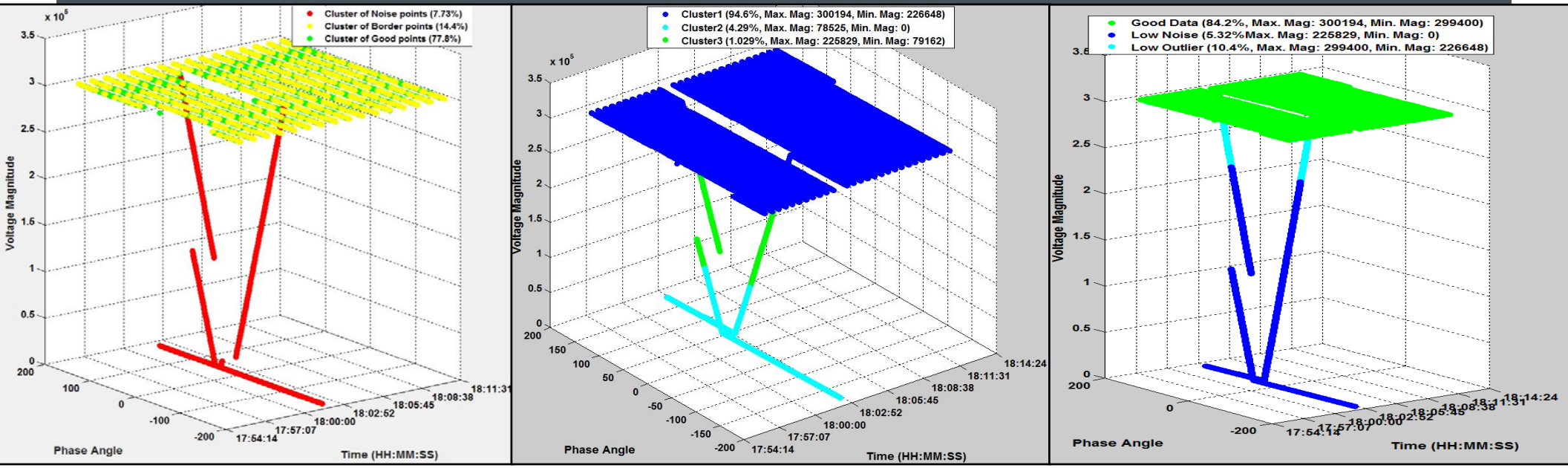


Figure 14: Clustering Schemes Applied on openPDC data under Light Load Conditions (a) DBSCAN, (b) k-means, (c) Multi-Tier

TEST SCENARIO: SLG FAULT CONDITION (SHORT-CIRCUIT)



(a) **(b)** **(c)**

Figure 15: Clustering Schemes Applied on openPDC data Under SLG Fault Conditions (a) DBSCAN, (b) k-means, (c) Multi-Tier

UNIVERSITY OF NORTH DAKOTA

DISTRIBUTION OF DATA POINTS

Load Condition	Noise Points (Red)	Border Points (Yellow)	Core Points (Green)
Normal	0.5	6.3	93.2
Heavy	0.078	8.96	90.5
Light	0.8	56.3	42.8
Fault	7.73	14.4	77.8

Table 1: % distribution of data points with DBSCAN

Load Condition	Cluster 1 (Blue)	Cluster 2 (Cyan)	Cluster 3 (Green)
Normal	27.1	36	36.7
Heavy	25.3	40.1	34.4
Light	32.7	27.7	39.4
Fault	94.6	4.29	1.02

Table 2: % distribution of data points with k-means

Load Condition	Low Noise (Blue)	Low Border (Cyan)	Good Points (Green)	High Border (Yellow)	High Noise (Red)
Normal	0	10.53	89.47	0	0
Heavy	24.7	5.2	70.04	0	0
Light	0	3.3	79.76	16.94	0
Fault	5.32	10.4	84.2	0	0

Table 3: % distribution of data points with multi-tier k-means

- Steady-state condition: Multi-tier k-means performs best.
- Heavy-load condition: DBSCAN performs best.
- Light-load condition: DBSCAN performs best.
- Fault condition: Multi-tier performs the best.

CONCLUSION

- An Integrated Software Suite (ISS) has been developed to apply decision-making data-mining algorithms on time-synchronized synchrophasor data.
- A novel, Multi-Tier variation of the k-means algorithm is presented, and its performance metrics are studied against common clustering techniques to classify and detect bad data, event detection, and alarm service applications.
- A comparative analysis has been carried out between the three data clustering algorithms, DBSCAN, k-means and the Multi-Tier k-means.
- It is believed that such a framework will enable the grid's system operators to utilize novel algorithms in order to enhance situational awareness about the grid. The framework is scalable and suitable for streaming time-series data sets.

FUTURE WORK

- Study application of forecasting algorithms like:
  - Time Series Data Analysis
  - Linear Regression
  - Exponential Smoothing
  - Holt's Model
- Topology based State Estimator
- Intrusion Detection and Mitigation Systems

REFERENCES:

[1] M. Panteli and D. S. Kirschen, "Situation awareness in power systems: Theory, challenges and applications," *Electr. Power Syst. Res.*, vol. 122, pp. 140–151, May 2015.

[2] A. G. Phadke, "Synchronized phasor measurements in power systems," *IEEE Comput. Appl. Power*, vol. 6, no. 2, pp. 10–15, Apr. 1993.

[3] A. J. Phadke, A. G., Thorp, *Synchronized Phasor Measurements and Their Applications*. 2008.

[4] J. Chen and A. Abur, "Placement of PMUs to Enable Bad Data Detection in State Estimation," *IEEE Trans. Power Syst.*, vol. 21, no. 4, pp. 1608–1615, Nov. 2006.

[5] J. Quintero and V. M. Venkatasubramanian, "Oscillation monitoring system based on wide area synchrophasors in power systems," in *2007, iREP Symposium - Bulk Power System Dynamics and Control - VII. Revitalizing Operational Reliability*, 2007, pp. 1–13.

[6] A. Pal, I. Singh, and B. Bhargava, "Stress assessment in power systems and its visualization using synchrophasor based metrics," in *2014 North American Power Symposium (NAPS)*, 2014, pp. 1–6.

[7] A. Pal, J. S. Thorp, T. Khan, and S. S. Young, "Classification Trees for Complex Synchrophasor Data," *Electr. Power Components Syst.*, vol. 41, no. 14, pp. 1381–1396, Oct. 2013.[19]

[8] Z. Li and W. Wu, "Phasor Measurements-Aided Decision Trees for Power System Security Assessment," in *2009 Second International Conference on Information and Computing Science*, 2009, vol. 1, pp. 358–361.

[9] "IEEE Guide for Phasor Data Concentrator Requirements for Power System Protection, Control, and Monitoring," 2013.

[10] S. Ma and J. L. Hellerstein, "Ordering categorical data to improve visualization," *Infovis-99*, no. 1, pp. 1–4, 1999.

[11] T. J. Overbye, D. A. Wiegmann, and A. M. Rich, "Human factors aspects of power system voltage contour visualizations," *IEEE Trans. Power Syst.*, vol. 18, no. 1, pp. 76–82, Feb. 2003.

[12] K. R. Žalik, "An efficient k′-means clustering algorithm," *Pattern Recognit. Lett.*, vol. 29, pp. 1385–1391, 2008.

[13] R. Cordeiro de Amorim and B. Mirkin, "Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering," *Pattern Recognit.*, vol. 45, no. 3, pp. 1061–1075, Mar. 2012.

THANK YOU…

Questions???

UNIVERSITY OF NORTH DAKOTA

K MEANS CLUSTERING SCHEME

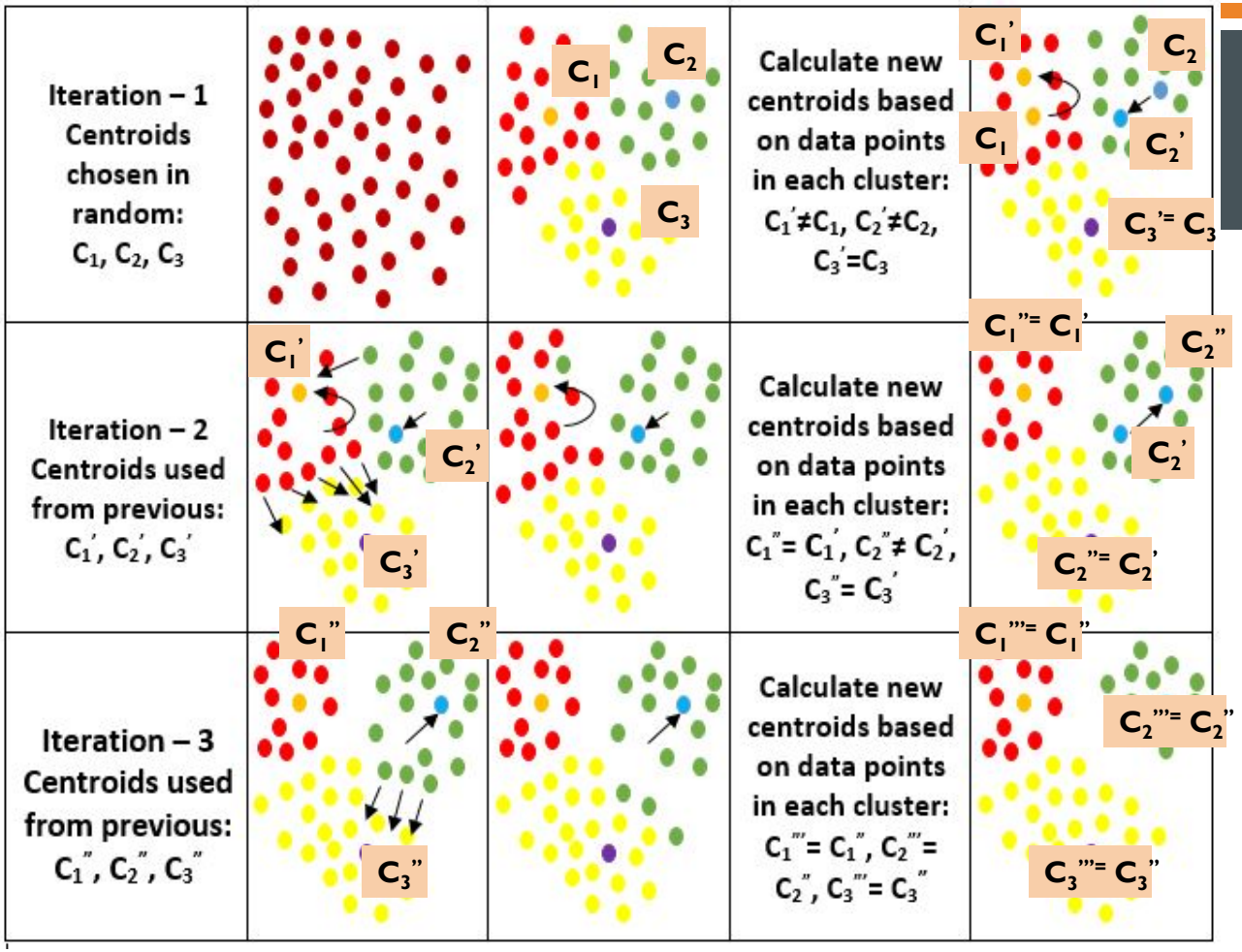
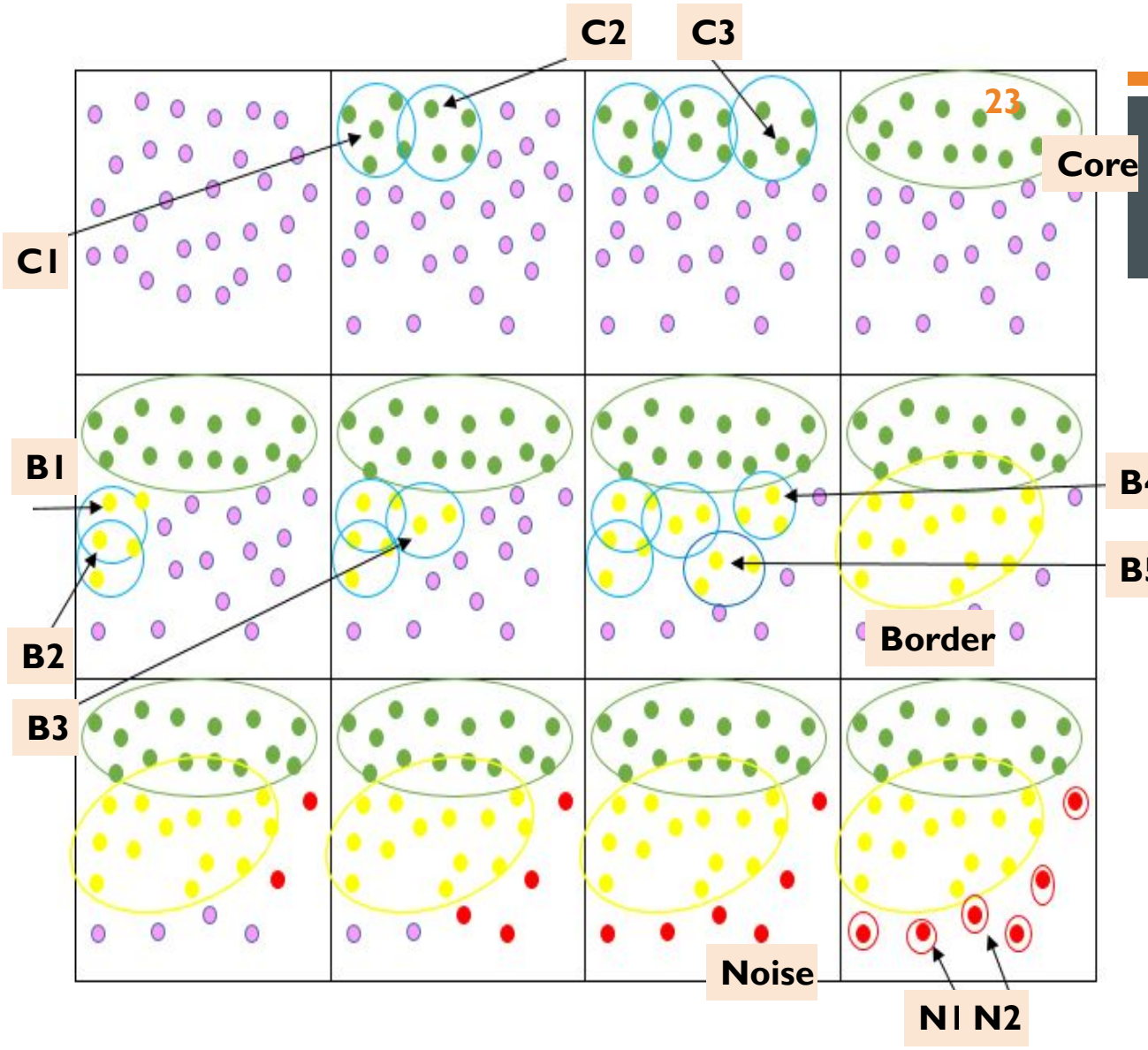


Figure : k-means Cluster formation

Distance Metric used: Euclidean

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

DBSCAN CLUSTERING SCHEME



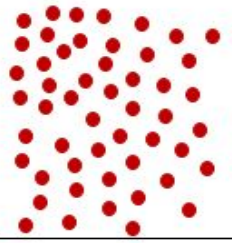
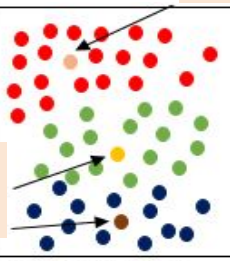
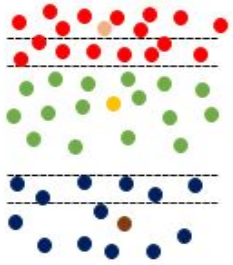
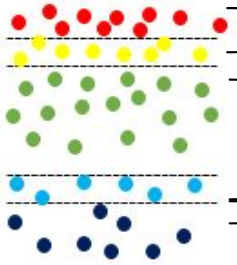
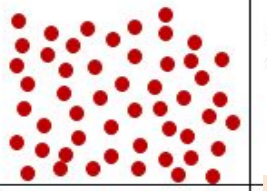
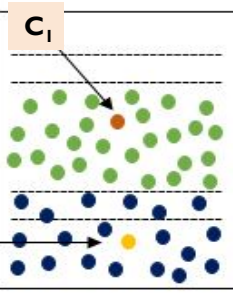
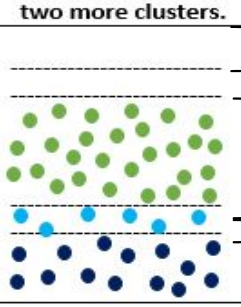
Distance Metric used: Euclidean

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Inputs for the Algorithm

X = Dataset
Eps = Min. distance between two points
D = Min. number of points required to make core

MULTI-TIER K-MEANS CLUSTERING SCHEME

STAGE – 1 Type – I Normal k-means algorithm is applied with 3 clusters and pre-defined values of centroids C_1, C_2, C_3		Running iterations until the dataset is clustered into three groups or less.		Based on the transmission line data, the three clusters are then modified to generate potentially two more clusters.
STAGE – 2 Type - I	Thresholds are setup to divide the dataset into five different cluster regions.		The regions are given a separate color to indicate the level of contingency: High noise(red), High border(yellow), core(green), low border(cyan) and low noise(blue)	
STAGE – 1 Type – II Normal k-means algorithm is applied with 3 clusters and pre-defined values of centroids C_1, C_2, C_3		Running iterations until the dataset is clustered into three groups or less.		Based on the transmission line data, the three clusters are then modified to generate potentially two more clusters.
STAGE – 2 Type - II	Thresholds are setup to divide the dataset into five different cluster regions.		The regions are given a separate color to indicate the level of contingency: High noise(red), High border(yellow), core(green), low border(cyan) and low noise(blue)	

Inputs for the Algorithm

X = Dataset
V = Expected voltage of Transmission line
S = Allowable range for the line voltage to fluctuate

Distance Metric used: Euclidean

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$