Spatio-temporal Data Streaming with Affinity Propagation(DSAP)

Nasrin E. Ivari and Monica Wachowicz

University of New Brunswick

Tamara Agnew and Patricia A.H. Williams

Flinders University

nasrin.eshraghi@unb.ca, monicaw@unb.ca, tamara.agnew@flinders.edu.au, trish.williams@flinders.edu.au

Outline

INTRODUCTION

BACKGROUND

METHODOLOGY

RESULTS

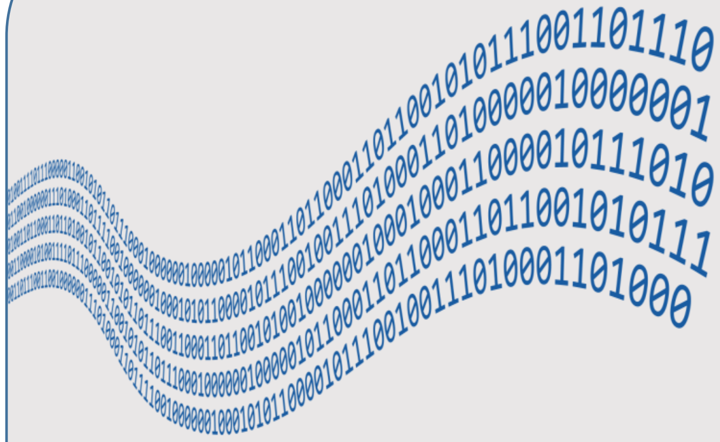
CONCLUSIONS

Introduction

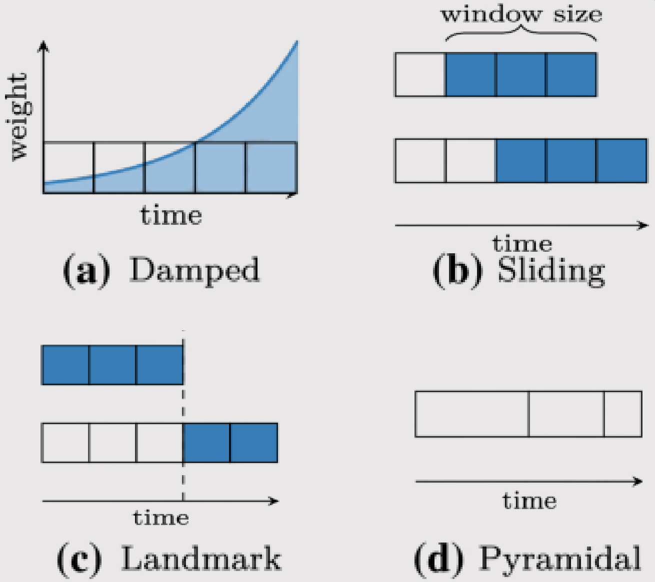
- Spatio-temporal data stream clustering is a growing research field due to the vast amount of continuous georeferenced data streams being generated by IoT devices.

- Very few attempts have been found in the literature for clustering data streams using the Affinity Propagation (AP) algorithm proposed by Dueck and Frey (2007).

Spatio-temporal Data Streaming

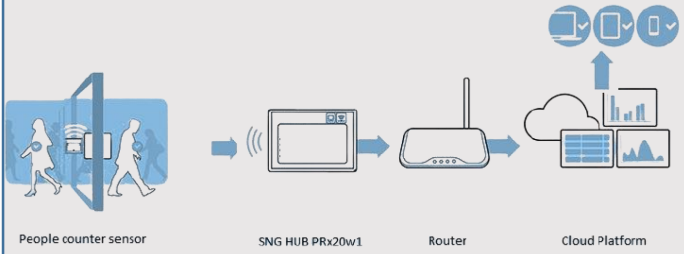


Spatio-temporal data streams are a continuous infinite sequence of data points where each data point contains sensor measurements, their location and a timestamp.



Optimizing data stream representation: An extensive survey on stream clustering algorithms

Different types of time windows can be used to gather spatio-temporal data streams

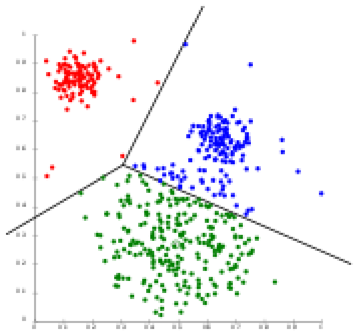


People counting sensors generate a large amount of spatio-temporal data streams.

Data Stream Clustering Methods

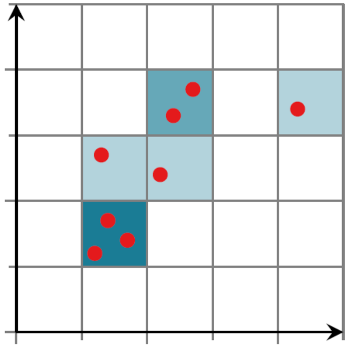
Data Stream Clustering Algorithms

Partitioning-based Stream Method



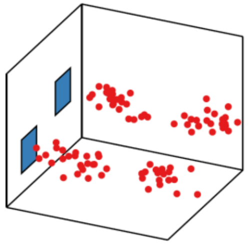
- Objects are grouped into some number of partitions, where each partition represents a cluster.
- Streaming AP, K-means Algorithm

Grid-based Stream Method



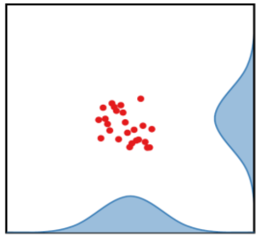
- Data space is split into a finite number of cells which form the grid structure.
- DStream Algorithm

Density-based Stream Method



- They are based on the connection between regions and density functions.
- DBSCAN algorithm

Model-based Stream Method



- They create an empirical model by fitting mathematical models to data.
- CluDiStram Algorithm

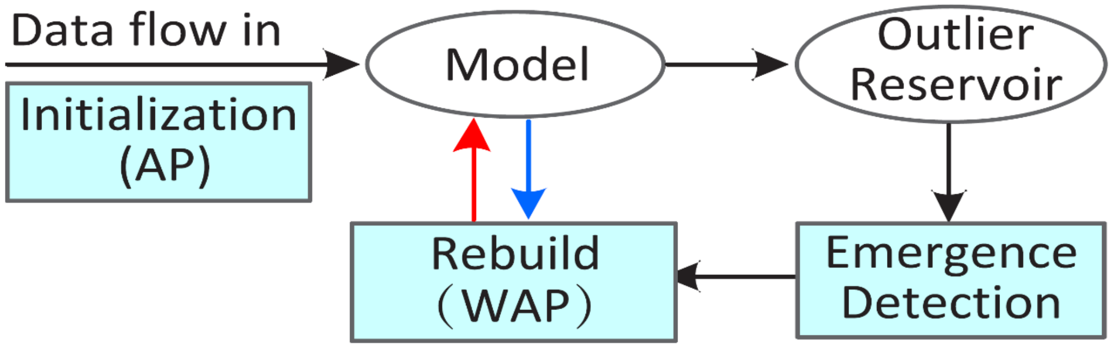
Previous Data Stream AP Clustering Algorithms

STRAP

- STRAP algorithm as an extended AP using sliding time windows for clustering text data streams

ISTRAP

- Affinity propagation with a decay density method using pyramid time window model
- Evaluating: MNIST database contains images of handwritten digits



Research Objectives

Develop a new data stream Affinity Propagation clustering algorithm (DSAP)

Apply the landmark time window model to handle spatio-temporal data streams

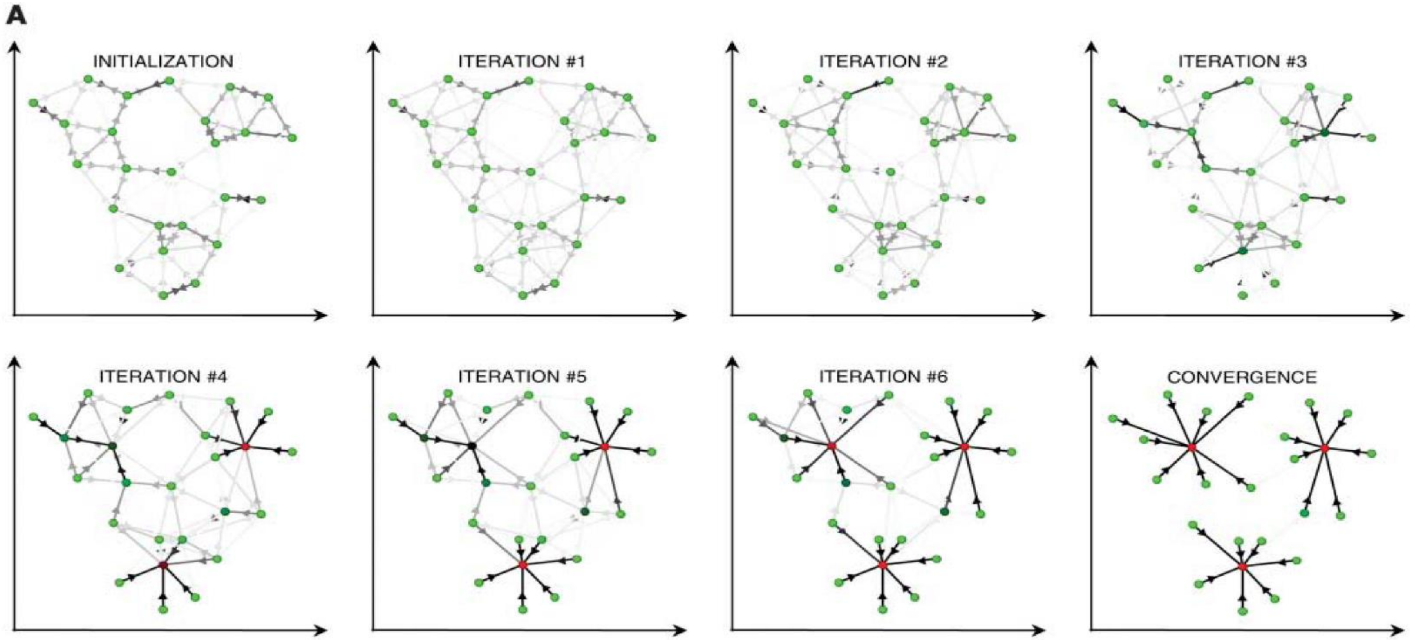
Apply DSAP to discover indoor mobility patterns that can be used to infer life style behavior

Evaluate the performance of DSAP using e-counter data streams

Data Stream Affinity Propagation(DSAP)

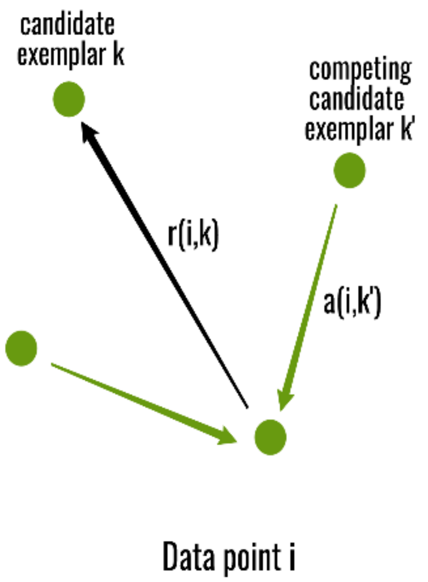
MAIN STEPS

- ❖ Determine the cluster centers
- ❖ Determine the optimal number of clusters for the data
- ❖ Algorithm stops when convergence is achieved.



Algorithm Breakdown: Affinity Propagation,
May 18, 2018 by Ritchie Vink

Matrices computed in DSAP

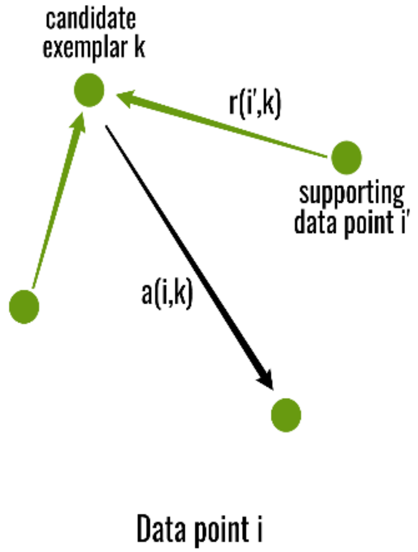


❖Similarity Matrix: Similarity between any instances

❖Responsibility Matrix: How well-suited point k is to be an exemplar for point I

❖Availability Matrix: Contains values that correspond to how available one object is to be an exemplar for another object

❖Criterion Matrix: Sum of the availability matrix and responsibility matrix, the highest criterion value is designated as the exemplar



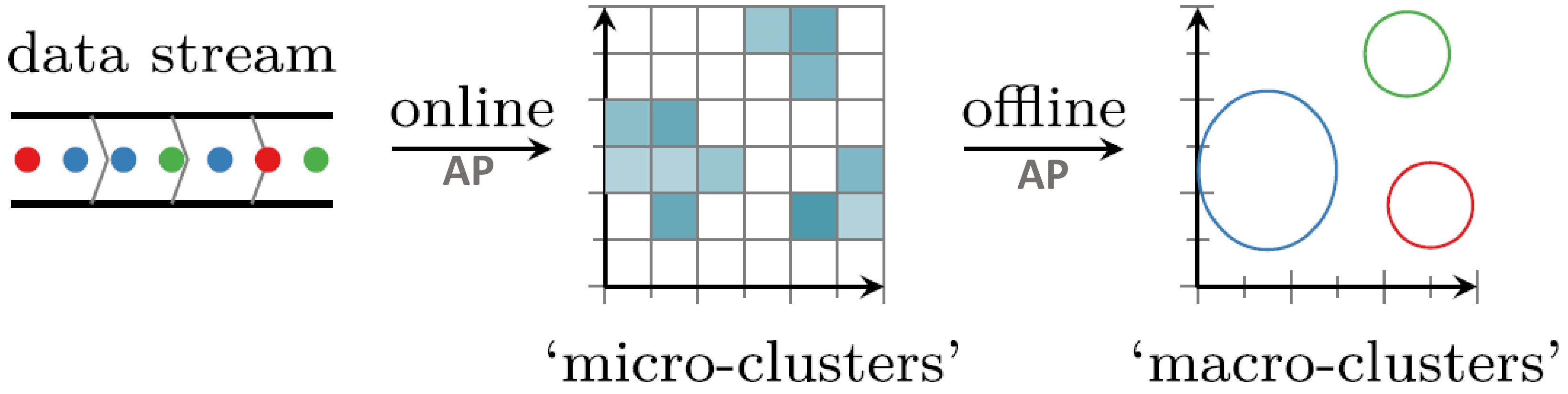
Hyper Parameters used in DSAP

- Damping Factor: It is the extent to which the current data point is maintained relative to incoming data points.
- Preference: For each data point that is more likely to be chosen as an exemplar
- Max_iter: Maximum number of iterations.
- Convergence_iter: the number of iterations with no change in the number of estimated clusters that stop the convergence.

DSAP: Online and Offline Phases

❖ Online phase uses a time window model to capture the data streams and compute micro clusters

❖ Offline phase re-cluster the micro-clusters to generate the macro-clusters after the entire stream data is processed



Carnein, M., & Trautmann, H. (2019). Optimizing data stream representation: An extensive survey on stream clustering algorithms. *Business & Information Systems Engineering*, *61*(3), 277-297.

Experiment

The experiment was conducted to observe the increase in physical activities of people with sedentary life styles using a method of motivation and educational **intervention.**

Dataset: e-counter

6 levels Tonsley building, University of Flinders, Adelaide, Australia

E-counter sensors: wireless infrared people counters digitally record the number of people who goes through the beam (PRx20W1 – PTx20-1)

From March 18 to June 23, 2019

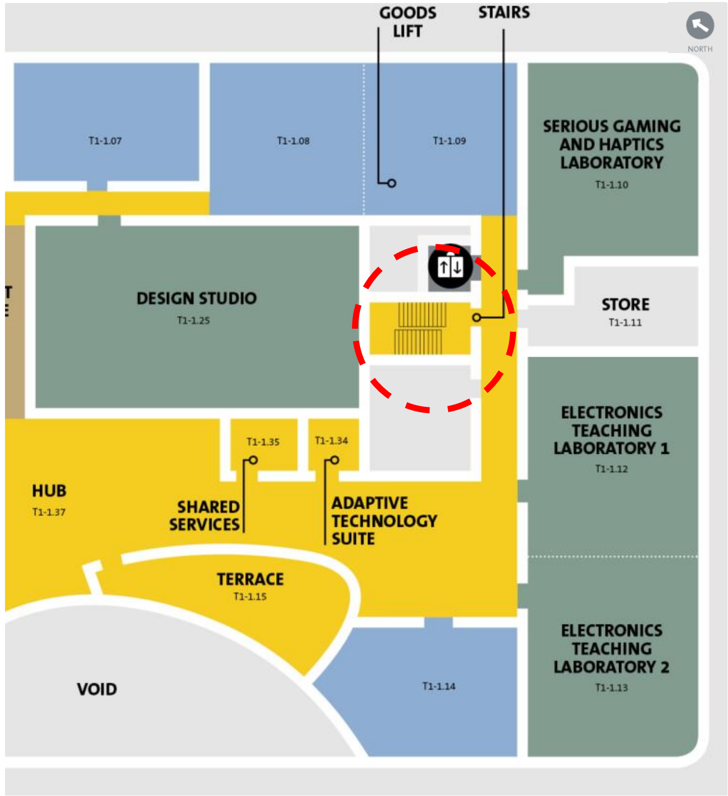
Dataset consists of nine columns: date, time, status, sensor, type, position, location, location code

Event-based dataset

Total records: 668000

E-counters Location in the Tonsley Building

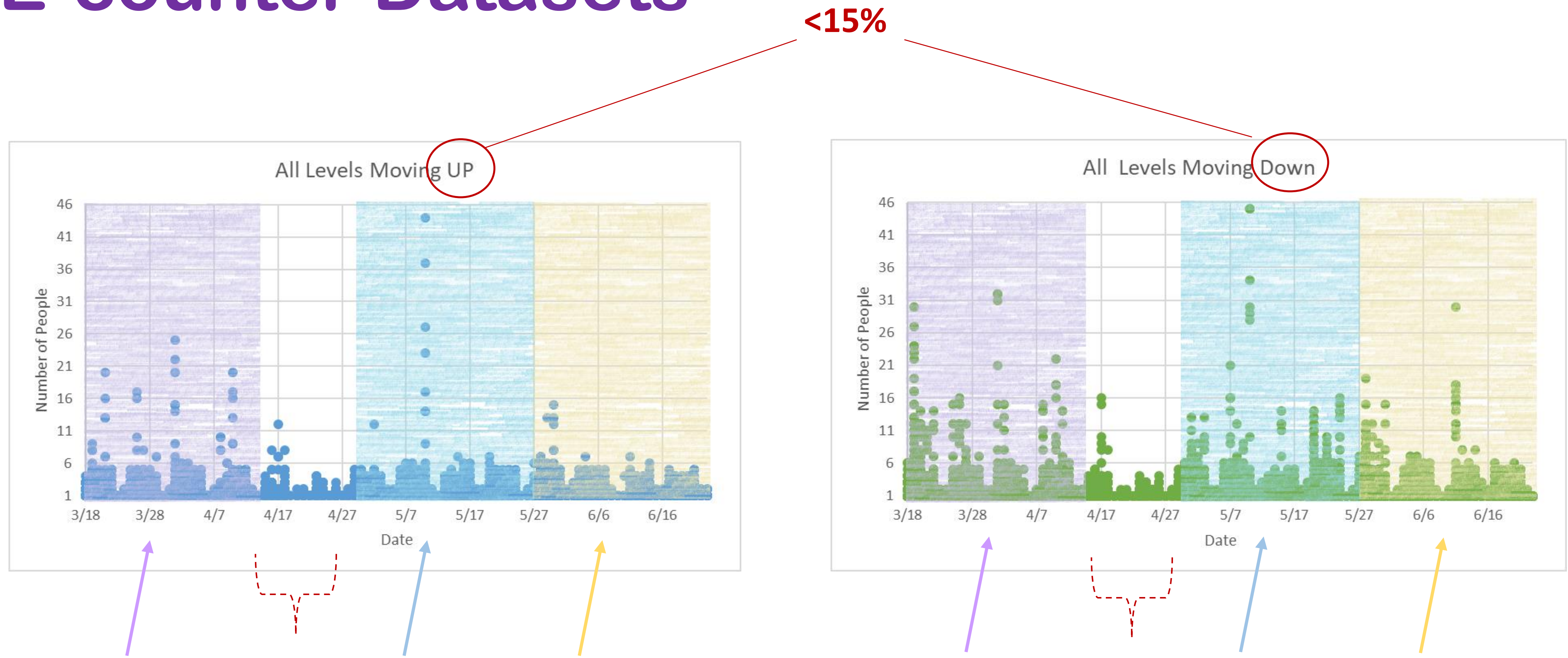
The sensors are placed at six locations: Level 2-1, Level 3-2 Center, Level 4-3 North and South, Level 5-4 North and South



Level 2-1 Stairs

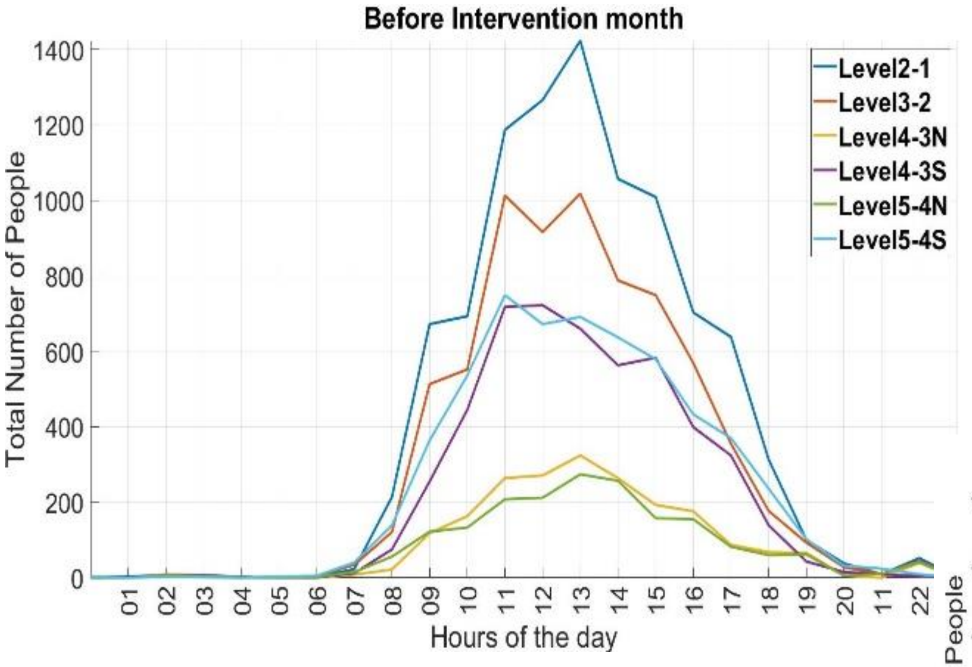
Level 4-3 North and South Stairs

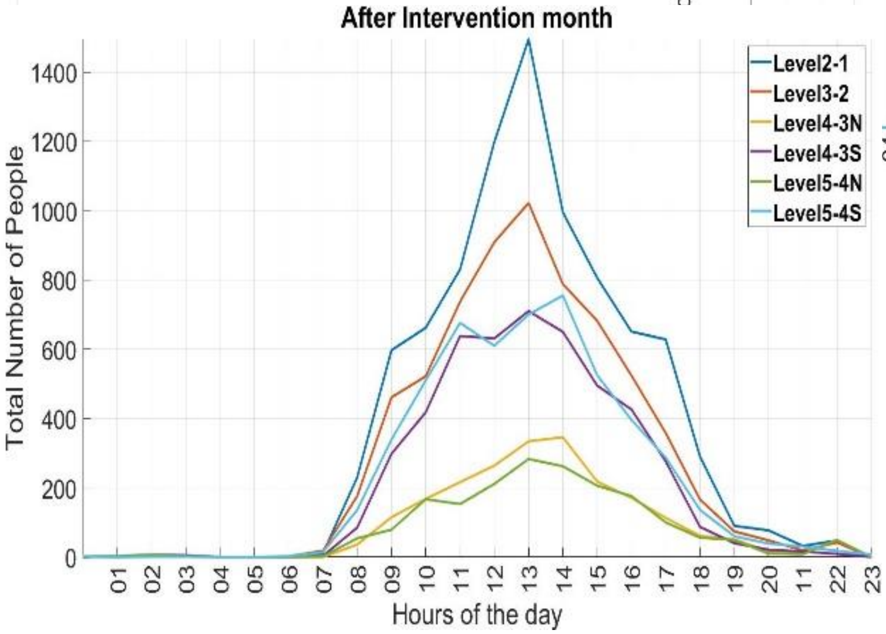
E-counter Datasets

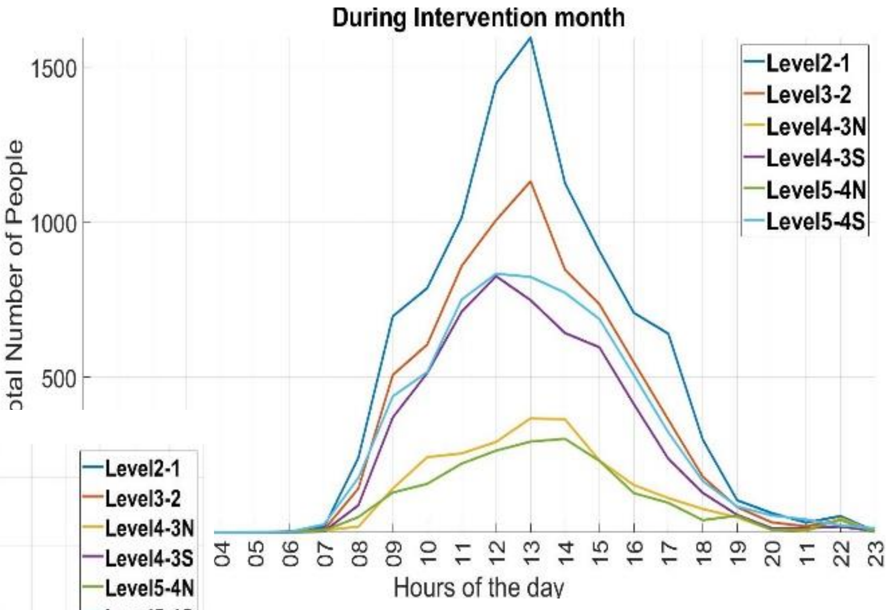


- Pre-intervention → March 18-April 14, 2019
- During-intervention → April 29- May 26, 2019
- After-intervention → May 27- June 23, 2019

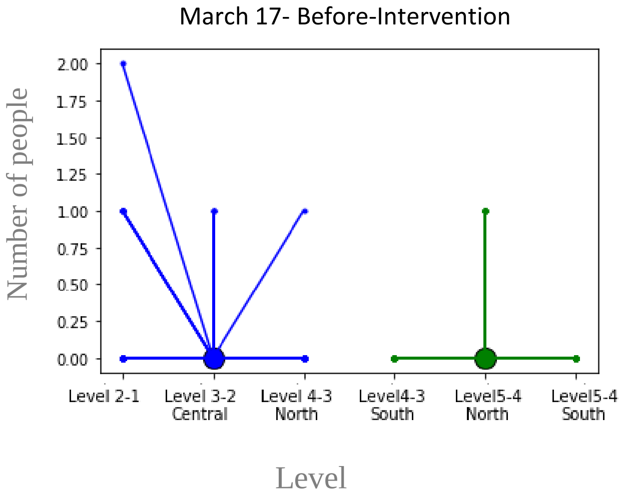
Accumulated hourly count of people for each level of the building during the entire experiment



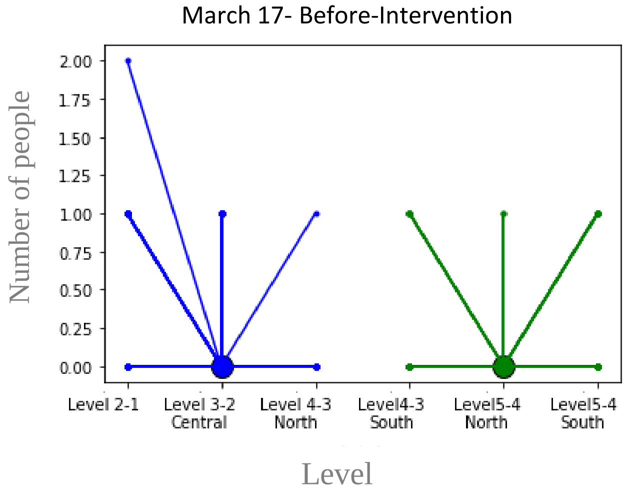




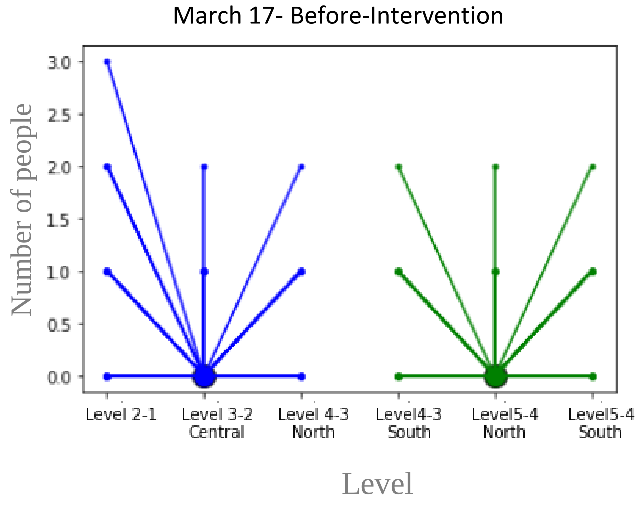
DSAP Results:
Morning Hourly Micro-clusters Before Intervention



Time: 8 a.m.

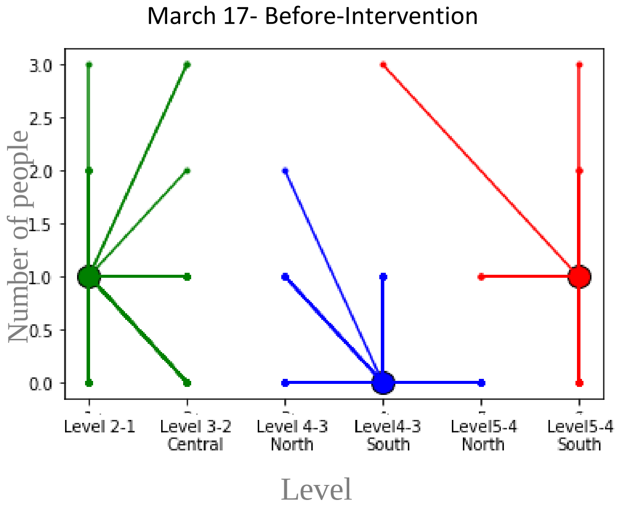


Time: 9 a.m.

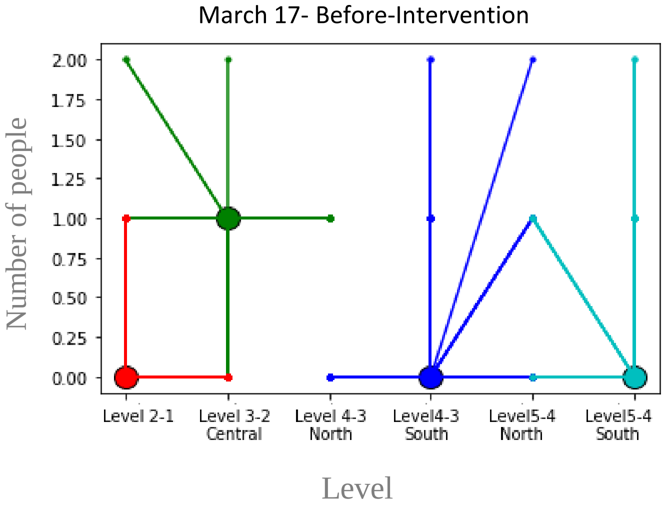


Time: 10 a.m.

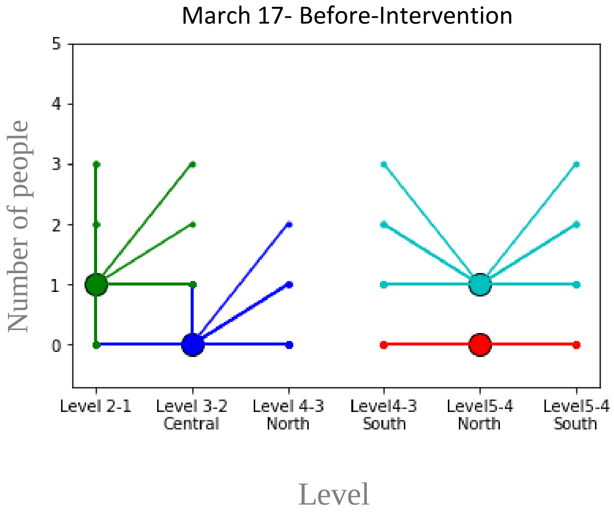
DSAP Results:
Lunch-time Hourly Clusters Before Intervention



Time: 11a.m.

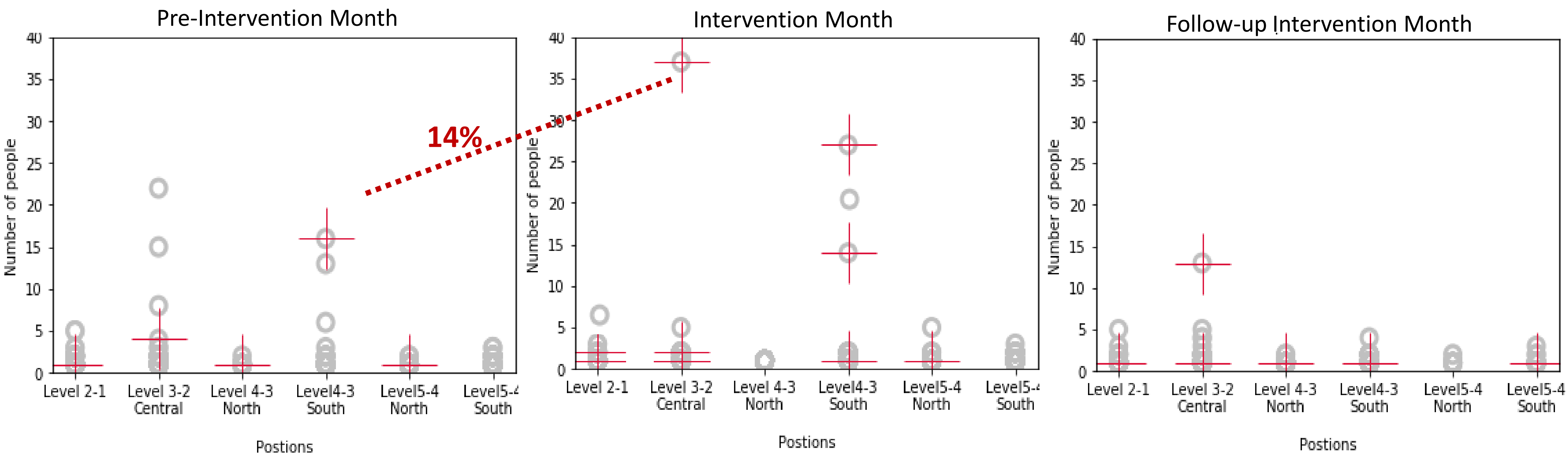


Time: 12 a.m.



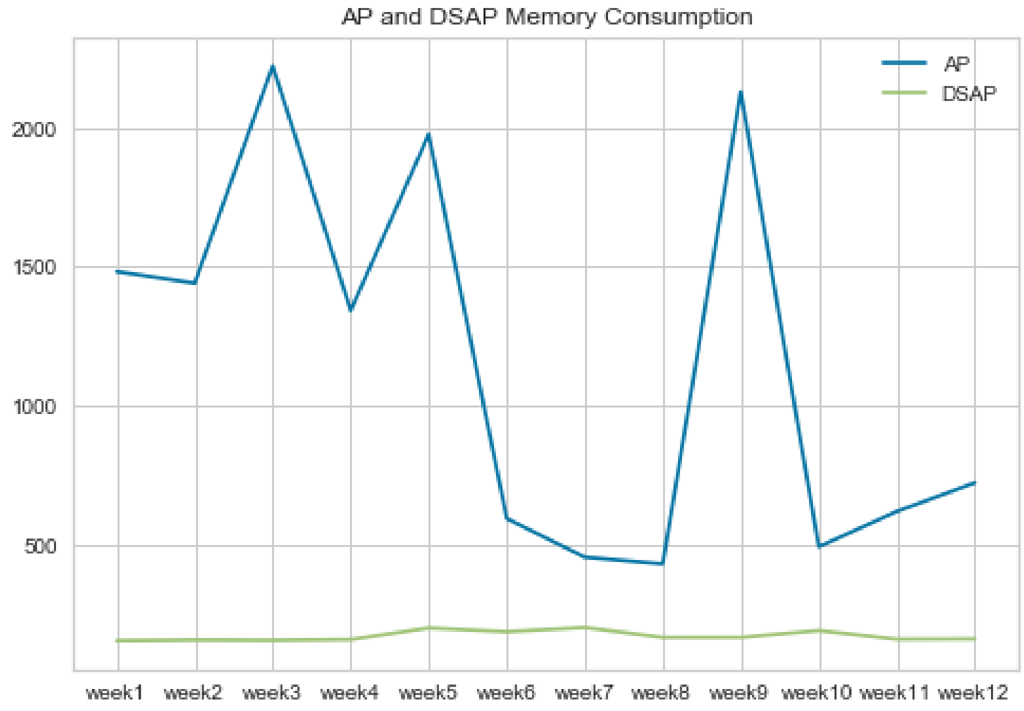
Time: 1 p.m.

DSAP Results:
Micro- and Macro Clusters During the Experiment



- The results are generated from 7 a.m. until 7 p.m.
- The micro cluster centers are represented by the grey circles and the macro clusters are shown with red crosses

DSAP Evaluation



1st week Before Intervention Month

	AP	DSAP
Processing time(S)	17.01	16.3
Memory (MB)	1482.65	156
Number of clusters	6	5
Silhouette Coefficient	.7	.5

1st week Intervention Month

	AP	DSAP
Processing time(S)	25	19
Memory (MB)	1975.7	201
Number of clusters	17	9
Silhouette Coefficient	.16	.5

1st week After Intervention Month

	AP	DSAP
Processing time(S)	25.54	15
Memory (MB)	2129	167
Number of clusters	24	6
Silhouette Coefficient	.23	.6

Silhouette Coefficient: Refers to a method of interpretation and validation of consistency within clusters of data how well each object has been classified. In other words, Means clusters are well apart from each other and clearly distinguished

Conclusions and Future Research Work

- We implemented a novel streaming AP algorithm **(DSAP)** using the landmark time window model for analyzing e-counter data streams.

- The DSAP algorithm is a flexible and can be easily applied on continuous, large spatio-temporal data streams for finding micro and macro-clusters.

- Small and medium volume of data streams are needed to maintain high speed performance when computing the micro-clusters. Towards addressing this issue, we will continue to explore other time window models

CISCO

People in Motion

www.people-in-motion-lab.org

Flinders UNIVERSITY

ADELAIDE · SOUTH AUSTRALIA

NSERC CRSNG

EST. 1785 UNB

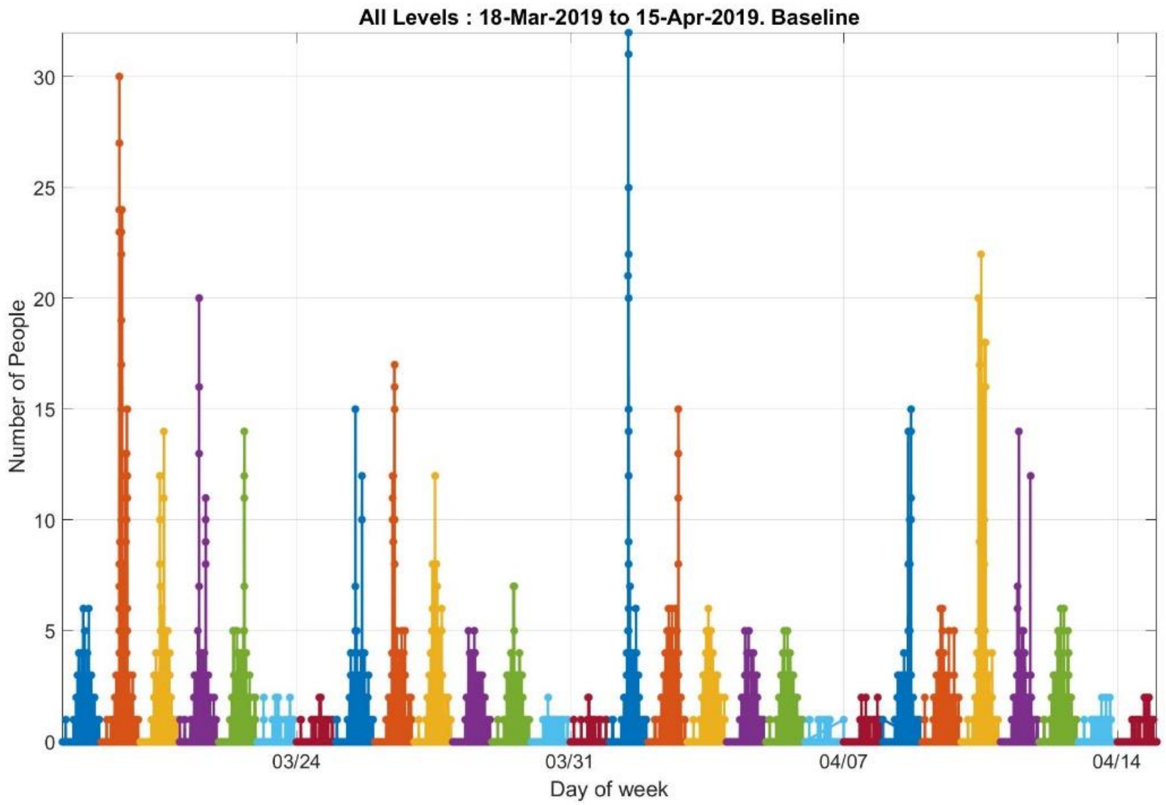
UNIVERSITY OF NEW BRUNSWICK

References

- Aggarwal, C. C., Philip, S. Y., Han, J., & Wang, J. (2003). A framework for clustering evolving data streams. In Proceedings 2003 VLDB conference (pp. 81-92). M. Kaufman.
- Cao, F., Estert, M., Qian, W., & Zhou, A. (2006). Density-based clustering over an evolving data stream with noise. In Proceedings of the 2006 SIAM international conference on data mining (pp. 328-339). Society for industrial and applied mathematics.
- Carnein, M., & Trautmann, H. (2019). Optimizing data stream representation: An extensive survey on stream clustering algorithms. Business & Information Systems Engineering, 61(3), 277-297.
- Dueck, D., & Frey, B. J. (2007). Non-metric affinity propagation for unsupervised image categorization. In IEEE 11th International Conference on Computer Vision (pp. 1-8).
- Sui, J., Liu, Z., Jung, A., Liu, L., & Li, X. (2018). Dynamic clustering scheme for evolving data streams based on improved STRAP. IEEE Access, 6, 46157-46166.
- Zhang, X., Furtlehner, C., & Sebag, M. (2008). Data streaming with affinity propagation. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 628-643). Springer.
- Zhang, J. P., Chen, F. C., Liu, L. X., & Li, S. M. (2013). Online stream clustering using density and affinity propagation algorithm. In 2013 IEEE 4th International Conference on Software Engineering and Service Science (pp. 828-832). IEEE.

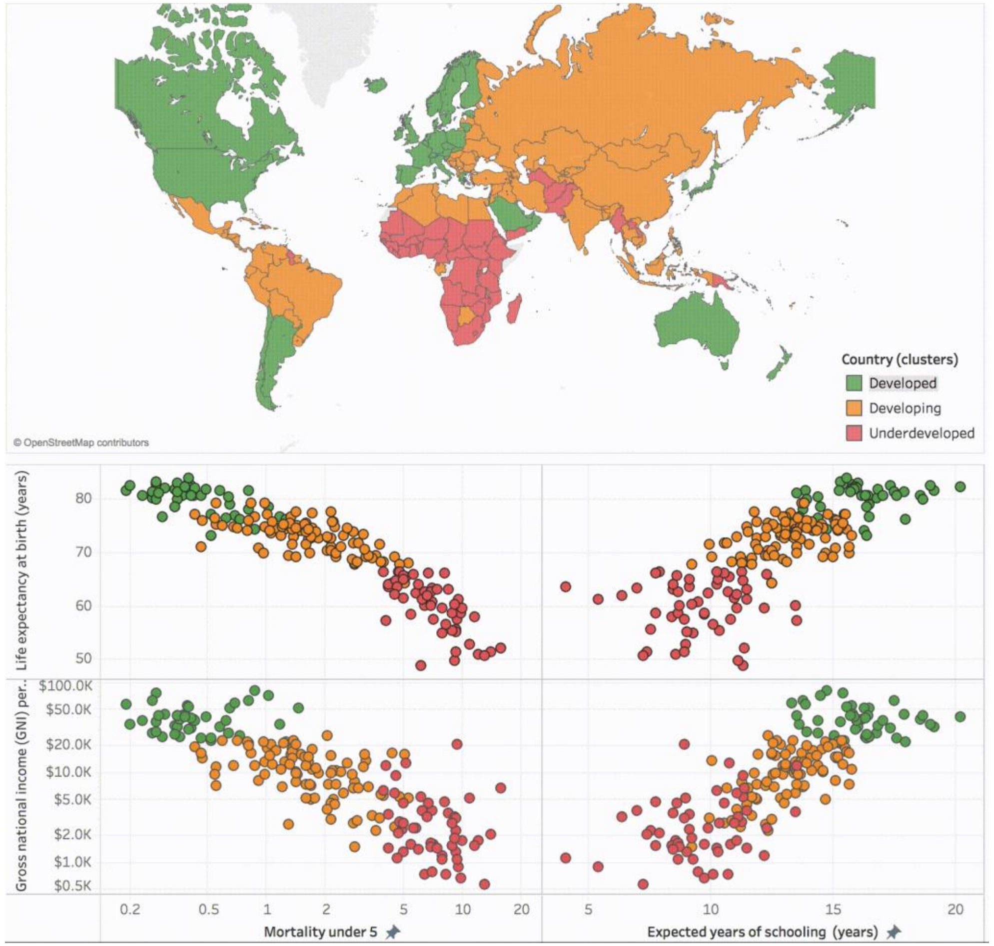
Appendix

Results: Data Distribution



Data Clustering

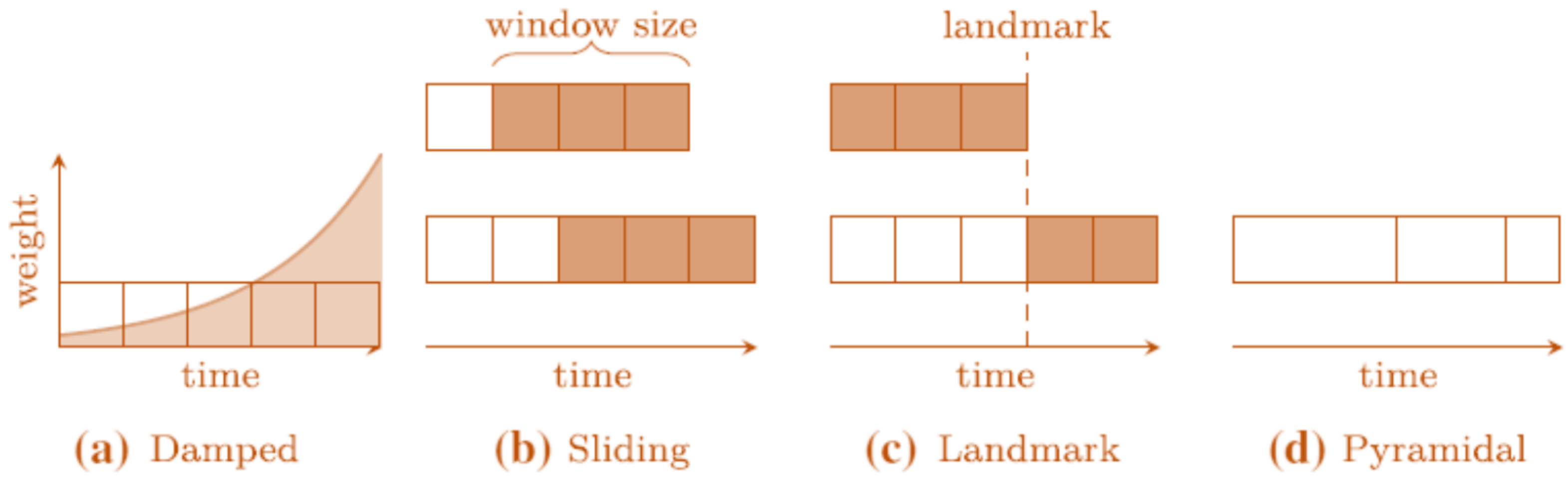
- **Cluster analysis** or **clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups.



https://www.tableau.com/about/blog/2016/7/uncover-patterns-your-data-tableau-10s-clustering-feature-56373

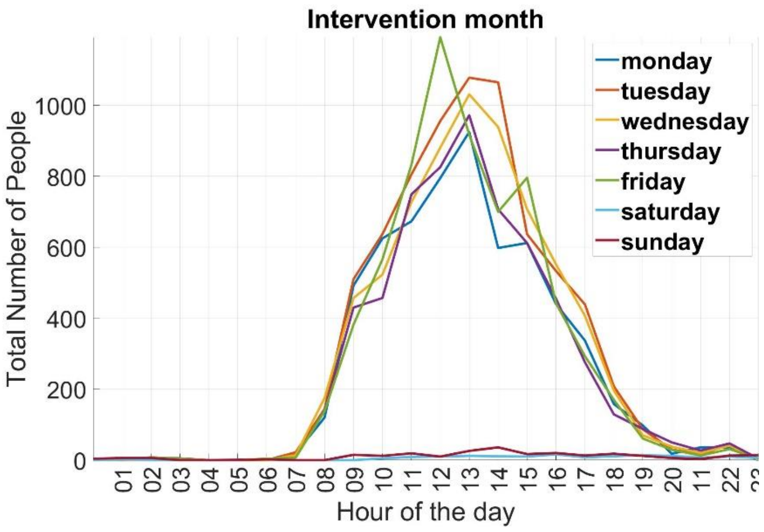
Time window Models

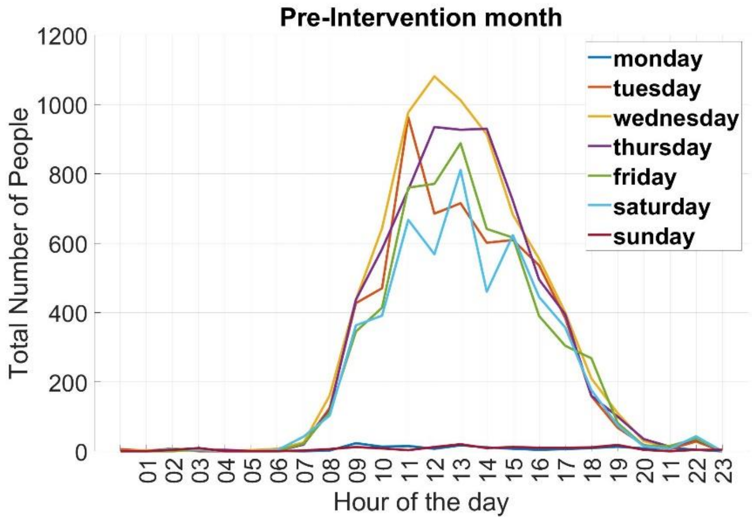
- Damped window: assigns a weight according to the number of observations
- Sliding time window: there is a fixed size of the window. As time passes, the window with the size w slides from the current time.
- Landmark time window: Clustering starts from a starting point called landmark to the current time
- Pyramidal time window: applies various granularity levels based on the novelty of data

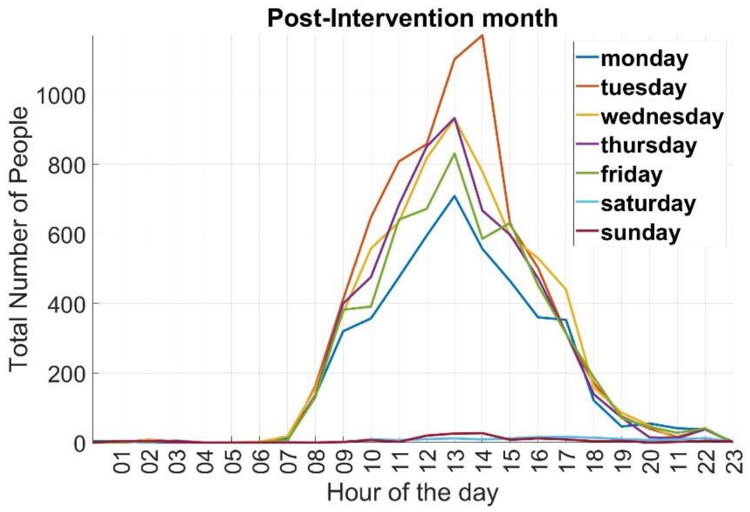


Time window Models(Aggarwal)

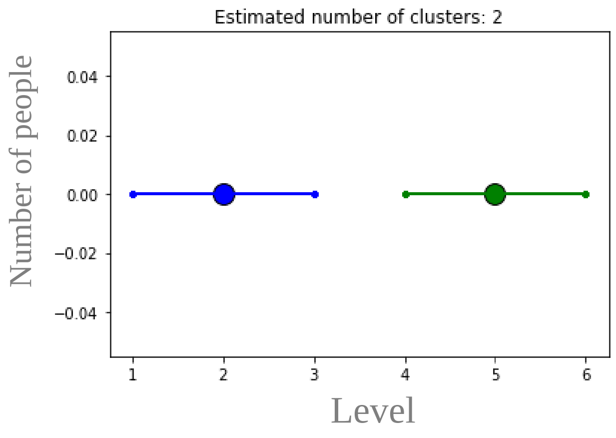
Accumulated hourly counting of people in the building for entire 3 month based on week days



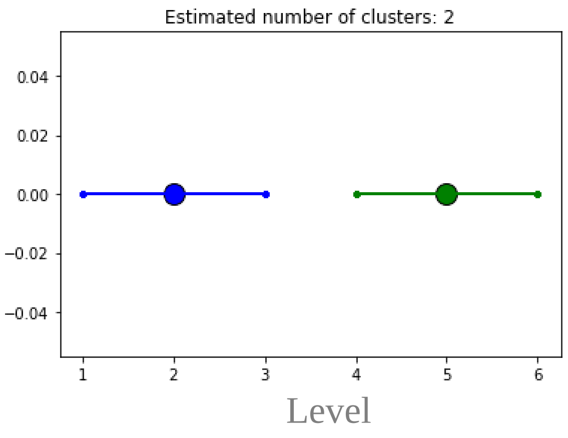




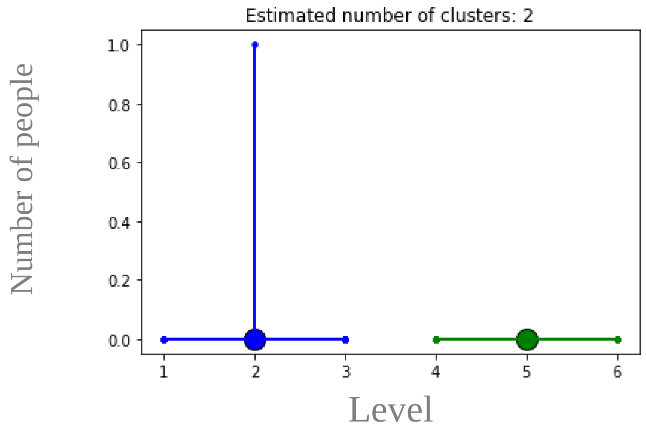
DSAP Results:
Early morning hourly micro-clusters before intervention



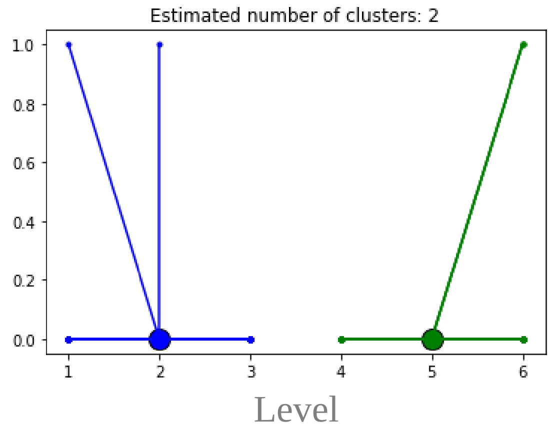
Time: 4am



Time: 5am



Time: 6am



Time: 7am

Implementation of DSAP

Algorithm 1 DSAP Algorithm

Data: Data Points: $E = (E_1, E_2, ..., E_n)$ for computing micro clusters;

Require hyper_parameters: preference, damping, max_iter, convergence_iter

Initialize: Landmark time window (size T_s = 60 minutes)

Similarity Matrix: S $\forall$ i, k: s(i, k) = 0

Availability Matrix: A $\forall$ i, k: a(i, k) = 0

Responsibility Matrix: R $\forall$ i, k: r(i, k) = 0

Function `Affinity_Propagation(`*Data_points*`)`**:**

$S \, \forall \, i, \, k\text{: } s(i,k) = -||x_i - x_k||^2$

while *r(i,k) and a(i,k)* $\neq$ *convergence* **do**

Updating R:

$r(i,\ k) \leftarrow s(i,k) - \max_{k' s.t. k' \neq k} \{\ a(i,\ k') + s(i,\ k')\ \}$

Updating A:

$(i,\ k) \leftarrow \min\{0, r(k,k) + \sum_{i' s.t. i' \notin \{i,k\}} \max\{0, r(i',k)\}$

non-diagonal A:

$a(i,\ k) \leftarrow \sum_{i' \neq k} \max(0, r(i',k))$

for *7a.m.* $\leq T_s \geq$ *7p.m.* **do**

Function `Affinity_Propagation(`*E*`)`**:**

Result: Set of cluster heads for computing macro clusters:

$P = (P_1, P_2, ..., P_n)$

Function `Affinity_Propagation(`*P*`)`**:**

Result: Macro Clusters:

$C = (C_1, C_2, ..., C_n)$