IDENTIFYING TEMPORAL USER BEHAVIOR THROUGH SMART CARD DATA

Mohammad Sajjad Ghaemi[1,2], Doctorate student

Bruno Agard[1,3], Professor

Vahid Partovi-Nia[1,2], Professor

Martin Trépanier[1,3], Professor

1 POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

2 GERAD
GROUP FOR RESEARCH IN DECISION ANALYSIS

3 CIRRELT
Interuniversity Research Center on Entreprise Networks, Logistics and Transportation

In this presentation

- **Introduction**
  - Why using data mining for smart card data?
- **Background**
  - Smart card data in public transport planning
- **Methodology**
  - Agglomerative Hierarchical Clustering
  - Distance calculation
  - Cluster identification
- **Results**
  - Case study
  - Cluster definition
  - Descriptive analysis
- **Conclusion**

Introduction

- A typical smart card automated fare collection system for public transit collects **millions of transactions** each day
- These systems are made for **revenue collection**, however they can be used to identify « card » behaviour for transport planning purposes *(→ strict privacy is kept)*
- However, to analyze the temporal behaviour of cards, advanced data mining technique must be used because:
  - The number of observations is **too large**
  - Classical distance-based techniques **are not always suitable** to clusterize temporal information
- In this project, we propose a **new method** of distance calculation + a **solving technique**

Background

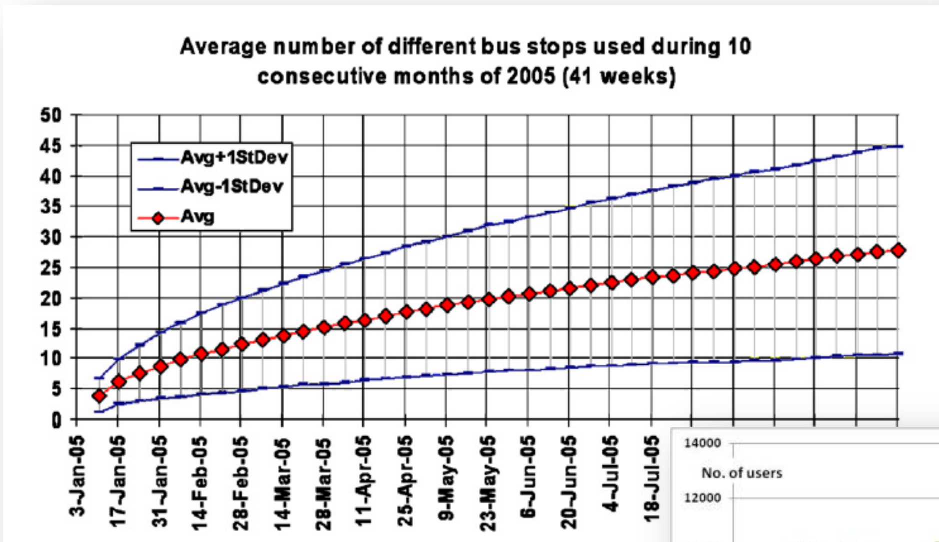
Smart card in public transit

- Through the years, the usefulness of smart card data for public transit planning has been demonstrated:
    - for **ridership** and **turnover** studies
    - for **behaviour detection** using classical data mining techniques
    - to identify destinations and **create OD matrices**
    - to evaluate **travel time** in subway systems
    - to examine the **impact of weather** on transit usage
    - to assess the **loyalty** of users
    - to calculate **KPIs** on demand and supply
- Many DM techniques were used on smart card data: classical k-means, DBScan, mixture of Gaussian distribution, etc.

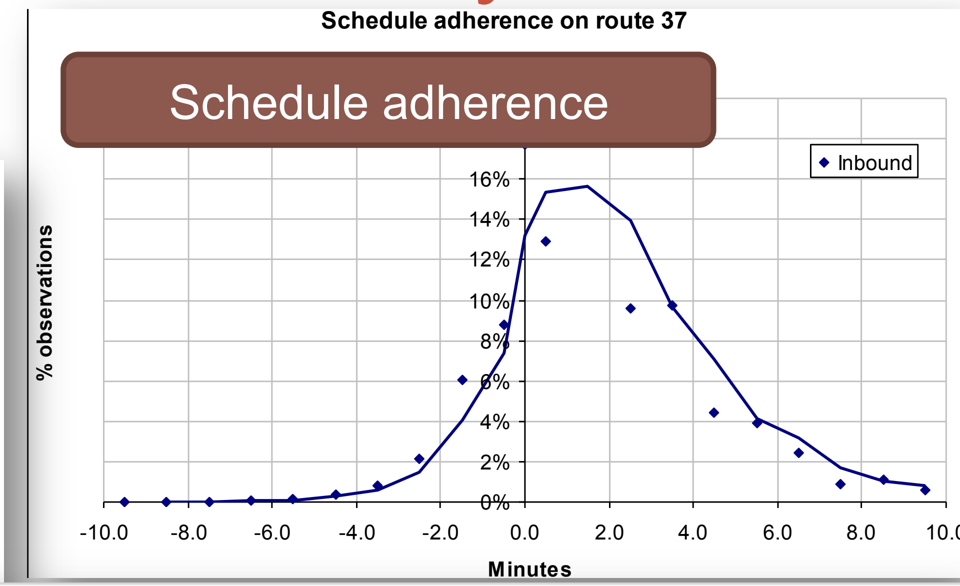
Background

Some examples of SC data analysis

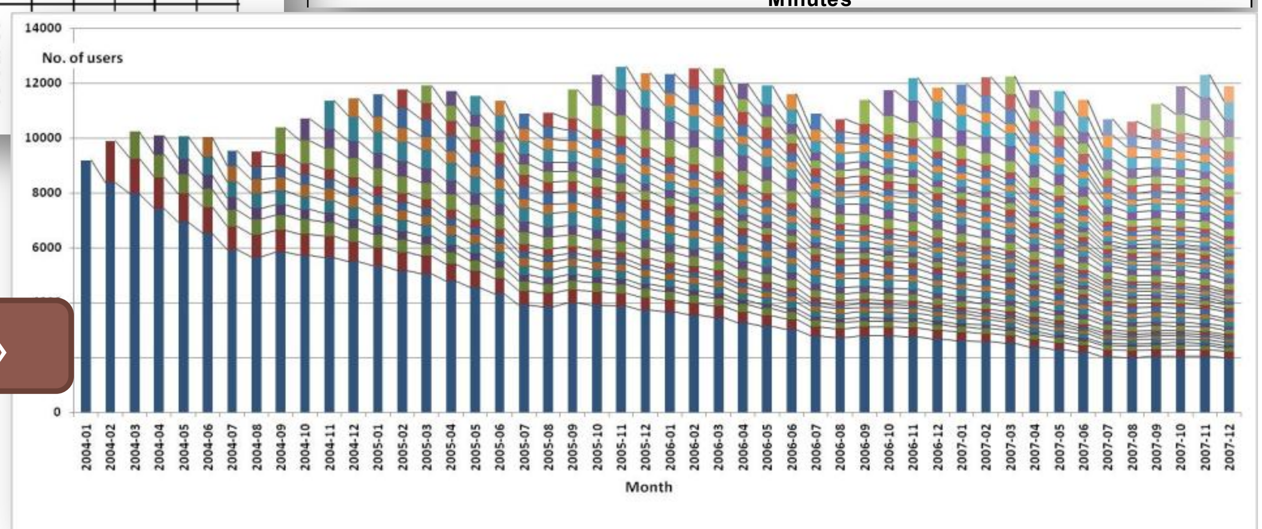
Learning of the bus network



Schedule adherence



Card « survival »



Methodology

Clustering method: AHC

- In our case study, because there are more than 400,000 observations, **we cannot use a classical k-means** in a reasonable computer calculation time (600 Gb memory needed!)
- We propose to use a model-based approach, a modified **Agglomerative Hierarchical Clustering** (AHC)
  - We start with a classical k-means with **1000 randomly selected observations** and consequently merges the rest with the closest cluster centers to end up with all data in clusters
  - The nested groups generated using a hierarchical clustering algorithm of data, are visualized through a **dendrogram** that shows the « distances » between observations
  - We use the dendogram to « **cut** » the observations into clusters

Methodology

Distance calculation

- When looking at temporal distribution of transactions, **distance calculation** between vectors is an issue

Hours of the day

Card-days	H1	H2	H3	H4	H5	H6	H7
1	1	0	0	0	0	0	0
2	0	1	0	0	0	0	0
3	0	0	1	0	0	0	0
4	0	0	0	1	0	0	0
5	0	0	0	0	0	1	0
6	0	0	0	0	0	0	1
7	1	1	0	0	0	0	0
8	1	0	1	0	0	0	0
9	1	1	1	0	0	0	0
10	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0
12	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1

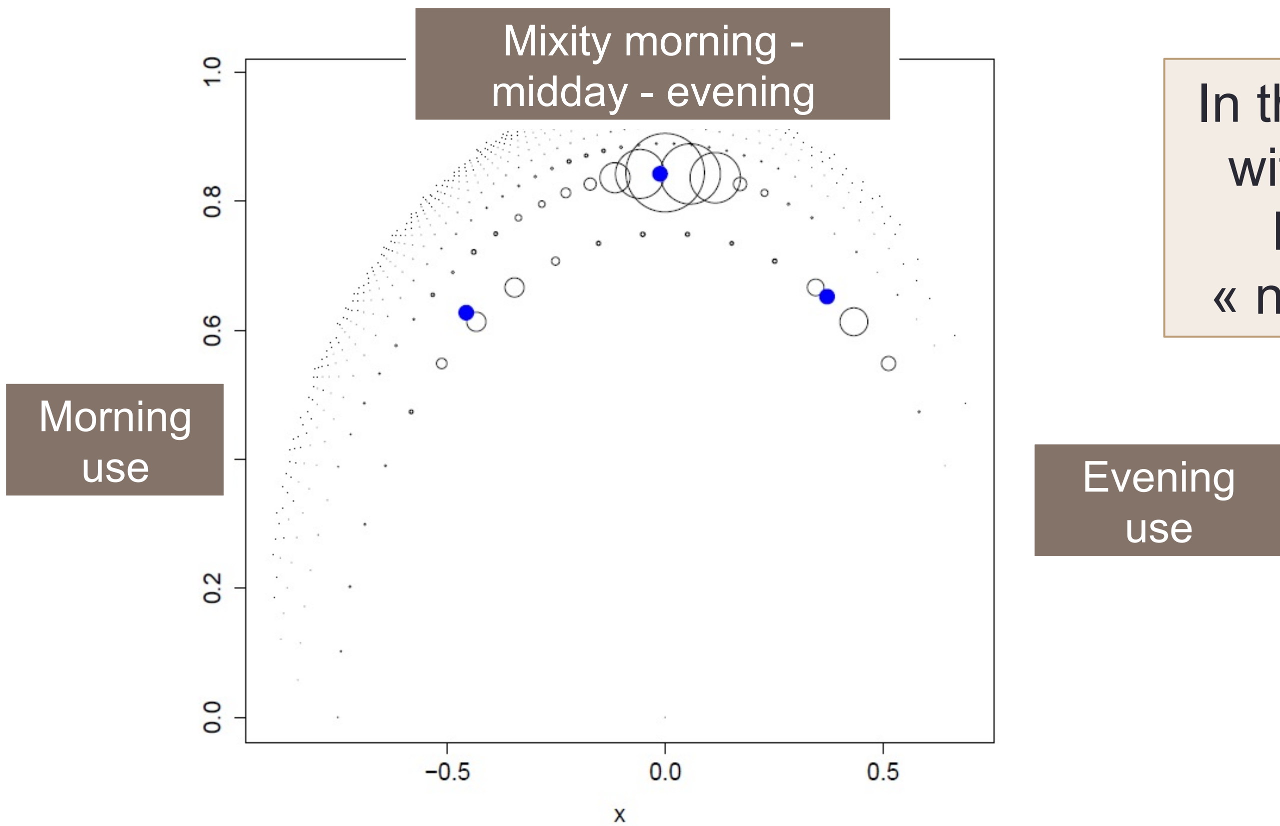
Distance	Euclidean	Manhattan
D(1,2)	$\sqrt{2}$	2
D(1,3)	$\sqrt{2}$	2
D(7,8)	1	1
D(7,9)	1	1

From a « transportation » point of view, D(1,2) should be smaller than D(1,3)!

Methodology

SCP method for distance

- To make results acceptable to transit planners, we propose to use a **Semi-Circle Projection** (SCP) of the vectors before calculating distances

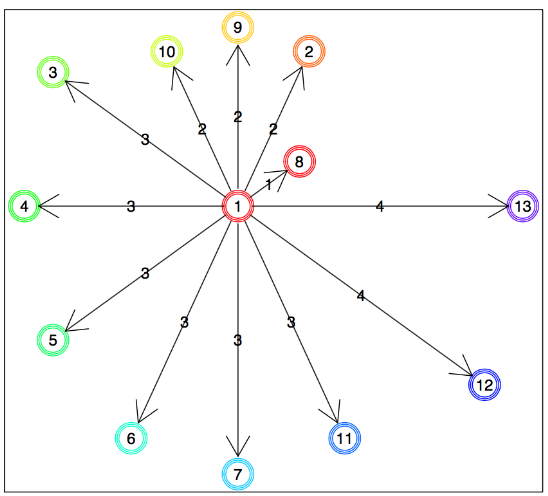


In this projection, users with similar temporal behaviour will be « near » to each other

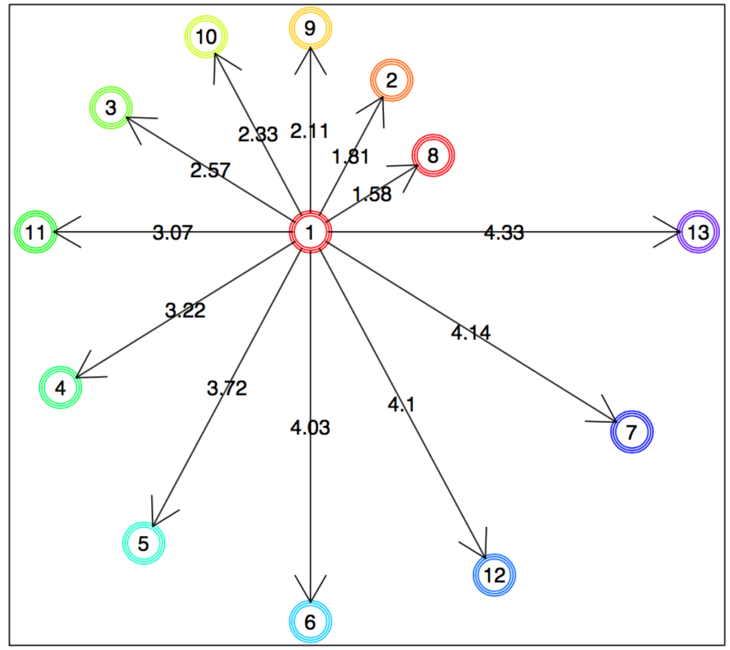
Methodology

Distance calculation

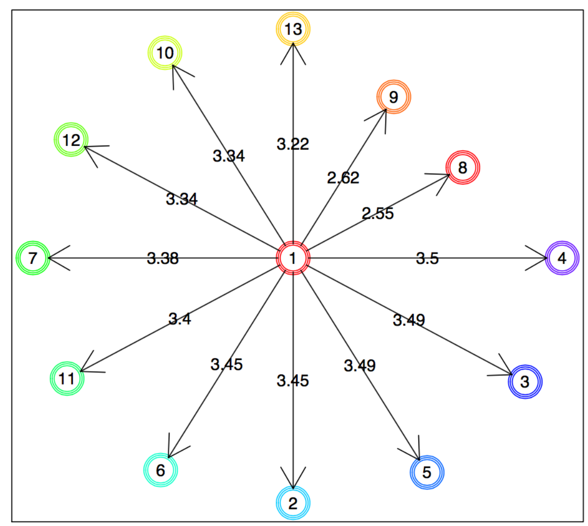
User	H_1	H_2	H_3	H_4	H_5	H_6	H_7
X_1	1	0	0	0	0	0	0
X_2	0	1	0	0	0	0	0
X_3	0	0	1	0	0	0	0
X_4	0	0	0	1	0	0	0
X_5	0	0	0	0	1	0	0
X_6	0	0	0	0	0	1	0
X_7	0	0	0	0	0	0	1
X_8	1	1	0	0	0	0	0
X_9	1	0	1	0	0	0	0
X_{10}	0	1	1	0	0	0	0
X_{11}	1	0	0	1	0	0	0
X_{12}	0	0	0	0	1	1	0
X_{13}	0	0	0	0	0	1	1



(a) Autocorrelation distance



(b) SCP distance

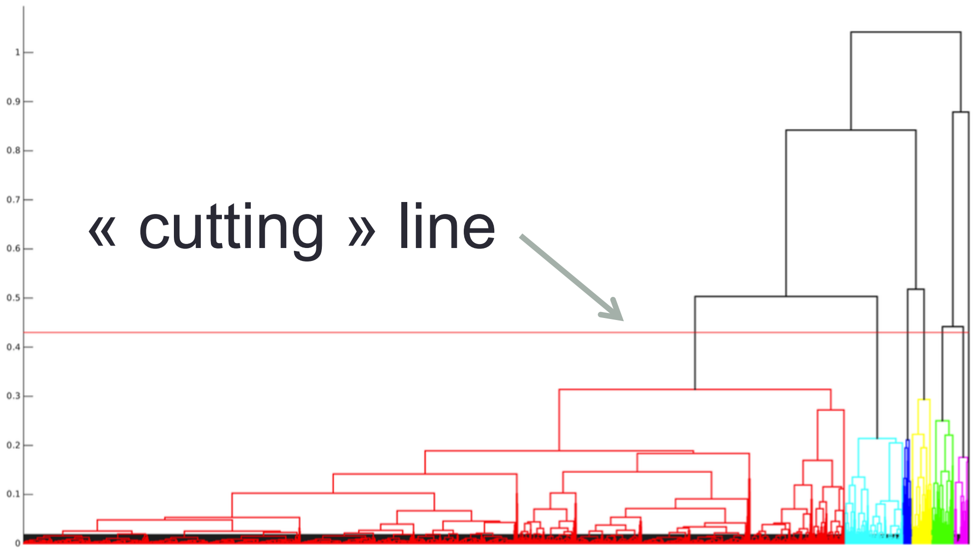


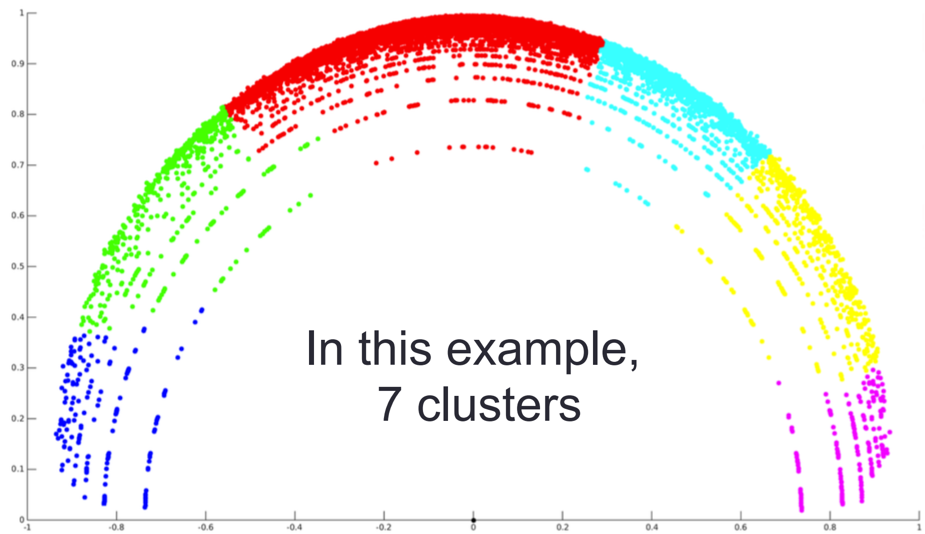
(c) Cross-correlation distance

Methodology

Cluster identification

- The **number of clusters** to be found is still an open research question; it depends on the level of resolution needed, we try to obtain **equilibrated clusters**

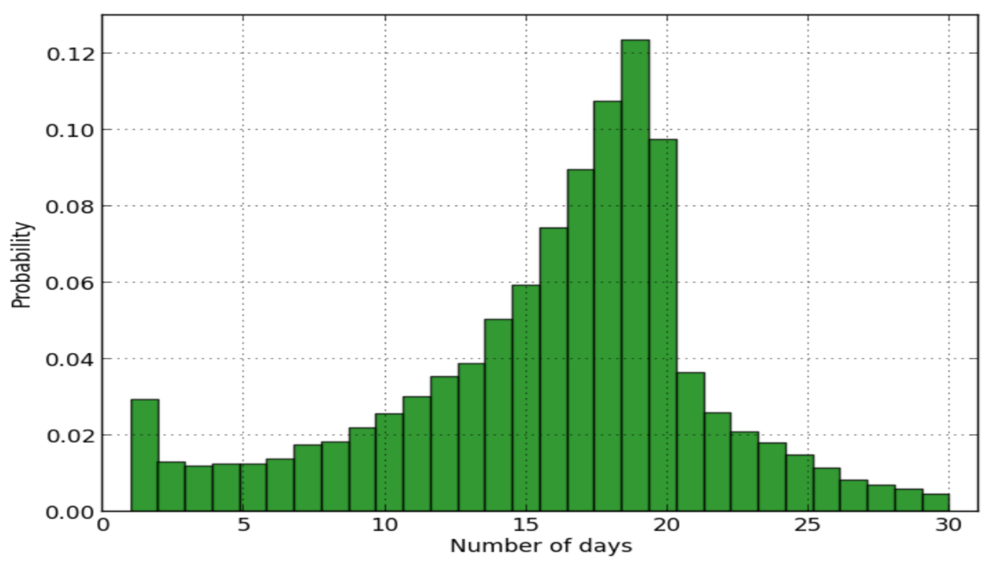




Results

Case study

- *Société de transport de l'Outaouais*, a mid-size authority (300 buses & 220,000 inhabitants)
- **One month** period (April 2009)
- 26,176 cards
- 753,016 transactions
- **416,076 card-days**

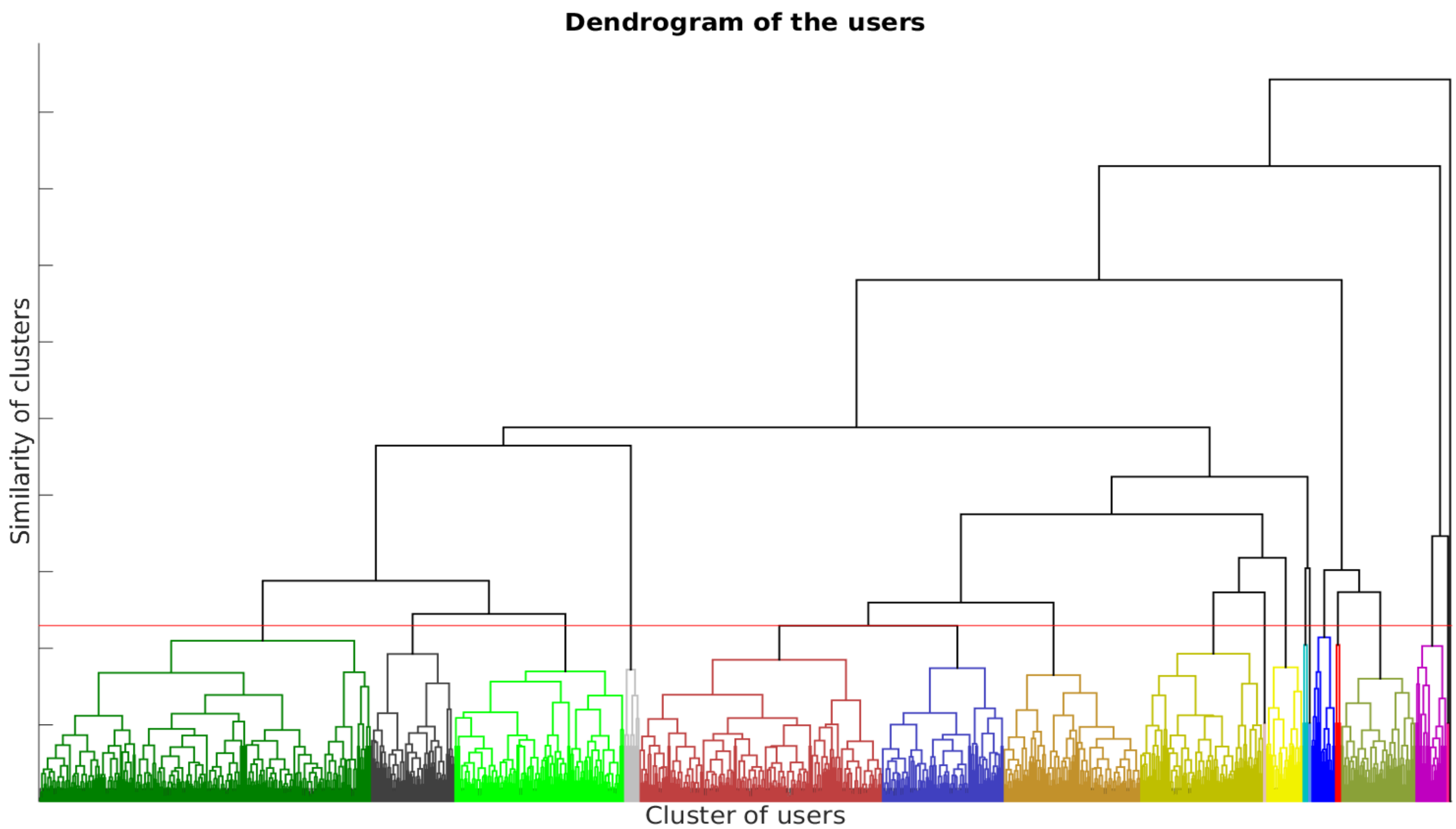


Most users (cards) travel 19 days per month

Results

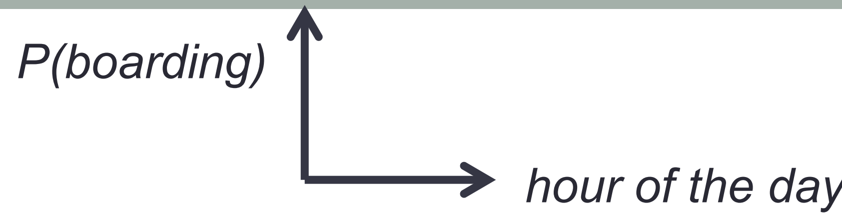
Dendogram

- 18 clusters were identified

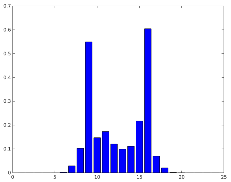


Results

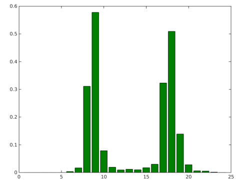
18 clusters



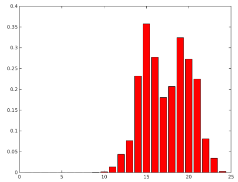
1



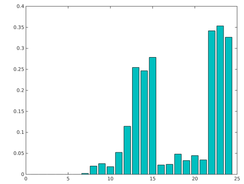
2



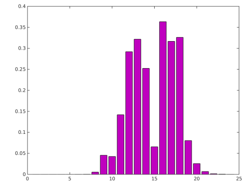
3



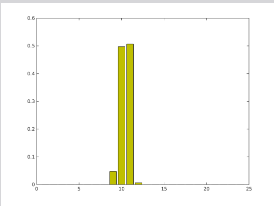
4



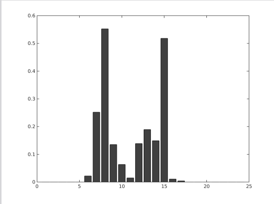
5



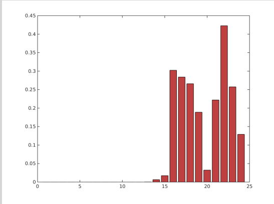
6



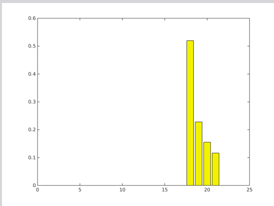
7



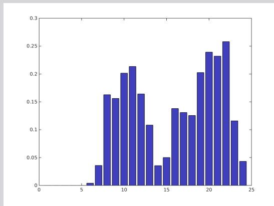
8



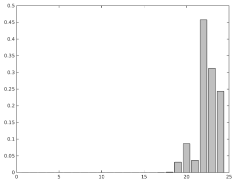
9



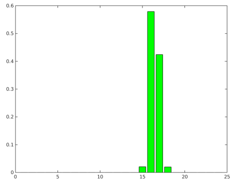
10



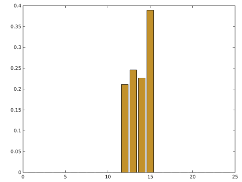
11



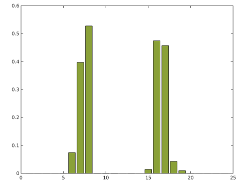
12



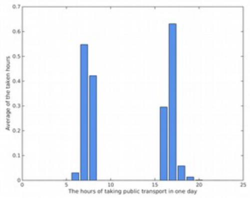
13



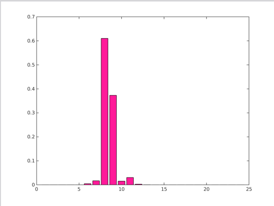
14



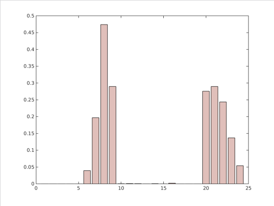
15



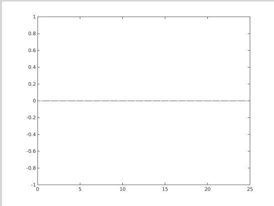
16



17



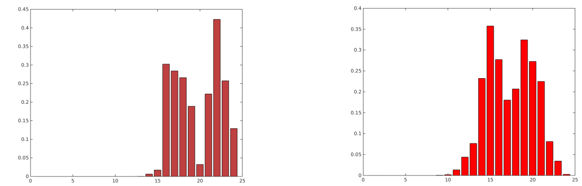
18



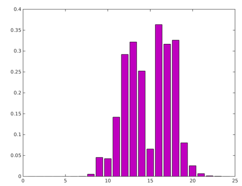
Results

Cluster characterization

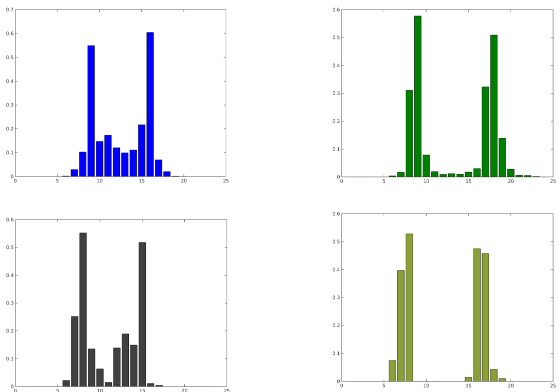
Late commuters



Midday

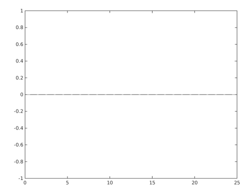


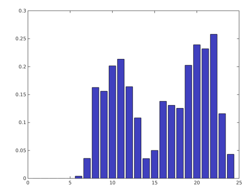
Regular commuters



Active?

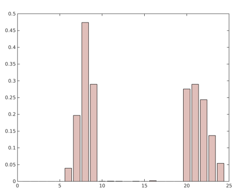
No!



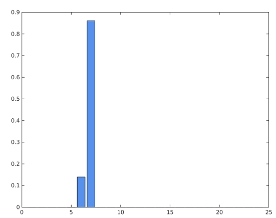


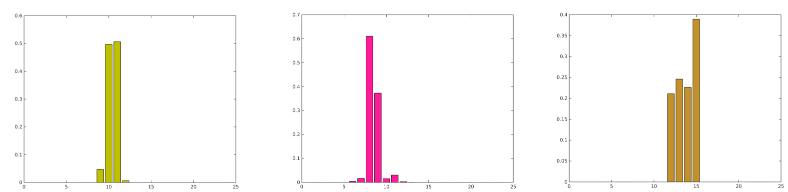
Yes!

Long day



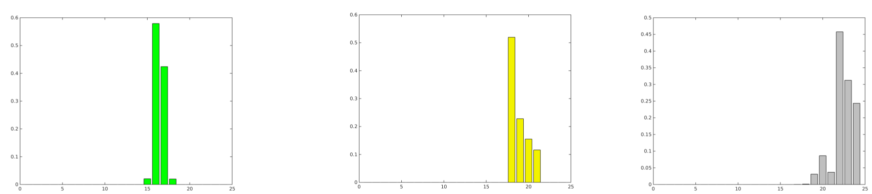
Single trip





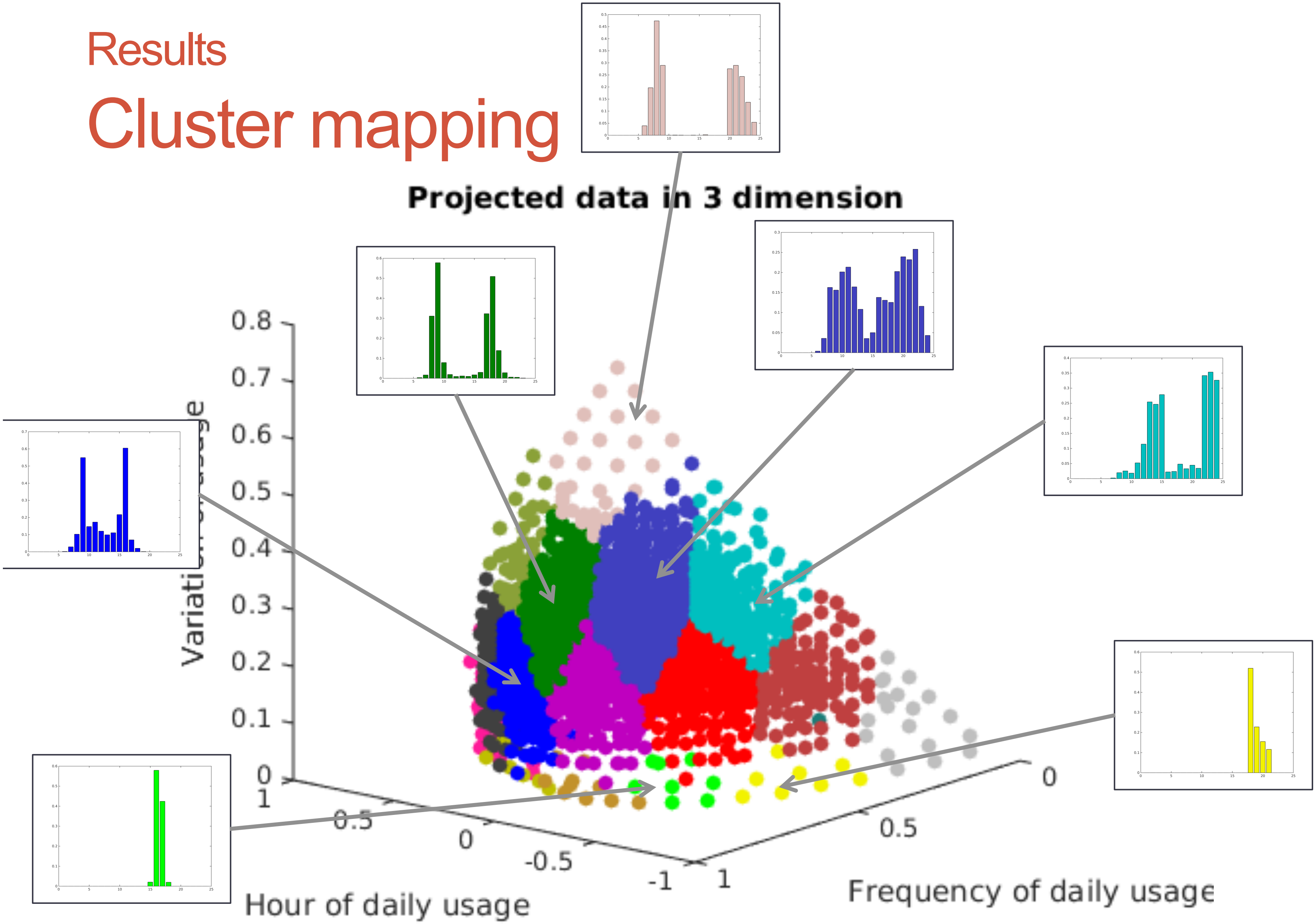
Earlier

Later

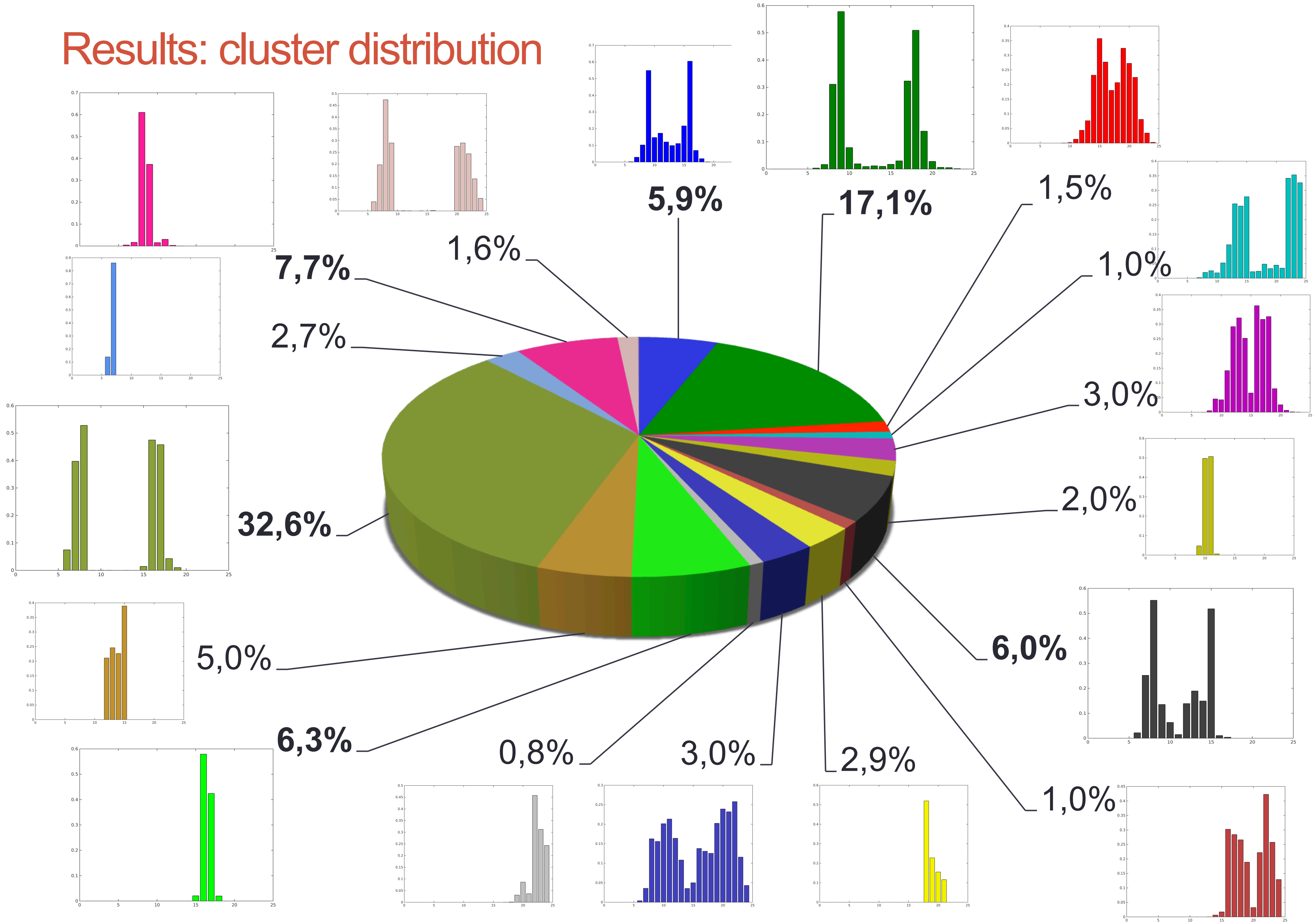


Results

Cluster mapping

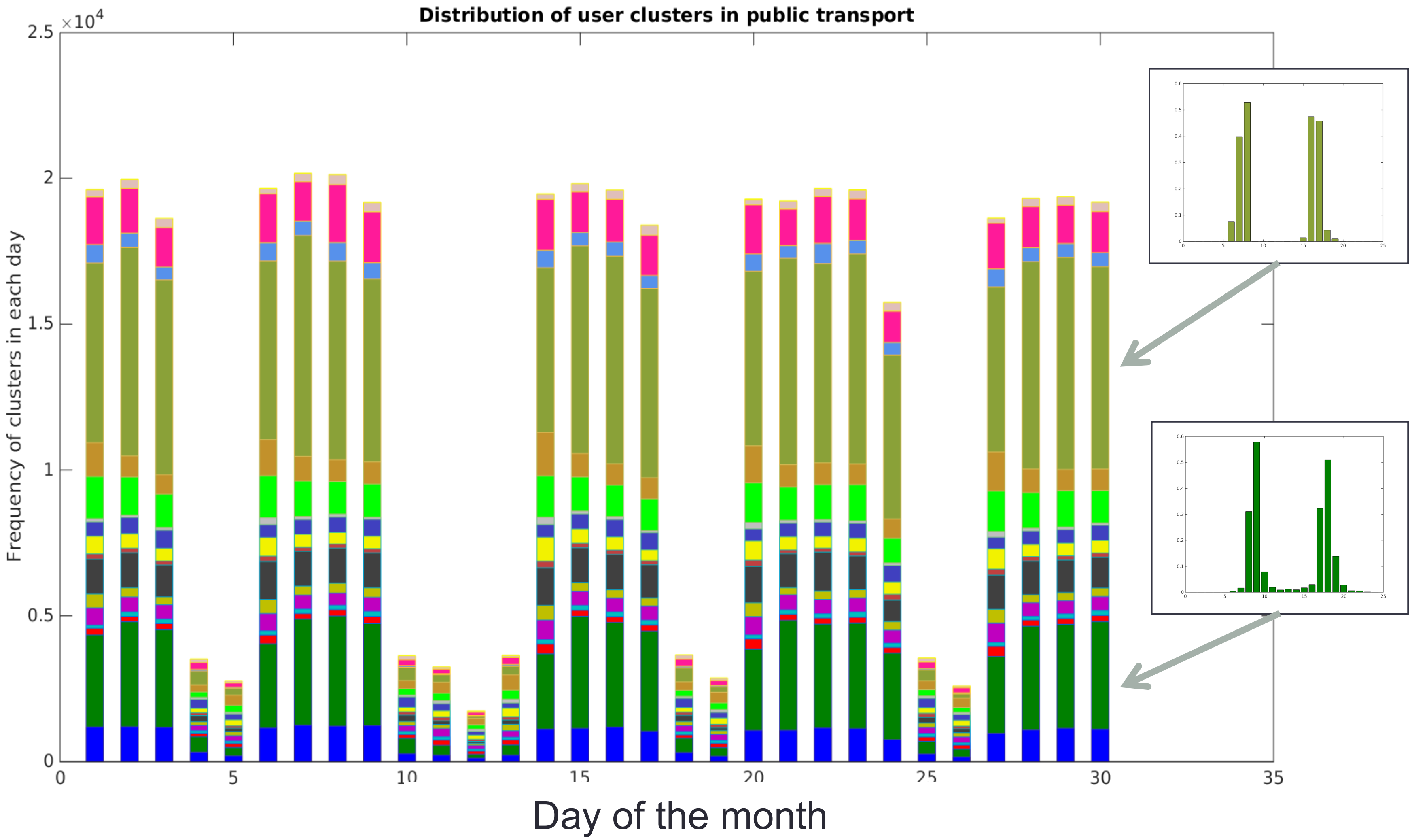


Results: cluster distribution

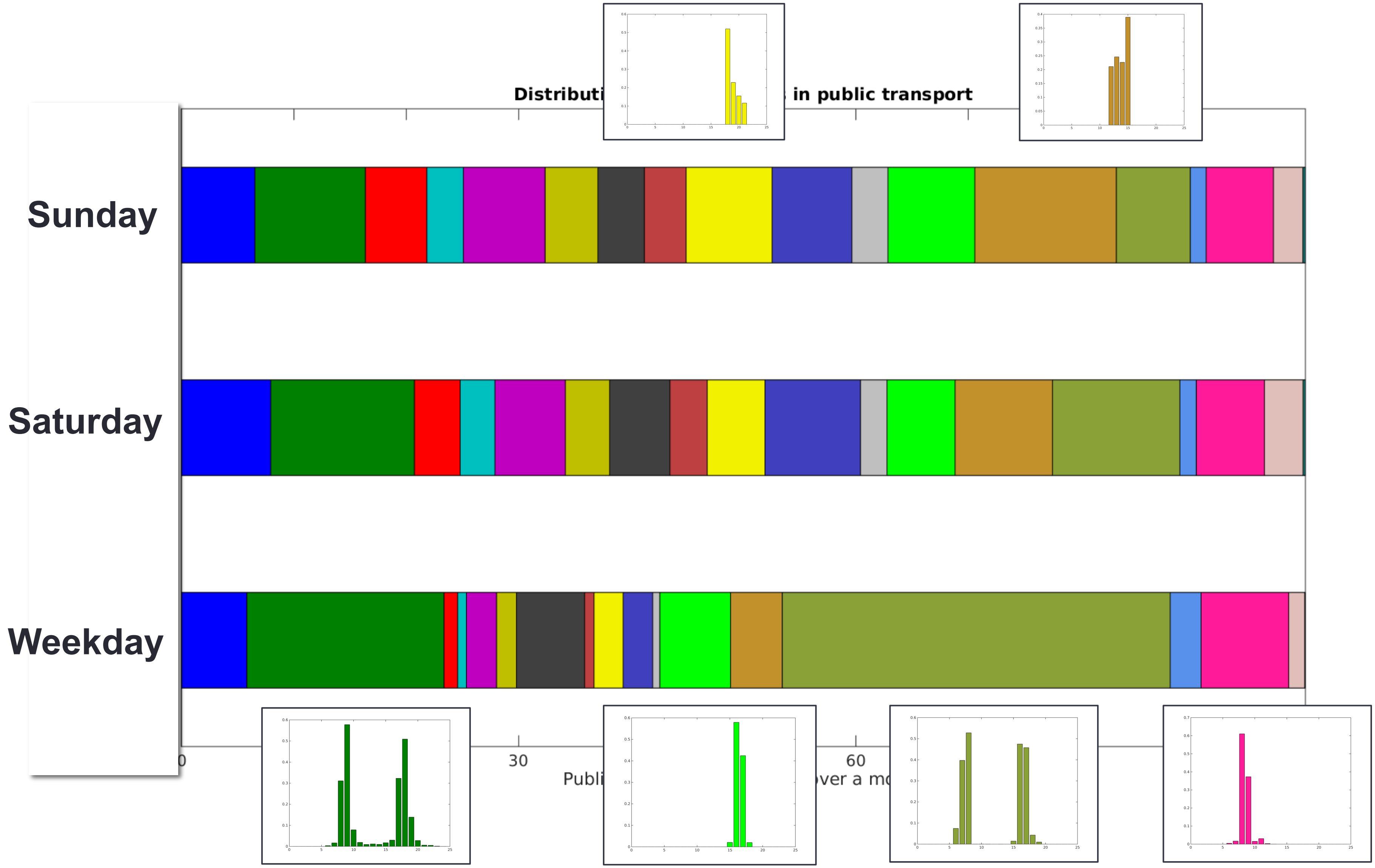


Results

Cluster distribution over the month



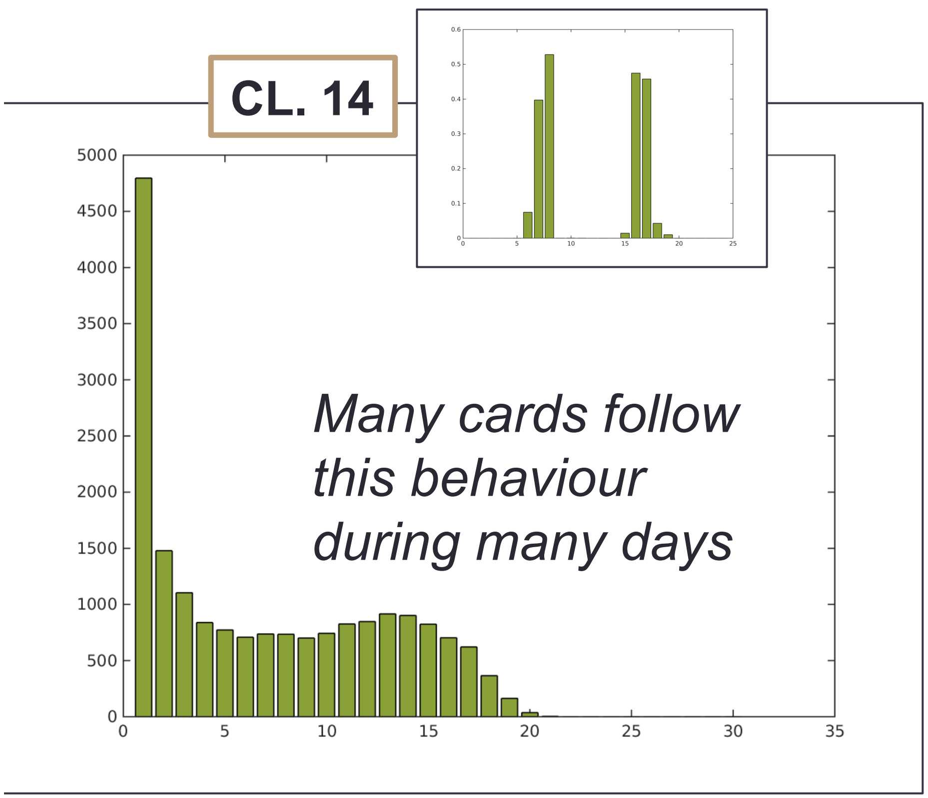
Results： Cluster distribution by average day

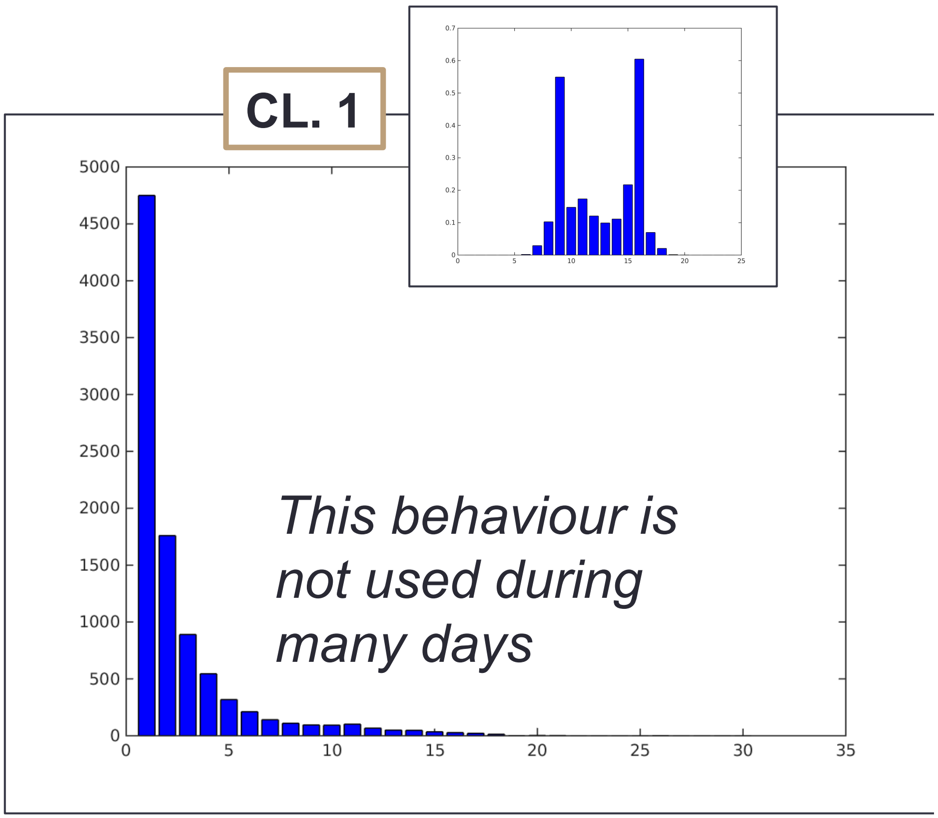


Results

Frequency of cards in cluster

- Clusters are made on card-days, so a card can be assigned up to 30 times to the **same cluster** in April 2009





Conclusion

- Smart card data is a **plentiful source of travel behaviour** knowledge of public transit users
- Number of observations explodes, so it is **difficult to apply classical data mining techniques**, we must find a way to tweak the existing methods
- Having a **good distance metric** is the key
- Once applied, the techniques help to **better understand** the type and the frequency of behaviours among cards

Perspectives

- Process **more data**
- Integrate the **spatial location** of boarding, not only temporality

Acknowledgements

STO Société de transport de l'Outaouais

THALES

NVIDIA®

CRSNG NSERC