Vectors of movement, a new approach to cluster multidimensional big data on mobility

Rafał Kucharski
Achille Fonzone, Arkadiusz Drabicki, Guido Cantelmo

Politechnika Krakowska, Poland

Oct 2018, TU Delft

Politechnika Krakowska
im. Tadeusza Kościuszki

Introduction

Politechnika Krakowska
im. Tadeusza Kościuszki

Problem

Synthesis

How to synthesize big & multidimensional mobility data into readable and meaningful form?

Comparison

How to determine if two mobility datasets are similar?

Clusterization

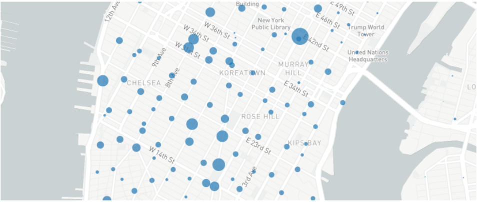
Can we identify groups of similar mobility?

Politechnika Krakowska
im. Tadeusza Kościuszki

Data

Trips

Stations of the system with their capacities



Trips made with one of 12 000 New York City bicycles to travel between 750 pick-up and drop-off stations spread over NYC.

Each of over 50M trips recorded since 2014 as:

$$T_i = \{O_i, D_i, t_i, \Delta t_i\}$$

, where:

O_i and D_i are pick-up and drop-off stations

t_i and Δt_i is pick-up time and trip duration.

Politechnika Krakowska
im. Tadeusza Kościuszki

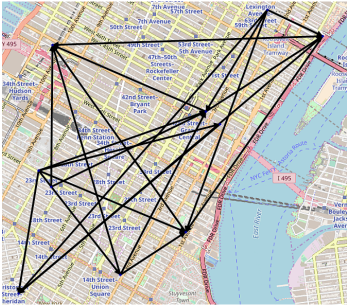
Data

Trip

$$T_i = \{O_i, D_i, t_i, \Delta t_i\}$$

Data

Mobility (trip set)



Mobility pattern

Set of trips, typically recorded over a given period of time (a day in case of this research)

$$M_i = \{T_1, T_2, \ldots, T_n\}$$

Politechnika Krakowska
im. Tadeusza Kościuszki

Problem

rephrased

Synthesis

How to synthesize big & multidimensional mobility data into readable and meaningful form?

$$M_i = \{T_1, T_2, \dots, T_n\}$$

Comparison (similarity measure)

How to determine if two mobility datasets are similar?

$$M_1 \approx M_2, M_1 \approx M_3, M_2 \approx M_3,$$

$$s(M_1, M_2) > s(M_2, M_3) > s(M_1, M_3)$$

Clusterization

Can we identify groups of similar mobility?

$$C_1 = \{M_1, M_3, M_5, \dots\}$$

$$C_2 = \{M_2, M_4, M_6, \dots\}$$

Method

Method

Objectives

Synthesis

Minimal dimension possible (max dimensionality reduction)

Comparison (similarity measure)

Formal distance (similarity) metrics needed to cluster.
Computationally light ($D \times D$ pairwise matrix precomputed).

Clusterization

Meaningful, revealing interesting groups, differences, valuable (explanatory, applicable, visual, ...)

Politechnika Krakowska
im. Tadeusza Kościuszki

Data synthesis

Center of gravity

For generic mobility pattern M we introduce center of gravity for origins:

$$O_M = E(O_i : i \in M)$$

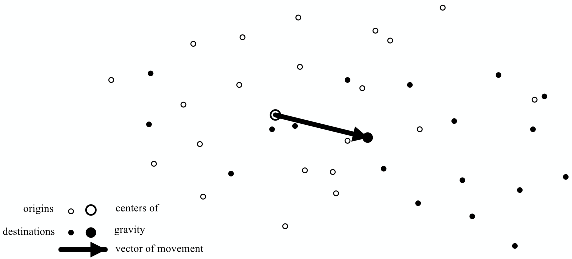
and destinations :

$$D_M = E(D_i : i \in M)$$

Vector of movement

spanned between centres of gravity:

$$\vec{V} = \overrightarrow{O_M D_M}$$



Politechnika Krakowska
im. Tadeusza Kościuszki

Data synthesis

From the daily mobility we analyse the trips of the

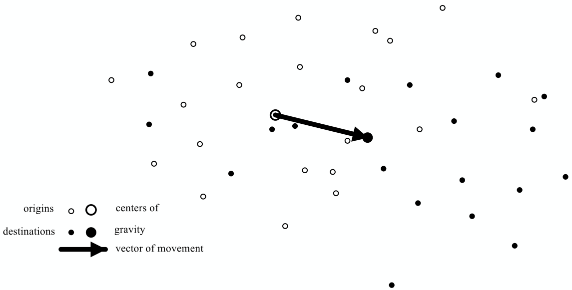
- AM
- PM

peaks separately.

Synthesis

Daily mobility pattern is the synthesized into two vectors of movement:

$$M \rightarrow \{\vec{V}_{AM}, \vec{V}_{PM}\}$$



Data synthesis

Synthesis

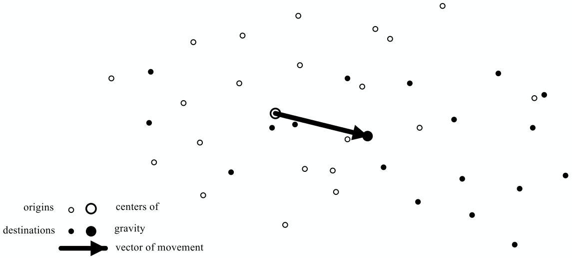
Daily mobility pattern is the synthesized into two vectors of movement:

$$M \to \{\vec{V}_{AM}, \vec{V}_{PM}\}$$

Hypothesis

Such synthetic representation of mobility is sufficient (will capture day-to-day differences in patterns).

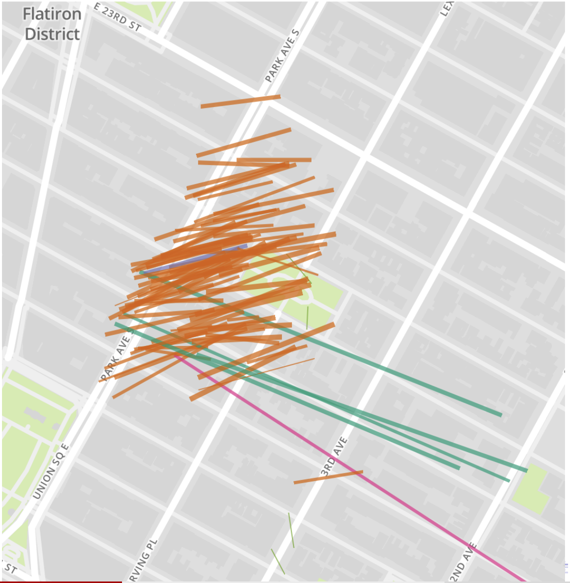
If it does we will use it, otherwise let's try more dimensions (e.g. number of trips, temporal profile, day-of-week, . . .)



Politechnika Krakowska im. Tadeusza Kościuszki

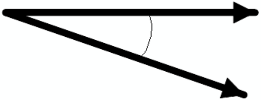
Similarity (inverse of distance) measure

when vectors are similar?



Similarity

cosine similarity



Cosine similarity

returns similarity from range 0 to 1, 1 for vectors of equal length and direction.
It is both direction and length sensitive, not location sensitive though.

$$s(\vec{V}, \vec{V}`) = \frac{\vec{V} \cdot \vec{V}`}{|\vec{V}||\vec{V}'|}$$

Politechnika Krakowska
im. Tadeusza Kościuszki

Similarity

pairwise distance

Finally, we can introduce the pairwise distance measure between two days:

$$d(V, V') = \alpha \cdot S(\vec{V}^{AM}, \vec{V'}^{AM}) + (1 - \alpha) \cdot S(\vec{V}^{PM}, \vec{V'}^{PM}),$$

with $\boldsymbol{\alpha}$'s being normalized weights, treated as a parameters of the procedure (we use default $\alpha = 0.5$ in the case-study).

Application

Such metrics can be applied for most of clustering packages.

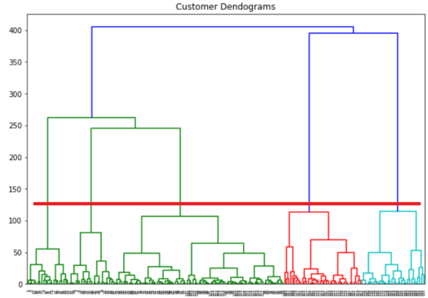
Politechnika Krakowska
im. Tadeusza Kościuszki

Clustering

unsupervised learning

Clustering

task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).



Clustering

quality, parameters, algorithm

Quality

Silhouette score - within-cluster consistence, compactness,
Calinski-Harabasz score - ratio between the within-cluster dispersion and the between-cluster dispersion
internal, self-validation without reference to unknown ground-truth

Parameters

- **nClusters** - arbitrary
- **pair-wise distance** - trail-and-error
- **algorithm**
  - various algorithms,
  - not very formalized - more procedural,
  - results sensitive to parameters,
  - we took `AgglomerativeClustering` from `python scikit-learn`

Politechnika Krakowska
im. Tadeusza Kościuszki

Clustering

validation

Good clustering

Clusters are valid when they well explain differences between groups. Hopefully not covered with distance measure - external validation.

We validate clustering by looking how it reproduces differences in:

- total number of trips (volume)
- temporal profile of trips
- day-type (holiday, working day, weekday)
- weather (hard to quantify)

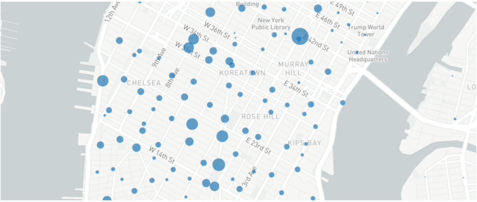
Politechnika Krakowska
im. Tadeusza Kościuszki

Results

Data

input

Stations of the system with their capacities

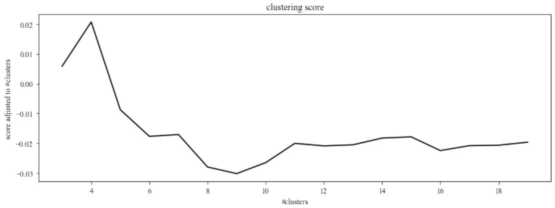


- 120 days
- 6 000 000 trips
- 1 GB of data
- preprocessed by the provider
- light .csv files (few redundant and heavy columns)
- `https://s3.amazonaws.com/tripdata/201804-citibike-tripdata.csv.zip`

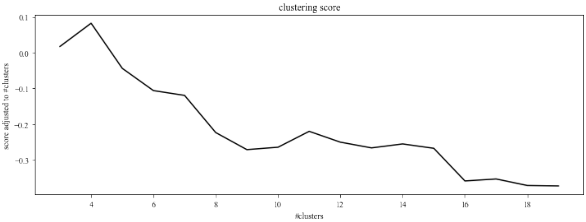
Politechnika Krakowska
im. Tadeusza Kościuszki

Results

determining cluster numbers



silhouette coefficient

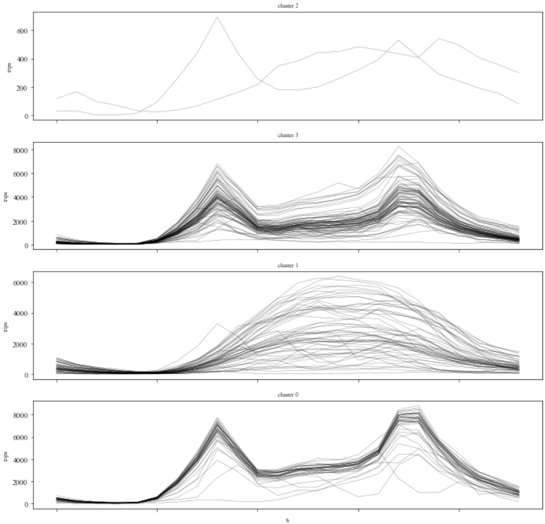


adjusted silhouette coefficient

Politechnika Krakowska
im. Tadeusza Kościuszki

Number of clusters

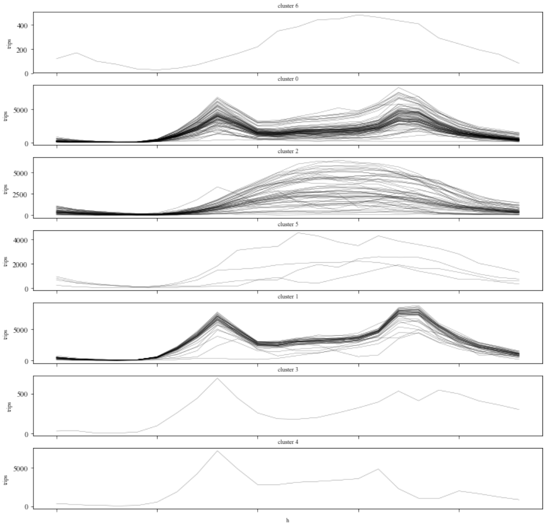
validation by temporal profiles



4 clusters

Number of clusters

validation by temporal profiles

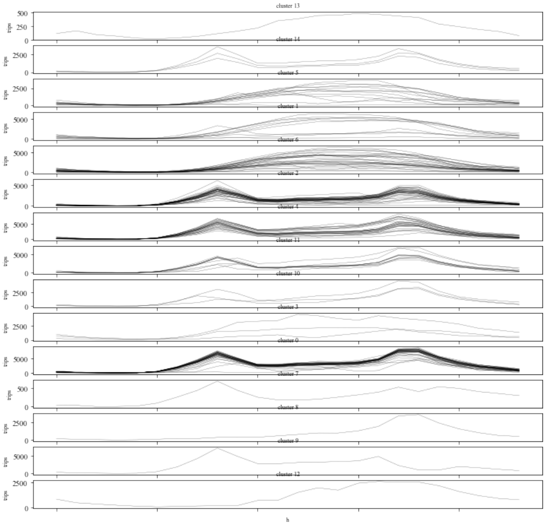


8 clusters

Politechnika Krakowska
im. Tadeusza Kościuszki

Number of clusters

validation by temporal profiles

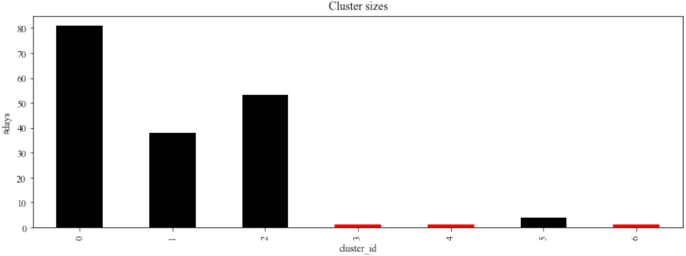


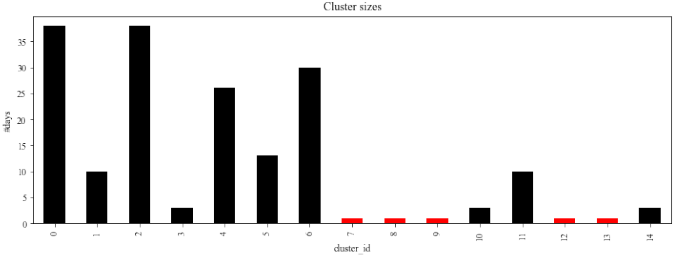
15 clusters

Politechnika Krakowska
im. Tadeusza Kościuszki

Number of clusters

clusters or outliers?





Politechnika Krakowska
im. Tadeusza Kościuszki

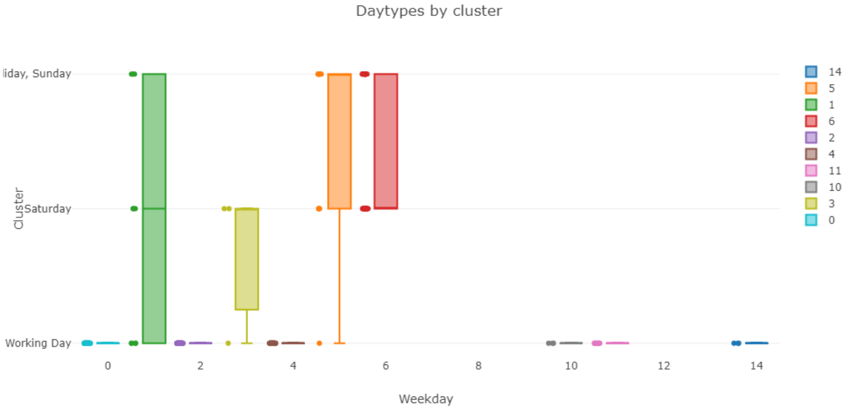
Validation

are weekdays captured?

Politechnika Krakowska
im. Tadeusza Kościuszki

Validation

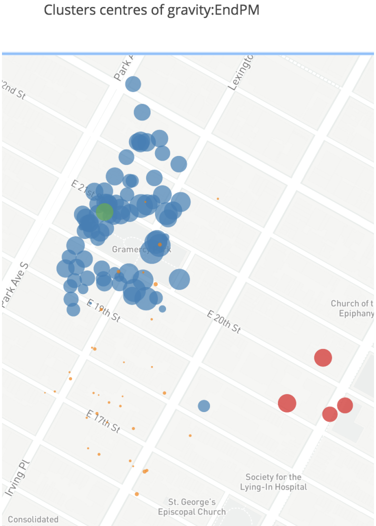
are holidays captured?



Politechnika Krakowska
im. Tadeusza Kościuszki

Validation

spatial and trip volume explanation

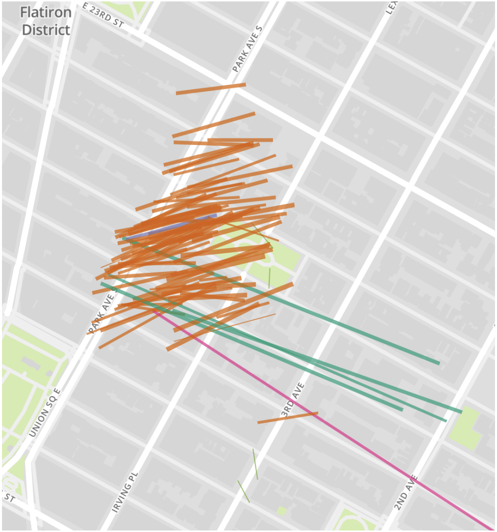


dot size is number of trips

Politechnika Krakowska
im. Tadeusza Kościuszki

Validation

vectors of movement - do they have explanatory power?



Politechnika Krakowska
im. Tadeusza Kościuszki

Conclusion

Politechnika Krakowska
im. Tadeusza Kościuszki

Conclusion

- big-data can on mobility can be synthesized (center of gravity, vector of movement)
- pair-wise distance between mobility patterns can be introduced (cosine similarity)
- thanks to this we can try to cluster mobility patterns
- vectors of movement compared with cosine similarity cluster mobility well and capture variation of: daytype, weekday, temporal profile, number of trips, spatial distribution of centres of gravity
- opensource GitHub repo `github.com/RafalKucharskiPK/clustering_mobility_data.git`

limitations

New York tested only, other cities might be strongly centric and vectors become null.
possibly some details may not be covered (single station profiles trip demand).

further directions

real-time prediction (R. Kucharski, G. Cantelmo, A. Drabicki - Submitted for MT-ITS2019 Kraków)

Thank you for your attention

Rafał Kucharski
rkucharski at pk.edu.pl
Politechnika Krakowska, Kraków, Poland

abstract deadline 31 Oct via www.mt-its2019.pk.edu.pl

Politechnika Krakowska
im. Tadeusza Kościuszki