# Symbolic networks and big data

## bike sharing data

Vladimir Batagelj

IMFM Ljubljana, IAM UP Koper and University of Ljubljana

**CRoNoS COST Action Working Groups Meeting**
Institute Tinbergen, Amterdam, Netherlands
1-2. September 2017

# Outline



1 Big data and aggregation
2 Kaggle
3 Data sets
4 Analyses
5 Conclusions
6 References

**Vladimir Batagelj**: vladimir.batagelj@fmf.uni-lj.si

**Last version of slides (September 1, 2017, 14 : 11):** SymNet.pdf

# Divide and conquer

A "standard" approach to deal with large structures is the *divide and conquer* strategy that (recursively) breaks down a large structure to smaller, manageable sub-structures of the same or related type.

Let $\mathcal{U}$ be a *set of units*, $u \in \mathcal{U}$ a *unit*, and $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$, $\emptyset \subset C_i \subseteq \mathcal{U}$ a *partition* of $\mathcal{U}$ – it holds: $\bigcup_i C_i = \mathcal{U}$ and $i \neq j \Rightarrow C_i \cap C_j = \emptyset$.

In classic data analysis the units are usually described as lists of (measured) values of selected variables ( properties, attributes)

$$X(u) = [x_1(u), x_2(u), \ldots, x_m(u)]$$

collected into a *data frame* $X$.

# Aggregation

The *aggregation* of a cluster $C$ is again a list of values

$$Y(C) = [y_1(C), y_2(C), \ldots, y_m(C)]$$

where $y_i C)$ is the aggregated value of the set of values $\{x_i(u) : u \in C\}$ – forming an *aggregated data frame* $Y$.
For example, for $x_i$ measured in a numerical scale their average is usually used

$$y_i(C) = \frac{1}{|C|} \sum_{u \in C} x_i(u)$$

Different aggregation functions are available – see Beliakov, Pradera, Calvo: Aggregation Functions, 2007.

# Notes

1. This kind of aggregation can produce a big loss of information.
2. It is not compatible with the recursive decomposition; but keeping $(|C|, y_i(C))$ is – for $C_1 \cap C_2 = \emptyset$ we have

$$y_i(C_1 \cup C_2) = \frac{|C_1|y_i(C_1) + |C_2|y_i(C_2)}{|C_1| + |C_2|}$$

3. The aggregated descriptions need not to be from the same "space" as the descriptions of units.

# Symbolic data analysis

Edwin Diday proposed in late eighties an approach, named symbolic data analysis (SDA), in which the (aggregated) descriptions can be *symbolic objects* (SO) such as: an interval, a list, a histogram, a distribution, etc. We get *symbolic data frames*. See Billard, Diday: Symbolic Data Analysis, 2012.

Using this approach we can reduce a big data frame into small, manageable symbolic data frame and preserve much more information. To analyze symbolic data frames new methods have to be developed – SDA.

# Discrete distributions

We found very interesting the representation with *discrete distribution*.

The range $V$ of a variable $x_i$ is partitioned into $k_i$ subsets $\{V_j\}$. Then $y_i(C) = [y_{i1}(C), y_{i2}(C), \ldots, y_{ik_i}(C)]$ where $y_{ij}(C) = |\{u : x_i(u) \in V_j\}|$.

The description based on a discrete distribution enables us to consider variables that are measured in different types of measurement scales and based on a different number of original (individual) units. It is also compatible with recursive decomposition.

# Kaggle

Some time ago I found on Kaggle

https://www.kaggle.com/benhamner/sf-bay-area-bike-share

a contest dealing with an analysis of data on bike sharing system in the San Francisco Bay Area. After some searching it turned out that similar data sets are available for several cities around the world (mainly in US).

# Some Open data sets on Bike Sharing Systems

## on my disk

| Bike sharing | City | data available | # of trips |
|---|---|---|---|
| Capital | Washington, D.C. | 2010/10-2016/09 | 14691090 |
| Hubway | Boston | 2011/07-2016/06 | 3930659 |
| Divvy | Chicago | 2013/01-2016/06 | 7867601 |
| Citi Bike | New York | 2013/07-2016/09 | 33319019 |
| BABS | San Francisco | 2013/08-2016/08 | 983648 |
| Healthy Ride | Pittsburgh | 2015/07-2016/09 | 118422 |
| Indego | Philadelphia | 2015/04-2016/09 | 673703 |
| NiceRide | Minnesota | 2010/06-2015/12 | 1808452 |
| Santander C. | London | 2015/01-2016/11 | 19212558 |

# Data about stations

The Stations file is a snapshot of station locations and capacities during the reporting time interval:

- Station ID
- Station name
- Lat/Long coordinates
- Number of individual docking points at each station

In some cases also the data about station elevantions are available.

North American Bike Share Association's open data standard – gbfs
General Bikeshare Feed Specification; Systems using gbfs.

Most of the systems provide a feed service returning a JSON file with current status of stations.
Divvy, Indego, CitiBike stations: info, status

# Reading station status in R

```
wdir <- "C:/Users/batagelj/data/bikes/philly"
setwd(wdir)
stat <- "https://gbfs.bcycle.com/bcycle_indego/station_status.json"
num <- 0
setInternet2(use = TRUE)
p1 <- proc.time()
while (num < 5){
   num <- num+1
   fsave <- paste('status_',as.character(num),'.json',sep='')
   test <- tryCatch(download.file(stat,fsave,method="auto"),
                    error=function(e) e)
   Sys.sleep(60)
   p2 <- proc.time()
   cat(p2 - p1,'\n'); flush.console()
   p1 <- p2
}
```

# Data about trips

Each trip is anonymized and includes:

- Bike number
- Trip start day and time
- Trip end day and time
- Trip start station
- Trip end station
- Rider type

In some cases additional data are available: Gender, Year of birth.

# Additional data sources

**Weather**

For cities in US we can get the weather data at NOAA, Quality Controlled Local Climatological Data

Precipitations, wind, temperature, humidity, pressure.

**Maps**

The ESRI shape files descriptions of maps can be found using Google.

Boston, Bay Area Cities, New York, Pittsburgh

## Large temporal and spatial network data.

There were some contests for analysing of bike sharing data. Some interesting observations were presented. Also some blogs and papers were written on this topic.

In December 2016 there were 100 hits in WoS to the query `"bike sharing system*"`.

# Analyses

Different overall distributions:
Pitts; Bay; Boston; NYC BSS

Impact of weather: temperature (day/night, winter), precipitations.

Cycles: year (temperature), week (working days/weekend), day (hours, parts of the day): week; days in a week

Other factors: subscriber/customer, trip duration, gendre, rider's age, speed, elevation: age

The moves of bikes among stations by the system can be recognized as those rides where the bike's next trip started at a different station from where the previous trip dropped off.
Arrivals/departures; Boston; Changes

Prediction: SF Bay Area: count prediction

# Analyses

We find especially interesting a blog by

Todd W. Schneider: A Tale of Twenty-Two Million Citi Bike Rides: Analyzing the NYC Bike Share System

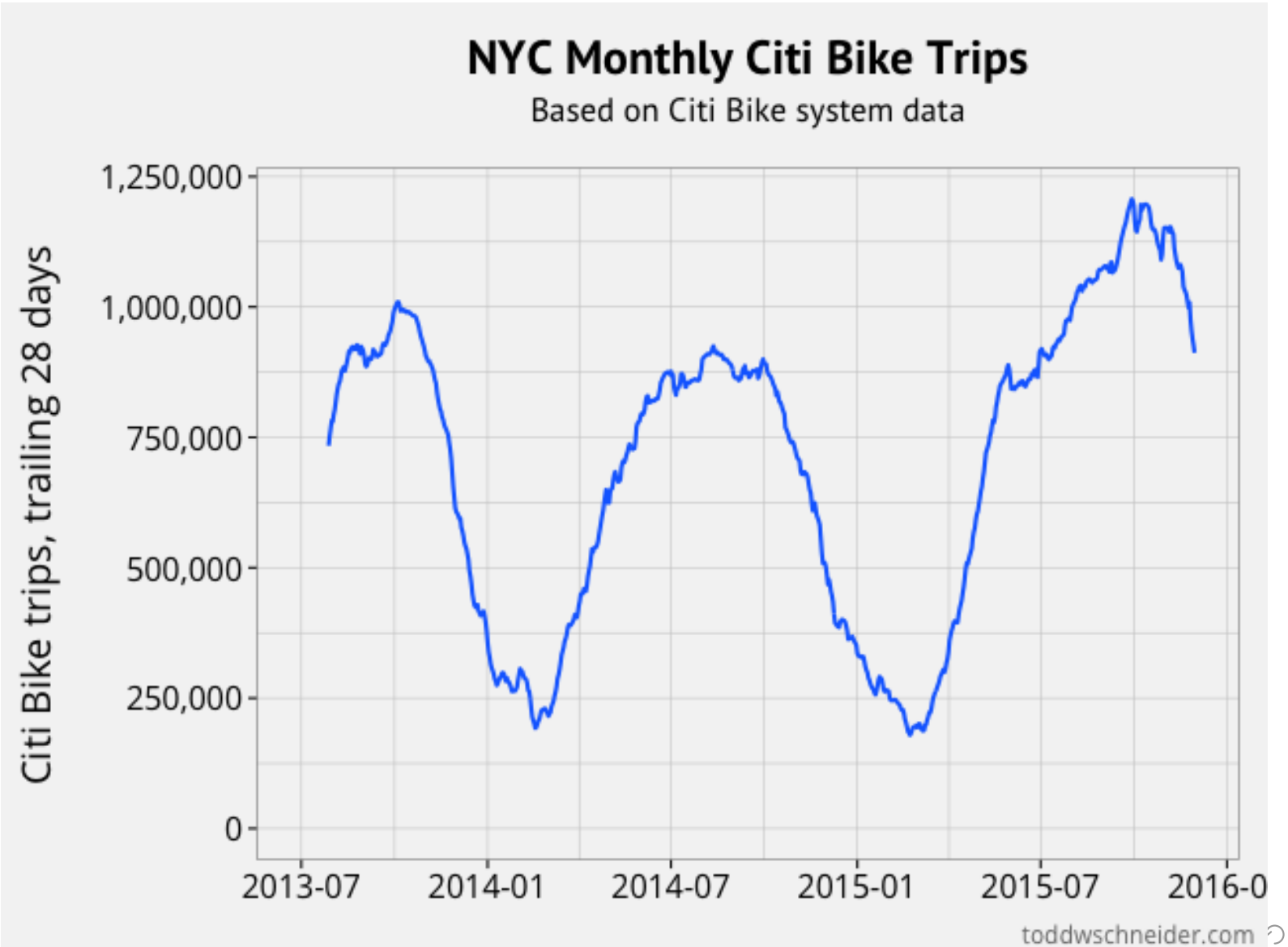
and

Jackson Whitmore: What's happening with Healthy Ride?, April 2016.

In the following slides we present some results from them.

# Year / Winter

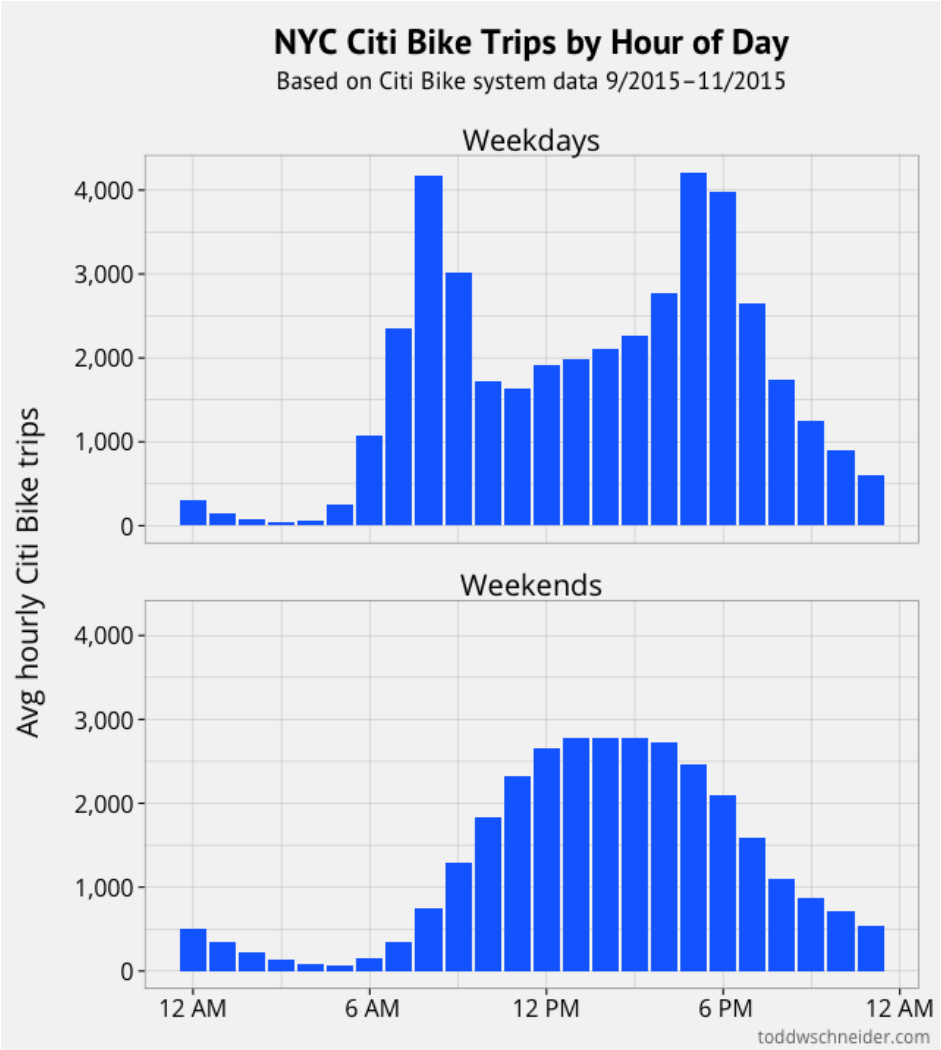
## by Todd W. Schneider



# Working days / Weekend

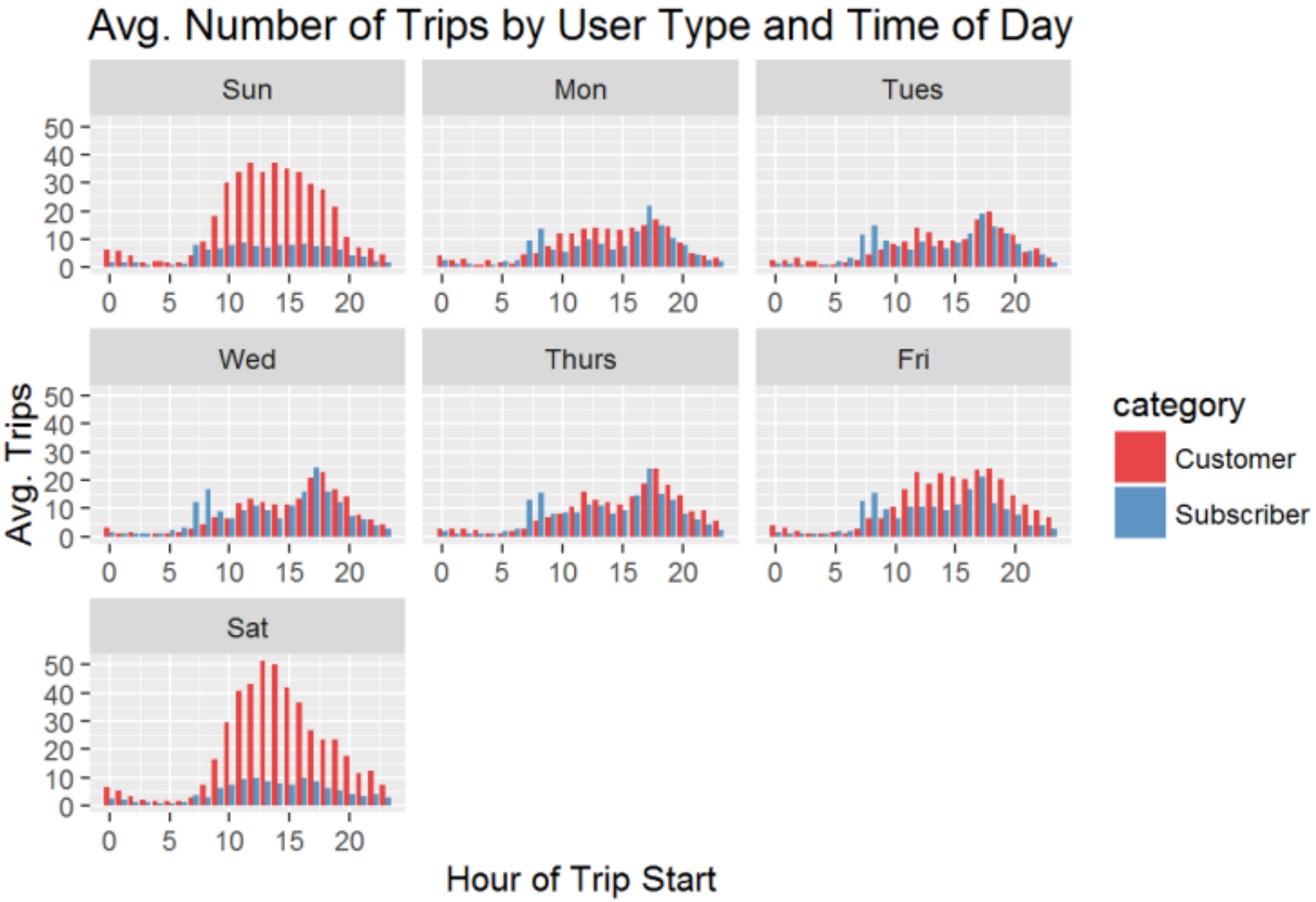
## by Todd W. Schneider





# Subscribers / Custumers

by Jackson Whitmore



# Bike sharing data and networks

The bike sharing data can be viewed as a spatial and temporal network:
Nodes – stations: name, location, capacity, (state)
Links – trips: from, to, start time, finish time, bike's id, rider type, gender, age

From this basic network we can construct several *derived* (aggregated) networks.

In most systems the data about nodes are static – fixed for longer period of time. It could be possible to collect these data using feeds.

Selecting an appropriate granulation (5 min, 15 min, 1 hour, part of a day, day, week, month, quartal, year) and some restrictions (rider type, gender, age, ...) we get the corresponding frequency distributions in nodes and on links.

# Symbolic networks

Assigning distributions to nodes and links we get a *symbolic network*.

There are different distributions on links:
**departures**: (# of trips starting in selected time interval),
**activity**: (# of trips active in selected time interval),
**duration**: (# of trips with duration in selected time interval), etc.

and in nodes, for example:

**departures**: the sum of link distributions for incident links,
**imbalance**, etc.

# Our analysis

NY Citi Bike one year data from October 2015 to September2016.
13 266 296 trips, 678 stations.

The Citi Bike system had an expansion in August 2015.

We constructed a departures network with daily distributions with half hour granulation.

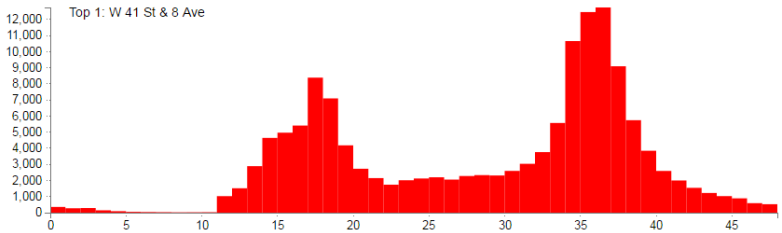
First we looked for extreme elements (links or nodes).

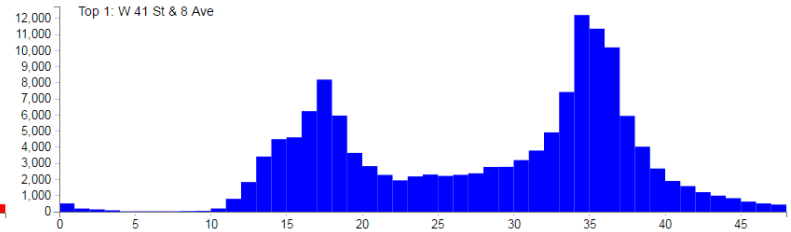
In a selected time interval:
$flow(u, v) = \#$ of trips starting in a node $u$ and finishing in a node $v$
$out(v) = \#$ of trips starting in a node $v$
$in(v) = \#$ of trips finishing in a node $v$
$flow(u, v; k) =$
$\#$ of trips starting in a node $u$ in the$k$-th half hour and finishing in a node $v$
...

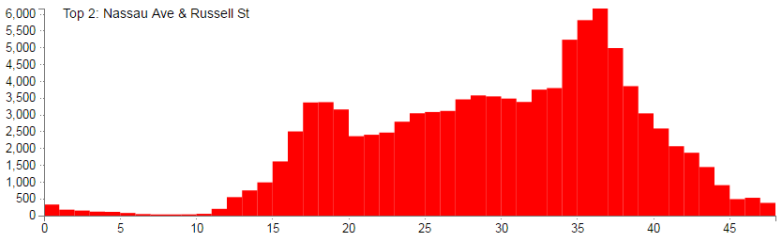
# The most active stations / Top 3

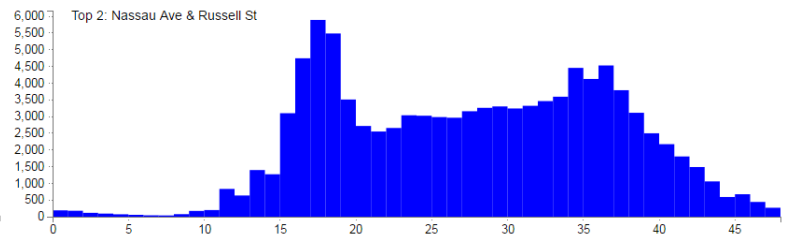
$activity(v) = out(v) + in(v)$

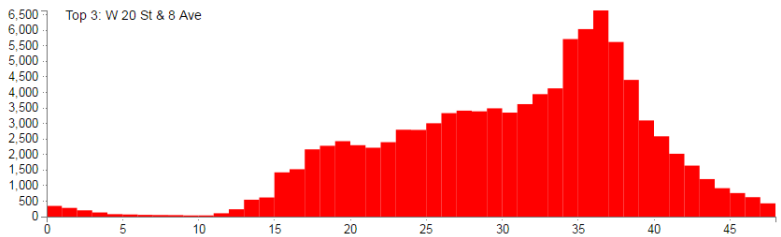


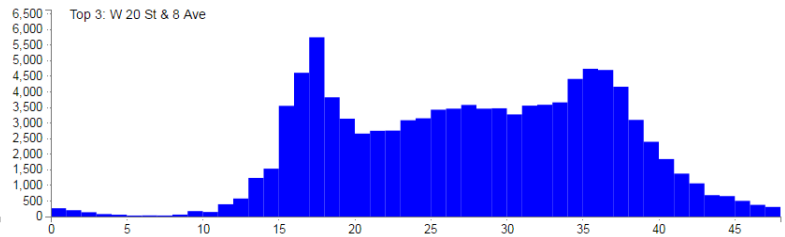



| n | station | trips | n | station | trips |
|---|---|---|---|---|---|
| 1 | W 41 St & 8 Ave | 281996 | 6 | W 45 St & 8 Ave | 170593 |
| 2 | Nassau Ave & Russell St | 203855 | 7 | W 38 St & 8 Ave | 164378 |
| 3 | W 20 St & 8 Ave | 200629 | 8 | E 14 St & Avenue B | 163962 |
| 4 | W 16 St & The High Line | 196414 | 9 | E 53 St & Madison Ave | 162828 |
| 5 | W 22 St & 8 Ave | 188394 | 10 | W 53 St & 10 Ave | 161931 |

# Imbalance

In a selected time interval:

$$diff(v) = out(v) - in(v)$$

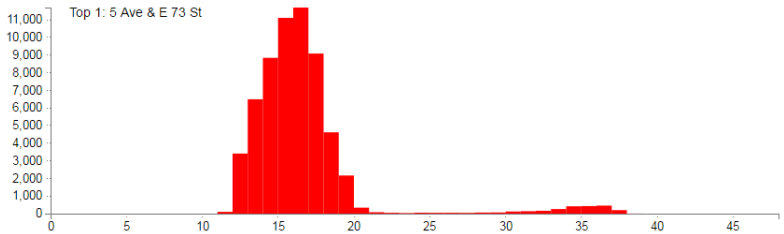
$$fDist(v) = \sum_{k=1}^{48} |out(v;k) - in(v;k)|$$

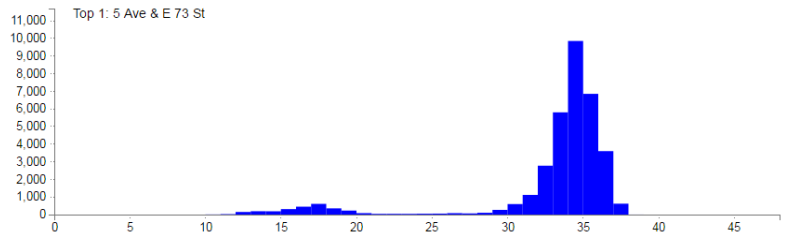
| n | station | out | in | diff | station | fDist |
|---|---|---|---|---|---|---|
| 1 | 5 Ave & E 73 St | 60524 | 34559 | 25965 | 5 Ave & E 73 St | 84703 |
| 2 | Van Vorst Park | 29962 | 14920 | 15042 | Fulton St & William St | 66453 |
| 3 | 8 Ave & W 33 St | 57127 | 67592 | -10465 | E 75 St & 3 Ave | 51297 |
| 4 | W Broadway & Spring St | 15217 | 23544 | -8327 | W 22 St & 8 Ave | 50530 |
| 5 | E 51 St & 1 Ave | 72651 | 80783 | -8132 | E 33 St & 2 Ave | 47893 |
| 6 | E 75 St & 3 Ave | 56302 | 48891 | 7411 | Water - Whitehall Plaza | 45554 |
| 7 | Catherine St & Monroe St | 36858 | 29455 | 7403 | E 51 St & 1 Ave | 34086 |
| 8 | E 45 St & 3 Ave | 48116 | 41601 | 6515 | W 37 St & 10 Ave | 33865 |
| 9 | Water - Whitehall Plaza | 71364 | 65638 | 5726 | Cambridge Pl & Gates Ave | 32562 |
| 10 | 6 Ave & Canal St | 23473 | 28451 | -4978 | E 16 St & Irving Pl | 30293 |

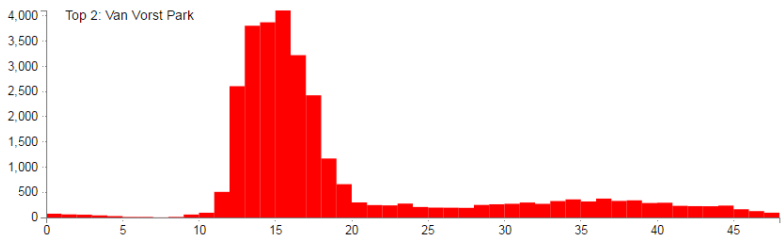
# Imbalance / diff

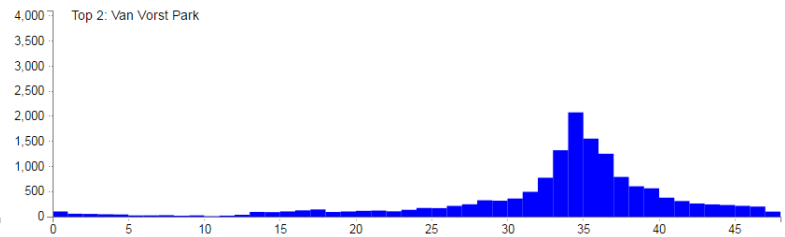
## Top 4

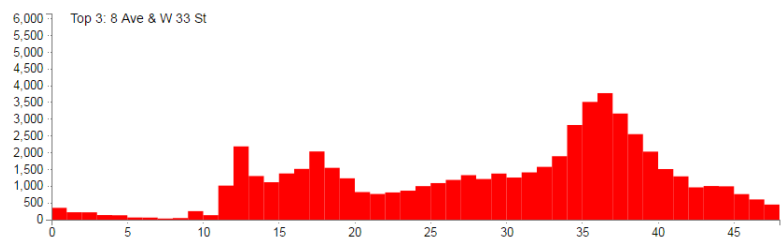


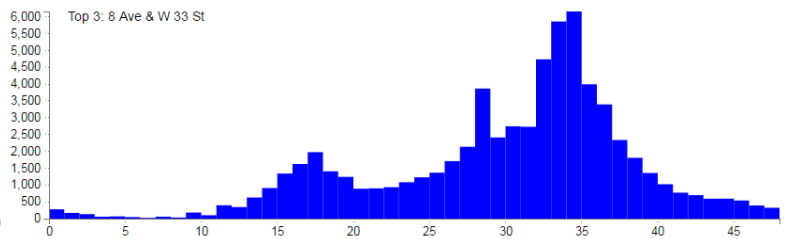


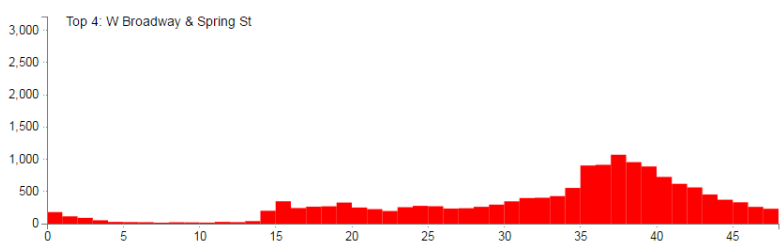



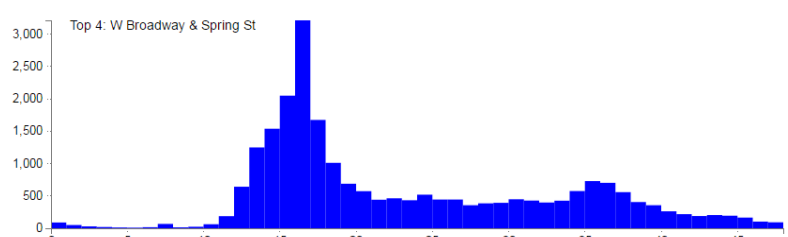



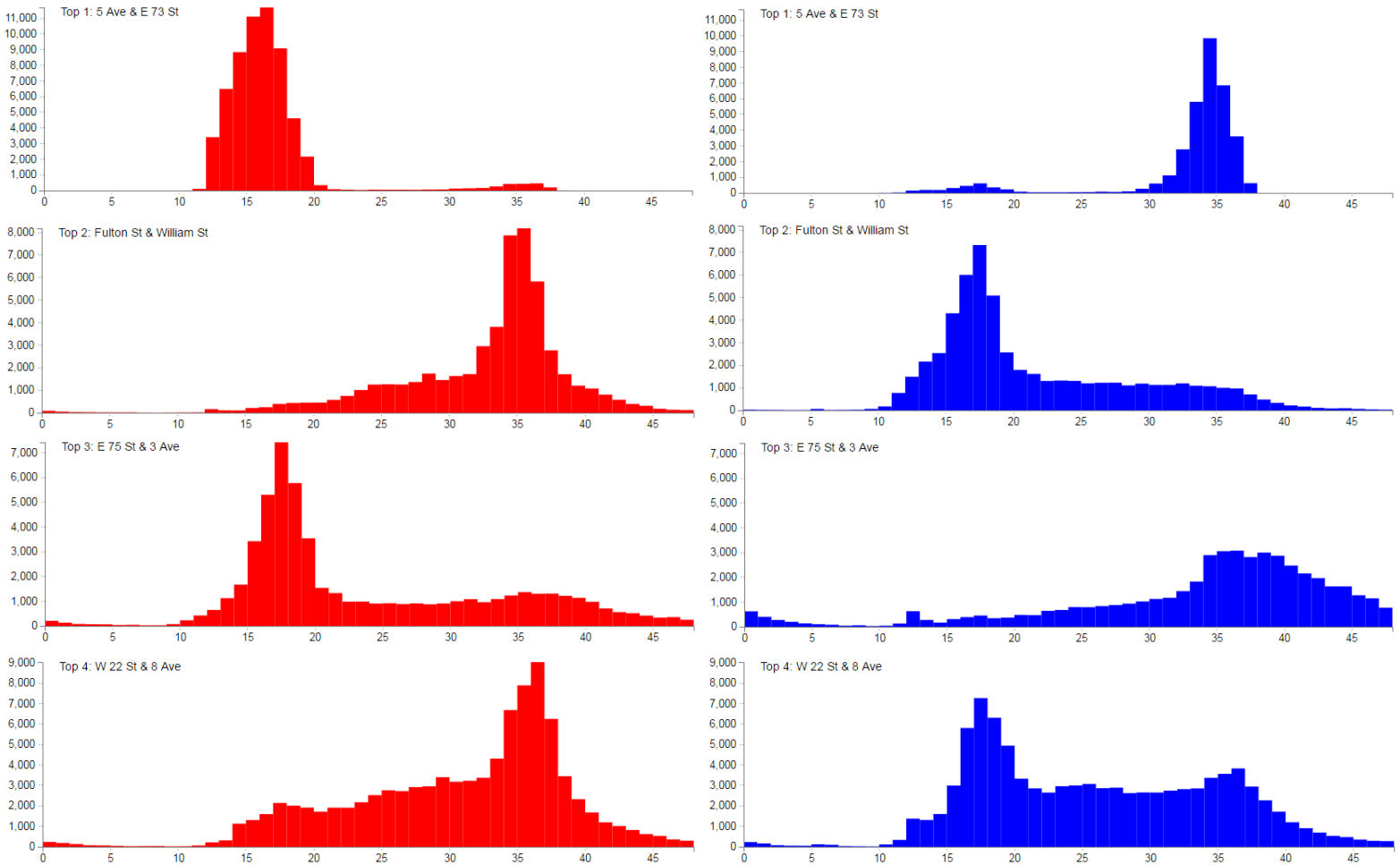



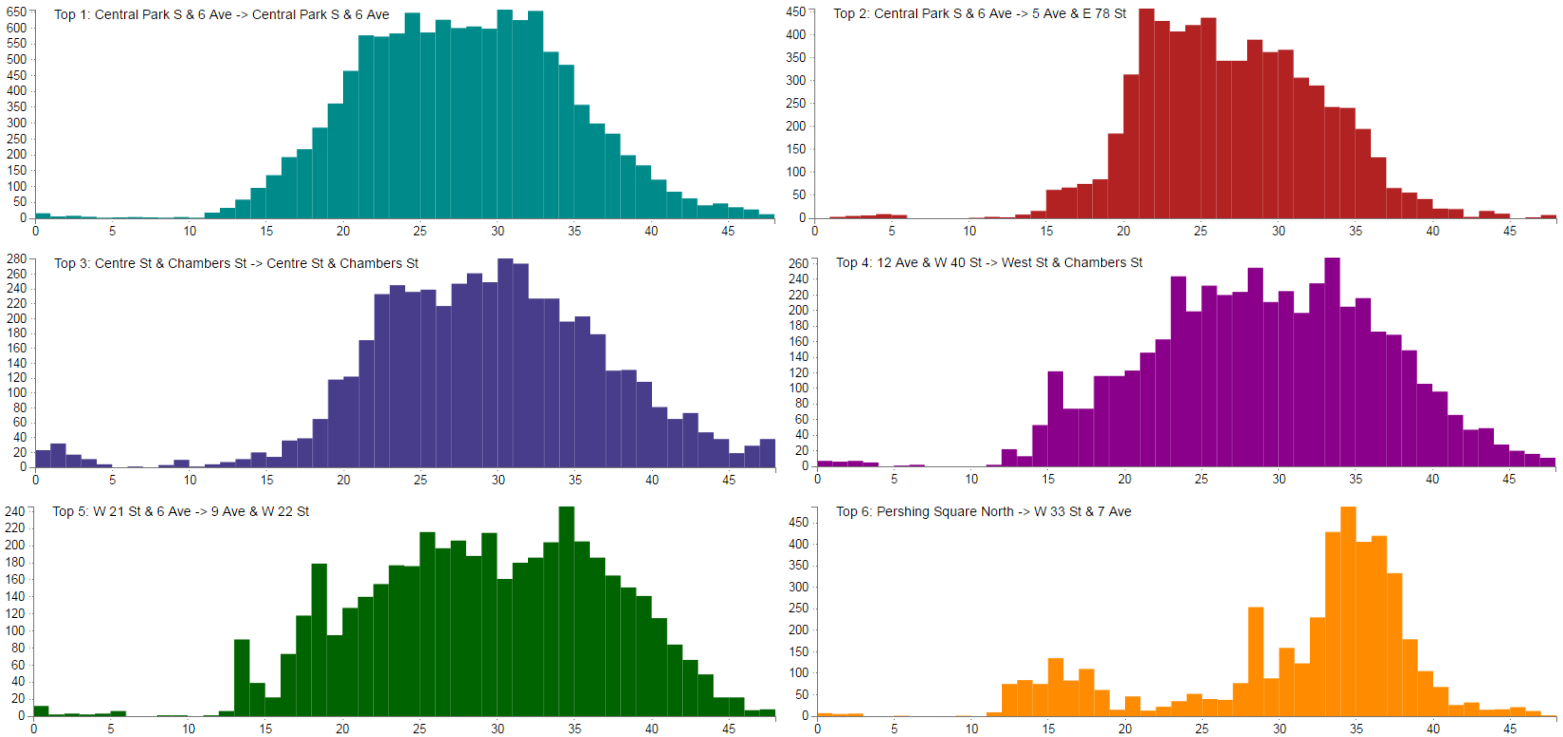
# Imbalance / fDist

## Top 4



# The largest flows / Top 6





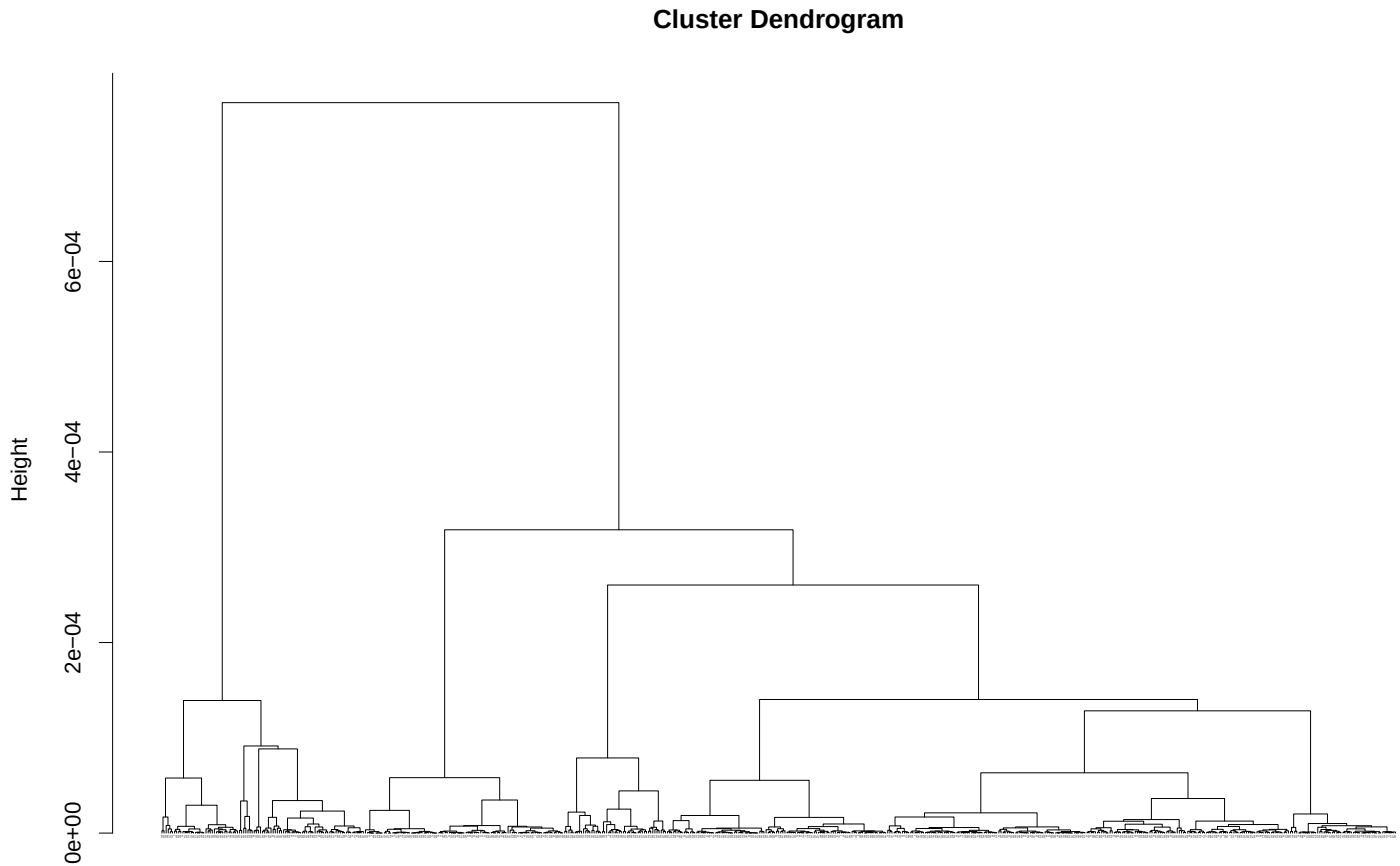
# Clamix – Clustering modal valued symbolic data

Two clustering methods for symbolic objects are implemented: the adapted leaders method and the adapted agglomerative hierarchical clustering Ward's method.
Clamix: R-forge, doc
Paper: V. Batagelj, N. Kejžar, and S. Korenjak-Černe. Clustering of Modal Valued Symbolic Data. ArXiv e-prints, 1507.06683, July 2015.

We clustered the set of 589 links with flow at least 1250. This gives us typical flow distribution shapes.

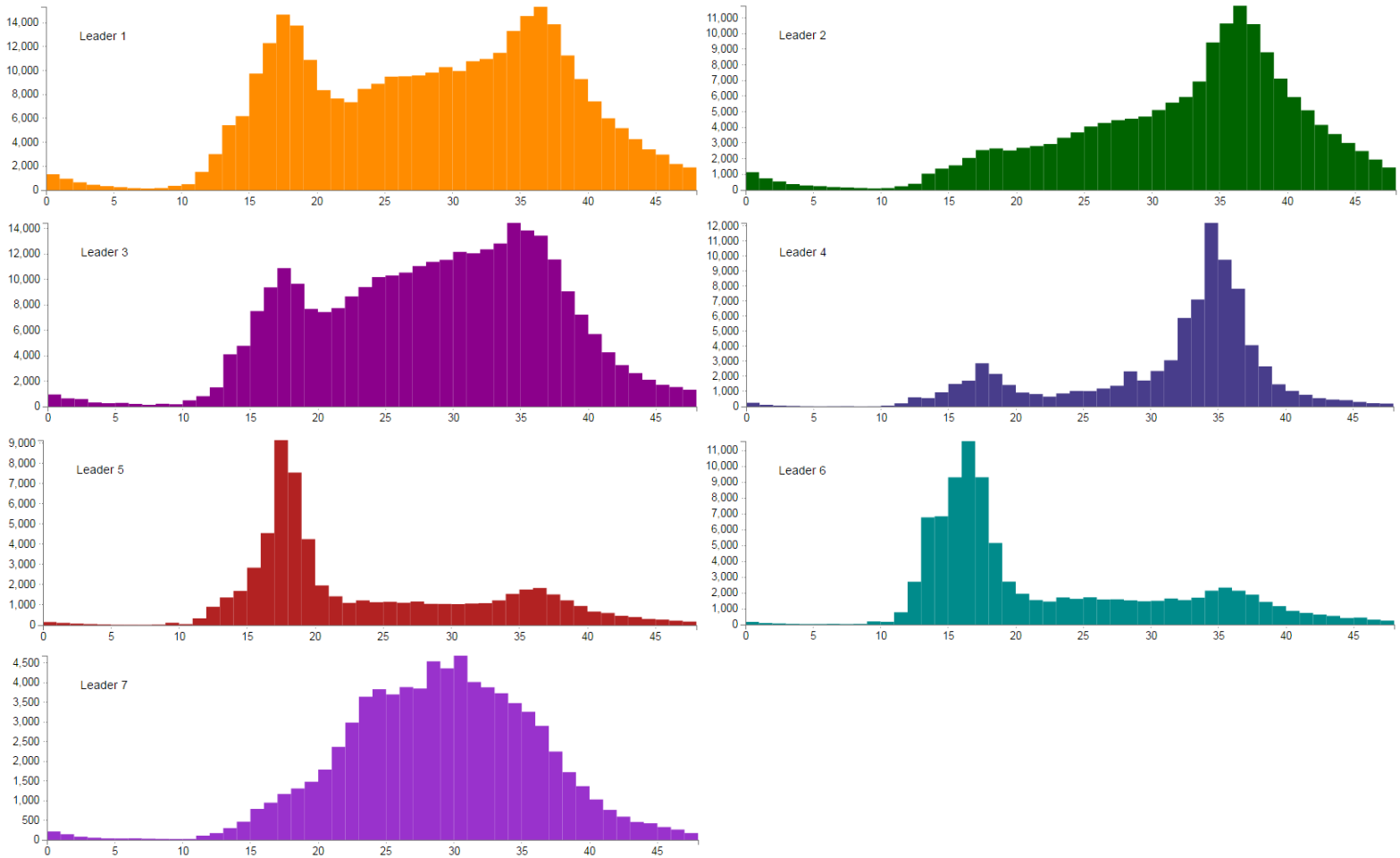
# Clustering of flows





# Clustering of flows / 7 clusters





# Conclusions

- bike sharing data are an interesting type of data,
- prepare some extended data sets; get or collect the dynamic stations data,
- additional analyses:
  - other symbolic objects: nodes (in and out distribution), links (subscriber, custumer distribution), ...
  - stability of distribution shape through time
  - ...
- compare bike sharing systems
- Taxi (Yellow and Green) and Uber data are available for New York.

# References I

1. V. Batagelj, N. Kejžar, and S. Korenjak-Černe. Clustering of Modal Valued Symbolic Data. ArXiv e-prints, 1507.06683, July 2015.
2. Gleb Beliakov, Ana Pradera, Tomasa Calvo: Aggregation Functions: A Guide for Practitioners (Studies in Fuzziness and Soft Computing). Springer, Berlin, Heidelberg 2007.
3. Lynne Billard; Edwin Diday (14 May 2012). Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley & Sons.
4. Bay Area Bike Share: San Francisco Bay Area - Kaggle challenge, Open data, challenge
5. Todd W. Schneider: A Tale of Twenty-Two Million Citi Bike Rides: Analyzing the NYC Bike Share System.
6. Jackson Whitmore: What's happening with Healthy Ride?, April 2016.

# Acknowledgments

This work was supported in part by the Slovenian Research Agency (research programs P1-0294 and research projects J7-8279 and J1-6720).

The author's attendance of the meeting was supported by the COST Action IC1408 – CRoNoS.