ANALYSIS OF
NYC SERVICE REQUEST CALLS TO 311

Outline

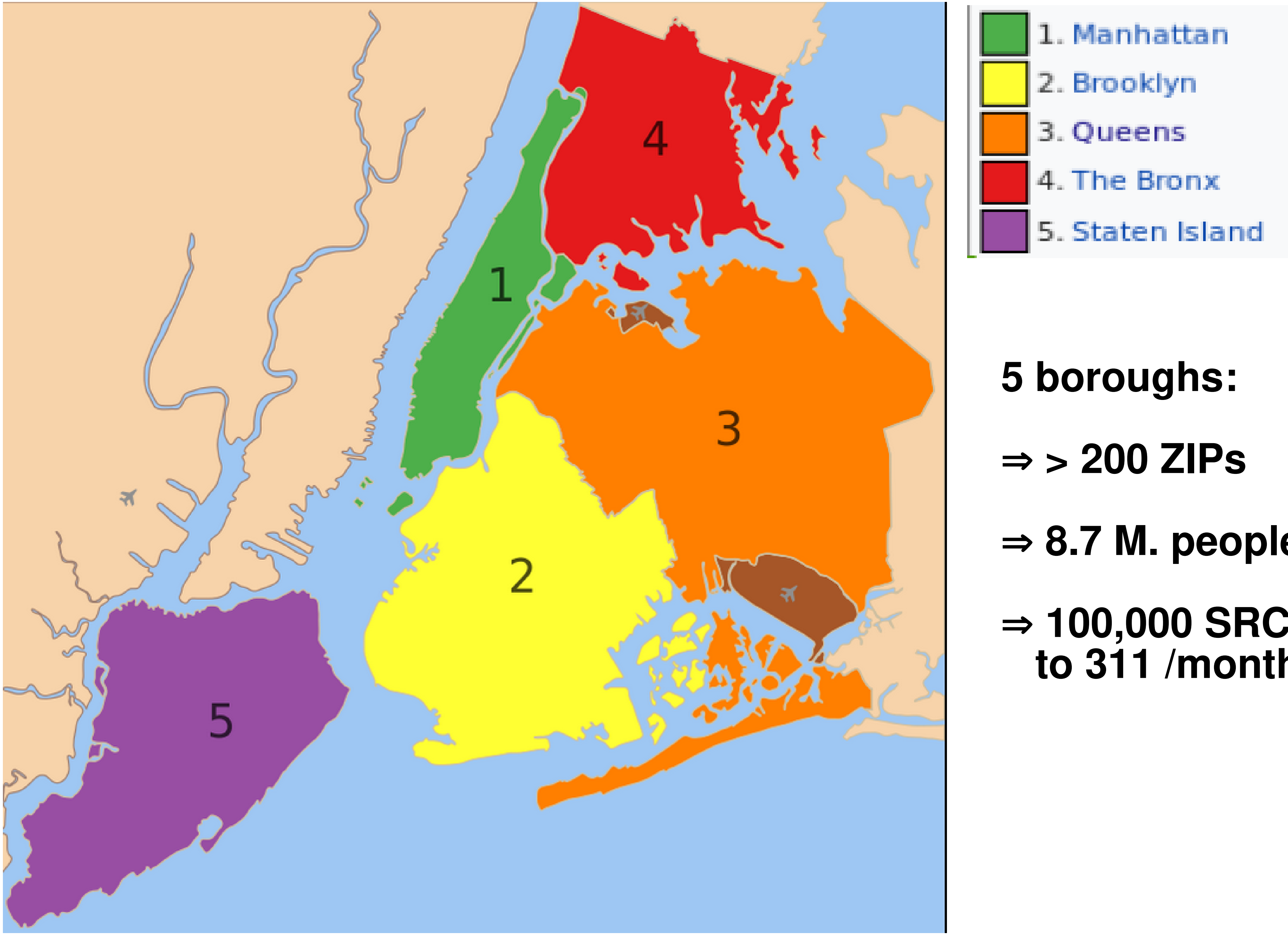
- Data ETL
- MVA
    - CA, PCA, MCA
    - Clustering, Tree classification
- Conclusions

By: **Cedric Bhihe** <cedric.bhihe@gmail.com> **Santi Calvo** <s.calvo93@gmail.com>

Data from:

- NYC OpenData
- NYPD DB
- IRS
- US Census Office

All material and code available at:
https://www.github.com/Cbhihe/nyc311



5 boroughs:

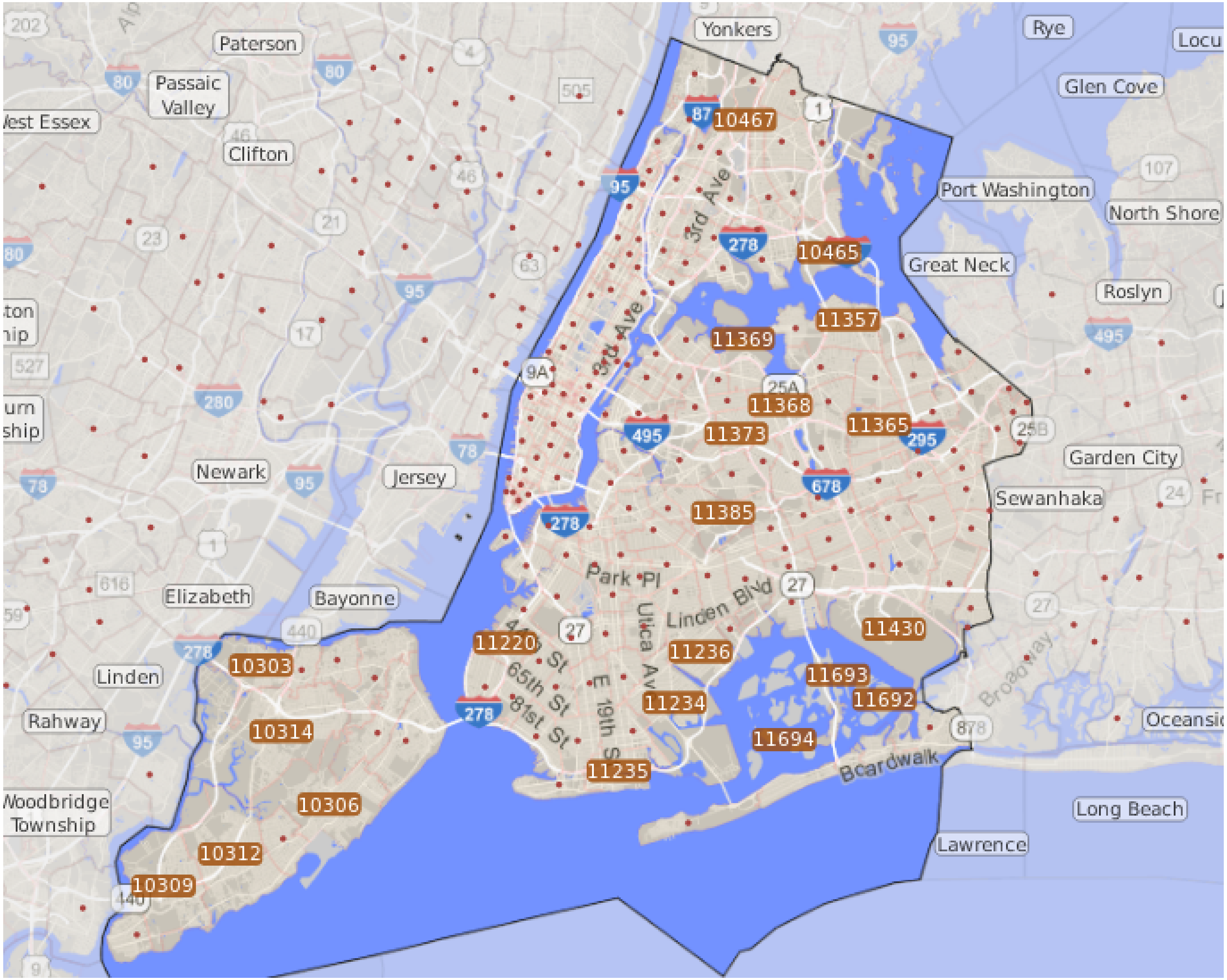
⇒ > 200 ZIPs

⇒ 8.7 M. people

⇒ 100,000 SRC to 311 /month

Objective(s):

- to explore data with MVA tools
- to extract features and descriptive information, so we may:
  - detect trends
  - optimize urban resources



Paterson
Passaic Valley
Clifton
Yonkers
Rye
Glen Cove
Port Washington
North Shore
Great Neck
Roslyn
Garden City
Sewanhaka
Newark
Jersey
Elizabeth
Bayonne
Linden
Rahway
Long Beach
Lawrence
10467
10465
11357
11369
11368
11373
11365
11385
11430
11220
11236
11693
11692
11234
11694
11235
10303
10314
10306
10312
10309
3rd Ave
Park Pl
Linden Blvd
Utica Av
Boardwalk
Broadway

Data ETL – extract

ZIP	Date	Complaint	Descriptor	Address	planeX	planeY	GPS
10463	2014-04-01	PAINT/PLASTER	CEILING	1 JACOBUS PLACE	1008544	258129	(40.875145400143914, -73.9121540
10463	2014-04-01	HEAT/HOT WATER	ENTIRE BUILDING	3631 CORLEAR AVENUE	1011190	261705	(40.884952723690205, -73.9025718
10038	2014-04-01	GENERAL	VENTILATION SYSTEM	41 JOHN STREET	982002	197700	(40.709317299970831, -74.0081084
10463	2014-04-01	FLOORING/STAIRS	FLOOR	110 TERRACE VIEW AVE	1008357	258602	(40.876444151331, -73.9128285342
10306	2014-04-01	Rodent	Condition Attracting Rodents	1934 NORTH RAILROAD	948358	144556	(40.563376021503466, -74.1291780
10007	2014-04-01	UNSANITARY CONDITION	GARBAGE/RECYCLING STORA	23 PARK PLACE	981818	199058	(40.713044638283755, -74.0087726
10463	2014-04-01	HEAT/HOT WATER	ENTIRE BUILDING	3810 BAILEY AVENUE	1012586	261482	(40.884336286878096, -73.8975242
10463	2014-04-01	HEAT/HOT WATER	ENTIRE BUILDING	2840 BAILEY AVENUE	1010532	257794	(40.874220233300874, -73.9049668
10463	2014-04-01	HEAT/HOT WATER	ENTIRE BUILDING	3810 BAILEY AVENUE	1012586	261482	(40.884336286878096, -73.8975242
10463	2014-04-01	HEAT/HOT WATER	APARTMENT ONLY	225 WEST 232 STREET	1010709	260028	(40.880351339751236, -73.9043180
10038	2014-04-01	HEAT/HOT WATER	APARTMENT ONLY	41 JOHN STREET	982002	197700	(40.709317299970831, -74.0081084
10463	2014-04-01	WATER LEAK	HEAVY FLOW	225 WEST 232 STREET	1010709	260028	(40.880351339751236, -73.9043180
10463	2014-04-01	PAINT/PLASTER	CEILING	1 JACOBUS PLACE	1008544	258129	(40.875145400143914, -73.9121540
10463	2014-04-01	HEAT/HOT WATER	ENTIRE BUILDING	3810 BAILEY AVENUE	1012586	261482	(40.884336286878096, -73.8975242
10463	2014-04-01	WATER LEAK	SLOW LEAK	1 JACOBUS PLACE	1008544	258129	(40.875145400143914, -73.9121540
10463	2014-04-01	PLUMBING	BASIN/SINK	110 TERRACE VIEW AVE	1008357	258602	(40.876444151331, -73.9128285342
10463	2014-04-01	PAINT/PLASTER	WALL	110 TERRACE VIEW AVE	1008357	258602	(40.876444151331, -73.9128285342
10463	2014-04-01	WATER LEAK	HEAVY FLOW	1 JACOBUS PLACE	1008544	258129	(40.875145400143914, -73.9121540
10463	2014-04-01	WATER LEAK	SLOW LEAK	1 JACOBUS PLACE	1008544	258129	(40.875145400143914, -73.9121540
10463	2014-04-01	UNSANITARY CONDITION	MOLD	110 TERRACE VIEW AVE	1008357	258602	(40.876444151331, -73.9128285342
10463	2014-04-01	HEAT/HOT WATER	ENTIRE BUILDING	140 VAN CORTLANDT A	1013473	261850	(40.885343438331205, -73.8943148
10463	2014-04-01	HEAT/HOT WATER	ENTIRE BUILDING	3810 BAILEY AVENUE	1012586	261482	(40.884336286878096, -73.8975242
10463	2014-04-01	HEAT/HOT WATER	ENTIRE BUILDING	3555 OXFORD AVENUE	1009450	262094	(40.886025546641576, -73.9088630
10463	2014-04-01	HEAT/HOT WATER	APARTMENT ONLY	3150 BAILEY AVENUE	1011414	259460	(40.878790215257844, -73.9017708
10463	2014-04-01	Rodent	Rat Sighting	2738 KINGSBRIDGE TER	1010952	256931	(40.871850303847744, -73.9034516
10463	2014-04-01	PAINT/PLASTER	CEILING	110 TERRACE VIEW AVE	1008357	258602	(40.876444151331, -73.9128285342
10463	2014-04-01	HEAT/HOT WATER	ENTIRE BUILDING	3810 BAILEY AVENUE	1012586	261482	(40.884336286878096, -73.8975242
10463	2014-04-01	HEAT/HOT WATER	ENTIRE BUILDING	3810 BAILEY AVENUE	1012586	261482	(40.884336286878096, -73.8975242
10463	2014-04-01	HEAT/HOT WATER	ENTIRE BUILDING	3810 BAILEY AVENUE	1012586	261482	(40.884336286878096, -73.8975242
10463	2014-04-01	DOOR/WINDOW	DOOR	110 TERRACE VIEW AVE	1008357	258602	(40.876444151331, -73.9128285342
10463	2014-04-01	Rodent	Condition Attracting Rodents	1934 NORTH RAILROAD	948358	144556	(40.563376021503466, -74.1291780
10463	2014-04-01	Rodent	Rat Sighting	534 WEST 175TH STREE	1002201	246971	(40.844535232977704, -73.9351199
10306	2014-04-01	HEAT/HOT WATER	ENTIRE BUILDING	3810 BAILEY AVENUE	1012586	261482	(40.884336286878096, -73.8975242
10033	2014-04-01	Rodent	Rat Sighting		954402	142870	(40.558770684605705, -74.1074178
10463	2014-04-01	Noise - Residential	Loud Music/Party	2341 RICHMOND ROAD	951527	149691	(40.577482843138505, -74.1177974
10306	2014-04-01	Noise - Residential	Banging/Pounding	2775 EAST 12 STREET	996114	151870	(40.583516220598898, -73.9572876
10306	2014-04-01	Noise	Noise: Construction Before/After Hours (NM1)		990265	220914	(40.773032188080042, -73.9782834
11235	2014-04-01	Noise	Noise: Alarms (NR3)		996858	163767	(40.616170043694414, -73.9545869
10023	2014-04-01	Noise	Noise: Construction Before/After Hours (NM1)		985895	205227	(40.729977235073379, -73.9940647
11230	2014-04-01						
10003	2014-04-01						

Data ETL – reduce

Period	Raw data's obs number	Obs # with missing ZIP	Obs # missing all location info	Service requests' modalities #	Unique ZIP
April 2014	81645	3206	2740	170	278
April 2015	101890	4231	3069	178	260

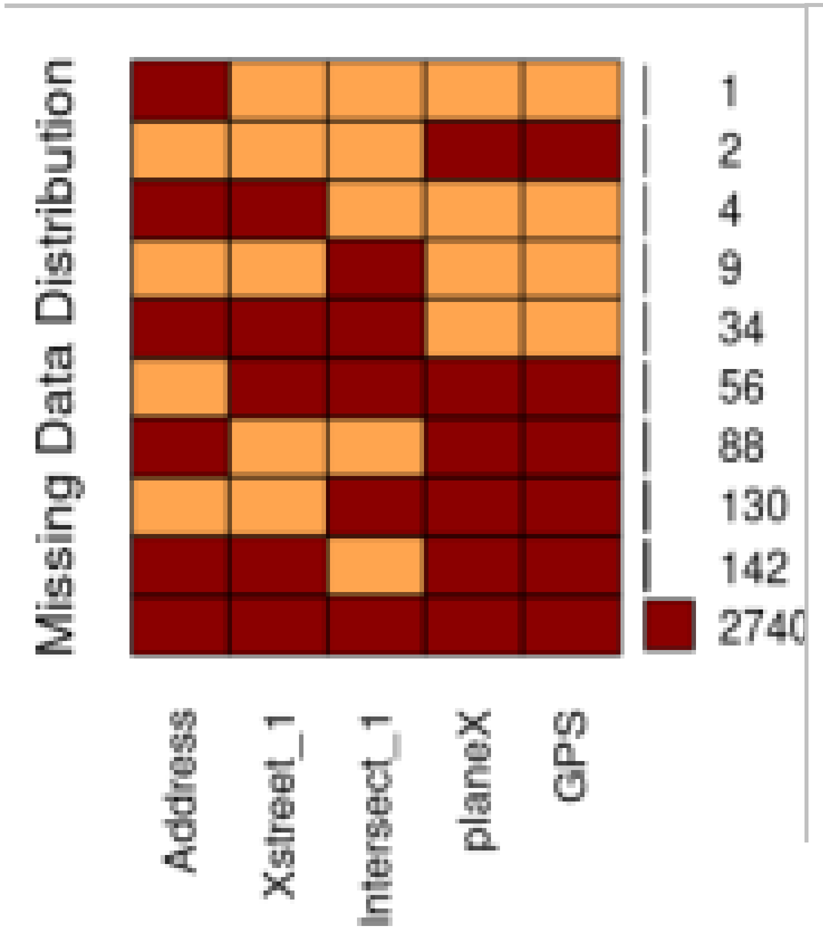
- Between 150 and 200 different SRCs' raw features
- Reduce to 13 features by consolidating calls' objects

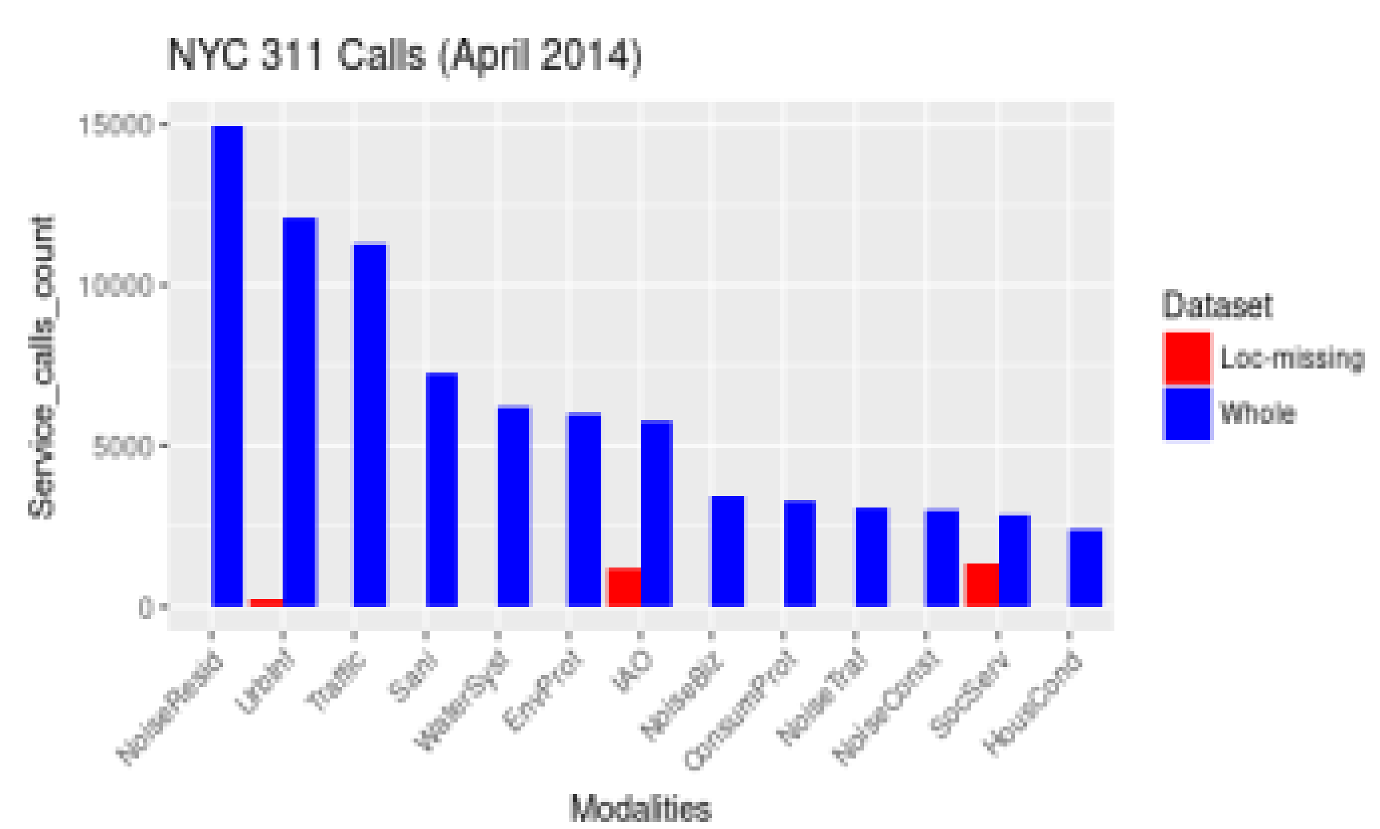
Data ETL – reduce

Service request calls' modalities	Modality description	Service request call frequencies		Change in rank from 2014 to 2015
		April 2014	April 2015	
NoiseResid	Residential Noise	19.00%	17.50%	–
UrbInf	Urban Infrastructure	15.00%	13.40%	↘
Traffic	Traffic related Issues	14.30%	17.20%	↗
Sani	Unsanitary Conditions	9.20%	10.50%	–
WaterSyst	Water Systems	7.80%	7.60%	–
EnvProt	Environmental Protection	7.60%	5.90%	–
IAO	Inspect, Audit, Order	5.80%	5.20%	↘
NoiseBiz	Commercial Noise	4.40%	4.90%	↘
ConsumProt	Comsumer Protection	4.20%	3.40%	↘
NoiseTraf	Traffic Noise	3.90%	5.40%	↗↗
NoiseConst	Construction Noise	3.80%	3.70%	↗
HousCond	Housing Conditions	3.10%	3.40%	–
SocServ	Social Services	1.90%	1.90%	–
	Total number of SRCs	78825	98649	↗↗

Data ETL – impute missings

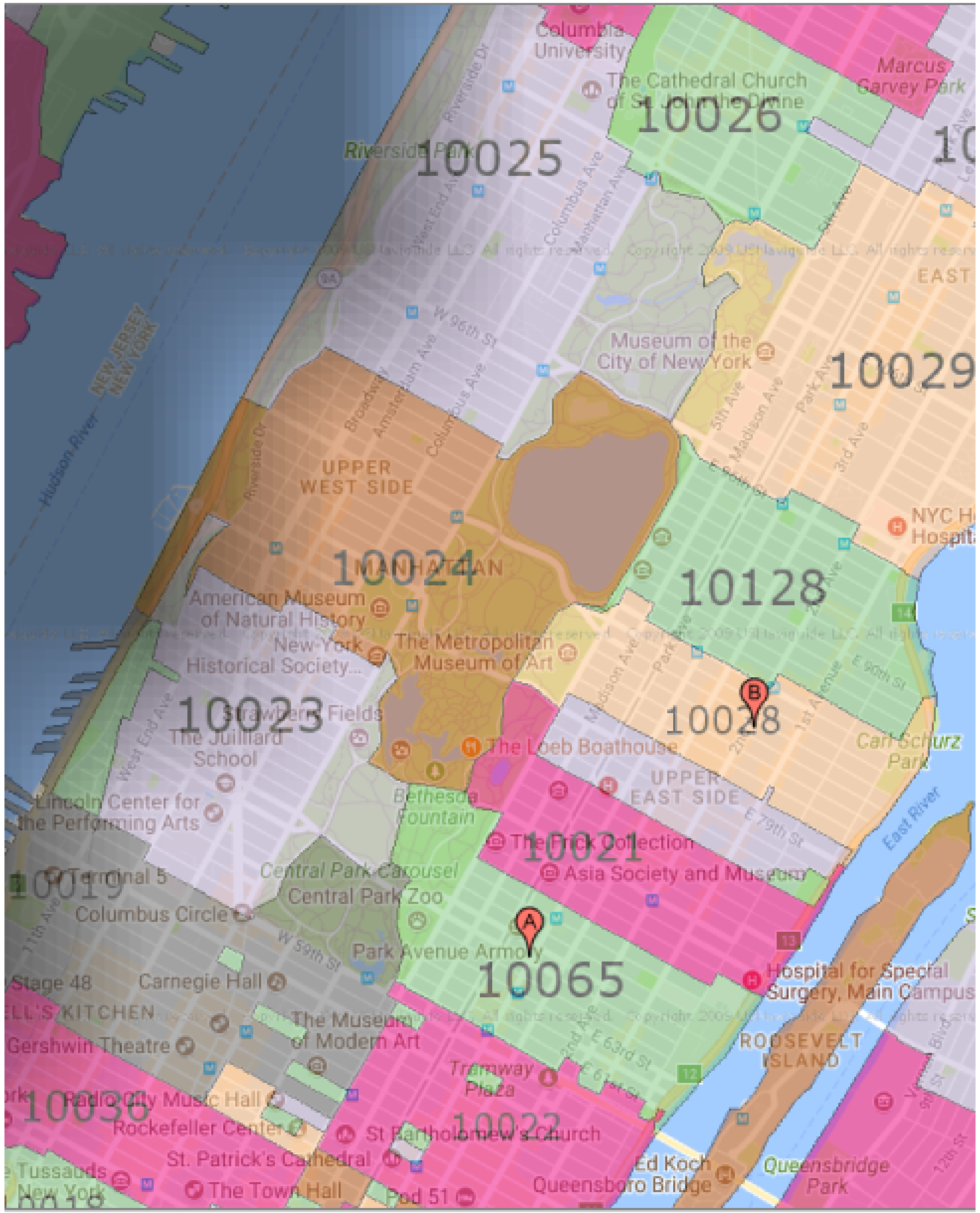
Period	Raw data's obs number	Obs # with missing ZIP	Obs # missing all location info	Service requests' modalities #	Unique ZIP
April 2014	81645	3206	2740	170	278





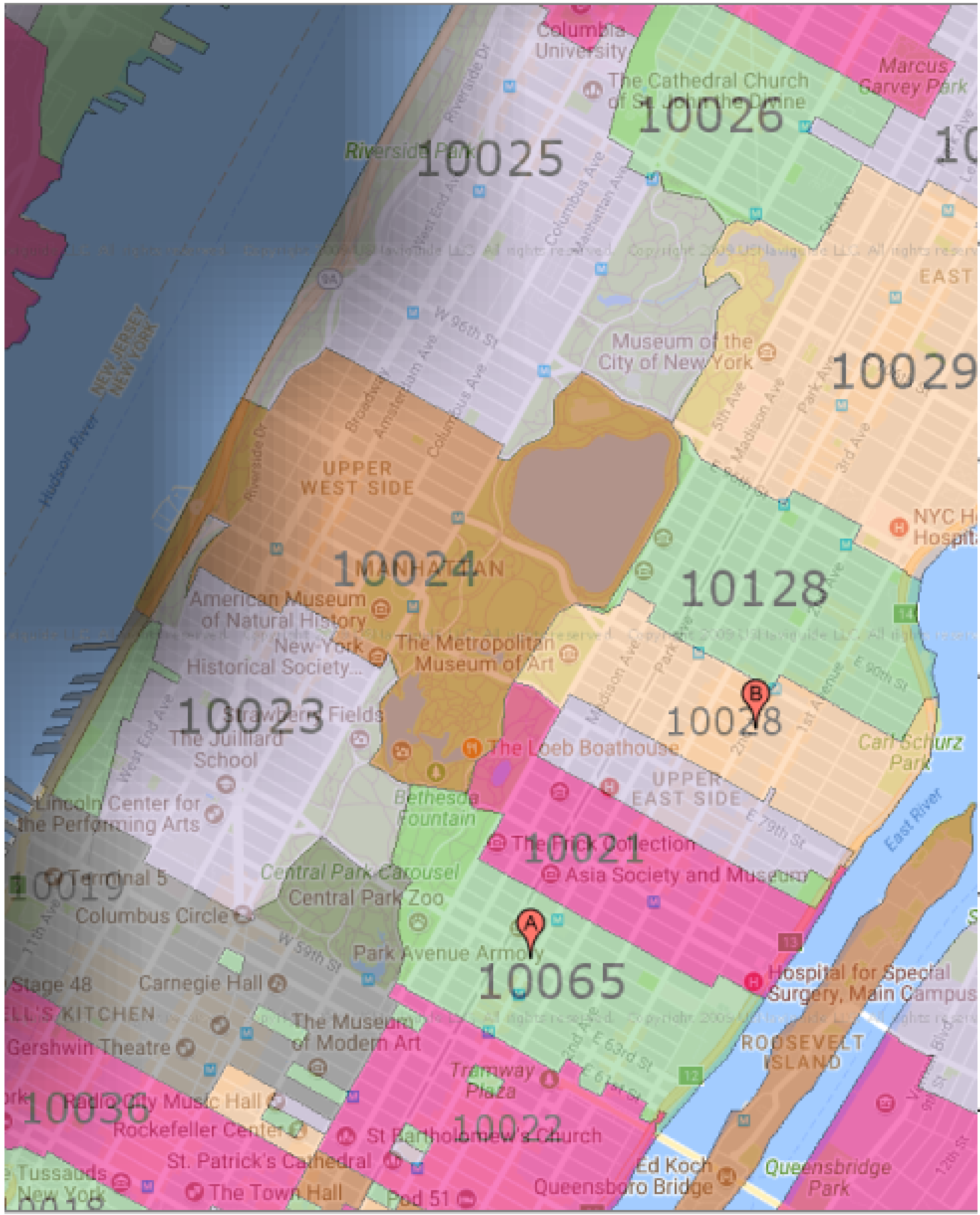
Data ETL – clean

ZIP code: 00083

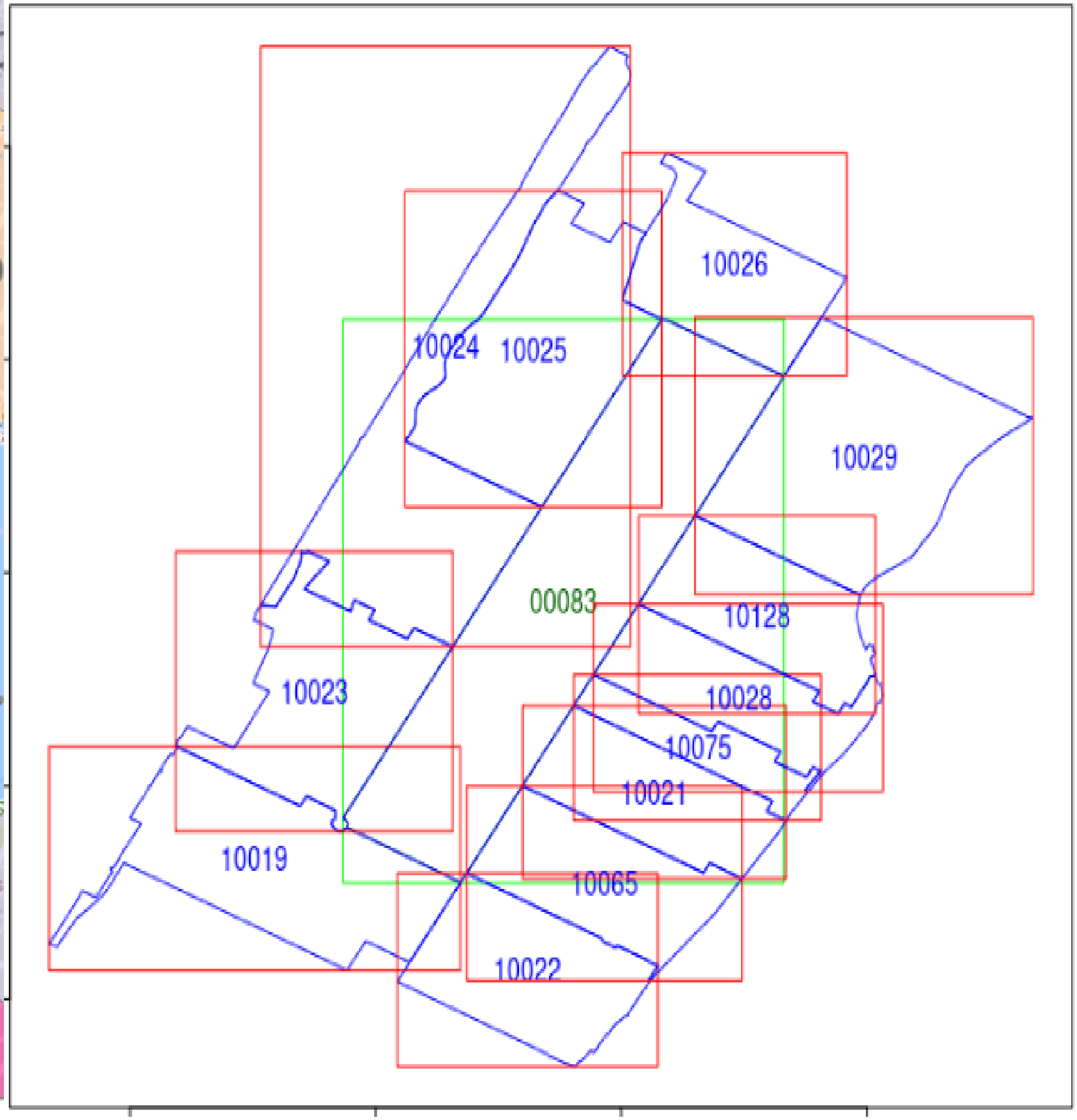


Data ETL – clean

ZIP code: 00083



NYC ZIP codes neighboring with "00083"



Data ETL – impute / clean

ZIP code: 10020

MVA – ZIP-SRC contingency table *(April 2014)*

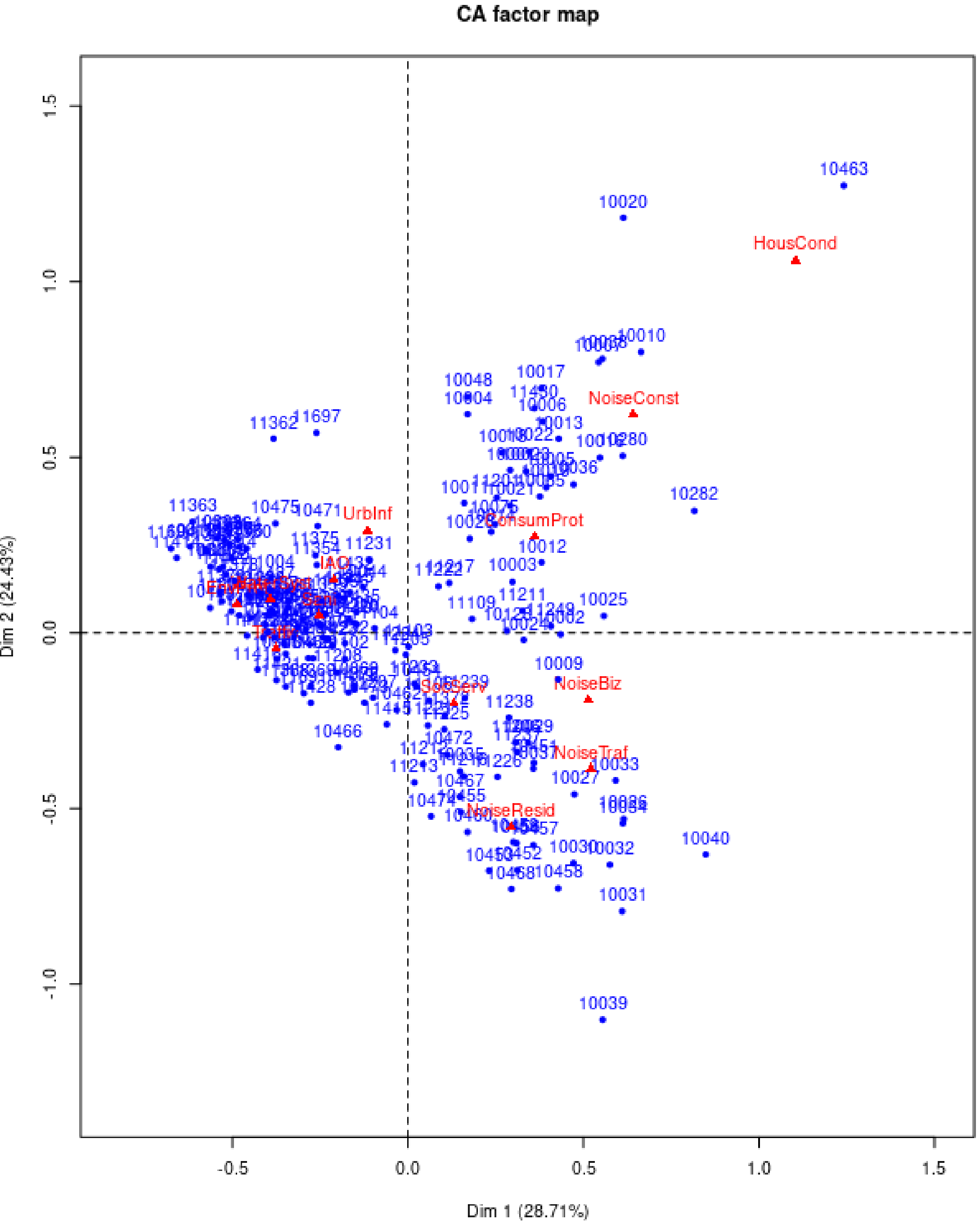
- Build frequency table
  (row profiles for ZIPs, column profiles for SRCs)

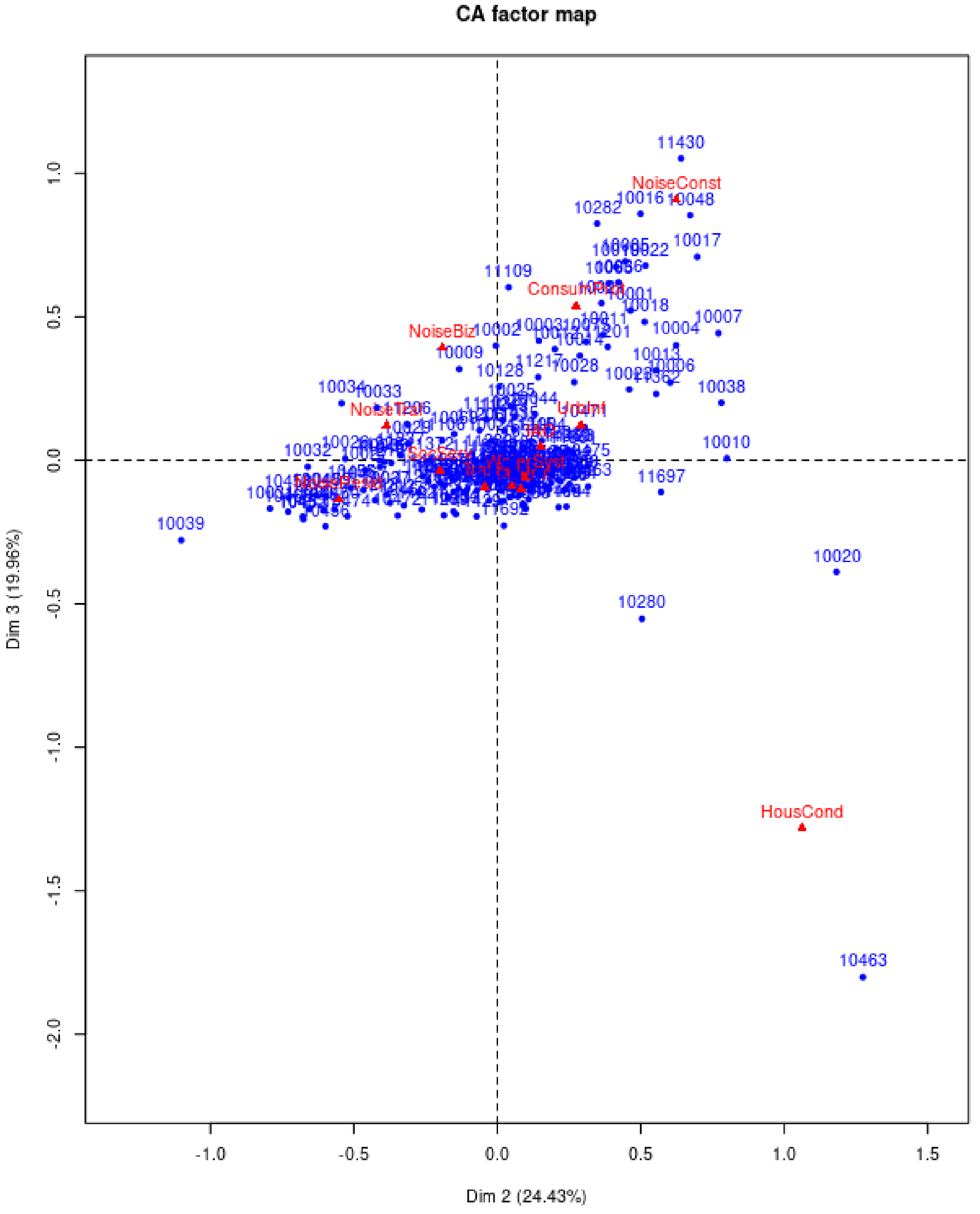
- Observe how 26 row marginals $<$ 5 / nbr_calls
  → Can we suppress them ? *(… χ^2-test of independence)*

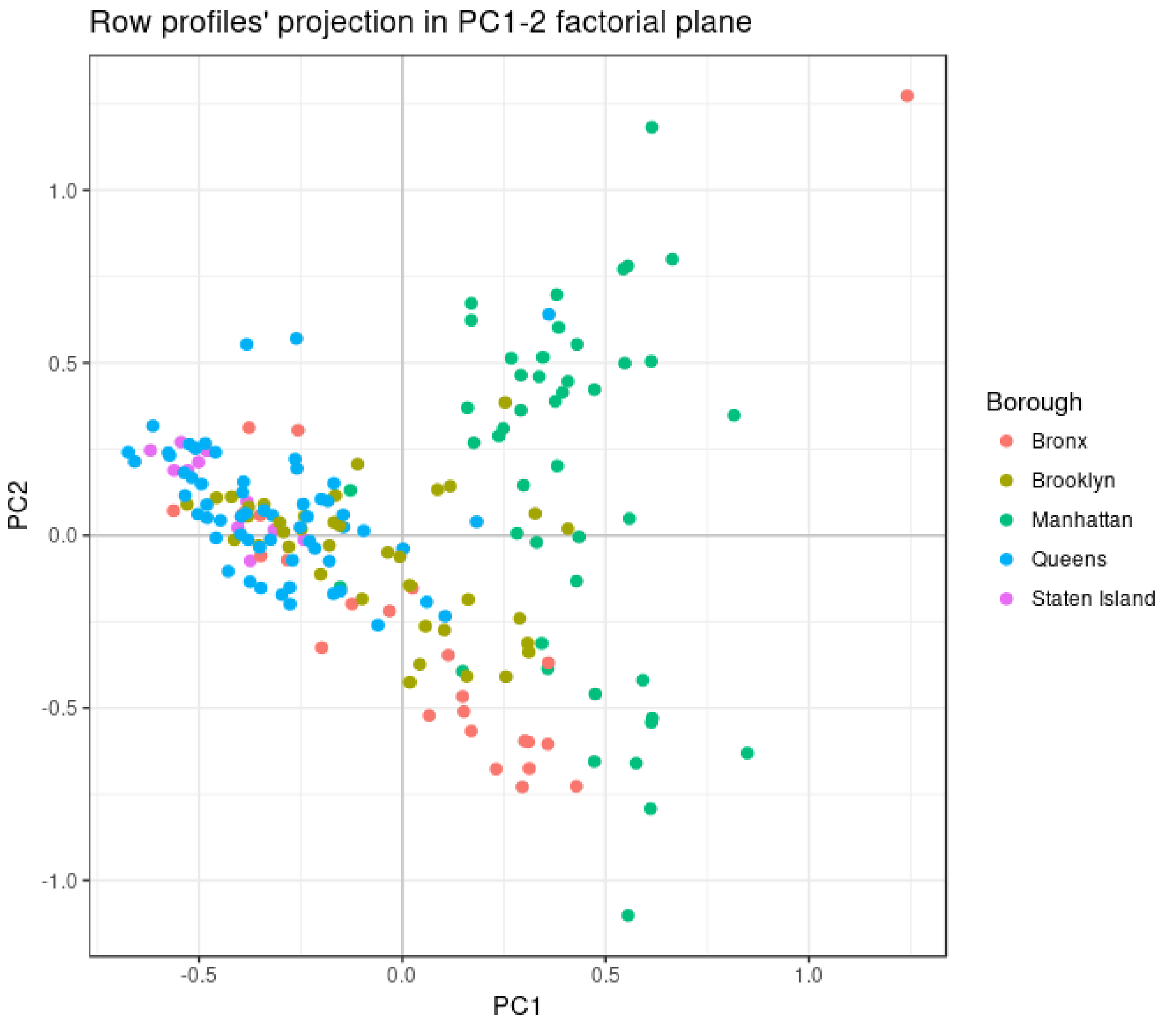
- Run CA with row marginals as row profile weights
  (*χ^2-metric)*

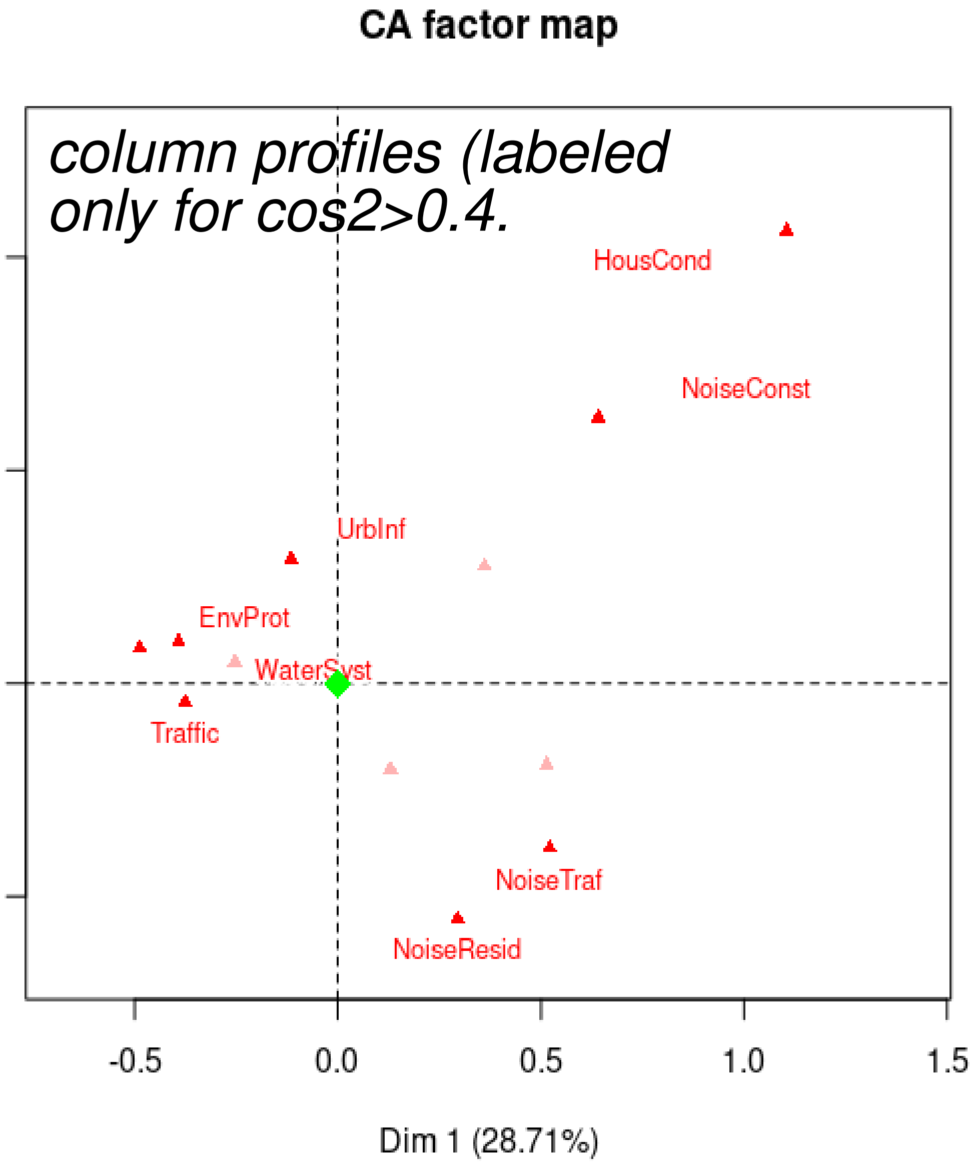
MVA – CA

(April 2014)



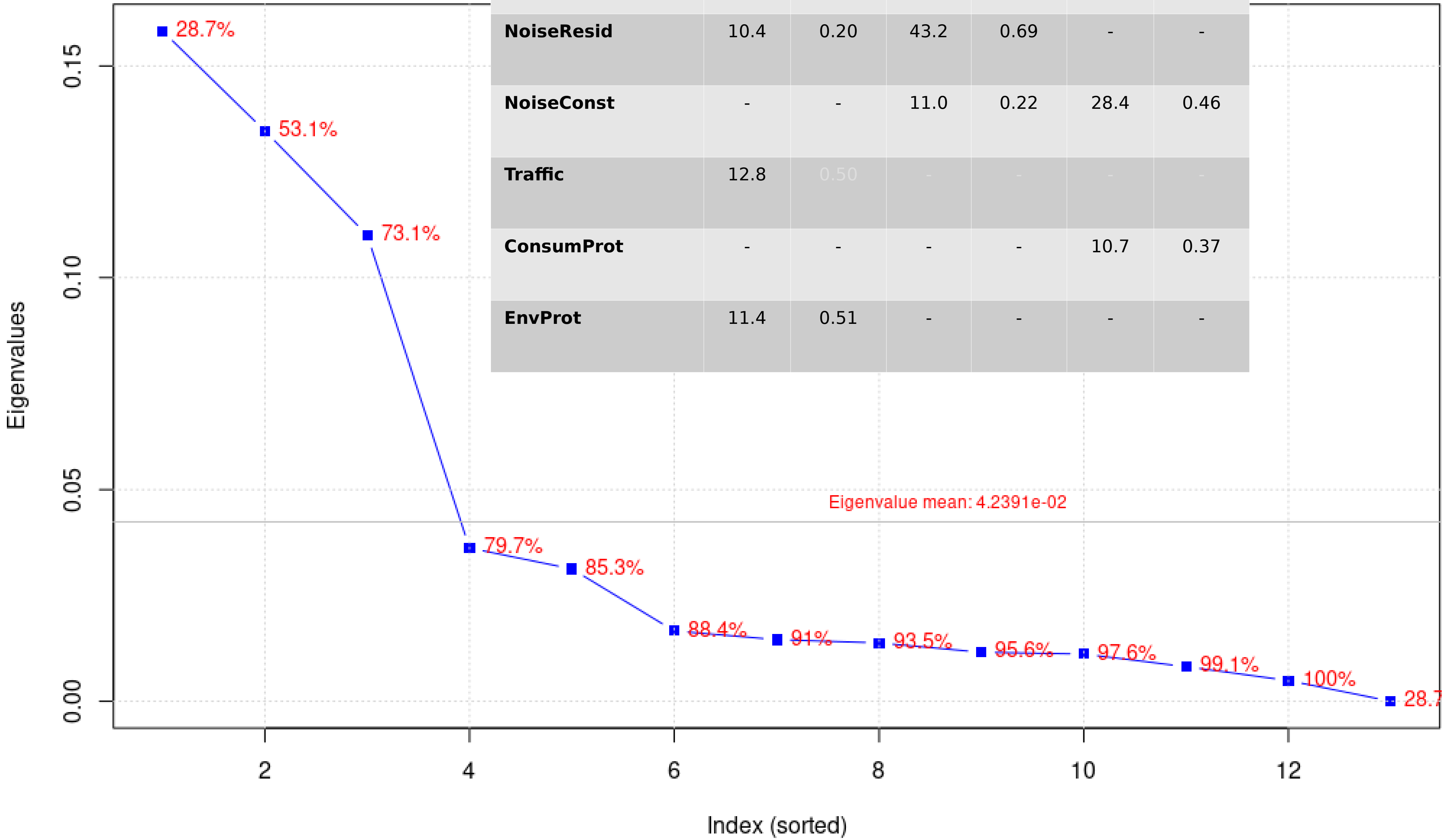



Row profiles' projection in PC1-2 factorial plane
PC2
PC1
Borough
Bronx
Brooklyn
Manhattan
Queens
Staten Island

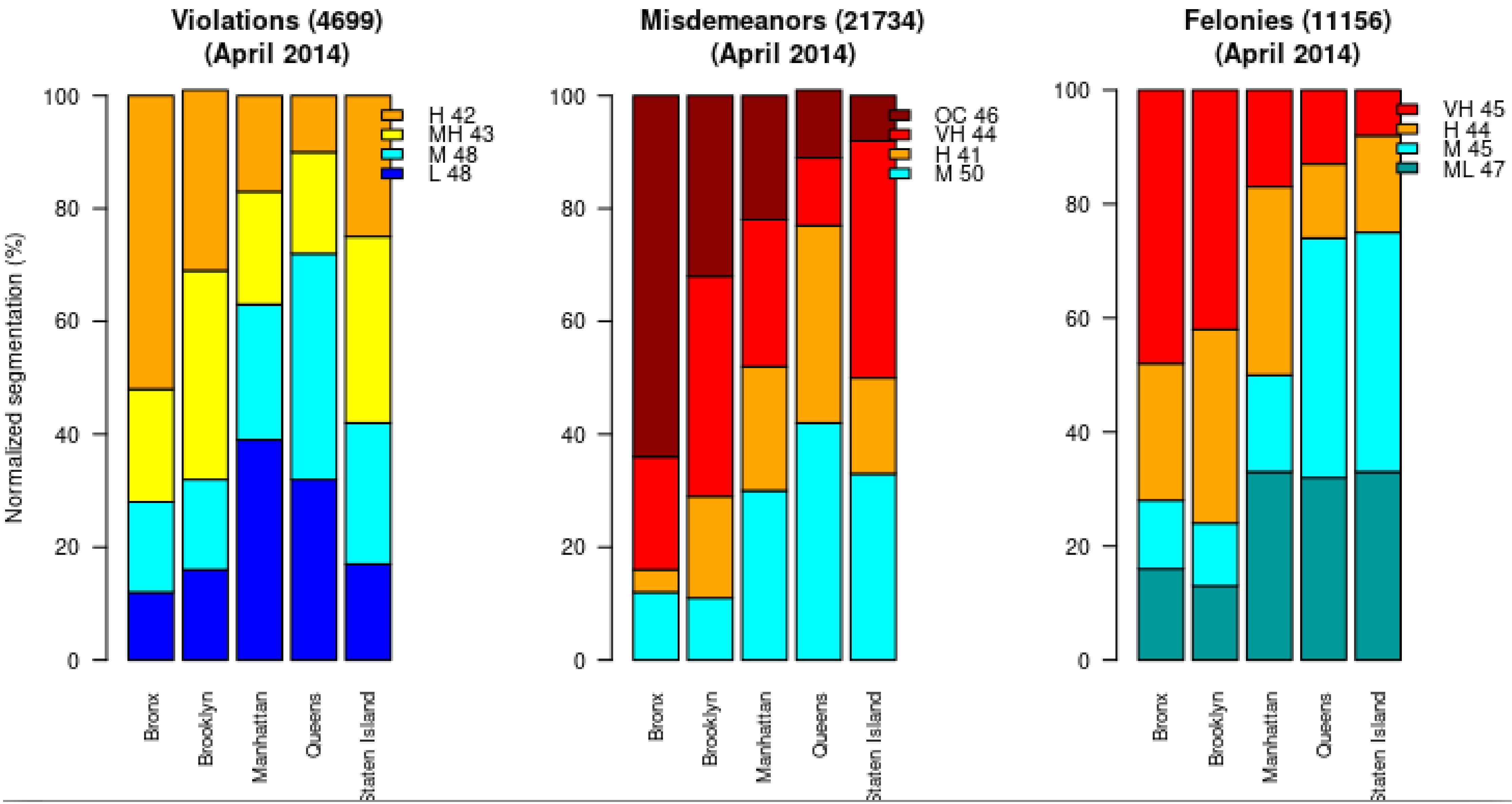

CA factor map
column profiles (labeled
only for cos2>0.4.
HousCond
NoiseConst
UrbInf
EnvProt
WaterSyst
Traffic
NoiseTraf
NoiseResid
Dim 1 (28.71%)

MVA – PCA

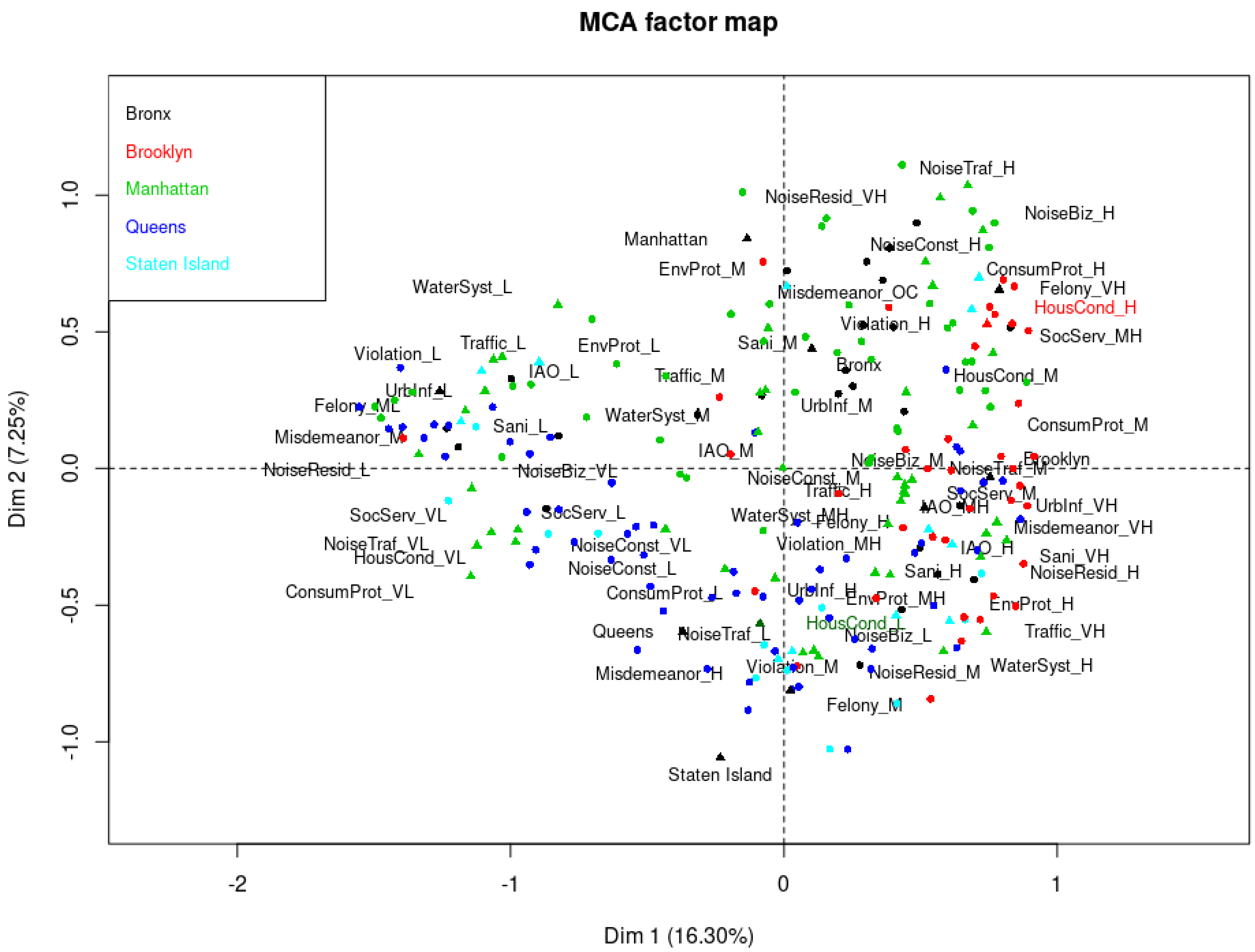
	Dim 1		Dim 2		Dim 3	
	ctr	*cos2*	*ctr*	*cos2*	*ctr*	*cos2*
HousCond	23.7	0.30	25.7	0.29	45.9	0.41
NoiseResid	10.4	0.20	43.2	0.69	-	-
NoiseConst	-	-	11.0	0.22	28.4	0.46
Traffic	12.8	0.50	-	-	-	-
ConsumProt	-	-	-	-	10.7	0.37
EnvProt	11.4	0.51	-	-	-	-



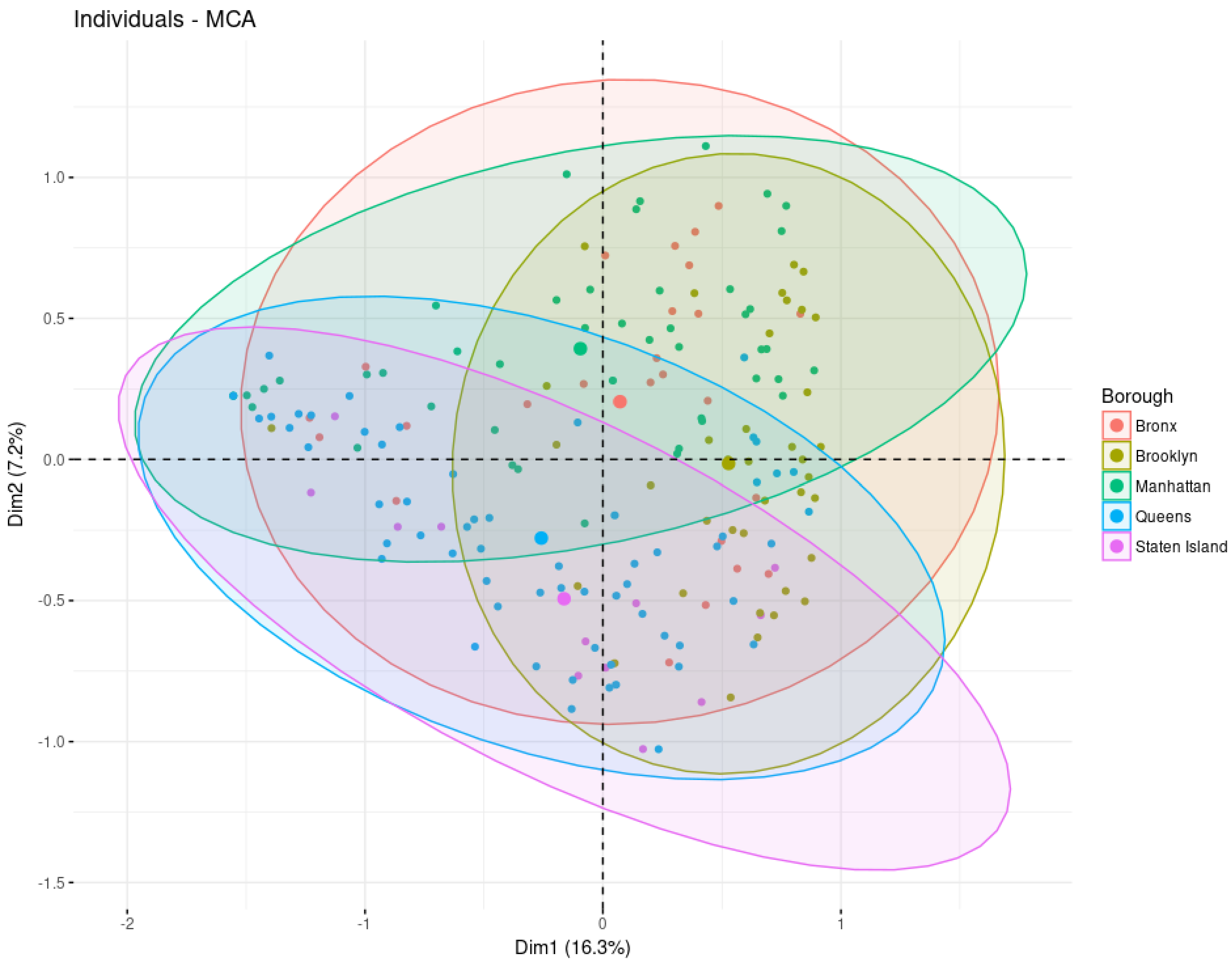
MVA – MCA



Violations (4699)
(April 2014)
Normalized segmentation (%)
100
80
60
40
20
0
H 42
MH 43
M 48
L 48
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Misdemeanors (21734)
(April 2014)
OC 46
VH 44
H 41
M 50
Felonies (11156)
(April 2014)
VH 45
H 44
M 45
ML 47



MCA factor map
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Dim 1 (16.30%)
Dim 2 (7.25%)
-2
-1
0
1
-1.0
-0.5
0.0
0.5
1.0
NoiseTraf_H
NoiseResid_VH
NoiseBiz_H
Manhattan
NoiseConst_H
EnvProt_M
ConsumProt_H
Felony_VH
HousCond_H
WaterSyst_L
Misdemeanor_OC
Violation_H
SocServ_MH
Violation_L
Traffic_L
EnvProt_L
Sani_M
Bronx
IAO_L
Traffic_M
HousCond_M
UrbInf_L
Felony_ML
UrbInf_M
WaterSyst_M
Sani_L
ConsumProt_M
Misdemeanor_M
IAO_M
NoiseBiz_M
Brooklyn
NoiseResid_L
NoiseBiz_VL
NoiseTraf_M
NoiseConst_M
Traffic_H
SocServ_M
UrbInf_VH
SocServ_VL
SocServ_L
WaterSyst_MH
Felony_H
IAO_VH
Misdemeanor_VH
NoiseTraf_VL
HousCond_VL
NoiseConst_VL
Violation_MH
IAO_H
Sani_VH
NoiseConst_L
Sani_H
NoiseResid_H
ConsumProt_VL
ConsumProt_L
UrbInf_H
EnvProt_MH
EnvProt_H
HousCond_L
Queens
NoiseTraf_L
NoiseBiz_L
Traffic_VH
Misdemeanor_H
Violation_M
NoiseResid_M
WaterSyst_H
Felony_M
Staten Island


Individuals - MCA
Dim2 (7.2%)
Dim1 (16.3%)
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2
-1
0
1
Borough
Bronx
Brooklyn
Manhattan
Queens
Staten Island

Hierarchical clustering

- In this section we present an attempt to clusterize our data set.
- This attempt is carried out by applying, in the following order:
  - Probabilistic clustering with k-means replications
  - Hierarchical clustering
  - Clustering consolidation using k-means.

Hierarchical clustering

- Selection of the optimal number of clusters:

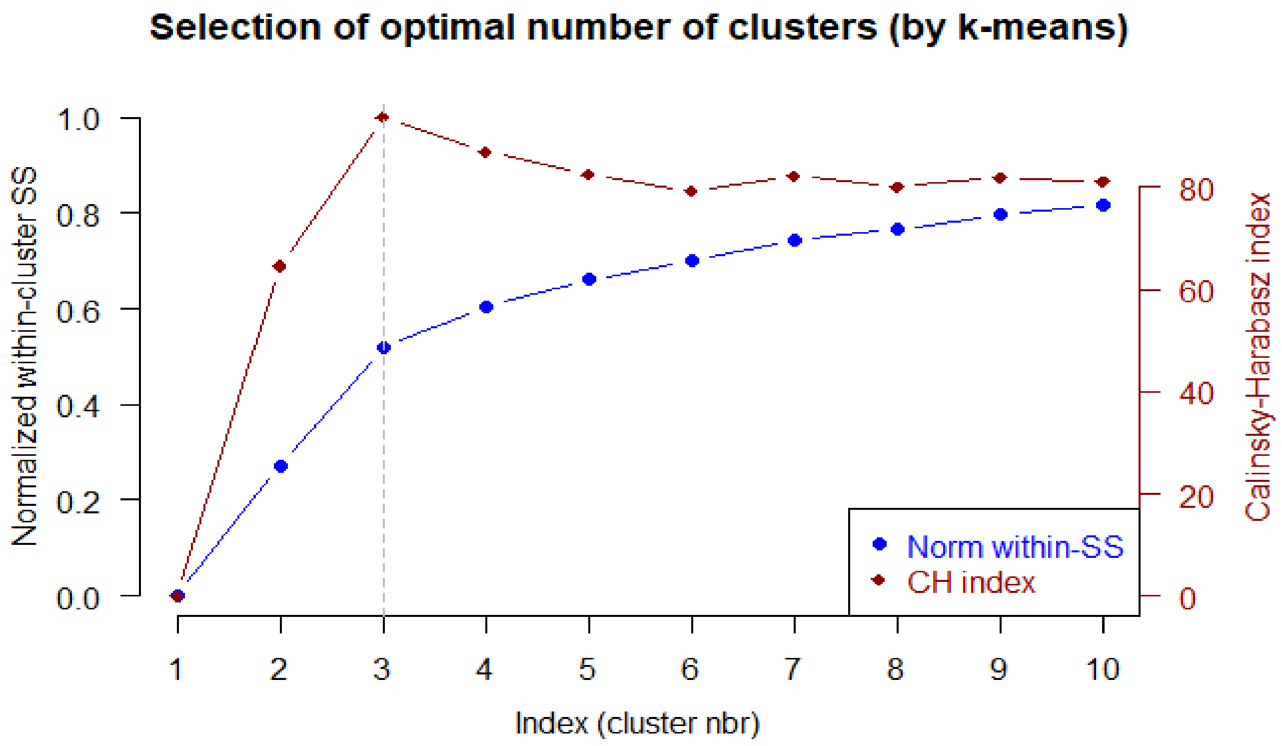
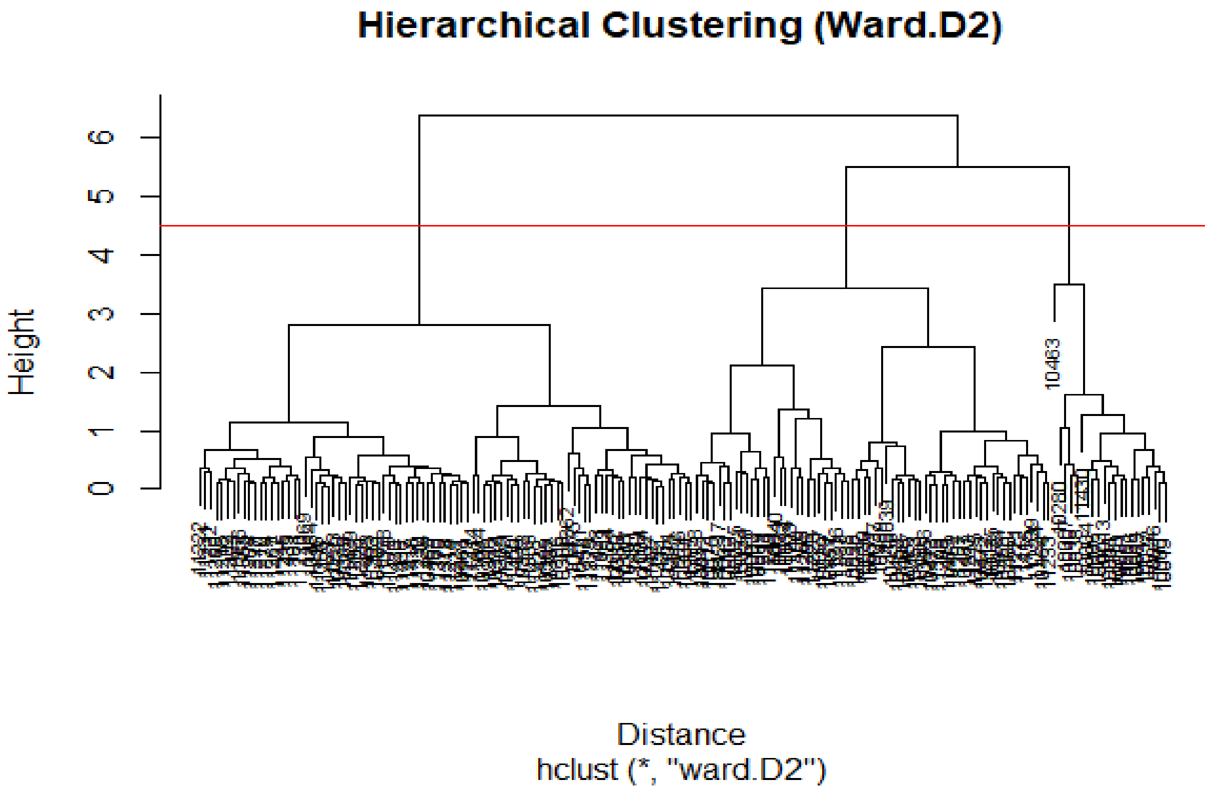


Figure 13. Selection of optimal number of clusters

Hierarchical clustering

- Selection of the optimal number of clusters:



Hierarchical clustering

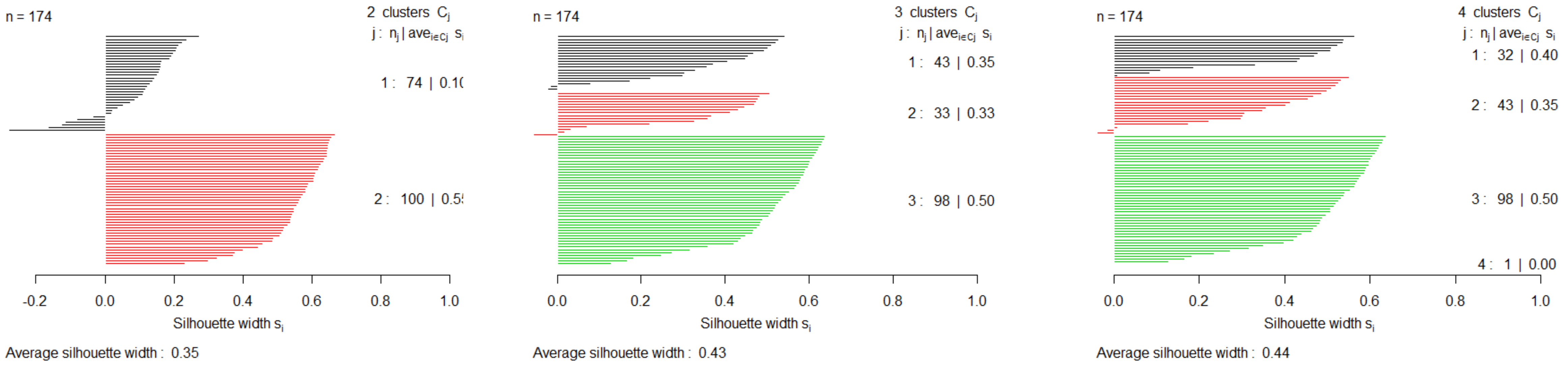
- Selection of the optimal number of clusters:

	G1	G2	G3
PC1	0,44	0,23	-0,36
PC2	0,58	-0,28	0,08
PC3	0,35	0,00	-0,06
PC4	-0,18	-0,02	0,01
PC5	-0,08	0,02	0,03

Table 7. Centroids of the clusters

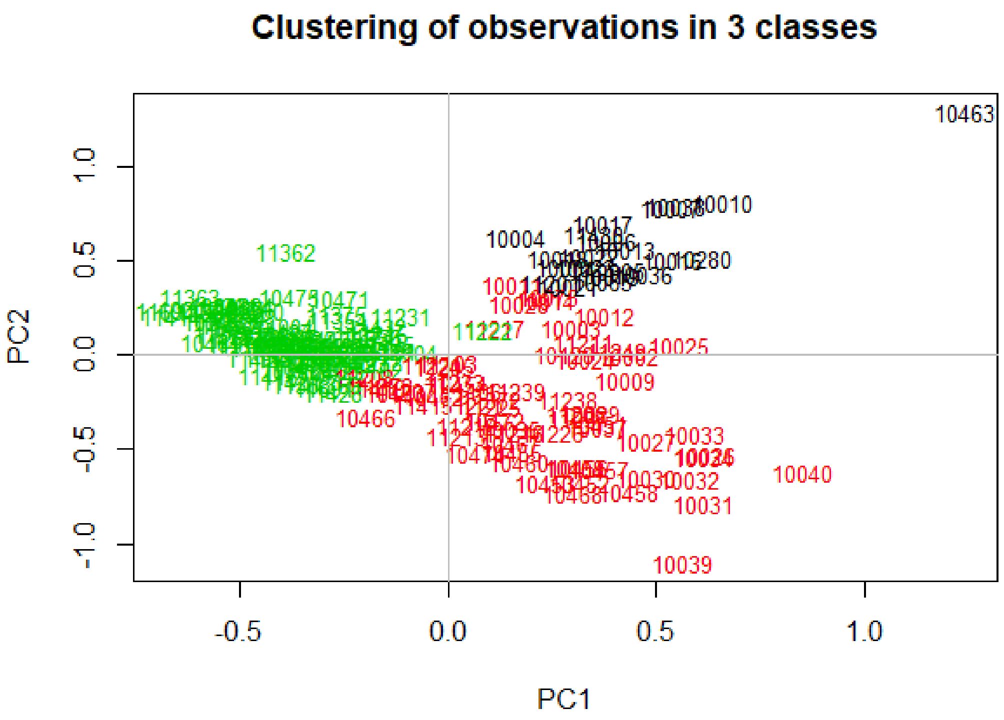
Hierarchical clustering

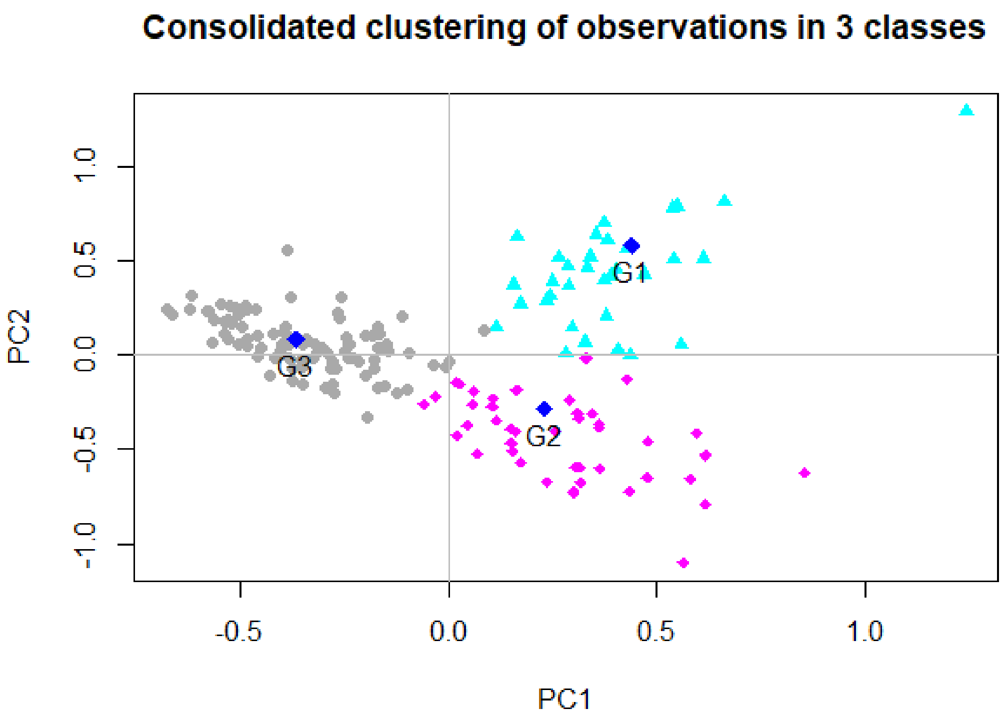
- Silhouette method:



Hierarchical clustering

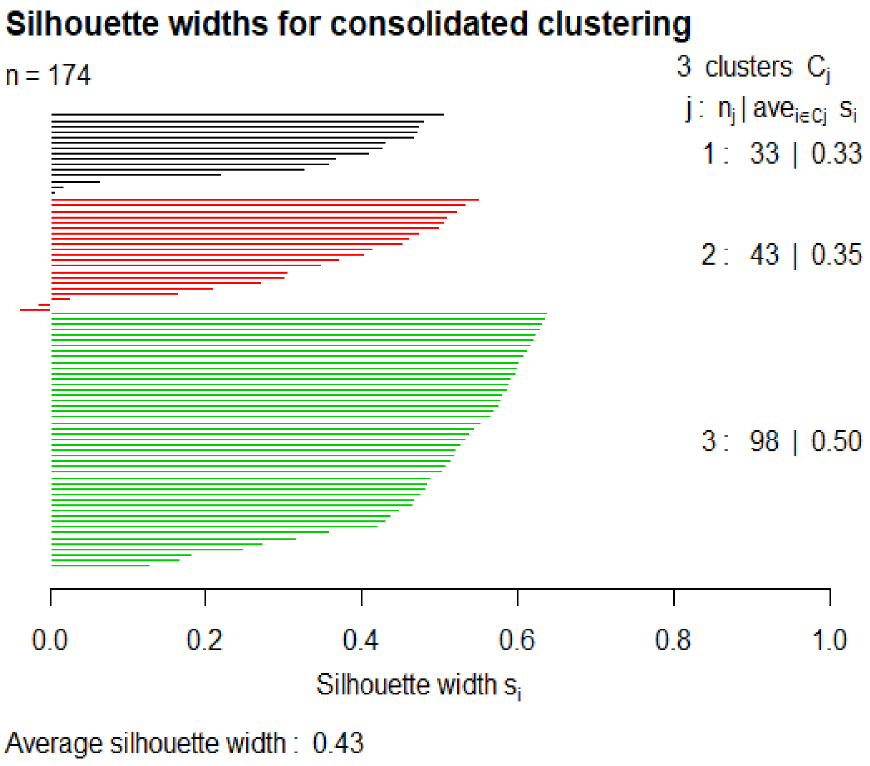
- Visualize partitions:





Hierarchical clustering

- Silhouette method after consolidation:



Hierarchical clustering

- Categorical description to interpret the clusters:
  - We reject the null hypothesis at the risk 0.05 of being wrong when the p-values <0.05.
  - Variables for which we reject H0 -> meaningful categorization:

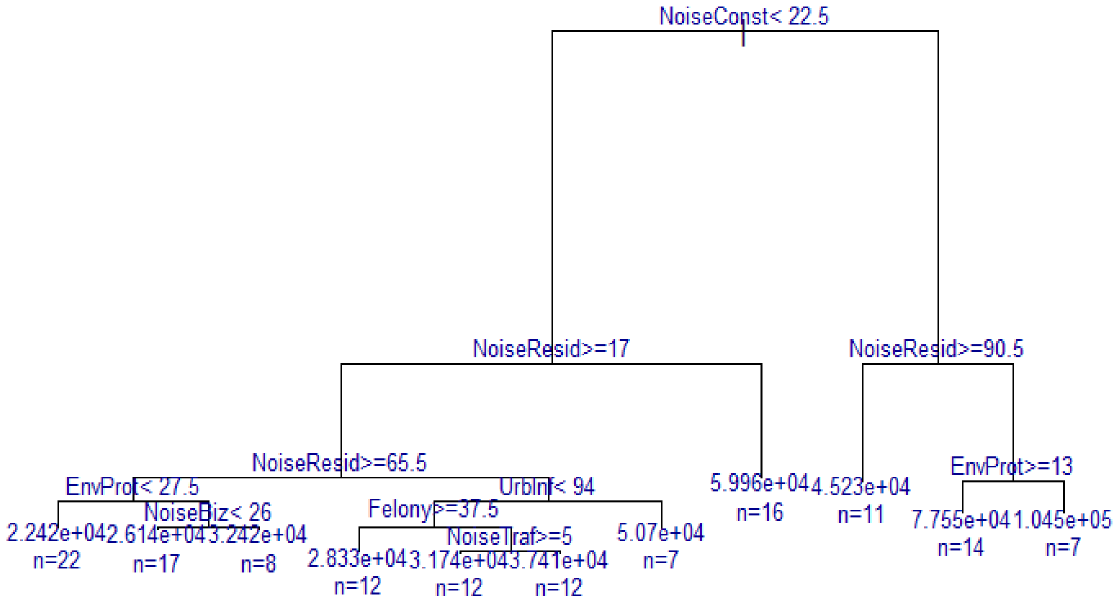
Cat1	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
NoiseConst	9,145613	59,242424	17,132184	44,958587	29,298499	5,93E-20
ConsumProt	8,694289	43,303030	18,385057	25,590582	18,236820	3,49E-18
UrbInf	4,484378	98,575758	67,804598	55,909575	43,662851	7,31268E-06
NoiseBiz	3,328495	33,787879	19,885057	29,805183	26,578197	0,000873166
HousCond	3,277523	39,090909	13,856322	108,106391	48,991520	0,001047222
NoiseTraf	2,004039	23,393939	17,465517	17,769465	18,823636	0,045065931
Sani	-1,960935	32,939394	41,557471	25,178080	27,965180	0,049886588
NoiseResid	-2,335256	58,606061	85,896552	48,646492	74,361347	0,019530032
WaterSyst	-2,425274	26,696970	35,454023	16,983923	22,975641	0,015296833
EnvProt	-3,539835	18,424242	34,379310	13,407439	28,680460	0,000400377
Traffic	-4,138211	33,757576	64,637931	21,908944	47,483221	3,50025E-05

Decision trees

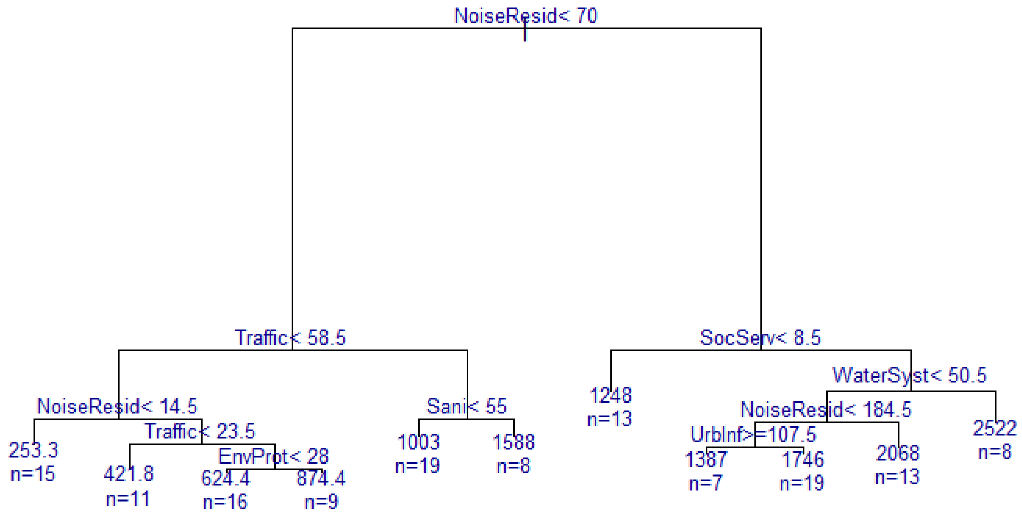
- In first place, we build the 2 possible decision trees:
  - One related to each of the 2 decision variables that we have (“medianInc” and “jlBenef”).
- Before building the trees we split the dataset in trainng (80% of individuals) and test (20% of individuals).

Decision trees

Fully grown decision tree for training data-set and "medianInc" as decision variable.

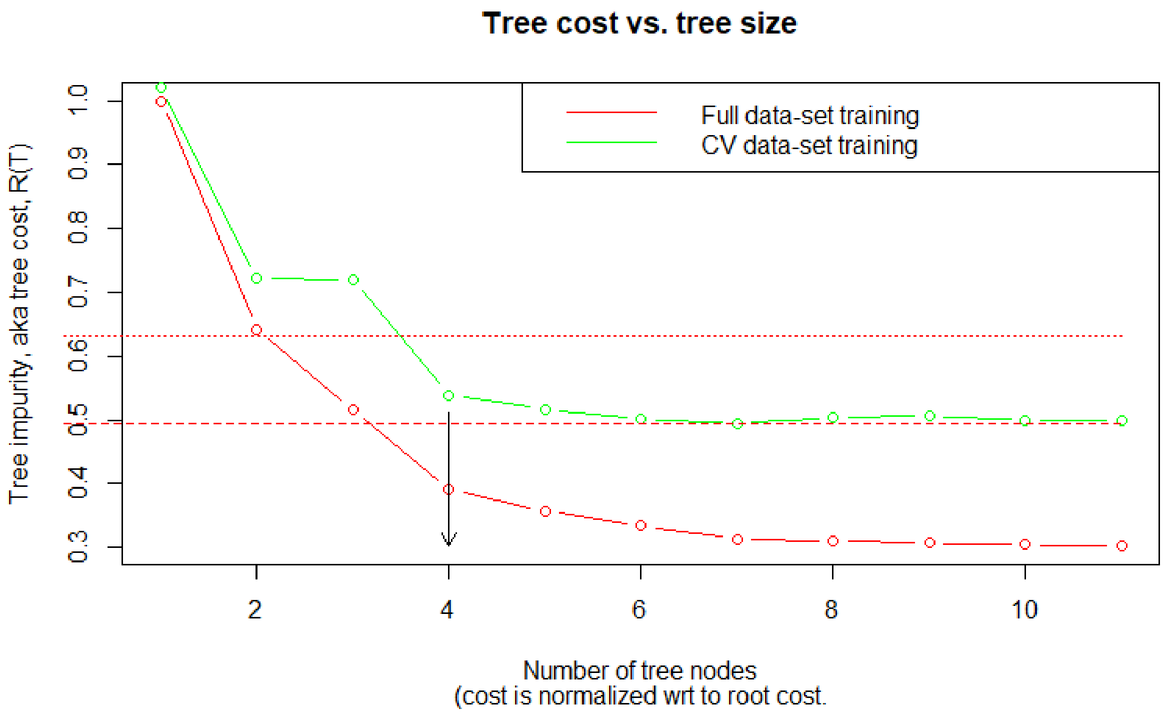


Fully grown decision tree for training data-set and "jlBenef" as decision variable.



Decision trees

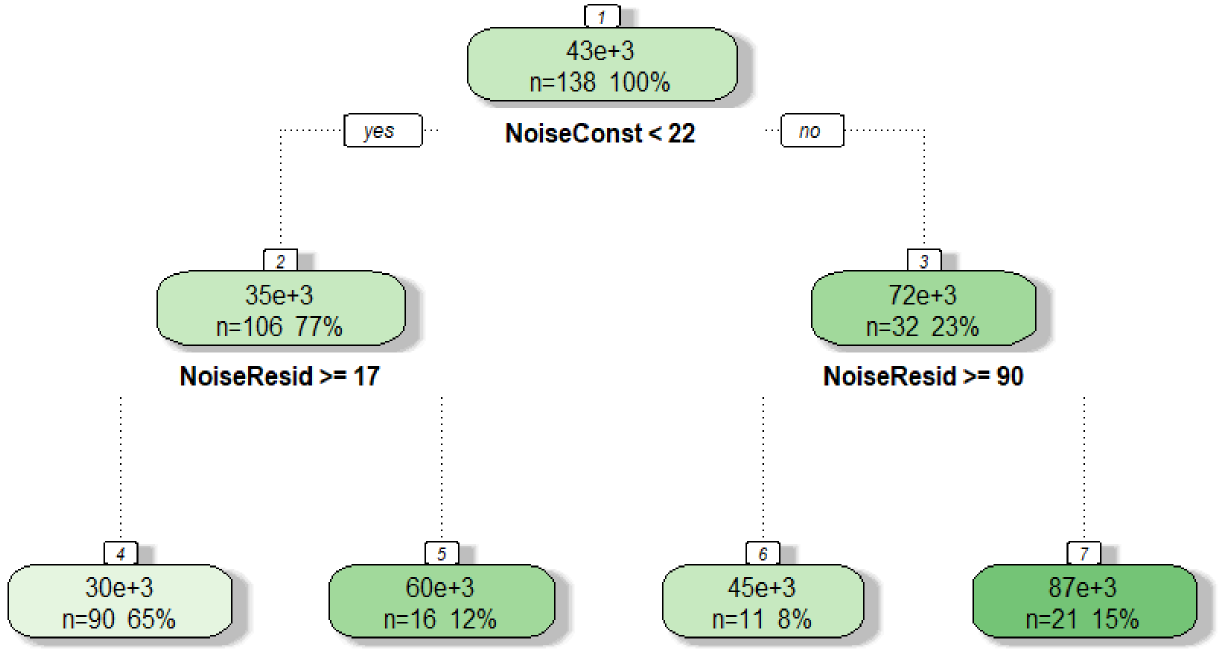
- CV normalized error mean and the whole data set based training error as a function of tree size:



- Red horizontal dashed line (below) -> minimum tree impurity (MTI) level
- Red dotted line (above) -> MTI + 1.
- Black arrow -> optimum number of nodes for post-pruning.

Decision trees

- Post-pruning:
  - Optimum complexity parameter -> αopt = 0.03489.
  - Now, we are able to post-prune the decision tree by using -> αopt.



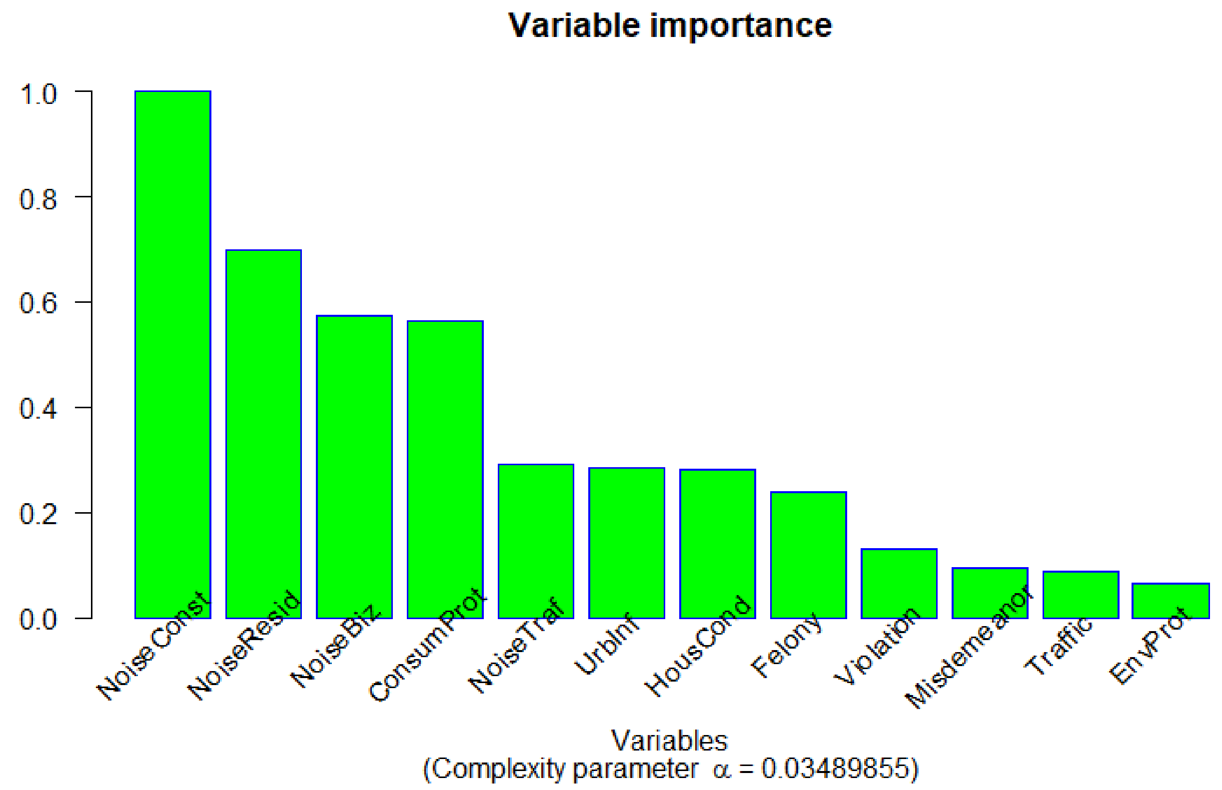
Decision trees

- Post-pruning:
  - Split rules:

```
Rule number: 4 [medianInc=30240.9222222222 cover=90 (65%)]

  NoiseConst< 22.5
  NoiseResid>=17

 Rule number: 7 [medianInc=86539.2380952381 cover=21 (15%)]
  NoiseConst>=22.5
  NoiseResid< 90.5

 Rule number: 5 [medianInc=59963.4375 cover=16 (12%)]
  NoiseConst< 22.5
  NoiseResid< 17

 Rule number: 6 [medianInc=45225.8181818182 cover=11 (8%)]
  NoiseConst>=22.5
  NoiseResid>=90.5
```

Decision trees

- Variable Importance for the optimally pruned decision tree :



Decision trees

- Predictions:

- Slice of the original results-table which contains the predictions for each value of "medianInc" in the test data set.

	4 predicted classes			
"medianInc" for Test-set	30240,9222	45225,8182	59963,4375	86539,2381
17992	1	0	0	0
18164	1	0	0	0
26143	1	0	0	0
26170	1	0	0	0
27102	1	0	0	0
27144	1	0	0	0
27203	1	0	0	0
27303	1	0	0	0
27331	1	0	0	0
27374	1	0	0	0
27898	1	0	0	0
90981	1	0	0	0
92955	0	0	0	1
93056	0	0	1	0
95992	0	0	0	1
97669	0	0	0	1
98024	0	0	0	1
110248	0	0	0	1
128571	0	0	1	0
185593	0	0	0	1
^^^^^^^^^^^	^^^^^^^^^^^	^^^^^^^^^^^	^^^^^^^^^^^	^^^^^^^^^^^
Total Freq	**97**	**10**	**15**	**18**

Questions?