# Multivariate statistical and semantic analysis of the 5-borough NYC urban environment

## EXECUTIVE SUMMARY

Author: Cédric Bhihe, PhD
<sedric.bhihe@upc.edu>

Principal Investigators: Prof. Jorge García-Vidal, PhD
<jorge@ac.upc.edu>
Prof. José Mª Barceló-Ordinas, PhD
<joseb@ac.upc.edu>

Department of Computer Architecture
Polytechnical University of Catalonia
Barcelona, Spain

### Motivation for this work: the issue

Large urban areas are complex environments, characterized by a large, heterogeneous set of co-varying variables, whose modalities coalesce in time and space as singular events. When put together, those events constitutes the tokens of *urban semantics*, a new idiom we propose to explore and analyze. The challenge consists in building a set of semantic analysis tools based on available static and dynamic data, whose results can be visualized in a user-friendly way. Ultimately at stake is for city officials and businesses alike to better grasp urban trends, their cause and their evolution with a degree of predictability. Their common objective is to make better decisions, in a timely way, to build better strategies and enact optimal policy on which to base their allocation of resources. In doing so they ultimately seek to maximize their return, be it for the benefit of the greater public or for that of private investors.

How are a city's parks and streets tended, how well is its garbage collected and its curbs swept, how is municipal regulation enforced and to what extent are those regulations heeded, where are shops and businesses, churches, kindergarten or even newspaper-stands and traffic lights established, or newly located as time passes, where does noise pollution arise and where do traffic jams occur within the city ? Every corresponding event is captured in constantly updated data-sets commonly found in many cities such as New York, Barcelona, Rome or Paris. The corresponding streams of data constitute the interwoven semantic threads of the urban fabric. Combined with temporal and topographical information about points of interest, as well as income, population and other statistics, we understand intuitively that urban semantics made of those innumerable tokens, do not only shape the perception of the city by its inhabitants and visitors alike, but also correlate with how crime is distributed in space and over time, where businesses establish themselves and property value increases or plummets. At the same time the state of a city reflects on how it is spoken of in social or more conventional media, in small or in significant ways, on matters of criminality, politics, private initiatives, social movements... All this feeds into how the urban environment is dealt with, with micro-decision on an individual basis, policy or strategy on a larger scale by public or private decision-makers.

### Multivariate analysis of urban semantics: the solution

A key issue, as often perceived by those poised to benefit from this work, is how to best interpret urban semantics, where every data point constitutes a token, in order to make allocations decisions. However today's paradox is that the out-pour of available smart-city data is often too much, too heterogeneous or too complex. Its sheer volume, velocity and variability all too often prevent available data from be usefully tapped and visualized both by the administrations of the very cities at the origin of the data and by organizations with a vested interest in tapping the same... As a result urban semantics remain an idiom difficult to understand. To overcome that predicament, and after extracting available urban data, we propose to automate data transformation to the extent possible, before conducting both linear and non-linear statistical analyses of its semantics. For this we may first use conventional methods based on the linear analysis of hidden data structures from principal component analysis, multiple correspondence analysis to clustering. Token proximity, represented both in time and space, will be the basis for spatial as well as associative inferences. Last we propose to visualize results in ways that may make untapped Big Data readily accessible to decision makers and conducive to the further elaboration of what-if scenarios.

The larger goal of this work is to design and implement a unified set of tool based on multivariate statistical learning methods and machine learning (together denoted *ML* hereafter). For the data scientist, this translates in at least three challenges:

- select and gather the data (Extract-Transform-Load (ETL) stage),
- analyze the data (ML) drawing on conventional and more modern techniques,
- visualize and present results in support of decision making processes.

### Stake-holders, current applicability and scope

The number of municipalities across the world, likely interested in better allocating their resources, is understandably large. The continued influx of people in cities make predictive management a sensitive must-have once reliable and user-friendly tools become available. The growing size of modern conurbations has direct consequences in terms of increasing complexity, when attempting to better understand or manage them. That makes the ability to manipulate big data and urban semantics in a customized way attractive both to business people seeking to maximize their ROI and to city officials seeking to maximize the well-being of city dwellers.

### Future applicability and pertinence

Cities becoming bigger and attracting more people year after year constitutes a trend, well consolidated over the past 100 years. The need for optimal resource allocations and complex commercial decision making should continue to assert itself, reinforced by increased environmental stress on dense cities due to global warming. It may be counterbalanced, at least in part, by the future possibility of large swaths of urban populations leaving their urban environments. This long term scenario (30 years in the future at least) finds its roots in both the global warming phenomenon and in soaring living costs in large cities. It does not constitute a proximate threat to this project, and should therefore not detract from its purported timeliness and usefulness.

### Preliminary results

Preliminary results for the New York City five-borough area are based on conventional multivariate statistical learning methods. Those methods were applied to a limited but rich set of data obtained through the NYC OpenData initiative, for the periods April 2010, April 2014 and April 2018. We focus more particularly on April 2014 but results are similar in nature for the two other periods 4 years before and after.

This is a summary of the complete report and annex to the complete report, simultaneously made available to the reader. We do not dwell on the ETL aspects of this research thrust as we have already and in most part overcome that challenge. Readers willing to invest time and effort in a more thorough approach of ETL, will find Section 2 of the complete report enlightening. It describes outlier detection, how to avoid bias in selecting data, data cleaning and consolidation.

The starting point for our analyses was raw data sets for:

- Service Request Calls (SRCs) to 311 placed by urbanites in the 5 borough new York city area
- Crimes reported to NYPD by 911 callers
- IRS returns with gross adjusted income and unemployment benefits reported on IRS returns.

SRCs consisted in 80,000 to 200,000 calls per month, while NYPD-gathered data oscillated between 36,000 and 160,000 calls a month, each of those 311 or 911 call characterized by more than 30 structured attributes as to the nature of the call, reported event location, etc. IRS data was fully included in our ETL stage but we choose not to report results in relation to it at this point. SRCs and crime reports represent the perception of urban issues by NYC inhabitants. They are event reports made spontaneously by individuals. As such they may be unwittingly replicated by different callers. In this sense the basis for this work is data which sheds light on the public perception or urban issues rather than on factual urban pathologies.

After data cleaning and preliminary feature extraction to reduce dimensionality of call modalities our cleaned data sets were frequency tables consisting of:

- between 170 and 220 ZIP codes, denoted individuals, observations or row profiles,
- 13 modalities of the categorical variable SRC, referred to as column-profiles,
- 3 modalities of the categorical variable “NYPD reported crime” (also column-profiles)

Modalities are summarized below. SRCs’ 13 modalities in particular, result from the selection and extraction of main issues out of 170 to 200 distinct calls’ objects. This process of selection and conversion to 13 resulting modalities is largely automated and permit to apply a common filter to all time periods. However it still requires human intervention to handle fringe cases where the object of 311 calls, based on its raw data description requires disambiguation.

| SRCs' modalities | Description |
|---|---|
| **HousCond** | Housing conditions: may include any aspect of public housing (heat, water, painting, blinds, odors, etc.) |
| **Sani** | Sanitary conditions: pests and rodents, graffiti, garbage not collected, public space odors, etc. |
| **NoiseResid** | Residential noise |
| **NoiseConst** | Construction noise |
| **NoiseBiz** | Commercial or business related noise |
| **UrbInf** | Urban infrastructure: street pot holes, dangling or missing street signs, traffic lights and public furniture |
| **Traffic** | Traffic related nuisance: illegal parking, blocked driveways, derelict vehicles, accidents, etc. |
| **NoiseTraf** | Traffic noise: may include boat, car, planes, motorbikes related noise. |
| **WaterSyst** | Water systems: rain water, water ducts issues, etc… |
| **ConsumProt** | Consumer protection: informing on bad business practices, cons by businesses, illegal business or commercial activity, false advertising, restaurant hygiene, etc. |
| **SocServ** | Social services: any claim, request or complaint related to social services to beneficiaries, etc. |
| **IAO** | Inspect-Audit-Order: includes all calls to request inpections, audits and the intervention of the authorities (not police) to re-establish due process and order, etc. |
| **EnvProt** | Environental protection: wild animals sighting, commerce of endangered species, improper disposal of controled chemicals, release of fumes and gas, trees and plants in need of care, etc. |

In addition the three crime modalities were: **Violation**, **Misdemeanor** and **Felony**.

In a first approach, Principal Component Analysis (PCA) and Correspondence Analysis (CA) on SRCs permitted extracting salient correlations between SRCs' modalities. We computed Inertia Explanatory Power (IEP) for all system's dimensions and for the system's significant dimensions (viz. Table 6 in the complete Report) , before applying the Varimax method for latent feature extraction (Figs. 1 & 2 below).

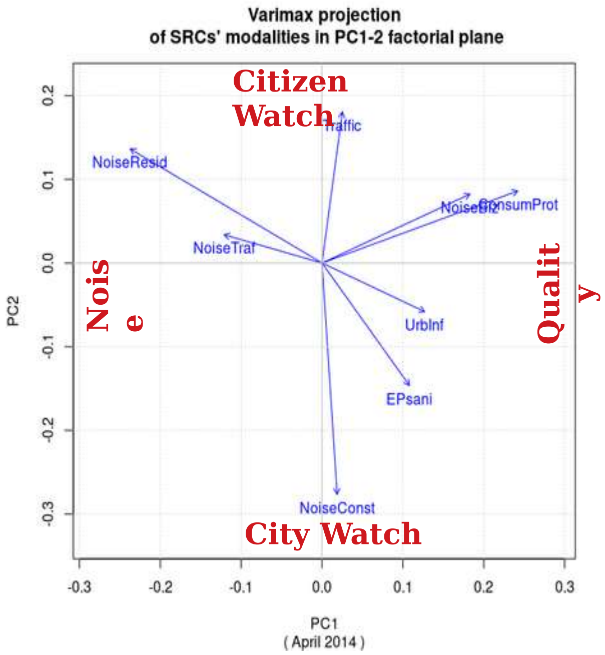

*__Figure__ 1: Maximized significance of projected variables in the rotated first factorial plane PC1-2, using the Varimax method and after feature extraction and selection (April 2014).*

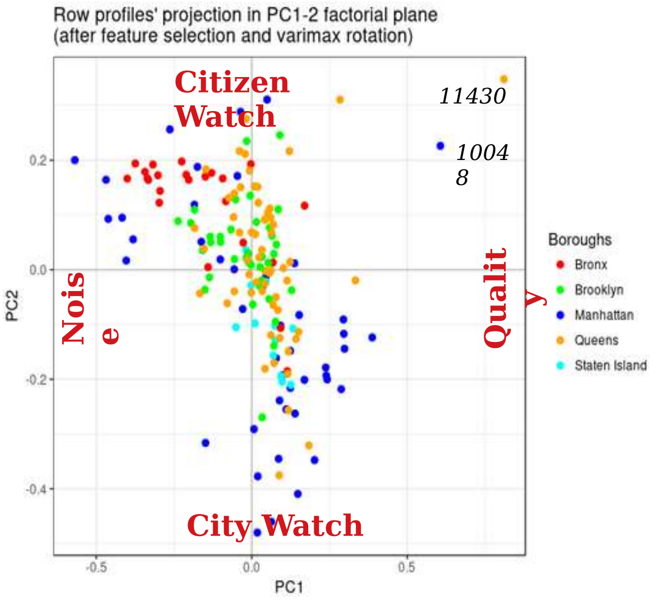

***Figure 2:** Orthogonal projections of individuals ZIP codes onto the varimax-rotated loading directions (PC1-2), color-coded per borough and after feature selection. (April 2014)*

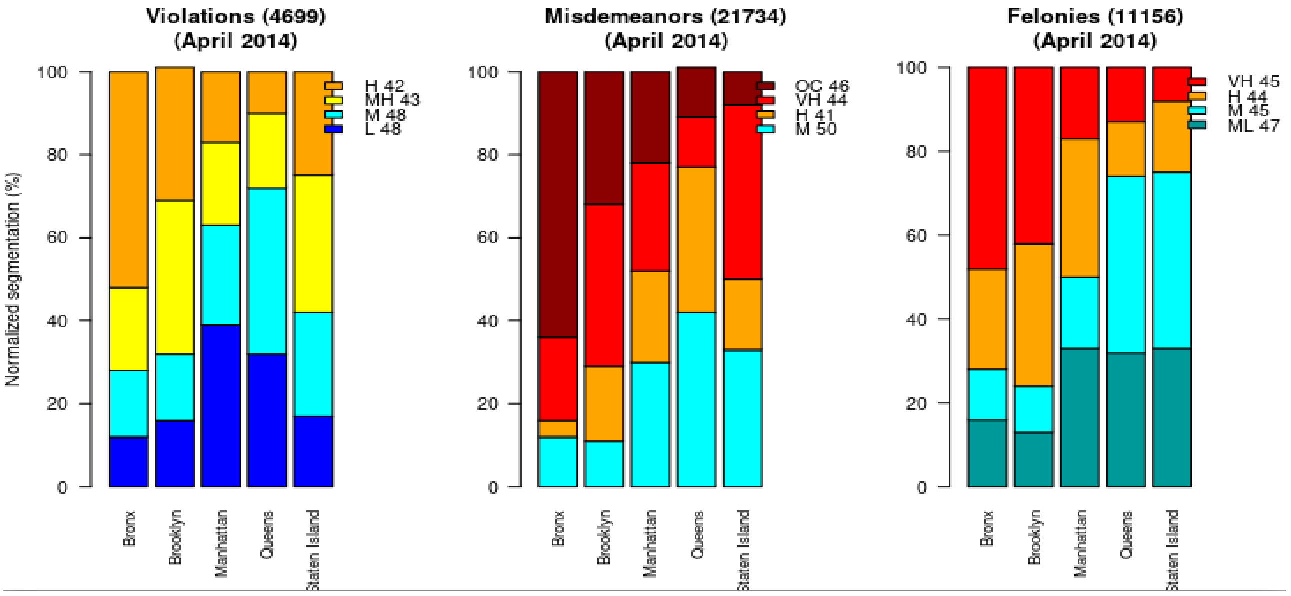

***Fig. 3***: *Borough crime index shown as normalized (to 100) segmentation for each crime modality as a function of borough. Legends show color-coded ordinal values followed by the bucket size (i.e. number of observed ZIP code areas) across boroughs*

In order to conduct Multiple Correspondence Analysis on SRCs' 13 modalities and crimes reports' 3 modalities, we elected to discretize each categorical variable modality frequencies in 4 bin levels with balanced bin memberships (viz. Section 3-2-1. of the complete Report). The result was (13+3) * 4 = 64 modality levels with a crime report segmentation per borough as shown in Fig. 3 below

A more accurate picture emerges from the discretization in 64 modality levels leading to MCA. Fig. 4 exhibits modality levels with quality of representation better than 20% and individuals projections in the first factorial plane (PC1-2).

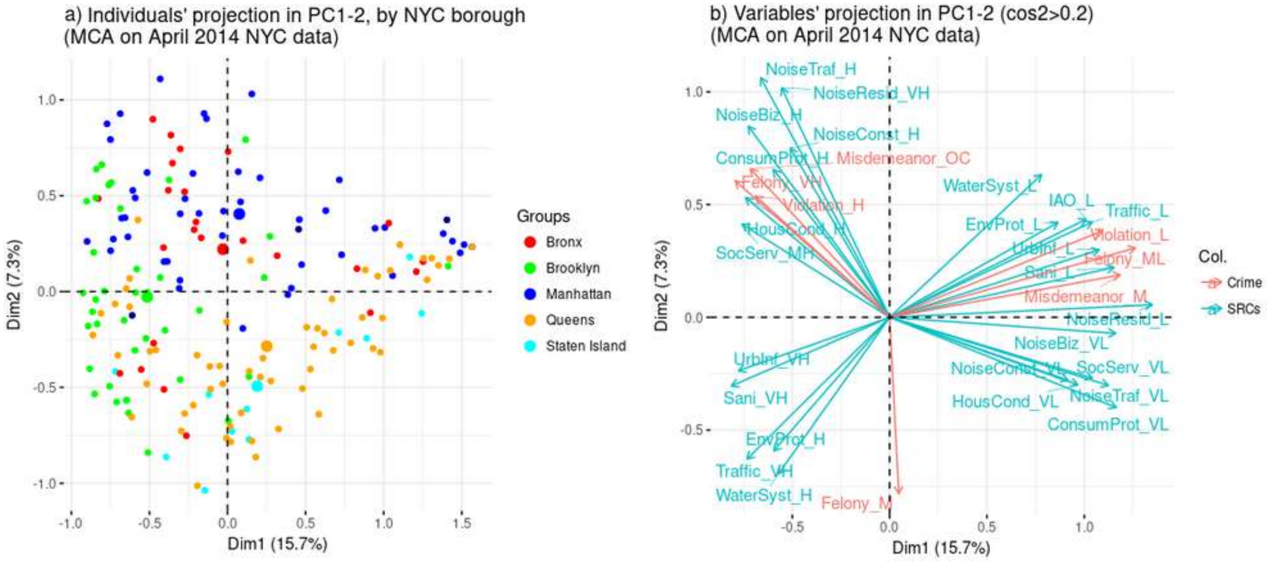

*Figure 4: MCA based representations in the 1st factorial plane (April 2014 NYC data) of variables' modalities' levels (categorical SRCs in turquoise, NYPD crime in red), for* ***a)*** *all individual ZIP codes,* ***b)*** *modality levels with representation better than 20%.*

To better understand the above joint representation of modality levels (as variables) and ZIP codes (as individuals) in the light of crime rate, we can visualize crime modality levels more finely to distinguish where each ZIP code and corresponding borough fits on the scale: low (L), medium-low (ML), medium (M), medium-high (MH), high (H), very high (VH) and out-of-control (OC) for each crime modality[1].

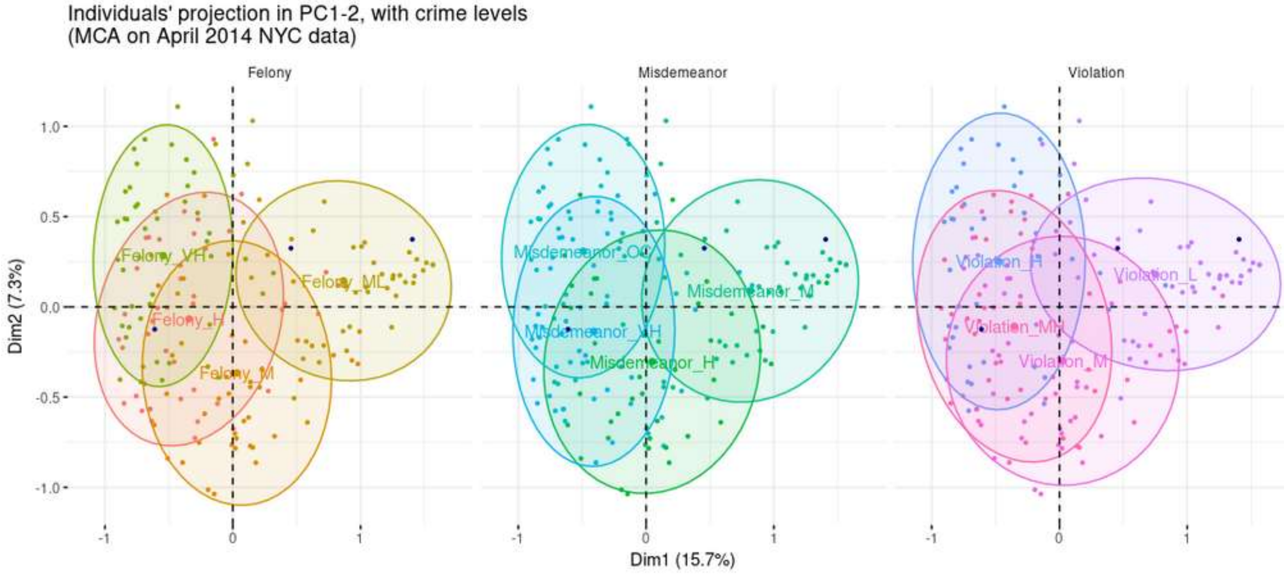

***Figure 5:*** *PC1-2 projections of individuals, color-coded according to their crime rates' modalities' levels. Ellipses are drawn for 75% confidence intervals in the $\chi^2$ sense, assuming standardized normal distributions for every crime modality's level.*

Common to all studied periods, crime levels increase gradually and clock wise, from the 1st quadrant on to the 4th, 3rd and 2nd where rates for violations, misdemeanors and felonies are H, OC and VH respectively. However ellipse sizes and orientations vary dramatically.

Based on MCA results, this preliminary survey included the Ward2 hierarchical clustering (HC) of data for each time period. We supplemented HC with k-means consolidation, a method which permits overcoming

| ***Cluster class*** | ***ZIP codes in class*** | ***IEP (%)*** |
|---|---|---|
| **1 (●)** | *33* | *20.3* |
| **2 (▲)** | *42* | *28.6* |
| **3 (■)** | *44* | *14.8* |
| **4 (+)** | *29* | *18.2* |
| **5 (☒)** | *29* | *18.2* |

[1] *The choice of terminology is completely arbitrary on the ana whole scale of reported crimes.*

difficulties in classification inherent to Ward2 HC (Section 3-3-2. of the complete report). It revealed 5 classes as shown right, with their IEP values, color-coded according to the scores representation in the first five dimensions. Fig.6 provides a partial view of the cluster distributions.

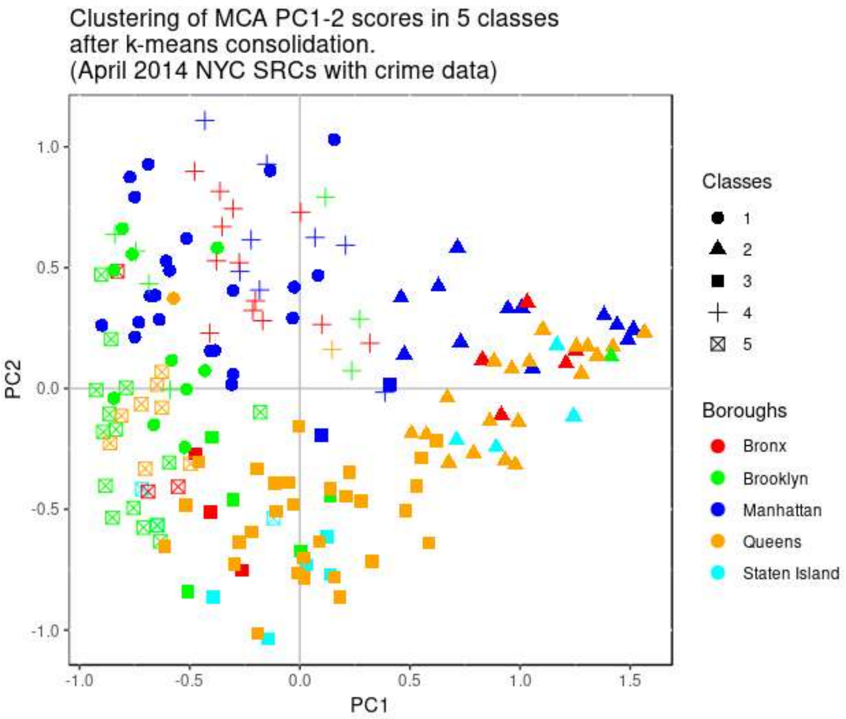

***Figure 6***: *Graph of scores in the first factorial plane after k-means consolidation, where each ZIP code is shown to belong to a specific cluster class, identified by a distinct symbol. ZIP codes are color-coded following their borough, following Fig. 4a.*

Common to all periods from 2010 to 2018, borough and class membership do not coincide.

**Bronx** (red symbols) ZIP codes are distributed among 4 distinct classes (4, 2, 5 and 3), class 4 being prominent.

**Brooklyn** (green symbols) ZIP codes are distributed among the 5 classes, class 5 being prominent.

**Manhattan** (blue symbols) ZIP codes are mainly seen in cluster classes 1 ,2, 4 and 3, in decreasing order of importance.

**Queens** (orange symbols) ZIP codes appears in clusters classes 3, 2 and 5, in decreasing order of importance.

**Staten Island** (cyan symbols) ZIP codes are divided between cluster 3, 2 and 5, in decreasing order of importance.

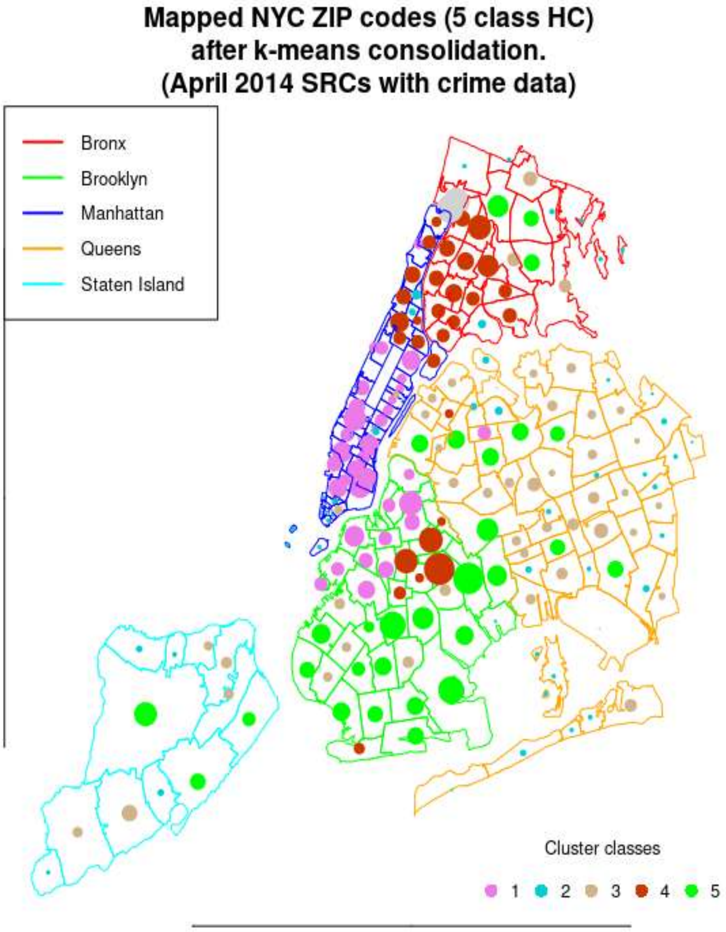
Mapped NYC ZIP codes (5 class HC)
after k-means consolidation.
(April 2014 SRCs with crime data)
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Cluster classes
1
2
3
4
5

***Figure 5***: *Topological mapping of the 5 class cluster obtained from Hierarchical Clustering (HC),* ***after k-means consolidation****. Gray colored dots are either outliers or ZIP codes areas otherwise not included in the analysis.*

A topological representation of class membership provides instant geolocation information. The semantic analysis of classes (i.e. understanding what they they stand for) is based on the analysis of which modality level contribute more significantly to their construction during HC. This includes over-represented and under-represented modality levels for each class. For the sake of clarity we only give a brief account of over-represented modality levels for the April 2014 period.

***Class 1***: 🟪 (orchid colored dots) High (H) incidence of SRCs: *"*`NoiseConst`*"*, *"*`ConsumProt`*"*, *"*`NoiseBiz`*"*, *"*`HousCond`*"* and *"*`NoiseTraf`*"*.
Crime related modalities do not play a significant role in the construction of that class

***Class 2***: 🟦 (blue colored dots)

Low (L) to very low (VL) incidence of SRCs, normally associated with dense urban areas, traffic, public housing, sustained street-level commercial activity ("*NoiseResid*=L", "*NoiseBiz*=VL", "*ConsumProt*=VL", "*HousCond*=VL", "*Sani*=L", "*Traffic*=L", "*NoiseTraf*=VL", "*UrbInf*=L", "*NoiseConst*=VL")
Crime related modalities are significant and show a moderate to low crime rate: "*Felony*=ML", "*Misdemeanor*=M", "*Violation*=L".

***Class 3***: 🟫 (tan colored dots)
Salient SRC modalities show medium to high concern for the condition of public places and urban infrastructure ("*Sani*=H", "*EnvProt*=MH", "*UrbInf*=H"),

- moderate residential noise related calls: ("*NoiseResid*=M"),
- low levels of complaint about factors normally associated with public housing, commercial areas, and construction work: ("*NoiseBiz*=L", "*HousCond*=L", "*NoiseTraf*=L", "*NoiseConst*=L", "*ConsumProt*=L")

Crime related modalities are significant with slightly higher level of incidence than for Class 2: "*Felony*=M", "*Misdemeanor*=H" "*Violation*=M".

***Class 4***: 🟥 (red colored dots)
Medium (M) to very high (VH) incidence of SRCs, usually associated with high population densities, public housing infrastructure and lower wealth: "*NoiseResid*=VH", "*UrbInf*=M", "*Sani*=M", "*ConsumProt*=M" "*NoiseTraf*=H", "*EnvProt*=M", "*Traffic*=H", "*SocServ*=MH".
Crime related modalities play a primordial role, with very high crime levels: "*Misdemeanor*=OC", "*Violation*=H", "*Felony*=VH".

***Class 5***:🟩 (green colored dots)
High incidence of SRCs: "*Traffic*=VH", "*Sani*=VH", "*EnvProt*=H", "*NoiseConst*=M", "*UrbInf*=VH", "*ConsumProt*=M", "*NoiseResid*=H", "*SocServ*=MH", "*NoiseTraf*=M"
Crime related modalities play a significant role in the construction of that class ("*Felony*=VH", "*Violation*=H", "*Misdemeanor*=OC"), at a level identical to that of class 4.

The main differentiating factors of Class 5, when compared to Class 4, are: "*Sani*=VH", "*EnvProt*=H" and "*UrbInf*=VH" as well as "*NoiseConst*=M", the latter being an SRC variable which plays no significant role in the construction of Class 4.

## Temporal evolution – On-going and future work

The study of how scores and loading directions (individual projections and variable representation in the factorial planes) evolve over time is made somewhat difficult by the fact that the projection basis changes for each data sets. For that reason a straightforward comparison of plots (say factorial plane PC1-2 from April 2010 with PC1-2 from April 2018) is unlikely to give us worthwhile answers as to the evolution of urban areas (ZIP codes here) described by any number of categorical variable.

The same applies to hierarchical clustering where algorithmic class membership optimization entails that class boundaries and modality level representation in them are also modified at each step. The uptake is that even if the same number of cluster classes were to be found for different data set periods, the construction of each class (i.e. their semantic meaning) would very probably be different.

However on-going work should shortly permit the joint scatter plot representation of distinct data sets based on the MCA representation basis of one of them. This will permit to visualize the trajectory of any ZIP code area of choice over time.

We propose to tackle those difficulties by applying shallow 2-layer deep Neural Network based models, designed to provide semantic information of proximity (in time and space) between words (urban events including POIs) in a bag of such words. In this upcoming effort, bags of words are the basis of an unsupervised learning process, where each word is ultimately represented by a vector in a n-dimensional space (n being quite large). Proximity of two vectors in that space indicate semantic proximity in both meaning and syntactic ordering, thereby reflecting not only proximity and perhaps correlation, but also importance in a given urban context and its temporal evolution.  This technique, Word2Vec, was created by a team led by Tomas Mikolov while at Google, ca. 2013.  Touted early on by its proponents as a necessary evolutionary step toward machine intelligence, its use to study urban semantics is unprecedented.