MASTER IN INNOVATION AND RESEARCH IN INFORMATICS

Specialty: Computer Networks and Distributed Systems

Multivariate statistical and semantic analysis of urban environments applied to New York City

Presented by: Cedric K. Bhihe
<cedric.bhihe@gmail.com>

Defended: February 2019

Advisor: Prof. Jorge García Vidal
Dept of Computer Architecture
Co-advisor: Prof. José Mª Barceló Ordinas,
Dept of Computer Architecture

FACULTAT D'INFORMÀTICA DE BARCELONA (FIB)
UNIVERSITAT POLITÈCNICA DE CATALUNYA (UPC)
BarcelonaTech

Foreword

A few years ago, Ben Wellington published an article[i] about mapping New York City's noisiest neighborhoods, soon followed by another one producing results on the hidden circumstances behind New York City's permanent traffic gridlock[ii]. Those two articles, published in the New Yorker, were meant for a wide (although somewhat upscale) readership. They also revealed that the author had used analytical and statistical methods based on a rich data base. That database is *NYC Open Data*[iii], a trove of information geographically and temporally more precise than census tract scale data as made publicly available by the US Federal Government. We tapped it. This report describes why, to what extent and how.

Large urban areas are complex environments, characterized by a large, heterogeneous set of co-varying quantities which put together (in time and in space) constitute a body of *urban semantics*. As such the fabric of cities is difficult to understand by both businesses and city government officials. In the face of sometimes conflicting priorities and difficult to grasp multi-dimensional issues, businesses and municipalities often rely on empirical data. That data (however incomplete) becomes the basis for intuitive, non-explicit and unverified correlations, the which lead to decisions or to corrective actions. We might well be able to do better on both counts. This work is an attempt to demonstrate how.

The data in question may be static[iv], or dynamic[v]. In the latter case it is harvested continuously by different agencies, municipal entities and social networks. For data to be accessible to us, it must be stored digitally in such a way that its posterior analysis is possible. Ultimately at stake is for city officials and business alike to better grasp urban semantics, their evolution and predictability. Decision-makers have common objectives: to make better decisions, to build better strategies and better policy, on which to base an optimal allocation of resources both in time and space.

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 tools become available. The increasing size of modern conurbations has direct consequences in terms of emerging complexity in grasping its semantics. That makes the ability to manipulate big data in an automated way attractive both to business people seeking to maximize their ROI and to city officials seeking to maximize the well-being of inhabitants under their responsibility.

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 long term 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 the global warming phenomenon and in the resulting soaring of living costs in large cities. It does not constitute an actual threat to this project, and should therefore not detract from its purported timeliness and usefulness.

The key issue, as it is often perceived by the potential beneficiaries of this work, is how to best interpret urban semantics, where every data point constitutes a manner of "word", in order to make allocations decisions.
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.

[i] https://www.newyorker.com/tech/elements/mapping-new-york-noise-complaints

[ii] https://www.newyorker.com/tech/elements/uber-isnt-causing-new-york-citys-traffic-slowdown

[iii] https://opendata.cityofnewyork.us/

[iv] Examples of static data include points of interest, topographical information, census statistics, IRS income reports, academic achievement by zone, etc.

[v] Examples of dynamic data include calls to 911, calls to 311, car traffic, weather, accidents, etc.

Table of Contents

1. Introduction

Since 2010, between 2,500 and 15,000 daily calls to 311 are recorded in New York City, NY (NYC). Those service request calls (SRCs) are logged with a slew of attributes (more than 50 fields are available per call), on the location of the incident, its nature (e.g. noise, public housing conditions, street potholes, stray animals, rodent sighting, ailing trees, barking dogs, unsanitary food establishments, uncivil behavior, parking violations, etc.). SRCs' attributes include time and date , as well as geo-location of the incident, reasons and object of the call. They are freely available[vi] on Internet by courtesy of the City government of New York.

Simultaneously the NYPD, New York's Police Department, registers over 1000 daily felonies, misdemeanors and violations[vii]. This affords the curious analyst a rich overview on the type of issues being reported, their location, and frequency. It is also an invitation to scrutinize possible correlations between the statistics of geo-located 311 SRCs and other factors such as population density, type of criminality, median income and IRS declared jobless benefits in income tax returns. We will restrict our geographical reach to ZIP code areas of neighborhoods in the 5 boroughs of NYC per Figure 1: Manhattan, Brooklyn, Queens, the Bronx, and Staten Island. All other ZIP codes are excluded.

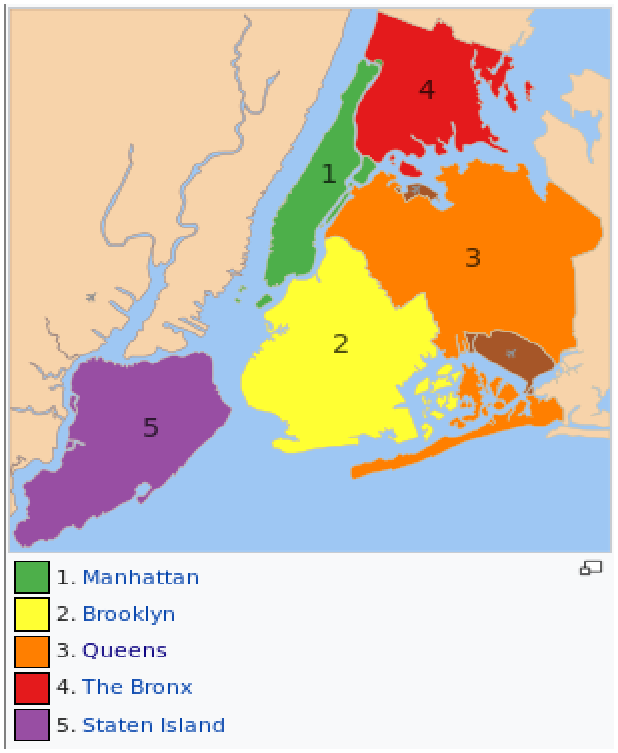


***Figure 1**: NYC's five historical boroughs (source: Wikipedia)*

We discuss and justify the choice of the ZIP code area as the right scale for this survey in a discussion at the beginning of Section 3.

In the end curiosity is what really subtends every human endeavor. More specifically in our case, the motivation to embark on this study was (i) to evaluate how much insight can be gained from realistic multidimensional data using classical multi-variate analysis (MVA) exploratory tools, (ii) to explore a relatively recent Machine Learning technique, Wordd2Vec, in a effort to capture the urban semantics from the same data.

In a first part (A), we present results based on Correspondence Analysis (CA), Principal Component Analysis (PCA), Clustering and Multiple Correspondence Analysis (MCA) to conduct data exploration, feature extraction and predictive modeling. Whenever suitable an effort is made to also offer a critical discussion of obtained results. In a second part (B), we present our implementation of Word2Vec and its concrete application…..

A less theoretically minded question is ultimately to reveal evolution patterns in the urban fabric of NYC. Our objective is to try to extract predictor-variables on the scale of a ZIP code[viii] area. Dealing with ZIP code tabulated areas is in general less precise than doing so with census tracts as ZIP codes topographical areas change frequently without a change in ZIP code. To boot there are approximately 50% more census tracts in the US as there are ZIP codes.

Possible applications are many:
- predict crime,
- link complaints about urban nuisance to certain neighborhoods and illustrate those neighborhoods in terms of social-economical categories,
- produce the basis reference model to help decide where to locate what business for maximum attractiveness to customers and return on investment for investors,
- optimize resources to better manage dense urban areas.

[vi] *https://data.cityofnewyork.us/Social-Services/311-Service-Requests/fvrb-kbbt*
[vii] *https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243*
[viii] *ZIP or "Zone Improvement Plan" is a territorial mapping used by the US Postal Service (USPS) for snail mail delivery since 1963.*

Although we provide a Table of Contents, a brief description of how this report is organized follows.

- In part A section 2, we present the protracted process of extracting data from various databases. This included cleaning it (in particular in terms of missing values) and modifying it from a time record format to a location oriented frequency table. Data cleaning, while not intrinsically or conceptually difficult, is a task laden with traps. It occupied over 170 hours of our time. This section sheds light on why and how. It can be skipped and the reader may go directly to the analysis of Section 3.

- Part A section 3, encompasses the multivariate data analysis including CA and PCA on NYC311 SRCs, Clustering and MCA on 2 categorical variables and a total of 16 modalities, plus 1 (illustrative) supplementary variable and 2 quantitative variables. The initial analysis is performed on the April 2014 data-set, the which constitutes our training data. Our testing or validation data are the April 2010 and April 2018 data-sets.

- Part A section 4, offers a general conclusion on obtained results and suggests new directions to pursue this multivariate analysis.

Due to external constraints imposed on this work, results produced in part A of this report were obtained exclusively by relying on custom R scripts. Part B, on the other hand, covers the implementation of advanced techniques (Word2Vec) and results from a computational implementation using Python 3. Notwithstanding external constraints, we cannot but warmly advise interested coders, not to code with R during the data cleaning phase. R is quirky at times, and has either scant or too much documentation to wade through at other times. Being FOSS, it does benefits from a community based ecosystem, and it is correct to say that the answer to many questions during development can be successfully crowd-sourced. This however does not normally include extremely specific situations, where the coder is largely left to her own device.

All in all data ETL can be performed with R, but many times it is awkward at best. The rest of the time it is mostly grueling and slow depending on the exact nature of the task. Many R proponents will readily swear under oath that the same is true of ETL with any alternative to R, but heed our dispassionate advice: if you have the choice between R and Python for ETL, pick Python to walk down the aisle and be forever thankful you did so.

All digital files (including input files, raw and processed data sets, scripts and result files) are made fully available to the reader, in a way which preserves the data structure and the files' hierarchical organization on any computing platform. Paths in adjoined scripts and occasionally in the body of this report are shown using Unix-like formats. However they can be transposed easily to any addressing format of the file system of your choice.

From the top containing folder "*`NYC311`*", the complete project's file tree is organized as follows. below means that we omit mention of some intermediate data files, obtained during the preliminary data processing phase. Those files are provided for the record. Their name usually starts with a time-stamp identifying the period to which they refer and ends with `__procXX.csv`, where XX is a double digit processing sequence identifier.

```
NYC311/
|____ Bibliography/
|____ Data/
|       |____ Geolocation/
|       |       |____ [7 shape files for NYC ZIP codes perimeter 2D drawing]
|       |____ 201x0400_nyc311_raw.csv
|       |____ 201x0400_nyc-crime-map_raw.csv
|       |____ 201x_zip-irs-exempt-unemp.csv
|       |____ [...]
|       |____ nyc_borough-zip.csv
|       |____ nyc311_00083-neighbors-common-border.csv # for ghost zip 00083 processing
|       |____ 20140x00_nyc_whole-data-set.csv # April 2014 data-set at start of analysis
|       |____ 201x0400_nyc_simple-whole-data-set.csv # April 2015 data-set at start of
analysis
|____ Report/
|____ Scripts/
        |__ 01_nyc311_input-parameters.R # defines basic period parameters and more
        |__ 02_nyc311_data-prep.R # clean up of raw data, serv. req. modalities reduction
        |__ 03_nyc311_missing-impute.R # NN-imputation or direct localization (GoogleMaps
API)
```

```
       |__ 04_apportion-ghost-zip_prep.R # prepare ghost ZIPs' obs apportionment to
neighbors
       |__ 05_nyc311_calls-by-zip.R # consolidates service request calls modalities per ZIP
       |__ 06_irs_median-inc-jobless.R # evaluate median income +joblessness per zip
       |__ 07_nypd_data-prep.R # clean up raw data, reduce crime modalities to 3
       |__ 08_nypd_crimes-by-zip.R # consolidates crime modalities per ZIP
       |__ 09_consolidate-by-zip.R # general consolidation
       |__ 10_apportion-ghost-zip_proc.R # apportion ghost ZIP's categorical counts
       |__ 11_apportion-ghost-zip_proc.R # apportion ghost ZIP's categorical counts
       |__ 12_apportion-ghost-zip_proc.R # apportion ghost ZIP's categorical counts
       |__ 13_apportion-ghost-zip_proc.R # apportion ghost ZIP's categorical counts
```

2. Data-sets

2-1. Terms and conditions of use

All raw data-sets used in this project are public and accessible for free under the US Freedom of Information Act[ix] (FOIA). Their use is regulated by the terms and conditions of use pertaining to each governing body responsible for their publication or production. Data dictionaries are generally made available in Appendix A, and the web pages harboring those terms are:

- http://www1.nyc.gov/home/terms-of-use.page
  for ZIP code centric and time-based NYC311 SRC data

- https://data.cityofnewyork.us/Business/Zip-Code-Boundaries/i8iw-xf4u
  for geometric ZIP code area boundary data

- https://www.irs.gov/statistics
  for ZIP code-centric income tax declaration data made available by the IRS

- https://www.census.gov/topics/income-poverty/income/data/tables/acs.html
  for ZIP code-centric unemployment benefit declared to the IR

2-2. Data scope and preparation – ETL

Data was generally available from various location on the web, from 2010 onward. We specialized our study to the months of April in 2010, 2014 and 2018 in order to be able to handle the corresponding volume of data. Raw files are available in ods and cvs formats at `NYC311/Data/`. Census data on population densities per ZIP code area was only available to us for the year 2016 and only for a limited number of ZIP code areas. We therefore do not include it in either one of our data-sets.

2-2-1. Duplicates, missings, and imputations

Every downloaded data-set was already fully labelled. A rapid inspection of raw data shows that "NA" (non-assigned / not-available) or erroneous values, referred to as “missings”, exist, but in such proportion that dealing with them was tractable. As described below, we either imputed, re-imputed, suppressed or researched missings by cross-referencing them between DBs, with the goal of avoiding issues of data bias.

– Service request calls (SRCs) to NYC 311

The two data sets *`yyyy0400_nyc311_raw.csv`* contain the raw data of NYC SRCs for yyyy={2010,2014,2018} as downloaded from *NYC Open Data*. That includes the call's object (description), date, time, ZIP codes and/or location (in several forms) of the reported matter and other less relevant information. We checked that data-sets contains SRCs (heretofore referred to as “dupes”) from different callers with the same object. Tracking down dupes is inherently complex

[ix] *The FOIA is a companion to the US Privacy Act of 1974 (5 U.S.C. 552a). Under the FOIA, anyone residing legally in the USA can make a request for a Federal Agency record.*

and we did not attempt it. More importantly, our study is concerned with people's spontaneous and independent tendency to call NYC 311 about aspects of their urban environment, which are important to them. In that sense dupes need not be eliminated; they are significant and represent a natural weighting for the data-set's observations. This will naturally influence observations' weights as represented later by marginals (row sums) in frequency tables.

Raw (unfiltered) data characteristics are shown in Table 1 for the April 2010, 2014 and 2018 data sets.

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

***Table 1**: Summary table of salient missings and other characteristics for raw NYC 311 SRCs data sets (before data cleaning). SRCs' modalities are available in the 2 files:* `Report/yyyy0400_nyc311_proc01_modalities.csv` *with yyyy={2014,2015}*

Figures 2 and 3 below represent missings for the period April 2014.

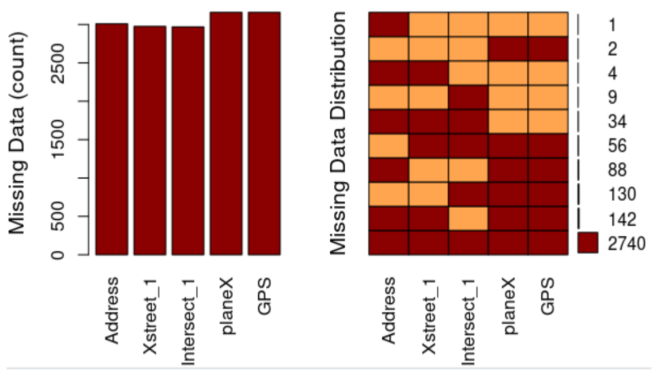


***Fig. 2**: Analysis of missings for April 2014 on the 3206 raw observations missing a ZIP code.*

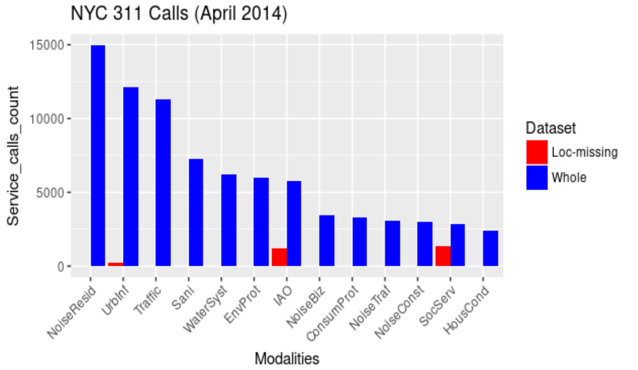


***Fig. 3**: NYC 311 SRCs' modality distribution for the whole data-set (blue) and for observations missing all location information.*

As sobserved from Figure 2 above, during the April 2014 period, 2740 observations or 3.4% of all observations, and 85.5% of the 3206 observations missing a ZIP code have no other geographic locator. Those observations cannot be attributed to any ZIP code and are therefore useless. Figure 3 compares the service request calls' modality distributions for observations missing all location information (including a ZIP code and denoted "*loc-missing*") and the whole data set. It is readily apparent that simply eliminating "*Loc-missing*" observations would disrupt our analysis in terms of the *SocServ* modality, while for other modalities the effect would be negligible.

For that reason, we proceeded to impute a ZIP code to the 466 RFC observations missing it in 2014, but not included in the *Loc-missing* subset of missings. In practice those observations miss a ZIP code but are nevertheless endowed with some other geolocation information:

- an address, and/or
- 2 cross-streets in the form of (Xstreet_1,Xstreet_2), and/or
- an cross-road in the form of (Intersect_1,Intersect_2), and/or
- planar (Euclidian) coordinates (planeX, planeY), and/or
- GPS coordinates (latitude and longitude)

Imputation was done by fully implementing automated requests to GoogleMaps, through its API, in R, for each one of the aforementioned cases. As a result more than 97% of all 466 observations missing a ZIP code could be imputed for the April 2014 data-set. The rest including the *Loc-missing* subset of observations were given the bogus ZIP code "99999" to be uses later as a supplementary observation.

As there is no structural difference between the April 2014 and April 2015 data-sets, graphical analysis results for missings were only shown for April 2014. From Table 1, in April 2015, 3069 observations or 3.0% of all observations, and 72.5% of all observations missing a ZIP code have no other geographic locator. Here again we treat missings following the same pattern and with a similar success rate as before.

– NYPD's crime reports for NYC's 5 boroughs

Crimes are reported according to 3 general categories, which coincide with the crime modalities used in our analysis. In decreasing order of severity, they are: **felonies**, **misdemeanors**, and **violations**. . They are described and instances listed in Appendix B per the NYPD's DB.

Data made publicly available by NYDP is completely devoid of ZIP information. However it does include planar localization and regular GPS coordinates. Because of the large amount of data involved in this study (close to 80,000 criminal observations) and of Google's imposed limitation on the number of queries (2500/day/account, as of 2018.04.30) , relying on our Google Maps API's implementation to impute a ZIP code to each crime was not deemed practical. We therefore developed two original algorithms to determine the ZIP code of each NYPD crime observation based on its planar (Cartesian) coordinates.

The first algorithm to be developed was based on nearest neighbor topological distance. It uses previously compiled ZIP code areas with planar and/or GPS coordinates for SRCs to NYC 311. The ZIP code of the 311 SRC closest in space to a crime's GPS or planar coordinates is imputed to the crime. This method is approximate and yield mixed results.

The second algorithm is exact and yields excellent results. It determines the ZIP code of every crime observation based on its planar coordinates and shape-formatted ZIP boundaries mapping data, downloaded from the *NYC Open Data* repository and made available to the reader under `Data/Geolocation/`.

The latter algorithm is general and is implemented in the form of a function, `whichBoxF()`, available at `Scripts/06_nypd_data-prep.R`. Its reaches its imputation target in more than 96% of all recorded observations. The rest, i.e. less than 4%, falls in the *missings* category and kept in supplementary observation with imputed bogus ZIP code "99999". Tables 2 below summarizes missing ZIP code "99999" imputation for crime data collected by NYPD in April 2014 and April 2015. A Chi square test of the NYPD crime data-sets' missings show that there is a significant association between missings and crime modalities. Simply suppressing missings would introduce a bias in the distribution.

2-2-2. SRCs' modality dimensional reduction

Service Request Calls' modality dimensional reduction was conducted by applying filters tailored to the semantics of the raw data's two columns: "Complaint", and "Descriptor".

The reduced modalities data-sets exhibit 13 modalities down from 170 and 178 (in Table 1, for April 2014 and April 2015 respectively) according to the description and distribution of Table 2. Noise related complaints remain the first reason for SRCs to 311 in NYC, with overall frequencies in noise related calls of 31.1% and 31.5% in 2014.and 2015 respectively.

April 2014	**Felony**	**Misdemeanor**	**Violation**	Total
non-missings	11,327	22,094	4,784	38,205
missings	481	985	64	1,530
Total	11,808	23,079	4,848	39,735

April 2015	**Felony**	**Misdemeanor**	**Violation**	Total
non-missings	11,669	22,080	5,010	38,759
missings	193	473	11	677
Total	11,862	22,553	5,021	39,436

Table 2*: Summary of misssings after imputation for the NYPD's crime datasets in NYC*

Table 3 is based on data after ZIP cleaning and missings imputation. SRC modality ranking change show that the perceived (and perhaps also real) traffic noise related SRCs increased markedly between April 2014 and April 2015.

2-2-3. ZIP code cleaning

At this data preparation stage, the data consists of a mixture of correctly formed and ill-formed ZIP code fields for each observation. An ill-formed ZIP code may be a code, which either does not have exactly 5 digits, or does not exists officially, or is otherwise not consistently found in US government DBs.

To easily associate ZIP codes and borough, we include a list of 200 ZIP codes and corresponding boroughs in Appendix C.

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	↗↗

Table 3*: SRCs' consolidated modalities after dimensional reduction. The right most column indicates changes in modality ranking from 2014 to 2015.*

For our purposes, ill-formed ZIPs include ZIP+4 codes of the form 11355-1024, where the last four digits identify a geographic segment or a PO box within the five-digit ZIP delivery area. In those cases we simply suppress string characters ranging from position 6 to the end.

Inadmissible ZIP codes also include ghost ZIP codes. One of them appears in our DBs as "00083". The NYC 311 service request call data-set includes it along with surrounding and overlapping ZIP codes. So do the NYPD's crime DB, and the topological ZIP code area boundary DB also found in the NYC Open Data repository. Within the NYC area it designates the Central Park area in Manhattan. But because it overlaps with other official ZIP code areas surrounding it, observations identified by that ZIP code should be instead apportioned to neighboring ZIP code areas. Figure 4a reveals the Zip mapping in that area, showing official ZIP code areas boundaries mapping Central Park in Manhattan. Surrounding ZIP codes are 10019, 10022, 10065, 10023, 10021, 10075, 10028, 10024, 10128, 10025, 10029, and 10026.

The use of ZIP code 00083 is incompatible with IRS and Census Agency DBs. To overcome that difficulty, we calculated the common boundaries between the 00083 ZIP code area boundary and surrounding ZIP areas boundaries.

Our goal is to apportion observations attributed to ZIP code 00083 to surrounding ZIP codes areas proportionally to the lengths of the boundaries they share, and in a way which should remain modality-neutral.

Figure 4b represents the Cartesian topological mapping, and Table 4 shows the computed proportion of common boundary lengths between Central Park's 00083 ghost ZIP and surrounding ZIP code areas. The algorithm developed can operate on arbitrary sets of ZIP codes.

After correcting for ill-formed ZIP codes, ghost ZIP codes, ZIP codes with zero surface area (i.e. corresponding to PO boxes), ZIP codes situated outside NYC's 5-borough area, we observed a little over 200 unique ZIP codes in our data sets (year in year out).

Finally we let a limited number of ZIP codes areas be absorbed by their "main neighbor", according to the following rationale:
whenever at least 75% of any two given ZIP codes' boundaries coincide, we apportioned the observations attributes of the ZIP code whodse area had the shorter overall boundary length to the longer boundary length neighboring ZIP code area.

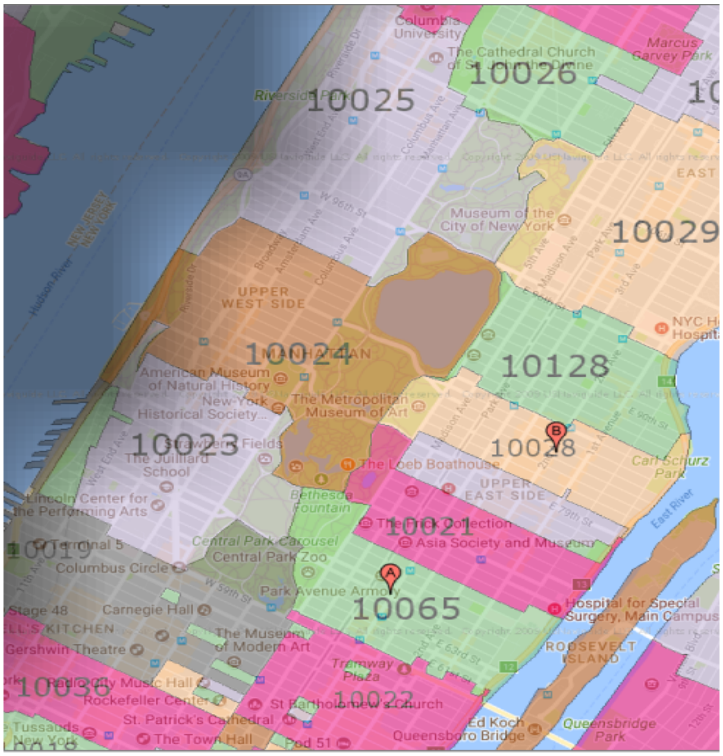


Fig. 4a: Detail of the ZIP code area map of Manhattan, showing how neighboring ZIP code areas pave Central Park piece-wise.

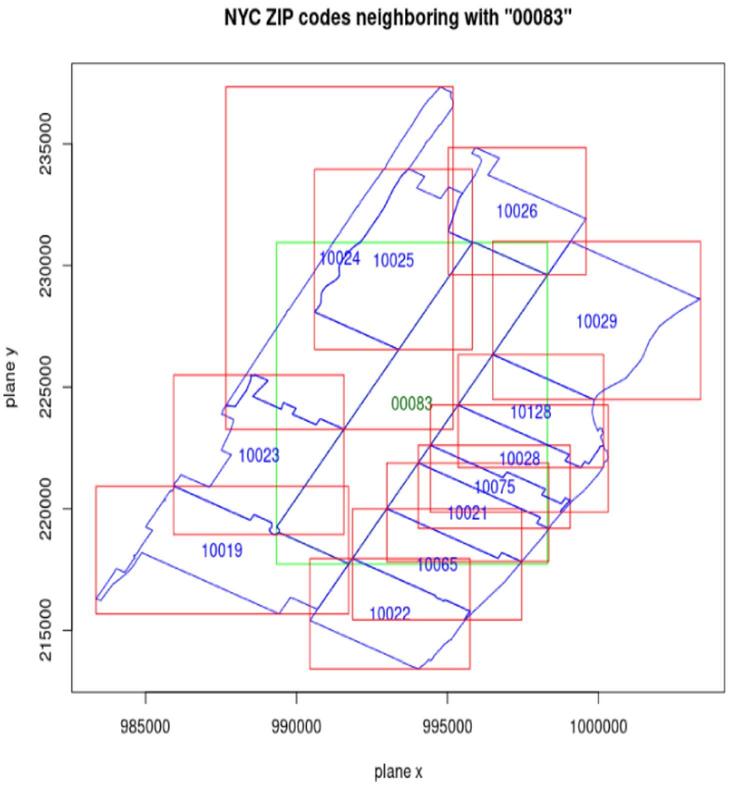


Fig.4b: New York City government ZIP code limits mapping Central Park in Manhattan and resorting to ghost ZIP code 00083 (at center in green).

ZIP code	Common boundary length (ft)	Common boundary length proportion (%)
00083	32,710.8	100.0
10019	2,651.4	8.1
10022	259.2	0.8
10065	2,341.4	7.2
10023	4,864.4	14.9
10021	2,150.5	6.6
10075	833.0	2.5
10028	1,889.7	5.8
10024	3,748.7	11.5
10128	2,371.1	7.2
10025	5,031.3	15.4
10029	3,749.0	11.5
10026	2,821.2	8.6

***Table 4**: 00083 ghost ZIP code area shared boundary analysis.*

The starting point for our statistical analysis of section 3 are the three data-sets located at:

```
NYC311/
|___ Data/
        |___ 20100400_nyc_simple-whole-data-set.csv
        |___ 20140400_nyc_simple-whole-data-set.csv
        |___ 20180400_nyc_simple-whole-data-set.csv
```

We noted a few cases (fewer than 30 observations per data set) of missing response variables (“`medianInc`” and “`jlBenefit`”) in each data set. This means that the Internal Revenue Service (IRS) chose not to make the corresponding ZIP code area’s tax return statistical data public. This should not pose a problem for the coming clustering analysis (classification) , but obviously does so in any regression-like approach.

So the reader may comfortably associate ZI3-1. Principal Components Analysis and Correspondence Analysis P codes to NYC boroughs, we provide a list of more than 200 ZIP codes and their corresponding boroughs in Appendix C.

3. Multi-Variate Analysis

The fact that we choose to conduct the analysis to come at ZIP code are scale and even comment its results at borough scale fits the objective of dimensionality reduction in Multi-Variate Analysis (MVA). However it should not detract from the fact that we could as easily conduct the same analyses not at ZIP code, but at block levels or even (on the basis of isolated events) at GPS coordinates level. At event scale for instance, every logged call, every reported crime, in general every included event would constitute a data point.

This in turn would force us to debate another important aspect of our analyses, in particular in the context of clustering analysis: as event would be the new individual data points, how appropriate weighted or unweighted event approaches are to reveal the semantics of urban dynamics ? In short, a weighted event approach treats all clusters equally, while an unweighted one considers that cluster classes are more or less important as a function of their memberships (i.e. their cardinality). In general, unweighted approaches are preferred unless there is reason to believe that observations should have different weights; e.g., perhaps classes of objects have been unevenly sampled ...

We chose to consider our data at ZIP code level to avoid that discussion, due to the fact that our categorical variables and their modalities are largely dissimilar in nature. A crime is not generally reported arbitrarily, i.e. following the whim or current state of mind of the person reporting it, but rather because, being a breach of social contract, failure to report it is in itself a punishable offense. By contrast a service request call to 311 by a NYC resident or visitor may or may not be placed without consequence. In all likelihood, certain city dwellers are more or less prone to report urban pathologies, visitors arguably much less still. This is turn may not solely be ascribed to the psychological make-up of any individual but probably also to a large number of external factors which influence and determine at least in part any potential caller’s decision to call. SRCs are in this sense arbitrary and are (at least from the perspective of an uninformed witness) placed at will. That however constitutes a completely different event sampling mechanism likely to introduce a bias at event scale. In our opinion and for the sake of simplicity, this was ample justification for considering our analysis at ZIP code area scale.

Our first approach was to consider the contingency table made of the *NYC 311 SRCs* categorical variable’s 13 modalities and 200+ zip codes seen as the modalities of a second categorical variable we name *Location*.

Among the zip codes the last one, “99999”, will either be overlooked or be treated as a supplementary observation.

We identify between 20 and 30 zip codes with row marginals smaller than 5/(sum of calls), where, e.g. for April 2014, the total number of calls so far retained in our analysis was about 78,700. We suppress those ZIP codes from our contingency table, on the grounds of they representing less than 0.2% of monthly SRCs (see footnote[x]). The resulting table for April

[x] *A χ^2-test of independence on the small contingency table made of ZIP codes to be suppressed and their RFCs’ modalities led us to reject the null hypothesis of independence. To that end, data was reduced so no zero valued marginals could perturb the test.*

2014 is made of 181 zip codes (row labels, row index *i*) and 13 SRC modalities (column labels, column index *j*).

Next we identify table cells where low frequency and (simultaneously) high contributions to the χ^2-statistic value for the test of association of the two categorical variables may perturb the subsequent analysis. We define as low cell count or low frequency any contingency table cell count smaller than 5. For every data set there are between 300 and 500 such cells. Based on the chi-square-test statistic:

$$\chi^2 = \sum_{i=1}^{N} \frac{\left(Count_{obs} - Count_{exp}\right)^2}{Count_{exp}}$$

we calculated the contribution of every low frequency cell to the overall χ^2 statistic value and found that for low frequency cells: (i) no contribution exceeds 1%, and (ii) only 1 contributions exceed 0.1%, for a 2-sided χ^2 test statistics of 43,338. As a result the Pearson chi-square test for significant association (dependence) between row & column categories is deemed appropriate. It leads to the clear rejection of the null hypothesis, with a p-value of the order of 10^{-4}:

> H_0: *"In the population, the two categorical variables are independent."*

The above p-value was computed from Monte-Carlo simulations with 10,000 replicates.

Inspecting marginals, we see that SRCs' modalities with lowest weight across zip codes are:
"SocServ" ($f_{.j} \approx 0.019$ for j=11), followed by HousCond ($f_{.j} \approx 0.030$ for j=1), and "NoiseConst" ($f_{.j} \approx 0.038$ for j=4).

3-1. Principal Components Analysis and Correspondence Analysis

3-1-1. PCA

From the contingency table made of the April 2014 SRC categorical variable's 13 modalities and 202 ZIP codes, we build a conditional frequency matrix, which we appropriately center based on a cloud centroid (of column marginals) with embedded χ^2 metric. We perform a PCA on that matrix excluding ZIP code "99999" as well as 21 other individual ZIP codes whose marginal row counts are smaller or equal to 5. 180 individual are left. We first include "10463" and then repeat the analysis considering it as a supplementary observation. The number of significant dimensions is 3, based on the criterion that the total explained inertia be at least 70%. Results are graphically summarized in Figures 5a, 5b and 5c.

When included in the analysis (as in Fig. 5a), ZIP code "10463" stands out as the biggest individual contributor to the construction of the 3 first dimensions with 17%, 22%, and a whopping 52% for PC1, PC2 and PC3 respectively. The ZIP code area roughly represents a one kilometer radius in the Bronx, known as Riverdale. Topologically neighboring ZIP areas are: 10467, 10468, 10471.

Riverdale has one of the highest population density in NYC with more than 30,000 housing units and more than 18,000 registered inhabitants per square kilometer. Understandably `HousCond` and other SRCs to NYC 311 are disproportionately large in Riverdale, when compared to other NYC areas.

Besides a noticeable change in cloud shape, a pronounced change takes place when we consider "10463" as a supplementary individual. It concerns principally the variable `HousCond`, whose:

- quality of representation in the first 3 dimensions (PC1, PC2, PC3), and
- contributions to the construction of dimensions

both plummet. Meanwhile the contributions and quality of representation of the other two main variables `NoiseResid` and `NoiseConst` are somewhat redistributed among dimensions or in some cases increased: e.g. for `NoiseResid` $\sum_{\alpha=1,2,3} \cos^2_\alpha$ goes from 0.94 to 0.98.

The first factorial plane (PC1-2) registers an increase in inertia explanatory power (from 54% to 64%) – Fig. 5a. Meanwhile PC2-3 and PC1-3 register a decrease from 48% to 43% and from 44 % to 37% respectively (see Fig. 5b and Fig. 5c).

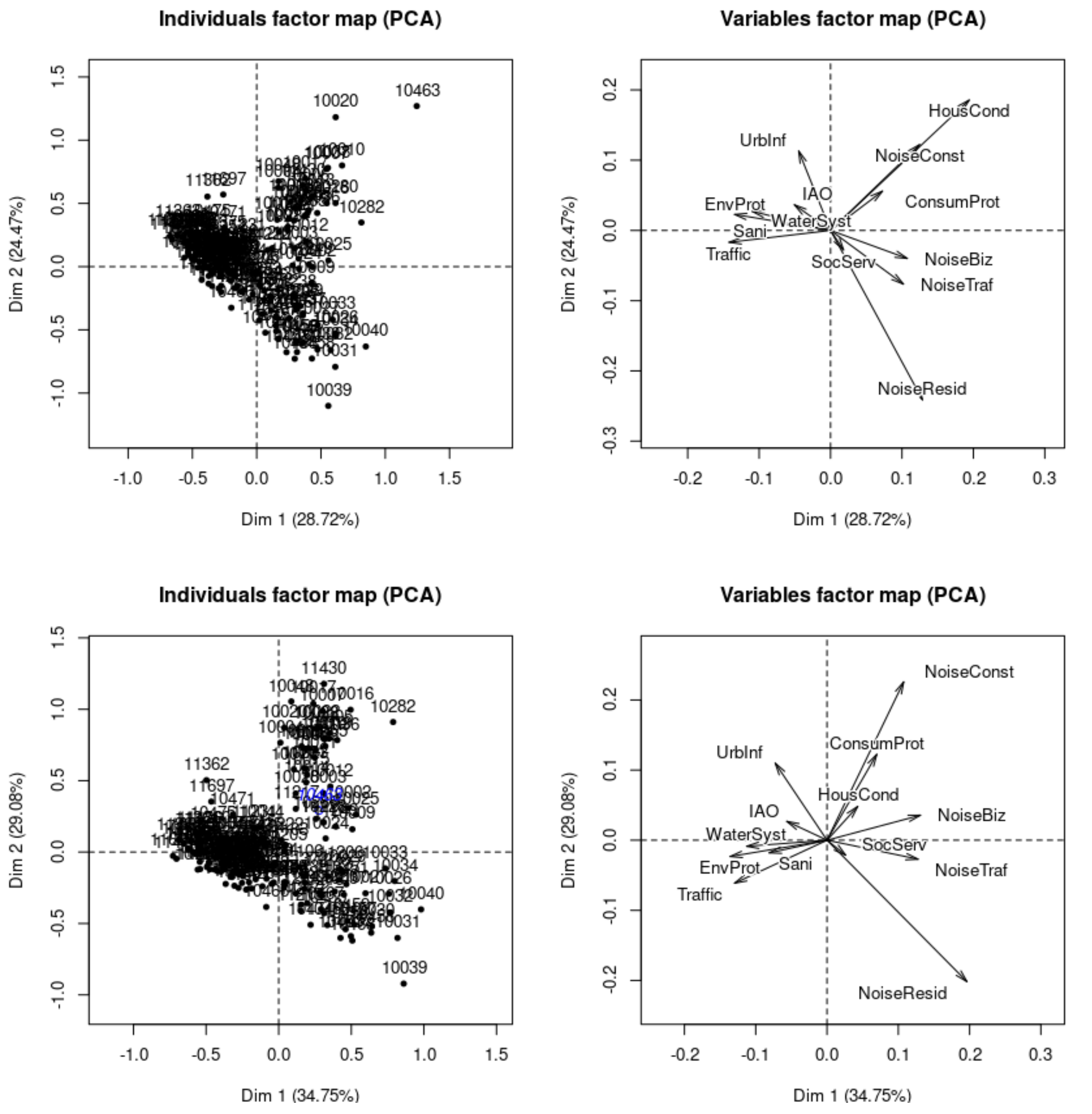


***Fig. 5a**: Weighted observations' and variables' first factorial plane (PC1-2) scatter plots obtained by PCA: (top) including and (bottom) excluding ZIP code "10463". The effect of the χ^2 metric is incorporated in the projections. Bogus ZIP code "99999" is not included and outlier "10463" when not included is a supplementary individual represented in blue on the left plot.*

We observe on the variable factor maps, considering "10463" as supplementary individual, that:

- `NoiseConst` and `ConsumProt` appear to be largely collinear, capturing together more than 32% of the data's dynamics. We decide not to join them however as no satisfactory justification was found to explain the apparent correlation.
- `EnvProt` and `Sani` are consistently collinear and may be merged into a new feature called `EPsani`, capturing more than 8% of the data dynamics.
- `EnvProt` and `WaterSyst` appear strongly correlated in planes PC1-2 and PC1-3, but anti-correlated in PC2-3. `WaterSyst` however exhibits a relatively poor Inertia Explanatory Power (see Table 5 below) with IEP < 5% for the retained significant dimensions. This justifies eliminating `WaterSyst` in an effort to decrease dimensionality.

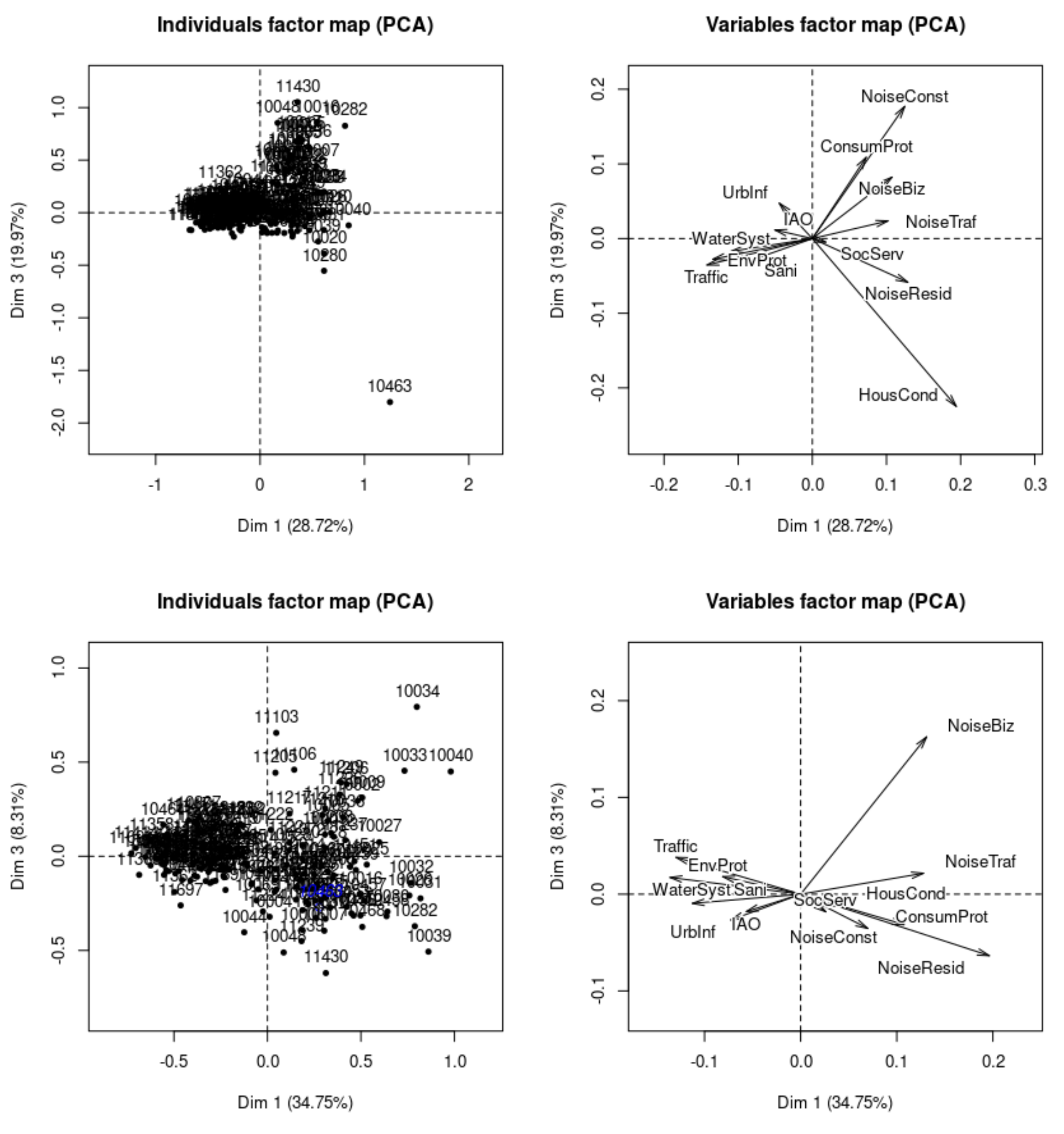


Fig. 5b*: Weighted observations' and variables' second factorial plane (PC1-3) scatter plots obtained by PCA: (top) including and (bottom) excluding ZIP code "10463". The effect of the χ^2 metric is incorporated in the projections.*

- `NoiseConst` and `HousCond` are largely collinear in planes PC1-2 and PC1-3, but anti-correlated in PC2-3. As previously for `WaterSyst`, `HousCond` being a weak variable with IEP < 5% for the retained significant dimensions, justifies doing away with `HousCond`.
- `IAO` and `SocServ` seem to play a negligible role in explaining variance and may be altogether dispensed with.

Among the largest contributors to the construction of the 3 first principal directions, we highlight the fact that `NoiseResid`, `NoiseConst`, `Traffic`, `NoiseTraf`, `ConsumProt` and `EnvProt` are all best represented in the first factorial plane (PC1-2).

By contrast `NoiseBiz` is best represented by the second factorial plane (PC1-3, see figure 5b above), where it plays a dominant role in the construction of the 3rd dimension, PC3.

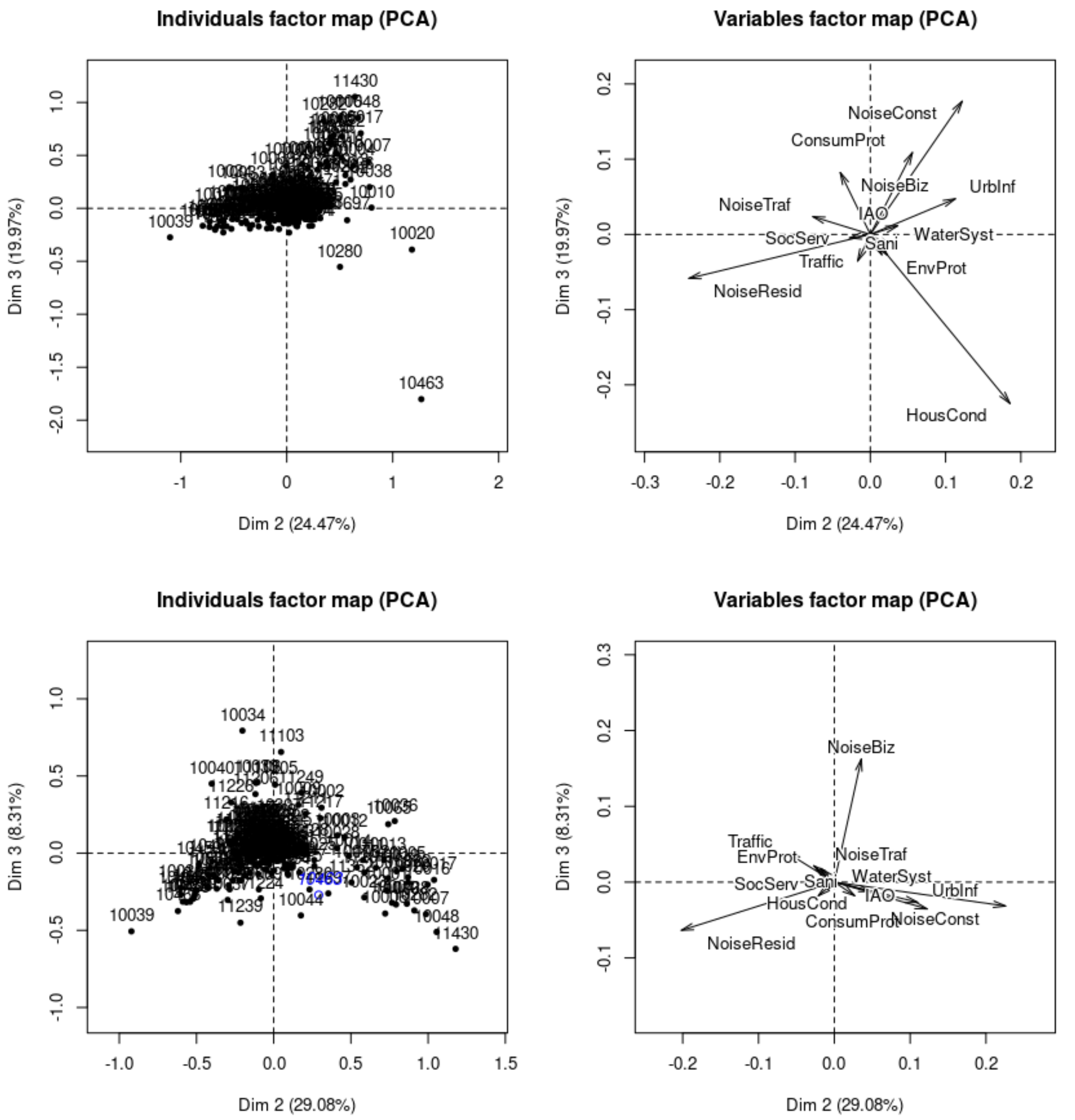


Fig. 5c*: Weighted observations' and variables' third factorial plane (PC2-3) scatter plots obtained by PCA: (top) including and (bottom) excluding ZIP code "10463". The effect of the* χ^2 *metric is incorporated in the projections.*

3-1-2. CA

We conducted a Correspondence Analysis (CA) with row marginals as row profile's weights, thereby incorporating the χ^2 metric effect into the row-profile cloud projection.

Distances <u>*between identically colored points*</u> are distances in the χ^2 sense to correct for the relative scarcity of factors. A red point (column profile) is a barycenter for the blue points (row profiles) expressing that column modality, weighted by said column, and vice versa.

Differently colored points may appear close, but no conclusion can be drawn from that apparent proximity on the graph. On the other hand, identically colored points, which are close together, have similar profiles.

Next Tables 5a and 5b exhibit the inertia explanatory power (IEP) for each SRC's modality, alternately considering all dimensions and only significant dimensions before and after feature selection and dimensionality reduction.

In Fig. 6, **row** (**blue**) and **column** (**red**) profiles are projected together as biplots, after feature selection and extraction, considering ZIP code "10463" as a supplementary observation.

SRCs' modalities	13D-IEP(%)	3D-IEP (%)	SRCs' modalities	8D-IEP(%)	2D-IEP (%)
HousCond	4.0	1.9			
Sani	4.2	2.3			
NoiseResid	19.0	26.0	***NoiseResid***	21.2	27.0
NoiseConst	15.4	19.6	***NoiseConst***	19.3	23.9
NoiseBiz	10.4	13.9	***NoiseBiz***	12.1	6.3
UrbInf	6.0	5.7	***UrbInf***	8.6	8.1
Trafic	8.9	6.9	***Trafic***	11.7	10.1
NoiseTraf	6.3	5.5	***NoiseTraf***	7.2	5.8
WaterSyst	6.2	4.0			
ConsumProt	7.1	6.5	***ConsumProt***	8.8	7.5
SocServ	1.6	0.5			
IAO	3.0	1.3			
EnvProt	7.9	6.0	***Epsani***	11.2	11.3

Table 5a (left) *features factors' inertia explanatory power,* ***before*** *feature selection, over all 13 dimensions and for 3 significant dimensions (shaded cells have IEP>5%).*
Table 5b (right) *shows the same after feature selection and dimensionality reduction.*

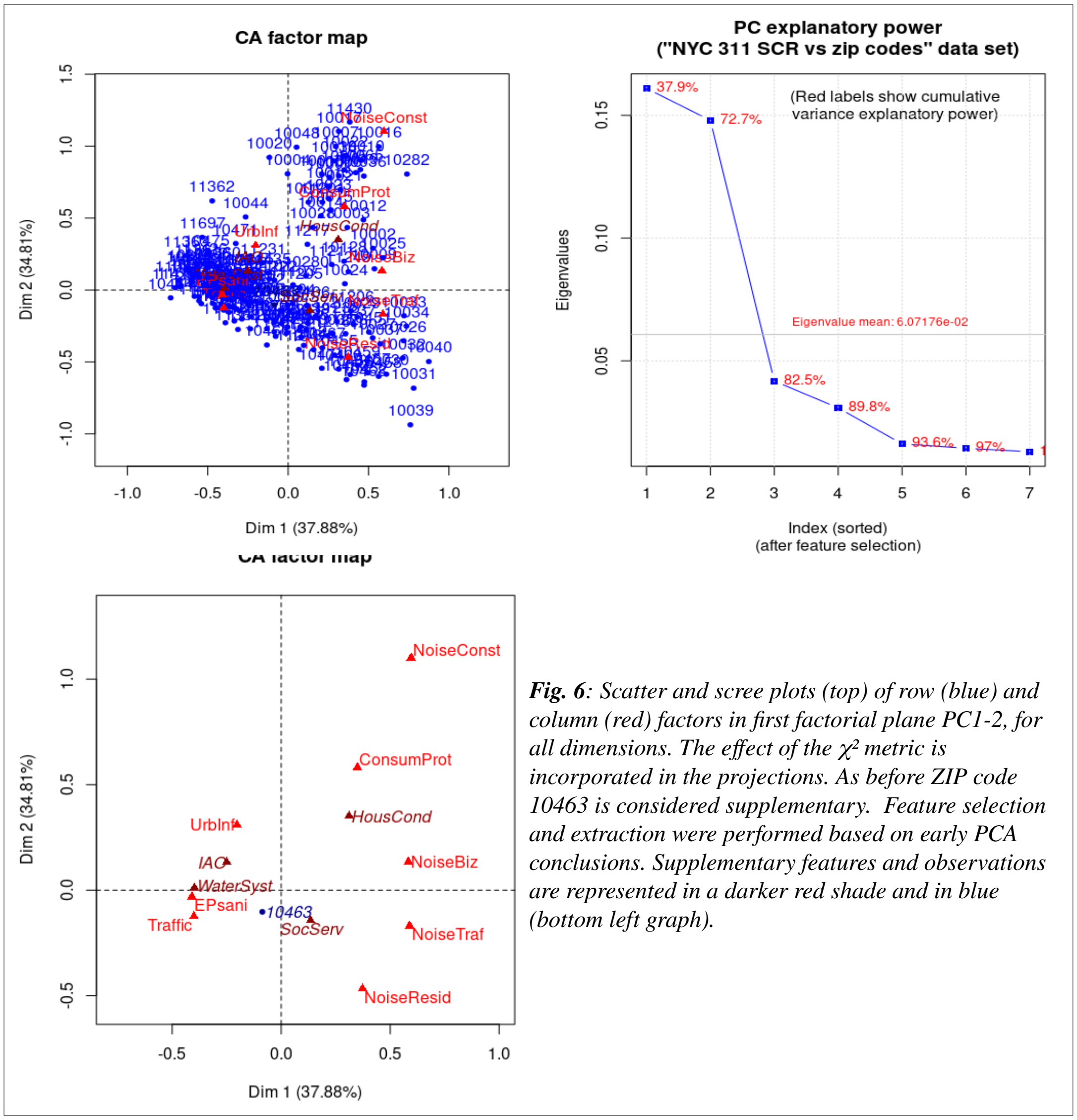


Fig. 6*: Scatter and scree plots (top) of row (blue) and column (red) factors in first factorial plane PC1-2, for all dimensions. The effect of the χ^2 metric is incorporated in the projections. As before ZIP code 10463 is considered supplementary. Feature selection and extraction were performed based on early PCA conclusions. Supplementary features and observations are represented in a darker red shade and in blue (bottom left graph).*

The scree plot reveals 2 significant dimensions with eigenvalues (in decreasing order of inertia representation): 0.16, 0.15 for a total explained variance of almost 73%.

Figures 7 and 8 exhibit variable and individual projections in PC1-2 after feature selection and extraction.

Figure 7 clearly shows that `NoiseResid` and `UrbInf` are anticorrelated. Areas of high incidence for `NoiseResid` SRCs exhibit low incidence of `UrbInf` related calls, as if populations beset by residential noise from neighbors were less prone than others to complain about surrounding urban infrastructure in their areas. The inverse may also hold as we make no hypotheses on causality between the two SRC's modality. Also worthy of note is the fact that `NoiseResid` and `NoiseConst` are very weakly correlated.

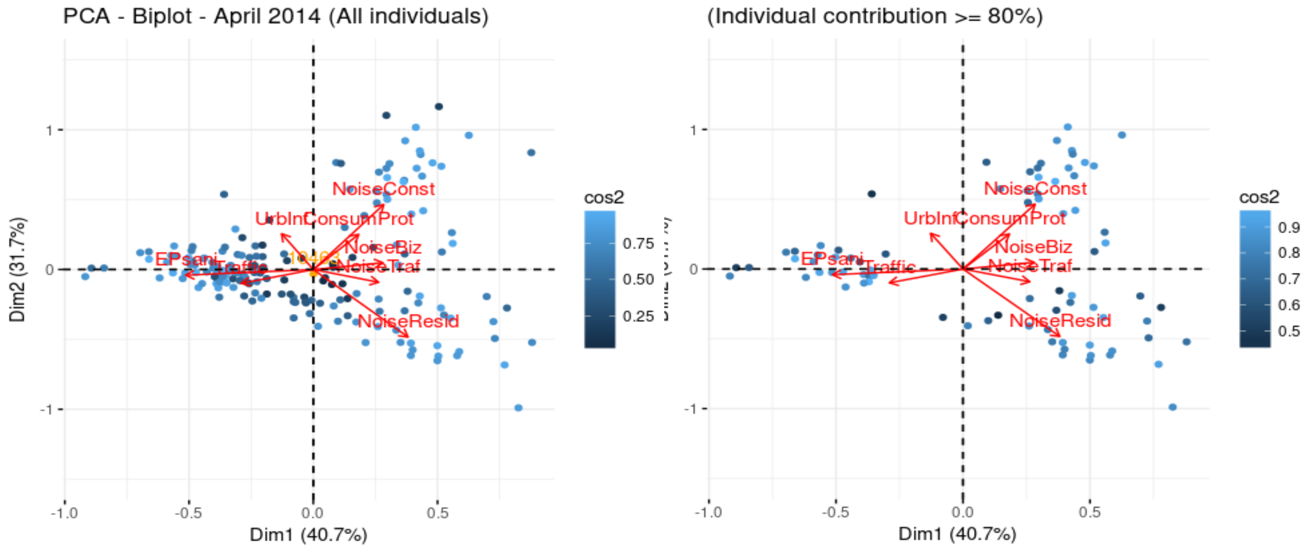


***Fig. 7**: Biplots in the 1st factorial plane PC1-2 for April 2014. The effect of the χ^2 metric is incorporated in the projections. ZIPs codes 99999 and 10463 are included (in gold) and both feature selection and extraction were performed based on previous PCA observations. Both figures show color-coded individuals, according to their quality of representation (cos2). The right plot is restricted to individuals whose contributions to the construction of the first 2 dimension is greater than 80%. Arrows represent direction of growth for features.*

From the borough-based color-coded visualization of scores in Figure 8, one further notes that:

- <u>Manhattan</u>'s make-up (dark blue dots) is heterogeneous appears characterized by `NoiseConst`, `NoiseBiz`, `NoiseTraf`, `NoiseResid`, and `ConsumProt`.
- Most of <u>Queens</u> (gold dots), and part of the Bronx (red dots) are consistent with higher incidences of `EPsani`, `Traffic`, and `UrbInf` related complaints,
- <u>Staten Island</u> (cyan dots) appears fully characterized by a majority of complaints under `Epsani`, and `Traffic`.
- In addition to the above, the <u>Bronx</u> (red dots) is also characterized by `NoiseResid` related complaints,

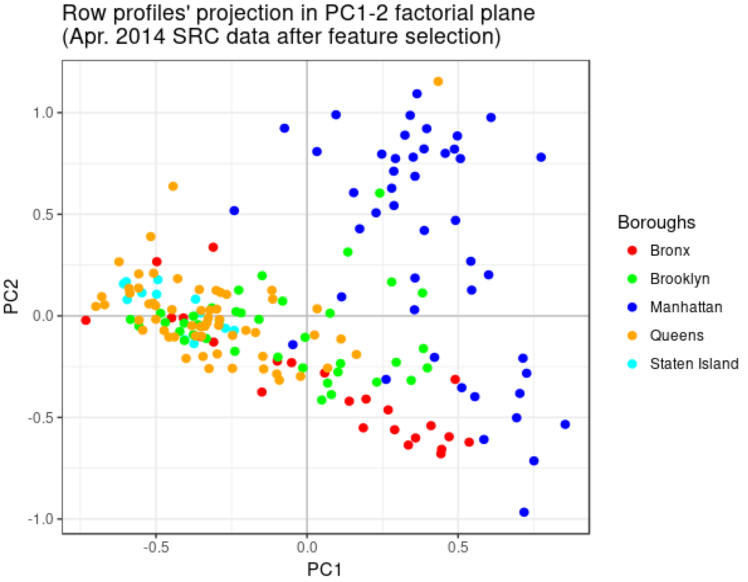


***Fig. 8**: First factorial plane map of individual ZIP codes for the period April 2014, color coded according to the NYC borough to which they belong.*

▪ Brooklyn's ZIP codes projections (green dots) are relatively difficult to interpret as they seems to simultaneously extend in all 4 quadrant, and is therefore representative as a borough of all SRCs' features and type of complaints.

Table 6 (below) summarizes SRC individuals' explanatory power per borough for all dimensions and for only the first factorial plane (i.e. for the 2 significant dimensions).

Borough	**number of ZIP codes**	**IEP all_dim (%)**	**IEP signif dim (%)**
Bronx	24	13.9	10.8
Brooklyn	38	16.2	9.8
Manhattan	46	46.2	37.1
Queens	59	17.2	9.3
Staten Isl.	12	6.0	4.4

***Table 6**: Inertia explanatory power by individual ZIP codes grouped by borough, computed over all dimensions (3rd column) and over the significant dimensions (4th column).*

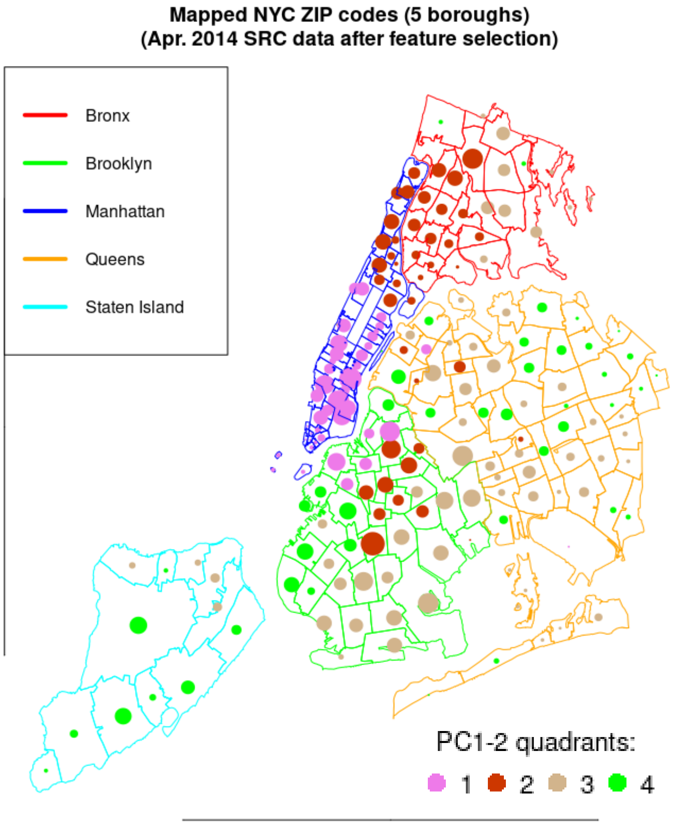

Mapped NYC ZIP codes (5 boroughs)
(Apr. 2014 SRC data after feature selection)
Bronx
Brooklyn
Manhattan
Queens
Staten Island
PC1-2 quadrants:
● 1 ● 2 ● 3 ● 4

***Figure 9**: Topographical representation of ZIP codes' projection quadrant in the first factorial plane (per Fig. 8) for the period April 2014. Dot sizes are proportional to the number of SRCs in a given ZIP code area*

The PCA based initial exploration of NYC's SRCs is almost concluded with a topographical map (Figure 9, right) of row individuals (i.e. ZIP codes), color-coded according to the position of their projection in the PC1-2 factorial plane, following Figure 7. Dot colors represent row profiles' (i.e. individual ZIPs') projections in the four PC1-2 quadrants:

1st quadrant (**orchid**),
2nd quadrant (**green**),
3rd quadrant (**tan**),
4th quadrant (**red**).

As previously noted Brooklyn covers the complete range of SCRs modalities, as shown by the fact that the borough contains dots of all four colors.

3-1-3. Varimax applied to PCA / CA results – Latent factor analysis

R's *varimax()* method (1) to find latent concepts, is a simple factor-oriented (i.e. *column-oriented*) structure rotation designed to maximize the sum of column-wise variances of the squared loadings, that is, the squared correlations between variables and factors. The approach aims at interpreting results in the reduced sup-space of the (in our case) two rotated significant directions. It does not *generally* preserve basis orthogonality, but may do approximately so for simple data structures. In such cases it brings further insight as demonstrated by the latent factor interpretation qualitatively subsumed and shown in red bold face type on Figure 10. In a nutshell, and bearing in mind the fact that NoiseBiz is poorly represented in the PC1-2 factorial plane, the newly rotated factors' projection shows that:

- Many modalities play a role in the construction of varimax-PC2. We observe that `Traffic` and `NoiseConst` are two *pure* and anti-correlated factors in varimax-Dim.2, quasi-absent from the construction of varimax-PC1

- Except for `Traffic` and `NoiseConst`, all other factors also play a role in the construction of varimax-PC1.

- The recurring SRCs, in particular in the borough of Manhattan, about construction noise (`NoiseConst`), appears to displace or be displaced by other noise related complaints to varying degrees and by SRCs about `Traffic` nuisance (outside traffic noise). In other words where construction noise related SRCs increase, all other complaint modalities tend to decrease and reciprocally, to varying extents, except for environmental protection and sanitation SRCs (`EPsani`), and for urban infrastructure (`UrbInf`) SRCs.

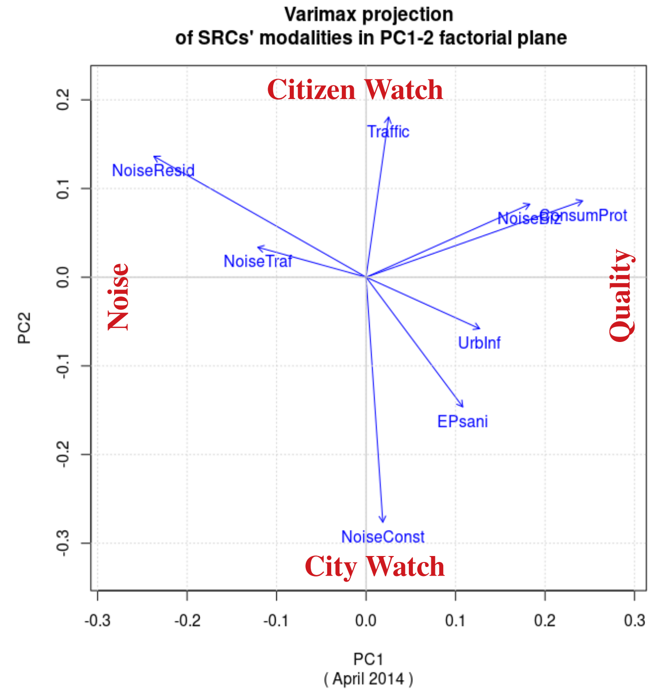


***Fig. 10**: Maximized significance of projected variables in the rotated first factorial plane PC1-2 (using the varimax method).*

Varimax-Dim.1:

That dimension reveals two tendencies among NYC dwellers and their ZIP code areas. Those most sensitized to noise either caused by car traffic during the day, or by residents at night. That group seems to report grievances under `NoiseTraf` and `NoiseResid` either with no correlation or anti-correlated with other SRC modalities. We call them the "**noise protesters**".

Opposite on Figure 10 are areas, where citizens tend to report substandard urban conditions or services in a way apparently anti-correlated with the perception by others of noise pollution. We dubbed members of this group the "**quality seekers**".

Varimax-Dim.2:

That dimension is consistent with NYC areas where inhabitants are primarily concerned by different form of urban pollution, such as: noise cause construction work, urban sanitation, environmental issues as well as an insufficiently well-maintained urban infrastructure. We dub this group: "**city watch**".

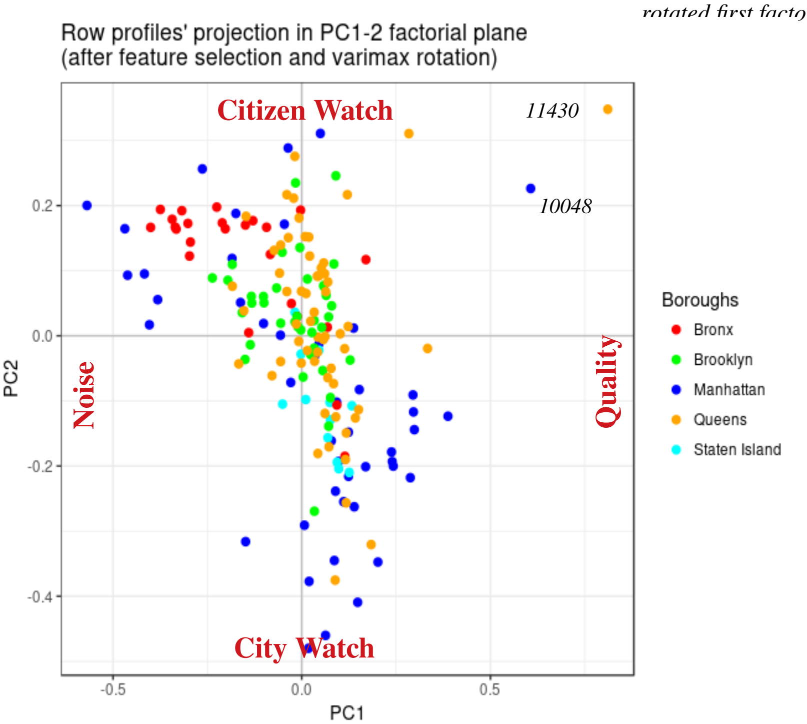


***Figure 11:** Orthogonal projections of individuals (ZIP code observations) onto the varimax-rotated loading directions, color-coded per borough and after feature selection.*

Opposite the "city watch" group, on Figure 10, are ZIP code areas, where citizens are more concerned about noise in their proximity and not caused by construction work, as well as consumer protection. Those people are also more likely to report parking violations than construction noise and are generally more sensitive to uncivil behavior as well as other nuisances directly caused by others. We dub this group: "**citizen watch**".

Two figures complete the presentation of our correspondence analysis for April 2014 SRCs data. Figure 11 revisits Figure 8, the projection of scores on the first factorial plane (PC1-2), after varimax-rotation of the loadings. Figure 12 further complements that by offering a topographically situated, color-coded representation of ZIP codes' quadrants following Figure 11.

From Fig.11 it is easy to observe that:

- 10048 (the Manhattan vanity ZIP code for the former World-Trade twin towers) and 11430 for Jamaica in Queens, resemble outliers due to very few SRCs originating in them, besides SRCs in the category "ConsumProt".
- Manhattan exhibits influences between "*City Watcher*" and a combination of "*Noise*" sensitized areas and "*Citizen Watch*"
- The Bronx is clearly dominated by residential noise related SRCs, in the graphical sector between "*Citizen Watch*" and "*Noise*".
- Brooklyn, Queens and Staten Island are dominated by varimax-rotated PC2, i.e. along latent factor axis defined by "*Citizen Watch*" and "*City Watch*".

The method used for obtain the above varimax-rotated scores, making possible the post-varimax visualization of scores (i.e. row-profiles projection along PC axes) requires a brief explanation (2). In varimax, loadings (i.e. eigenvectors scaled by the square roots of their respective eigenvalues) are rotated. In other words, eigenvectors obtained from the covariance matrix on scaled observations are not directly rotated. In fact, rigorously speaking, varimax rotation does not generally produce orthogonal loading vectors (even though the varimax rotation is often referred to as an orthogonal transformation). The upshot is that the orthogonal projections of individuals onto the rotated loading directions, that is the varimax rotated scores, cannot be computed in a straightforward way. To find them, one can use varimax-rotated loadings, multiplying the scaled (i.e. in our case merely centered) data by the transposed pseudo-inverse of the rotated loadings.

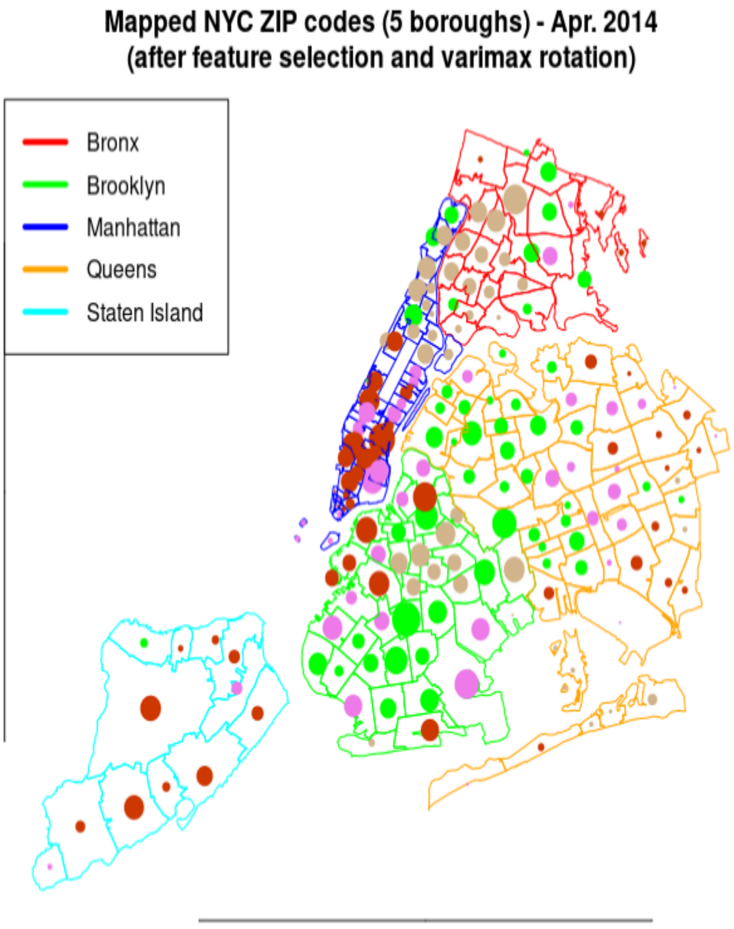

Mapped NYC ZIP codes (5 boroughs) - Apr. 2014
(after feature selection and varimax rotation)
Bronx
Brooklyn
Manhattan
Queens
Staten Island

Figure 12: *Topographical representation of the latent factors at play in NYC's five borough area. As before in Figure 9, boroughs' ZIP code areas are drawn following the color coded legend provided in the upper-left corner of the map.*

Finally Figure 12 below illustrates the topographical representation of SRCs according to the previous latent factor analysis, i.e. after varimax rotation. Proposed latent factors (in the sense of Figure 11) found dominant for each ZIP code are shown by means of color coded solid dots. Dots' diameters are proportional to the number of total SRCs originating in the ZIP code for that period. Each latent factors' sector of dominance is defined as the positive or negative varimax-rotated PC directions ± 45º. They correspond to 90º cones, whose apices coincide with the projected cloud's centroid on the first factorial plane and whose axes of symmetry are the positive or negative rotated PC directions.

Row profiles' (i.e. individual ZIPs') projections, i.e. scores, in the varimax-rotated PC1-2 plane are color-coded according to the latent factor's cone they fall into:

- PC1+ cone	*(**orchid**)*	*QUALITY*
- PC2+ cone	*(**green**)*	*CITIZEN WATCH*
- PC1- cone	*(**tan**)*	*NOISE*
- PC2- cone	*(**red**)*	*CITY WATCH*

- As previously noted Brooklyn covers the complete range of SCRs modalities, as shown by the fact that the borough contains dots of all four colors.

- The Bronx (unsurprisingly at this point) is dominated by the 2 latent factors "*NOISE*" and "*CITIZEN WATCH*".

- Manhatttan does too to a lesser extent, but is clearly divided between down and midtown on one hand and uptown on the other hand. Down- and midtown are areas where dominant factors are "*QUALITY"* and "*CITY WATCH"* while uptown is clearly dominated by "*NOISE*" and "*CITIZEN WATCH*". This seems to reflect changes as much in residents' concerns and perception, as in the individual behaviors at the origin of SRCs.

- Queens is geographically divided between two wide areas: the west side, facing Manhattan and bordering Brooklyn and the east side facing the ocean and bordering Nassau county. The first one is characterized by the "*CITIZEN WATCH*" factor, while the second seems more focused on concerns about "*QUALITY*".

- Finally the population of Staten Island, as before, demontrates its focus on urban conditions, as captured by the latent factor "*CITY WATCH*".

3-2. Multiple Correspondence Analysis (MCA)

3-2-1. Discretization of data

To supplement our previous CA on SRCs to NYC 311 per location, we now add NYPD crime data (calls to 911) in the form of 3 modalities in increasing degree of gravity: violations (4,699 counts), misdemeanors (21,734 counts) and felonies (11,156 counts). Those events were recorded during the sole month of April 2014 and were distributed over 181 zip codes and 5 boroughs. In order to conduct MCA we discretized our multivariate contingency table so that every modality (column) is now expressed in the form of ordinal values related to 4 buckets (bins) of similar sizes.

Discretization makes losing some information unavoidable. However it also allows us to extract 2 way contingency tables involving crime modalities and NYC boroughs. How we went from frequencies (counts) to ordinal variables is shown next for the sample consisting of NYPD's records of 21,734 misdemeanors in April 2014. Sample quartiles corresponding to the distribution of counts per ZIP code were:

```
Min.  1st Qu. Median   Mean   3rd Qu.  Max
0.0     32.0    87.0   120.1   179.0   705.0
```

Based on quartiles, the chosen bucket intervals were: < 33 – 87 – 178 – >178. Corresponding ordinal variable values are summarized in tabular form below. They may differ for non crime related variables (SRCs) where ordinal values may refer to a different count scale. This however is not detrimental to the correct analysis.

Bin upper bound	2~3	6~16	20~33	~38	85~91	150~180	> 180
Ordinal variable value	VL	ML	M	MH	H	VH	OC
Interpretation	Very low	Medium low	Medium	Medium high	High	Very high	"Out of Control"

An exception is made for the treatment of the SRC categorical variable "HousCond". Its count spread is such over several different time periods of interest that rather than setting a fixed bin scale we just report quartile intervals for each time period.

3-2-2. Analysis of crime segmentation across NYC boroughs

The normalized crime segmentation per borough, for each crime modality is shown next, in Figure 13.

For each crime modality, Fig. 13 is interpreted in terms of the relative proportions of ZIP code areas in each borough belonging to a low, medium, high or very high crime count bucket. For instance, for the 4860 misdemeanors committed in Manhattan ZIP code areas in April 2104:

- 30% of ZIP code exhibited a medium (M – cyan) crime count, (14/46)
- 21% of ZIP codes exhibited a high (H – orange) crime count, (10/46)
- 26% of ZIP codes exhibited a very high (VH – red) crime count, (12/46)
- 21% belong to the bucket of extremely large crime counts, dubbed "out of control" (OC – dark red), (10/46)

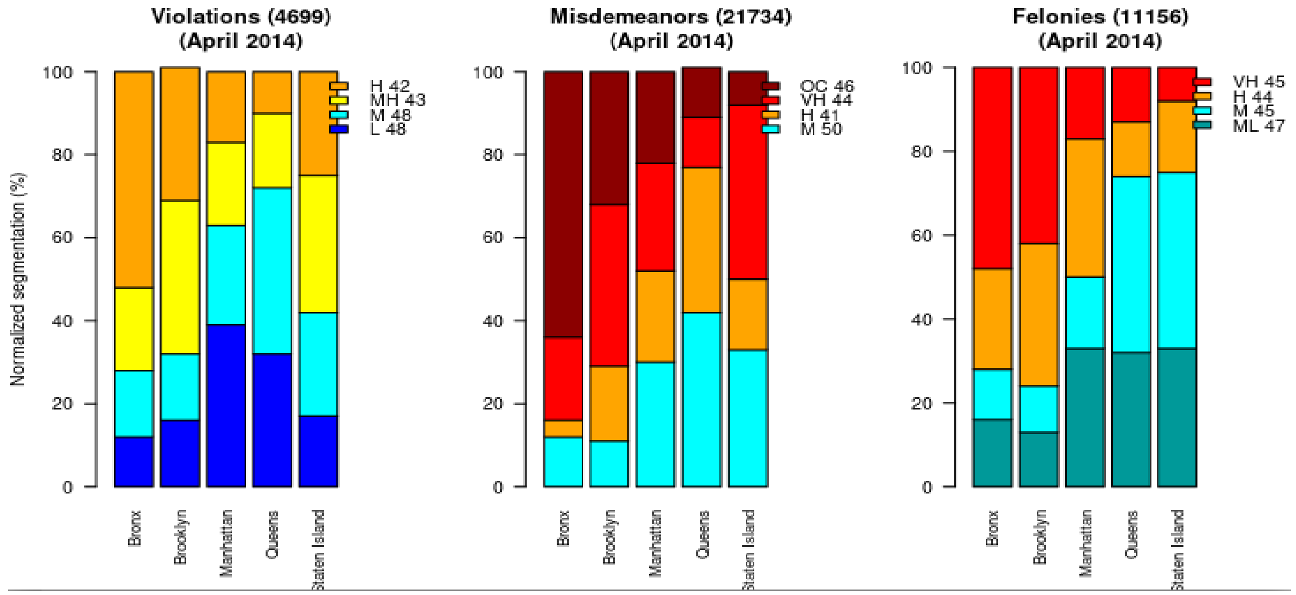


***Fig. 13**: Borough crime index shown as normalized (to 100) segmentation for each crime modality as a function of borough. Legends show color-coded ordinal values followed by the bucket size (i.e. number of observed ZIP code areas) across boroughs.*

Figure 13 does not inform us on the number of crime committed in each borough, but rather on the distribution of hot "crime spots" within each borough. The Bronx clearly shows a tendency to concentrate high crime areas, across all crime modalities, when compared to other boroughs. It is followed in that by Brooklyn, Manhattan, Queens and Staten Island, in the cases of misdemeanors and felonies.

3-2-3. MCA

We extend our previous Correspondence Analysis (CA) results to include:

- the categorical variable `Crime` Calls to NYPD 911) whose three modalities are described earlier, in Section 3-2-2,
- the two quantitative variables: `medianInc` (median income) and `jlBenef` (jobless benefit).

The resulting Multiple Correspondence Analysis is based on the indicator matrix method. It specifies:

- rows "99999" (bogus ZIP code), "11430" (JFK airport, Queens), "10463" (Riverdale, the Bronx) as supplementary individuals, and
- columns `medianInc` and `jlBenef` as quantitative supplementary variables.

Figure 14 shows how all crime modalities ("`Violation`", "`Misdemeanor`" and "`Felony`") are particularly correlated with the (almost super-imposed) SRCs modalities "`HousCond`" and "`Traffic`", indicating that areas where housing conditions are poor and traffic violations reported by inhabitants are numerous also have a higher crime incidence in all three crime modalities. On the figure, the dashed gray line represents the direction of crime growth.

To note the two first PCs account in MCA for far smaller fraction of system inertia than their counterparts in CA or PCA (namely 23% vs. 73%). It is a normal consequence of the multiplicate increase in dimensionality used to carry out our Multiple Correspondence Analysis.

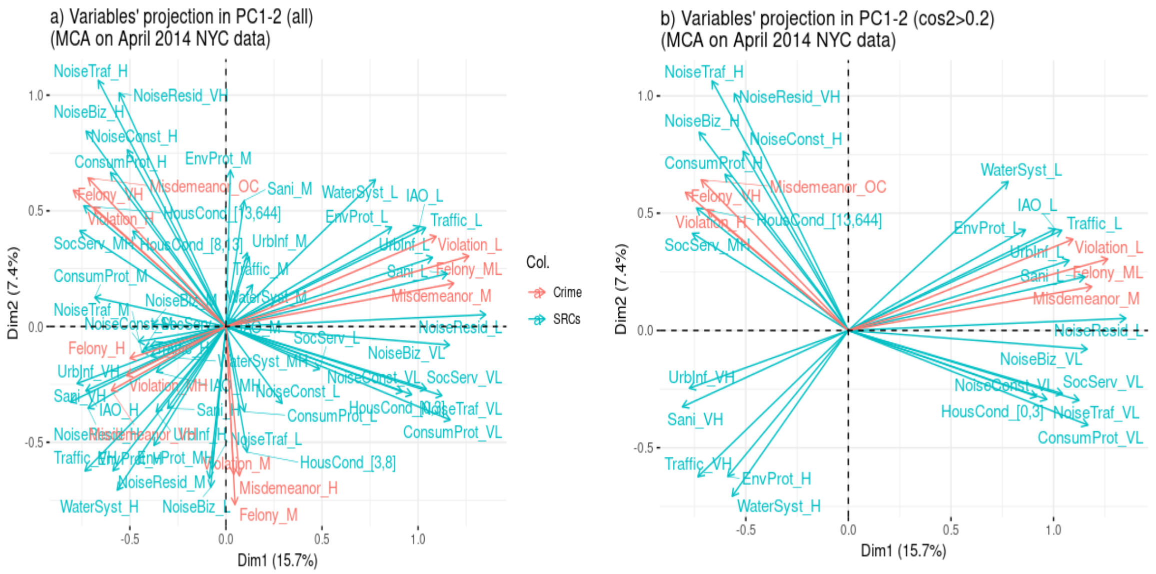


Figure 14: *MCA based representations in the 1st factorial plane (April 2014 NYC data) of variables' modalities' levels (categorical SRCs in turquoise, NYPD crime in red), for **a)** all variables and **b)** modality levels with quality of representation better than 20%.*

Figure 15 exhibits row profiles', i.e. individuals' projection in the 1st factorial plane. The quality of representation and the contribution to the construction of PC axes are generally seen as poor compared to PCA and CA results due to the inherently higher dimensionality of the MCA technique. The centroids of individuals belonging to a borough are indicated by a larger diameter colored dot. Interpretation of their relative positions in the 1st factorial plane is related to that of PC1 and PC2.

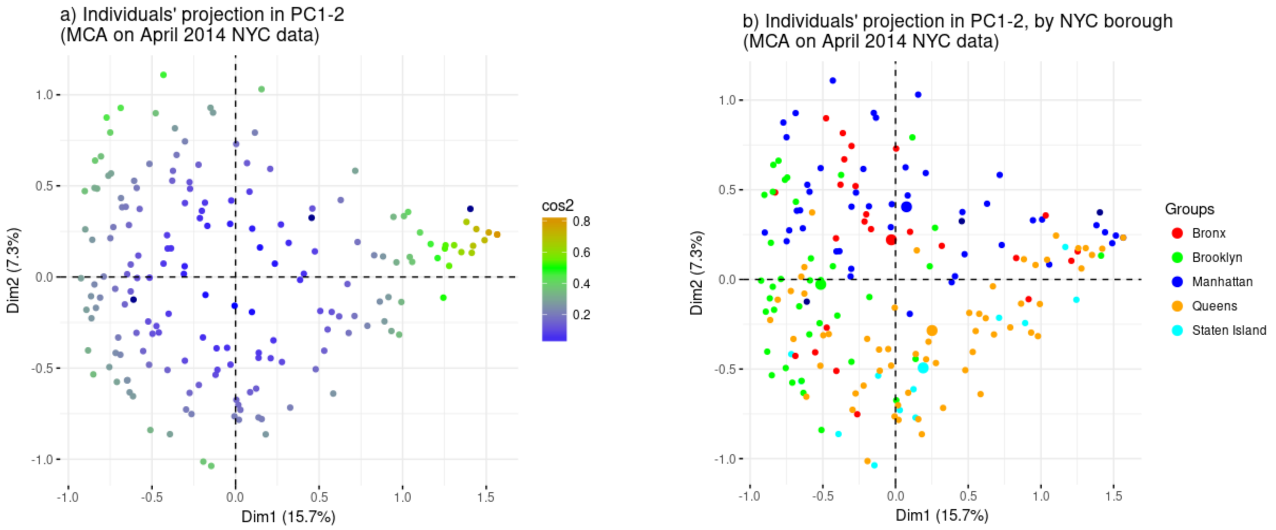


Figure 15: *PC1-2 projection for Abril 2014 of individuals (row profiles): **a)** color-coded according to quality of representation (cos2), **b)** color-coded according to NYC borough.*

Figure 16 below exhibits individual projections in the first factorial plane (PC1-2) along with crime levels per crime modality (***felony***, ***misdemeanor***, and ***violation***). Modalities' levels are represented by abbreviations as denoted before (Section 3-2-2): low (L), medium-low (ML), medium (M), medium-high (MH), high (H), very high (VH) and out-of-control (OC) crime counts, the choice of terminology being completely arbitrary on the analyst's part. It is only meant to cover the whole scale of reported crimes counts in every category during the month of April 2014. No matter the modality of crime, its rate increases clock wise, with Quadrant 1 containing the lowest crime rate observations and Quadrant 2 the highest.

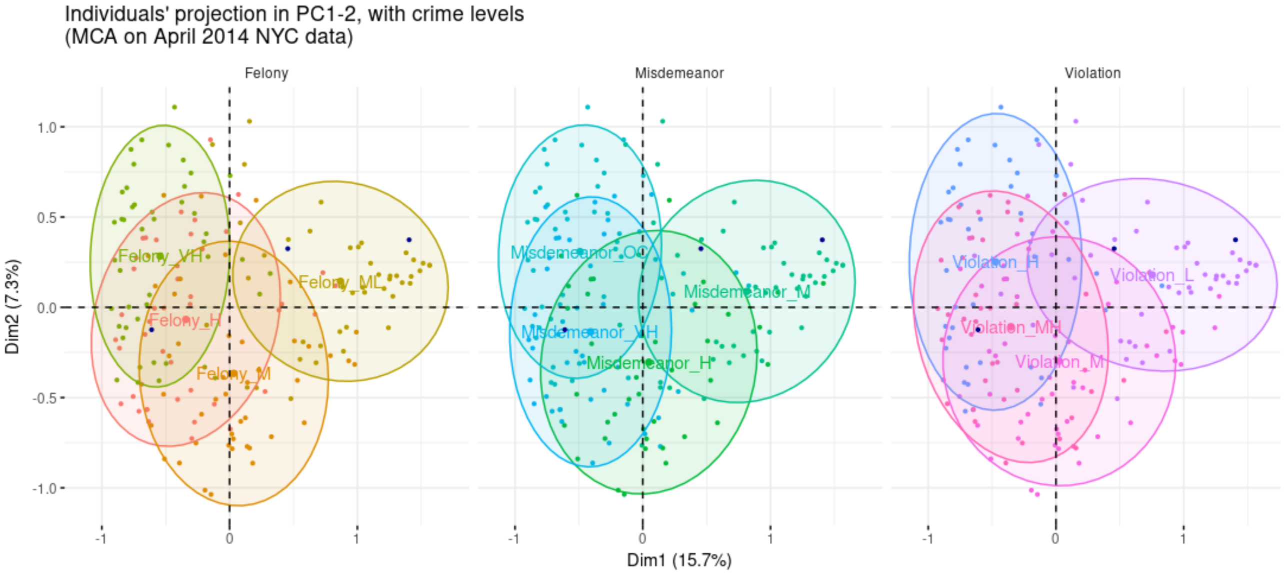


Figure 16: *PC1-2 projections (April 2014 NYC data) of individuals, color-coded according to their crime rates' modalities' levels. Ellipses are drawn for 75% confidence intervals in the χ^2 sense, assuming standardized normal distributions of observations for every crime modality's level.*

Figures 15 and 16 confirm and enrich a number of early conclusions drawn from PCA and CA.

▪ Manhattan consists of two parts (1st and 2nd quadrants of Fig. 15b): downtown and midtown Manhattan (dark blue dots), characterized by relatively low frequencies of SRCs (across all modalities of SRCs) and by the statistically lowest crime rate in NYC.

▪ Meanwhile uptown Manhattan borders the Bronx (red dots), and shares many traits with it in all crime modalities and in many SRCs. From the view point of urban planning it is a transition area between very different neighborhoods of the 5 borough metropolitan area. Going from south to north, Manhattan transitions from low to very low frequency SRCs neighborhoods to areas where complaints related to poor public housing conditions, noise (in particular but not only residential noise), traffic nuisance and reported occurrences of crimes are at their statistical highest.

▪ The 2nd Quadrant of Figures 15 and 16 covers mainly uptown Manhattan, South and West Bronx, Central Brooklyn as well as a few isolated ZIP codes belonging to Queens, for a total 51 ZIP codes out of 177. Those are the most violent areas in NYC in April 2014 , with:

- a ***violation*** sum-total of 1691 representing 36% of all reported violations in the NYC area
- a ***misdemeanors*** sum-total of 8943 representing 40% of all reported misdemeanors.
- a ***felony*** sum-total of 4294 representing 37% of all reported felonies

Those urban areas are perceived by callers to NYC-311 as being in poor keep and as the place of uncivil or disorderly behaviors to boot.

▪ 3rd and 4th quadrants are intermediate ones between highest and lowest crime rates, also between highest and lowest incidences of SRCs. The 3rd quadrant corresponds to West and East Brooklyn, a large swath of Queens plus the North-East part of the Bronx. Meanwhile the 4th quadrant reflect mainly the rest of Queens and Staten Island.

Figures 14a, 15 and 16 are combined in the form of a biplot in Figure 17, where the usual color code is used to identify the borough of each plotted individual ZIP code, and crimes' modalities' levels are indicated in purple.

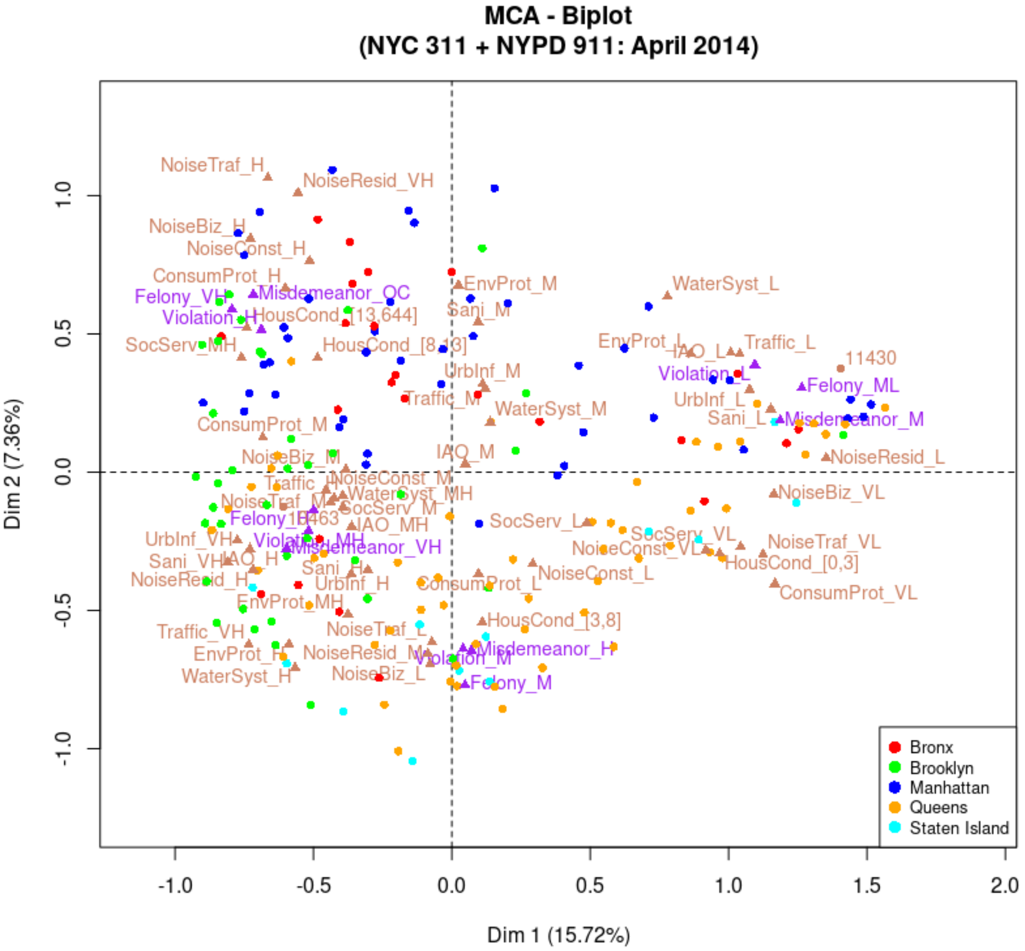


Figure 17: *MCA based biplot representations in PC1-2 and for Abril 2014 of categorical variables and individuals, color-coded according to their borough. The brown dots are the 3 supplementary individuals "99999" (bogus ZIP), "10463" (Riverdale, Bronx) and "11430" (JFK airport, Jamaica, Queens).*

We define Figure 17's quadrant associations loosely and as follows:

- **1st Quadrant**:

- very low to medium counts of SRC's modalities "`EnvProt`", "`WaterSyst`", "`Traffic`", `UrbInf`", "`IAO`", "`Sani`" and a high count of "`ConsumProt`" SRCs.

- low incidence of reported `violations`,
- medium incidence of reported `misdemeanors`,
- medium-low incidence of `felonies`

The 1st quadrant is representative of urban pockets across the five boroughs, associated with the latent factor ***Quality***. It shows that a fairly appeased crime scene correlates well with a much reduced frequency of calls to NYC's 311.

- **2nd Quadrant**:

- medium to very high counts of SRC's modalities "`NoiseResid`", "`NoiseConst`", "`NoiseTraf`", "`NoiseBiz`", "`HousCond`", "`Traffic`", "`ConsumProt`", "`SocServ`".

- high incidence of reported `violations`,
- out-of-control reported `misdemeanors`, i.e. an incidence rate so high as to dwarf other areas in NYC.
- very high incidence of `felonies.`

The 2nd quadrant characterizes mid and uptown Manhattan, most of the Bronx and central Brooklyn, by far areas with the highest crime rate among the five boroughs, and associated with the latent factors ***Noise*** and ***City Watch***.

▪ **3rd Quadrant**:

- medium to high counts of SRC's modalities "*NoiseResid*", "*WaterSyst*", "*EnvProt*", "*ConsumProt*", "*IAO*"
- high to very high counts of SRCs' modalities "*Sani*", "*UrbInf*", "*Traffic*", "*Sani*"

- medium-high incidence of reported *violations*,
- very high incidence of reported *misdemeanors*,
- high incidence of reported *felonies*

The 3rd quadrant concerns small pockets in the Bronx and a significant part of Brooklyn as well as four Staten Island's ZIP codes and a sizable area of Queens. It appears associated with the latent factors ***City Watch*** as far as Staten Island ZIP codes are concerned,and with ***Citizen Watch*** for ZIP codes associated with either the Bronx or Brooklyn.

▪ **4th Quadrant**:

- very low to low counts of SRC's modalities "*NoiseResid*", "*HousCond*", "*NoiseConst*", "*NoiseTraf*", "*NoiseBiz*", "*ConsumProt*", "*SocServ*".

- medium incidence of reported *violations*,
- high incidence of reported *misdemeanors*,
- medium incidence of *felonies*

The 4th quadrant concerns the rest of Staten Island and most of Queens and appears to be associated with latent factors ***Noise*** and ***Citizen Watch***.

3-3. Clustering analysis

To further explore the underlying stucture(s) in our April 2014 NYC data set, we carry out probabilistic clustering on the row profiles of our previous MCA data matrix, using replicated *k*-means (3) partitioning. Next we deploy agglomerative Hierarchical Clustering, a well known bottom-up grouping method. Finally, we consolidate our crisp clustering results using *k*-means.

Clustering consists in grouping objects or observations in non-overlapping groups, clusters or classes, based on some criterion of proximity, similarity or likeness. Its purpose is to help understanding complex information by reducing its dimensionality. The concept we put to work here is based on spherical cluster classes (multi-dimensional Euclidian proximity or distance), separable in such a way that the mean observables' value in a class converge towards the class' centroid. It ensues that clusters are expected to be of similar size, for the assignment to the nearest cluster class center to be the correct assignment.
Being a Euclidian distance based classification process, *k*-means considers variance of observations, but not covariance between observations and cluster classes thereof. Superseeding this naive approach would be possible in a number of ways (4), e.g. with:

- the *Gaussian Mixture* based on the *expectation-maximization* algorithm, which maintains a probabilistic assignment to cluster classes (Bayesian soft clustering) and a multivariate normal (MVN) distribution instead of the mean;
- the *Partitioning Around k-Medoids* (PAM), a heuristic algorithm reminiscent of *k*-means but which makes use of arbitrary non-Euclidian distances such as the Manhattan (L1) distance, the Jaquart distance, the cosine similarity, etc.

Still, in spite of its shortcomings, k-means does comply with the objectives of our preliminary MVA.

3-3-1. Probabilistic k-means and hierarchical clustering

We first deployed a probabilistic clustering analysis using twenty *k*-means replications and a number of clusters to be ascertained in the range 2~10. Individuals to be clustered are embedded in an Euclidean space defined by the factorial coordinates or "scores" of our observations, derived from the MCA results of Section 3.2.

The standard *k*-means algorithm is a heuristic process. Replication is made necessary by the fact that (a) its result depends on initial conditions, i.e. the choice of the *k* initial centers, and (b) the algorithm does not guarantee a global optimum.

For every experiment consisting of 20 replicas each, we calculate two ratios, and use them as criteria to ascertain the optimal number of cluster classes, allowing for 3 random starts per replication.

- *SSB/SStot*, the between-cluster *sum of squares* or variance, denoted SS_B over the total SS", referred to as the "normalized within-cluster SS criterion", where SS represents inertia (variance) and is calculated relative to the relevant cluster centroid for each computed cluster.

- the *Calinsky-Harabasz index* consisting of the ratio of *between-cluster SS*, and *within-cluster SS*, denoted SS_W, corrected by the number of clusters, k, and observations, n: $\frac{SS_B/(k-1)}{SS_W/(n-k)}$

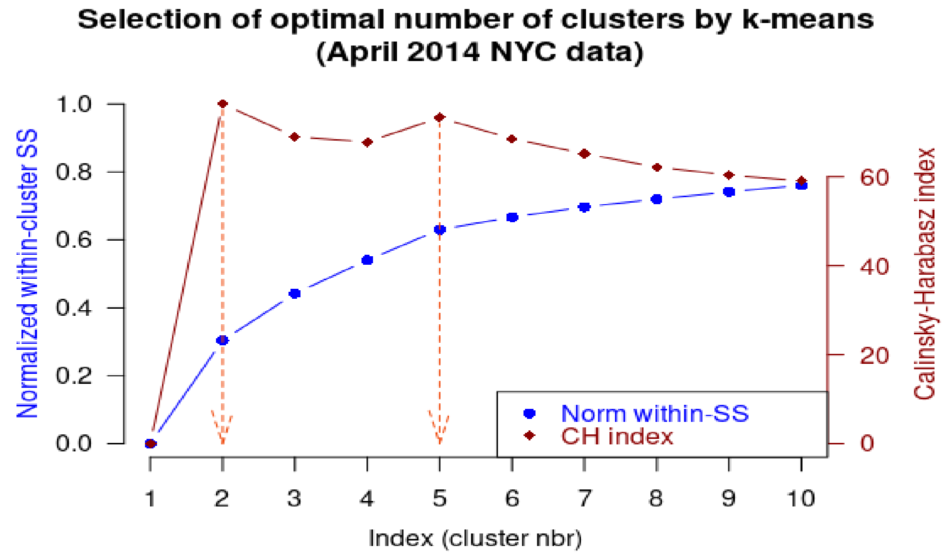


***Figure 18**: Graph of the two index criteria used to qualify our probabilistic search for the optimal number of clusters (from MCA results on April 2014 NYC data).*

Whereas the first ratio (blue line) increases continuously as the number of trial clusters rises, we observe in Figure 18 that the CH index (red line) gives us two local optima for 2 and 5 clusters each, signalled by dashed vertical arrows pointing toward abscissae 2 and 5.

The above result is further qualified by the Cluster S*ilhouette* method, the which given k clusters and n individuals *"i"* computes:

- *a(i)* the distance of i to all individuals of the same cluster class
- *b(i)* the lowest distance of i to all individuals of any other cluster class

and finally the ratio $s(i) = \frac{b(i)-a(i)}{max[a(i),b(i)]} \in [-1,1]$

where *s(i) > 0*, *s(i) = 0 or s(i) < 0* when *i* is correctly allocated, close to the decision boundary or allocated to the wrong cluster respectively. Average[*s(i)*] over the whole clustered data set is a measure of clustering quality as exemplified in Figure 19 below for 2 and 5 clusters.

The 2 cluster silhouette shows an imbalance in the numbers of allocated individuals between the 2 clusters. It also sports a large number of erroneously classified individuals (ZIP codes). Its average cluster silhouette width is: 0.29

The 5 cluster silhouette contrasts in that its shows a well balanced distribution of observations among clusters and very few allocation errors. This contributes to the better *average cluster silhouette width* of 0.31.

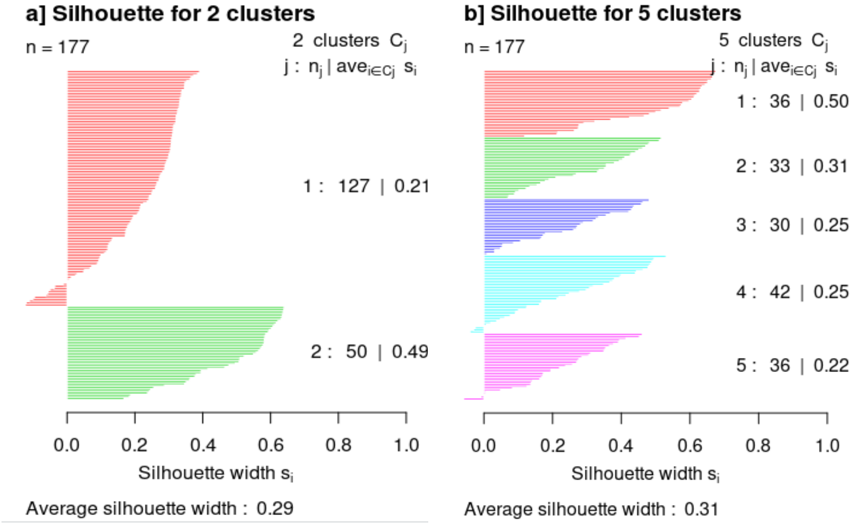


***Figure 19**: Cluster silhouette for a) 2 and b) 5 cluster classes, out of a population of 177 individual observations for April 2014 NYC (SRCs + crime) data.*

We further substantiate the preliminary finding of 5 cluster classes, rather than just 2 (trivial result), by representing the dendrogram built from hierarchical clustering, in Figure 20. Conceptually, considering ZIP code level agglomeration is consistent with nested hierarchies in Ward-similarity[xi] based HC.

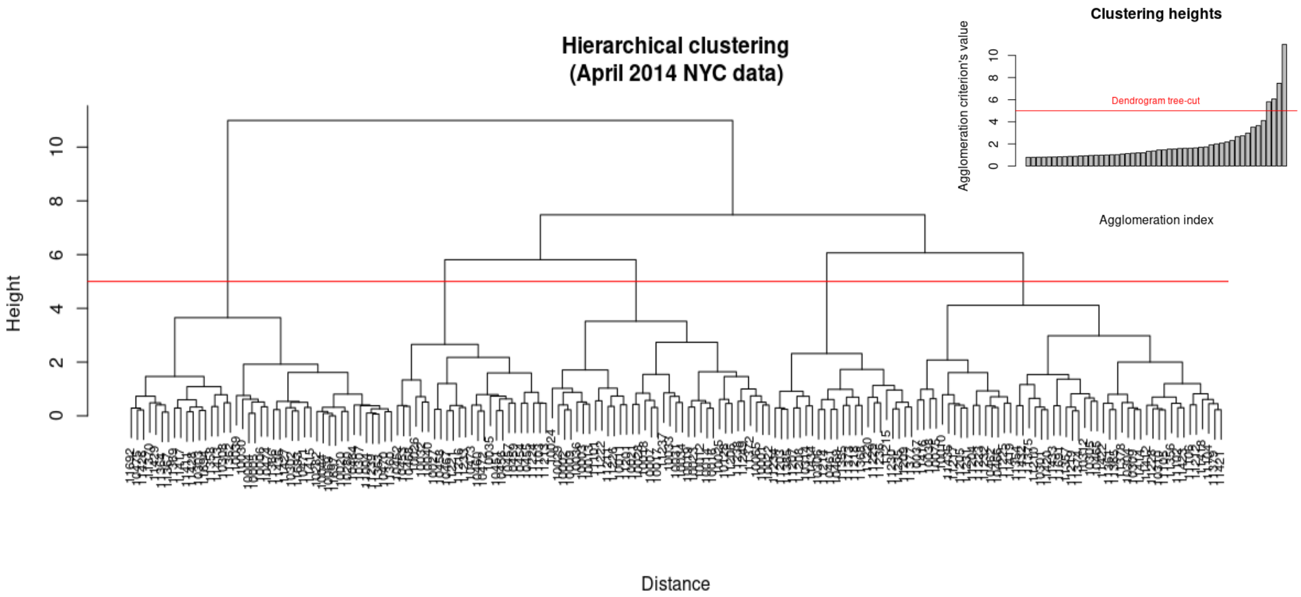


***Figure 20**: HC dendrogram based on a Ward-2 distance matrix deterministically calculated from MCA scores. The most appropriate tree pruning corresponds to 5 clusters, per the horizontal red line. The inset (top-right) offers another graphical view of the agglomeration's criterion's values for each one of the 50 last merge-operations.*

Figure 21 hereafter exhibits a PC1-2 factorial plane projection of observed ZIP codes according to cluster and to borough. Appendix D contains the same plot with fully labelled observations.

[xi] *For the Ward's method, the proximity between two clusters (or two cluster-classes) is defined as the increase in the squared error that results when two clusters are merged.*

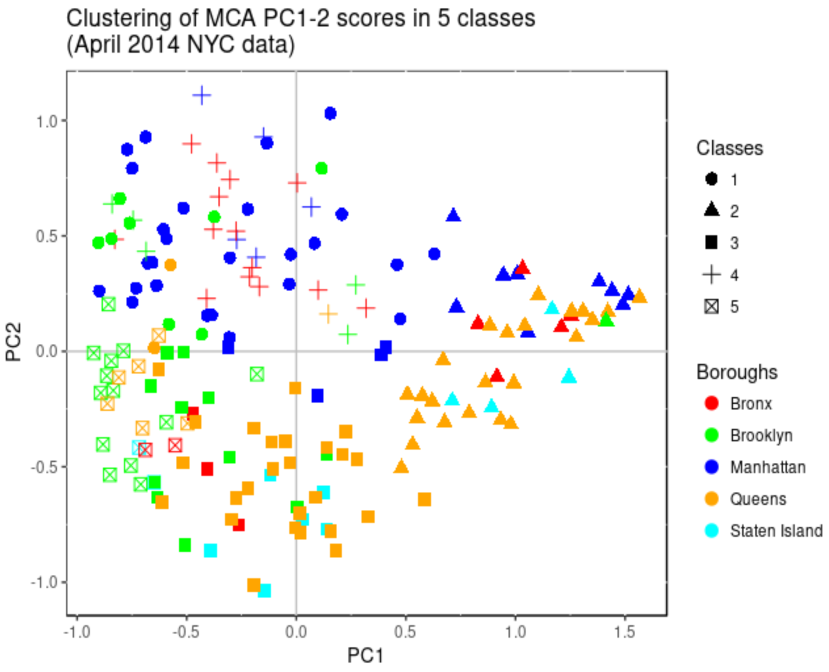


***Figure 21**: Grap of scores in the first factorial plane, where each ZIP code area is shown to belong to a specific cluster class, identified by a distinct symbol. Observations are color-coded following Figs. 11, 15b, 17 and others.*

Boroughs are not uniformly clustered.

The Bronx (red symbols) is shown to belong to 3 distinct clusters (3, 4 and 5).

Brooklyn (green symbols) ZIPs are primarily distributed in clusters 1,3 and 5.

Manhattan ZIP code areas (dark blue symbols) are mainly seen in cluster 1 and 2 with very few ZIP observations classified in 3 and 4.

Queens (orange symbols) appears in clusters 2,3 and 5.

Staten Island ZIP codes (cyan symbols) are divided between cluster 2 and 3.

Before completing this analytical sequence for the period April 2014, we show a comparative summary in terms of Inertia Explanatory Power (IEP), as obtained from CA/PCA without crime data on one hand and from MCA with crime data on the other hand.

The inclusion of crime statistics brings about a shift in IEP per borough, as evidenced by Table 7.

- The Bronx and Staten Island conserve their significance in terms of IEP.
- Meanwhile Manhattan loses its statistically prominence and goes from 46% to 29% of EIP over the same period.
- Brooklyn and Queens in turn gains in significance and go from 16 to 24% and from 17 to 28% respectively.

If one considers service request calls (SRCs) and crime report as "statistical events", then the petulance of Manhattan dwellers, when it comes to reporting urban nuisance by calling NYC 311, appears "statistically diluted" by the crime rates in the two neighboring counties of Queens and Brooklyn.

Borough	*number of ZIP codes*	*IEP all_dim PCA (%)*	*IEP 5_dim MCA (%)*
Bronx	*24*	*13.9*	*13*
Brooklyn	*38*	*16.2*	*23.4*
Manhattan	*46*	*46.2*	*29.1*
Queens	*59*	*17.2*	*28.1*
Staten Isl.	*12*	*6.0*	*6.5*

***Table 7**: Inertia explanatory power by ZIP codes, grouped by borough, computed over all dimensions (col. 3 - with PCA, SRCs w/o crime data) and over the 5 most significant significant dimensions (col. 4Th – with MCA, SRCs w/ crime data).*

At this stage hierarchical clustering as conducted by us also shows (viz. Fig. 21) that at least 3 dimensions are at play. This is made obvious by the fact that the 5 detected clusters appear (to a large extent) superimposed in the first factorial plane, but not so when looked at in 3 or more dimensions.

Figure 22 offers a basic 3D perspective for the five cluster classes clearly separated in 3D space.

The inset (bottom right) in Figure 22 above offers a perspective in the PC1-3 plane (i.e. phi=0° and theta=90°) with the 5 orange colored cluster class centroids. It reveals the importance of the third factorial dimension (PC3) for out of 1st factorial plane structure, and exhibits clearer cluster separation in 3D space.

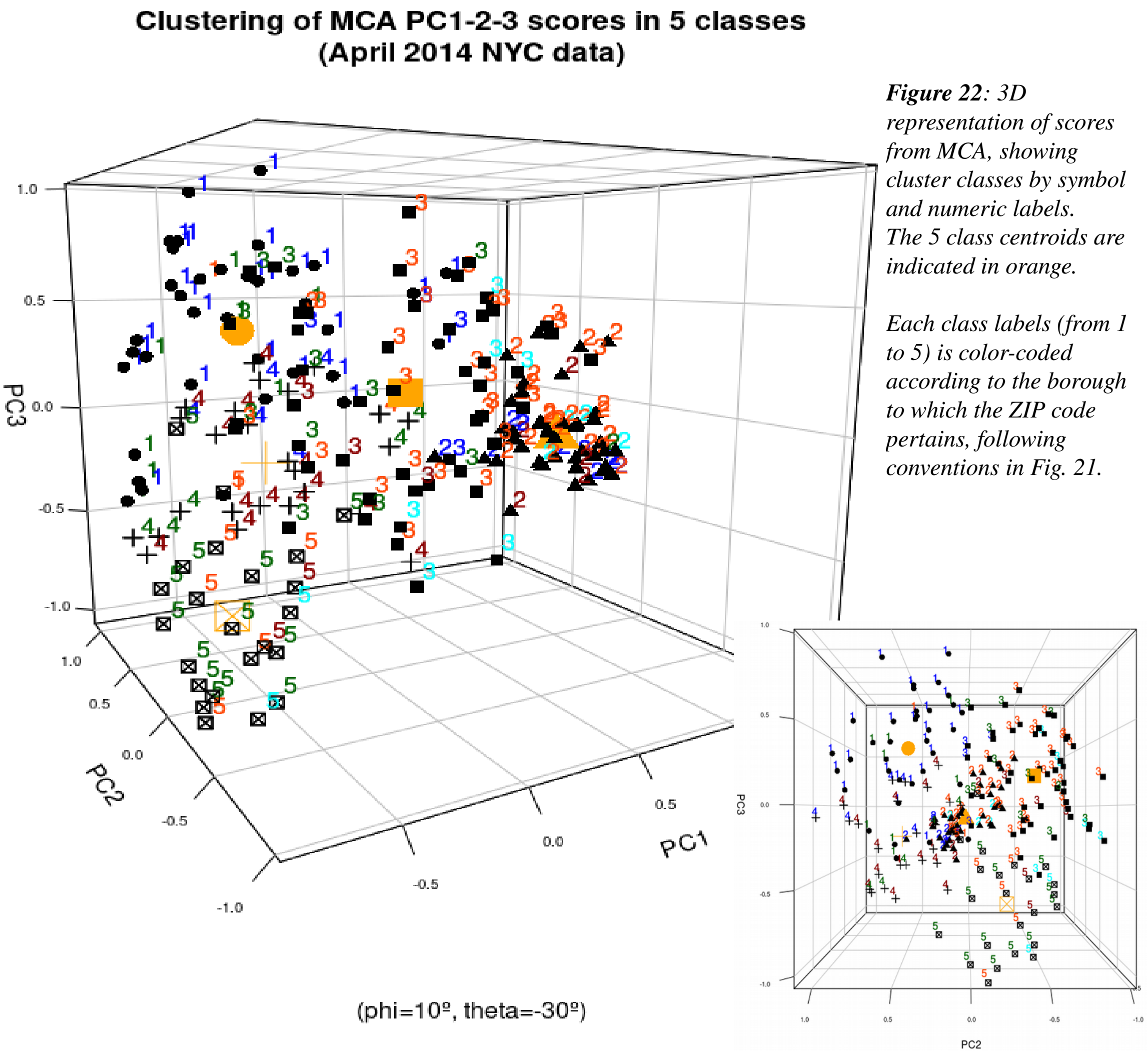


Figure 22: 3D representation of scores from MCA, showing cluster classes by symbol and numeric labels. The 5 class centroids are indicated in orange.

Each class labels (from 1 to 5) is color-coded according to the borough to which the ZIP code pertains, following conventions in Fig. 21.

3-3-2. Clustering with k-means consolidation

As pointed out before, we first deployed a probabilistic approach to ascertain the optimal number of clusters describing our data set. The class centroids resulting from hierarchical clustering are then used as seeds to conduct a new k-means computational optimization of classes. The so-called "clustering consolidation" technique permits overcoming to some extent the curse of "merges being final" in Ward2-based Hierarchical Clustering. A qualitative explanation follows.

Although the minimum sum-of-squares criterion is used in HC with Ward2, ensuring that no merge occurs if the system's resultant Sum of Square is not minimized, it is in fact possible to merge observations (or groups thereof) even though the merged "points" may be closer to another cluster's centroids than to the centroid of its current cluster. In that sense the sequence of successive merges in HC is path-dependent and therefore not optimal. By using centroids so obtained to conduct a new *k*-means optimization, a reshuffling of observations occurs about them, while the same centroids continuously updates their positions until convergence. This yields improved clustering results and remedies the difficulty inherent in performing HC with Ward2.

Figure 23 exhibits the modified scatter plot of MCA scores in the 1st factorial plane, *after consolidation*, using both color codes and character symbols of Figure 21.

As expected class members are somewhat redistributed. Most notably classes 1 (purple) and 3 (tan) appear to lose membership, while the numbers of others either remain stable or grow. By this consolidation method, the quality index is: $I_b = \dfrac{SS_W}{SS_W + SS_B}$ It can only improve. In the present case, its value increases from 60.3 to 62.8.

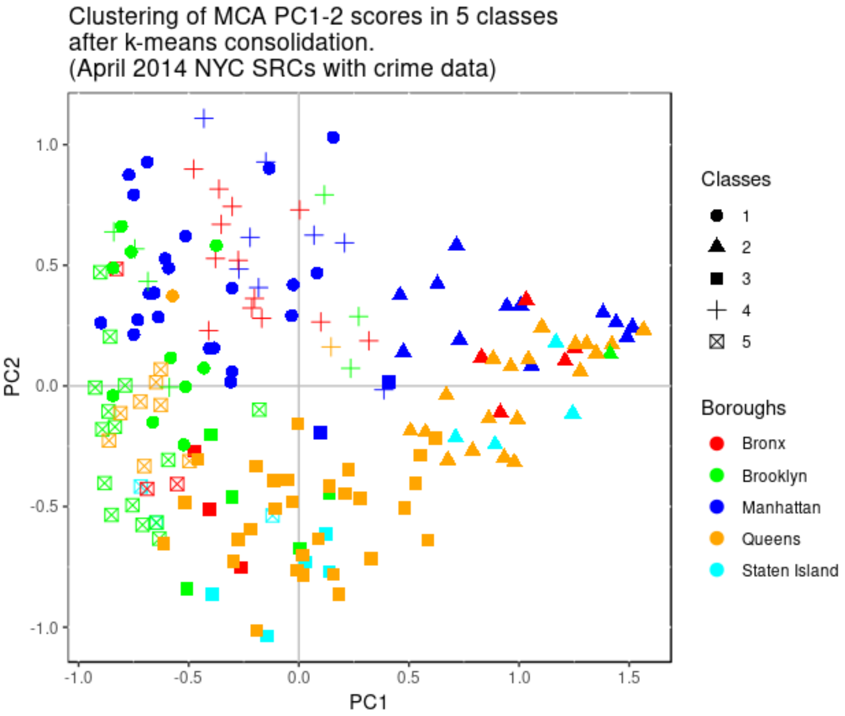


***Figure 23**: Graph of scores in the first factorial plane after k-means consolidation, where each ZIP code area is shown to belong to a specific cluster class, identified by a distinct symbol. Observations are color-coded following Fig. 21.*

Each class' member ZIP codes are topologically mapped in Figure 23 after consolidation. As before (in Figure 21) boroughs are not uniformly clustered. Rather:

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.

Additionally each cluster class' size and IEP is listed in Table 8 below before and after consolidation. Cluster classes' colors correspond to numeric labels and to character shapes (but not to colors) in Figure 22 and to cluster class hues (bottom right legend) in Figure 23.

As the main effect of *k*-means consolidation, we see that cluster class 3 (square symbols in Figures 21 and 23 and tan-colored dots in Figure 24) loses a lot of inertia representativeness: it goes from 21.8 to 14.8% in favor of classes 4 and 5, which increase each to 18.2%.

	Before consolidation		***After consolidation***	
Cluster class	***number of ZIP codes in class***	***IEP 5_dim (%)***	***number of ZIP codes in class***	***IEP 5_dim (%)***
1 (●)	*36*	*21*	*33*	*20.3*
2 (▲)	*43*	*28.6*	*42*	*28.6*
3 (■)	*50*	*21.8*	*44*	*14.8*
4 (+)	*25*	*13.5*	*29*	*18.2*
5 (☒)	*23*	*15.1*	*29*	*18.2*

***Table 8**: Inertia explanatory power by ZIP codes observation scores (SRCs w/ crime data), grouped by cluster class, computed over the 5 most significant significant dimensions , before and after k-means consolidation.*

Figure 24 is reminiscent of Fig 9, where ZIP code areas characterized by noise (in particular residential), traffic nuisance, a high crime rate and poor housing conditions (now signalled by dark red dots) seem better circumscribed than before as Cluster class 4.

■ Recalling that dot diameters are directly proportional to the frequency counts in a given ZIP code area, we see that adding crime statitics to SRCs has a balancing effect on dot size across cluster classes.

■ We nonetheless distinguish class 2 (cyan colored dots as the group of ZIP codes with the lowest incidence of recorded events and class 3 as the second lowest.

■ As previously noted, Manhattan now appears statistically comparable to other boroughs, if not in terms of the nature of recorded events, then at least in terms of event frequency.

The same topological representation, obtained from Hierarchical Clustering **without** *k*-means consolidation, is provided in Appendix E. It differs from Figure 24 in significant aspects.

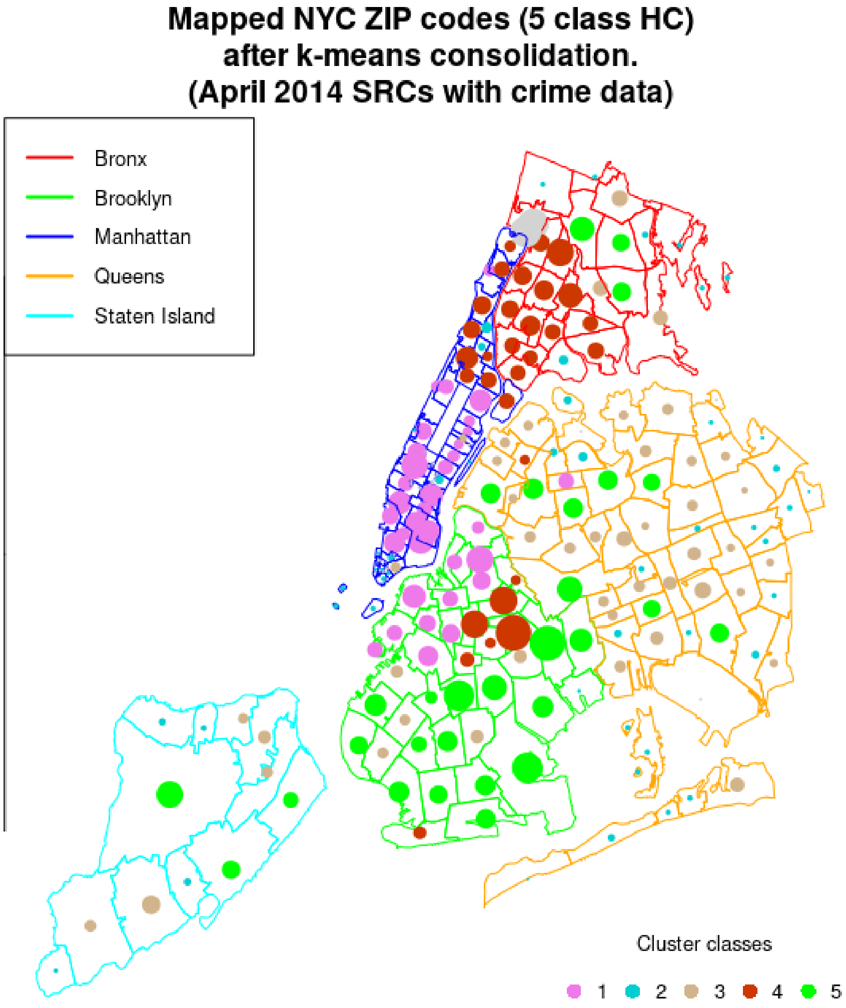

Mapped NYC ZIP codes (5 class HC)
after k-means consolidation.
(April 2014 SRCs with crime data)
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Cluster classes
● 1 ● 2 ● 3 ● 4 ● 5

Figure 24*: 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.*

To go beyond the mere visual inspection of the distribution cluster classes in the NYC geography, we need to identify quantitatively which factors, and among them which categorical variable modality, significantly contribute to the construction of clusters. For that we execute tests of independence between cluster classes (5 classes mean 4 DoFs) and each categorical variable (4 modalities means 3 DoFs). Results are summarized in Table 9, where we see that the null hypothesis (H0: “there is no relationship between the 2 tested categorical variables.”) is rejected for each variable but to various degrees.

Factor	**p.value (χ^2 independence test)**
ConsumProt	2.17e-29
Felony	6.69e-29
NoiseConst	3.84e-28
NoiseResid	1.07e-27
NoiseBiz	4.36e-27
Misdemeanor	2.36e-26
Sani	5.20e-26
Traffic	1.55e-23
Violation	1.12e-21
NoiseTraf	3.47e-21
UrbInf	6.34e-21
HousCond	6.31e-20
IAO	2.91e-17
EnvProt	1.32e-16
SocServ	2.87e-16
WaterSyst	4.60e-10

Table 9*: Variable relative significance in the construction of cluster classes. A smaller p-value means a higher significance.*

The variables most related to the formation of classes are, in decreasing order of significance:

- ConsumProt,
- Felony,
- NoiseConst,
- NoiseResid
- NoiseBiz
- Misdemeanor
- Sani

covering 3 orders of magnitude in values of p-values.

Inversely the least related one are WaterSyst, SocServ, EnvProt, and IAO, as was already intimated during the PCA, CA analyses.

Figures 24 and 25 below provide elements to interpret the latent class semantics. Figure 24 in particular show how certain variable modalities may play a role, i.e.:

- be significantly over-represented (in blue) simultaneously in up to 3 classes,
- be significantly under-represented (in red) simultaneously in up to 2 classes.

This induces a certain complexity of interpretation, better unraveled by Figures 25 on the basis of which, we propose a summary view of each class profile.

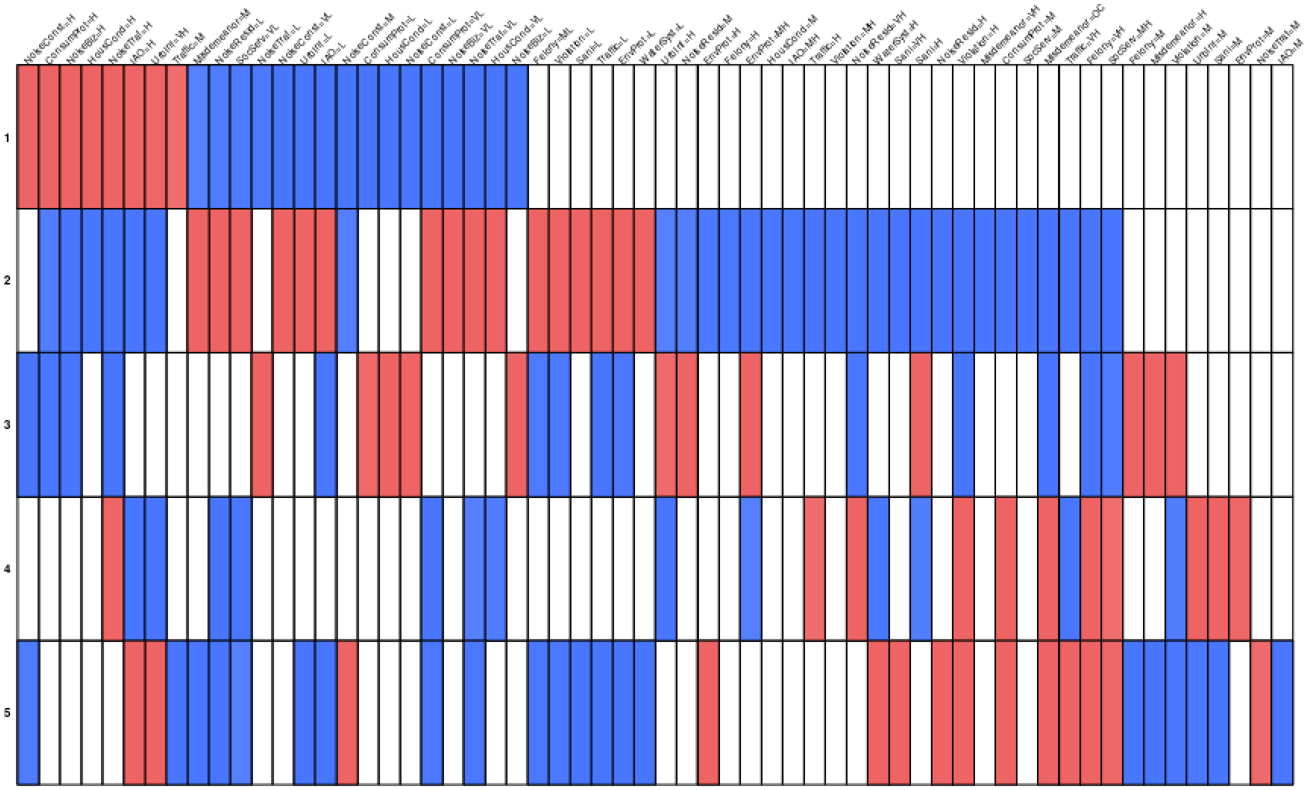


Figure 24*: Matrix view of over-represented (blue) and under-represented (red) modalities in each class (rows 1 to 5).*

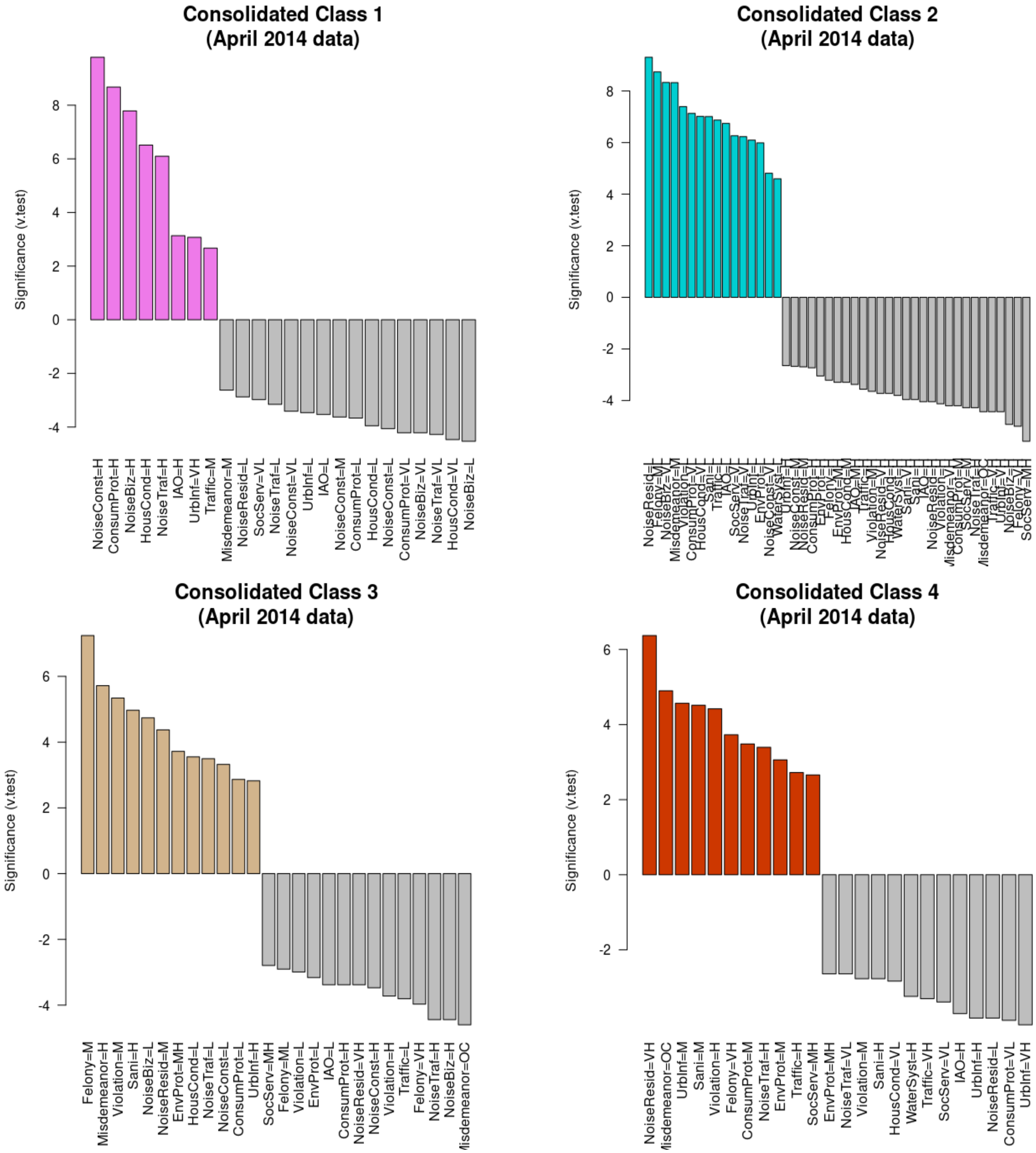


Figure 25*:Significance of modalities for classes 1,2 3 and 4 (NYC SRCs and crime data). Over-represented modalities are color-coded per class following the convention adopted for Table 8, while under-represented modalities are shown in gray.*

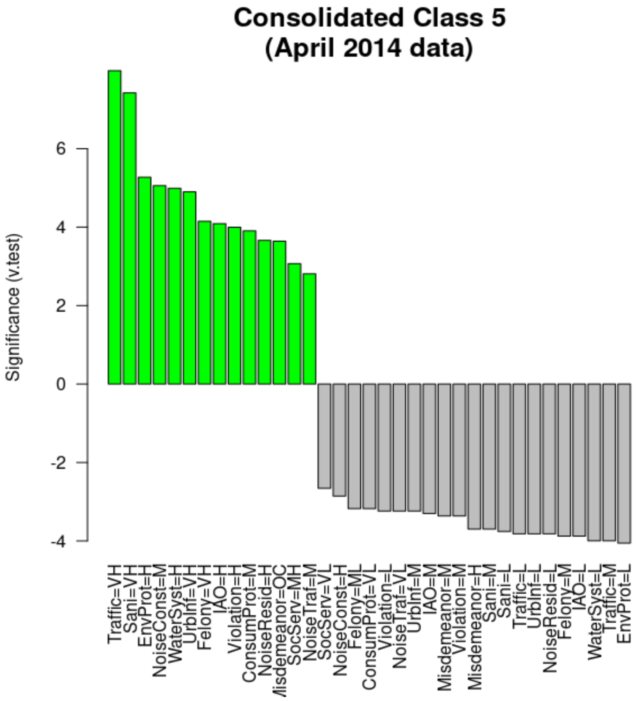


Figure 25 (continued)*:Significance of modalities for class 5 (NYC SRCs and crime data). Over-represented modalities are color-code following Table 8, while under-represented modalities are shown in gray.*

<u>Class 1</u>: 🟪
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.

<u>Class 2</u>: 🟦
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".

<u>Class 3</u>: 🟫
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".

<u>Class 4</u>: 🟥
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".

<u>Class 5</u>: 🟩
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.

4. Conclusions

The ultimate objectives of this work were to show (a) how the statistical analysis of SRCs enriched with crime rate provides a means to interpret the results of urban management policies. (b) how urban trends in the development, disappearance or displacement of pathologies can be extracted from the systematic study of successive time windows. This is not only to try to understand crime and its statistical correlations, but also to help political decision-makers and urban managers assign budgetary and human resources based on more reliable data models.

The composite data set used in this work is formed by data originating from different digital sources. Data pre-processing and generally speaking ETL at data mining stage occupied about 65% of our time and required more than 5000 lines of R code. We finally produced an automated ETL pipeline capable of processing complex, composite data almost unattended.

Data mining revealed that data is all too often incomplete, sometimes wrong, or statistically unreliable. This was the case for the continuous variable `jlBenef`, short for "jobless benefits" (source: IRS), which proved patchy at best, and was disregarded in later stages of this study. Meanwhile the other imported continuous variable `medianInc`, short for "median income" (source: IRS) was kept for later forays. Statisticians are keenly aware that skewing data, introducing bias(es) is a *caveat emptor,* when dealing with any type of statistical analysis. Caution must be applied not to introduce data biases prior or during the analysis by truncating data for instance. Thus, at various stages of this work, statistical tests (e.g. independence tests) were conducted to put our choices in perspective.

In the first half of our data exploration, we showed how we gain limited information from classical approaches such as CA / PCA and even MCA. Data structure subjected to dimensionality reduction was observed. Based on the determination of directions of maximum variance coupled with feature selection and extraction, latent semantics were proposed at different stages of the work. We confirmed a well-known result, namely that low frequency cells have a dramatic impact, in particular on CA / PCA results. This led us to gradually and carefully rid our data set from such spurious effects.

In the second half or this preliminary study, we deployed generic tools of clustering (unsupervised statistical learning). Our approach was based on the exact same data set as before (April 2014), augmented by crime rate statistics from NYPD. It also revealed that observations (ZIP codes) "10463" (Riverdale in the Bronx) and "11430" (JFK Airport in Queens) are meaningful observations (or "row-profiles") and at the same time true outliers, with considerable influence on our analytical results when included. They were therefore excluded from reported clustering results.

Our latent semantics analysis results, at the end of Section 3, point toward 5 cluster classes. They do not coincide with NYC's five boroughs, but rather with particular traits of the local geography and of the residents' socio-economic makeup inside those boroughs. Factors and intensity levels of those factors' modalities, instrumental in the statistical construction of those 5 classes, were elucidated. We showed how dramatic the effect of the addition of crime rate was, as it shifted the focus away from Manhattan and onto other boroughs, in terms of variance explanatory power..

All "statistical events" are neither socially equal nor equally acceptable. It becomes critical however, to discuss on their relative importance or statistical weight, when their joint analysis aims at designing balanced city budgets in a complex urban environment – e.g. as that of NYC. The scope of this work did not include a complicated discussion on the merits of (non-)uniform event weighting. For that reason we chose to deal with SRC data enriched with NYPD crime statistics and IRS data, aggregated at ZIP code area level without weighting. That however is strictly equivalent to weighting cluster classes according to their membership or boroughs according to how many observed ZIP code areas they contain. We are aware that this may also introduce a degree of (at this juncture unqualified) bias. Again, this issue was knowingly left unattended.

TODO *[Insert conclusive remarks on temporal evolution here]*

In order to further strengthen our multivariate analysis, we would welcome new data sources aggregated per ZIP code, with:

- population density,
- hospital beds per 10,000 inhabitants
- reliable income data,
- age distributions,
- academic achievements levels

This would allow us to embark on predictive classification or regression, for instance on continuous income level variable. That in turn would be usefully supplemented by a measure of quality, based not on a traditional confusion table, but on a multi-target (or -output) regression method such as support-vector-machine (SVM) based regression. As they stand, results presented in this document could also benefit from a comparison with results obtained with the more diversified toolkit of Machine Learning. Random Forest would be a prime candidate to extend this work.

As our primary objective is ultimately to further our understanding of the woes (social and urban) of large conurbations such as New York City, our immediate plan for further work is to lay the basis for the semantic analysis of urban areas based on events and Points of Interests (POIs), using the Word2Vec technique.

REFERENCES

1. Kaiser H. F., “The varimax criterion for analytic rotation in factor analysis”. Psychometrika, 23, p187–200 (1958).

2. @amoeba. “PCA - How to compute varimax-rotated principal components in R?” [Internet post: Cross Validated] (2017). Available from: https://stats.stackexchange.com/questions/59213/how-to-compute-varimax-rotated-principal-components-in-r

3. “k-means clustering” in Wikipedia [Internet] (2018) [cited 2018 Oct 23]. Available from: https://en.wikipedia.org/w/index.php?title=K-means_clustering&oldid=864895100

4. Kumar V., “Cluster Analysis: Basic Concepts and Algorithms” (Chapt 8), in “Introduction to Data Mining”, 2nd Ed., U. Minnesota, p. 487–568 (2019). (Series: What’s new in computer science). Available from: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

APPENDICES

Appendix A: Data-set's variables' dictionaries

NYC 311 Service Request Calls – Raw Data Dictionary

Column Name	Description
Unique Key	Unique identifier of a Service Request (SR) in the open data set
Created Date	Date SR was created Date in format MM/DD/YY HH:MM:SS AM/PM
Closed Date	Date SR was closed by responding agency. Date in format MM/DD/YY HH:MM:SS AM/PM
Agency	Acronym of responding City Government Agency
Agency Name	Full Agency name of responding City Government Agency
Complaint Type	This is the fist level of a hierarchy identifying the topic of the incident or condition. Complaint Type may have a corresponding Descriptor (below) or may stand alone.
Descriptor	This is associated to the Complaint Type, and provides further detail on the incident or condition. Descriptor values are dependent on the Complaint Type, and are not always required in SR.
Status	Status of SR submitted: Assigned, Canceled, Closed, Pending, +… (Prior column indicates most frequent)
Due Date	Date when responding agency is expected to update the SR. This is based on the Complaint Type and internal SLAs. Date in format MM/DD/YY HH:MM:SS AM/PM
Resolution Action Updated Date	Date when responding agency last updated the SR. Date in format MM/DD/YY HH:MM:SS AM/PM
Resolution Description	Describes the last action taken on the SR by the responding agency. May describe next or future steps.
Location Type	Describes the type of location used in the address information
Incident Zip	Incident location zip code, provided by geo validation.
Incident Address	House number of incident address provided by submitter.
Street Name	Street name of incident address provided by the submitter
Cross Street 1	First Cross street based on the geo validated incident location
Cross Street 2	Second Cross Street based on the geo validated incident location
Intersection Street 1	First intersecting street based on geo validated incident location
Intersection Street 2	Second intersecting street based on geo validated incident location
Address Type	Type of incident location information available (Values: Address; Block face; Intersection; LatLong; Placename)
City	City of the incident location provided by geovalidation.
Landmark	If the incident location is identified as a Landmark the name of the landmark will display here
Facility Type	If available, this field describes the type of city facility associated to the SR
Community Board	Provided by geovalidation.
Borough	Provided by the submitter and confirmed by geovalidation.
X Coordinate (State Plane)	Geo validated, X coordinate of the incident location.

Y Coordinate (State Plane)	Geo validated, Y coordinate of the incident location.
Latitude	Geo based Lat of the incident location
Longitude	Geo based Long of the incident location
Location	Combination of the geo based lat & long of the incident location
Park Facility Name	If the incident location is a Parks Dept facility, the Name of the facility will appear here
Park Borough	The borough of incident if it is a Parks Dept facility
School Name	If the incident location is a Dept of Education school, the name of the school will appear in this field. If the incident is a Parks Dept facility its name will appear here.
School Number	If the incident location is a Dept of Education school, the Number of the school will appear in this field. This field is also used for Parks Dept Facilities.
School Region	If the incident location is a Dept of Education School, the school region number will be appear in this field.
School Code	If the incident location is a Dept of Education School, the school code number will be appear in this field.
School Phone Number	If the facility = Dept for the Aging or Parks Dept, the phone number will appear here. (note - Dept of Education facilities do not display phone number)
School Address	Address of facility of incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept
School City	City of facilities incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept
School State	State of facility incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept NY
School Zip	Zip of facility incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept
School Not Found	Y' in this field indicates the facility was not found (Y; N; BLANK)
School or Citywide Complaint	If the incident is about a Dept of Education facility, this field will indicate if the complaint is about a particualr school or a citywide issue. (Y; N; BLANK)
Vehicle Type	If the incident is a taxi, this field describes the type of TLC vehicle.
Taxi Company Borough	If the incident is identified as a taxi, this field will display the borough of the taxi company.
Taxi Pick Up Location	If the incident is identified as a taxi, this field displays the taxi pick up location
Bridge Highway Name	If the incident is identified as a Bridge/Highway, the name will be displayed here.
Bridge Highway Direction	If the incident is identified as a Bridge/Highway, the direction where the issue took place would be displayed here.
Road Ramp	If the incident location was Bridge/Highway this column differentiates if the issue was on the Road or the Ramp.
Bridge Highway Segment	Additional information on the section of the Bridge/Highway were the incident took place.
Garage Lot Name	Related to DOT Parking Meter SR, this field shows what garage lot the meter is located in
Ferry Direction	Used when the incident location is within a Ferry, this field indicates the direction of ferry
Ferry Terminal Name	Used when the incident location is Ferry, this field indicates the ferry terminal where the incident took place.

NYPD Crime Reports – Raw Data Dictionary

CMPLNT_NUM	Randomly generated persistent ID for each complaint
CMPLNT_FR_DT	Exact date of occurrence for the reported event (or starting date of occurrence if CMPLNT_TO_DT exists)
CMPLNT_FR_TM	Exact time of occurrence for the reported event (or starting time of occurrence if CMPLNT_TO_TM exists)
CMPLNT_TO_DT	Ending date of occurrence for the reported event if exact time of occurrence is unknown
CMPLNT_TO_TM	Ending time of occurrence for the reported event if exact time of occurrence is unknown
RPT_DT	Date event was reported to police
KY_CD	Three digit offense classification code
OFNS_DESC	Description of offense corresponding with key code (KY_CD)
PD_CD	Three digit internal classification code (more granular than Key Code)
PD_DESC	Description of internal classification corresponding with PD code; more granular than Offense Description (OFNS_DESC).
CRM_ATPT_CPTD_CD	Crime completion indicator (completed, attempted but failed, interrupted prematurely)
LAW_CAT_CD	Level of offense (felony, misdemeanor, violation)
JURIS_DESC	Jurisdiction responsible for incident. Either internal (Police, Transit, Housing) or external (Correction, Port Authority, etc."
BORO_NM	The name of the borough in which the incident occurred
ADDR_PCT_CD	The precinct in which the incident occurred
LOC_OF_OCCUR_DESC	"Specific location of occurrence in or around the premises (inside, opposite of, in front of, at the rear of)
PREM_TYP_DESC	Specific description of premises (grocery store, residence, street, etc.)
PARKS_NM	Name of NYC park, playground or greenspace of occurrence if applicable (state parks are not included)
HADEVELOPT	Name of NYCHA housing development of occurrence if applicable
X_COORD_CD	X-coordinate for New York State Plane Coordinate System, Long Island Zone (NAD 83) in units of feet (FIPS 3104)
Y_COORD_CD	"Y-coordinate for New York State Plane Coordinate System, Long Island Zone (NAD 83) in units of feet (FIPS 3104)
Latitude	"Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) "
Longitude	"Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

IRS Statistics of Income per ZIP code– Raw Data Dictionary

IRS Documentation Guide (year 2014)

Contents

A. Overview
B. Nature of Changes
C. Population Definitions and Tax Return Addresses
D. Disclosure Protection Procedures

E. File Characteristics
F. Selected Income and Tax Items
G. Endnotes

A. Overview

The Statistics of Income (SOI) division bases its ZIP code data on administrative records of individual income tax returns (Forms 1040) from the Internal Revenue Service (IRS) Individual Master File (IMF) system. Included in these data are returns filed during the 12-month period, January 1, 2015 to December 31, 2015. While the bulk of returns filed during the 12-month period are primarily for Tax Year 2014, the IRS received a limited number of returns for tax years before 2014 and these have been included within the ZIP code data.

B. Nature of Changes

The following changes have been made to the Tax Year 2014 ZIP Code data:

- Two new variables have been added for volunteer prepared returns: volunteered income tax assistance (VITA) and tax counseling for the elderly (TCE) prepared returns.
- Five new variables, related to the Affordable Care Act (ACA), have been added to the data: Excess advance premium tax credit repayment, Total premium tax credit, Advance premium tax credit, Health care individual responsibility payment, and Net premium tax credit. Please refer to section F for a complete list of variables and their corresponding names.

C. Population Definitions and Tax Return Addresses

- ZIP Code data are based on population data that was filed and processed by the IRS during the 2015 calendar year.
- State totals may not be comparable to State totals published elsewhere by SOI because of specific disclosure protection features in the ZIP code data.
- Data do not represent the full U.S. population because many individuals are not required to file an individual income tax return.
- The address shown on the tax return may differ from the taxpayer's actual residence.
- State codes were based on the ZIP code shown on the return.
- Excluded were tax returns filed without a ZIP code and returns filed with a ZIP code that did not match the State code shown on the return.
- Excluded were tax returns filed using Army Post Office (APO) and Fleet Post Office addresses, foreign addresses, and addresses in Puerto Rico, Guam, Virgin Islands, American Samoa, Marshall Islands, Northern Marianas, and Palau.

D. Disclosure Protection Procedures

SOI did not attempt to correct any ZIP codes on the returns; however, it did take the following precautions to avoid disclosing information about specific taxpayers:

- ZIP codes with less than 100 returns and those identified as a single building or nonresidential ZIP code were categorized as "other" (99999).
- Income and tax items with less than 20 returns for a particular AGI class were combined with another AGI class within the same ZIP Code. Collapsed AGI classes are identified with a double asterisk (**).
- All number of returns variables have been rounded to the nearest 10.
- Excluded from the data are items with less than 20 returns within a ZIP code.
- Excluded from the data are tax returns with a negative adjusted gross income.
- Excluded are tax returns representing a specified percentage of the total of any particular cell. For example, if one return represented 75 percent of the value of a given cell, the return was suppressed from the tabulation. The actual threshold percentage used cannot be released.

E. File Characteristics

The ZIP code data are available in three formats:

(1) Individual state excel files—14zp##xx.xls (## = 01-51; xx = AL-WY)
(2) A comma separated file (.csv) with AGI classes —14zpallagi.csv
(3) A comma separated filewithout AGI classes(The AGI_STUB variable has been set to zero for this file)—14zpallnoagi.csv

For all the files, the money amounts are reported in thousands of dollars.

F. Selected Income and Tax Items

STATEFIPS	The State Federal Information Processing System (FIPS) code
STATE	The State associated with the ZIP code
ZIPCODE	5-digit Zip code
AGI_STUB	Size of Adjusted Gross Income (AGI) 1 = $1 under $25,000 2 = $25,000 under $50,000 3 = $50,000 under $75,000 4 = $75,000 under $100,000 5 = $100,000 under $200,000 6 = $200,000 or more
N1	Number of returns
...	**...**

G. Endnotes:
For complete individual income tax tabulations at the State level, see the historic table posted to Tax Stats at http://www.irs.gov/uac/SOI-Tax-Stats---Historic-Table-2.
Does not include returns with adjusted gross deficit.
The "Number of volunteer prepared returns" shows counts of returns prepared by IRS-certified volunteers to taxpayers with limited income, persons with disabilities, limited English speaking taxpayers, current and former members of the military, and taxpayers who are 60 years of age and older.
“Qualified dividends” are ordinary dividends received in tax years beginning after 2002 that meet certain conditions and receive preferential tax rates. The maximum qualified dividends tax rate is 15%.
Includes the Alaskan permanent fund, reported by residents of Alaska on Forms 1040A and 1040EZ's.
This fund only applies to statistics in the totals, and the state of Alaska.
Earned income credit includes both the refundable and non-refundable portions. The non-refundable portion could reduce income tax and certain related taxes to zero. The earned income credit amounts in excess of total tax liability, or amounts when there was no tax liability at all, were refundable. See footnote 6 below for explanation of the refundable portion of the earned income credit.
The refundable portion of the earned income credit equals total income tax minus the earned income credit. If the result is negative, this amount is considered the refundable portion. No other refundable credits were taken into account for this calculation.
Income tax reflects the amount reported on Form 1040 line 56. It also includes data from Form 1040A and 1040EZ filers.
“Total tax liability” differs from “Income tax”, in that “Total tax liability” includes the taxes from recapture of certain prior-year credits, tax applicable to individual retirement arrangements (IRA's), social security taxes on self-employment income and on certain tip income, advanced earned income payments, household employment taxes, and certain other taxes listed in the Form 1040 instructions.
[10] Reflects payments to or with-holdings made to "Total tax liability". This is the amount the tax filer owes when the income tax return is filed.
[11] The amount of over-payments the tax filer requested to have refunded.

Appendix B: NYPD's crime categorization

Crime modalities are: felony, misdemeanor, and violation.

FELONY is the most serious of offenses and gives rise to a more thorough classification. Felonies are lettered, with Class A being the most serious and Class E being the least serious. They are also divided into a smaller sub category; violent and non violent. In the state of NY, a non-violent, Class D felony would call for 1 to 4 years of probation. However, a violent Class D felony would automatically require a prison sentence of at least 2 years. What characterizes each felony as violent or non-violent is usually the presence of a weapon (possession of a firearm) or bodily harm to another person (aggravated assault/battery). A Class A Felony (e.g a 1st degree murder) is punishable by life in prison, with or without parole, depending on the circumstances.

MISDEMEANOR is the second type of criminal offenses, less severe than felonies but more serious than violations. Misdemeanors can carry up to a year in jail. In addition to jail time, a person convicted of a misdemeanor can also be subject to fines, probation, community service or restitution (victim compensation). A classic case of a misdemeanor would be simple assault, possession of a small amount of marijuana, or driving under the influence.

VIOLATION (also known as "infractions") is a minor offense. A speeding ticket, public intoxication, or jaywalking are some of the many petty offenses that could fall under the umbrella of violations. Violations are punishable by fines primarily, and do not result in jail or prison time.

In the subsequent listings, a number following a label within each category indicates the degree of the charge within that category, i.e. sub-categorization for judicial purposes.

Felonies

RAPE 1 *(means "1st degree rape", i.e. generally speaking rape under the threat of a deadly weapon, etc.)*
LARCENY,GRAND BY OPEN/COMPROMISE CELL PHONE ACCT
LARCENY,GRAND BY OPEN CREDIT CARD (NEW ACCT)
RAPE 3
FRAUD,UNCLASSIFIED-FELONY
LARCENY,GRAND BY DISHONEST EMP
BURGLARY,RESIDENCE,NIGHT
SEX CRIMES
RAPE 2
LARCENY,GRAND BY BANK ACCT COMPROMISE-REPRODUCED CHECK
SODOMY 1
LARCENY,GRAND BY THEFT OF CREDIT CARD
LARCENY,GRAND BY FALSE PROMISE-NOT IN PERSON CONTACT
LARCENY,GRAND FROM RESIDENCE, UNATTENDED
SEXUAL ABUSE
LARCENY,GRAND FROM BUILDING (NON-RESIDENCE) UNATTENDED
COERCION 1
PUBLIC ADMINISTRATION,UNCLASSI
COMPUTER TAMPER/TRESSPASS
LARCENY,GRAND FROM OPEN AREAS, UNATTENDED
LARCENY,GRAND BY IDENTITY THEFT-UNCLASSIFIED
BURGLARY,RESIDENCE,UNKNOWN TIM
BURGLARY,RESIDENCE,DAY
LARCENY,GRAND BY FALSE PROMISE-IN PERSON CONTACT
TAMPERING 1,CRIMINAL
RAPE 1, ATTEMPT
LARCENY,GRAND BY CREDIT CARD ACCT COMPROMISE-EXISTING ACCT

```
LARCENY,GRAND BY BANK ACCT COMPROMISE-TELLER
FORGERY,ETC.,UNCLASSIFIED-FELO
NY STATE LAWS,UNCLASSIFIED FEL
CRIMINAL CONTEMPT 1
LARCENY,GRAND BY BANK ACCT COMPROMISE-ATM TRANSACTION
LARCENY,GRAND BY ACQUIRING LOST CREDIT CARD
MISCHIEF,CRIMINAL,    UNCL 2ND
ARSON 2,3,4
RECKLESS ENDANGERMENT 1
MISCHIEF, CRIMINAL 3 & 2, OF M
LARCENY,GRAND OF VEHICULAR/MOTORCYCLE ACCESSORIES
LARCENY,GRAND FROM STORE-SHOPL
LARCENY,GRAND BY BANK ACCT COMPROMISE-UNCLASSIFIED
LARCENY,GRAND BY ACQUIRING LOS
LARCENY,GRAND FROM VEHICLE/MOTORCYCLE
LARCENY,GRAND OF AUTO
BURGLARY,COMMERCIAL,NIGHT
LARCENY,GRAND FROM RETAIL STORE, UNATTENDED
BURGLARY,COMMERCIAL,UNKNOWN TI
LARCENY,GRAND FROM PERSON,PICK
LARCENY,GRAND OF MOTORCYCLE
LARCENY,GRAND BY EXTORTION
WEAPONS POSSESSION 3
FORGERY,DRIVERS LICENSE
LARCENY,GRAND FROM PERSON,PERSONAL ELECTRONIC DEVICE(SNATCH)
ROBBERY,OPEN AREA UNCLASSIFIED
LARCENY,GRAND FROM NIGHT CLUB, UNATTENDED
CONTROLLED SUBSTANCE,INTENT TO
ASSAULT 2,1,UNCLASSIFIED
CONTROLLED SUBSTANCE,POSSESS.
ROBBERY,DWELLING
IMPRISONMENT 1,UNLAWFUL
STRANGULATION 1ST
LARCENY,GRAND FROM EATERY, UNATTENDED
STOLEN PROPERTY 2,1,POSSESSION
LARCENY, GRAND OF AUTO - ATTEM
BURGLARY,TRUCK NIGHT
ROBBERY,PERSONAL ELECTRONIC DEVICE
BURGLARY,UNCLASSIFIED,NIGHT
LARCENY,GRAND OF BICYCLE
ARSON, MOTOR VEHICLE 1 2 3 & 4
WEAPONS POSSESSION 1 & 2
CONTROLLED SUBSTANCE, SALE 5
FORGERY,M.V. REGISTRATION
ASSAULT 2,1,PEACE OFFICER
ROBBERY,COMMERCIAL UNCLASSIFIED
FORGERY-ILLEGAL POSSESSION,VEH
ROBBERY,RESIDENTIAL COMMON AREA
LARCENY,GRAND FROM PERSON, BAG OPEN/DIP
CONTROLLED SUBSTANCE,SALE 1
BRIBERY,PUBLIC ADMINISTRATION
IMPERSONATION 1, POLICE OFFICER
MARIJUANA, SALE 1, 2 & 3
ROBBERY,PUBLIC PLACE INSIDE
```

```
MENACING 1ST DEGREE (VICT NOT
CRIMINAL MIS 2 & 3
ROBBERY, PAYROLL
ROBBERY,HOME INVASION
CONTROLLED SUBSTANCE,SALE 3
LARCENY,GRAND FROM PERSON,PURS
THEFT,RELATED OFFENSES,UNCLASS
LARCENY,GRAND FROM PERSON,UNCL
ROBBERY,CAR JACKING
AGGRAVATED HARASSMENT 1
BURGLARY,COMMERCIAL,DAY
LARCENY,GRAND BY BANK ACCT COMPROMISE-UNAUTHORIZED PURCHASE
ROBBERY,POCKETBOOK/CARRIED BAG
CONTROLLED SUBSTANCE, POSSESSI
UNAUTHORIZED USE VEHICLE 2
CONTROLLED SUBSTANCE, INTENT T
BURGLARY,TRUCK DAY
MARIJUANA, POSSESSION 1, 2 & 3
ROBBERY,OF TRUCK DRIVER
CRIMINAL DISPOSAL FIREARM 1 &
CONTROLLED SUBSTANCE,SALE 2
LARCENY,GRAND BY OPEN BANK ACCT
BURGLARY,UNCLASSIFIED,UNKNOWN
FORGERY,PRESCRIPTION
SODOMY 2
GAMBLING 1,PROMOTING,BOOKMAKIN
AGGRAVATED CRIMINAL CONTEMPT
ROBBERY, CHAIN STORE
FALSE REPORT 1,FIRE
ROBBERY,PHARMACY
ROBBERY,LICENSED MEDALLION CAB
STOLEN PROPERTY-MOTOR VEH 2ND,
LARCENY,GRAND OF TRUCK
ROBBERY,LIQUOR STORE
LARCENY,GRAND FROM PERSON,LUSH WORKER(SLEEPING/UNCON VICTIM)
BRIBERY, POLICE OFFICER
ARSON 1
TRESPASS 1,CRIMINAL
ROBBERY,UNLICENSED FOR HIRE VEHICLE
CONTROLLED SUBSTANCE, SALE 4
ROBBERY,BICYCLE
OBSCENE MATERIAL - UNDER 17 YE
ROBBERY,BANK
ROBBERY,NECKCHAIN/JEWELRY
LARCENY,GRAND PERSON,NECK CHAI
ROBBERY,BODEGA/CONVENIENCE STORE
DRUG PARAPHERNALIA, POSSESSION
CUSTODIAL INTERFERENCE 1
ESCAPE 2,1
PROMOTING A SEXUAL PERFORMANCE
BURGLARY,UNCLASSIFIED,DAY
ROBBERY,GAS STATION
MENACING 1ST DEGREE (VICT PEAC
USE OF A CHILD IN A SEXUAL PERFORMANCE
```

CONSPIRACY 2, 1
SEX TRAFFICKING
INCOMPETENT PERSON,KNOWINGLY ENDANGERING
TAX LAW
MANUFACTURE UNAUTHORIZED RECOR
MISCHIEF, CRIMINAL 3&2, BY FIR
ROBBERY,ON BUS/ OR BUS DRIVER
ROBBERY,ATM LOCATION
LARCENY,GRAND FROM TRUCK, UNATTENDED
OBSCENITY 1
CHILD ABANDONMENT
INTOXICATED DRIVING,ALCOHOL
HOMICIDE, NEGLIGENT, VEHICLE,
MAKING TERRORISTIC THREAT
BURGLARY,UNKNOWN TIME
KIDNAPPING 2
BAIL JUMPING 1 & 2
FACILITATION 3,2,1, CRIMINAL
SOLICITATION 3,2,1, CRIMINAL
END WELFARE VULNERABLE ELDERLY PERSON
AGGRAVATED SEXUAL ASBUSE
LARCENY,GRAND FROM PIER, UNATTENDED
ROBBERY,BAR/RESTAURANT
SODOMY 3
SUPP. ACT TERR 2ND
LARCENY, GRAND OF MOPED
LARCENY,GRAND FROM BOAT, UNATTENDED
SALE SCHOOL GROUNDS 4
KIDNAPPING 1
ROBBERY,CHECK CASHING BUSINESS

Misdemeanors

ASSAULT 3
LARCENY,PETIT FROM BUILDING,UN
FRAUD,UNCLASSIFIED-MISDEMEANOR
AGGRAVATED HARASSMENT 2
SEXUAL ABUSE 3,2
CRIMINAL MISCHIEF 4TH, GRAFFIT
SEXUAL MISCONDUCT,INTERCOURSE
CRIMINAL MISCHIEF,UNCLASSIFIED 4
MISCHIEF, CRIMINAL 4, BY FIRE
MISCHIEF, CRIMINAL 4, OF MOTOR
LARCENY,PETIT OF LICENSE PLATE
CHILD, ENDANGERING WELFARE
UNAUTHORIZED USE VEHICLE 3
VIOLATION OF ORDER OF PROTECTI
PUBLIC ADMINISTATION,UNCLASS M
LARCENY,PETIT BY CREDIT CARD U
CUSTODIAL INTERFERENCE 2
LARCENY,PETIT FROM OPEN AREAS,
NY STATE LAWS,UNCLASSIFIED MIS
LARCENY,PETIT FROM STORE-SHOPL
FORGERY,ETC.-MISD.

LARCENY,PETIT FROM AUTO
STOLEN PROPERTY 3,POSSESSION
LARCENY,PETIT BY FALSE PROMISE
CONTEMPT,CRIMINAL
LARCENY,PETIT BY CHECK USE
BRIBERY,COMMERCIAL
MENACING,UNCLASSIFIED
OBSTR BREATH/CIRCUL
ADM.CODE,UNCLASSIFIED MISDEMEA
LARCENY,PETIT OF VEHICLE ACCES
LEWDNESS,PUBLIC
CONTROLLED SUBSTANCE, POSSESSI
MARIJUANA, POSSESSION 4 & 5
WEAPONS, POSSESSION, ETC
INTOXICATED DRIVING,ALCOHOL
TRESPASS 2, CRIMINAL
THEFT,RELATED OFFENSES,UNCLASS
ACCOSTING,FRAUDULENT
MARIJUANA, SALE 4 & 5
LARCENY,PETIT OF MOTORCYCLE
LARCENY,PETIT OF BICYCLE
RECKLESS ENDANGERMENT 2
LEAVING SCENE-ACCIDENT-PERSONA
IMPERSONATION 2, PUBLIC SERVAN
RESISTING ARREST
TRAFFIC,UNCLASSIFIED MISDEMEAN
LARCENY,PETIT BY ACQUIRING LOS
TRESPASS 3, CRIMINAL
LARCENY,PETIT FROM TRUCK
IMPRISONMENT 2,UNLAWFUL
BURGLARS TOOLS,UNCLASSIFIED
THEFT OF SERVICES, UNCLASSIFIE
LARCENY,PETIT FROM BOAT
LARCENY,PETIT BY DISHONEST EMP
RECKLESS ENDANGERMENT OF PROPE
TAX LAW
UNAUTH. SALE OF TRANS. SERVICE
PETIT LARCENY-CHECK FROM MAILB
IMPAIRED DRIVING,DRUG
ASSEMBLY,UNLAWFUL
BAIL JUMPING 3
FALSE REPORT UNCLASSIFIED
RECORDS,FALSIFY-TAMPER
SEXUAL MISCONDUCT,DEVIATE
PROSTITUTION, PATRONIZING 4, 3
SALE OF UNAUTHORIZED RECORDING
DRUG PARAPHERNALIA,   POSSESSE
CHILD,ALCOHOL SALE TO
GAMBLING 2,PROMOTING,UNCLASSIF
CHECK,BAD
FALSE REPORT BOMB
LARCENY, PETIT OF AUTO - ATTEM
RECKLESS DRIVING
AGRICULTURE & MARKETS LAW,UNCL

TAMPERING 3,2, CRIMINAL
PROSTITUTION 4,PROMOTING&SECUR
GENERAL BUSINESS LAW,TICKET SP
LARCENY,PETIT OF BOAT
POSSESSION HYPODERMIC INSTRUME
ALCOHOLIC BEVERAGE CONTROL LAW
GAMBLING, DEVICE, POSSESSION
STOLEN PROP-MOTOR VEHICLE 3RD,
CHILD,OFFENSES AGAINST,UNCLASS
LARCENY,PETIT OF AUTO
PUBLIC SAFETY,UNCLASSIFIED MIS
LARCENY, PETIT OF MOPED
DOG STEALING
DIS. CON.,AGGRAVATED
RIOT 2/INCITING
MENACING,PEACE OFFICER
JOSTLING
PERJURY 3,ETC.
ESCAPE 3
PUBLIC HEALTH LAW,UNCLASSIFIED
COMPUTER UNAUTH. USE/TAMPER
FALSE ALARM FIRE
NUISANCE,CRIMINAL,UNCLASSIFIED
WOUNDS,REPORTING OF
LARCENY, PETIT FROM COIN MACHINE

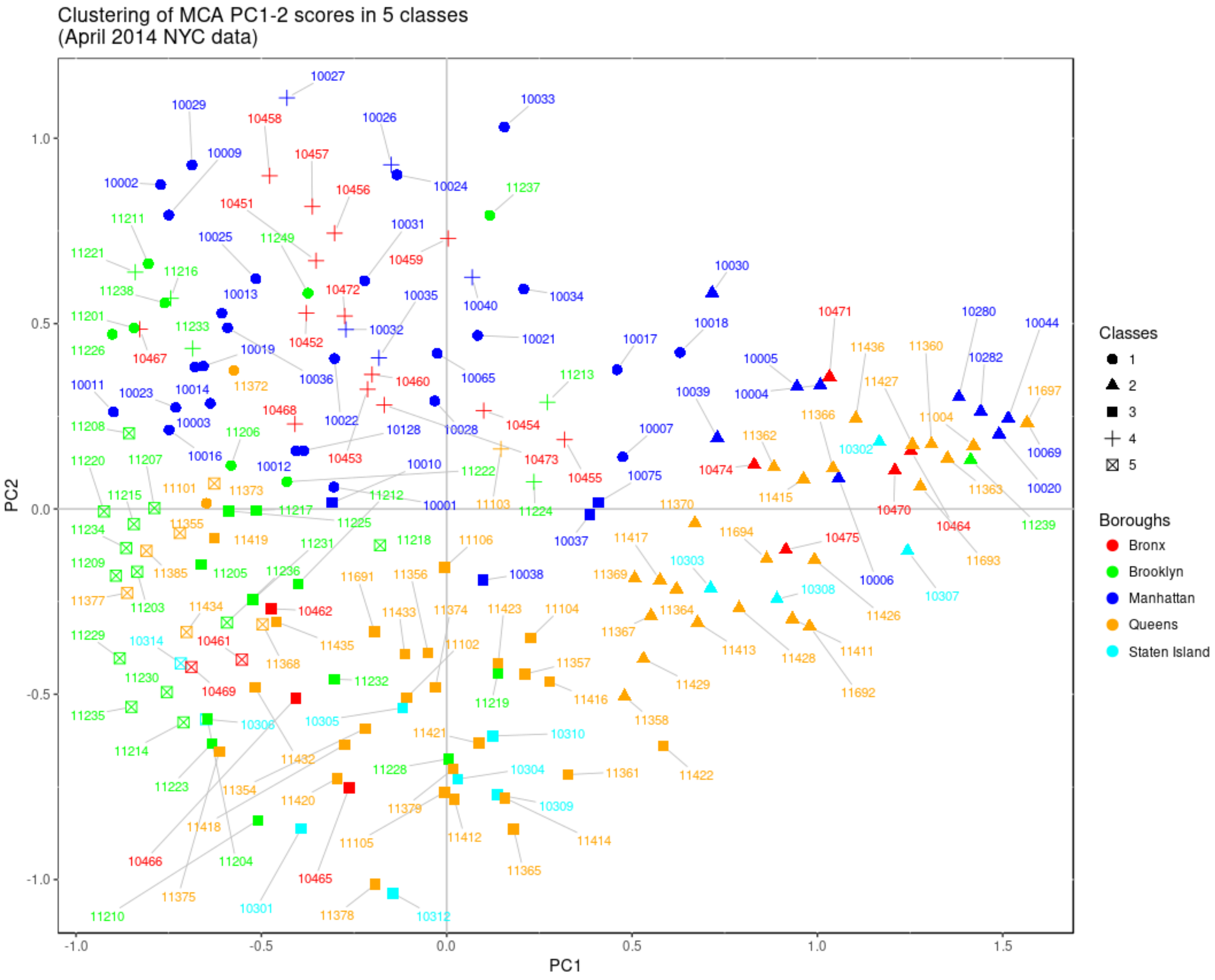
Violations

HARASSMENT,SUBD 3,4,5
HARASSMENT,SUBD 1,CIVILIAN
MARIJUANA, POSSESSION
ALCOHOLIC BEVERAGES,PUBLIC CON
THEFT OF SERVICES- CABLE TV SE
POSSES OR CARRY A KNIFE
ADM.CODE,UNCLASSIFIED VIOLATIO
PEDDLING,UNLAWFUL
TRESPASS 4,CRIMINAL SUB 2
DISORDERLY CONDUCT
IMITATION PISTOL/AIR RIFLE
PARKR&R,UNCLASSIFIED VIOLATION
NY STATE LAWS,UNCLASSIFIED VIO
APPEARANCE TICKET FAIL TO RESP
IMITATION PISTOL/AIR RIFLE
TRAFFIC,UNCLASSIFIED INFRACTION
LOITERING,GAMBLING,OTHER
ENVIRONMENTAL CONTROL BOARD
INAPPROPIATE SHELTER DOG LEFT
EXPOSURE OF A PERSON
UNDER THE INFLUENCE OF DRUGS

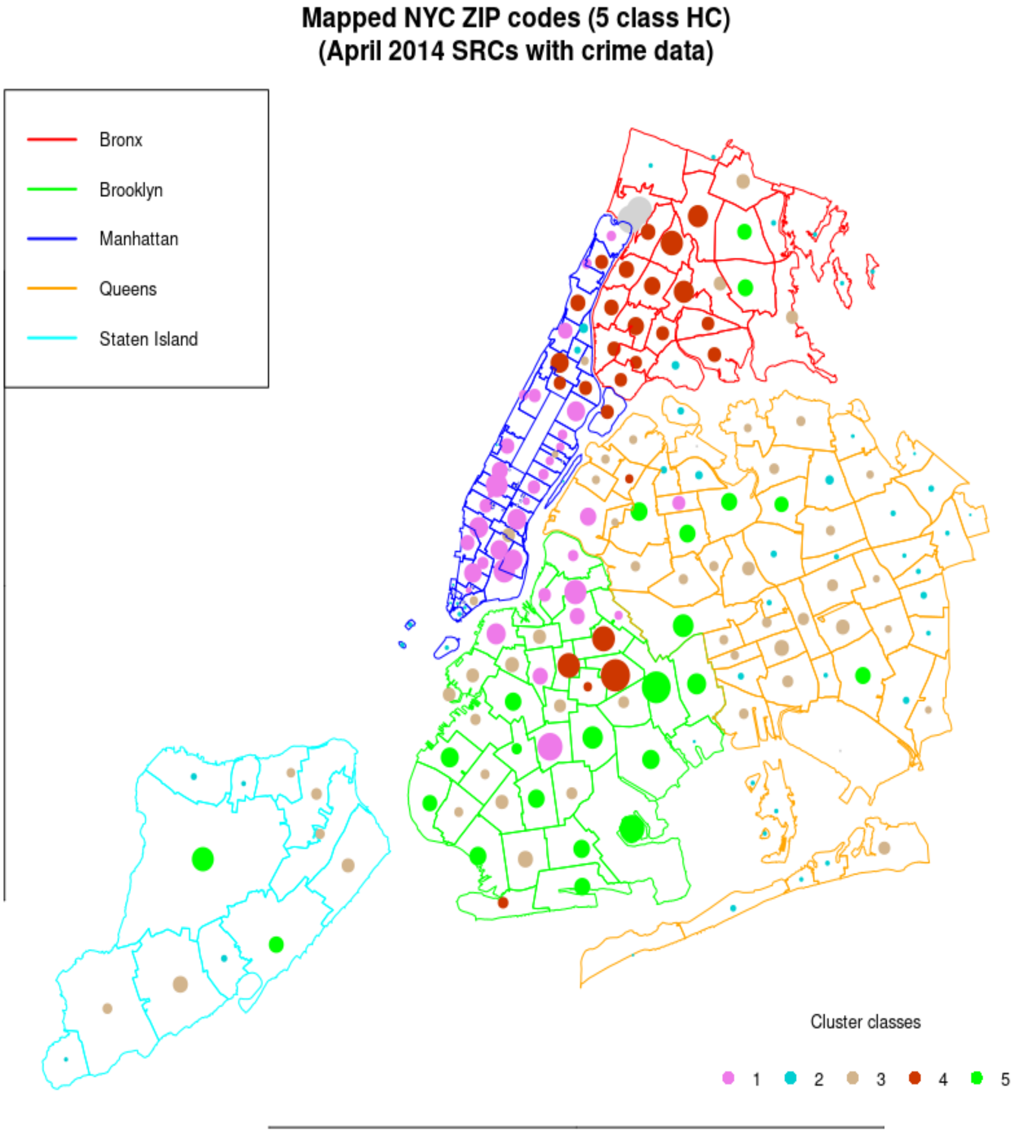
Appendix C: Index of ZIP codes and New York city boroughs

ZIP	Borough
10001	Manhattan
10002	Manhattan
10003	Manhattan
10004	Manhattan
10005	Manhattan
10006	Manhattan
10007	Manhattan
10009	Manhattan
10010	Manhattan
10011	Manhattan
10012	Manhattan
10013	Manhattan
10014	Manhattan
10016	Manhattan
10017	Manhattan
10018	Manhattan
10019	Manhattan
10020	Manhattan
10021	Manhattan
10022	Manhattan
10023	Manhattan
10024	Manhattan
10025	Manhattan
10026	Manhattan
10027	Manhattan
10028	Manhattan
10029	Manhattan
10030	Manhattan
10031	Manhattan
10032	Manhattan
10033	Manhattan
10034	Manhattan
10035	Manhattan
10036	Manhattan
10037	Manhattan
10038	Manhattan
10039	Manhattan
10040	Manhattan
10041	Manhattan
10044	Manhattan
10045	Manhattan
10048	Manhattan
10065	Manhattan
10069	Manhattan
10075	Manhattan
10103	Manhattan
10107	Manhattan
10111	Manhattan
10112	Manhattan
10118	Manhattan
10119	Manhattan
10129	Manhattan
10162	Manhattan
10163	Manhattan
10167	Manhattan
10170	Manhattan
10172	Manhattan
10178	Manhattan
10203	Manhattan
10259	Manhattan
10278	Manhattan
10280	Manhattan
10281	Manhattan
10282	Manhattan
10301	Staten Isl.
10302	Staten Isl.
10303	Staten Isl.
10304	Staten Isl.
10305	Staten Isl.
10306	Staten Isl.
10307	Staten Isl.
10308	Staten Isl.
10309	Staten Isl.
10310	Staten Isl.
10312	Staten Isl.
10314	Staten Isl.
10451	Bronx
10452	Bronx
10453	Bronx
10454	Bronx
10455	Bronx
10456	Bronx
10457	Bronx
10458	Bronx
10459	Bronx
10460	Bronx
10461	Bronx
10462	Bronx
10463	Bronx
10464	Bronx
10465	Bronx
10466	Bronx
10467	Bronx
10468	Bronx
10469	Bronx
10470	Bronx
10471	Bronx
10472	Bronx
10473	Bronx
10474	Bronx
10475	Bronx
11004	Queens
11101	Queens
11102	Queens
11103	Queens
11104	Queens
11105	Queens
11106	Queens
11109	Queens
11201	Brooklyn
11202	Brooklyn
11203	Brooklyn
11204	Brooklyn
11205	Brooklyn
11206	Brooklyn
11207	Brooklyn
11208	Brooklyn
11209	Brooklyn
11210	Brooklyn
11211	Brooklyn
11212	Brooklyn
11213	Brooklyn
11214	Brooklyn
11215	Brooklyn
11216	Brooklyn
11217	Brooklyn
11218	Brooklyn
11219	Brooklyn
11220	Brooklyn
11221	Brooklyn
11222	Brooklyn
11223	Brooklyn
11224	Brooklyn
11225	Brooklyn
11226	Brooklyn
11228	Brooklyn
11229	Brooklyn
11230	Brooklyn
11231	Brooklyn
11232	Brooklyn
11233	Brooklyn
11234	Brooklyn
11235	Brooklyn
11236	Brooklyn
11237	Brooklyn
11238	Brooklyn
11239	Brooklyn
11249	Brooklyn
11251	Brooklyn
11354	Queens
11355	Queens
11356	Queens
11357	Queens
11358	Queens
11359	Queens
11360	Queens
11361	Queens
11362	Queens
11363	Queens
11364	Queens
11365	Queens
11366	Queens
11367	Queens
11368	Queens
11369	Queens
11370	Queens
11371	Queens
11372	Queens
11373	Queens
11374	Queens
11375	Queens
11377	Queens
11378	Queens
11379	Queens
11385	Queens
11411	Queens
11412	Queens
11413	Queens
11414	Queens
11415	Queens
11416	Queens
11417	Queens
11418	Queens
11419	Queens
11420	Queens
11421	Queens
11422	Queens
11423	Queens
11426	Queens
11427	Queens
11428	Queens
11429	Queens
11430	Queens
11432	Queens
11433	Queens
11434	Queens
11435	Queens
11436	Queens
11451	Queens
11691	Queens
11692	Queens
11693	Queens
11694	Queens
11695	Queens
11697	Queens
99999	bogus ZIP

Appendix D: ZIP codes projection in PC1-2 after MCA and k-means/HC clustering (April 2014 data)


Clustering of MCA PC1-2 scores in 5 classes
(April 2014 NYC data)

Appendix E: Topological representation after MCA and clustering, without consolidation – April 2014 NYC SRCs+crime data


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