International Journal of Architectural, Civil and Construction Sciences
ISSN: 2415-1734
Vol:14, No:11, 2020

Spatial Data Science for Data Driven Urban Planning: The Youth Economic Discomfort Index for Rome

Iacopo Testi, Diego Pajarito, Nicoletta Roberto, Carmen Greco

***Abstract*—**Today, a consistent segment of the world's population lives in urban areas, and this proportion will vastly increase in the next decades. Therefore, understanding the key trends in urbanization, likely to unfold over the coming years, is crucial to the implementation of sustainable urban strategies. In parallel, the daily amount of digital data produced will be expanding at an exponential rate during the following years. The analysis of various types of data sets and its derived applications have incredible potential across different crucial sectors such as healthcare, housing, transportation, energy, and education. Nevertheless, in city development, architects and urban planners appear to rely mostly on traditional and analogical techniques of data collection. This paper investigates the prospective of the data science field, appearing to be a formidable resource to assist city managers in identifying strategies to enhance the social, economic, and environmental sustainability of our urban areas. The collection of different new layers of information would definitely enhance planners' capabilities to comprehend more in-depth urban phenomena such as gentrification, land use definition, mobility, or critical infrastructural issues. Specifically, the research results correlate economic, commercial, demographic, and housing data with the purpose of defining the youth economic discomfort index. The statistical composite index provides insights regarding the economic disadvantage of citizens aged between 18 years and 29 years, and results clearly display that central urban zones and more disadvantaged than peripheral ones. The experimental set up selected the city of Rome as the testing ground of the whole investigation. The methodology aims at applying statistical and spatial analysis to construct a composite index supporting informed data-driven decisions for urban planning.

***Keywords*—**Data science, spatial analysis, composite index, Rome, urban planning, youth economic discomfort index.

I. Introduction

OVER the last two centuries humanity experienced a phenomenon of urbanization that has never happened before. Today, 55% of the world's population lives in urban areas, a proportion that is expected to increase to 68% by 2050 [1]. Rome is no exception, the Italian capital, home to 200,000 citizens in 1870, skyrocketed to approximately three million current inhabitants. It is during the enormous urban growth at the end of the 19th century that a master-plan for Rome, named PRG (Piano Regolatore Generale) was adopted for the first time. The PRG was subjected to many alterations throughout time, starting with its first version in 1870, never really implemented, to its final elaboration and current form dating back to 2016. The conception of the master-plan responded mostly to the basic needs of providing and regulating the new infrastructures - building housing, highways, schools, public amenities, parking lots, etc. Therefore, since from the beginning, planners relied on specific tools and techniques to apply urban standards such as defining heights, built area percentage, social housing proportion, etc. Nowadays the city development is involved in a paradigm shift: moving from the infrastructure to the content of it. The incredible amount of data produced by objects, buildings and municipalities is crucial to comprehend not only how humans are moving throughout cities but also their preferences, expenses, emissions, social and economic status.

Over the last three years alone, 90% of the data in the world were generated and the amount created daily is expected to rise steadily throughout the years to come. This is one of the motivations why the urban planning realm is likely to be strongly influenced by this so-called 'Big Data' wave. As a result, an increase in relevance is dedicated to the professionals capable of processing and analyzing, through statistical procedures, this massive quantity of information, working in close collaboration with multidisciplinary teams. Another consideration, worth noting at the beginning of the discourse, is regarding the importance of open data sources [2] as necessary requirement for any advancement in the planning field. There are multiple reasons to concentrate the attention on specific datasets with potential urban purposes. A common example of that is mobility. Applications like Google maps, Waze, TomTom possess millions of user's origin, route and destination opening the opportunity to investigate how infrastructures are used and discover potentials or critical issues. Another instance is mobile data, through which telephone companies might uncover citizens' movement patterns, allowing to optimize public services in the city. In the specific case the dissertation has the objective of displaying a tangible application of data science methodology to the urban planning realm.

The paper presents a composite statistical index, the youth (18-29 years old) economic discomfort (YED) index, using social, economic and demographic information to support data-driven urban planning. The testing ground is represented by the city of Rome but the methodological approach aims at introducing a prototypical strategy reproducible on other urban contexts.

II. The Dataset

In order to obtain a composite index to describe the

Iacopo Testi is with the University of Washington Rome Center, Piazza del Biscione 95, RM 00186 Italy (e-mail: iacopo.testi@gmail.com).
Diego Pajarito is with the Institute for Advanced Architecture of Catalonia IAAC, Barcelona, BC 08005 Spain
Nicoletta Roberto is with the Municipality of Rome, Piazza di Cinecittà, 11, RM Italy.
Carmen Greco is Partner within the company Greco Real Estate by Greco Mario S.N.C, Belvedere Marittimo, CS 87021 Italy.
All authors are with University of Rome Tor Vergata, Master in Data Science, Information Engineering Department, Via del Politectico,1 Rome.

phenomenon of YED, different datasets containing information on the socio-economic conditions of 18-29-year-olds in Rome City have been identified.

A. Data Search and Selection

The first phase of the research is to identify which kind of data can describe our phenomenon and then identify the ones relevant and available. First screening is to inspect information from public official sources, such as National Institute of Statistics (ISTAT) and public administrations. Otherwise data are selected from other sources, such as private entities or websites.

The second phase is to select data to verify the degree of complexity and accessibility of the sources. For this purpose, a preliminary check list is evaluated to adopt the appropriate means for data collection in terms of costs and time constraints.

B. Data Selected and Individual Indicators

Among the big amount of data available, six datasets have been selected to construct elementary indices (Table I). Datasets collected do not respect temporal homogeneity (span from 2011 to 2019); however they refer to the same decade and it possible to assume that could be compatible for the purpose of this work.

From each dataset shown in Table I, six elementary indices are obtained: (i) the percentage of young people, (ii) the gross monthly income (in range 18-29 years) in euro, (iii) the mean price of a monthly room rental in euro; (iv) the percentage of social housing, (v) the superficial density of restaurants and cafes in counts per km^2 (vi) the unemployment rate (express in percentage). Each index is a vector containing 155 values one for each urban zone (UZ) of Rome. Details on data process to obtain each single index are presented in the next section. The matrix, composed by the six vectors of simple indicators per 155 UZ record, is built up and saved as .csv (comma separated values). Finally, to calculate the composite index, a code is written by using the Python programming language, which has the big advantage to be equipped of numerous free libraries. In this work the following libraries are used for calculation purposes: *pandas, pyplot, numpy, mathplotlib, math, seaborn* (see Appendix).

TABLE I
DATASET SELECTED FOR COMPOSITE INDEX

Dataset	Year	Source
Population (18-29 years)	2018	Statistics Office – Registry source - Rome Municipality
Gross Incomes (18-29 years)	2016	Statistics Office Rome Municipality Siatel data - Agenzia delle Entrate – Economic Resources Department
Room rental price	2019	Scraping on websites: idealista.it; immobiliare.it
Social housing	2011	Processing on Rome Municipality data and ATER housing Istat
Cafes, Restaurant	2019	Open data portal Rome Municipality Id dataset: c_h359:D.757
Unemployment	2011	ISTAT

III. Data Processing

Since data are collected from different sources, a specific data processing is applied to each dataset obtaining homogenous and sensible data.

To acquire room price, data are collected scraping real estate websites to compile a nonprobability sample. In this case data need to be interpreted since they are conditioned to the language of ads posted. Particular attention is paid to build an accurate processing to calculate geolocation of the rooms. From scraping websites information, the address of each ad is collected and it is converted into latitude and longitude coordinates via an API service such as OpenCage (geocoding of 2500 locations per day are free). Using QGIS software and shape files of the UZ of Rome (source Open Data of Roma Capitale) it is possible to geolocate and calculate the mean of room prices of all ads within each UZ. The same process of converting addresses to geographic coordinates is also applied to find the position of cafeterias and restaurants in Rome. The dataset available contains all the commercial activities in the capital and the associated addresses. After filtering only cafes and restaurants, addresses for each record is converted in latitude and longitude coordinates and assigned to each UZ (Fig. 1). Lastly, the percentage of the commercial activities extracted per km^2 is obtained summing all records in the same UZ and dividing by the UZ surface.

Another important phase of data processing is devoted to filtration of data. It mainly consists of elimination of repetitions "nan" values, and incomplete data. Furthermore, datasets having a spatial resolution greater than UZ are selected, like social housing and gross incomes. In fact, information for these two refers to municipal area that includes several UZ. In these cases records for each UZ are replicated for all UZs belonging to the same municipal area. After data processing each dataset is saved as *.csv* file ready to be analyzed.

Fig. 1 Mapping Rome city divided in UZ. Black dots are geolocations of cafes and restaurants

IV. Methodology

A. Introduction to the Methodology Used

To build the YED index the following steps are applied:

1) *Development of a theoretical framework*: at first, It is defined the economic discomfort index among youth aged between 18 and 29 years.
2) *Elementary indices selection:* We selected six elementary indices (see Section II *B*), depending on both theoretical

and analytical relevance. The index polarity compared with the analyzed phenomenon allowed us to evaluate the actual concordance or discordance of direction.

3) *Multivariate analysis:* It is defined as the relationship among single elementary indices and it evaluated their suitability to describe YED.
4) *Imputation of missing data:* It is identified as the input matrix of missing data, with the hot deck imputation technique, in order to avoid possible defects in the index definition.
5) *Elementary index standardization:* Three standardization methods are performed, in particular: Relative index, z-scores and MPI [3] in order to obtain three different youth discomfort indices and subsequently evaluate which one displays the best robustness in the description of the phenomenon.
6) *Weighting:* Following a subjective approach, it is attributed to each elementary index a weight equal to 1.
7) *Aggregation:* With the aim of reducing the multiplicity of the indices and subsequently obtain one unique vector, two different approaches are followed: one using arithmetic mean, by which is synthesized the standardized relative and z-scores matrices, and the other one using the latter with penalizing function (MPI) [3].
8) *Robustness analysis:* The most influencing elementary index is defined by reproducing output value for all the adopted construction procedures.
9) *Result presentation and diffusion:* The discomfort index is divided into five classes (high, moderately high, medium, moderately low, low) and with the use of HTML, CSS and JavaScript is developed an interactive mapping tool that reveals the highest rate of youth discomfort for the central areas of Rome and a decreasing one by moving towards the suburbs.

B. Preliminary Analysis

A set of elementary indicators is selected based on phenomenon significance and for their analytics capacity in the observation of correlation levels between them. Except for room monthly rentals and unemployment rate (see Section II *B*), all others simple indicators show a positive polarity, consequently they are in opposite relationship with the nature of the phenomenon (negative polarity).

From a practical point of view, the choice of these indicators must consider both the data availability and their completeness. Particularly, due to the second aspect described, it is appropriate to highlight that our input index matrix, in particular the rental room indicators, showed a lack of data that could cause distortion in the design of the index. In order to solve this obstacle, the hot deck imputation technique has been used. Through the use of the software QGIS, geolocation of the interested urban areas and their adjacent areas has been possible. The monthly rental room data of the adjacent urban areas were known. Successively, the average of the known data was calculated and linked to the lacked data of these areas. An extract of calculation of imputing missing data is shown in Table II. The exploration analysis of the input matrix has been made through the average calculation of the standard deviation, minimum, maximum variation coefficient, quantile, summation of single vector element (see Appendix).

The quantitative analysis of the different elementary index, preliminary selected, has been made in order to analyze the relationship between them. From the figure, that shows the correlation between each elementary index [4], available in the Appendix, it is highlighted the absence of correlation between the different elementary indicators that compose the theoretical framework. On the consequence the lack of redundant information of correlation between them was demonstrated. The assumption implies that the selected components are appropriate to the examined phenomena description.

TABLE II
EXAMPLE OF MISSING DATA IMPUTATION

UZ	11Y	16X	9C
Price of rental room (adjacent UZ)	450€ (10X)	390€ (16B)	442€ (09B)
Price of rental room (adjacent UZ)	400€ (11X)	476€ (16D)	406€ (10B)
Price of rental room (adjacent UZ)	318€ (12H)	523€ (18A)	400€ (11X)
Mean	389.3€	436€	416€

C. Index Construction

Once identified the components of the theoretical framework, the single index values are aggregated in a composite index to measure the YED.

Firstly, in order to compare the indicators between them, dimensions are neutralized and polarities reversed with opposite direction from the analyzed phenomenon. As previously explained, four elementary indices of the input matrix showed a positive polarity, therefore it was necessary to perform a linear transformation, which consists in a difference between the maximum value, taken from the index, and its each unit, with the advantage of not varying the distance between the different urban areas. This transformation was used for the following standardization methods: relative indices, z-scores [4] and MPI [3].

The first method allowed transforming the value of each unit of the elementary index in a value in the range 0-1. In detail, taken into account an elementary index, i.e. the unemployment rate, the urban areas with a value close to 1 will have a high unemployment rate (worst case) and, on the contrary the urban areas with a value close to 0 will have a low unemployment rate (better case).

The second method allows obtaining new variables with mean value equal to 0 and standard deviation equal to 1, thus making our indices not depending on relative variability. The disadvantage of this method is the poor readability of the result. For example, in the case examined, the urban areas 1A and 1B, that initially have an unemployment rate equal to 7.75% and 12.50%, following the standardization they show a value of -0.383 and 0.962, respectively. These results show that a low and a high unemployment rate are associated to a negative (unfavorable case) and positive (favorable case) values respectively, which is in contrast with the definitive nature of the index.

The third method allowed obtaining new standardized variables with mean value equal to 100 and standard deviation

equal to 10. The advantage of this method is an improvement of the result readability since the standardized values are in a range between 70 and 130 with reference average vector equal to 100. Considering the previous example, the urban areas 1A and 1B show the new standardized values of 96.173 and 109.621 when the unemployment rate is low (favorable case) or high (unfavorable case) respectively.

At the end of the standardization process three new input matrices are acquired. Subsequently, a "weight" system is defined for the weighting of the chosen indices based on their relevance in the description of the phenomenon. Considering a subjective approach and evaluating the results of the multivariate analysis, that shows an absence of the correlation between the different indices, the same weight was established for all indicators equal to 1.

Regarding the synthesis methodology, two different synthesis approaches have been used, the first one uses the arithmetic mean to synthetize the standardization matrices with the relative indices and z-scores, the second one uses the arithmetic mean with the penalizing function (MPI) to synthetize the standardization matrix with the MPI. In the first case, the substitutability of the elementary indices is considered, that is a deficit of one index can be compensated by a surplus of the other. On the contrary, in the second case a non-substitutability between the different components is considered and the compensations are only in part allowed.

In conclusion, Fig. 2 shows the three vectors obtained that will be analyzed qualitatively, with respect to their frequency distribution, whose descriptive analyses are shown in Table III.

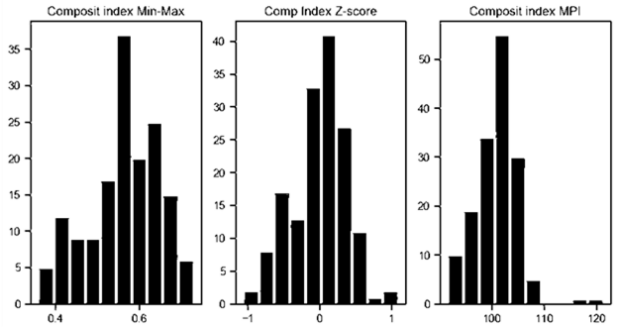

Fig. 2 Histograms of the three youth discomfort indices.

TABLE III
STATISTIC VALUES FOR YED INDEXES CALCULATED BY DIFFERENT METHODS

Statistic Value	Min-max	z-scores	MPI
mean	0.56	0.00	101.02
std	0.09	0.39	4.05
min	0.36	-1.06	91.60
max	0.73	1.10	121.17
25%	0.52	-0.22	98.85
50%	0.57	0.07	101.43
75%	0.63	0.25	103.23

As expected, Fig. 2 shows a distribution that is similar to the Gaussian one, for what concerns the YED indices, obtained by a z-scores and MPI standardization, and a bigger variability of the indices received by the standardization of relative indices. Also, it is clear that the frequency peak is reached for above average values, as shown in the histograms, where higher values than 0.5, 0, or 100 are respectively found. It follows that, more than 40/50 UZ are in economic discomfort youth situation. Afterwards, the influence analysis of the youth discomfort index is performed. This allows to relate, through scatterplot, the composite index with its initial input matrix removing the elementary indicators one by one, in order to observe the indicator with highest influence on the final result. This method is applied on each of the three vectors, following this order of YED index standardized by MPI, z-scores and, finally, relative indices. Therefore, six scatterplots, for each composite index showing the composite index itself on x-axis, and the composite index calculated using input matrix without an elementary indicator on y-axis. So, precisely: the public housing percentage "ATER" (marked in blue); the monthly room rental price (marked in orange); the whole population (marked in green); the monthly income (marked in red); the services (marked in purple); and the unemployment rate (marked in brown). From a qualitative analysis based on correlation index for each scatter plots, it can be concluded that YED index calculated by MPI method, is mostly influenced by the unemployment rate, while composite index calculated by z-scores it is not shown marked influence by an elementary indicator, and the last index calculated by relative index shows, instead, the greatest influence by the average income. In order to decide the composite index to be adopted a subjective approach, made by two steps, is used: in the first one it is evaluated the goodness of fit relative index. In the second one it is excluded the composite index that does not guarantee an easy reading of the results. From the first step it is shown how the index obtained by standardization with relative indices has a less robust elementary indicator, and so, a less goodness of fit than the other two composite indices. Therefore, it does not return a stable index. In the second step, instead, comparing the standardized index in z-scores and the one in MPI, in order to obtain an index that guarantees an easy reading of the results, we excluded the first one, because it gives a low value of youth discomfort, a negative score that contrasts the defining nature of the phenomenon.

V. RESULTS AND DISCUSSIONS

The distribution of YED index shown in Fig. 2 and Table III gives an overall trend of the composite index, however it is interesting understand its spatial distribution over the UZ of Rome. Considering the complex calculated by MPI method the map of its spatial distribution is depicted in Fig. 3. The map shows that medium discomfort affects the Nord West part of the city, and moderately high values are found around the center of the city. Outskirt zones, especially in the South part (close to the coast) of Rome are affected by moderately low and low YED index values. An interactive mapping tool is created to explore results and it is free available at [10]. By querying on each UZ a window containing single indexes and composite index is opened.

Fig. 3 Map of YED index over UZ of Rome

The results mentioned above clearly show that this methodological approach can be used to great effect by urban planning professionals. Admittedly, the composite index presented, might potentially be considered as a decision-making tool for both public and private entities. In particular, based on the discomfort level of Rome different UZ, localization and priority of intervention can be defined. A common example of that are specific social and economic policies that may apply to certain areas and population target of the city. Furthermore, considering exact citizens groups, in need of social and financial assistance, can determine the conception of innovative building typologies, addressing those discomforts. Besides, the indicator provides urban insights that can be further implemented by more granular and accurate analysis. Observing the modus operandi applied, it is intelligible the approach scalability and it is easily predictable how this can impact various other urban layers, such as mobility, energy consumption, pollution and socio-demographics just to mention a few. For instance, a publication produced by the urban planning bureau 300.000 km/s clarifies how urban fabric fosters transfer and innovation in Barcelona. [5] The consistency of the YED index lies also in the synthetic presentation of its outcome. It is indeed fundamental to generate comprehensible visualizations for a not specialized audience in order to communicate effectively the results. Lastly, it is worth mentioning the significance of both the skillset required by the contemporary planners and the necessity of multidisciplinary teams as indispensable qualification to design our future urban areas.

VI. Limitations and Conclusions

After the analysis performed and the research presented, we can confidently state that the limitations encountered were various. Advancing the urban planning sector surely requires close attention to the primary data sources. The datasets, if produced a priori by different stakeholders, might terminate being unproductive. Therefore a sharp scope should drive the data collection and storage, implying a mutual cooperation throughout the entire data workflow, from retrieval to application. The aforementioned process is also undoubtedly linked with the scale and granularity of data, necessary to center the objective of the investigation. An exemplary application of that is the 'Trash I Track' experiment [6], produced by the MIT Senseable City Lab, that displays a case study where 3000 objects, destined to disposal, were implemented with GPS sensors. All the garbage routes and destinations were identified, thus revealing the critical issues of the disposal system, allowing urban planners to potentially intervene in the discarding procedure. In the specific case of this paper, the precision of the YDE index might have been considerably improved if all datasets divided by age at the scale of the UZ where available and statistically robust. Furthermore, an accurate dataset of housing prices and job opportunities would outline a more intelligible picture. Another central topic to examine is certainly the importance of data uniformity across different city departments. The effort of the Rome municipality, especially in the last years, to tailor and gather all information within a unique city data platform, is visible. The recent updates of the Open data portal, the exploration of 5G technologies as enablers of the so-called smart city and the new cartographic infrastructure, are drivers of innovation and digital transformation. Nevertheless, the homogeneity and consistency of data sources between different departments, that we observe in cities such as New York, San Francisco or Barcelona, just to name a few, is still considered a role model to be pursued. Furthermore, in order to achieve fruitful results, the cooperation between city departments and private entities is today indispensable. As a matter of fact, the majority of the citizens information are retained either by global corporates such as Amazon, Google, Facebook, Apple or minor companies like Tim, Wind, TomTom, Mytaxy, etc. For most of those organizations data is a business asset, which explains the limitations in accessibility to their sources. Within those massive datasets, there is multiple information that might potentially reshape urban planning as traditionally practiced. Namely, in 2016 the collaboration between MIT and Uber illustrated the possibility to rethink mobility, as we know it today, with a remarkable environmental impact [7]. Moreover, the alliance between MIT Media Lab, Foursquare and Cuebiq generated 'the Atlas of Inequalities' showing economic inequity and segregation across various American cities [8]. In this scenario municipalities interpret a fundamental role in mediating between public and private interests, improving services for their citizens and simultaneously preserving the companies' returns. Lastly, the study in discussion is to be considered inclusive but not exhaustive in terms of both methodology and data sources. Surely several statistical procedures might be further refined and integrated, such as the imputation of missing data, aggregation approaches, principal component analysis and correlations with other similar indexes. Additionally to the datasets already implemented in the YED index, multiple others might be the objective of supplementary

explorations. Percentage of young people living with their parents, number of job offers per urban area or unemployment proportion for age groups, are just few representative cases of the wide variety of information that might potentially be included in the composite index proposed. Moreover, the elaboration, through its results, displays the possibilities not only to evaluate the current scenario but also to assess eventual urban interventions, like policy making or different building typologies, to be included in further studies. In conclusion, the research's scope is to demonstrate an innovative and contemporary approach to urban planning through data science and statistical procedures that refers back to the principles outlined in the General Theory of Urbanization written in 1867 by Ildefonso Cerdà [9].

APPENDIX

The workflow written in Python and markdown code, that includes the complete analysis related to the YED starting from the vector of elementary indicators is reproducible and available at the following github repository: https://github.com/IacopoTesti/Spatial_Data_Science_Rome.

ACKNOWLEDGMENT

All authors would like to thank the Municipality of Rome for the data provided by the Open Data Portal, the professor of the Tor Vergata Master in Data Science, Matteo Mazziotta (ISTAT – National Institute of Statistics) for the support as a supervisor for the construction of the composite index. Furthermore the author Iacopo Testi wishes to thank all the members of the organization IaaC (Institute for Advanced Architecture of Catalonia), with a specific mention of the coordinator of the Master in City and Technology Alex Mademochoritis and the doctor Diego Pajarito for all the precious insights provided during the construction of this work.

REFERENCES

[1] United Nations, *World Population Prospects*, 2019
[2] X. Liu, Y. Song, K. Wu, J. Wang, D. Li, and Y. Long, "Understanding urban China whith open data", *Cities*, vol. 47, pp. 53-61, 2015
[3] M. Mazziotta, A. Pareto, "Un indice sintetico non compensativo per la misura della dotazione infrastrutturale: un'applicazione in ambito sanitario", *Rivista di Statistica Ufficiale*, vol. 1, pp. 63-79, 2011.
[4] OECD, *Handbook on Costructing Composite Indicators Methodology and user guide*, OECD Publications, pp. 63-65:83-86-102, Paris 2008.
[5] Santamaria-Varas, M. and Martinez-Diez, P. (2016) "How urban fabric fosters knowledge transfer and innovation: the example of Barcelona", 51st International Society of City and Regional Planners Congress: Cities Save the World: Rotterdam, Netherlands: October, 19-23.
[6] Boustani *et al*., "Investigation of the waste-removal chain through pervasive computing" in *IBM Journal of Research and Development*, vol. 55, no. 1.2, pp. 11:1-11:11, Jan.-March 2011.
[7] M. M. Vazifeh, P. Santi, G. Resta, S. H. Strogatz, and C. Ratti "Addressing the minimum fleet problem in on-demand urban mobility", *Nature*, pp. 534-538, 2018.
[8] Xiaowen Dong, Alfredo J. Morales, Eaman Jahani, Esteban Moro, Bruno Lepri, Burcin Bozkaya, Carlos Sarraute, Yaneer Bar-Yam, Alex Pentland, "Segregated Interactions in Urban and Online Space", *Physics and Society*, 2020.
[9] I. Cerdà, "General Theory of Urbanization", 2018.
[10] https://iacopotesti.github.io/Spatial_Data_Science_Rome/