Zooming into Socio-economic Inequalities: Using Urban Analytics to Track Vulnerabilities - A Case Study of Helsinki

Ylenia Casali
TU Delft
y.casali@tudelft.nl

Nazli Aydin
TU Delft
n.y.aydin@tudelft.nl

Tina Comes
TU Delft
t.comes@tudelft.nl

ABSTRACT

The Covid19 crisis has highlighted once more that socio-economic inequalities are a main driver of vulnerability. Especially in densely populated urban areas, these inequalities can drastically change even within neighbourhoods. However, conventionally such vulnerabilities are analysed at city or district scale. As such, new methods with higher granularity are needed to zoom into the spatial patterns locally.

Machine learning techniques enable us to extract detailed spatial information from geo-located datasets. In this paper, we present a prototypical study that uses Principal Component Analysis (PCA) to analyse the distribution of labour and residential characteristics in the urban area of Helsinki, Finland. The main goals are twofold: 1) identify patterns of socio-economic activities, and 2) study spatial inequalities and related vulnerabilities. Our analyses use a grid of 250x250 meters that covers the whole city of Helsinki, thereby providing a higher granularity than the neighbourhood-scale.

The study yields four main findings. First, the descriptive statistical analysis detects inequalities in the labour and residential distributions. Second, relationships between the socio-economic variables exist in the geographic space. Third, the first two Principal Components (PCs) can extract most of the information about the socio-economic dataset. Fourth, the spatial analyses of the PCs identify differences between the Eastern and Western areas of Helsinki, which persist since the economic crisis in the 1990s, indicating clear path-dependencies. Future studies will include further datasets related to the distribution of urban services and socio-technical indicators.

Keywords

Inequality, socio-economic patterns, vulnerability, PCA, GIS, urban analytics, Helsinki.

INTRODUCTION

Technology and globalization have shaped our cities at an unprecedented pace. Socio-economic changes have always defined the growth of cities and the distribution of resources in urban areas. Those changes, essentially, create differentiation in the spatial and social structure of a city and lead to urban segregation, social inequalities and diverging development, posing a threat to social cohesion and stability (Tammaru et al., 2016. Musterd et al., 2017). At the same time, inequalities influence the vulnerability and resilience to crises and disasters (Donner et al., 2008). Therefore, if urban planning and crisis management authorities would like to reduce vulnerability and identify potential hotspots before a crisis strikes, understanding the spatial distribution of inequalities and identifying hotspots are of paramount importance.

Traditionally, indicator-based studies to analyse vulnerability are only available at the level of a city (e.g., Tapia et al., 2017) or different infrastructures (Comes & Van de Walle, 2014). Urban dynamics are conventionally analysed by mapping the geographical distribution of households in urban areas using surveys or census data (Morales et al., 2019), which are extremely cumbersome to collect. As such, what is needed is a method that makes use of the pervasive data that is accessbile publicly and openly that provides high spatial granularity below neighborhood scale.

In this paper, we propose using Machine Learning (ML) techniques for urban analytics to shed a new light on socio-economic urban vulnerabilities with unprecedented accuracy and depth, complementing other methods. It is undeniable that ML studies have helped unveiling patterns in urban data. Yet some of these algorithms are difficult to interpret (Mohanty and Vyas, 2018). ML studies often do not link algorithms with the highly interdisciplinary problems of urban planning and vulnerability (Grekousis, 2019). Therefore, we aim to contribute to advancing the state of the art of urban ML by detecting socio-economic inequalities and their spatial distribution via an analysis of changes in the principal components (PCs). The results can help decision-makers to identify vulnerable areas and track inequalities at a high spatial resolution.

We showcase our approach by using the city of Helsinki as our case study area. In the 1990s, Helsinki transitioned to a new phase of economic development, resulting in a new distribution of the labour market, and thereby creating socio-economic differences between the eastern and western areas of the city (Vaattovaara and Kortteinen, 2003). With this prior information on hand, our study analyses whether those socio-economic inequalities between the east and the west can still be detected, and how the socio-economic conditions are distributed. In sum, the purpose of this study is to 1) investigate how principal component analysis (PCA) can identify patterns of inequality in the socio-economic data of Helsinki in 2016, and 2) examine how spatial changes of PCs identify areas with the highest concentration of values, which represent different socio-economic attractiveness levels.

BACKGROUND

Recently, there is an increase in crisis and risk management studies that use machine learning for various purposes, such as predicting flood risk maps of cities (Eini et al., 2020. Darabi et al., 2019); improving the delineation of flood areas from satellite data (Palomba et al., 2020); assessing damages due to wildfires (Oliveira et al., 2017); studying the role of social factors in hurricanes damages (Szczyrba et al., 2020); identifying the factors that affect the vulnerability of communities during drought, flooding, illness and crop-disease events (Knippenberg et al., 2019); monitoring influenza outbreaks using social media (Allen et al., 2018); or predicting patient volumes at mass gatherings (Serwylo et al., 2011).

In a nutshell, there are four main categories of ML algorithms: (i) supervised (data is labeled), (ii) unsupervised (data is unlabeled), (iii) semi-supervised (only a small portion of data is labeled for training) and (iv) reinforcement learning (data is analyzed and labeled on the flight). Selecting the ML algorithm depends on the problem properties, such as prediction, classification or clustering (Grekousis, 2019). In this study, we used an unsupervised ML technique to derive information from data without imposing any label on the classes of data.

We focus here on Principal Component Analysis (PCA), which was first introduced by Pearson (1901) and Hotelling (1933). Although the literature on PCA spans many disciplines, the use of PCA for the analysis of vulnerability is limited. In 2008, Cutter and Finch investigated the social vulnerability of the US population to cope with hazards by extracting PCs of demographic data in different years. Then, Holand and Lujala (2013) studied two social vulnerability indexes in the main municipalities of Norway. Stafford and Abramowitz (2017) compared the PCA and k-means clustering method by studying the social vulnerability due to sea-level rise in a metropolitan area in Virginia. In 2020, Dong et al. characterized vulnerability against urban flooding by analysing PCA in neighbourhoods. They used PCs to create a disruption tolerance index based on census data, which identified the accessibility to healthcare services during floodings. There are more applications in urban studies, where PCA helps to investigate socio-economic inequalities. Moser and Scott (1961) published a prime example of the social-economic urban inequalities between British towns by using PCA analyses. Lalloue' et al. (2013) investigated social health inequalities in metropolitan areas in France and used principal components (PCs) to extract socio-economic indices at neighbourhood scales. Wang and Zhang (2017) employed PCA to examine the social inequality of public leisure space provision in Shenzhen. Reades et al. (2019) studied gentrification in London by measuring socio-economic status with PCs.

The advantage of using PCA is that it offers an alternative to the otherwise subjective variable selection by objectively simplifying a large number of variables into a few uncorrelated factors. In this way, the factors that influence vulnerability are identified on a more elevated level than when represented by individual variables. Furthermore, the method circumvents the problem with multicollinearity and facilitates increased comprehensiveness (Holand and Lujala, 2013). Moreover, Stafford and Abramowitz (2017) discussed the advantages of PCA compared to k-means clustering method in vulnerability analyses. They concluded that one of the strengths of the PCA is to provide a way of ranking communities in terms of vulnerability which cluster analyses do not. In particular, indices derived by PCA are more appropriate than cluster analyses for academic research where the ability to provide a continuous measure of vulnerability is important (Stafford and Abramowitz, 2017).

While as such PCA has been successfully applied to a wide range of topics and data types (Joliffe and Cadima, 2016), many studies focused on using PCA at highly aggregated levels, i.e., neighbourhoods or city units. What is missing, however, is an application of PCA for detailed analysis at a higher spatial resolution. Analyses at a spatial scale lower than neighbourhood can increase the granularity and contribute to developing site-specific analyses to help urban risk management.

Moreover, most applications of PCA do not discuss the relationship with spatial attributes. Geospatial data distinctively characterizes the relations between socio-demographic structure and a corresponding spatial location in the urban environment. Therefore, analyzing the PCA within the spatial context will add to a more comprehensive understanding of the dynamics of urban areas than purely statistical analyses. These results can help urban planners and risk managers to identify the vulnerabilities of urban areas at a more detailed level, and thereby better prepare for crises.

METHODOLOGY

In this study, we analyse the changes in socio-economic data in Helsinki by using principal component analysis (PCA). We chose Helsinki as a case study as it showed different economic and demographic patterns in the western and eastern areas that result from a transition period in the 1990s (Vaattovaara & Kortteinen, 2003). In particular, our study investigated the PCs' changes in the whole city and the western and eastern areas separately.

Our approach allows us to look at the local variability of urban socio-economic variables by using a classification technique that does not require a priori information about the original data. Meaningful information from the dataset is extracted to study the specific urban context without following any prior assumptions or biases. Our approach presents a flexible way to analyze case studies with assumed local variability of data.

DATA

Our study area is the city of Helsinki, Finland, defined by its administrative boundaries. The total area is approximately 213.8 km2 with a population of approximately 628,208 in 2016 (Statistical Yearbook of Helsinki, 2016). The administrative boundaries and district division of Helsinki are obtained in shapefile format by the Helsinki map service of the municipality of Helsinki. Our dataset consists of two sets of statistics data: the enterprises and establishments' statistics (Työpaikkaruudukko tietokuvaukset) and the population statistics (Väestöruudukko) in 2016. Those are openly accessible via the Helsinki Region Infoshare portal (HRI), which is administered by the municipality of Helsinki. Both datasets are on grids of 250x250m cell size in shapefile formats that cover all of Helsinki's built environment.

We use four socio-economic variables to capture the interplay between the labour demand and the implications for residential areas and housing. In 2019 Reades et al. also used four variables to measure socio-economic status by using PCA. Previous studies used census data to study social vulnerability with PCA (Cutter and Finch, 2008. Dong et al., 2020). Although census data represent social and economic disparities of residents in cities, they fail to represent business attractiveness. For the labour demand, we chose the enterprises and establishments' statistics which provide information about the number of enterprises and the number of employees at each grid cell. The number of enterprises shows the number of business and commercial activities, while the number of employees shows the number of people working in a specific grid cell. The main difference between the two variables is that the number of employees is related to the number of commuters in an area, while the number of enterprises to its local economy. To study the social and residential implications, we selected the population statistics containing the number of residents and the total size of living space in the area unit. The latter is a proxy for the number of accommodations since the living space size and the number of accommodations increase proportionately in a specific area.

Therefore, we chose the four variables that represent the distribution of residential and employment extent in Helsinki. A location with high residential and employment values represent busy hotspots where people live and work within the urban area. Low values show less developed areas in terms of socio-economic activities. This means that areas with high socio-economic values attract more social and economic activities. Lower socio-economic values, in turn, represent higher vulnerability in the same context.

Principal Component Analysis

The Principal Component Analysis (PCA) is a statistical technique used as an exploratory tool for data analysis. The purpose of PCA is to reduce the dimensionality of a dataset, while preserving as much statistical information, or variability, as possible. This means identifying new variables that are linear functions of those in the original dataset, which maximize the variance and are uncorrelated with each other (Jolliffe and Cadima, 2016). These new variables are called principal components (PCs). The higher the degree of correlation among the original variables in the data, the fewer components required to capture common information (Vyas and Kumaranayake, 2006).

The original dataset is defined by the data matrix X, which is formed by p vectors $x_1, x_2, \ldots, x_p$ of n observations. Linear combinations are given by $\sum_{j=1}^{p} \quad a_j x_j = Xa$, where a is a vector of constants $a_1, a_2, \ldots, a_p$ and $a'a = 1$. The weighting factors a_j are selected such that the variance of Xa is maximized. This means to maximize $a'Sa - \lambda(a'a - 1)$, where S is the covariance matrix of X and λ is a Lagrange multiplier. Differentiating with respect to the vector a, and equating to the null vector, produces the equation $Sa = \lambda a$ (Joliffe and Cadima, 2016). The elements of the eigenvectors a_k are called PC loadings, and the elements of the linear combinations Xa_k are called PC scores. The PC loadings indicate the effect of each original variable on the new principal component. The PC scores are the values that each element of X would score on a given PC.

In spatial data analysis, PC scores can be mapped as they correspond to each vector of observations at each spatial location of the data set. Raster PC maps are often used to produce composite indices that describe a certain subset of data with particular properties (Demšar et al., 2012). The PC score maps show the distribution of values of each principal component at each location reflecting the combination of several variables. Hence, variations in spatial distributions of scores indicate the existence of different patterns in the data.

Analytical Framework

For this project, our analyses contain the following three main steps (see Figure 1). First, we pre-processed the original data (labour and residential shapefiles) by using ArcGIS software. To avoid data gaps, we select only the cells that are in common to all four socio-economic variables by using the built-in ArcGIS selection tool. Each cell has a specific ID that we use to join two grid shapefiles. In this way, we create a unique Helsinki shapefile.

Second, we evaluate the PCA on the processed data in Matlab by importing the attribute table of Helsinki shapefile that is exported as a text file to build a matrix. Each column of the matrix represents one of the four socio-economic variables that we use and an additional column to store the cell IDs. We analyze the correlations between the variables. Then, we standardize the data for the PCA by multiplying for the mean and dividing every column by the standard deviation. This standardization is necessary for the input data to be independent of their respective scales. After conducting the PCA analysis in Matlab, the scores are saved as text files with each score result associated with the corresponding cell ID. This allows us to easily integrate the results back into the ArcGIS by joining cell IDs with the existing Helsinki cell shapefile.

Third, we conduct the spatial analysis in ArcGIS with the resulting PCA scores. In order to compare the results of different PCs, all scores are rescaled to the interval [0, 1] such that maximum PC score is mapped to 1 and the minimum to 0. Then, we classify the values in five classes by using the Jenks Natural Breaks method (Jenks, 1967), which is a data clustering method to derive the best arrangements of values into classes. The classified results are represented as maps in ArcGIS both for the city as a whole as well as for the Eastern and Western parts, respectively.

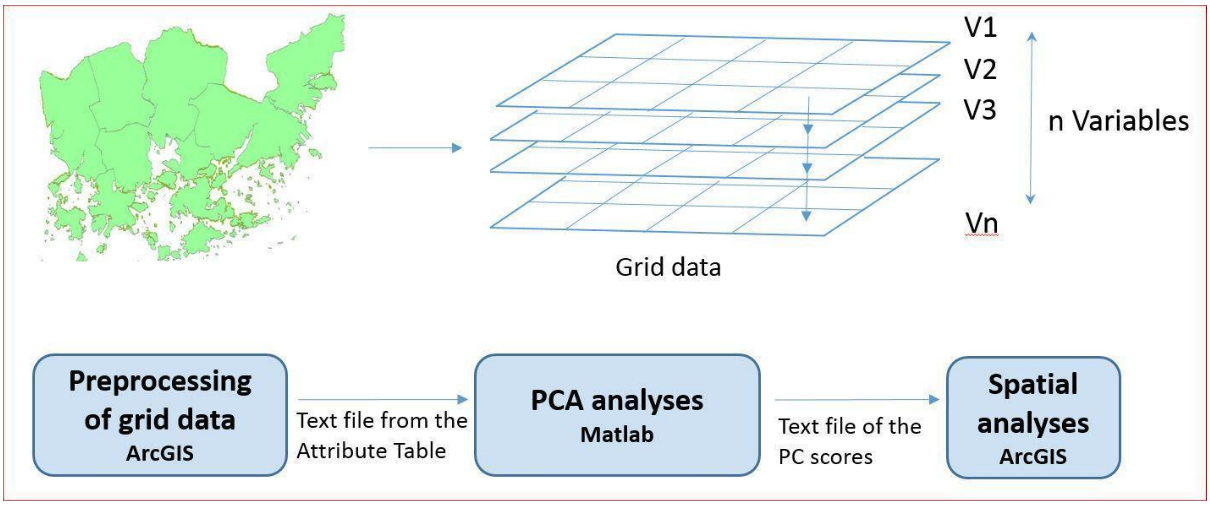
V1
V2
V3
n Variables
Vn
Grid data
Preprocessing of grid data
ArcGIS
Text file from the Attribute Table
PCA analyses
Matlab
Text file of the PC scores
Spatial analyses
ArcGIS

Figure 1. Workflow diagram. We used ArcGIS for processing the spatial data and Matlab for the PCA analyses.

RESULTS

Descriptive Analysis

First, we analyse the distribution of the four selected socio-economic variables in Helsinki: the number of residents, the size of the living space, the number of enterprises and employees. Locations with high values in socio-economic variables represent areas that attract activities within the city. We investigate the patterns of the four socio-economic variables by studying their probability distributions, spatial distributions and correlations coefficients. Table 1 shows the basic statistics.

Table 1. Basic statistics of the socio-economic variables.

	Residents	Living space [m2]	Enterprises	Employees
Mean	318	38	22	106
Standard deviation	332	14	51	361
Skewness	2.3	2.6	7.5	7.5

The distributions of the number of enterprises and employees are more concentrated towards small values than in the residents and living space distributions. Figure 2 shows the distributions of the four variables. The distributions of the number of enterprises and employees are similar, while the residents and living space variables show different results.

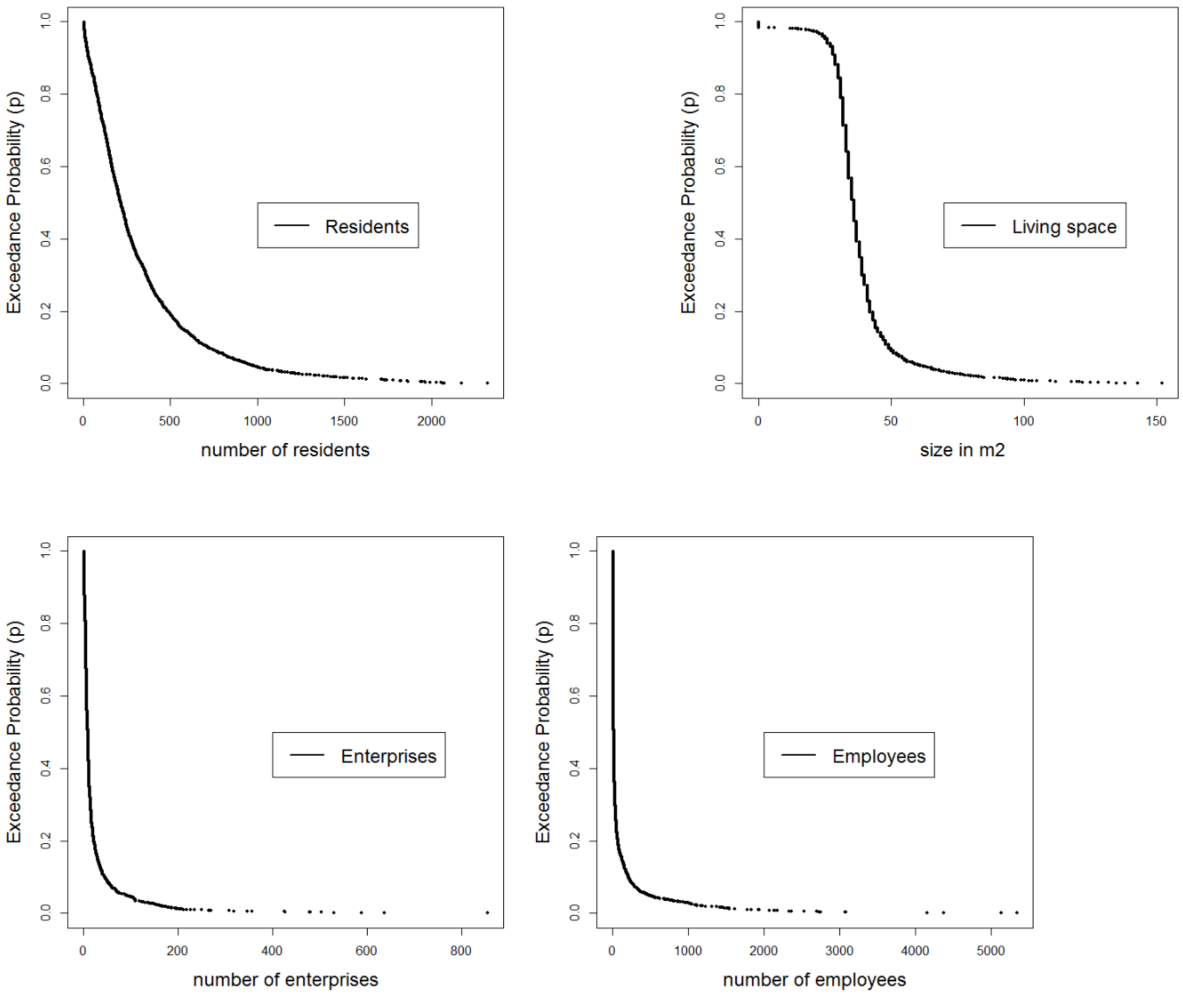

Figure 2. Distribution of four socio-economic variables (e.g. number of residents, enterprises, employees and size of the living space) of Helsinki urban area in 2016.

We classify the socio-economic variables into five categories with the Jenks Natural Break classification method, which is a built-in ArcGIS classification to create thematic maps. The resulting classification yields three patterns of spatial distribution. Figure 3 shows the resulting spatial distributions.

For the residential aspects, the first pattern corresponds to the spatial distribution of residents in Helsinki. Locations with a high number of residents are situated in the south-western area of Helsinki, close to the city center (Figure 3.A). Sites of medium number of residents are spread in the northern and western areas. The second pattern corresponds to the spatial distribution of the living space (Figure 3.B). Here, the locations of middle values spread in the whole city area, whereas values with the highest and lowest amount of living space size are located in fewer sites. The sites with most per capita living space are found mostly in the southern coastal areas, but only partially corresponding to the areas with few residents.

For the economic aspects, we identified a similar spatial distribution of the number of enterprises and employees (Figure 3.C and 3.D). The highest values are clustered in the core of the south-western area of Helsinki in accordance with the central district, while lower values spread from this area in the neighbouring districts. The lowest values are mostly located in the northern and eastern areas, which forms a larger cluster of low values than for the residential aspects.

Overall, residents, enterprises and employees form larger spatial clusters with high values around specific locations. In contrast, high living space values are more scattered throughout the city. The south-western area forms a socio-economic hub that attracts the highest number of people to live and work. When comparing the spatial distributions of residents and employees, we identify areas with high values of residents and the lowest values of employees. Those areas are located in the northern and the whole eastern districts of Helsinki. This means that those areas are mostly residential with a limited labour attractiveness. As a consequence, residents in those areas might travel for longer commuting time to reach the areas more attractive for companies.

To consolidate these findings, we analyse the correlation coefficients to examine the relationships between the four variables (see Table 2), all of which show p-value < 0.01., indicating statistically significant results. The number of residents is correlated with the number of enterprises and employees by a coefficient of 0.4 and 0.1 respectively. Enterprises and employees are strongly correlated by a coefficient of 0.8. Whereas for the living space, we find a slightly inverse correlation with all other variables. These results confirm that enterprises and employees distributions follow similar trends in the city, while the distribution of residents has a stronger relationship with the number of enterprises than the other variables.

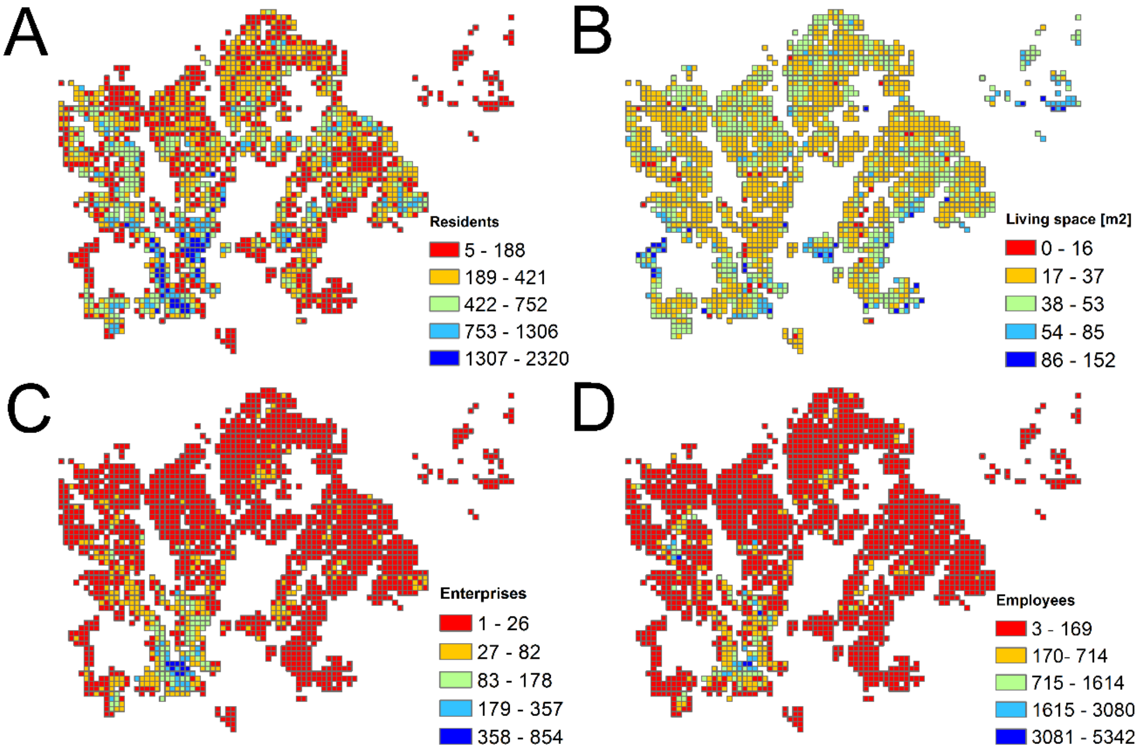

Figure 3. Spatial distribution of four socio-economic variables (A number of residents, B the size of the living space, C number of enterprises, and D number of employees) of Helsinki urban area in 2016. Values were classified into 5 classes by using the Jenks Natural Break classification method.

Table 2. Results of the correlation analysis for the four socio-economic variables.

	Residents	**Living space [m2]**	**Enterprises**	**Employees**
Residents	1	-0.2	0.4	0.1
Living space [m2]		1	-0.1	-0.1
Enterprises			1	0.8
Employees				1

Principal Component Analysis

Results for the Entire Helsinki City Area

The second part of our analysis examines clusters of socio-economic attractiveness levels in the entire city of Helsinki. We apply PCA to extract the underlying structure of the socio-economic dataset. Each principal component is a new variable that is a linear function of those in the original dataset. The first and second PCs account for 76% of the total variance, which implies that the first two PCs alone contain most of the information of the original data set.

The PC's loadings indicate how the original variables contribute to the PCs. We report those loadings values in Table 3. Loadings of the first PC are (sorted in order of importance): 0.7 with the enterprise variable, 0.6 with the number of employees, 0.4 with the number of residents and -0.2 with the living space variable. This is an indication that the number of enterprises has a strong effect on the first PC, followed by the number of employees and residents, while the living space shows the lowest and negative correlation with the first PC. The loadings of the second PC show a different trend (again, in order of importance): 0.8 for the living space variable, 0.4 for the number of employees, 0.2 for the number of enterprises and -0.5 for the number of residents. This means that the living space is the leading variable for the second PC, while the other variables had weaker effects on it.

Table 3. Principal component explained variance and loadings. The loading coefficient explains the correlation between each socio-economic variable and the kth PC. Each column contains coefficients for one principal component, which are ordered of descending component variance.

	PC1	**PC2**	**PC3**	**PC4**
Explained variance [%]	49.6	26.5	19.4	4.5
Residents	0.4	-0.5	0.7	-0.3
Living space [m2]	-0.2	0.8	0.6	-0.1
Enterprises	0.7	0.2	0	0.7
Employees	0.6	0.4	-0.3	-0.6

PC scores are the values that each cell would score on a given PC in Helsinki. Figure 4.A shows the spatial distribution of the normalized scores for the first PC, classified by using the Jenks Natural Break classification method. The highest scores form a cluster located in the south-western area of the city. Scores mostly decrease around this central cluster and in a northerly direction. Indication that the central area was the most important area for the first PC. Figure 4.B illustrates the scores for the second PC, dominated by the living space variable. The highest values are still clustered in the south-western peninsula, but also scattered in the northern and eastern regions in smaller areas. The medium to lowest scores are located throughout the city. This means that the central area still maintained its importance even by changing the PC loadings.

To compare the changes between the Eastern and Western parts of Helsinki, we analyse the number of cells for each class of scores in the East and West (Table 3). The total number of cells in Helsinki is 1944, of these 612 are in the Eastern and 1332 in the Western areas. Then, we calculate the rate of cells in the East and West areas by dividing for the total number of cells in the area for each class, see Table 4.

We find that the number of cells in the 5th class, i.e., the highest values, has the lowest total count in both PCs with 1% in the West and 0% in the East for the first PC, dominated by enterprises and employees. For the second PC, the share of the 5th class increases to 2% in both East and West. Therefore, the first PC is more selective in assigning high values than the second PC.

For the lowest scores, the first PC shows a striking concentration of 64% of the lowest class of scores are in the East of Helsinki (vs 50% in the West). This indicates that areas with less importance in terms of the socio-economic variables are in this part of the city. The second PC, dominated by living space, shows more balanced results in between the East and the West. In particular, the high scores, corresponding to the 4th class, have a higher percentage in the East than the West of Helsinki.

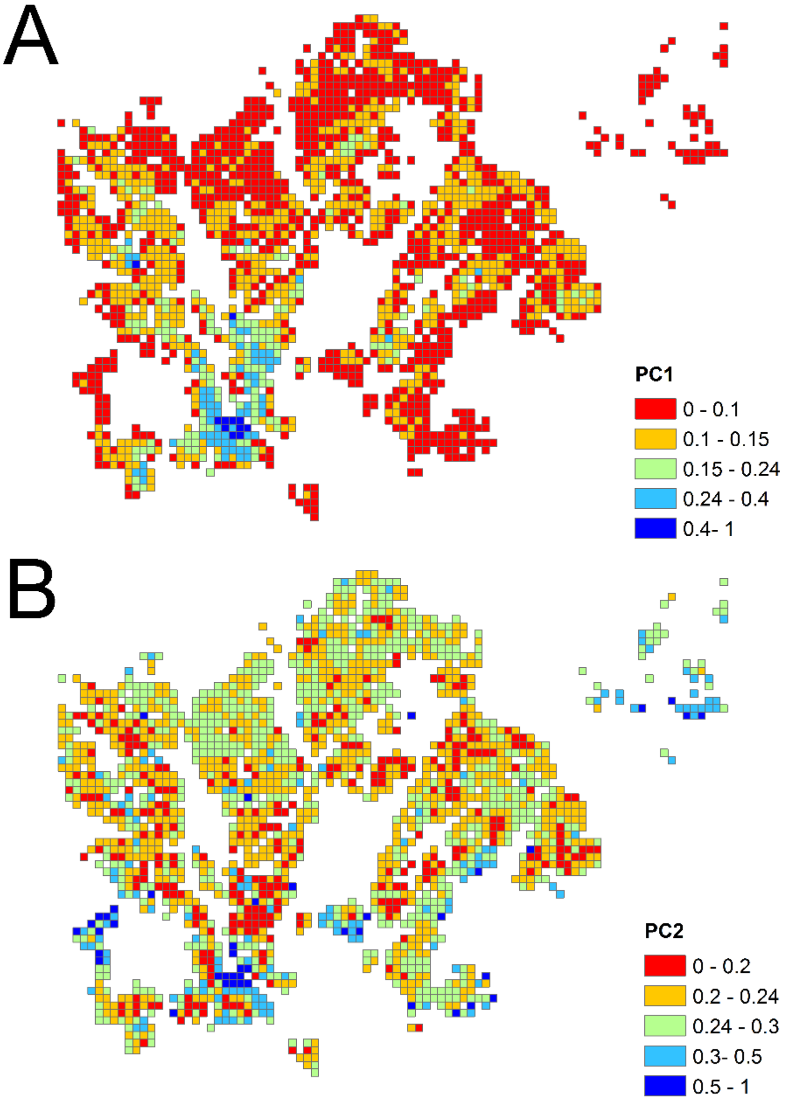

Figure 4. Spatial distribution of the Principal Component (PC) scores in Helsinki. 4.A represents the results for the first PC and the 4.B for the second PC. The values show how much each cell would score on a given PC. High cell values mean a higher attractiveness of the socio-economic variables in that specific location. The scores are normalized to [0,1] and then classified with the Jenks Natural Breaks method. The highest values for both PCs are located in the south-western peninsula, close to the central areas of Helsinki.

Table 4. The number of cells for each class of the PC scores in the East, West and whole Helsinki. The scores are firstly ranged between 0-1 intervals and then classified in 5 classes. We use the Jenks Natural Breaks method to classify the values. The 1st class corresponds to the lowest values, while the 5th class to the highest values. The rate is calculated by dividing the number of cells in a specific class by the total number of cells in each area.

First PC					
Classes	**Whole**	**East**	**West**	**East rate**	**West rate**
1st	1062	392	670	64%	50%
2nd	652	197	455	32%	34%
3rd	149	21	128	3%	10%
4th	70	2	68	0%	5%
5th	11	0	11	0%	1%

Second PC					
Classes	**Whole**	**East**	**West**	**East rate**	**West rate**
1st	307	90	217	15%	16%
2nd	802	238	564	39%	42%
3rd	647	210	437	34%	33%
4th	145	59	86	10%	6%
5th	43	15	28	2%	2%

Results for the West and East of Helsinki Comparison

In the third part of our analysis, we investigate the West and East areas of Helsinki individually. As striking socio-economic differences between the Eastern and Western parts of Helsinki have been found after the 1990s (Vaattovaara and Kortteinen, 2003), we now aim to analyse if those socio-economic differences have prevailed until 2016.

The first and second PCs explain 80% of the total variance in the East of Helsinki and 75% in the West of Helsinki. These results show that the first two PCs store most of the information representing the original socio-economic variables. We analyze the loadings to study which variable most influenced the two PCs. Table 5 shows the loadings for the first two PCs in the East and West areas. The results yield similar trends to the ones for Helsinki as a whole. The living space variable is still inversely correlated in the first PC and then strongly correlated in the second PC. In particular, the loadings in the western areas slightly differ from the Helsinki loadings. The changes are only in the values of the loadings and not in the importance ranking of the variables to PCs.

Table 5. First two principal component loadings of the eastern and western areas analysed separately.

	East		**West**	
	PC1	**PC2**	**PC1**	**PC2**
Residents	0.4	-0.6	0.4	-0.4
Living space [m2]	-0.4	0.6	-0.2	0.8
Enterprises	0.6	0.3	0.7	0.2
Employees	0.6	0.5	0.6	0.3

Then, we analyse the spatial distributions of the PC scores. Figure 5.B and 5.D illustrate the results in the eastern part of Helsinki. In the first PC, the highest scores are located in the central area of the eastern part which are also mainly surrounded by high scores. Whereas, in the second PC (see Figure 5.D), the distribution is similar to the second PC for the entire Helsinki (see Figure 4.B). High scores are located even in the extreme southern of the eastern districts. Figures 5.A and 5.C show the PC scores in the West part of Helsinki. Spatial distributions for the first and second PCs are similar to those found in the whole Helsinki (see Figure 4). A cluster of high scores is located in the south-western peninsula, similarly to the previous results, which confirm the attractiveness of the south-western areas. We notice fewer cells classified with the highest scores in the West part than in the whole of Helsinki.

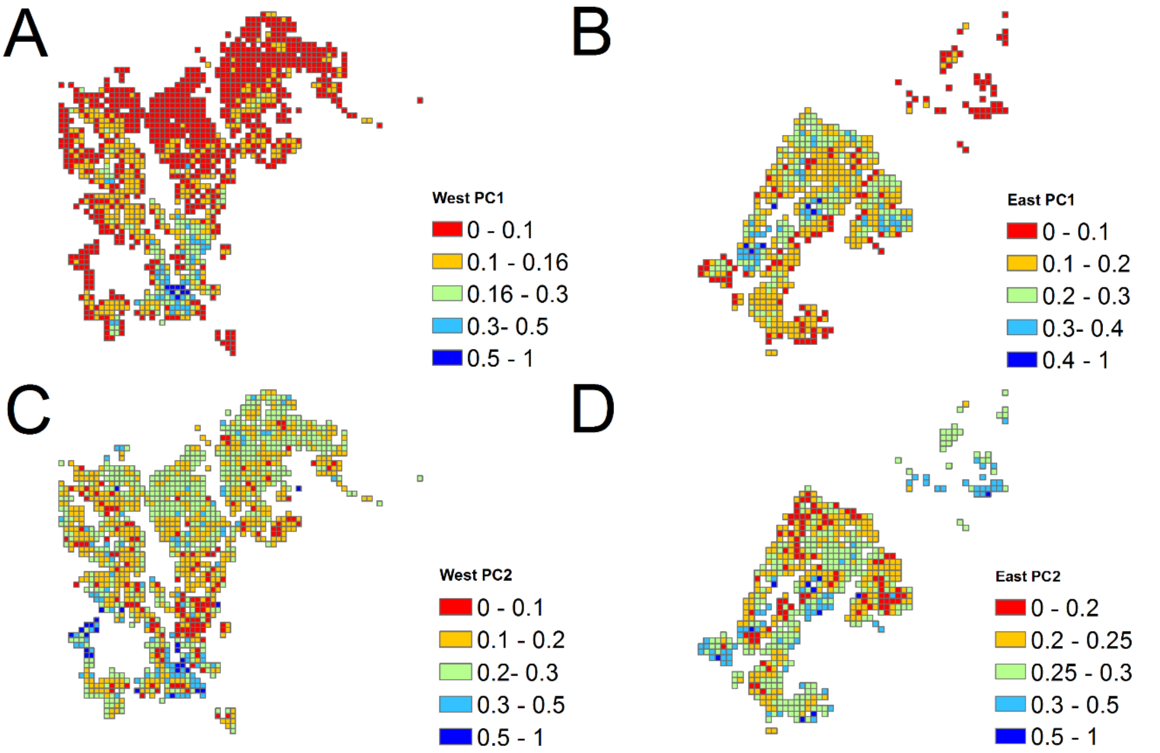

Figure 5. Spatial distribution of the PC scores in the West and East of Helsinki. 5.A and 5.B represents the results for the first PCs, 5.C and 5.D the second PCs. The score values range between 0-1 interval and then are classified using the Jenks Natural Breaks method. High cell values mean a higher combination of the socio-economic variables in that specific location.

DISCUSSION AND CONCLUSION

In this study, we investigate socio-economic patterns and their spatial distributions in the city of Helsinki based on four key variables (number of enterprises, employees, residents and surface of living space). We use PCA to identify the patterns and examine the changes in high PC values in the spatial context at a high level of granularity.

Our study yields four main findings. First, the descriptive statistical analysis provides us with the spatial distributions for four socio-economic variables. We find that the number of enterprises and employees show similar and complementary spatial distributions, while different patterns are observed for the residents and living space. The highest values of the number of enterprises and employees are located in the south-western area of Helsinki, while the highest values of the number of residents and living space are located in different areas of the city. These results confirm the findings of Vaattovaara and Kortteinen (2003), who showed how social differences between the Western and Eastern areas of Helsinki rose during the economic crisis of the 1990s. Even though Helsinki experienced since then an economic upswing by focusing on knowledge-related industries, the spatial patterns from the 1990s are still persistent, even though the city has invested reducing inequalities in the residential sector (Nilamo, 2020). For economic activities, this is in line with the findings of Inkinen & Kaakinen (2016) showing that the number and intensity clusters of knowledge-intensive business services diminish when the distance to the core center of Helsinki increases.

Second, all the four socio-economic variables are correlated with each other. The number of enterprises and employees are strongly correlated by a coefficient of 0.8, "residents and enterprises" and "residents and employees" are correlated by a coefficient of 0.4, and 0.1, respectively. The living space is negatively correlated with all the variables by coefficients between 0.1-0.2. These results confirm that relationships between the socio-economic variables exist in the geographic space.

Third, the first two PCs account for at least 75% of the total variance in all cases (i.e., the entire city area, or the west and east parts separately). Moreover, the PC loadings show similar results in all analyses. The first PCs are positively correlated with the number of enterprises, employees and residents, while the second PCs are strongly positively correlated with the living space.

Fourth, the spatial analyses of the PCs identify patterns of change between the Eastern and Western areas of Helsinki. We find that the high values of PC scores (4th and 5th classes) are mostly clustered in the western peninsula for our analysis of the entire Helsinki city area. The separate analysis results showed that in the Western areas, there is a constant cluster of highly important cells in the south-western peninsula. These findings confirm the observations of Vaattovaara and Kortteinen (2003) that there are differences between the eastern and western districts from a social-economic perspective that still persists, more than 30 years after the economic transformation. The western areas are the most important for residential and labour-related aspects.

Our findings have implications for decision-makers, planners and the scientific community. Decision-makers can use the PCs to identify socio-economic clusters in cities and to aim to address social inequalities, especially in preparedness for or response to crises. Planners can use these highly localized clusters as an indicator of possible vulnerabilities and how they distribute in urban areas. They can identify strategic plans to make communities and neighbourhoods more sustainable and resilient. In particular, planners can use our framework for better preparedness to understand critical vulnerability hotspots and lower the social and economic risk in case of crises. For the scientific community, our study contributes to extending the ML techniques to investigate urban socio-economic patterns and inequalities, which have utmost importance to evaluate and improve the resilience of cities.

For the scientific community, this study contributes to social vulnerability studies, and connect it to the (urban) planning literature. Studying social vulnerability is important because the impacts of hazard events are falling disproportionately on the most vulnerable people in society (Copeland et al. 2020, Howell and Elliott 2018). Moreover, planning scholars maintain that multiple sectors of urban planning (e.g., transportation, conservation, housing) often ignore the uneven impacts of hazards on socially vulnerable populations (Berke et al., 2019. Anguelovski et al., 2016). Our results extend the discussion of Cutter and Finch (2008), who emphasized the importance of using a flexible approach including place-specific local variability while studying social vulnerability rather than a 'one-size-fits-all' approach. Researchers from different disciplines already contributed to foster the discussion around these themes from analysing the urban planning of cities to making surveys and models (Rufat et al., 2015. Anguelovski et al., 2016). Our methodology complements this research by allowing urban planners and crisis managers to extract information at a high granularity from data available, instead of investing in additional data collection. As such, this approach is especially useful under the stress and time pressure that is characteristic for crises (Comes et al., 2020).

This study has some limitations as it serves as a prototypical demonstration of using PCA for urban vulnerability. First, we focus on four key socio-economic variables. Data on income levels, or urban services (e.g., access to education and healthcare), or real estate prices would complement our analysis to improve the detection of demographic vulnerability. Importantly, PCA can extract information from larger datasets. Therefore, future research will focus on using more variables as input to the analyses. Second, we studied socio-economic spatial distribution and ignored the spatial distribution of infrastructures, buildings and land use in Helsinki. Future research will go beyond the socio-economic and include socio-technical and socio-environmental indicator sets (e.g., transportation networks, smart infrastructures; green and blue spaces). To study the distribution of the urban service at a local scale, we identify areas with similar residential density but different average incomes to compare the urban service densities between them.

REFERENCES

Anguelovski, I., Shi, L., Chu, E., Gallagher, D., Goh, K., Lamb, Z., Reeve, K., & Teicher, H., (2016). Equity Impacts of Urban Land Use Planning for Climate Adaptation: Critical Perspectives from the Global North and South. *Journal of Planning Education and Research*, 36(3), 333 –348.

Berke, P., Yu,S., Malecha, M., & Cooper, J. (2019). Plans that Disrupt Development: Equity Policies and Social Vulnerability in Six Coastal Cities. *Journal of Planning Education and Research*, 1 –16.

Comes, T., & Van de Walle, B. (2014). Measuring disaster resilience: The impact of hurricane sandy on critical infrastructure systems. *Proceedings of the 11th ISCRAM Conference*, University Park, Pennsylvania, USA.

Comes, T., Van de Walle, B., & Van Wassenhove, L. (2020). The coordination-information bubble in humanitarian response: theoretical foundations and empirical investigations. *Production and Operations Management,* 29(11), 2484-2507.

Copeland, S., Comes, T., Bach, S., Nagenborg, M., Schulte, Y., & Doorn, N. (2020). Measuring social resilience: Trade-offs, challenges and opportunities for indicator models in transforming societies. *International Journal of Disaster Risk Reduction,* 51, 101799.

Cutter, S.L., & Finch, C. (2008). Temporal and spatial changes in social vulnerability to natural hazards. *PNAS*, 105(7), 2301-2306.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A.S., & McLoone S. (2013). Principal Component Analysis on Spatial Data: An Overview. *Annals of the Association of American Geographers*, 103:1, 106-128.

Donner, W., & Havidán, R., (2008). Population Composition, Migration and Inequality: The Influence of Demographic Changes on Disaster Risk and Vulnerability, *Social Forces*, 87(2), 1089-1114.

Dong, S., Esmalian, A., Farahmand, H., & Mostafavi, A. (2020). An integrated physical-social analysis of disrupted access to critical facilities and community service-loss tolerance in urban flooding. *Computers, Environment and Urban Systems*, 80: 101443.

Grekousis, G. (2019). Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis. *Computers, Environment and Urban Systems*, 74, 244-256.

Jenks, George F. (1967). The data model concept in statistical mapping. *International Yearbook of Cartography*, 7, 186-190.

Jochem, W. C., Bird, T. J., & Tatem, A. J. (2018). Identifying residential neighbourhood types from settlement points in a machine learning approach. *Computer Environment Urban Systems*, 69, 104-113.

Holand, I.S. & Lujala, P. (2013). Replicating and adapting an index of social vulnerability to a new context: a comparison study for Norway, *The Professional Geographer*, 65, 312–328.

Howell, J., & Elliott., J. (2018). As Disaster Costs Rise, So Does Inequality, *Socius: Sociological Research for a Dynamic World,* 4, 1-3.

Inkinen, T., & Kaakinen, I. (2016). Economic geography of knowledge-intensive technology clusters: Lessons from the Helsinki Metropolitan Area. *Journal of Urban Technology*, 23(1), 95-114.

Knippenberg, E., Jensen, N., & Constas, M. (2019). Quantifying household resilience with high frequency data: Temporal dynamics and methodological options. *World Development*, 121, 1-15.

Lalloue', B., Monnez, J., Padilla, C., Kihal, W., Le Meur, N., Zmirou-Navier, D., & Deguen, S. (2013). A statistical procedure to create a neighbourhood socioeconomic index for health inequalities analysis. *International Journal for Equity in Health* 12(21).

Mohanty, S. & Vyas, S. (2018). How to Compete in the Age of Artificial Intelligence: Implementing a Collaborative Human-Machine Strategy for Your Business. APRESS, New York, USA.

Morales, A. J., Dong, X., Bar-Yam, Y., & Sandy Pentland, A. (2019). Segregation and polarization in urban areas. *Soc Open Sci*, 6(10), 190573.

Moser, C.A., & Scott W. (1961). British Towns. A statistical study of their socio economic differences. Oliver & Boyd, Edimbourg and London, UK.

Musterd, S., Marcińczak, S., Van Ham, M., & Tammaru, T. (2017). Socioeconomic segregation in European capital cities. Increasing separation between poor and rich. *Urban Geography*, 38(7), 1062-1083.

Niitamo, A. (2020). Planning in no one's backyard: municipal planners' discourses of participation in brownfield projects in Helsinki, Amsterdam and Copenhagen. *European Planning Studies*, 1-18.

Palomba, G., Farasin, A., & Rossi, C. (2020). Sentinel-1 Flood Delineation with Supervised Machine Learning. *Proceedings of the 17th ISCRAM Conference*, Blacksburg, VA, USA.

Reades, J., De Souza, J., & Hubbard, P. (2019). Understanding urban gentrification through machine learning. *Urban Studies*, 56(5), 922–942.

Rufat, S., Tate, E., Burton, G.C., & Maroof, A.S. (2015). Social vulnerability to floods: Review of case studies and implications for measurement. *International Journal of Disaster Risk Reduction,* 14(4), 470-486.

Serwylo, P., Arbon, P., & Rumantir, G. (2011). Predicting Patient Presentation Rates at Mass Gatherings using Machine Learning. *Proceedings of the 8th International ISCRAM Conference*, Lisbon, Portugal.

Stafford, S. & Abramowitz, J. (2017). An analysis of methods for identifying social vulnerability to climate change and sea level rise: a case study of Hampton Roads, Virginia. *Natural Hazards*, 85, 1089–1117.

Statistical Yearbook of Helsinki 2016 (2016). City of Helsinki. https://www.hel.fi/uutiset/en/tietokeskus/the-statistical-yearbook-of-helsinki-2016. Accessed date: 31 July 2020.

Szczyrba, L., Zhang, Y., Pamukcu, D., & Eroglu, D. (2020). A Machine Learning Method to Quantify the Role of Vulnerability in Hurricane Damage. *Proceedings of the 17th ISCRAM Conference*, Blacksburg, VA, USA.

Tammaru, T., Marcińczak, S., Van Ham, M., & Musterd, S. (2015). Socio-Economic Segregation in European

Capital Cities: East meets West (Regions and Cities). Routledge publisher, 1 ed. UK.

Tapia, C., Abajo, B., Feliu, E., Mendizabal, M., Martinez, J. A., Fernández, J. G., Laburu, T., & Lejarazu, A. (2017). Profiling urban vulnerabilities to climate change: An indicator-based vulnerability assessment for European cities. *Ecological indicators*, *78*, 142-155.

You, H., & Yang, X. (2017). Urban expansion in 30 megacities of China: categorizing the driving force profiles to inform the urbanization policy. *Land Use Policy*, 68, 531-551. doi:10.1016/j.landusepol.2017.06.020

Vaattovaara, M., & Kortteinen, M. (2003). Beyond Polarisation versus Professionalisation? A Case Study of the Development of the Helsinki Region, Finland. *Urban Studies*, 40(11), 2127-2145.

Vyas, S., & Kumaranayake, L. (2006). Constructing socio-economic status indices: How to use principal components analysis. *Health Policy and Planning*, 21, 459–468. https://doi.org/10.1093/heapol/czl029

Wang, Q., & Zhang, Z. (2016). Examining social inequalities in urban public leisure spaces provision using principal component analysis. *Quality & Quantity*, 51, 2409-2420.