Visualization of the spatial distribution of population of Mérida, Yucatán

Gonzalo J. Manrique Arevalo
Data Engineering
Universidad Politécnica de Yucatán
Km. 4.5. Carretera Mérida — Tetiz
Tablaje Catastral 4448. CP 97357
Ucú, Yucatán. México
Email: st1809095@upy.edu.mx

Dr. Gonzalo G. Peraza Mues
Universidad Politécnica de Yucatán
Km. 4.5. Carretera Mérida — Tetiz
Tablaje Catastral 4448. CP 97357
Ucú, Yucatán. México
Email: gonzalo.peraza@upy.edu.mx

Ing. Fís. Didier Omar Gamboa Angulo
Universidad Politécnica de Yucatán
Km. 4.5. Carretera Mérida — Tetiz
Tablaje Catastral 4448. CP 97357
Ucú, Yucatán. México
Email: didier.gamboa@upy.edu.mx

Abstract

At first, the project was focused on the visualization of road networks in the city of Mérida, but in the process, change of course towards the study of INEGI data, the study of the population between the city of Mérida and Kanasín, as well as other important factors such as the economic and educational level. In the study I focus specifically on the study of the total population in the city of Mérida and Kanasín, where python was the main tool together with geopandas, to perform the different data analyzes. The visualizations were taken by the hand with Folium to obtain an interactive view of the maps, and different techniques and mathematical formulas with which we could observe different behaviors in the data, and optimize the visualizations depending on our data, to finally take data from past years to obtain new information and compare how the population in the city evolved and which sectors were the ones that resulted in population growth or decrease.

Index Terms

Quantitative data classification, EDA, Pre-processing Freedman Diaconis, Geopandas, Python, Folium, Choropleth, Density, Distribution, Redistribution

Visualization of the spatial distribution of population of Mérida, Yucatán

I. Introduction

The most important thing about a place are its inhabitants, who make it up with their different languages, traditions and customs that give it the special characteristics that identify it.

But people are not always the same or stay in the same place; they are born, they die, they leave their parents' house, they return, or they are permanently located in others; for those reasons, their number will never be the same from one to another.

The analysis of the factors that guide the location and population patterns of the territory, serves to support the development of public policies that strengthen regional and urban planning, therefore, it is important to know and analyze the geographical, economic, social, and political determinants, etc. that influence demographic dynamics, and that are translated into various realities, ranging from the dispersion of the population in small rural localities, to the marked concentration of large cities and metropolises.

The study of the population is a field of knowledge that has the purpose of providing information on the demographic characteristics of the communities, and their relationships with social, economic and environmental contexts that shape local development processes, regional and national. Conceptual frameworks, data, and population analysis have several applications in social research, such as the inclusion of the problem of the structure and of population change within multidisciplinary social studies; the use of fonts secondary information; and in the identification and projection of socio-economic offers and demands that qualify national, sectoral and regional planning.

It is important to know these figures because that way you can plan what is needed or will be needed (for example, how many schools, hospitals, transport or sources of work require a place).

That is why periodic censuses are carried out in each country, a census Population is precisely that: the population count. In our country, censuses are carried out every ten years.

Other things that allow knowing the censuses are:

A. The distribution of the population

That is, where the people were obtained and how they are distributed in a territory.

B. The population density

It refers to the number of people who inhabit a place and the square kilometer is taken as a reference of space, although in this study the square hectare was taken into account.

II. Objectives

1) The analysis of the official data of the INEGI will be carried out, to later create visualizations of the urban population.
2) The objective of the project is the creation of a board visualization of geospatial distribution of urban and population data of Mérida, Yucatán.
3) The participant must perform a data integration from different sources of information related to urban, population, and economic dynamics. Pre-process the information and create a visualization of the results with different information layers.

III. State of the Art

Over the years it has opened different studies related to population, my main job was to create visualizations that could be useful for various subsequent studies, which could be useful for governments or private organizations with different objectives in mind.

In 2004, a population study was carried out by Victoria Cramer, Svenn Torgersen and Einar Kringlen(Quality of Life in a City: The Effect of Population Density)[9], where there are a series of concepts and operational definitions of quality of life. In the study, the objective has been to develop a global and complete quality of life index, and to relate the sub-indices and the global index with various sociodemographic variables, somatic health and population density in the residential area. The sample consisted of 2,066 individuals between 18 and 65 years of age from the common population. Seven subscripts were developed. They constituted a factor with moderate intercorrelations between the subscripts. Good somatic health, living in a stable relationship with a partner, preferably married, in a less densely populated area, having a good education, a good income, and being a younger woman were the independent statistical determinants of overall quality of life. However, several subscripts were related to different sociodemographic variables. Age was related in an opposite way with different subscripts. The study shows the importance of what type of quality of life is investigated. To our knowledge, this is the first study of the effect of population density on quality of life.

For the year 2013, a study of Quality of life was carried by

Omar Fassio, Chiara Rollero and Norma De Piccoli (Health, Quality of Life and Population Density: A Preliminary Study on "Contextualized" Quality of Life)[10], where Quality of life concerns individual (physical and psychological health), interpersonal (social relationships) and contextual (environment) aspects, which are both subjective and objective. In considering contextual characteristics, empirical findings have demonstrated that people's relation to their living environment is a key issue for their well-being. However, until now literature has paid little attention to population density as an element affecting quality of life. The present study aimed at assessing the predictive role of population density on the several domains of quality of life, along with socio-demographic characteristics and physical diseases. Participants were 344 subjects living in the Northern Italy area. A questionnaire with WHO Quality of Life Brief Scale, a checklist of chronic diseases and a socio-demographic form was used to collect data. Results showed that population density influences psychological, relational and environmental quality of life. Theoretical and policy implications are discussed.

And as we can see, a strong point of population studies is in contexts of the well-being of the population, as well as economic and educational factors, which we can study in a more concrete way through analysis and visualizations carried out by the hand of scientists. data, experts and amateurs in the area, which with a little more study and research can be taken to a higher level.

IV. Methods and Tools

During the development of the project, different methods and tools were implemented for the analysis and visualization of the INEGI data. For the development of the visualizations and the dashboard, several courses [1 to 5] were needed, and previous knowledge acquired during the data engineering career, such as data pre-processing, to perform EDA's, and different classification techniques for the optimization of the visualization of maps. The data set used for this project were extracted from the Geostatistical Framework of the official INEGI website[6], and from the 2020 and 2010 Population and Housing Census data, from the INEGI official website[7], which were later treated by data pre-processing techniques for optimal use.

The tools needed to make this project are the following:

1) One Computer(Windows OS)
2) Python 3.8.5
3) Jupyter Lab
4) GitHub
5) Microsoft Teams
6) Discord

Our main IDE was Jupyter lab, using the Python 3.8.5 kernel, and the libraries that were used are the following:

1) Matplotlib
2) Contextily
3) Geopandas
4) Pandas
5) Pysal
6) Numpy
7) Mapclassify
8) Seaborn
9) Folium
10) Pyproj
11) Requets

Also, once having the data sets ready for the visualizations, the data went through different diagrams and tests to have the correct parameters according to the data we had.

V. Development

In order to successfully develop this project, it was necessary sary to establish the parts that were to be elaborated, below are the subsections of the project to be developed

A. Data Cleaning of 2010 INEGI data

Thanks to the previous work of colleague Alejandro Puerto Castro, I had the Shapefile ready to work, corresponding to the year 2020, so for the year 2010, I had to carry out all the pre-processing of the datasets and shapefile of the INEGI so that it could be handled in a optimal way. As mentioned, the data was provided by the official pages of INEGI. This task was quite tedious since the data provided by the government is usually very dirty and difficult to handle, so it requires a lot of work.

	CVEGEO	CVE_ENT	CVE_MUN	CVE_LOC	CVE_AGEB	geometry	nom_ent	nom_mun	nom_loc	pobtot
0	3100100010107	31	001	0001	0107	POLYGON ((2759997.300 2436729.986, 2759993.895...	Yucatán	Abalá	Total AGEB urbana	779.0
1	3100100010111	31	001	0001	0111	POLYGON ((2760549.860 2436839.396, 2760669.221...	Yucatán	Abalá	Total AGEB urbana	1066.0
2	3100100010126	31	001	0001	0126	POLYGON ((2759643.636 2435989.581, 2759623.462...	Yucatán	Abalá	Total AGEB urbana	15.0
3	3100100010130	31	001	0001	0130	POLYGON ((2760229.363 2437069.166, 2760235.194...	Yucatán	Abalá	Total AGEB urbana	6.0
4	3100100010145	31	001	0001	0145	POLYGON ((2760125.637 2436981.633, 2760125.466...	Yucatán	Abalá	Total AGEB urbana	7.0

5 rows × 199 columns

Fig. 1. Dataframe of 2010 year

B. EDA of 2020 and 2010 datasets INEGI

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.
In this case, it was very useful to be able to know the correct parameters necessary for the realization of the map visualizations. Two necessary points to keep in mind were:

1) Freedman Diaconis Rule: For a set of empirical measures sampled from some probability distribution, the Freedman-Diaconis rule is roughly designed to minimize the integral of the squared difference between

the histogram (that is, the relative frequency density) and the density of the theoretical probability distribution.

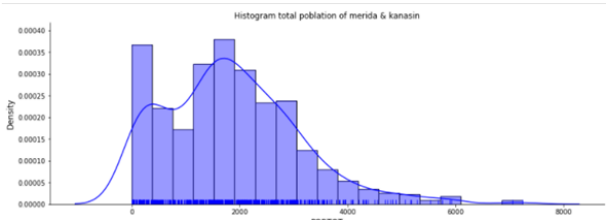

Fig 2. Histogram total population Mérida and Kanasín

2) Quantitative data classification: Data classification considers the problem of dividing attribute values into complete and mutually exclusive groups. The precise way in which it is done will depend on the measurement scale of the attribute in question. For quantitative attributes (ordinal, interval, ratio scales), the classes will have an explicit order.
The different classification schemes are derived from your definition of class boundaries. The choice of the classification scheme must take into account the statistical distribution of the attribute values.

As a special case of clustering, the definition of the number of classes and the class boundaries pose a problem to the map designer. Freedman-Diaconis rule was to be optimal, however, the optimally necessitates the specification of an objective function. In the case of Freedman-Diaconis, the objective function is to minimize the difference between the area under estimated kernel density based on the sample and the area under the theoretical population distribution that generated the sample.
This notion of statistical fit is an important one. However, it is not the only consideration when evaluating classifiers for the purpose of choropleth mapping. Also relevant is the spatial distribution of the attribute values and the ability of the classifier to convey a sense of that spatial distribution.

For map classification, one optimally criterion that can be used is a measure of fit. In PySAL the “absolute deviation around class medians”(ADCM) is calculated and provides a measure of fit that allows for comparison of alternative classifiers.

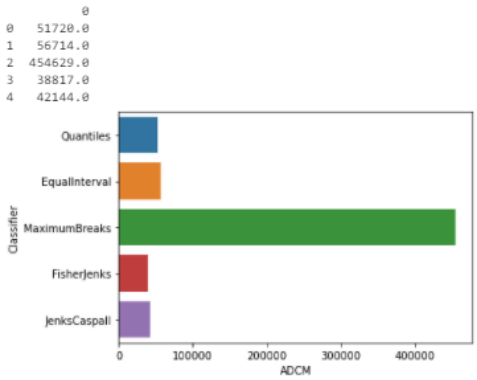

Fig. 3. Comparison of different classifiers

This Comparison of different classifiers was carried out to know which schema parameter would be the best to use, depending on the dataset used for the visualizations of choropleth maps, for this case the data from the 2010 and 2020 INEGI census .
As expected, the Fisher-Jenks classifier outnumbered all other k = 19 classifiers with an ACDM of 38,817. The jenks-caspall classifier works fine. The Maximum break classifier has a very poor setting.

C. Population density 2010 2020

Population density, sometimes also called relative population, refers to the average number of inhabitants of a country, region, urban or rural area in relation to a given surface unit of the territory where that country, region or area is located.

The visualizations were originally made with geopandas. It was necessary to reproject the projected coordinate system, to a espg that was in the measurements I wanted to handle, in my case I used 4485, Mexico ITRF92 / UTM zone 12N, which is in meters. The data used from the data set was the total population, the CVEGEO(Concatenated geostatistical key) and creating a new column under the name of area, which was calculated thanks to the .area function, which provides us with geopandas, which uses the geometry of the shapefiles as parameters. As we well know, the population density is based on people per square kilometer, but for our case, it was done per square hectare for better visualization.
In the case of the dashboard, the visualizations were made again with Folium.

	CVEGEO	POBTOT	geometry	area	pop_den
0	3105000010027	21	POLYGON ((2754849.653 2488201.053, 2754860.000...	193766.758498	1.083777
1	3105000010031	253	POLYGON ((2771776.890 2478590.322, 2771690.909...	157205.277026	16.093607
2	3105000010120	1444	POLYGON ((2764143.656 2479289.504, 2764145.335...	368451.299357	39.191068
3	3105000010154	186	POLYGON ((2752865.731 2469623.554, 2752868.219...	547488.520235	3.397332
4	3105000010169	1580	POLYGON ((2761880.766 2478332.690, 2761889.991...	987275.032911	16.003646

Fig. 6. Dataframe 2020

	CVEGEO	pobtot	geometry	area	pop_den
0	3105000010120	1642.0	POLYGON ((2764143.656 2479289.504, 2764145.335...	368451.299357	44.564913
1	3105000010169	1618.0	POLYGON ((2761880.766 2478332.690, 2761889.991...	987275.032910	16.388544
2	3105000010188	1709.0	POLYGON ((2764110.007 2478349.070, 2764166.774...	482177.592003	35.443372
3	3105000021A	1666.0	POLYGON ((2760966.319 2476809.061, 2760907.648...	568825.022770	29.288444
4	3105000010224	926.0	POLYGON ((2761747.489 2477244.065, 2761744.829...	971559.192461	9.531071

Fig. 7. Dataframe 2010

D. Population comparison

One of the important visualizations to make was the comparison between both years. To be able to observe the change that has occurred between 2010 and 2020 in terms of the population. For this case, the areas that had a decrease in population were separated and two new columns were created, with the growth and decrease in population. The problem was that the areas with a population decrease were in negative numbers, so it was not possible to graph them, so it was separated into two new columns, applying the absolute value to the negative areas and thus being able to graph in a

positive way the growth and decrease of the population and obtain a good visualization of the map of density differences.

```
meridaNeg.head()
```

	CVEGEO	POBTOT2010	POBTOT2020	geometry	Difference
11	3105000010296	992.0	999.0	POLYGON ((2760896.157 2476434.241, 2760893.852...	7.0
23	3105000010440	739.0	760.0	POLYGON ((2761443.980 2474867.516, 2761440.637...	21.0
50	310500001078A	982.0	1455.0	POLYGON ((2758565.513 2467497.722, 2758573.397...	473.0
52	3105000010826	3751.0	3855.0	POLYGON ((2763542.504 2470705.961, 2763540.798...	104.0
65	3105000011523	1539.0	1795.0	POLYGON ((2760660.349 2482741.857, 2760773.799...	256.0

```
meridaPos.head()
```

	CVEGEO	POBTOT2010	POBTOT2020	geometry	Difference
2	3105000010120	1642.0	1444.0	POLYGON ((2764143.656 2479289.504, 2764145.335...	198.0
4	3105000010169	1618.0	1580.0	POLYGON ((2761880.766 2478332.690, 2761889.991...	38.0
5	3105000010188	1709.0	1538.0	POLYGON ((2764110.007 2478349.070, 2764166.774...	171.0
6	310500001021A	1666.0	1571.0	POLYGON ((2760966.319 2476809.061, 2760907.648...	95.0
7	3105000010224	926.0	694.0	POLYGON ((2761747.489 2477244.065, 2761744.829...	232.0

Fig. 8. Dataframe difference positive and negative among 2020 and 2010

For this case, it was only necessary to subtract the population of the year 2010 minus the population of the year 2020. Save it in a new column called Difference, and separate the positive and negative values into two new columns, and then apply the absolute value to the negative values and save them in two different dataframes, which would be graphed in a single choropleth.

E. Population redistribution within the city

For the calculation of the population redistribution, the only difference to the simple population difference between years is that the formula applied should be: ”the populations of the year 2010 divided over the total population of that year, to later subtract the populations of the year 2020 that were equally divided over the total population of that year ”. (Formula: d2 / total2 - d1 / total1)

```
meridaNeg.head()
```

	CVEGEO	POBTOT2010	POBTOT2020	geometry	Difference
50	310500001078A	982.0	1455.0	POLYGON ((2758565.513 2467497.722, 2758573.397...	0.000356
65	3105000011523	1539.0	1795.0	POLYGON ((2760660.349 2482741.857, 2760773.799...	0.000063
66	3105000011542	1173.0	1715.0	POLYGON ((2760506.936 2481992.726, 2760509.643...	0.000401
67	3105000011557	804.0	1307.0	POLYGON ((2761966.613 2482435.025, 2761969.269...	0.000409
68	3105000011561	1212.0	1582.0	POLYGON ((2761057.169 2480685.336, 2761045.339...	0.000221

```
meridaPos.head()
```

	CVEGEO	POBTOT2010	POBTOT2020	geometry	Difference
2	3105000010120	1642.0	1444.0	POLYGON ((2764143.656 2479289.504, 2764145.335...	0.000413
4	3105000010169	1618.0	1580.0	POLYGON ((2761880.766 2478332.690, 2761889.991...	0.000247
5	3105000010188	1709.0	1538.0	POLYGON ((2764110.007 2478349.070, 2764166.774...	0.000394
6	310500001021A	1666.0	1571.0	POLYGON ((2760966.319 2476809.061, 2760907.648...	0.000311
7	3105000010224	926.0	694.0	POLYGON ((2761747.489 2477244.065, 2761744.829...	0.000355

Fig. 9. Dataframe redistribution positive and negative among 2020 and 2010

VI. Results

In this section the final and most important maps made in geopandas will be presented, the results to folium are presented in the dashboard due to its interactive nature.

A. Total population of Mérida and Kanasín

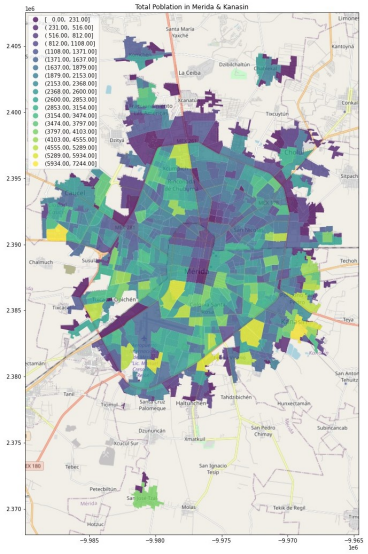

Fig. 10. Total Population of Mérida and Kanasín

Map of the total population of Mérida and Yucatan after performing the quantitative data classification study and see what would be the best classification method for the data we have, in this case the Fisher-Jenks classifier dominates all the other classifiers k = 19 with an ACDM of 38,817. (2020)

B. Population density Mérida and Kanasín 2010

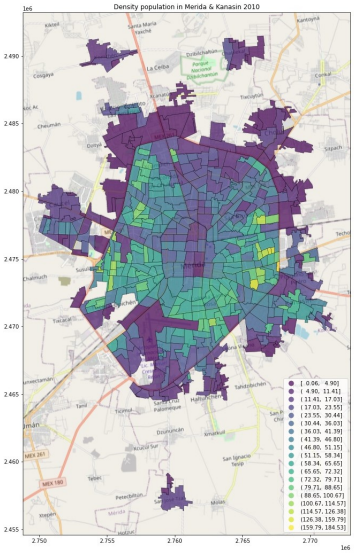

Fig. 11. Density population of Mérida and Kanasín 2010

Population density in Mérida and Kanasin; Inhabitant per square hectare, from the year 2010. (Total population 871,620)

C. Population density Mérida and Kanasín 2020

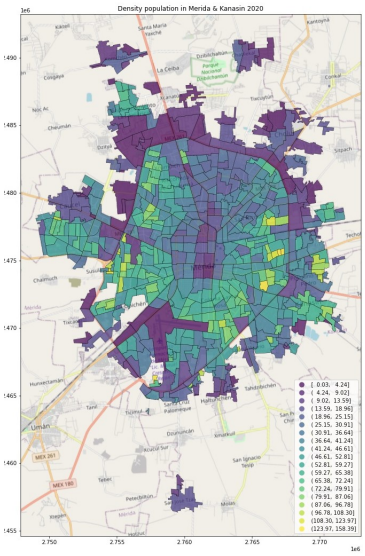

Fig. 11. Density population of Mérida and Kanasín 2020

Population density in Mérida and Kanasín Inhabitant per square hectare, from the year 2020. (Total population 1,097,152)

D. Population difference between 2010 and 2020

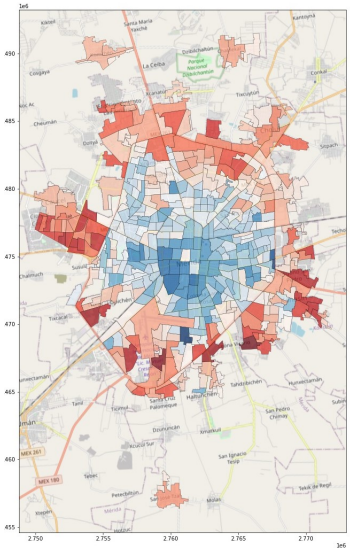

Fig. 12. Population difference between 2010 and 2020

E. Population redistribution between 2010 and 2020

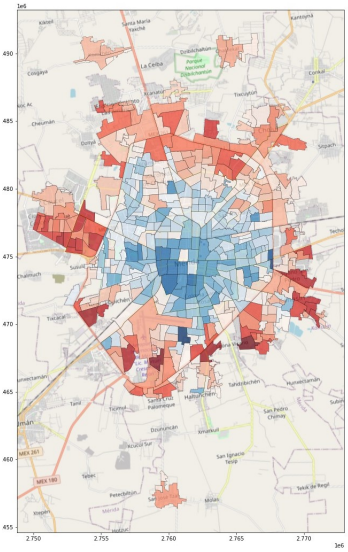

Fig. 13. Population redistribution between 2010 and 2020

VII. Conclusion

The study of the population is a field of knowledge that has the purpose of providing information on the demographic characteristics of the communities, and their relationships with the social, economic and environmental contexts that configure the local development processes, regional and national. Conceptual frameworks, data, and population analysis have several applications in social research, such as the inclusion of the problem of the structure and of population change within multidisciplinary social studies; the use of fonts secondary information; and in the identification and projection of socioeconomic offers and demands that qualify national, sectoral and regional planning. With this paper, the knowledge and tools necessary for compression could be provided. of the relationship between the population and other sectors such as social, economic and educational. The implementation of these visualizations in folium to a dashboard, is also useful for the dissemination of this data. This is just the beginning for a more elaborate future work, where it could not only be implemented for a small area such as the city of Mérida, but for the entire state of Yucatán, and the entire country. In the same way, it will be very useful for the future development of the main objective of this project which is the visualization of road networks.

Acknowledgment

Thanks to Professor Gonzalo Peraza for giving me this opportunity to work with him. It is worth expressing greater gratitude to Professor Didier Gamboa, who, thanks to his classes at UPY, were very helpful and that most of the techniques implemented and tools used were part of his teaching in previous courses. Also thanks to UPY for his great support during this project. On the other hand, thanks to my colleagues Walter Vives and Adrian Carmona, for their advices to improve during this project.

REFERENCES

[1] Cs.nju.edu.cn, 2021. [Online]. Available: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest.

[2] "Advanced plotting with Bokeh — Intro to Python GIS documentation", Automating-gis-processes.github.io, 2021. [Online]. Available: https://automating-gis-processes.github.io/CSC18/lessons/L5/advanced-bokeh.html.

[3] "Interactive maps with Bokeh — Geo-Python - AutoGIS documentation", Automating-gis-processes.github.io, 2021. [Online]. Available: https://automating-gis-processes.github.io/2017/lessons/L5/interactive-map-bokeh.html.

[4] "Choropleth Mapping", Geographicdata.science, 2021. [Online]. Available: https://geographicdata.science/book/notebooks/05choropleth.htmlcomparing-classification-schemes.

[5] "Interactive maps with Bokeh — Geo-Python - AutoGIS documentation", Automating-gis-processes.github.io, 2021. [Online]. Available: https://automating-gis-processes.github.io/2017/lessons/L5/interactive-map-bokeh.html.

[6] I.(INEGI), "Marco Geoestadístico", Inegi.org.mx, 2021. [Online]. Available: https://www.inegi.org.mx/temas/mg/Descargas.

[7] I.(INEGI), "Censo Población y Vivienda 2020", Inegi.org.mx, 2021. [Online]. Available: https://www.inegi.org.mx/programas/ccpv/2020/Datos$_a biertos$.

[8] C. Población, "La distribución territorial de la población", gob.mx, 2021. [Online]. Available: https://www.gob.mx/conapo/acciones-y-programas/la-distribucion-territorial-de-la-poblacion.

[9] V. Cramer, S. Torgersen, E. Kringlen, "Quality of Life in a City: The Effect of Population Density, "Springer Link", [103-116], 2004.

[10] O. Fassio, C. Rollero, N. De Piccoli, "Health, Quality of Life and Population Density: A Preliminary Study on "Contextualized" Quality of Life", "Springer Link", [479-488], 2013

[11] Fahey, T., Whelan, C. T. (2005). First European quality of life survey: Income inequalities and deprivation. Luxembourg: Office for Official Publication of the European Communities.

[12] Costa, G. (2008). Geografia della salute in contesti urbani. In G. Nuvolati M. Tognetti Bordogna (Eds.), Salute, ambiente e qualità della vita nel contesto urbano [Health, environment and quality of life in the urban context] (pp. 97–150). Milano: Franco Angeli.