SPATIAL DATA: Homework 3

Sergio-Yersi Villegas

March 2022

Studying the urban nature and children's recreation areas around primary schools in Barcelona (extensions in blue)

Basis of our work

For this extension of the second homework, we will be adding the remarks provided in the feedback and we will try to comply with the clarifications that were made in the task for this third homework. More specifically, as we will see throughout the report, we will be discussing a little more the implications of the results with respect to the research question (which we will also try to clarify) in a way that we explain what are we trying to get from the data. It is worth mentioning that for the purpose of this homework, which is supposed to advance as much as possible as if it was our term paper (taking into consideration that this is being made individually) and improve homework 2, we will be performing the specified requirements of the task. Therefore, let's start with our report.

In this report we have followed the framework of [1], where they perform an analysis over the green spaces surrounding children's schools in the city of Barcelona. In this paper, they set the hypothesis that schools and their environments might play a relevant role in the reduction of residential disparities in the access to urban nature, given the substantial amount of time that children spend in school settings on a daily basis. Thus, their overall goal is to spatially assess the amount and main components of school green infrastructure within and around a sample of primary schools in the city of Barcelona, Spain, and to examine the equity implications of its distributional patterns. Furthermore, there have also been many other studies regarding how the vegetation cover of a city may somehow have an effect on the children, besides the social inequalities that have been studied in [1]: in [2], they investigated relationships between vegetation and academic achievement as indicated by high school graduation rates across social and environmental contexts in the continental US, therefore assessing this variation and the potential for urban vegetation to support academic attainment. Moreover, [3] analyzes how several building factors influence student performance outcomes, including cognitive skills, standardized test scores and rates of absenteeism. Then, in [4] they studied public elementary schools in the Commonwealth of Virginia to examine correlations between school ground vegetation and outside recess, since exposure to green landscapes within schools may allow children to regain focus, suppress impulses and pay more attention in class.

Therefore, we have found in these aforementioned papers the motivation to perform a study on the topic, since we have seen the huge scope and different approaches that this field can have. More specifically, we have focused on the relevance of Barcelona as case study area since it is supported by an ambitious urban greening plan that is currently in development. One key target that the City Council has outlined is to increase urban green space per resident until 2030, with likely greening effects on school environments. The city's plan aims, among other goals, to enhance the suitability of schoolyards and public green spaces as playful and socially inclusive areas for children.

For the purpose of this extended homework, we will not be replicating the work done in [1], where they end up performing a regression and a k-means clustering algorithm. These advanced computations, or other different and complex approaches, could have been performed in the case that this would have ended up being the topic of the term paper. Alternatively, in order to comply with the main remarks given in the feedback, we will be better describing our approach, by both clarifying our concrete research

question and hypothesis, so that we can discuss with more detail the implications of our results. The goal of homework 2 was to provide an analysis over which sub-divisions of Barcelona have the best-suited conditions for children to go to school to, in terms of their vegetation index and a measure between the schools and children's recreation areas of the neighborhood. This was motivated by the previous paper examples we saw, since both the surroundings and the spaces within schools can have an impact on children in many different levels. Hence, as we saw in homework 2, we will be constructing some of the spatial computations we have learnt in the past sessions and end up with the best sub-divisions of the city, according to the previous stated criteria. In addition, we will be also adding data regarding the real estate market in Barcelona per each sub-district, so that we can compare the housing prices $\left(\frac{€}{m^2}\right)$ with the results we obtained; since schools in Barcelona take into account the proximity of each family's house to the school itself, we will try to extract some conclusions on whether it is a privilege, or not, to have more desirable school surroundings that do not have a negative effect on children's academic performance.

The chosen software to work with has been *Python*, where the main essential libraries we have used to work with spatial data are *GeoPandas* and *Rasterio*. The *Jupyter Notebook* with the done work will be one of the published files.

Data

The data source from which we have obtained all our files is the *Barcelona's City Hall Open Data Service* (`https://opendata-ajuntament.barcelona.cat/en`). Furthermore, we will be working with the projected coordinate reference system of $EPSG : 25831$. Therefore, the scale of all plots will be defined within the following values of these coordinates system, where $x \equiv$ horizontal axis and $y \equiv$ vertical axis:

$$x \in [420500, 436000]$$

$$y \in [4574000, 4591500]$$

We specify it here so that there is no need to display it in every figure, so that we output is displayed nicer and better. Besides, in the *Jupyter Notebook* we can check that these are the correct values of the scale.

- Firstly, the raster data used describes the vegetation cover of the city of Barcelona, based on the Normalized Vegetation Difference Index (NVDI), seen from the sky: the file name is *'2017_ndvi.tif'* (`https://opendata-ajuntament.barcelona.cat/data/en/dataset/cobertura-vegetal-ndvi`).
- Secondly, regarding the vector data, we have used three different data files.
  - To begin with, the administrative units of the city of Barcelona, so that we can eventually represent the vegetation cover raster data by each selected sub-district: the file name is *'Unitats_Administratives_BCN_GeoJSON'* (`https://opendata-ajuntament.barcelona.cat/data/en/dataset/20170706-districtes-barris/resource/cd800462-f326-429f-a67a-c69b7fc4c50a`).

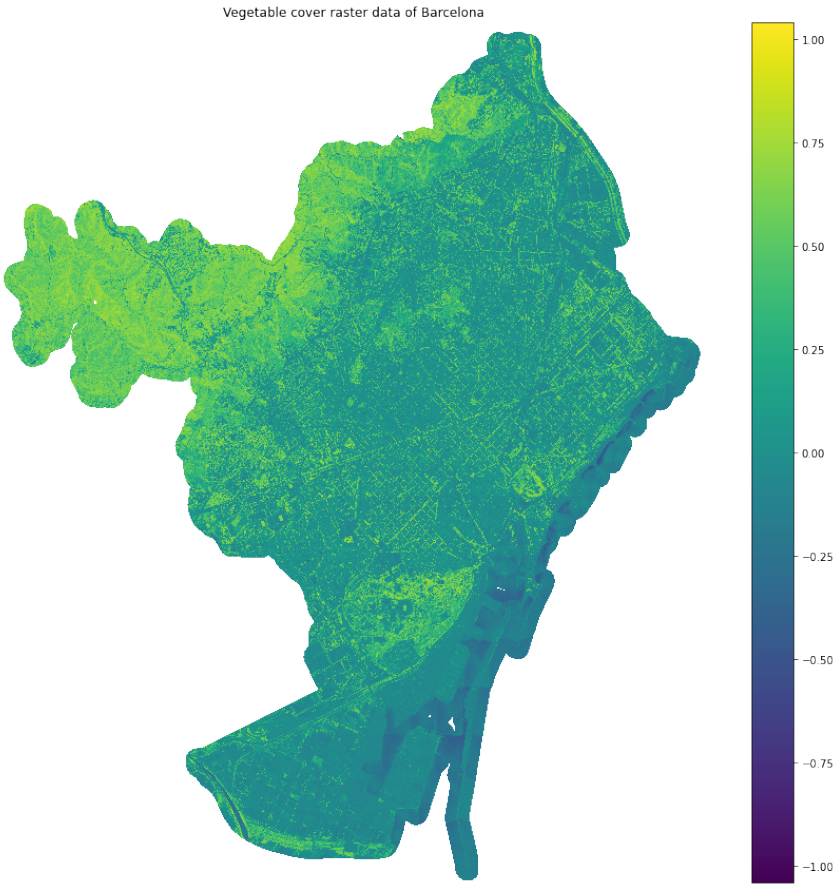

Figure 1: vegetation cover of Barcelona as raster data.

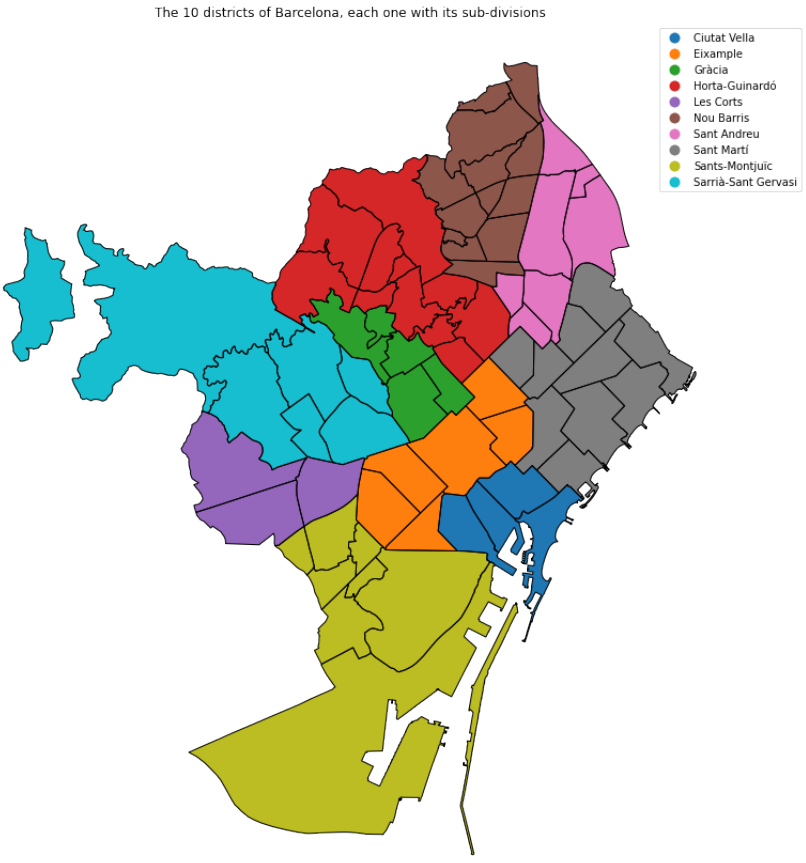

Figure 2: Administrative units of Barcelona colored by the 10 districts of the city, each one with its sub-divisions.

– Then, the educative centers of the city of Barcelona that offer regulated educative courses: the file name is *'opendatabcn_educacio_ensenyament_reglat-js.json'* (`https://opendata-ajuntament.barcelona.cat/data/en/dataset/educacio-ensenyament-reglat/resource/94c7ea5d-c0b3-4482-bea8-6d5023844798`).

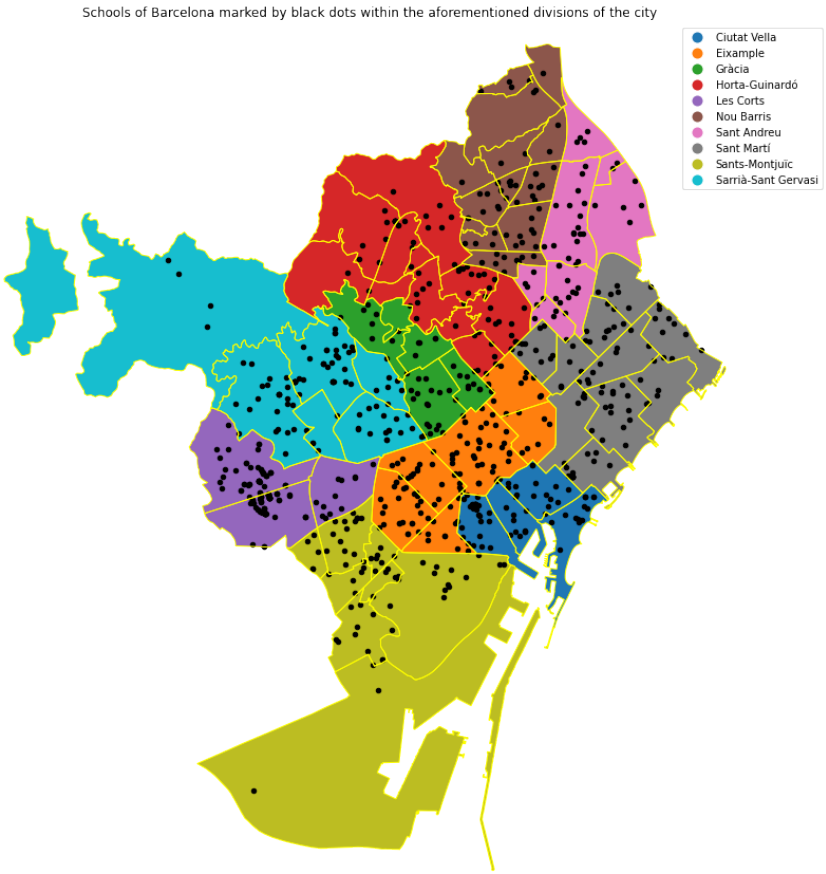

Figure 3: Schools of Barcelona marked by black dots within the aforementioned divisions of the city.

– Moreover, the children's play areas in the city of Barcelona: the file name is *'opendatabcn_cultura_espais-infantils-js.json'* (`https://opendata-ajuntament.barcelona.cat/data/en/dataset/culturailleure-espaisinfantils/resource/1d94653e-33a4-4ca2-94f6-83de6f3014fb`).

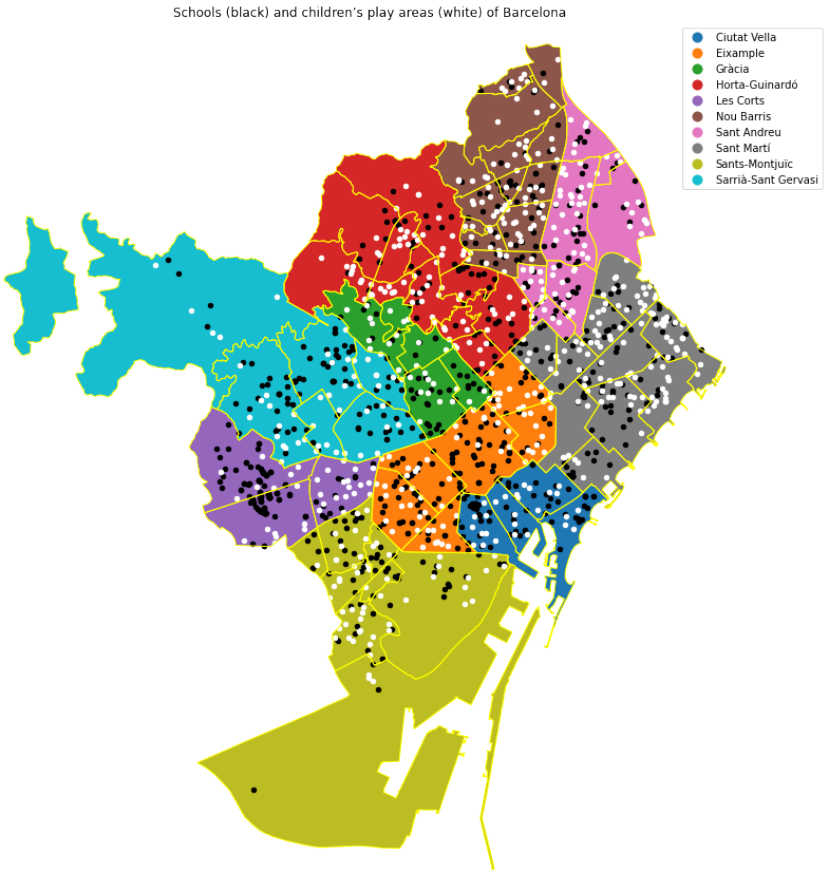

Figure 4: Schools (marked by black dots) and children's play areas (marked by white dots) of Barcelona within the aforementioned divisions of the city.

It is worth noting that the *.json* files, originally being regular *Pandas' DataFrames*, have been converted to *GeoPandas' DataFrames* since they had the required coordinates stored as a dictionary. These steps (and more), which can be seen in detail in the aforementioned *Jupyter Notebook*, have been performed in order to be able to construct the upcoming spatial computations.

- Finally, the real estate market prices of the city of Barcelona, from 2013 to 2021: the file name is *'house_market_bcn.csv'* (`https://ajuntament.barcelona.cat/estadistica/angles/Estadistiques_per_territori/Barris/Habitatge_i_mercat_immobiliari/Mercat_immobiliari/Habitatge_segona_ma/t01.htm`).

Construction

GeoDataFrame constructor

Firstly, as we already stated, in order to have the *.json* data as *GeoDataFrames*, we will be creating the corresponding 'geometry' column (which uniquely differentiates a *GeoDataFrame* from a regular *DataFrame*), since the data contains the coordinates (in the three different systems of reference) as dictionaries. In order to do that, we apply our custom function, which returns a list with the values of a given dictionary, to each row. Then, once we have the coordinates in a list, we apply the *Point()* function from the *shapely.geometry* package, which creates a geographical 'point' object to identify the values in a map. Finally, with this new column, we can already transform the data into a *GeoDataFrame* with the right coordinates. As we said, these steps are more easily understood by looking at the *Jupyter Notebook*.

Overlay

Then, in order to have the raster data directly related with the vector data of the administrative units of Barcelona, we will overlay the data by computing the zonal statistics of our raster data. Hence, we will now be able to plot the administrative units of the city represented with the color corresponding to the index of vegetation cover they have, computed by the zonal statistics. In particular, with the **mean** as the chosen statistic to use, below we can see the difference between plotting our vector data on top of the raster data and plotting all the data together with this statistic.

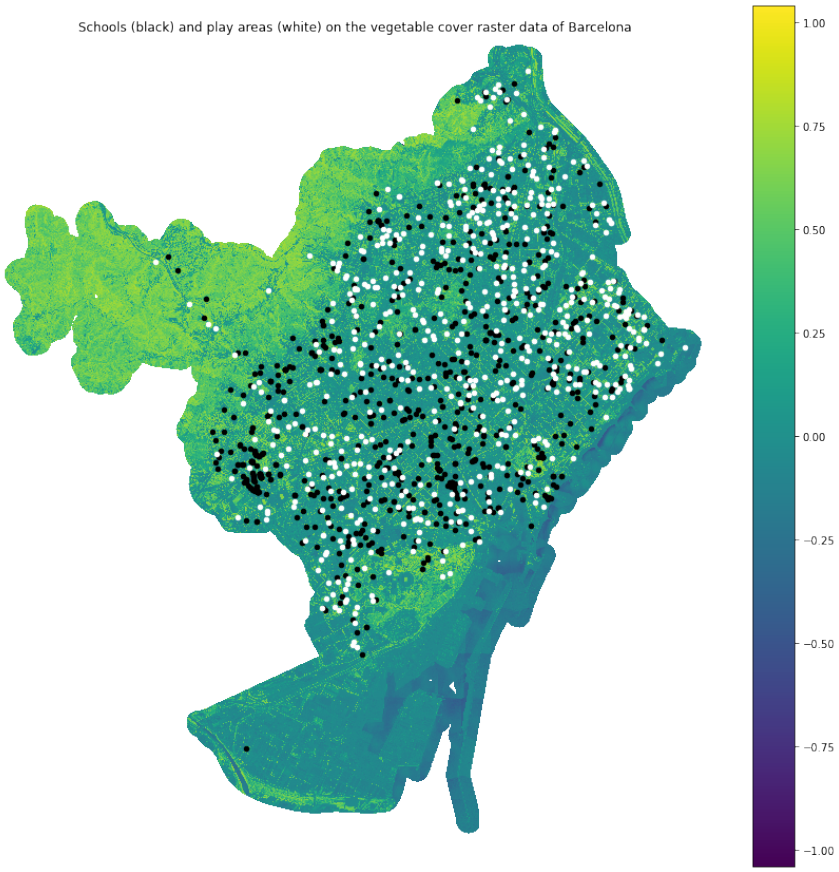

Figure 5: Schools (black dots) and children's play areas (white dots) of Barcelona shown on top of the vegetation cover raster data.

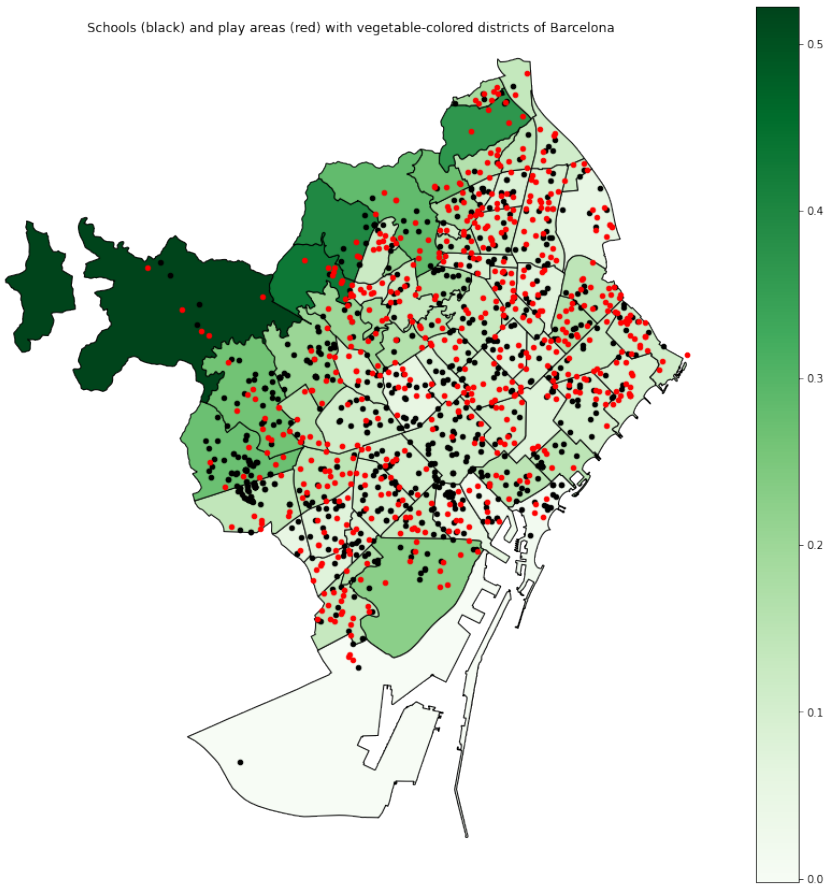

Figure 6: Schools (black dots) and children's play areas (red dots) of Barcelona within the sub-divisions of the city, now colored by the **mean** vegetation cover index.

Buffer

Now, we join the data preserving the geometry of the schools (points), and not the polygons from the districts. Hence, we now have the information regarding the administrative units of the city of each school. In order to be able to plot the buffers around each school, we create a new *GeoDataFrame* where now the geometry of each row (corresponding to each school) is a buffer with a radius of 300 meters, which we have defined by previously changing the projected coordinates reference system to $EPSG : 32634$. Once created, we convert it again to the original coordinates, so that we can plot it without any problems.

Therefore, below we plot the buffers around the schools of Barcelona and the children's play areas within the correctly colored administrative units of the city according to the **mean** index of vegetation cover they have. Hence, we can see the schools containing play areas in a radius of 300m.

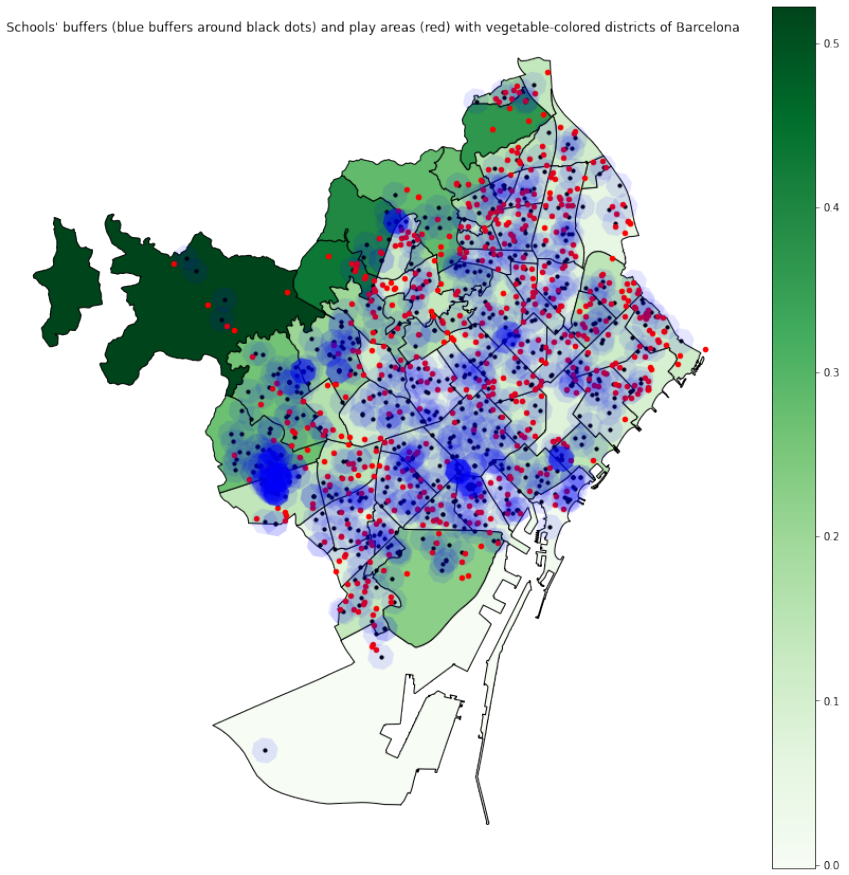

Figure 7: Schools' buffers (blue buffers around black dots) and children's play areas (red dots) of Barcelona within the sub-divisions of the city, now colored by the **mean** vegetation cover index.

Final results

Finally, we can compute the amount of schools and play areas that are contained in each sub-division of the city. Therefore, together with the **mean** vegetation index of each of these divisions, we can create a final *DataFrame* with the name of the sub-division, the **mean** vegetation index and the ratio between the amount of play areas and the number of schools of each sub-division. Thus, we can create two sorted *DataFrames*, each one corresponding to the measure used to sort the results (**mean** vegetation index and ratio). Now, we can filter each sorted *DataFrame* by just taking the sub-divisions that are also above a chosen threshold on the other measure; i.e., when sorting by **mean** vegetation index, we keep those that have a ratio of play areas per school higher than 2 and, when sorting by ratio, we keep those that have a **mean** vegetation index higher than 0.25. Hence, by joining those results with the intersection of them, we see that the sub-divisions that are above both mentioned thresholds are:

Sub-division	Mean vegetation index	Play areas per school
Sant Genís dels Agudells	0.430919	4.5
Torre Baró	0.368097	2.5
Canyelles	0.268988	5

Table 1: Sub-divisions that are above a **mean** vegetation index of 0.25 and a play area per school ratio of 2.

Moreover, in order to exactly rank these results, we could get the top-5 sub-divisions by computing one "score" column, which multiplied the **mean** vegetation index and the ratio of play areas per school. As we can see in Table 2, where we have displayed the top-5 sub-divisions, computing this score also yields us the 3 sub-divisions that were the only ones above both thresholds. This column has been scaled with a min-max normalization, which follows the expression below:

$$\hat{s_i} = \frac{s_i - \min(s)}{\max(s) - \min(s)} \tag{1}$$

where s is the vector with all the scores, s_i is each score and $\hat{s_i}$ is each normalized score, which are within $\hat{s_i} \in [0, 1]$. Besides, in order to assess how these results are linked to the real estate market prices of the city, we will add the housing price, $\left(\frac{€}{m^2}\right)$, of each sub-division appearing in the following top-5. In that way, we will be able to see if it is a privilege, or not, to live closer to schools with the best suited conditions in terms of green and play areas surrounding the location. Note that we have also added a normalized value of the housing prices, so that it can be easily understood if each price is high or low compared to the rest.

Sub-division	**Score**	$\left(\frac{€}{m^2}\right)$	**Scaled housing price**
Sant Genís dels Agudells	1.000000	2.461375	0.207485
Canyelles	0.693933	NaN	NaN
Can Baró	0.485061	2.899500	0.301105
Torre Baró	0.475173	NaN	NaN
Vallvidrera, el Tibidabo i les Planes	0.337516	3.791000	0.491603

Table 2: Top-5 sub-divisions according to the computed score between the **mean** vegetation index and play area per school ratio. House price and normalized house price has been included.

Therefore, according to the both measures we obtained in our spatial analysis, this would be the top-5 neighborhoods for children to go to school to in the sense that, as it was stated in [2], [3] and [4], their school's surroundings were not detrimental to their academic performance. First of all, the missing information is already stated in the link pasted above for this data (see little remark after the table from the link): the open-data source from the city of Barcelona lacks the values for some of the sub-divisions. Therefore, this is an issue in order to assess the results, since two of the top-5 sub-divisions has missing information. Then, for the ones we do have the data, we can see that they are not even above half of the most expensive price. Furthermore, by checking the whole data (see notebook), not being above the half is not because there is one sub-division that is much more expensive than the others, but because there are actually many other sub-divisions above. More precisely, there are more than 20 sub-divisions above Vallvidrera, el Tibidabo i les Planes (it is the same sub-division with a compound name), which is the one with the highest house pricing among the top-5 above. Therefore, from these results, it does not seem to be that the neighborhoods with more green and play areas are the most expensive ones.

Finally, while the complete table with the results for each sub-division can be checked in the *Jupyter Notebook*, here we will display the 5 neighborhoods with the highest housing price, each one with their corresponding scores.

Sub-division	Mean vegetation index	Play areas per school	Score	$\left(\frac{€}{m^2}\right)$	Scaled housing price
Diagonal Mar i el Front Marítim del Poblenou	0.092713	4.000000	0.192184	6.170222	1.000000
Pedralbes	0.271265	0.142857	0.021121	5.833666	0.928084
les Tres Torres	0.161047	1.250000	0.104853	5.431223	0.842089
la Dreta de l'Eixample	0.105092	0.297297	0.017253	5.418777	0.839430
Sarrià	0.262925	0.600000	0.082419	5.372778	0.829600

Table 3: Top-5 sub-divisions according to the house price, together with the values for the **mean** vegetation index, the play area per school ratio and the computed score between the previous ones.

From these results in Table 3, we can observe that the most expensive sub-divisions of the city actually have low scores regarding their green and play areas. Together with what we observed in Table 2, we can conclude that, according to our analysis, having the best-suited conditions for children to go to school to, in terms of their vegetation index and a measure between the schools and children's recreation areas of the neighborhood, does not seem to be positively correlated.

	mean_veg_cover	play_areas/school	veg_play_score	euro_per_m2
mean_veg_cover	1.000000	0.254882	0.624523	0.012207
play_areas/school	0.254882	1.000000	0.816368	-0.330861
veg_play_score	0.624523	0.816368	1.000000	-0.269402
euro_per_m2	0.012207	-0.330861	-0.269402	1.000000

Figure 8: Correlation matrix between the covariates of study.

What's more, as we can see in Figure 8, by taking the correlation matrix between the covariates we studied, we can even see some negative correlation between the house prices and both the ratio of play areas per school and the final computed score.

Conclusions and further work that could be done

Therefore, concluding, for the study we realized, it does not seem to be a privilege (money-wise) to live in a neighborhood with well-suited for children to grow in. Hence, the academic performance of children would not be influenced by their parents' money income limitations and so, there would not be social disparities. However, this study has been limited by several key factors: the resources we had (such as the missing data for some sub-divisions, which at least it was just for a few of them, as it can be checked in the notebook), the available time and the individuality of the work. With better conditions, this work could be further expanded by: obtaining better data that yielded better results; perform more complex computations such as the ones done in [1], where they execute a k-means clustering algorithm in order to group the schools of Barcelona according to several measures; perform more comparisons with other estimators applied to the data; and, finally, if it was available, get some data on the grades of the schools so that we could assess the predictions made on whether the children's academic performance was actually influenced by the studied factors.

References

[1] F. Baró, D. A. Camacho, C. P. Del Pulgar, M. Triguero-Mas, and I. Anguelovski, "School greening: Right or privilege? examining urban nature within and around primary schools through an equity lens," *Landscape and Urban Planning*, vol. 208, p. 104019, 2021.

[2] C. B. Hodson and H. A. Sander, "Relationships between urban vegetation and academic achievement vary with social and environmental context," *Landscape and Urban Planning*, vol. 214, p. 104161, 2021.

[3] D. Vakalis, C. Lepine, H. MacLean, and J. Siegel, "Can green schools influence academic performance?," *Critical Reviews in Environmental Science and Technology*, vol. 51, no. 13, pp. 1354–1396, 2021.

[4] K. L. Arbogast, B. C. Kane, J. L. Kirwan, and B. R. Hertel, "Vegetation and outdoor recess time at elementary schools: What are the connections?," *Journal of Environmental Psychology*, vol. 29, no. 4, pp. 450–456, 2009.