Reducing contamination indicators in Spain by city mobility improvements based on open data

Data Life Cycle

Cédric Prieels
Eduardo Ruiz Ruiz
Fernando Solar Iglesias
Nicolò Trevisani

Table of contents

1. Introduction

Nowadays, it is widely accepted that the high levels of contamination around the different countries in the world are causing critical problems, not only on our planet like the global warming or ozone layer depletion, but also on their individual's and animals health like respiratory and heart problems or habitat changes. One of the most important causes of air pollution is the combustion of fossil fuels, produced by the aggressive use of private transport in our day-to-day transportation. This is mainly a consequence of the lack of available public infrastructure. [1]

In order to reduce the problem previously mentioned, this project aims to find the optimal place in Spain to introduce a new bike lane based on transportation usage and air quality conditions. Along the different sections of this document, details and procedures about the public and official published data used are given. The process of combining different data sources results on the production of new information and a final conclusion.

2. Information about data and its format

This section describes a detail description of the data planned to produce and the one that currently exists, as well as its processing and management.

2.1. Description of data to be produced

The first step of the project is to identify where is the best place in Spain to place a new bike lane. To do so, we will understand in which city in Spain the low rate of bike usage can be related to a lack of bike lanes. In particular, we will use a national-based inquiry asking the reasons why citizens do not use bikes, focusing on:

- A complete net of bike lanes is missing;
- There is too much traffic;
- There are not enough parking sites dedicated to bikes;
- Personal security.

Once the city has been selected, we plan to collect different kinds of data to comprehensively evaluate its needs in terms of traffic conditions. This will help us to find the most contaminated and less connected by bus place, in order to introduce a bike line in this region. The main sources of data will be:

- Sensors
    - Parking;
    - Pollution;
    - Traffic.

- Data from buses:
    - People that get on/off a bus and where;
    - Time bus takes to cover a given distance;
    - Time to wait for the bus, if available.

- Polls:
    - Citizens opinion.

2.2. How data will be acquired

Data will come from different sources.

For the selection of the city, the “Instituto Nacional de Estadística” page provides open access to the results of several nationally inquiries. As many of them are created by grouping the results per autonomous community, these may need to be further integrated by additional polls, for example specifically designed for the autonomous community that in average shows the highest need for a stronger bike lane net.

Then, we will search among the open-access data created by the city we select. This will be our main source of data, as this is a perfect way to get information about the existing transportations in the city in a reliable and efficient way, without spending any money nor spending time creating experimental devices and sensors.

As we realize this is probably not enough to get all the data we need, we also consider two other sources of data, as we will explain later on. Basically, we plan to create and distribute several polls to the inhabitants of the target city (observational data), probably online in order to limit the cost of such a process. This will be a good way to learn more about the people already living in the city, learning about their complaints and things that they think the city is doing right, in order to see what to improve.

Finally, a limited number of additional sensors might be needed in the city to gather data that is not currently available in its open data portal (experimental/observational data).

Data will be acquired using

- Sensors;
- Polls;
- Open source data.

It might be needed to add some additional sensors since the open source data already available is probably not enough for our needs, which will take time. Indeed, we will need to first of all develop the sensors we need, install them, and finally take data for at least a year, in order to account for the seasonal variation of the pollution in the city and gather statistics. Redacting and propagating the polls will also take time, but we should have the data back in a matter of a few weeks, and the analysis of this data should be quite quick (cf. Gantt diagram).

2.3. How data will be processed

Open source data will be accessed and read by a Python script, reading directly the CSV files accessible online, or moving to CSV pieces of information that may be accessible only in other formats, so that we can have a homogeneous dataset.

The workflow that we plan is summarized as follows:

1. Download data from online sources in CSV or JSON format;
2. Create structured data based on polls using Google Forms (allowing us to export the data in a CSV format as well);
3. Integrate data from different sources in an online work repository based on open source platforms such as Jupyter Notebook and GitHub;
4. Perform a statistical analysis on data using Python;
5. Extract a set of advices on how to improve mobility in our target city, also supported by graphics to ease the interpretation, based on the contamination observed in the city.

2.4. File formats

The open data is available in several formats, such as JSON and CSV, which are perfect candidates to be read from a Python script, since some built-in functions in Python allow us to read and transform them directly.

About naming conventions, English will be the base language when it comes to label data. File names will not only include its title but also the source of its data.

2.5. Quality assurance & control during sample collection, analysis, and processing

Open source data is already controlled before being published. We plan in any case to introduce tests to find possible outliers or issues in the data, and to clean this data by removing unnecessary and unavailable data, for example.

For the polls, we plan to put some control questions to understand if a person answered sensibly and if its answers can be considered in the analysis.

When it comes to control the development phase, which involves analysis and processing, Github will be used to track all the changes made by the team, in order to detect human mistakes, but most important to have the code safely stored and always up to date.

More details will be given in the next section related to data management.

2.6. Existing data

We used two datasets coming from open data repositories:

- "Main reasons why people older than 16 who usually do not move by bicycle or on foot, grouped by autonomous community", from the Spanish National Statistics Institute (INE).
    - Available in: https://www.ine.es/jaxi/tabla.do?type=pcaxis&path=/t25/p500/2008/p04/l0/&file=04017b.px
- Air quality in Malaga in 2018, provided by the Malaga open data repository.
    - Available in: https://datosabiertos.malaga.eu/recursos/ambiente/calidadaire/2018.json

Both datasets, as provided by the corresponding repositories, were already in a good shape and almost ready to be used for the analysis. Nevertheless, some small curation operations were needed:

- **INE dataset**: used commas to define decimals, so that the standard comma-separated csv was useless. Luckily, the dataset was also provided in the point-comma separated format: by simply replacing in this dataset the commas with points, Python was able to properly read it.

- **Air quality dataset**: in the Spanish open data repository, only the metadata were provided. Among them, the URL of the JSON file with the information about air quality. This file was then converted to csv to make it easier to read by Pandas. Coordinates are provided such that 4 points define a polygon, where the measurement of the air quality has been performed. Since the polygons are in general small with re respect to the size of the city, we eventually use only the first point, for simplicity. Since the dataset contains both numerical and categorical pieces of information about substances present in the air, it has been split, to treat the two kinds of data separately.

2.7. How data will be managed in short-term

Collaborative repositories will be used in order to guarantee a correct data management. This virtual facility is planned to be hosted in a private environment where some precautionary measures will be followed by IT administrators:

- **Backups**: With the aim of minimizing the impact that technical issues or service outages could cause to data, daily backups to a mirror server will be scheduled. This will allow us to restore a stable state when data is inaccessible, corrupted or even lost.

- **Version control**: The well-known Git control version will be used to get a trustworthy log of changes made in files, code and repositories.

- **Security**: All the platforms used in this project will be monitored and secured in order to protect the data that is collected and processed. This will require to set up customized firewall and network rules, as well as standard security procedures for data storage and transmission.

- **Protection**: Authorization to the platforms mentioned will be possible through personal credentials that will be secured with standard and proven methods like two-factor authentication. There will be also documentation and training about good practices when being online to reduce the risk of security breaches due to human factors.

3. Metadata content and format

This section shows the standards and schema used for defining the dataset metadata associated with this project.

3.1. What format will be used for the metadata and what metadata are needed

Dublin Core is the format chosen for describing the metadata associated with this project as it is a well-known standard commonly used in this scenario. It will be represented using the Extensible Markup Language (XML) and following the official schema provided.

These are the metadata fields needed in this project and its details according to the mandatory schema [2].

- **Title**: The name given to the resource.
- **Creator**: The entity responsible for making the resource.
- **Subject**: The topic of the resource, represented using keywords separated by a semicolon.
- **Description**: A brief text describing the dataset content.
- **Publisher**: The entity responsible for making the resource available.
- **Date**: As a best practice, the W3CDTF profile of ISO 8601 [W3CDTF] recommended encoding scheme is used: YYYY-MM-DD.
- **Type**: It could be a collection, dataset, event, image, interactive resource, service, software, sound, text or physical object. In this case, it will be a dataset.
- **Format**: The file format according to official MIME types.
- **Identifier**: An unambiguous reference to the resource within a given context.
- **Language**: A language of the resource which it is not applicable for this case.
- **Relation**: A related resource.
- **Coverage**: The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.

- **Rights and access**: Information about rights held in and over the resource.
- **Accrual method and policy**: The method for adding data to the collection and the policies involved.
- **Audience:** The kind of people target associated with the resource.

3.2. How metadata will be created and/or captured

Metadata will be manually created using the Dublin Core XML schema, based on the information available on the dataset's web sites. In this case, two metadata files were generated about each source origin.

INE:

- **Title**: Porcentaje de personas de 16 y más años que usualmente no se desplazan caminando o en bicicleta, por comunidad autónoma de residencia y motivos por los que no lo hacen
- **Creator**: Instituto Nacional de Estadística
- **Subject**: desplazamiento;españa;bicicleta;caminando;encuesta
- **Description**: Resultados de la encuesta de Hogares y Medio Ambiente del 2008.
- **Publisher**: Instituto Nacional de Estadística
- **Date**: 2008
- **Type**: Dataset
- **Format**: text/csv
- **Format extent**: 1.9 kB
- **Identifier**: 04017b
- **Language**: es_ES
- **Coverage temporal**: 2008
- **Coverage espacial**: España

- **Rights**: Se permite reutilización con cita
- **Access rights**: Abierto
- **Accrual method**: Encuesta
- **Accrual policy**: Cerrado
- **Audience education level**: Educación obligatoria

Air Quality Malaga:

- **Title**: Calidad del aire 2018
- **Creator**: Urban Clouds
- **Subject**: contaminación;aire;calidad;málaga
- **Description**: Datos del 2018 asociados a la calidad del aire en Málaga. Los atributos que contienen _APP se refieren a medidas tomadas con Appmosfera. Los atributos que contienen _M se refieren a medidas tomadas con SMAQ_mobile. Los atributos que contienen _F se refieren a medidas realizadas con SMAQ_fija. Los gases que no contienen ninguno de los atributos anteriores (p. ej. o3, o3_level,....) se corresponden al cálculo agregado de SMAQ mobile y SMAQ fijas. Las medidas se toman en ug/m3 (microgramos / metro cúblico). Atributos con el texto global (p.ej. "iuca.level_APP_global") corresponde a los globales calculados de cada dispositivo o el global de todos los dispositivos. En los IUCA (Indice Urbano de Calidad del Aire) no se tiene en cuenta PM1
- **Publisher**: Datos Abiertos Ayuntamiento de Málaga
- **Date**: 2018-05-14
- **Type**: Dataset
- **Format**: application/json
- **Format extent**: 450 kB
- **Identifier**: 99d54259-ab9c-4fda-a58e-d96988821282
- **Language**: es_ES
- **Coverage temporal**: 2018

- **Coverage espacial**: Málaga
- **Rights**: Reconocimiento-CompartirIgual 3.0 España (CC BY-SA 3.0 ES)
- **Access rights**: Abierto
- **Accrual method**: Sensores
- **Accrual policy**: Cerrado
- **Audience education level**: Educación obligatoria

4. Policies for access, sharing and re-use

The open data is already available online for everyone to use, and do not include any personal information, only the data collected by sensors in the city, which should therefore not induce any kind of intellectual property, copyright nor ethnical and privacy issues. Additionally, the data is already published on the internet, so we don't need to worry about distributing this data for an eventual re-use in the future. However, if we do need at some point to gather additional data by adding sensors in the city, then we will need to distribute this new data in the same way that is already currently done, meaning that we will just need to add it to the current open data database, using the same copyright and ways to distribute the data as already available.

The data used for this work has in any way been published, after the needed curation steps, in Zenodo. Additionally, both the code used to analyze the datasets and the datasets themselves have been published to Github.

- **Mobility project in Zenodo**: https://zenodo.org/communities/mobilityproject/
- **Dataset analysis**: https://github.com/NTrevisani/DataScienceMaster/tree/master/DataLifeCycle/TrabajoGrupo

These measures ensure that the datasets and the code will be safely stored and easily accessible for anyone interested in consulting them.

The concern is more about the polls data that we will collect. In this case, we will need to first of all be careful not to include any question relating to race, religion or personal data, to avoid any ethnical and privacy issue.

The best way to proceed will be to completely anonymize the data received, by assigning it a random ID instead of the name of the person who responded to the survey, in such a way to be compliant with regulations in vigor in Spain about personal data collection. Then, once collected, the data will be analyzed by ourselves and might be made available as a public open dataset in one of the existing open data platforms in order to distribute it to the general public, in case someone ever needs to perform a similar study in the future. In this case, the randomization of the personal identifiers of the people that filled the survey will be mandatory.

This dataset will be published using the Creative Commons copyright, forcing the eventual future users to cite our work, without any further requirements.

5. Long-term storage and data management

The data analyzed in the first part of this project is already available and maintained by the city of Malaga on the long term. They are directly responsible for the preservation and maintenance of this open data on the long term and of the portal distributing the data itself, according to the usual rules.

On the other hand, the data gathered by our polls might be stored online using cloud solutions already offered by large companies such as Google, for a minimal cost, at least at the beginning of the project, since this data is only expected to be used by ourselves at first. If we see that the data collected has a significant value, or that many people did respond to the survey, that we will definitely consider publishing this data for an eventual re-use by other teams of different projects in the future. This will be done first of all by making sure that the data is completely randomized and do not include any personal data, before publishing this data to Zenodo or any other online platform to publicly share the data for a minimal cost to the general public. This would allow us in the same time to define our own metadata and privacy policy, for example, and to obtain a DOI for our dataset.

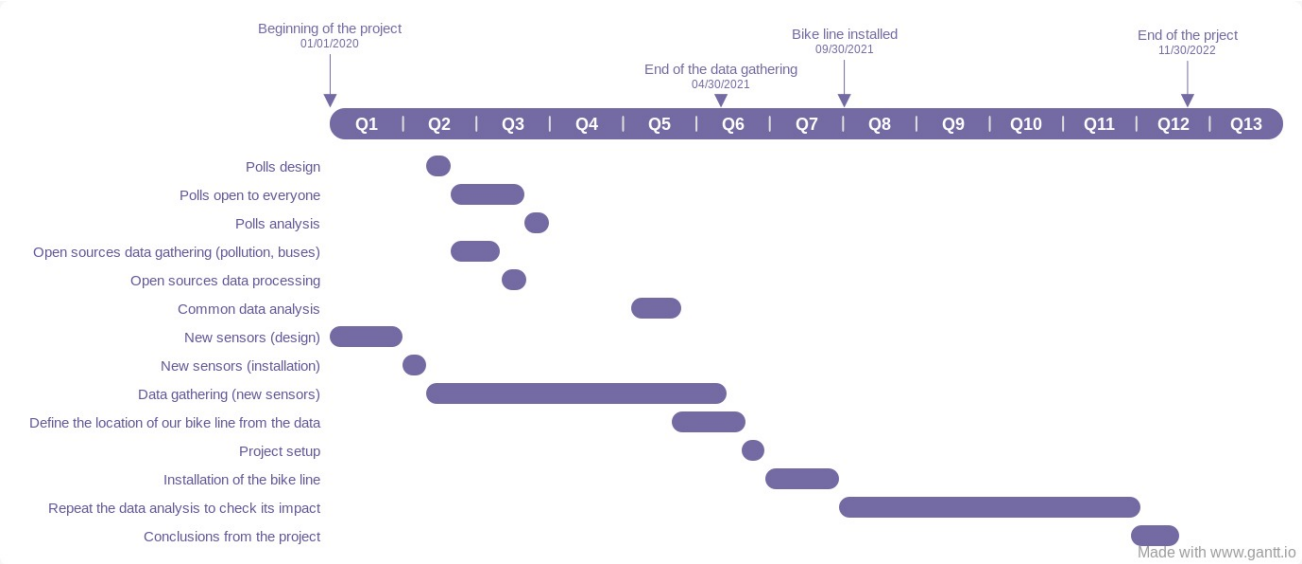
If we need to install additional sensors in the city, we would probably call the same company, Urban Clouds, as the one that already installed the data quality sensors in the city of Malaga. This means that the data gathered by these new sensors will be provided in the same format as the data already available, meaning that we should be able to add it to the Malaga open data portal in a straightforward way. They will then guarantee themselves the long-term availability of this data to everyone, in the same way that is already done, as this does not suppose any additional work and cost from them.

6. Gantt chart

This diagram is usually used in order to represent the time distribution of the different tasks that will have to be performed in order to complete our project.

This project is divided in several main tasks, the first being the design of some sensors we need to add in the city of Malaga, and the design of the polls, that can be performed at the same time. Then, a few months will be dedicated to the analysis of the data that has been gathered through several ways, during one year, since we want to make the air quality does not depend on the season of the data taking.

After this analysis, the actual bike line itself will be set up, and additional measurements will be performed in order to check the new results obtained after the bike line installation. In this way, we will test the impact that our project had on the air quality and traffic situation in the city.



7. Budget

7.1. Anticipated costs

Let's now estimate the budget of this project, depending on its different phases:

- **Data gathering**: Most of the data we need is already accessible on the Malaga open data portal, so is therefore free to access. This data has been gathered by Urban Clouds, a company specialized in the development of sensors for the measurement of air quality data in different cities. The only fee associated to the analysis of this data will be the salary of the different people involved in their analysis, probably not more than two people, for an estimated time of three months (if we consider a good salary of 1500€ per month, this means that the cost estimated for this operation is around de 9000€).

The polls we plan to develop in order to gather the opinion of the people living in the city should be quite simple and straightforward to implement, even though we need make sure the data received is reliable by implementing control questions as well. Depending on the number of responses received, the design of the polls and the analysis of this data should not take more than 2 months, as a job done by a single person (3000€). Free platforms do exist to distribute the poll online, but we probably will have to advertise it as well to get as many responses as possible (1000€), even though a smart use of social media might help us reduce this cost.

If needed, depending on the quality of the open data, we might also need to develop our own sensors and install them in the city, which will of course delay the execution of this project. In this case, two options are available: either we buy already existing sensors from a company such as Urban Clouds, which might be the fastest and easiest solution since they already provided many sensors for this particular city. The other possibility would be to design ourselves our own sensors, which should take at least 4 months with a team of specialized people (~25.000€, for the salaries and the material needed to build our prototypes). In our case, the first solution seems more viable, especially since Urban Clouds offers the possibility to rent sensors instead of buying them. If more sensors are needed, we could then just rent them for a year, to get seasonal data of the contamination in the city, which is expected to be much cheaper, especially since many sensors are already installed and we would therefore only need to add a few of them on strategic locations. Additionally, this company is providing the maintenance of the sensors as well.

- **Actual implementation.** The most expansive part of this program will be the actual implementation of this bike lane. Once the location of this bike lane is decided, we would need to actually build it in the city. This will be the most expensive part of this project. Depending on the location of this lane, the actual engineering works are expected to last for a few months, while several people would need to do the physical work of installing this line. Its installation will probably also result in some

traffic closures, even though limited, has to be taken into account. Depending on the type of bike lane we decide to put (which will depend on its actual location, once the data analyzed), the price of such an engineering work has been estimated [3] to be between 5.000 and 500.000€ for each kilometer build (in our case, the cheapest case will be probably be more than enough in a city, since our plan is to use an existing road and change it into a bike line). If we imagine that the line we build will be around three kilometers long, which is reasonable in a "small" city such as Malaga, we can then estimate its cost to around 25.000€, accounting for eventual delays and/or issues with its actual construction.

In total, the project should then not cost more than 50.000€, accounting for the data gathering, the eventual installation of new sensors in particular areas of the city and the actual construction of the new bike line.

7.2. How costs will be paid

The budget for this project will have to come from the "Ayuntamiento de Malaga", since our plan is to install the bike line in their city, and they will be the one beneficiating from the improvement in the air and in the mobility in the city.

8. Data analysis

The data analysis has been carried out using the Jupyter Notebook that can be found at this Github link and where all the details are given:

https://github.com/NTrevisani/DataScienceMaster/blob/master/DataLifeCycle/TrabajoGrupo/TrabajoGrupo.ipynb

The analysis has been divided in two steps:

1. Selection of the city where two build a bike lane;
2. Choice of the part of the city which best fits the construction of the new bike lane.

8.1. City selection

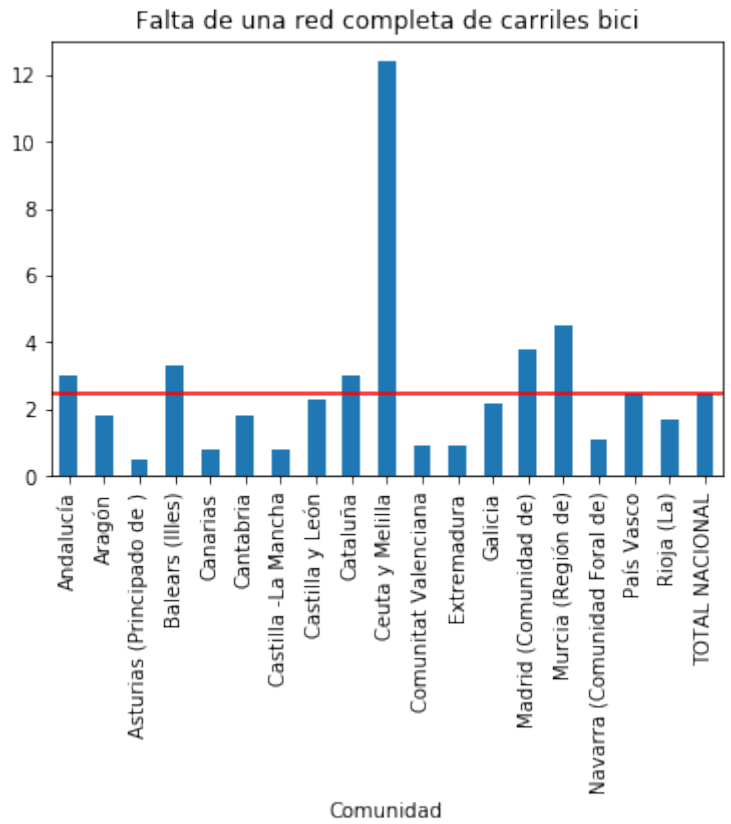
For the first point, we relied on a dataset directly provided by the "Instituto Nacional de Estadística", summarizing the reasons why people from the age of 16 does not move on foot or by bicycle, grouped by autonomous community.

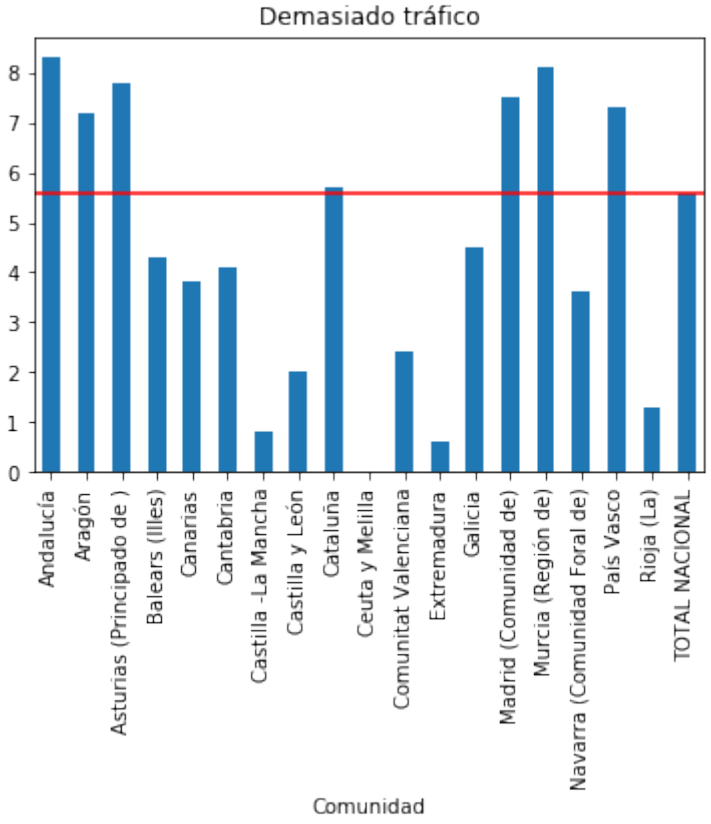
As already mentioned, we focused on the reasons most correlated with the lack of bike lanes:

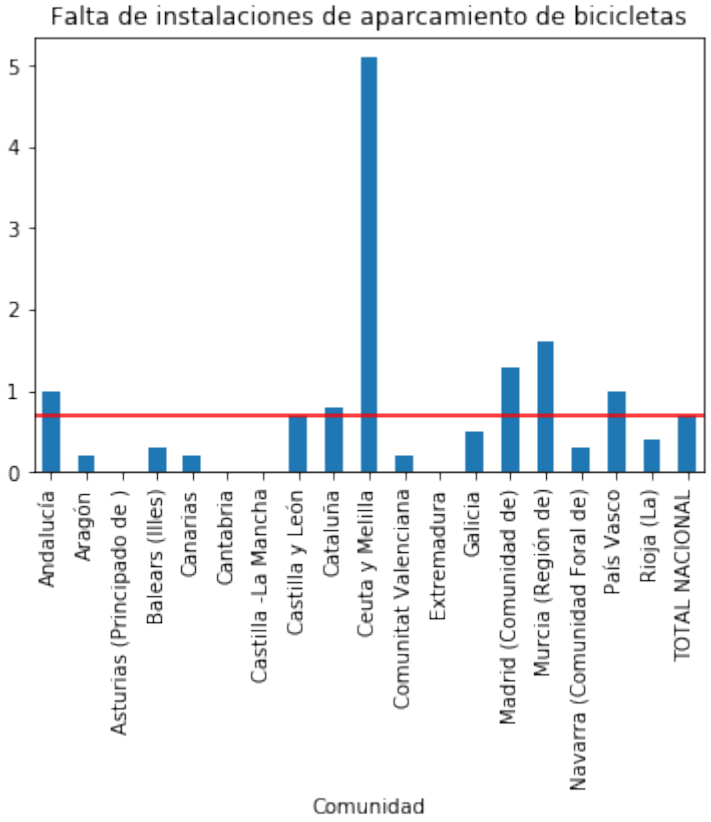
- A complete net of bike lanes is missing
- There is too much traffic
- There are not enough parking sites dedicated to bikes
- Personal security

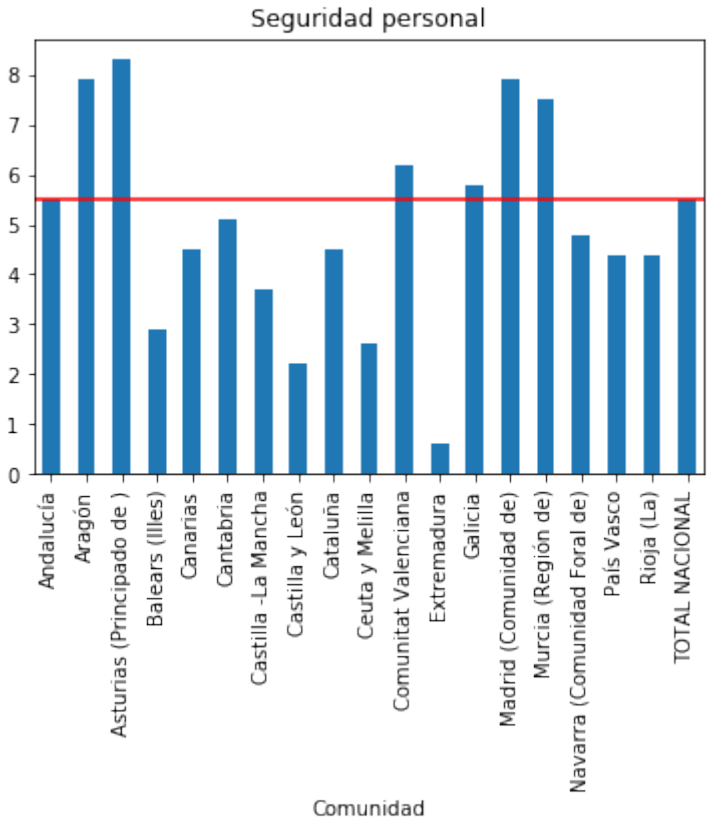
As can be seen looking at the plots below, where the y axis corresponds to the percentage of people saying they refuse to take the bike for a particular reason, such as the lack of bike lane in the first plot, even if Ceuta and Melilla show a very high interest in having a complete network of bike lanes and more bicycle parking sites, high levels of traffic and personal security are well below the national average (red line).

Considering the other autonomous community, Andalucía is among the ones that systematically have an attention level higher than the national average for the criteria we used. On top of this, and this is quite important for this work, the city of Malaga provides open-access data for air quality.




Demasiado tráfico
0
1
2
3
4
5
6
7
8
Andalucía
Aragón
Asturias (Principado de)
Balears (Illes)
Canarias
Cantabria
Castilla -La Mancha
Castilla y León
Cataluña
Ceuta y Melilla
Comunitat Valenciana
Extremadura
Galicia
Madrid (Comunidad de)
Murcia (Región de)
Navarra (Comunidad Foral de)
País Vasco
Rioja (La)
TOTAL NACIONAL
Comunidad


Falta de instalaciones de aparcamiento de bicicletas
0
1
2
3
4
5
Andalucía
Aragón
Asturias (Principado de)
Balears (Illes)
Canarias
Cantabria
Castilla -La Mancha
Castilla y León
Cataluña
Ceuta y Melilla
Comunitat Valenciana
Extremadura
Galicia
Madrid (Comunidad de)
Murcia (Región de)
Navarra (Comunidad Foral de)
País Vasco
Rioja (La)
TOTAL NACIONAL
Comunidad



8.2. Area selection

To select the area of the city where we want to install a new bike lane, we relied on the fact that the pollution level is the key: a highly-polluted area of the city is hosting too much traffic. If it is possible to move part of the fossil-fueled traffic to bicycle, through the incentive of a safe road, fully dedicated to bicycle, the air quality would benefit.

For this reason, we used a dataset provided by the Malaga government and with open access, listing for many substances present in the air:

- The numerical quantity, in the proper units, of the substance;
- A qualitative (categorical) indicator associated to the measured quantity of the substance, in terms of air goodness.

For each measurement, the geographic coordinates are also provided, so that it is eventually possible to build a sort of map of the city with the air quality information.

Looking at the data, it appears clear that using the numerical quantities is not straightforward:

- They may present different magnitudes
- It is not trivial to operate on them to extract results (e.g. get the average level of pollution of a point);
- A basic knowledge of the substances and of the quantities considered as dangerous or unhealthy is needed.

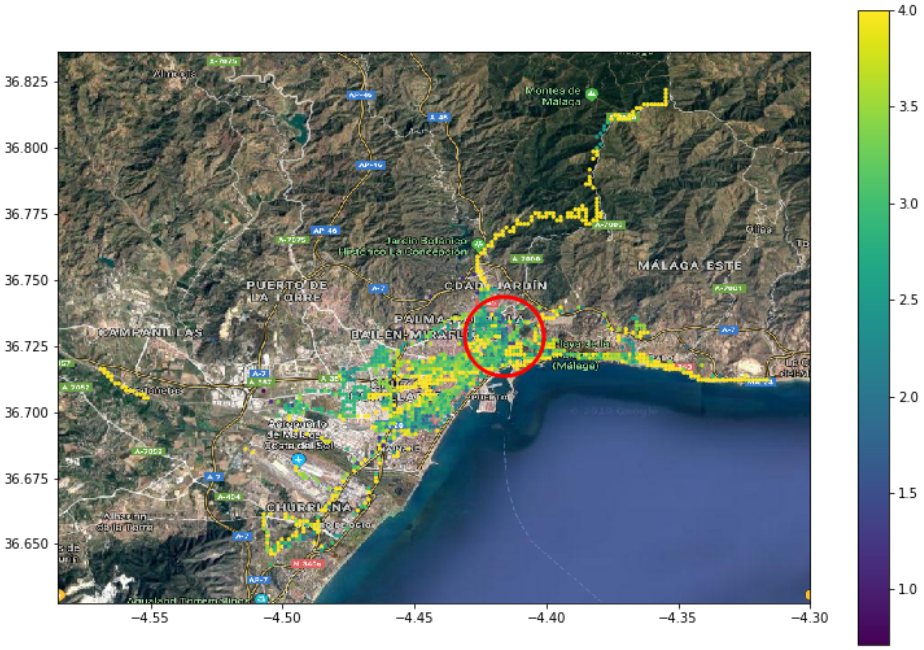
It is much easier, in this sense, to operate on the categorical data, since they already contain the information calibrated with respect to the recommended level of each substance in the air. On the other hand, a small curation work is needed to be able to perform an analysis on them.

Specifically, we have to associate to each qualitative indicator, a number, as follows:

- good = 4;
- moderate = 3;
- unhealthy-low = 2;
- unhealthy = 1;
- unhealthy-high = 0.

Extracting the average value of this air-quality indicator in every (x,y) point provided, it is thus possible to get a summary map of the air quality.

The new bicycle lane will be placed in the zone in which the air appears globally worse.



From the graphics, it looks like the best place to put a bicycle lane is in the center of the city.

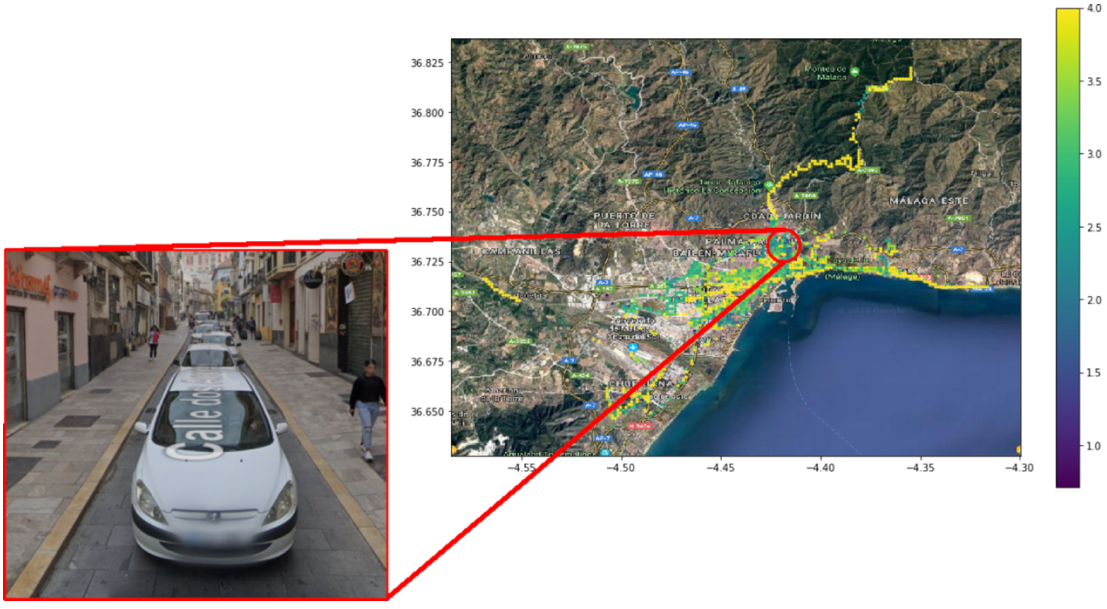
9. Conclusions

The data analysis performed in this project gives us a real solution that could be applied to reduce the contamination in the city of Malaga.

According to the data analyzed in this project, we find first of all that the best city to be studied in the context of this project was Malaga, because of the contamination in this city along with the traffic issues reported by its inhabitants and found in the first dataset analyzed. These results point that Malaga (Andalucía) is a city for which the traffic conditions can be improved, and for which we do already have open data available about the contamination in several locations, thanks to sensors which have already been installed a few years ago, therefore reducing the cost of such project.

According to the open data found on the Malaga open data portal, and our analysis of this data, the best place to put a bike line would be in the North area of this city, as shown in the next figure. Indeed, in this region, the air was considered to be quite bad in 2018, and the traffic seems to be a problem. There is moreover one small road already built that could be changed easily into a bike line for a minimal cost, which would definitely help a lot the air quality in the area while reducing noise contamination, and will probably not affect the traffic in the city in any way given the size of the street and the communication around it.

This project can definitely be improved if we do get some official funding at some point, by introducing the idea of asking the people what they think about the traffic condition in their city through the use of polls. The installation of additional sensors in the city might therefore be needed depending on the results of these polls, in order to pinpoint exactly the best location for the bike lane to be put.



References

[1] https://www.conserve-energy-future.com/causes-effects-solutions-of-air-pollution.php

[2] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[3] https://peopleforbikes.org/blog/protected-bike-lanes-do-not-cost-1-million-per-mile/