Data Management Planning: Santander Mobility

Data Life Cycle

Cédric Prieels
Eduardo Ruíz Ruíz
Fernando Solar Iglesias
Nicolò Trevisani

1.1. Description of data to be produced

We plan to collect different kinds of data, to comprehensively evaluate the needs of Santander in terms of traffic conditions. 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.

- Polls:
  - Citizens opinion.

1.2. How data will be acquired

Data will come from different sources. First of all, we will use open data created by the city of Santander and already accessible by everyone online. This will be our main source of data, as this is a way to get information about the transportations in Santander in a reliable and efficient way, without spending any money nor spending time creating experimental devices.

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 Santander (observational data), probably online in order to limit the cost of such a process. This will be a great way to learn a bit 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 the "Santander datos abiertos" portal (experimental/observational data).

Data will be acquired using

- Sensors;
- Polls;
- Open source data.

It might be needed to add some additional sensors depending on the project, which will take time. Redacting and propagating the polls will also take time, but we should have the data back in a matter of a few weeks.

1.3. 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 summarised 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 Santander, also supported by graphics to ease the interpretation.

1.4. 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, due to the fact that 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.

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

Open source data from Santander are already controlled before being published. We plan in any case to introduce tests to find possible outliers or issues in the data.

For the polls, we plan to put some control questions to understand if a person answered sensibly and 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.

1.6. Existing data

Open source data will mainly come from the Santander page located at: http://datos.santander.es/.

Here, several sources of information can be found, as:

- TUS schedule: http://datos.santander.es/dataset/?id=programacion-tus;
- Traffic data and traffic light control: http://datos.santander.es/dataset/?id=datos-trafico;
- Bike lane: http://datos.santander.es/dataset/?id=carril-bici.

As explained in the previous workflow section, this data will be combined with ours in order to create a new dataset.

1.7. 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 inaccesible, 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 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.