Synthetic data Design

To recreate street level pollution measured by mobile sensors (using as a reference the data that Reby sent us) we divide this process in two phases -the real data acquisition and the data actually measured by mobile sensors.

“Real Data”

1. Split the map in M square zones. The number M can vary depending on the necessities of the team. The lower the M is, the more uniform will be the map of pollution.
2. We give each of these M zones a mean value (e.g. 300 +- alpha, where alpha is a random value between a defined range). This way we have different mean values for each zone but similar values inside each zone.
3. Now each zone has a unique value, and so all scooters found in the same zone has this value. As this is not real, we must add bias between scooters inside a zone.
4. Each scooter has 3 NO2 sensors and 3 CO sensors. Real values would show the same quantity in each sensor of the same scooter.
5. To finish with the real data, we create a function which uses Poisson distribution , in order to increment pollution during rush hours and reduce it by night. This way we can reproduce the real 24 hour-periodicity of pollution.

5.1 Here we have thought of importing data from Base Stations to use this measured level of pollution and the date which the measure was done to compute time depending values around the map.

5.2 To simulate the diffusion as we move away from the base station we have thought of implementing a Gaussian distribution.

“Measured Data”

1. We create a function which introduces errors in each sensor, so the 3 sensors may measure different but similar values.
2. We create a function which adds a probability in each sensor to measure an extremely wrong value (for example in one scooter sensor 1 measured a CO level of 330, sensor 2 310 but sensor 3 measured 500, which is clearly a wrong measured level)