

|  |
| --- |
|  |

|  |
| --- |
| Big Data Programming Project – 5011 CEM |
| Low Level O3 Concentration Study |
|  |

# Introduction

The ambit of this project was the visualization of the data set provided. Therefore, there was a need for the creation of pictures or figures that could illustrate the different data points in that data set.

Knowing what the data was and what it represented the answer as obvious, the plot of the same data should be over the map of Europe, and to make the visualization easier there should be the delimitation of the countries, coastlines and seas. What overs the map is a color variance according to the numbers encountered in the data set.

To calculate better and verify the changes between each model provided it was also required to calculate the mean concentration of low level O3 daily on a mean position of Portugal, a bar chart was then generated to represent these models’ differences in values.

# Specification

In terms of specification the project ran smoothly according to what was expected on the appended spec sheet.

It requires a medium/low end PC to run so it is mainly accessible by any modern computer, with a standard dual core CPU that has at least a speed of around 2.0 GHz, it is recommended to use at least 8Gb of RAM since MatLab seems to be a bit memory intensive and when running multiple figures the extra RAM helps (I had to upgrade my system during the project), and around 5 Gb of storage available since the usage of MatLab was rather simple and no external libraries where needed to be used in the installation process. It is also recommended to use an Ubuntu based OS since it was the one used in the making of the project, but if another OS is available then the user just needs to refer to the MathWorks website and follow the installation instructions for the target OS.

The main disadvantage of the whole project was in terms of some visualization disabilities the user may have, such as color blindness. At the current most stable build there are no visual aids to fight that so that would possibly be a limitation of this project because of a remarkable time constraint due to poor time management and a small part of the COVID-19 outbreak.

Nothing that was out of scope was attended to since there was not even time to build all the ideas that were thought in the back for the project.

In terms of project management Notion was used since it provided an HTML project that simplifies the whole readability of the different figures, tables and tools used in the portfolio. Version Control with Git and Github was also a smooth operation.

# Code

The code was separated into two main points of focus, visualization and comparison.

Starting with the visualization the models are separated in different NetCDF files, each one contains a 3-dimensional matrix with a latitude, longitude and time variables, the value contained in each position of the 3D matrix is the value of the O3 concentration on that geographic position at that given time. So, after extracting these variables from the file, we start building our map, setting green for land, red for cities, the coastlines as black lines and the rivers in blue. To extract a mean out of the 3D matrix it is turned into a 2D matrix with the mean() function, it basically creates a mean for each point by using the third column in the matrix which is the time. Now we just need to plot the data as we find it in the matrix. Being X our longitude and Y our latitude we plot the new 2D matrix variable having each X and Y being colored on the map. It uses a color grading from blue to yellow has the concertation increases.

On the comparison code there was the need of joining each model on the same bar chart, so for each model a code was created that would take a latitude and longitude of three points in Portugal, one in the North, one in the South and one in the Centre. Then we loop through the longitude and latitude to try and find the correct indexes for the values per part of Portugal. We also loop through the 25 possible hours to create the mean per hours in those positions. Then a mean of the 3 positions is calculated and we get the mean concentration of O3 in Portugal. By joining all of the models code we just plot it on a bar chart with each respective model in the label.

In general both codes work as supposed, there were no matters left to address and therefore, no extra code creation would be required.

# Results and Conditions

To understand the outputs there must be an explanation beforehand.

For the figures the values are color coded, so a bluer color means that the value of low level O3 is lower, the more yellow it gets higher is the concentration. There must also be a general knowledge on how the Europe map looks and where the countries are since no country name labels exist.

The following figure is an example of how a mean of all the 25 hours would look if plotted (model: Chimere):

A picture containing text

Description automatically generated

By looking at this map it doesn’t really make sense that the eastern and northern European countries have more concentration than the rest of Europe, since the industry in Europe is well scattered and in terms of air pollution it would make sense for countries with more populational density to have more O3 concentration. This might due to a several number of factors such as the models’ inaccuracy, flow of winds, or just by it being a mean of all the values in the 25 hours.

The other type of output would be the differences between models by visualizing a bar chart. Here we have a concentration mean per hour in Portugal per each model:

A picture containing drawing

Description automatically generated

As we can see the models differ a little, which might again have several reasons for this occurrence, might be human error, sensors inaccuracy or just miss calculations or inefficiency of the calculation algorithms and equations. This type of data can help the models to try to generate more accurate results.

# Future Work

The project has a lot of areas to improve but the most significant one is in terms of accessibility, the figures should at least have different versions according to different color blindness dystrophy, or a controller in the figure that allows for size changing or other visual aids.

The MatLab code would need to be changed in order to change the color scheme but creating new figures would follow the same schematic that was used in the figures already created, it would just need a different color scheme. To improve the size of the image or create a controller in the figure I am not certain which process to follow since I didn’t research on it due to time constrains.

# Summary

The overall project objectives were achieved, it is possible visibly deduct some information out of the map and with the GIF image, that will be included in the code folder as an appendix, we can understand how the time of the day and the industry working times affect the overall concentration of the substance. We also are able to judge the models by looking at differences in values they have, without a base model that we would think as correct it’s impossible to securely chose the most accurate models but it is possible to extract the values an calculate a mean and then choose the model that looks closest to this same mean.

It lacks a bit of accessibility functionality but that leaves a safe room to work on in the future maintaining a ongoing relation with the client providing extra features and constant bug fixing and support.

# Appendices

## Specification Sheet (double click on the icon bellow).

