Effect of meteorological variables and air quality on SaRS-CoV-2 transmission

Master thesis presented by

Tania Villalba Sanchez

to obtain the Master's degree in Chemical Engineering from the Universitat Rovira i Virgili.

Company Supervisor: Carmen M. Torres Costa

URV Tutor: Manuel Martínez del Álamo

Tarragona, February 2022

I would like to express my gratitude to my supervisors Carmen and Manuel for the useful comments, remarks and engagement through the learning process of this master thesis.

Table of contents

Nomenclature

PROCEED transversal epidemiological prediction for the evolution management and use of resources in pandemics project

COVID-19 Coronavirus disease 2019

ABS *Àrees bàsiques de salut*

API Application programming interfaces

INE *Instituto nacional de Estadística*

XEMA *Xarxa d'Estacions Meterològiques Automàtiques*

XZP *Zones de qualitat de l'aire*

ZVPCA *Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica*

WAS Water, Air & Soil Technological Unit

CDTI Center for Industrial Technological Development

T Temperature

HR Relative humidity

RS Solar radiation

Y dependent variable in regression equation

X_k independent variable k in regression equation

β_k relative weight of variable k in regression equation

β_0 constant value in regression equation

ε random component in regression equation

q ratio of population stay-at-home

m ratio of population that make 1 or more trips

r ratio of population that make trips inside the same area

p ratio of population that make trips to other areas

delta difference on population that stay-at-home according to baseline from the 17th of February to 01st of the March 2020.

Summary

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. COVID-19 can be severe and has caused millions of deaths around the world and in addition chronic health problems in some who have survived the illness. Hence, the study of epidemiological models has been deepened in order to minimize the impact produced by the transmission of SARS-CoV-2.

This Final Master Project (FMP) is carried out within the framework of PROCEED, a public project funded by Center for Industrial Technological Development (CDTI) which involves the creation of an augmented epidemiological model that will support decision-making throughout the epidemiological management cycle. PROCEED is developed in EURECAT by different technological units with diverse expertise, each of them addressing different factors that may affect COVID-19 incidence.

The project focuses on Catalonia as region of interest to develop and validate this augmented epidemiological model and refers to the environment (meteorological and air quality conditions) as part of the many aspects investigated in PROCEED. Taking into account the information in the scientific literature, the meteorological and atmospheric pollution parameters may have an effect on COVID-19 incidence. The aim is to develop a computational model to predict the SARS-CoV-2 propagation associated to atmospheric environmental factors. Among the assessed atmospheric environmental factors, the meteorological variables include temperature and relative humidity of the air, wind speed, precipitation, and solar radiation; and air quality parameters involving the concentration of major pollutants in air such as suspended particles (PM10 and PM2.5), SO_2, NO_2 and O_3, among others.

The prediction model and its validation have been studied based on two time periods, lockdown and rest of the pandemic, to better isolate the effect of atmospheric environmental variables from mobility factors.

Hence, the lockdown period covers from the 1st of March 2020 to the 15th of May 2020 and the pandemic period covers from the first day in lockdown to the 15th of March 2021.

The model developed during this FMP includes epidemiological, meteorological and air quality data from public available databases.

Overall, it has been demonstrated that meteorological and atmospheric pollutants concentrations have correlate with the propagation of COVID-19 during the lockdown and pandemic periods. Concretely, the correlation analysis can reduce the variables used in the prediction model to temperature, relative humidity and solar radiation as meteorological parameters, and concentrations of NO, NO_2, PM10, PM2.5, SO_2, CO and O_3, as air quality parameters.

The best model to predict COVID-19 has been given by a specific *Àrees bàsiques de salut* (ABS). Moreover, it has been demonstrated an improvement taking in consideration the model to predict COVID-19 through a common vector composed by meteorological and atmospheric pollutants simultaneously by ABS.

In order to reconstruct better the COVID-19 propagation by a common framework through the air vector based on meteorological and air quality data parameters, a mobility factor has been introduced into the model to predict COVID-19 cases based on atmospheric environmental variables. The model has been validated by comparing the COVID-19 predicted cases to the actual ones.

1. INTRODUCTION AND BACKGROUND

The COVID-19 pandemic produced by the transmission of SARS-CoV-2 has spread rapidly and most of the worldwide population has been affected due to insufficient planification and coordinated pandemic response.

SARS-CoV-2 was officially announced by the World Health Organization on 31st of December 2019 and has already affected more than 120 million people throughout the world. The pandemic has rapidly expanded in most of the countries. In the first four months of 2020, more than 3 million cases of infections had been confirmed and more than 210.000 deaths. By the end of April, 212 countries, territories or areas had reported confirmed cases of COVID-19. In total, there are 2.147.090 cases of infection and 16.348 deaths reported in Catalonia in the period from start date of pandemic to February 2022.

In order to carry out efficient and coordinate epidemiological interventions it is crucial to have tools that optimize the management of resources necessary to control the pandemic and minimize negative impacts on society. Facing that situation, the study of epidemiological models from diverse knowledge fields has been deepened to minimize the impact produced by the fast transmission of SARS-CoV-2. Thus, these studies can contribute to complement and improve the decision-making in the different stages of COVID-19 incidence, resulting in a significant repercussion on the management of disease outbreaks and new pandemics.

Hence, the project on augmented epidemiological prediction for the evolution management and use of resources in pandemics (PROCEED) aims at generating a transversal epidemiological model from a multidisciplinary perspective with the cooperation of different Technological Units of Eurecat exploring multiple typologies of data: clinical data from the health system, population mobility from telephony data, social environment information from social networks use and telematic data, atmospheric environmental parameters from meteorological and air quality stations, and data related to the presence of viral load in waste water. Specifically, the tasks developed in the presented Master Thesis tackled the potential contribution of atmospheric environmental parameters to a computational model to predict the risk of COVID-19 propagation as part of the PROCEED project.

A literature review has been performed in order to analyze the information sources related to epidemiological, meteorological and air quality parameters, as well as to update the state of the art in the field of COVID-19 propagation and its relation with atmospheric environmental parameters.

Regarding the hypothetical influence of meteorological and air quality parameters on COVID-19 incidence.

Preliminary studies have been recently published correlating air pollution and meteorological parameters with the number of people affected by COVID-19 and their mortality (Lorenzo et al. (2021); Zhao et al. (2021); Huang et al. (2020) and Zoran et al. (2021)). Specifically, Lorenzo et al., (2021) noticed a significantly positive association between the air pollutants concentration as nitrogen dioxide (NO_2), suspended particles (PM2.5) and temperature with COVID-19 number of cases. Conversely, PM10, ozone (O_3), sulfur dioxide (SO_2), carbon monoxide (CO), rainfall and humidity were significantly associated with lower confirmed of COVID-19 infections.

Similar results were found by Zhao et al. (2021) in reference to the emerging role of air pollution and meteorological parameters in COVID-19 evolution, which corroborates the fact that the air pollution exposure is related to the increase of COVID-19 cases and that,

additionally, the changes produced by the meteorological variables are also a significant factor. Concretely, Zhao et al. (2021) found a positive correlation for suspended particles, O_3 and wind speed while temperature has a negative correlation. Inconsistent results were found in case of SO_2, carbon monoxide (CO), carbon dioxide (CO_2) and relative humidity. Regarding the implications of air pollution on the evolution of COVID-19 cases, Marques et al., (2021) presented a review of previous studies concluding that there is an evident and clear association between air concentration of certain pollutants (PM2.5, PM10, O_3, NO_2, SO_2 and CO) and the negative incidence and severity of the COVID-19 cases.

Other publications, such as Brandt et al. (2020), mention the evidence linking air pollution exposure with the COVID-19 severity, due to the direct affectation to lungs and the contribution to cardiopulmonary disease. It was observed a positive association between pollution measurements and COVID-19 fatality rates in China.

To conclude, considering the several scientific articles on the association between meteorological parameters and air pollution in the transmission of SARS-CoV-2, the initial hypothesis is that there is an influence of meteorological and air pollutants parameters on the incidence in COVID-19 may be true. It is relevant to mention that the previous studies were based on data from different locations that differ on meteorological, socioeconomic and geographical conditions. Moreover, the investigation was carried out applying different methodologies and over a different population, so it is not easy to extract straightforward conclusions.

Most publications concluded that low temperature and humidity favor the transmission, while in other studies a link has been established between the concentration of pollutants in air and the impact of COVID-19.

Taking into account the information found in the literature, the meteorological parameters studied in the present thesis have been wind speed, temperature, relative humidity, rainfall and solar radiation. Regarding the air pollution parameters, the pollutants have been analyzed, namely, suspended particles (PM1, PM2.5, PM10), Cl_2, Cl, H_2S, As, Pb, NO_X, NO_2, NO, HCl, SO_2, O_3, C_6H_6, CO, BaP, Ni, Cd and Hg.

2. SCOPE OF THE PROJECT AND SPECIFIC OBJECTIVES

The aim of the project is to develop a model for studying the propagation of SARS-CoV-2. In order to reach that goal, epidemiological, meteorological and air quality data were analyzed including lockdown and pandemic time frames in Catalonia. Hence, lockdown period covers from 1st of March 2020 to 15th of May 2020 and pandemic period from the first day of lockdown to 15th of March 2021.

Then, an existing draft model based on municipalities was modified to obtain COVID-19 and meteorological and atmospheric pollution data correlation by *Àrees bàsiques de salut* (ABS) and to develop a regression fit.

The specific objectives in this project are the following:

1. Analysis of the sources of epidemiologic, meteorological and atmospheric variables and literature review to update the information related with the topic.

2. Implement automatic input data extraction (clinical, meteorological and air quality data) through web scraping.

3. Automation of epidemiological, meteorological and pollutants concentration in air data in the current model assessing COVID-19 with meteorological and atmospheric pollution by web scraping.

4. Adapt geographical resolution of the model assessing COVID-19 correlation with air quality and meteorological data based on ABS.

5. Screening of the meteorological and air quality parameter, with the identification of the variables with higher dependency.

6. Selection of scenarios to develop the regression fit. This includes setting the time and geographical scope, and the offset (the time lapse between the day of the measurement of the atmospheric environmental parameters and the day when COVID-19 cases were registered).

7. Applying linear regression fit for the reconstruction of the COVID-19 cases in the selected scenarios and with the selected atmospheric environmental variables.

8. Develop a common framework to predict propagation of COVID-19 through the air vector based on atmospheric environmental parameters and investigate its relative contribution by introducing the effect of mobility data in the model during the lockdown and pandemic periods under the studied scenarios.

3. STUDENT'S ROLE IN COMPANY

Eurecat is a technological center in Catalonia which is an innovative and differential technology provider whose aim is to cover the innovation needs and boost technological competitiveness in the business world. The industrial technology provider in Catalonia offers business services which are applied Research and Development, technology services and consulting, trainings, product and service innovative development, promotion and dissemination about the innovative technology.

The field of knowledge are industrial such as chemical technology, digital, biotechnology and sustainability areas.

Concerning the sustainability area, Eurecat develops and optimizes technologies and processes to enhance the water, soil, air and waste, energy and environmental impact.

Focus on the Water, Air & Soil Technological Unit also named WAS, it enables to develop and optimize innovative technology and process for optimizing water, treating industrial emission in air and soil and groundwater management and associated resources such as simulation and modeling capabilities for treatment processes or developing models with generic languages. Moreover, this Technological Unit provides solutions for complex and specific problems in a wide range of sectors including water treatment, chemicals, food and agriculture, pharmaceuticals and materials processing and capital equipment.

It should be pointed out that Eurecat mission is becoming a key agent in public-private cooperation within the research and innovation area.

The internal practices in Eurecat were developed in WAS, specifically in the modelling line with expertise in computer-aided process engineering. WAS is involved in the PROCEED project, a public project also developed by diverse Technological Units of Eurecat with different specialized knowledge in order to create an augmented epidemiological model that predicts the COVID-19 propagation. The internal practices scope is in reference to atmospheric environmental variables as part of diverse aspects investigated in the PROCEED project.

4. METHODS: COMPUTATIONAL MODEL

The computational model is composed by their processing, correlation analysis and model regression and fit which is programmed in Python using Spyder as IDE tool (Anaconda distribution).

4.1 Data extraction automation

In order to automate and simplify the data extraction process from different websites, web scraping has been carried out.

Web scraping, web harvesting, or web data extraction is a data scraping technique used for extracting data from websites. The term typically refers to automated processes software that may directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser It is a form of copying in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

In this case, Pandas library is used to extract the data and store in a dataframe and Socrata library for web scraping. Hence, application programming interfaces (API) of Socrata open access data allow programmatic access to a dataset, including the ability to filter, search and add data. The communication with API is done through HTTP, and coded in Python to obtain the data in xml and csv format using Pandas library.

Therefore, web scraping can perform a task in a few lines of code saving time, the data is not lost and easy to call another code which are the main benefits of implementing this technique to obtain information.

Hence, epidemiological, meteorological and atmospheric pollution data are needed to carry out the study of the effect of meteorological variables and air quality on SaRS-CoV-2 transmission.

The data will be analyzed from lockdown and pandemic periods in Catalonia so, the website used is Catàleg de *Dades Obertes de Salut* corresponding to *Generalitat de Catalunya*. As aforementioned, lockdown period is established from 1st of March to 15 of March 2020 and pandemic period is analyzed from 1st March 2020 to 15 of March 2021.

Regarding the epidemiological data, the COVID-19 record cases in Catalonia identified as positive are extracted by diagnostic test or meteorological study by day, municipality and sex from *Departament de Salud i del Servei Català.*

In reference on meteorological data, it is registered in the station of la *Xarxa d'Estacions Meterològiques Automàtiques* (XEMA) from *Servei Meteorològic de Catalunya* have been used. It should be noted that XEMA variables selected for the study have been wind speed, temperature, relative humidity, precipitation and solar radiation due to the fact that these are the most relevant variables according to the literature review carried out in Section 1. Meteorological data record contains these variables that are measured with a frequency of every hour per day.

Concerning the study of the data of pollutants in air, two types of measurements are taken into account, according to the *Catàleg de Dades Obertes de Salut*. Thus, the concentration level of air pollutants is measured at the automatic and manual measurements points of the *Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica* (ZVPCA). The air pollutants concentration data to be studied are suspended particles (PM1, PM2.5 and PM10), SO_2, NO_X and O_3, among other.

Pandas library was used to configure the data into a dataframe (data matrix), which will make it easier to read and analyze the data. Regarding Socrata is the package which allow access in open data resources from the organization, in this case, from *Dades Obertes de Salut*. Thus, the code set the variable to the URL of the *Catàleg de Dades Obertes de Salut* website of *Generalitat de Catalunya*. The code use "get" for the next line to insert the specific URL endpoint in order to specify the epidemiological, meteorological and air pollutant data and at the same time, focus on the data extraction with filtering by date in the period of lockdown or pandemic.

In the case of the meteorological web scraping, the sorted applied has been for dates and variables of interest due to the fact that the large amount of information.

To end up, the sorted data in the interests' websites displays a dataframe which will be an easy way to import filtered data.

Meteorological, epidemiological and manual and automatically air quality web scraping can be seen in Section 8.1.

4.2 Geographical resolution by ABS

The current model assessing COVID-19 and meteorological and air pollutant concentration data correlation are in reference to the municipalities, so in order to work in parallel with the different technological units which are composed by PROCEED, *Àrees bàsiques de salut* (ABS) have to be the geographical resolution for the model.

Àrees bàsiques de salut (ABS) is the elementary territorial unit in Catalonia through which primary health care services are organized. This specific geographical resolution is composed by districts in urban areas or by one or more municipalities in rural areas.

Hence, the relation between the health region, health sector and ABS are extract taking into account the *Institut d'Estadística de Catalunya* (Indescat).

In order to adapt the model assessing COVID-19 and meteorological and air pollutants concentration data correlation with ABS, it is required a document file where the relationship between municipalities and ABS is established.

4.3 Data pre-processing

The technique of web scrapping extracts the information directly from the interested website in the unstructured format, but it helps to collect this unstructured input and convert it in a structured form. However, the different extracted data require pre-processing to work easily and reduce the large number of records.

Moreover, meteorological, air quality and epidemiologic data have to be related with the ABS geographical resolution before to be introduced in a correlation model.

In the subsections below the data treatment of meteorological, air pollutants concentration and epidemiologic data are shown.

4.3.1 Meteorological data

In this part is relevant to create an excel file with the correspondence between stations, municipalities and code and description ABS which will be the basis of preprocessing code to obtain the data by ABS.

Input files are meteorological web scrapping and the Excel file named *Location_ABS_Meteo_stations.xls*.

In order to obtain the meteorological data with ABS, a preprocessing code is done. Basically, inputs files mentioned above are imported in the code and it checked the format of the data import and if it is necessary, the format to numeric or datetime are modified.

In this case, meteorological data has valued every 30 minutes per day, so in order to merge the data in the future, it is required to change the date in format of day, month and year independent of the time to make it easy the process.

Hence, the code to associate the meteorological data and ABS geographical resolution is started to create a dataframe. This is filled in with the ABS code and corresponding description, providing a desirable dataframe in csv format, named *Meteo_ABS_webscrapping_v1.csv*.

Finally, the output format of the meteorological data processing is a csv file which will be the input document in a COVID-19 and meteorological data correlation code in Python.

Similar procedure is applied for extracting meteorological data during lockdown and pandemic, the unique difference will be in the date selection depending on analysis period.

The flowchart of the explained meteorological data preprocessing can be seen in Figure 4.1 and the code in Python called *DataPreProcesingMeteo_ABS* in Section 8.2.1

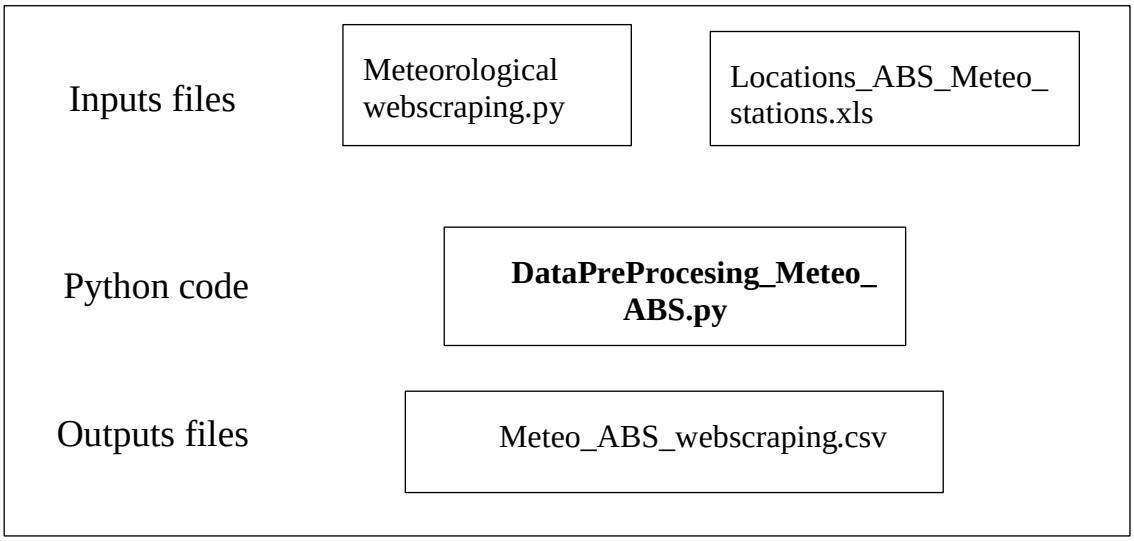

Figure 4.1 Flowchart of the data preprocessing code.

4.3.2 Air quality data

The processing in air quality data has been more complex than for meteorological data due to the fact that these data are composed by automatic and manual measurements. This implies processing data from automatic and manual stations individually, and then apply a data merge code unify both measurements in a single formatted data input file.

A file reflecting the reciprocation between the air pollutant stations and municipalities was based on the metadata of the datasets of the *Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica (XVPCA)* for each of the quality air regions (*Zones de qualitat de l'aire (ZQA))* according to *Medi Ambient i Sostenibilitat* in *Generalitat de Catalunya.* By this way, it can relate the data extract of the air quality web scraping with the air pollutant station and INE code and finally, with the appropriate ABS.

Other inputs files which are imported in the processing code are the web scrapings of air quality data for automatic and manual measurements which are named *Auto_cont_webscrapping.py* and *Manual_cont_webscrapping.py*. As it mentioned above, it is required to check the format of each parameter in web scraping in order to manipulate easily the data.

Concerning the part of the preprocessing code of automatic measurement of air quality parameters, data of the specific contaminant are measured every hour, so in order to obtain a daily value, the average of these hours has been done.

Once the automatic data are in format, it is needed to do a for-loops in order to relate and obtain the desirable outputs files.

Thus, a pair of for-loops are executed, the first loop in range of INE/Municipality code which objective is compared every INE code from automatic web scraping and INE code from the input file which related the municipalities, INE code and ABS. The comparison is done for every single INE code and then, the corresponded ABS is added. The second loop is in range of the pollutants in air variables, following the same procedure that in the first loop, the air pollutant parameters from web scraping is compared with the list of all contaminants.

In order to simplify and order the dataframe from the two principal loops, a third for loop is executed, concretely the function of this loop is organized the dataframe by dates and joined it by the same date all the data, taking into account that in the case of air pollutants values have diverse values in the same date, so the average is done.

Finally, the dataframe of automatic processing part is exported in a excel file being this the automatic air quality data output file where the data are organized by data and air pollutant parameter and related with ABS.

On the other hand, another part of the preprocessing code has been the air quality processing measuring manually which is essential to change the format in the data import by the web scraping. Regarding the date, it is distributed the air pollutant measurements each day which is composed by 31 columns in the dataframe, so in order to reduce the dataframe and have the data that follow the same rule that in other data (day-month-year), for-loop is executed. This for-loop has been in range of the manual air pollutant dataframe and consisted in recopilate the data of each day and created a corresponding completed date which is filled with the respective pollutant in air concentration value.

Once the date has been in format, the same for-loops of the automatic part in this preprocessing code are executed, it means that, the first loop has been to relate the INE code of the manual web scraping and added the corresponding ASB code and description and the second one is related air pollutants concentration data with their respective values of concentration. Hence, the last for loop is joined the data by day and applied the average in air pollutant concentration data in case of having different values in the same data.

Before the exportation of excel file, it is relevant to make the data filter due to the fact to be consistent with the automatic air quality output data.

To end up in the manual processing part, a second output file is obtained in the preprocessing code, in this case, the manual air quality output data.

To sum up this part, the processing code is composed by automatic and manual measurement which are executed individually in order to get the automatic and manual air quality output data file separately.

This output files are called *Contaminacion_atmos_2020_2021_manual_ord_V1.xls* and *Contaminacion_atmos_2020_2021_auto_ord_V1.xls.*

As a way to understand the procedure explain above, a flowchart in processing code of air quality data is indicated in Figure 4.2.

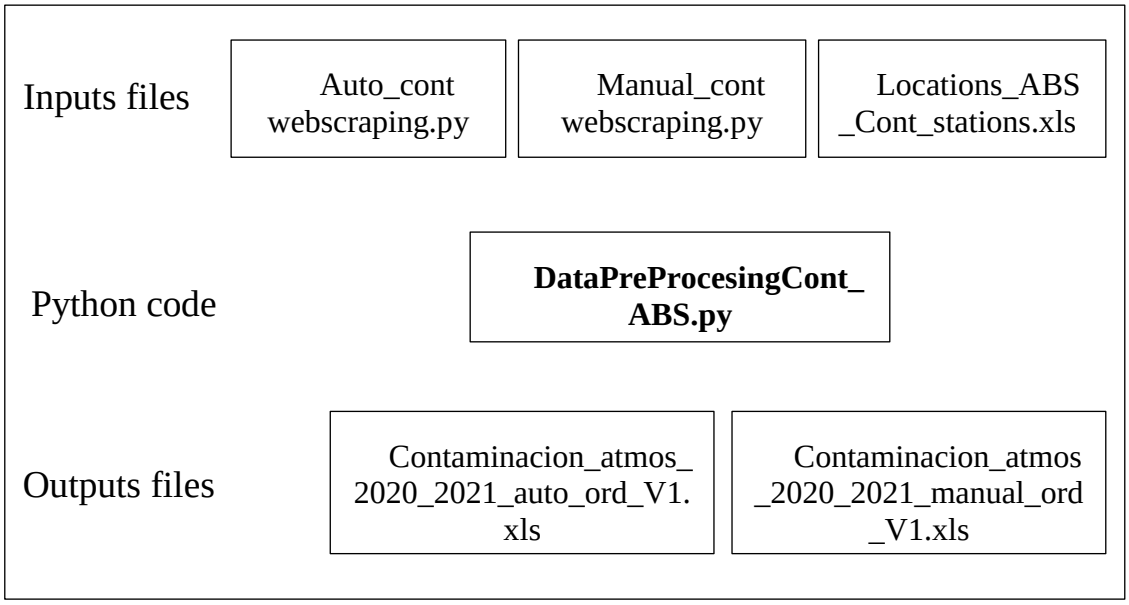

Figure 4.2 Flowchart of the air quality data processing code.

The explained code is called *DataPreProcesingCont_ABS* can be seen in Section 8.2.2.

Concerning, the data merge code is necessary to merge the output excels files obtained in the processing code, automatic and manual air quality data files.

Thus, inputs files will be the excel file of automatic and manual air pollutants measurements from the air quality preprocessing code.

The first step is created a dataframe with the columns of interest in this case, ABS code, ABS name, air pollutants and value of concentration merge to fill out with the two inputs files. The merge is consisted in compare the list of pollutants in air and ABS code are extracted from both inputs' files with the data of the automatic and manual excel file from the processing code.

Once both requirements are achieved, the code is calculated the average of the air pollutants concentration data from automatic and manual data and inserted the respective ABS description.

To do so, for-loops of ABS code and air pollutants data are compared to automatic data and manual data and linked the value of air pollutant concentration data used an average. In the case of relating both inputs file and one of them not have a pollutant in air concentration value, the average will not do, and the unique value will be selected.

To end up, the dataframe created for the merge values is converted to an excel file which has the automatic and manual measurements of air quality merge (*Contaminacion_atmos_2020_2021_MERGED_V1.xls*). In Figure 4.3, it can be displayed the flowchart of this merge code.

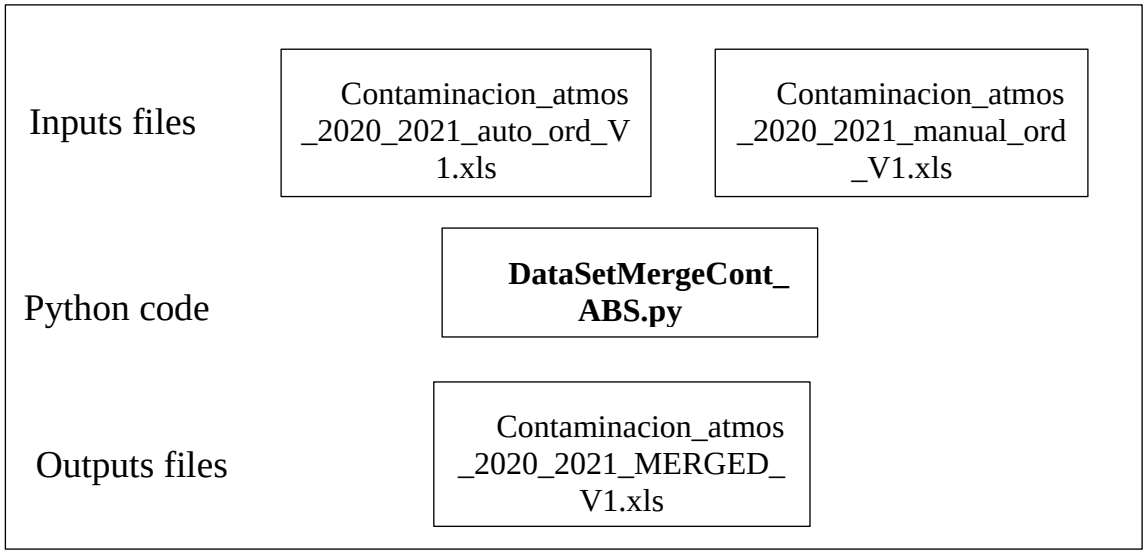

Figure 4.3 Flowchart of atmospheric pollutants data merge code.

The merge code in air pollutant's part named *DataMergeCont_ABS* can be seen in Section 8.2.3.

The mentioned procedure for atmospheric pollutant is the same for lockdown and pandemic periods, the difference will be the selected dates for the analysis.

4.3.3 Epidemiological data

The data processing in epidemiological data has been easy due to the fact that it is possible to obtain *Registre de casos de COVID-19 a Catalunya per Area bàsica de salut (ABS) i sexe* in *Catàleg de Dades Obertes de Salut* website. These data show for each day the number of COVID-19 cases identified, ABS and gender in Catalonia. In case where it has no been possible to identify the ABS of the person identified as a positive case, the value of the variable "ABS Description" is "Unclassified".

Data filter is applied in epidemiological web scraping in order to obtain the data in the different studied periods, lockdown and pandemic. After that, the data are in object format, so it is essential to modify the format in numbers and the date in datetime.

4.4 Correlation models

The aim of this section is to create a model assessing COVID-19 incidence and meteorological and atmospheric pollutant concentration data correlation for all *Àrees bàsiques de* salut (ABS) in Catalonia.

In order to study more accurately, the mentioned models between meteorological and air pollutants parameters with COVID-19 incidence are done separately.

The subsection below explains the procedure followed for the studies of COVID-19 incidence and meteorological variables and COVID-19 incidence and air pollutants parameters in lockdown period. Moreover, it should be pointed out that the explained codes are valid for pandemic period by selecting a different date.

4.4.1 Correlation of COVID-19 cases and meteorological data

Correlation of COVID-19 incidence and meteorological data has been essential to know the influence of changes in meteorological variables that can be affect the incidence of COVID-19.

In order to reproduce the model correlation, it is done by calculating the coefficient between COVID-19 incidence and each of the meteorological variables varying the offset of the meteorological variables up to 30 days for all ABS in Catalonia.

The meteorological variables studied are wind velocity, temperature, relative humidity, rainfall and solar radiation.

Input files in the COVID-19 incidence and meteorological variables correlation code have been epidemiologic web scraping (*ABS_clinical_webscraping.py*) which is indicated the classification of the number of cases by ABS and document file in csv named *Meteo_webscrapping_v1* from the processing meteorological code where is described the values of the meteorological data for each ABS and the Location_ABS_Meteo_stations excel file.

COVID-19 incidence and meteorological data correlation model is started with the definition of the parameters requires for the code. In this case, it is needed to define the analysis period which is the start and end date in lockdown or the pandemic period depending on when the study is focused. Moreover, it is defined a series of offsets in order to study the correlation varying the offsets. At the same time, a dataframe is created in order to store meteorological data by ABS and date with datasets of desirable offset, finally, this dataset is exported to an excel file which are named *Timeseries_Conf_Offset_Meteo_offset save.out.xlsx*.

After that, input files are called from preprocessing output, it means that, epidemiological web scraping and Meteo_webscrapping_v1 csv document and it is defined the meteorological variables to taking into account which are mentioned above. Then, dataframes are created in order to store the desirable data at the end of the code, in this case, the INE code, Municipality, ABS name and description, offset or temporal deviation to found the variables performance (from 0 to 30 days) and the meteorological variables (30 wind velocity, 32 temperature, 33 relative humidity, 34 rainfall and 35 solar radiation).

The model includes the following tasks:

- Setting the time frame, locations specified in ABS, selected meteorological data and offset range.
- For each ABS extract and order the clinical and meteorological data according to the time frame from pre-processing input data files.
- Calculation of the Pearson correlation for each offset.
- Arrange the output data in a matrix (municipalities, ABS, Offset and each meteorological variable), obtaining for each ABS and offset combination with correlation coefficients of each meteorological variable.
- Import the output data in an excel file called *test_AllABS_meteo_alloffsets.xls*.

The flowchart of the procedure explained about the COVID-19 incidence and meteorological variables model correlation can be seen in Figure 4.4 and the code, named *Meteo_correlation_ABS.py*, is displayed in Section 8.3.1

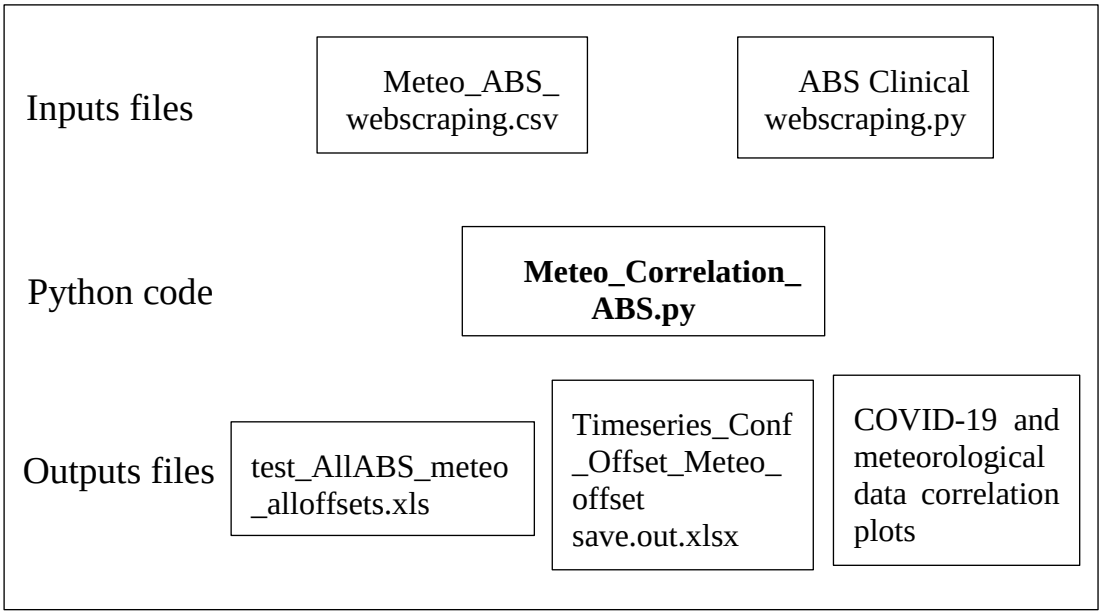

Figure 4.4 Flowchart of COVID-19 incidence and meteorological data correlation code.

Correlation of COVID-19 incidence and air pollutants data follow the same procedure to correlation of COVID-19 incidence and meteorological data. The unique difference is the variables. Air pollutants parameters taking into account in this model are suspended particles (PM1, PM2.5, PM10), Cl_2, Cl, H_2S, As, Pb, NO_X, NO_2, NO, HCl, SO_2, O_3, C_6H_6, CO, BaP, Ni, Cd and Hg which are the pollutants that composed by the data extract from automatic and manual measurements of air quality in Catàleg de *Dades Obertes de Salut* website. After the study, some of these pollutants in air will be ruled out depending on their influence in the incidence of COVID-19.

To end up, COVID-19 incidence and air pollutants data correlation code flowchart can be seen in Figure 4.5.

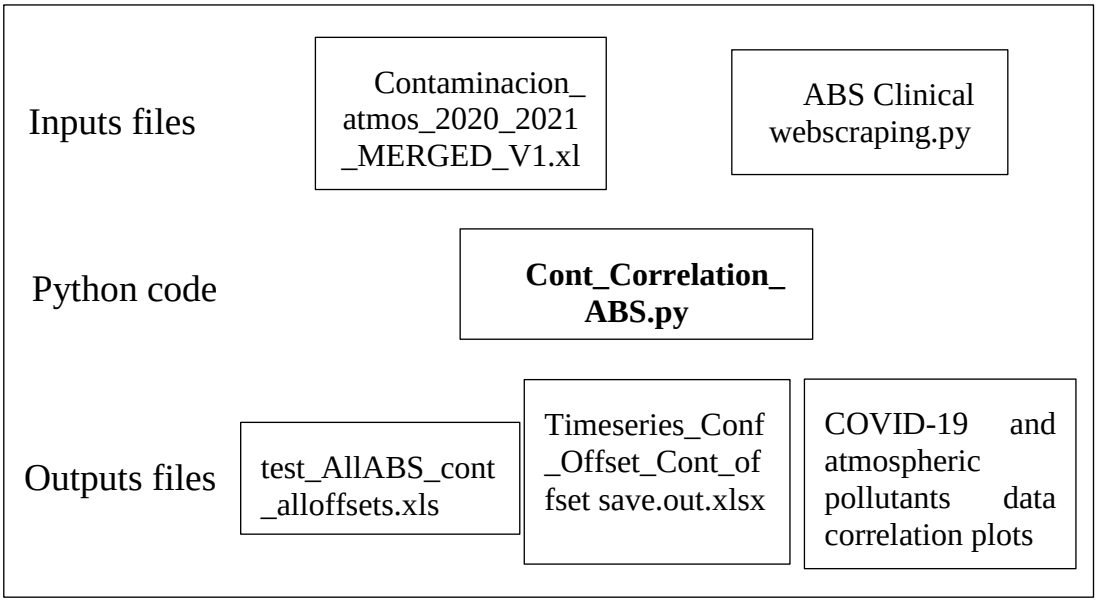

Figure 4.5 Flowchart of COVID-19 incidence and air quality data correlation code.

4.5 Offset analysis

The purpose of the offset analysis is to assess the existence of offset patterns that enables the selection of a unique offset that best fit in a potential predictive model.

To do so, a Python script was created to obtain the maximum value of correlation filtering for the maximum value correlation in absolute value. This study is done individually depending on the variables to be analyzed, it means that, two codes are done, one for the meteorological data and another for atmospheric pollutant data.

It is relevant to explain that the meteorological and atmospheric pollutant offset analysis codes follow the same procedure, but the unique difference is the inputs files and the variables to take into account. Thus, the general code is explained specifying the differences.

The model includes the following tasks:

- Import the matrix from correlation models to set offset combination with correlation coefficients of each variable by ABS.
- Setting the time frame, ABS locations and selected meteorological or air quality data.
- Calculation of maximum offset in absolute value for each ABS and variable.
- Arrange the output data in a matrix obtaining the maximum offset with correlation coefficients of each variable by each specific ABS.
- Import the output data in an excel file.

To end up, the flowchart of the procedure is displayed in Figure 4.6 where to the left-hand side the meteorological part is shown and in the right-hand side the air quality parameter.

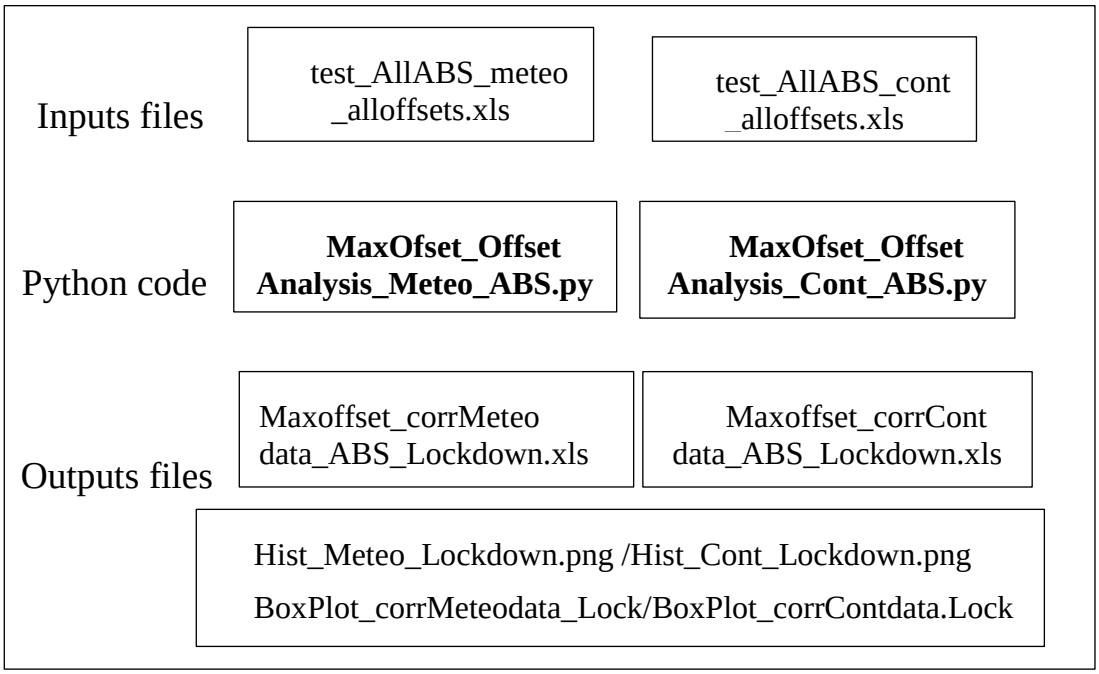

Figure 4.6 Flowchart of maximum values of correlation model and offset analysis from maximum correlation code (Left-hand side: meteorological correlation model study and Right-hand side: Air pollutant correlation model).

In order to understand both procedures, the code *MaxOffset_OffsetAnalysis_Meteo_ABS.py* is shown in Section 8.4.1 which can be used for *MaxOffset_OffsetAnalysis_Cont_ABS.py* changing the input files and variables.

The selection of unique offset for each ABS is done by a code named *OffsetSelection_Meteo_ABS.py* can be found in the Section 8.4.1 which are the same that *OffsetSelection_Cont_ABS.py* and the flowchart is in the Figure 4.7 where it can be seen selected offset code by meteorological and air quality data model correlation.

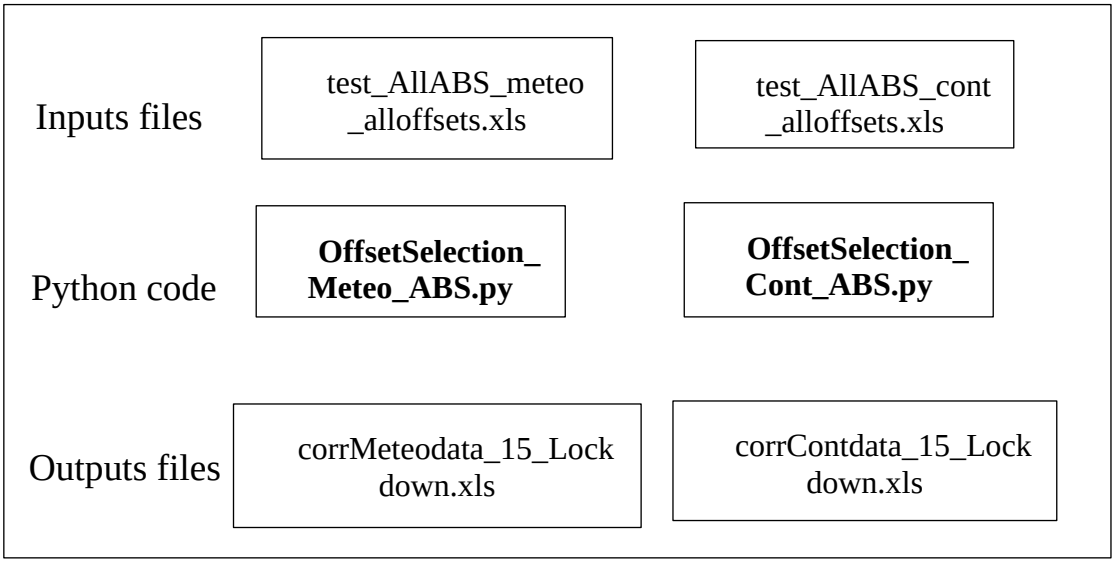

Figure 4.7 Flowchart of selected offset code in reference to meteorological and air quality model correlation. (Left-hand side: meteorological offset selection and Right-hand side: Air pollutant offset selection).

4.6 Multiple linear regression

Once a specific offset is selected, with the correlation models, the atmospheric environmental parameters (meteorological and air quality) were screened, and the offset was selected. The next step is applied a multiple linear regression to the construct a prediction model.

To analyse the possible scenarios the regression fit is applied separately, starting with meteorological variables and adding complexity

The regression fit is composed by a processing code due to the fact that it is needed to process the data in a certain way and the regression fit code is done.

In the following subsection, it will explain the procedure done for this part, highlighted that the model assessing COVID-19 with meteorological and air pollutant data carry out separately. Moreover, it is relevant to develop a common framework with the environment and air pollution to predict the COVID-19 incidence.

In order to evaluate the data during pandemic and lockdown, the aforementioned multiple linear regression model is valid for both cases by applying different analysis period.

4.6.1 Processing regression code

Processing regression code is required in order to introduce the data extract from these code in the regression fit code.

As it mentions before, processing regression code is the same for the meteorological and atmospheric pollutant model regression, the unique difference is the input files and the variables to take into account.

Basically, the aim of this code is imported the input data, processing these data in a specific way and merge the possessing data.

The meteorological processing regression code input data is composed by the *Location_ABS_Meteo_stations.xlsx* which are the relationship between the municipalities, ABS and the meteorological station in Catalonia, the *Timeseries_Conf_Offset_Meteo_15_out.xls* data extract from the code of regression model (Section 4.4.1) which are the values of the variable, COVID-19 cases by ABS and each variable with the selected offset and the *epidemiological web scraping*.

Regarding the input data of the atmospheric pollutant processing regression code, the excel file *Location_ABS_Cont_stations.xls*, *Timeseries_Conf_Offset_Cont_15_out.xls* and the *epidemiological web scraping* is used.

After that, variables had to be defined, in case of meteorological variables will be temperature, relative humidity and solar radiation, and referring to air pollutant variables will be defined suspended particles (PM10 and PM2.5), SO_2, NO, NO_2, CO and O_3.

Hence, the processing regression code is started with the adequacy of the epidemiological data. The epidemiological web scraping by ABS is imported and modified the format of the dataframe in case of date to datetime and object to a numeric. Then, a dataframe is created in order to group the COVID-19 case by ABS and date by the summatory of them and the columns are renamed to make it easier the posterior merge of the data.

Concerning the processing of the meteorological data extract from the *Timeseries_Conf_Offset_Meteo_15_out.xls*, input all the data in the sheets and gather together the ABS code and air pollutants parameters making a list of sets.

A dataframe is created by each meteorological and ABS which are collected variable value and the respective ABS code by a for loop, then it is related with ABS code to obtain a dataframe ordered with date. A compilation of that is done in order to obtain a dataframe with date, value of each meteorological variable and related to ABS code.

In reference to the air pollutants code, the epidemiological processing data is carried out identically and air pollutants parameters are done similar to the meteorological processing but using the pollutants in air variables.

Finally, the dataframe created for COVID-19 data is merged with the last dataframe related with the meteorological variables, getting the dataframe necessary to introduce in the Regression code is exported to excel (*TimeSeries_Meteo_Conf_Offset_15_Regresion.xlsx*).

Following the same procedure, it is obtained a desirable dataframe in reference to the air pollutants variables which is named *TimeSeries_Cont_Conf_Offset_15_Regresion.xlsx*.

In order to see in more details, the procedure of Processing Regression code regarding the meteorological part is in Section 8.5.1.

The processing regression flowchart can be seen in Figure 4.8 and Figure 4.9 the meteorological processing regression code (*PreprocessingRegression_Meteo_ABS.py*) and air quality processing regression code *(PreprocessingRegression_Cont_ABS.py*), respectively.

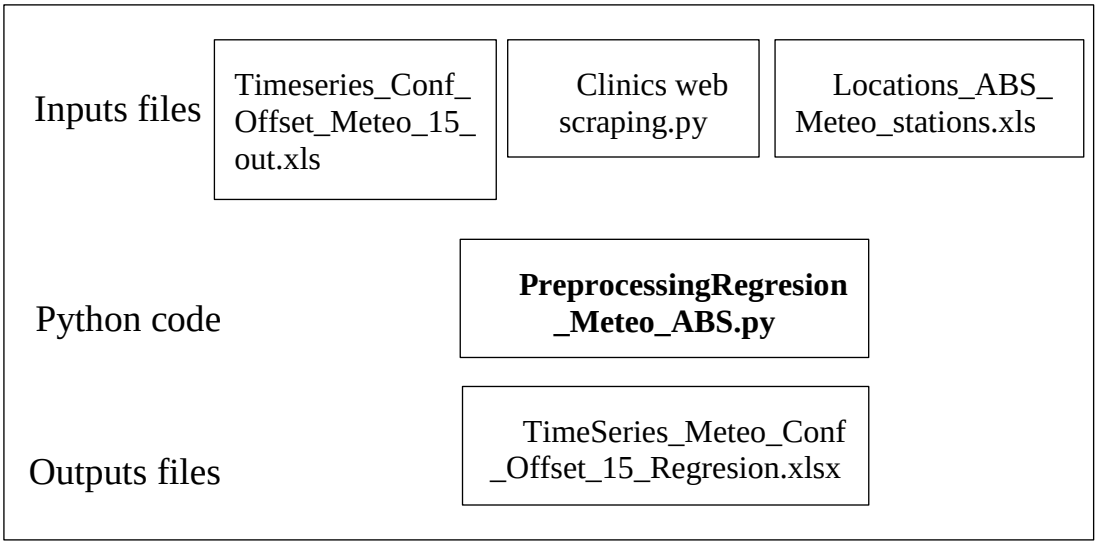

Figure 4.8 Flowchart of meteorological preprocessing regression code.

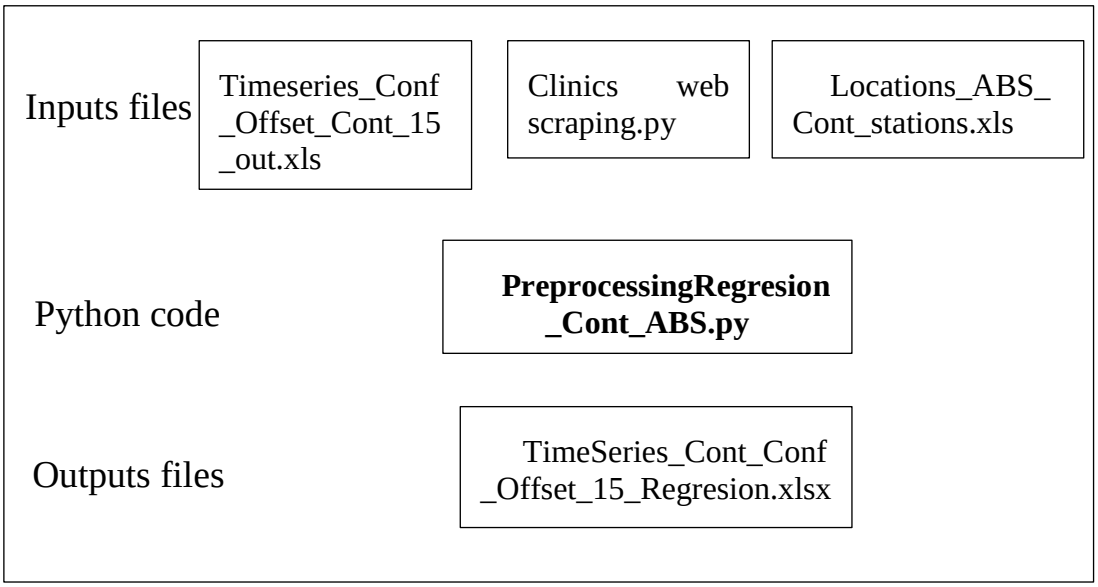

Figure 4.9 Flowchart of air quality preprocessing regression code.

4.6.2 Model to predict COVID-19 propagation based on meteorological and atmospheric pollutants

As it mentions before, the objective in this section is develop a regression fit concretely a multiple linear regression between the COVID-19 incidence and meteorological data and COVID-19 based on atmospheric pollution data. Hence, it can end up in a reconstructed COVID-19 data from regression fit which will be compared with original COVID-19 cases in order to predict COVID-19 cases in case of modifying atmospheric environmental variables in the model.

A multiple linear regression is a statistical technique that use variables in order to predict the outcome of a response variable. This is based on the assumption of there is a linear relationship between the dependent and independent variables and the fact of the independent variables are not highly correlated with each other.

This statistical method is used to model the linear relationship between the independent variables and response of the dependent variables, it means that the independent variables is the parameter necessary to calculate the dependent variables or the outcome.

The equation of the model is shown in Equation 4.1, where Y is the dependent variable, X_k is the independent variable each of which is related with the regression coefficients, β_k which indicated relative weight of this variable in the general equation. Moreover, the equation includes a constant β_0 and a random component ε which collects everything that the independent variables are unable to introduce in the model.

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k + \varepsilon \quad (4.1)$$

Hence, it can be taken advantage of a possible model derivation with the Beta coefficients also named standard regression coefficients of the regression model which can be based in standard scores and therefore, these are directly comparable to each other without the necessity of a constant β_0 and a random component ε. These specific coefficients provide a useful information about the relative importance of the independent variable in the regression equation, it means that, the standard regression coefficient has more importance in the equation which the higher value is in absolute value.

In this specific case, the multiple linear regression with standard regression coefficients is based on the values of the meteorological or air pollutants variables which are used to predict the COVID-19 incidence.

In order to obtain this relationship between the variables, a multiple linear regression is done by a Python code. In this instance, it will explain a multiple linear regression to obtain the standard regression coefficients, which is applied in the same way for COVID-19, and meteorological model regression fit and COVID-19 incidence and air pollutants data model regression fit, the difference will be the input data and the independent variables, it means that, the meteorological or air pollutants parameters.

First of all, it is needed to import a StandardScaler and Linear_model modules in Python in order to do the statistical modeling so, standardize the data values into a standard format and implement the linear regression in order to get the standard regression coefficients or beta coefficients and finally the COVID-19 incidence equation with meteorological and air pollutants variables.

Hence, the input file is used meteorological data from the processing regression code, it means that, the document file named *TimeSeries_Meteo_Conf_Offset_15_Regresion.xlsx* and for the atmospheric pollution data is *TimeSeries_Cont_Conf_Offset_15_Regresion.xls*. These files have collected all the data necessary to do the multiple linear code. Moreover, variables are defined in MultipleLinearRegression_Meteo_ABS code the meteorological variables and MultipleLinearRegression_Cont_ABS, the air pollutant parameters.

Once the input data is imported, an organization of the data is required in order to complete the dataframe in the empty spaces, sort the columns and eliminate those that contain the ABS code and date, creating a dataframe with the interested data to proceed with the code.

A standardize data part is started for the purpose of transform the data independent for each column and the given distribution which is done based on the dataset of each value which will have the mean value subtracted and then, divided by the standard deviation of the entire dataset.

In the code, it is simplified the standardize part with a StandardScaker function and a fit.transform function along the last data to transform and standardize it. To end up this part, this resultant data is related with the ABS code and date.

Thus, a multiple linear regression part of the code is initialized, define the linear model regression and in a range of the standardized data is obtained the standard regression coefficients for each of the variables evaluated. Then, a dataframe is created to make the summatory of the product between the standardize variables data and the standard regression coefficients in order to get a dependent variable (COVID-19 incidence).

Following the results, the predicted standardized COVID-19 incidence obtained with the regression fit is compared with the standardized data of COVID-19 cases from the epidemiological data of *Departament de Salud i del Servei Català* in order to evaluate the multiple regression fit.

To do so, a coefficient of determination also called R^2 scored is used to evaluate the performance of the multilinear regression model and reflect the goodness of fit. This coefficient can have a result between 0 and 1, it means that, values closer to 1 indicate that the fit model to the variable that is intended to be applied for the specific case will be greater. On the contrary, the results are close to the value of 0, the fit model will be lower and because of this, the model will be less reliable.

Regarding the application of the determination coefficient (R^2) in the code, a r2_score function from the skearn.metrics in Python is imported.

This function evaluated the original standardized COVID-19 cases data with the data obtained by the application of the regression fit model.

In order to visualize this result, a plot is displayed in order to see the tendency of the original standardized data in reference to the standardized data COVID-19 of each ABS by date.

A second plot is done to see the standardization between the variables data and the standardized variable data for each specific variable and ABS.

Mentioned codes are named *MultipleLinearRegression_Meteo_ABS.py* in case of meteorological regression fit and *MultipleLinearRegression_Cont_ABS.py* for atmospheric pollution regression fit. Thus, *MultipleLinearRegression_Meteo_ABS.py* can be found in Section 8.5.2.

In Figure 4.10 and 4.11 can be seen the meteorological multiple linear regression code flowchart and the air pollution, respectively.

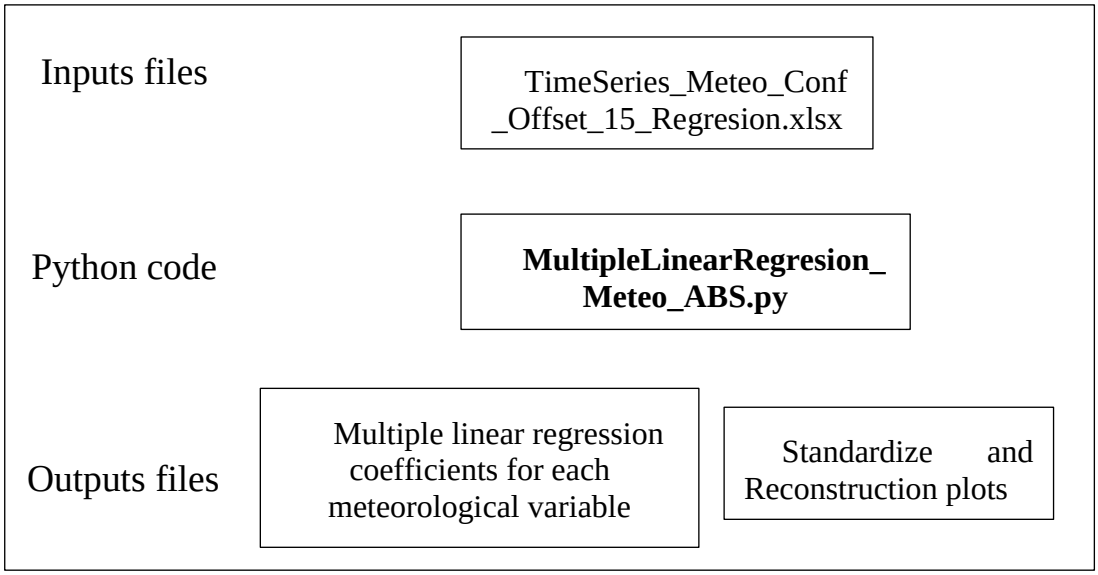

Figure 4.10 Flowchart of model to predict the propagation of the COVID-19 through the meteorological variables.

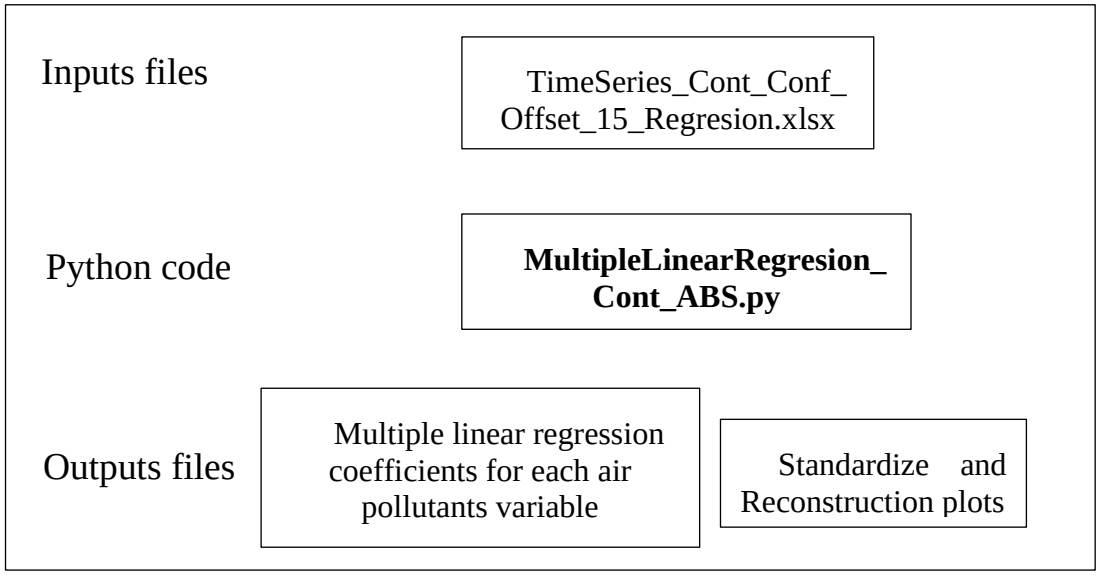

Figure 4.11 Flowchart of model to predict the propagation of the COVID-19 through the atmospheric pollutant variables.

4.6.3 Model to predict COVID-19 propagation based on atmospheric environmental parameters

The objective in this regression model is to simulate a common model to predict the propagation of COVID-19 through the vector associated with the environmental and pollution in air.

Hence, the multiple linear regression code is the same that it explained in the last section but taking into account the meteorological and air pollutions variables together.

Therefore, the input data are extract from the resultants excels files of meteorological and air pollutant processing regression code, *TimeSeries_Meteo_Conf_Offset_15_Regresion.xlsx* and *TimeSeries_Cont_Conf_Offset_15_Regresion.xls*, respectively.

These are required to merge the data by date and ABS code in order to get a general dataframe whose columns are composed by the air pollutants parameters and meteorological variables and the number of COVID-19 cases.

The common multiple linear regression code between air pollutants parameters and meterological variables are called *MultipleLinearRegression_MeteoCont_ABS.py*. The flowchart of this common multiple linear regression model can be displayed in Figure 4.12.

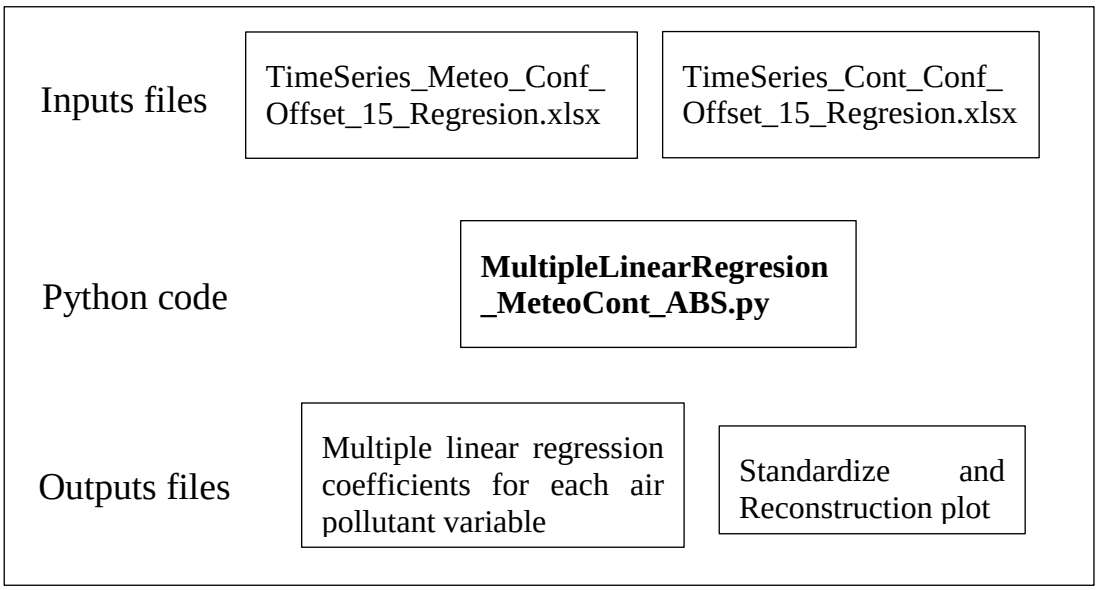

Figure 4.12 Flowchart of common framework model to predict the COVID-19 propagation through meteorological and air pollutant parameters.

4.6.4 Model to predict COVID-19 propagation based on mobility improved with atmospheric environmental parameters

Once a model to predict COVID-19 propagation based on atmospheric environmental variables is evaluated, it will do a study about the effect of mobility data in the common framework model during the lockdown period and pandemic period.

Mobility data is a relevant factor which must take into account in the spread of COVID-19 according to the recent studies.

Many researchers have reported a significant positive relationship between mobility and the COVID-19 positive cases. For instance, Zhu et al. (2020) research about the mediating effect of air quality on the association between human mobility and COVID-19 infection in China have investigated the relationship between mobility and COVID-19 infection and evaluated the role of air quality on this association. This mentioned study concludes that limiting human mobility highlighting lockdown period have contributed to the reduction in COVID-19 cases that it has a close relation with the affection of air quality data.

In order to assess this factor, a common framework to predict propagation of COVID-19 through the air vector associated with atmospheric environmental variables explained in section 2.6.3 is carried out by introducing the effect of mobility data during lockdown and pandemic periods.

The mobility data is composed by five parameters ratios which are organized by a geographical area. This area or also named ‘source’ represents the INE or municipalities code specifying the district.

The parameter which describes the mobility data are population stay-at-home ratio (q), population that make 1 or more trip ratio (m), ration of population that make trips inside the same area (r), population that make trips to other areas ratio (p) and difference on population that stay-at-home to the guideline of month before (delta).

In order to work in *Àrees bàsiques de* salut, an ABS will be selected and related with a source near to the district. Concretely, it is evaluated in a unique ABS, so by locations, in this case for ABS 403 – Barcelona-8L and ABS 192 – Sabadell.2. It means that, Barcelona-8L ABS corresponds to a health sector *Barcelona Nous Barris* which is located in a Barcelona district *8 Nous Barris* and the source will be INE code and district, so Barcelona-8L source is 801908. Follow a similar approach, Sabadell-2 ABS are from health sector *Vallès Occidental Est* which is complex sector with other ABS, so Sabadell-2 is located in *Can Puiggener* and corresponding district for this sector is 2, so Sabadell-2 ABS correspond to source 818702.

To do so, a Python code named *MultipleLinearRegression_MeteoCont_ABS.py* is used by adding the Catalonian mobility data from Big Data & Data Science unit of Eurecat in PROCEED project.

Thus, the input data are the excel files which are the same that in section 4.6.3 and the mobility data on the specific ABS. So, all the data is merged to form a general dataframe with the desirable data.

In this part, the python code is called *MultipleLinearRegression_AmbMobility_ABS.py* which can execute and assess the results verifying the contribution of mobility data.

The Python code *MultipleLinearRegression_AmbMobility_ABS.py* are shown in the Section 8.5.3 and the flowchart of the multiple linear regression model to predict COVID-19 spread and atmospheric environmental variables with mobility data can be seen in Figure 4.13.

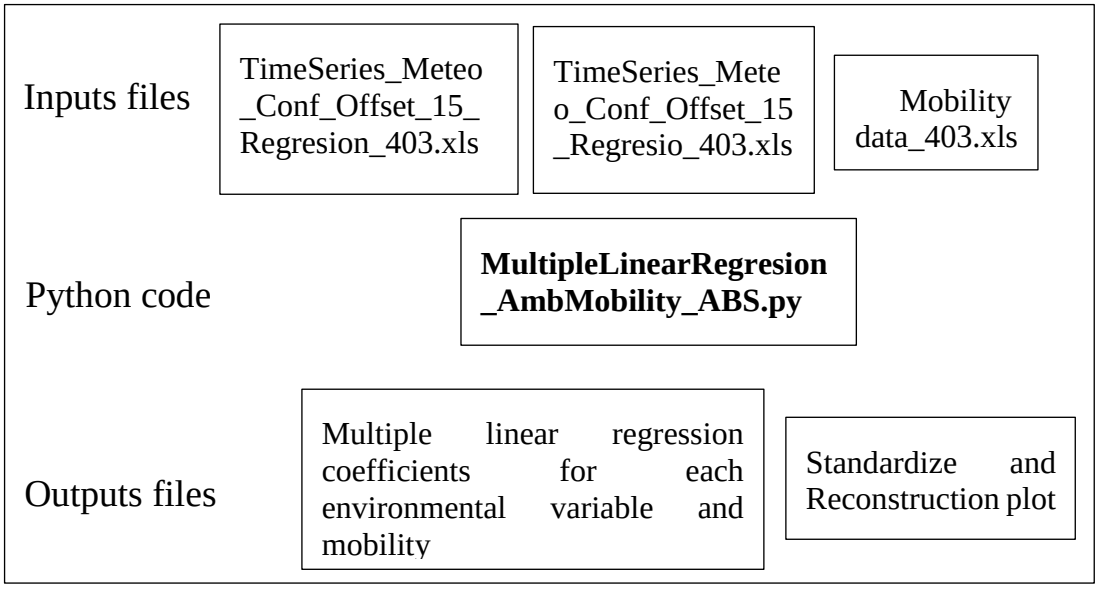

Figure 4.13 Flowchart of model to predict the COVID-19 propagation through meteorological and air pollutant parameters with mobility data.

5. RESULTS AND DISCUSSION

5.1 Correlation analysis

This section is aimed to provide the contribution of a set of meteorological and air pollutant variables in COVID-19 spreads during the lockdown and pandemic periods.

5.1.1 Correlation of COVID-19 cases and meteorological data in lockdown period

The results of the model assessing COVID-19 incidence and meteorological data correlation during lockdown period by ABS will be commented. The meteorological variables evaluated are wind velocity, temperature, relative humidity, precipitation and solar radiation.

Concerning the wind velocity variable, it is showed a negative and positive correlation, emphasizing the positive correlation predominance in case of higher correlation between an offset 21 and 25. This can be possible due to the fact that the existence of clouds with contaminated droplets which are maintained with the wind speed of approximately 15m/s according to recent investigations. Another possible relationship between wind velocity and COVID-19 incidence may be that the higher values of wind speed is produced a greater transmission of COVID-19 in different directions and a close association with the presentation of pneumonic symptoms, among others.

The wind speed evidence as an influencing factor in COVID-19 cases is limited and not conclusive because, apart from the wind velocity, direction influences among other factors. For this reason, it cannot be demonstrated that wind speed has a positive or negative significant relationship with the COVID-19 spread and therefore the number of confirmed COVID-19 cases.

Concerning the temperature parameter, a clear negative correlation can be seen, that is, the COVID-19 propagation is less with an increase in temperature. The highest correlation can be appreciated in population with more than 150,000 inhabitants such as Barcelona, Sabadell and an offset 0.

An evident correlation between the temperature parameter and the number of COVID-19 cases in offset 0 is demonstrated. This may be due to the fact that temperature increase by default due to the spring season in the lockdown period and at the same time, an increase in COVID-19 cases is experimented because of the fact that the period which the COVID-19 pandemic is started. In other words, it is observed the same trend in a natural way for that reason, in offset 0 is appreciated a high influence between temperature variable and COVID-19 propagation.

In case of relative humidity variable, the correlation obtained is appreciated positives and negatives which are closer to the same value taking into account an absolute value. However, the higher correlations are negative with an offset around 23 and 25 days. It can explain the fact that the temperature by defect of the spring season goes up in lockdown period, so there is a greater capacity to saturate the steam and the results is less relative humidity regarding an increase of COVID-19 cases.

Another important fact is the infection transmission by the respiratory droplets and aerosols which are closely related with the humidity, it means that, the stability of the respiratory virus membrane is higher, increasing the viability and transmission capacity.

Thus, the relative humidity may be related to COVID-19 transmission by respiratory droplets, but it is not possible to establish a definitive relation with the propagation of COVID-19 incidence due to the fact that other transmissions routes are not affected.

In reference to the rainfall parameter, the obtained correlation is reached positives and negatives values, highlighted that the higher correlations are positives. The results are not conclusive because the values are different, and an offset range cannot be established. The evidence is limited to confirm that rainfall can have a positive or negative impact on the COVID-19 incidence because of Catalonia is a place where no predominate the rainfall.

To sum up this part, there is not enough evidence to affirm that rainfall could be a limiting factor for the COVID-19 propagation.

The last parameter to be evaluated is the solar radiation where a negative correlation is observed as a tendency in the case of the highest correlation in offset 7 days. Hence, the negative correlation between the COVID-19 propagation can be appreciated such as the solar radiation is low, so the COVID-19 cases will be high.

Temperature in lockdown period goes up due to the fact that is spring so the solar radiation should be high. This fact is because the rays fall vertically and travel less distance to the atmosphere and more energy is concentrated in a smaller surface, so the temperature is probably high. In this case, probably the temperature is not enough high to increment the solar radiation and consequently, the decrease of COVID-19 cases. Moreover, it should be pointed out that the analyses period is lockdown which is the initial period of the pandemic.

In order to visualize the explained tendency, the correlation of each variable in ABS 197 which corresponds to Sabadell-6 and the correlation in ABS 295 – Hospitalet de Llobregat-8 can be seen in Figure 5.1 and Figure 5.2, respectively.

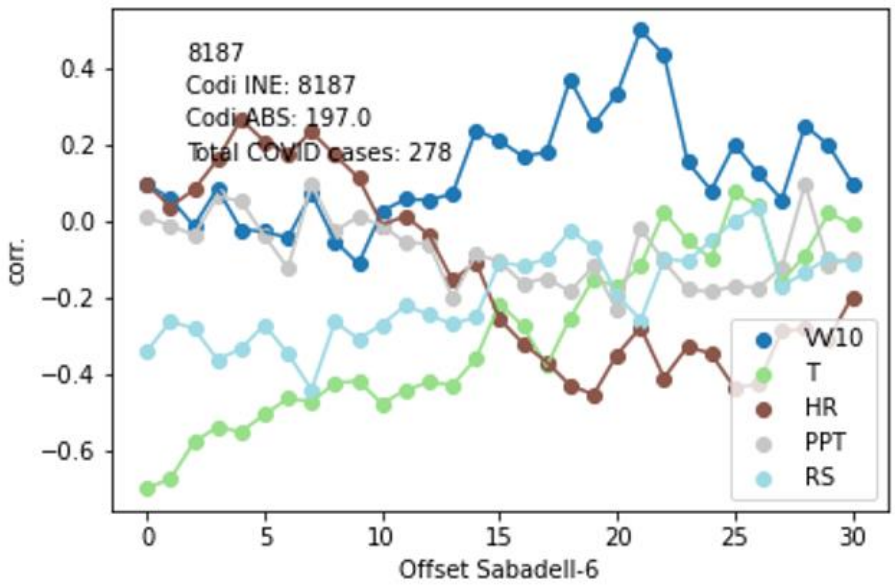

Figure 5.1 Correlation between COVID-19 incidence and meteorological data for the ABS Sabadell -6 and different offsets during lockdown period.

In Figure 5.1, it can be appreciated meteorological tendencies explained above. With respect to the temperature and solar radiation variables are a clear negative correlation along the 30 days of the offset evaluated. The relative humidity reaches the maximum correlation with negative values and an offset around 20.

In this case, the rainfall is around a correlation between 0 and -0.2, which should be pointed out that not prevail a clear correlation. The wind velocity has a positive correlation in reference to COVID-19 propagation and reaches the highest value in offset between 20 and 25 days.

The Figure 5.2 corresponds to the Hospitalet de Llobregat-8 which is another representation of the meteorological variables studied in order to contrast the results.

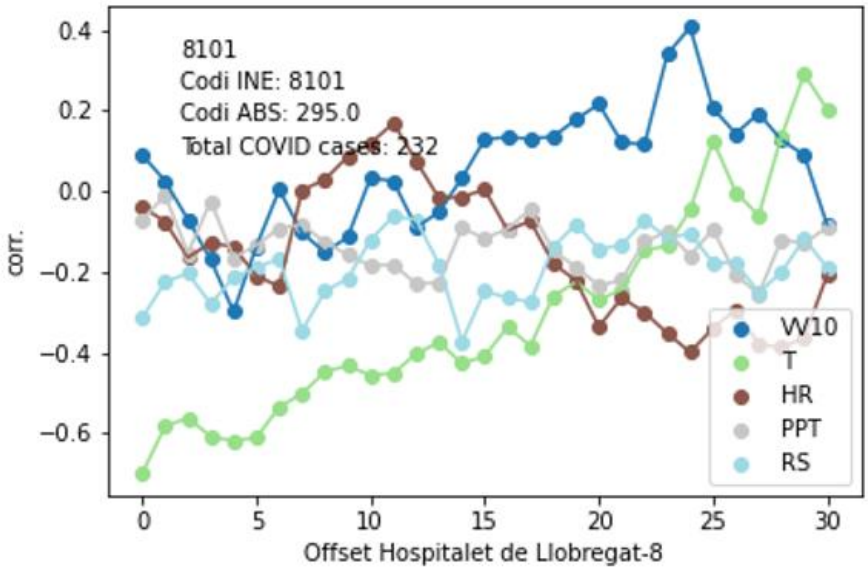

Figure 5.2 Correlation between COVID-19 incidence and meteorological data for the ABS Hospitalet de Llobregat-8 and different offsets during lockdown period.

Solar radiation and temperature are always in negatively correlated with the transmission of COVID-19, where the temperature has a high regression in offset 0 and in offset around 15 for the solar radiation. Concerning the relative humidity, it gets to the highest correlation between the offsets 24 and 28. Rainfall has values around zero, so there is no evidence to affirm the impact of this variable in the incidence of COVID-19 due to the rare rainfall in Catalonia. The wind velocity has the same tendency that in Figure 5.1, where it arrived at a high correlation in offset 24.

Finally, it should be noted that the behavior of the majority meteorological variable has a cyclical conduct each 7 days, for instance, the solar radiation, represented in Figure 5.1 and Figure 5.2, has a tendency to go up the first seven days, once the seven days passed, the tendency goes down.

To conclude this study, temperature, relative humidity and solar radiation are demonstrated which are an influencing factor in COVID-19 propagation. Concerning wind velocity and rainfall, the evidence is limited to establish a relation between these meteorological variables and COVID-19 incidence, thus wind velocity and rainfall can be ruled out of the study.

5.1.2 Correlation of COVID-19 cases and meteorological data in pandemic period

In this section, it will be mentioned the influence of selected meteorological variables versus COVID-19 spread in pandemic period which are from 1st of March 2020 to 15th of March 2021. The selected meteorological variables are temperature, relative humidity and solar radiation as aforementioned in the result from lockdown period.

In reference to temperature, a negative correlation in reference to COVID-19 confirmed cases are obtained during the pandemic period, so in a general, temperature has a predominance to increase the propagation of COVID-19 when temperature decrease.

The highest correlation takes place in offset 0 which there is more influence between temperature and COVID-19 incident, the same fact occurred during the lockdown period. It should be pointed out that the correlation reaches low values in comparison with lockdown period, for instance the highest correlation in lockdown period was -0.70 and during pandemic -0.54. This fact is an expected result because of the changes in season during the pandemic period, detected COVID-19 variants and the intervention of other factors such as mobility, which is key factor that changes during lockdown and pandemic periods.

In regard to relative humidity factor, there are positive and negative correlations which are closer to similar value during the pandemic period, due to the fact that the weather changes around the period. It follows the similar tendency that in lockdown period but obtaining a low correlation value between relative humidity and COVID-19 infection. Moreover, it can appreciate a clear cyclical tendency around each 7-day offset.

The last parameter analyzed is the solar radiation which has a similar performance in comparison with temperature. Hence, the highest correlation between COVID-19 positive cases and solar radiation are negative, reaching the highest value of -0.52. Therefore, a cyclical domain is presented in function of the offsets in the pandemic period.

With the aim of visualize the performance of meteorological variables, Figure 5.3 and Figure 5.4 can be plotted which are related to ABS 252 - Terrasa-F and ABS 46 - Barcelona-5E respectively.

In Figure 5.3 can be seen COVID-19 and meteorological model correlation in Terrasa-F during the pandemic. In this case, a clear negative correlation can observe in reference to the three meteorological variables in function of COVID-19 propagation.

The relative humidity oscillates around the same values with a cyclical tendency along the 30 days of offset and present a low correlation between COVID-19 confirmed cases. In reference to temperature and solar radiation can be seen an analogous disposition and also a cyclical performance over the offset. Both variables reached a high correlation in comparison with relative humidity.

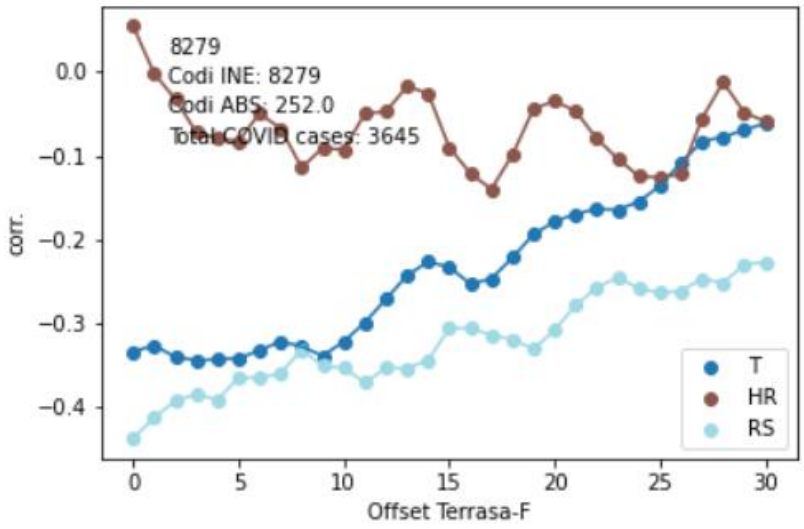

Figure 5.3 Correlation between COVID-19 incidence and meteorological data for the ABS Terrassa-F and different offsets during pandemic period.

In Figure 5.4 model evaluating COVID-19 propagation and meteorological data correlation in Barcelona-5E during pandemic can be seen.

All meteorological variables in Barcelona-5E ABS have the same tendency than Terrasa-F ABS but it reaches higher correlations. This can be due to the fact that Barcelona has a high population and the meteorological condition change depending on the location.

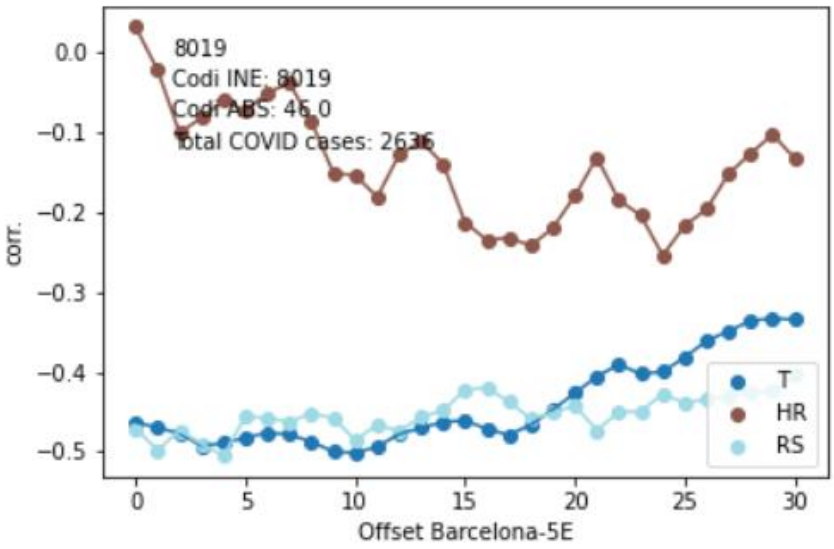

Figure 5.4 Correlation between COVID-19 incidence and meteorological data for the ABS Barcelona-5E and different offsets during pandemic period.

It is concluded that meteorological variables evaluated have a close relationship with COVID-19 propagation during the studied period and present a cyclical tendency each 7 days. The pandemic correlation is lower than in lockdown period due to influence of other factors such as seasonality, mobility and detected COVID-19 variants.

5.1.3 Correlation of COVID-19 cases and air quality data in lockdown period

The aim of this section is to evaluate the model assessing COVID-19 propagation and air quality data correlation in reference to the ABS. The air pollutants parameters evaluated in the model are the suspended particles such as PM1, PM2.5 and PM10, nitrogen dioxide (NO_2), nitrogen oxide (NO), nitrogen oxides (NO_x) ozone (O_3), sulfur dioxide (SO_2), carbon monoxide (CO) nickel (Ni), chlorine (Cl), dichloride (Cl_2), cadmium (Cd), mercury (Hg), lead (Pb), arsenic (As), benzene (C_6H_6), benzo-alpha-pyrene (BaP), hydrogen chlorine (HCl) and hydrogen sulfide(H_2S).

It can be appreciated that some pollutants in air have no correlation between the COVID-19 incidence and the air quality data correlation, these are chlorine, cadmium, mercury, hydrogen chlorine and benzo-alpha-pyrene. This can be because it is unfeasible to filter the air pollutants data in the way that avoid those with no correlations. Moreover, it can occur that these pollutants are in a minor quantity in the air and are not relevant in the influence of the COVID-19 propagation.

Assessing the influence of various air pollutants on the transmission of COVID-19, such as nickel and benzene, there is insufficient evidence to establish a relation between them and the capacity to spread COVID-19. It is obtained that the highest correlation is positive compared to nickel and benzene and that correlations correspond just to the Barcelona area. Therefore, it can be mentioned that the behavior of these air pollutants is not conclusive and accurate to establish that nickel and benzene have been a positive impact in the COVID-19 incidence.

With respect to air pollutants parameters such as suspended particles (PM1), lead, arsenic and hydrogen sulfide were not significant correlated with COVID-19 infection rate due to the fact that the obtained correlation oscillates between negative and positive tendency, besides the fact that both values are extremely close to each other taking into account the absolute value.

In this way, it is not possible to determine a definitive relationship between these air pollutants in the COVID-19 spread. Concerning the rest of air pollutants parameters, such as suspended particles (PM2.5 and PM10), nitrogen oxides (NO, NO_2), sulfur dioxide (SO2), carbon monoxide (CO), they have significant positive correlation with the confirmed cases of COVID-19, while ozone has a negative association with COVID-19 incidence.

Suspended particles, concretely PM2.5 and PM10, have a positive correlation with COVID-19 incidence, and both follows the same tendency. The highest correlation in both air pollutants is around 0,65 which correspond a range of offset from 27 to 30 days in ABS of the area such as Barcelona, Hospitalet de Llobregat and Prat de Llobregat.

Another shared trend of diverse air pollutants is composed by the nitrogen oxides, that is, NO and NO_2, which have a positive contribution to COVID-19 cases. The offsets which correspond to the highest correlation oscillate between 13 and 15 days and the achieved regression is approximately 0,65 in ABS, for instance, el Prat de Llobregat, Vic and Badalona. Also, it is mentioned that sulfur dioxide (SO2), carbon monoxide (CO) is positively correlated with COVID-19 cases, it means that it has a significant negative impact of COVID-19 propagation.

Regarding the sulfur dioxide air pollutant, the highest correlation reaches the 0.65 with a respective offset of approximately 22 days in Badalona, Granollers and Hospitalet de Llobregat ABS.

In reference to the air pollutant carbon monoxide, the correlations obtained are positive and negative, which are closer to the same value, however, the positive correlation prevails from the negative influencing in the spread of COVID-19.

Finally, it is important to highlight that the ozone air pollutant is difficult to come to the conclusion because of it has a positive or negative influence on the COVID-19. That fact can be because of the ozone is a gas which is not emitted directly but it is produced from other air pollutants, that can be reacted with other substances in presence of natural light, so the behavior can be change radically. Thus, ozone has a different behavior compared to other atmospheric pollutants in the studied model correlation data.

Some examples of the model assessing COVID-19 incidence and atmospheric pollutant data correlation are displayed, concretely the ABS 341 which belongs to Badalona-7A and ABS 383 Barcelona-3H, which can be seen in Figure 5.5 and Figure 5.6.

It can be seen that air pollutants parameters in general have the same tendency except the ozone and the hydrogen sulfide. The remainder of air pollutants, which are not displayed in the graphic, are some that does not have data or those that does not have a measurement according to the pollutant station in these specific ABS.

In Figure 5.5, the nitrogen oxides (NO, NO_2, NO_x) follows the same tendency and these values are closer to each other reaching a high correlation in offset 13. It should be noted that nitrogen oxides (NO_x) are a redundant air pollution because it is the summation between nitrogen monoxide and nitrogen dioxides, so it can be ruled out. In the same way, suspended particles (PM1, PM2,5 and PM10) have a same disposition getting a maximum value of correlation in offset 17.

Sulfur dioxide has a positive correlation regarding the COVID-19 propagation and the tendency is similar to the other pollutants in air, but in this specific case, it arrives at the highest correlation around 0.6 with an offset of 22. On the contrary, hydrogen sulfide is the air pollutant whose highest correlation is lower compared to the others air pollutants parameters. Finally, the ozone is the unique air pollutant which has a negative correlation in reference to the COVID-19 spread because it is a pollutant which is produced by a reaction with other air pollutants.

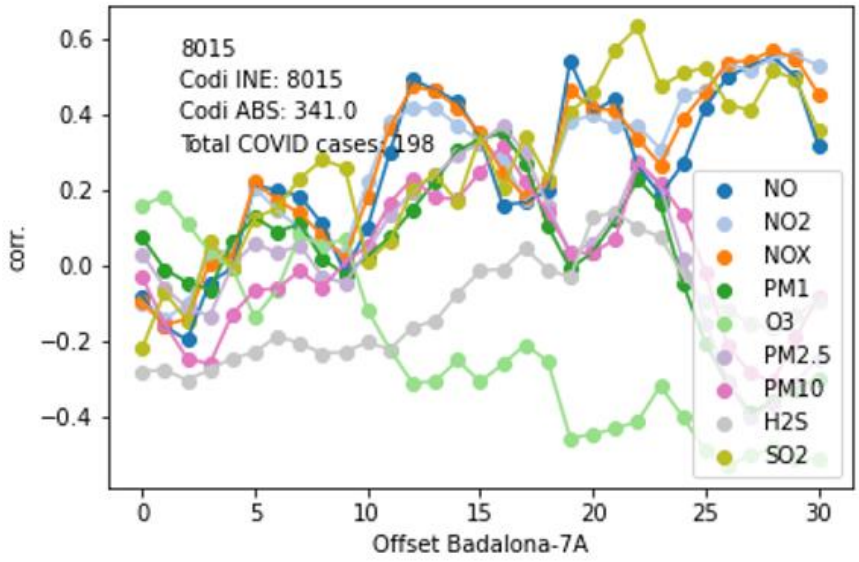

Figure 5.5 Correlation between COVID-19 incidence and atmospheric pollutant data for the ABS Badalona-7A and different offsets during lockdown period.

Taking into account Figure 5.6 in reference to the Barcelona-3H it can be appreciated the same disposition that in Figure 5.5. Nitrogen oxides and suspended particles are closer to each other, confirming a positive tendency associated with an incrementation of numbers COVID-19 cases. Moreover, the highest tendency in reference to suspended particles, nitrogen oxides, benzene, carbon monoxide and sulfur dioxide are around 13 days of offset. On the other hand, ozone air pollutant has a contrary disposition with a negative correlation concerning the COVID-19 transmission.

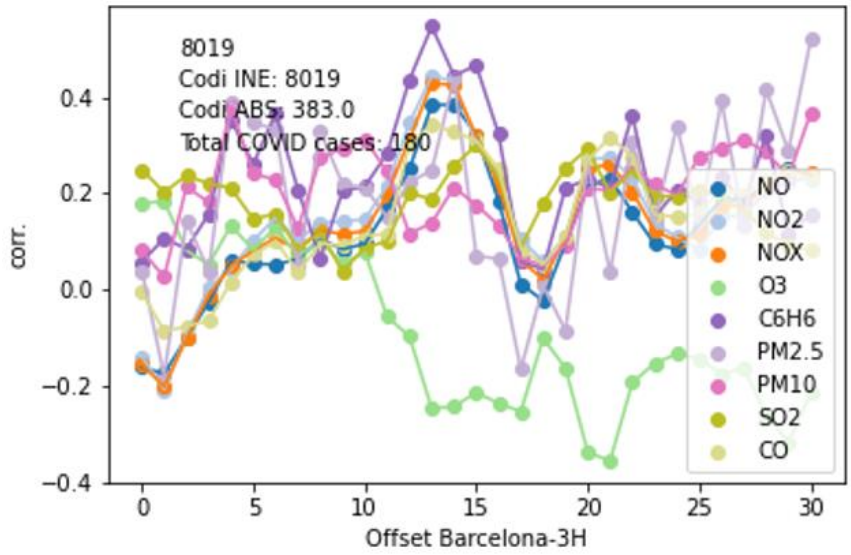

Figure 5.6 Correlation between COVID-19 incidence and atmospheric pollutant data for the ABS Barcelona-3H and different offsets during lockdown period.

To end up this point, there is a diversity in air pollutants performance highlighting that the mayor air pollutant has a positive tendency, but a low correlation in comparison with meteorological variables. Moreover, ozone is a particular pollutant in air, which is difficulty to see its performance in a clear way due to the fact that it is an emitted gas composed by reaction of others. The nitrogen oxides are ruled out of this study because of the redundancy with nitrogen monoxide and nitrogen dioxide.

5.1.4 Correlation of COVID-19 cases and air quality data in pandemic period

Model assessing COVID-19 incidence and atmospheric pollution data correlation in pandemic period will be discussed. Air pollutant variables to be commented are nitrogen oxides, suspended particles, carbon monoxide, sulfur dioxide and ozone.

In pandemic period, the general disposition of the air pollutants along the offset are similar but the correlation obtained is lower in comparison with lockdown. Therefore, the atmospheric pollutants follow a cyclical tendency of each seven-day offset where the tendency is repeated.

The majority of air pollutants parameters have a positive correlation in relation to COVID-19 propagation, it means that, an increment of COVID-19 positive cases is produced by an incrementation of specific pollutants in air, except of ozone which has a contrary tendency. However, in pandemic is much difficult to establish a common pattern because the data present more dispersity.

For instance, model assessing COVID-19 propagation and atmospheric pollutant data correlation during pandemic period in ABS 247 - Terrasa-A and ABS 293 – Hospitalet de Llobregat-6 can be plotted in Figure 5.7 and Figure 5.8, respectively.

Figure 5.7 shows the air pollutant tendency in Terrasa-A, which has reached a low correlation around 0.3. This is insufficient evidence to establish a pattern relation between them and capacity to spread COVID-19. Moreover, the air pollutants disposition is cyclical and around the same values make it difficult the conclusions in pandemic period.

Nitrogen oxides have reached the higher positive correlations and ozone the negative one, however, sulfur dioxide, carbon monoxide and PM10 reaches a 0.10 correlation, which is not enough evidence to affirm that these air pollutants variables are an influencing factor in COVID-19 propagation.

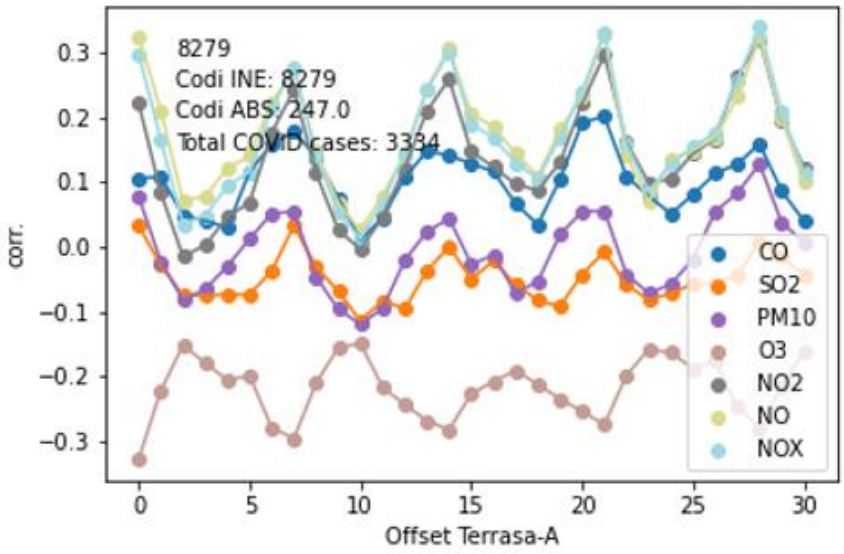

Figure 5.7 Correlation between COVID-19 incidence and atmospheric pollutant data for the ABS Terrasa-A and different offsets during pandemic period.

In Figure 5.8 can observed the correlation of each air pollutant over the offsets in reference to the Hospitalet de Llobregat-6 ABS. The correlation is lower in comparison to lockdown period and in this case, the air pollutants parameters are more disperse, hence it is complicated to define a pattern.

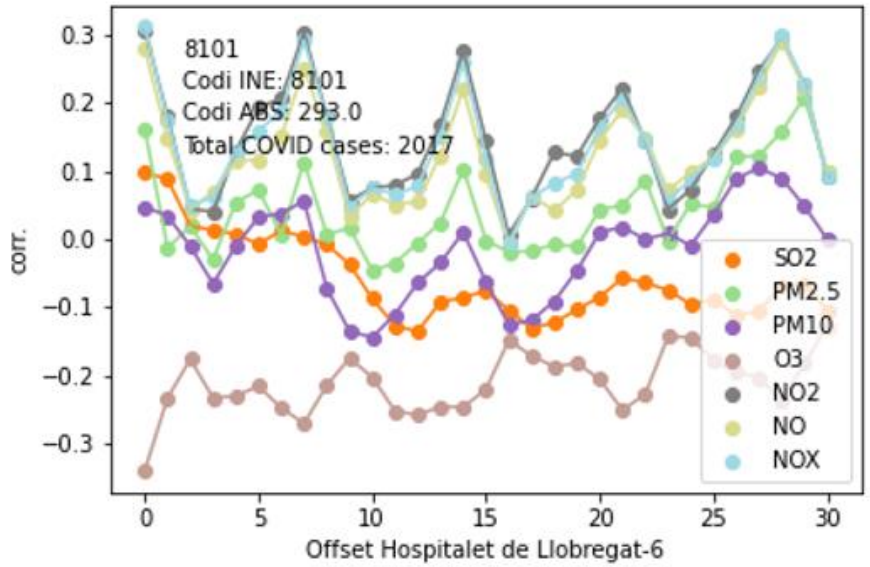

Figure 5.8 Correlation between COVID-19 incidence and atmospheric pollutant data for the ABS Hospitalet de Llobregat-6 and different offsets during pandemic period.

To conclude, model evaluated COVID-19 propagation and air pollutants data correlation are studied in lockdown and pandemic period which has a positive correlation over offset and on the contrary, a negative correlation with ozone. The disposition is cyclical each offset of seven days and lower correlation are obtained in pandemic period due to the data dispersion and probably the influence of COVID-19 variants, but it cannot be clearly appreciated the relation to each air pollutant parameter because some of the are emitted and influenced in different factors.

5.2 Offset analysis

A study to evaluate the influence of the offset in COVID-19 propagation and meteorological and atmospheric pollutants in lockdown and pandemic period has been performed. Hence, the objective is to choose an offset with a higher correlation coefficient in reference to COVID-19 and atmospheric environmental variables data correlation.

5.2.1 Offset analysis based on meteorological

As aforementioned in section 5.2, frequency histogram and box plot can be seen in order to analyses the COVID-19 incidence and meteorological offset in lockdown and pandemic.

The frequency histogram indicates the frequency distribution of the offset in case of the meteorological variable has a higher correlation coefficient in order to visualize in which offset the variable has more frequency. Hence, Figure 5.9 and Figure 5.10 show the frequency histogram and box plot of COVID-19 and meteorological variables in lockdown period.

Figure 5.9 displays the histogram of the offsets corresponding to the highest correlations for each of the meteorological variables considering all ABS, the y axis corresponds to the frequency and in x axis the offset (0-30 days). It can be appreciated that the offset is kept around a value for a specific variable except the case of temperature variable that it cannot achieve a specific conclusion with its regular distribution.

However, it can be visualized, that the offset of the rest variables tends to have a different value depending on the variables, for instance, relative humidity has the highest regression in offset 25 days and solar radiation at 10 days.

It can end up; meteorological variables have a different offset value, where it seems to have the highest correlation.

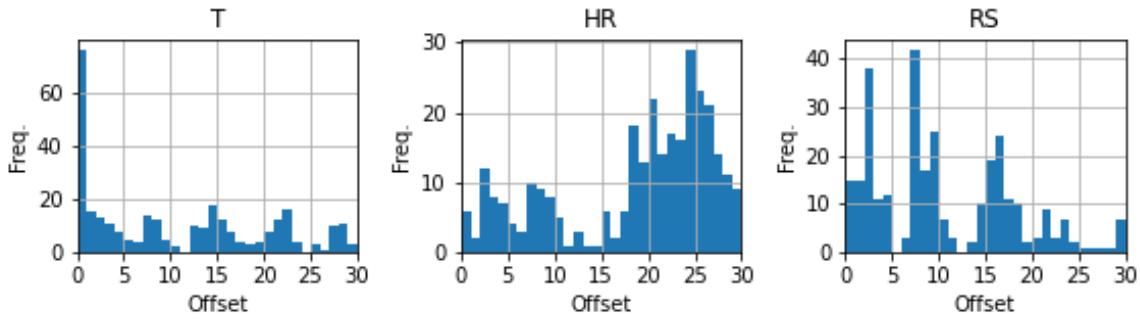

Figure 5.9 Histogram of the offsets associated to the higher correlation values for the meteorological parameters considering all ABS during lockdown period.

Specially, the boxplot obtained for the evaluation of meteorological correlation model is displayed in Figure 5.10. It can be seen a similar distribution than in Figure above of the boxplot with the meteorological variables which has a common range of offset with different variables are remain. This offset range with the highest correlation by meteorological variable is from offset 12 to 20.

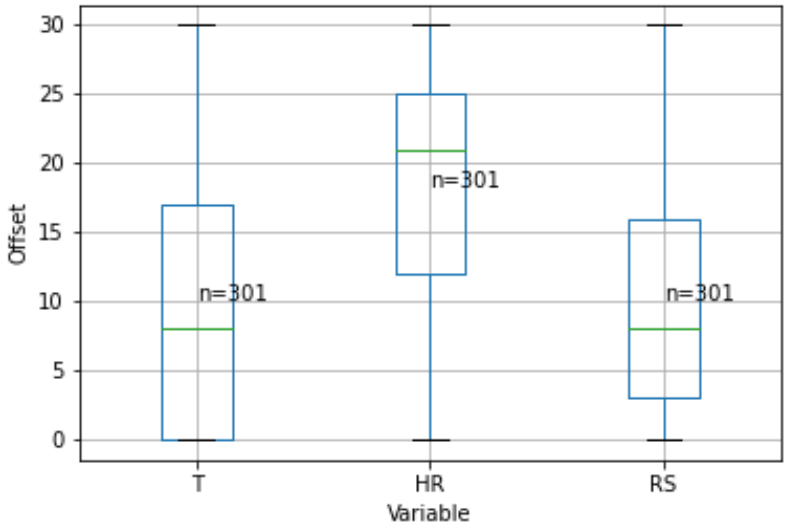

Figure 5.10. Boxplot diagram of the offsets with maximum correlation values considering all ABS during lockdown period.

Concerning the pandemic period, the frequency histogram and box plot in regard to COVID-19 incidence and meteorological data correlation can be seen in Figure 5.11and Figure 5.12. As it can observed in Figure 5.11, the frequency distribution of the meteorological variables with high correlation in pandemic period has more dispersion around the offsets and it is not clear to specify an offset for meteorological variables. Moreover, the offset which has more frequency are 3 for temperature parameter and 25 in case of relative humidity. In case of solar

radiation, the results are dispersed over the offsets, and it is not possible to decide a specific offset.

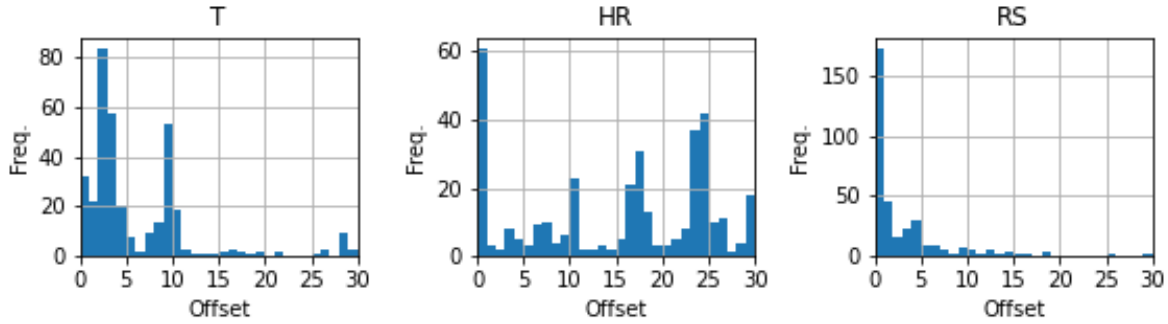

Figure 5.11 Histogram of the offsets associated to the higher correlation values for the meteorological parameters considering all ABS during pandemic period.

Regarding the box plot in pandemic, it can be appreciated that values have more dispersion than in lockdown due to the fact that in some meteorological variables there are values which differ significantly for the remainder of datasets.

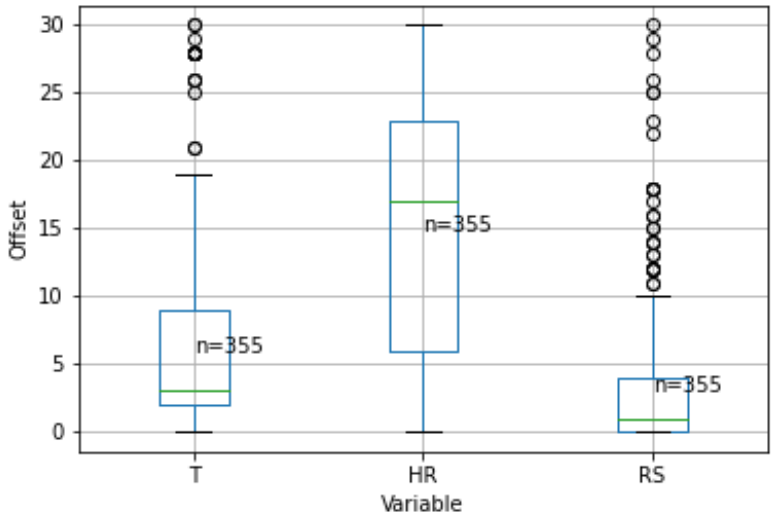

Figure 5.12 Boxplot diagram of the offsets with maximum correlation values considering all ABS during pandemic period

To conclude, offset results regarding the COVID-19 and meteorological variables in lockdown and pandemic period are not conclusive due to the dispersion of values and it is not possible to establish a specific offset. Moreover, offset range appreciated in the lockdown are from 15 to 20, but in case of pandemic, it is not plausible to determine the range.

5.2.2 Offset analysis based on air quality data

Following the same approach than in section 5.2.1, COVID-19 propagation and atmospheric pollutant offset study during lockdown and pandemic period is done by frequency histogram and box plot.

The lockdown frequency histogram is displayed in Figure 5.13 which are the distribution of the offset frequency related with the highest correlation of air pollutants variables in function of the offsets in the range from 0 to 30 days.

It can be seen that frequency histogram plot in some pollutants in air are empty, it can be due to the fact that these are minor air pollutants parameters, and no correlation model between COVID-19 incidence and these air pollutants data is produced. Hence, it can be ruled out these air pollutants variables and focus on the rest of the air pollutant, that are the major pollutants such as suspended particles (PM10 and PM2.5), SO_2, NO, NO_2, CO and O_3.

The selected pollutants in air are the most relevant pollutants which has evidence of influence in COVID-19 incidence according to the Scientist publications mentions in bibliographic research (Section 1).

In this case, the offset of the majority of the air pollutants parameters are in a range from 10 to 15 days, concretely, offset 13 where there is the offset with the highest correlation by pollutants in air variables. However, suspended particles in particular PM10 and PM2.5 and NO are around the offset 25-30 and in the particular case of the ozone, the evidence is limited to establish an offset.

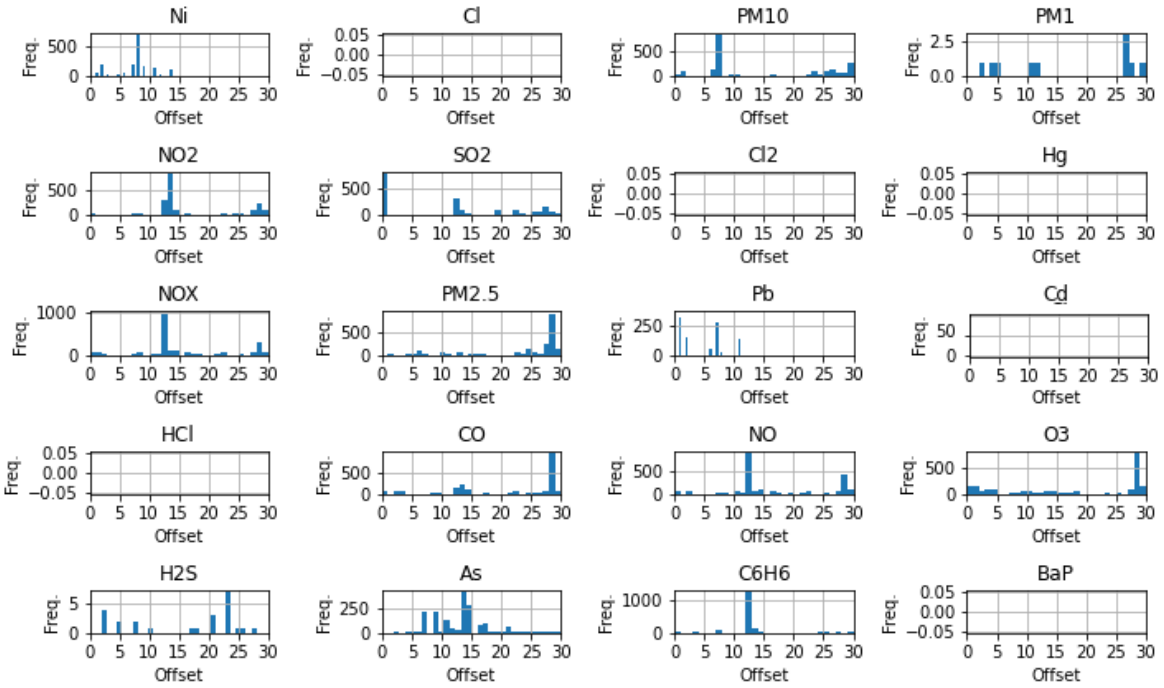

Figure 5.13 Histogram of the offsets associated to the higher correlation values for air pollutants parameters considering all ABS during lockdown period.

Concerning the boxplot during lockdown, Python offset analysis code is run in order to obtain only the selected air pollutants variables, so the Boxplot for the selected air pollutants variables is shown in Figure 5.14. It can be observed the boxplot about the air pollutant variables has a similar performance relating to Figure 5.14 and thus, a unique range of offset can be extracted which pollutant in air variables will have the highest correlation concerning COVID-19 influence. The offset range with high correlation in case of pollution air parameters can be from 13 to 20 days.

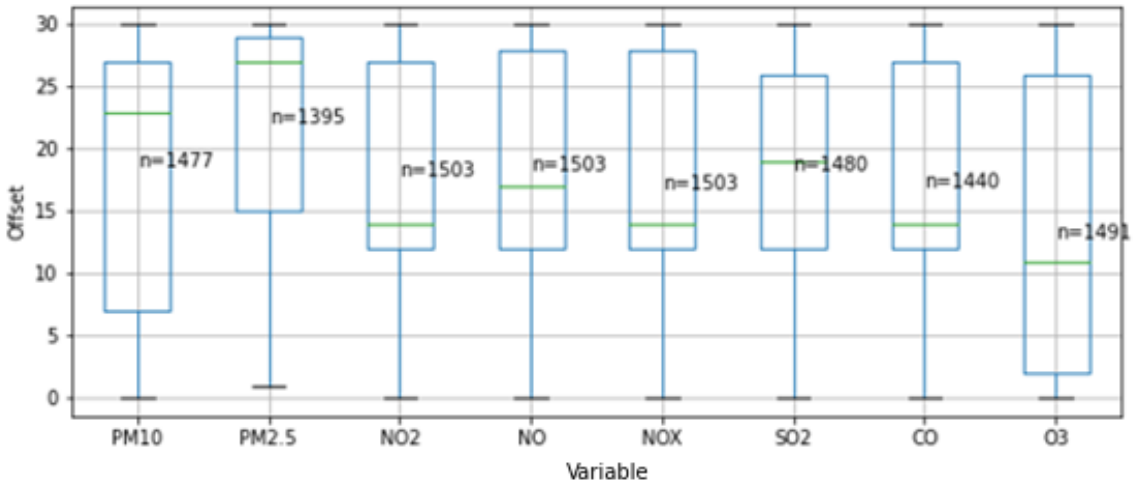

Figure 5.14 Boxplot diagram of the offsets with maximum correlation values considering all ABS during lockdown period

Regarding results from pandemic, frequency histogram and box plot about COVID-19 propagation and atmospheric pollutant data correlation can be displayed in Figure 5.15 and Figure 5.16 respectively.

As Figure above shows, air pollutants data distribution displays a frequency dispersion data along the offset, making challenging determine an offset which dataset are highly correlated with COVID-19 spread.

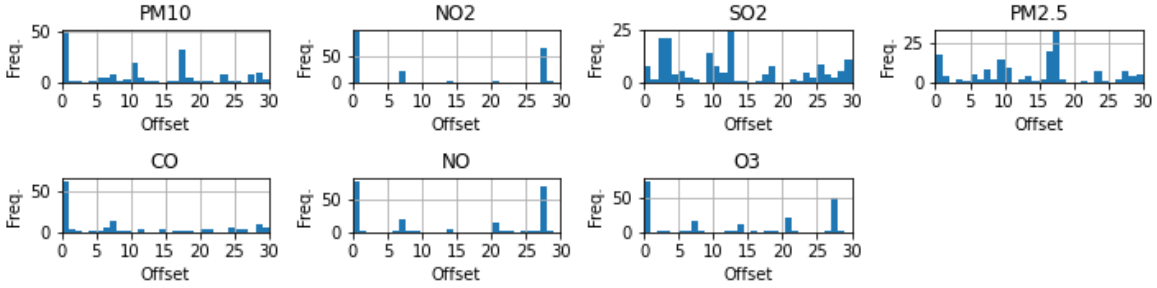

Figure 5.15 Histogram of the offsets associated to the higher correlation values for air pollutants parameters considering all ABS during pandemic period.

Additional alternative to examine atmospheric pollutant data during pandemic is by box plot which can be plotted in Figure 5.16. It can end up with a common offset in a range from 10 to 17, nevertheless, this range is not accurate with a specific pollutant in air due to the fact some of them present a large disparity data along the offsets.

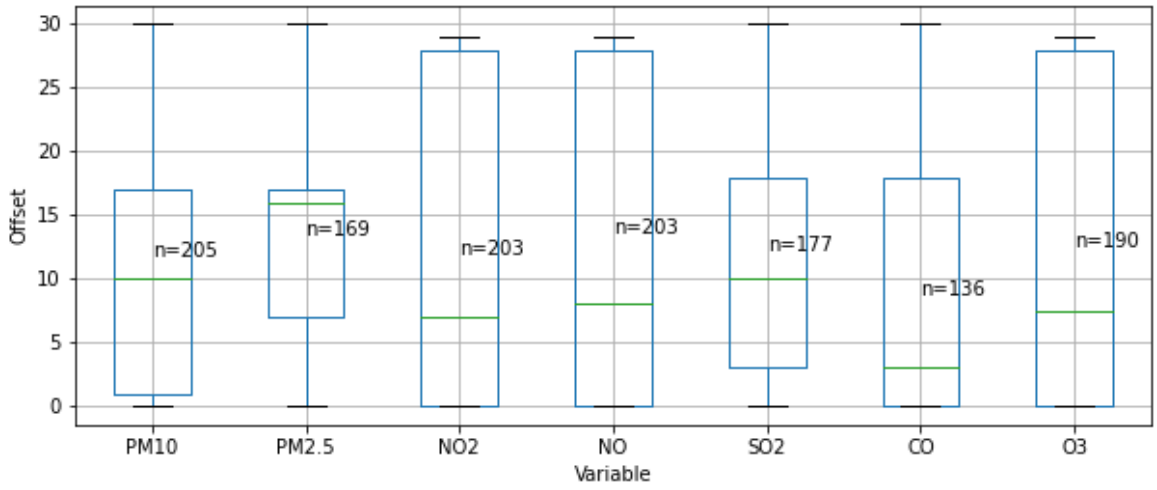

Figure 5.16 Boxplot diagram of the offsets with maximum correlation values considering all ABS during pandemic period

Finally, COVID-19 confirmed cases and atmospheric pollutant offset findings during lockdown and pandemic cannot appreciate an explicit trend to affirm the offset selected. There is a common offset range in both periods, but it is not conclusive to validate that it will offset chosen.

5.2.3 Selected offset

The goal of this section is to make out the selection of an offset to be used in the regression fit of COVID-19 incidence and the atmospheric environmental parameters that were selected during the correlation analysis. To this aim, a literature review was carried out to complement the analysis explained in previous sections since no clear pattern was identified.

According to Marks et al. (2021) publication, the incubation period was defined as time from the first exposure to symptom onset, hence the mean incubation period has estimated between 5 to 7 days for the SARS-CoV-2 transmission. The article highlights that the initial virial load can change significantly the incubation time, so the article concludes that the incubation time for participants with a high load was 5 days and for patients with a low viral load, 7 days.

A similar approximation was found by Zaki et al. (2021) in their review article about estimation of the COVID-19 incubation period. The manuscript mentions that the majority of publications estimate the incubation period of the virus in 7.8 days by average, with a median of 5.01 days.

In reference to COVID-19 with long-term use of glucocorticoids article by Yuanyuan Han et al. (2020), it is mentioned that the average incubation period is from 2 to 14 days and mostly from the range of 3-7 days. However, it pointed out that this period may be longer in some patients.

Dhouib et al. (2021) have published the article in reference to the incubation period during the pandemic COVID-19: a systematic review and meta-analysis which provides the evidence of incubation period for COVID-19 and shows that it can exist an incubation period up to 14 days. Moreover, this study proportionate studies done in China which the mean and median incubation period were 8 days and 12 days, respectively and in different parametric models, the 95th percentiles were in the range from 10.3 to 16 days.

To sum up, it has not been elucidated that a common offset can be established in COVID-19 incidence and atmospheric environmental variables during pandemic and lockdown because of the absence of clear trend. According to the presented offset study and the literature review, the offset selected was 15, although this parameter can be considered as a free variable to set properly in the model considering a certain scenario.

In the selection of the offset other factors are taken into consideration, for instance, the offset may not only depend on the COVID-19 incubation time and thus on the variants of COVID-19, but also in the lapse between the occurrence of symptoms (infection) and when the positive case in a specific ABS is notified.

5.3 Regression fit

With the aim of finding a proper model to predict COVID-19 propagation based on atmospheric environmental variables, several models need to be analyzed in lockdown and pandemic periods.

Hence, in this part is relevant to discuss the results from the multiple linear regression in different scenarios in order to predict the propagation of COVID-19 and meteorological variables, COVID-19 incidence and air pollution data and the common framework COVID-19 through the air vector composed by meteorological and air pollution parameters.

The scenarios listed below have been considered:

1st Scenario: Regression fit is carried out based on data from all ABS.

2nd Scenario: Regression fit is carried out based on data from the specific ABS where the reconstructed cases are to be calculated.

3rd Scenario: Regression fit is carried out based on data from ABS regions with more than 200.000 inhabitants.

By analyzing the results of reconstructed COVID-19 cases from regression fit and real COVID-19 cases, it will be selected the best approximation to predict COVID-19 propagation and atmospheric environmental variables during lockdown and pandemic periods.

5.3.1 Regression fit based on meteorological parameters during in lockdown period

A study to evaluate the adjustment and influence of meteorological variables (temperature, relative humidity and solar radiation) in the transmission of COVID-19 has been performed in the diverse scenarios during lockdown period.

As aforementioned, the multiple linear regression is done with the meteorological selected variables in the different scenarios, hence, standardize regression coefficients of each scenario can be found in Table 5.1.

Table 5.1 Standardized regression coefficients to predict COVID-19 incidence for meteorological variables in each scenario during lockdown period.

	1st scenario	2nd scenario (ABS 403)	2nd scenario (ABS 192)	3rd scenario
Temperature	-0.0732	-0.2771	-0.00519	-0.1425
Relative humidity	-0.1064	-0.1779	-0.3499	-0.1080
Solar radiation	-0.1004	-0.2103	-0.3889	-0.1075

Regarding the first scenario, the multiple linear regression is done with meteorological selected variables and all ABS, hence, it can end up with the standardize regression coefficients in reference to temperature, relative humidity and solar radiation which form the model equation to explain the influence of COVID-19 propagation.

The independent variable that predicts better the propagation of COVID-19 is the relative humidity due to the fact that it is the highest regression value in the model in absolute value. The negative signs are indicated that low relative humidity will produce an increment of COVID-19 spread.

Solar radiation is another variable which is closed to predict the behavior of COVID-transmission, highlighting the case that less levels of solar radiation, COVID-19 propagation will increase. Moreover, a low temperature is produced by more tendency COVID-19 spread, although this relation in this case is lower.

Concerning the second evaluated scenario, it is reproduced the multiple linear regression for a unique ABS in lockdown, so the ABS selected is Barcelona-8L which correspond to ABS code 403. In addition, it is followed the same procedure for ABS 192 – Sabadell-2 in order to contrast the result for another ABS.

Table 5.1 shows standardize regression coefficient of the temperature, relative humidity and solar radiation which explain the influence with COVID-19 propagation in case of second scenario for ABS 403 – Barcelona 8L.

Temperature is the best variable that predict the COVID-19 spread in the second scenario with a value of -0.2771, then the solar radiation with -0.2103 and finally, the relative humidity at -0.1779. The tendency of the independent variables is the same that in the first scenario, it means that, at low temperature, relative humidity and solar radiation, it will produce an increase of COVID-19 cases.

In reference to ABS 192 – Sabadell-2 in second scenario, solar radiation and relative humidity presents a high regression with the COVID-19 cases prediction. On the contrary, temperature parameter has a low magnitude in absolute value, so this variable has not a relative importance in reference to the COVID-19 incidence in the regression equation model.

After analyzing results in the second scenario, it can be observed that high values of independent variables are obtained in comparison with the first scenario. This can be due to the fact that the input data import is uniquely of the specific ABS and coefficients are more exactly and adjust the regression fit that in the first case.

In reference to the third scenario, a multiple linear regression is computed by applying a filter which provides municipalities' ABS with more than 200.000 population. Table 5.1 shows the standardized regression coefficients of the third scenario in order to predict the behavior of the propagation COVID-19.

As it is observed in the table above, the standard regression coefficients are similar than in the first scenario except the temperature regression coefficient. This can be because of the ABS which correspond to a municipality with high population have more influence in the regression coefficients than in ABS with minor inhabitants. Hence, the independent variable which has more precision in COVID-19 propagation is the temperature, follows to the relative humidity and solar radiation.

It is important to highlight coefficient signs which are indicated the disposition of the meteorological independent variables in the three scenarios are the same, that is, the COVID-19 incidence increase when the temperature, relative humidity and solar radiation are low.

In order to evaluate the model to predict COVID-19 transmission based on meteorological regression fit in different scenarios, Figure 5.17 is displayed taking into account ABS 403 - Barcelona-8L data to reproduce the reconstruct COVID-19 cases and compare the adjustment between the reconstructed COVID-19 data and original COVID-19 cases extract from *Dades Obertes de Salut.*

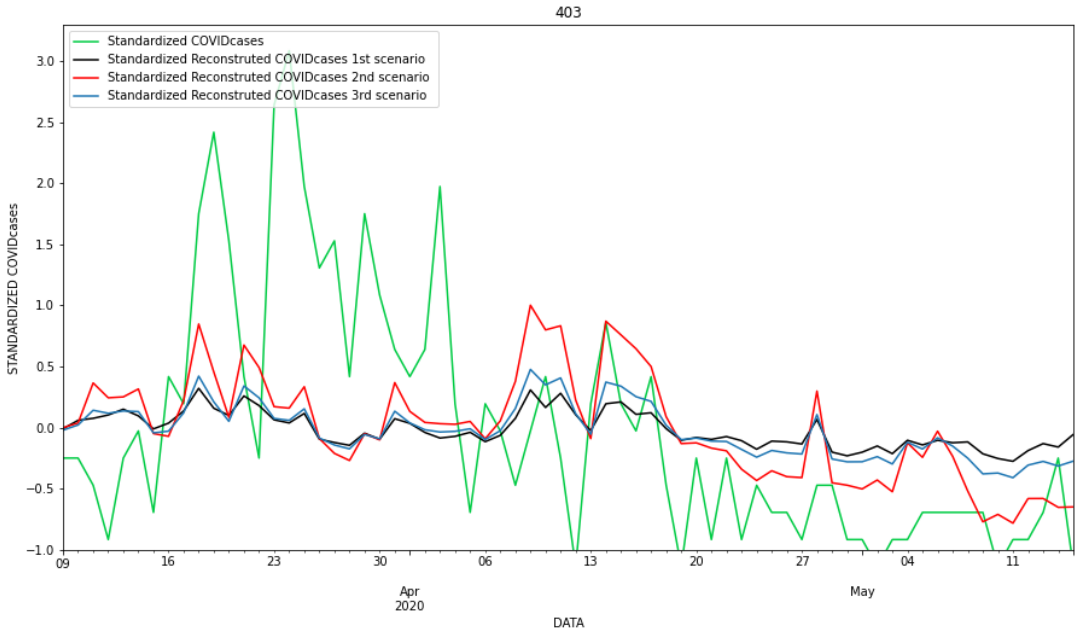

Figure 5.17 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on meteorological data (standardized values) for the three scenarios in the ABS "403-Barcelona-8L" during lockdown period.

As it can appreciated in Figure 5.17 the model which predict better the COVID-19 cases with the meteorological influence is the second scenario, it means, model to predict COVID-19 transmission based on meteorological variables by a specific location (ABS). However, it can be seen some uptick COVID-19 cases in start period of lockdown so the goodness of fit will be a bad performance in this part.

With the aim of quantify the model adjustment, a determination coefficient is calculated in each scenario. Thus, the coefficient determination in the first scenario is 0.4245, that is, more

than a third of the COVID-19 propagation is explain by the set of selected independent variables. However, the goodness of fit between reconstructed COVID-19 data and original in the second scenario is 0.5937 which reflects a considerable improvement that in the first case.

Finally, the Barcelona-8L determination coefficient in case of the third scenario is 0.5300 which is better value in comparison with the first scenario. It can conclude that the mentioned scenarios are behave in a similar way due to the fact that ABS with a high population have a directly influence in the multiple linear regression and in the standardize regression coefficients of the model.

As it mentioned above, it can confirm that the best scenario to predict COVID-19 cases by meteorological influence is the scenario with a unique ABS.

In Figure 5.18, it can be displayed the model to predict COVID-19 propagation and meteorological regression fit for ASB 192 – Sabadell-2 in order to reproduce the analyzed scenarios and see the influence of a different ABS.

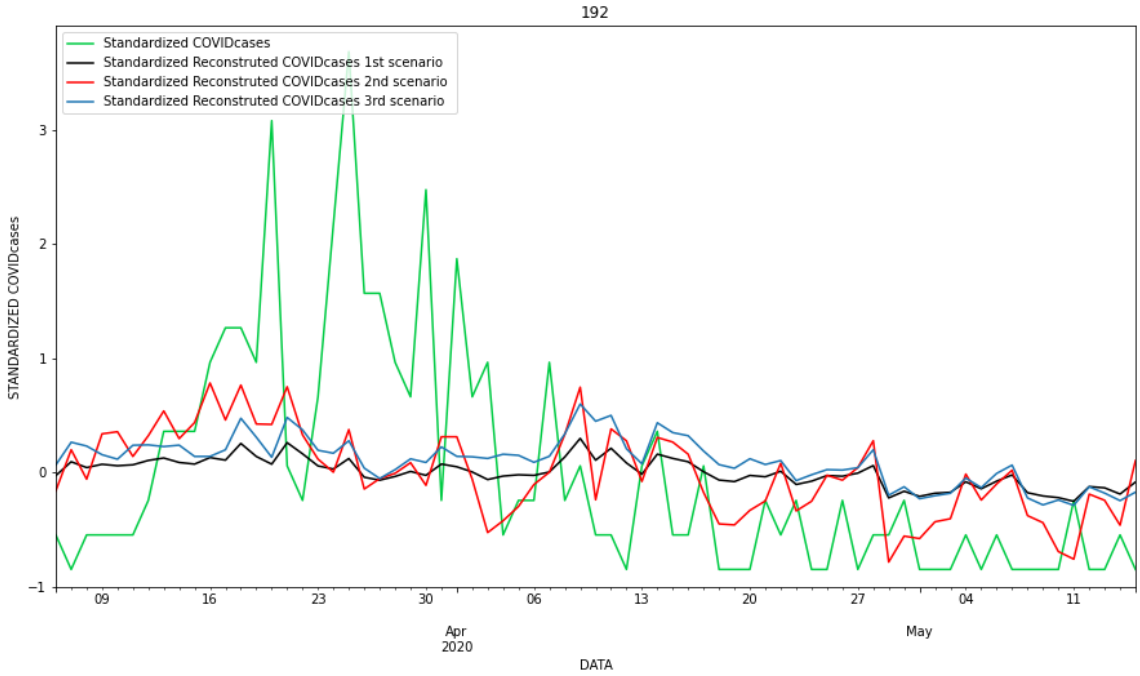

Figure 5.18 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on meteorological data (standardized values) for the three scenarios in the ABS "192 – Sabadell-2" during lockdown period.

It is clearly observed that the model which is adjusted better in comparison to original COVID-19 confirmed cases by Sabadell-2 ABS is the second scenario, nevertheless, the first scenario is closed to the second scenario trend. It should be noted that the adjustment is better at the end of lockdown period than at starting dates when the COVID-19 cases increased.

In order to verify both adjustments quantitatively, coefficient determination is estimated for each scenario. Hence, real COVID-19 cases and reconstructed COVID-19 data adjustment in the studied scenarios for Sabadell-2 ABS are 0.5389, 0.5689 and 0.4236 in each scenario, respectively. It can come to the same conclusion than ABS- 403- Bacelona-8L, so the second scenario by specific locations can reach a better adaptation in order to predict the COVID-19 positive cases.

Regarding the ABS influence in second scenario, Barcelona-8L ABS achieved a better goodness of fit in the model in comparison with Sabadell-2 ABS, it can be due to the fact that this specific district in Barcelona is bigger than Sabadell-2 ABS, thus high amount of data is obtained and more accurate the mathematical model will be.

To sum up, it can be appreciated the best scenario to predict COVID-19 incidence is second scenario analyzed by an ABS, it means that, the model to predict the incidence of COVID-19 based on meteorological variables in lockdown period has the best approximation to original COVID-19 cases by locations.

5.3.2 Regression fit based on meteorological parameters during in pandemic period

With the aim of predict COVID-19 cases with the influence of meteorological data, a model assessing COVID-19 incidence and meteorological regression fit in pandemic period is done and commented.

As mentioned before, three scenarios are needed to evaluate and find the best adjustment between COVID-19 cases from *Dades Obertes* and reconstructed COVID-19 data from regression fit. Thus, Table 5.2 shows the standardize regression coefficients for each scenario which are the components to the regression fit equation during pandemic period.

Table 5.2 Standardized regression coefficients to predict COVID-19 incidence for meteorological variables in each scenario during pandemic period.

	1st scenario	2nd scenario (ABS 403)	2nd scenario (ABS 192)	3rd scenario
Temperature	-0.0686	0.0311	0.1000	-0.0007
Relative humidity	-0.1809	-0.2296	-0.5183	-0.1941
Solar radiation	-0.2248	-0.4129	-0.2606	-0.2915

In regards to the first scenario evaluating the data of all ABS in Catalonia, it can observe that meteorological variables in pandemic have the same influence in COVID-19 propagation than in lockdown period, it means that, an intensification in COVID-19 confirmed cases are generated by a decrease in temperature, relative humidity and solar radiation. It should be pointed out that the meteorological variable which has more influence in COVID-19 cases is the solar radiation, follow by a relative humidity. However, in reference to temperature variable are not high correlated with COVID-19 cases.

Concerning the second scenario, it is done COVID-19 incidence and meteorological regression fit in a specific ABS during pandemic period. Following the procedure, second scenario is reproduced for ABS 403 - Barcelona-8L and ASB 192 – Sabadell-2, during pandemic period.

Thus, the relevant parameter in ABS 403 is solar radiation which is the variable that predicts better the COVID-19 cases, follow by relative humidity. Both variables are negative correlated in reference to COVID-19 spread, on the contrary, temperature parameter.

With the intention to check the second scenario performance, a COVID-19 and meteorological regression fit for ABS 192 – Sabadell-2 is executed and the regression coefficients for the model are shown in Table 5.2.

The relative humidity is the variable which predict better the model COVID-19 incidence and meteorological regression fit in Sabadell-2 ABS, due to its highest magnitude. Solar radiation and temperature are lower impact in comparison with solar radiation. In this specific case, a growth in COVID-19 positive cases can be related with low relative humidity and solar radiation and increasement of temperature.

The standardized regression coefficients magnitude differs from Barcelona-8L ABS and Sabadell-2 ABS in second scenario during pandemic, it can be due to the fact that difference in geographical area has an effect in some meteorological parameters.

Regarding the third scenario, COVID-19 propagation and meteorological regression fit is done for ABS from municipalities more than 200.000 population during pandemic period. Thus, the standardize regression coefficient which form the model to predict COVID-19 based on meteorological variables in pandemic can be seen in Table 5.2.

Hence, meteorological variables are negative correlated with the prediction of COVID-19 spread in case of pandemic period. The variable which predicts in a better way the COVID-19 cases is the solar radiation and, in case of temperature, it cannot observe a correlation in this specific case.

The standardized regression coefficient in pandemic is different in comparison with lockdown period. It is expected these dissimilarities due to the fact that intervention of different factors like the seasonality, detected COVID-19 variants during pandemic period, mobility factor which highly decrease with localities restrictions, among other factors.

Figure 5.19 displayes the model regression fit tendency of the model to predict positive cases with the influences of meteorological variables of each scenario taking into consideration ABS 403 - Barcelona-8L during pandemic.

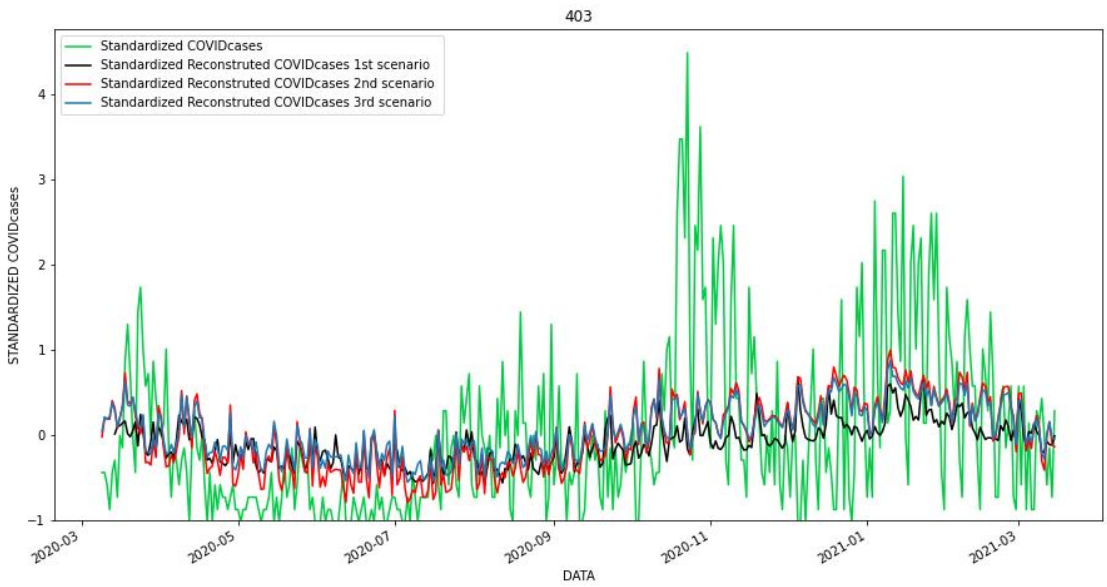

Figure 5.19 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on meteorological data (standardized values) for the three scenarios in the ABS “403-Barcelona-8L” during pandemic period.

As the Figure 5.19 shows, the scenarios are followed by the same performance along the pandemic period, and it is difficult to select the best scenario. Moreover, it should be pointed

out that original COVID-19 cases present some representative pick during pandemic period which indicate an increase in COVID-19 cases may be due to the fact impact on detected COVID-19 variants, among other factors. In these specific picks, the regression fit model does not achieve the original positive cases trend.

In order to estimate the adjustment formed by the regression fit data and real COVID-19 cases in pandemic to decide the best model, determination coefficient is calculated to quantify the adjustment for Barcelona-8L ABS in the analyzed scenarios.

Hence, the determination coefficient to calculate the goodness of fit between the reconstructed and original COVID-19 data is 0.5131, 0.5833 and 0.5586 for first, second and third scenario, respectively. The best approximation model to predict COVID-19 confirmed cases based on meteorological variables is the scenario evaluated with a unique ABS.

Following the same procedure, in Figure 5.20 can be seen reconstructed COVID cases with original COVID cases from ABS 192 – Sabadell-2 during pandemic.

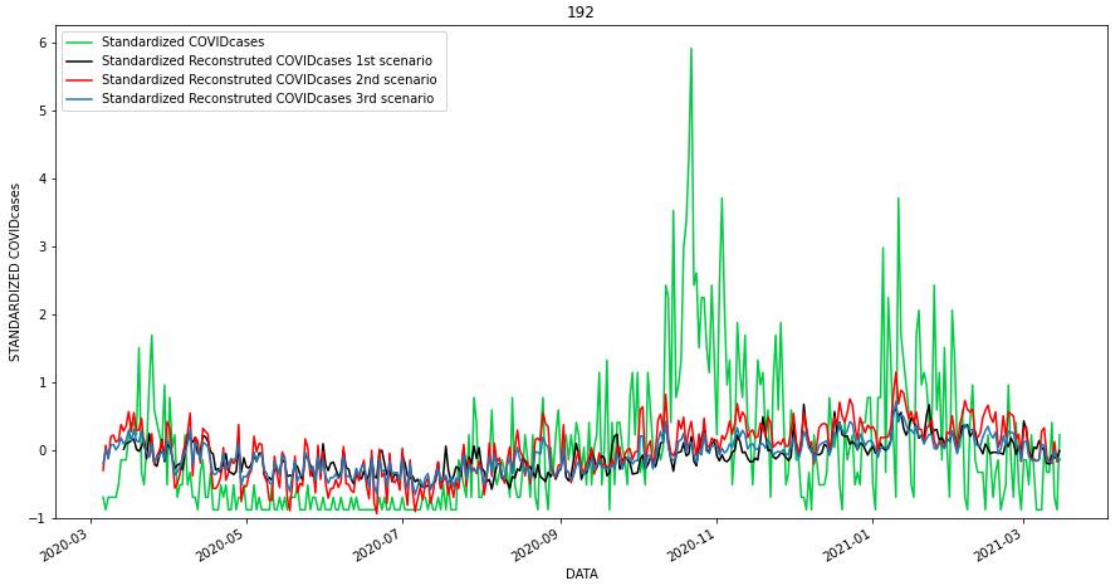

Figure 5.20 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on meteorological data (standardized values) for the three scenarios in the ABS "192 – Sabadell-2" during pandemic period.

Figure 5.20 is followed by a similar trend in comparison with Figure 5.19, nevertheless, minor modification in original data is experimented in this specific location. There is not enough evidence to select the scenario which predicts better original COVID-19 propagation.

With the goal to decide the best scenario, it is essential to quantify the adjustment between real COVID-19 confirmed cases and reconstructed COVID-19 data from the predicted regression fit. Hence, a determination coefficient is calculated for Sabadell-2 ABS which are 0.5543, 0.5748 and 0.5671. A close value between the different scenarios is obtained, nevertheless, the best performance model to predict COVID-19 and meteorological regression fit is the second by specific location.

Regarding the second scenario with different evaluated ABS, it can conclude with the fact that Barcelona-8L ABS adjustment is slightly better than in Sabadell-2 ABS. A possible

observation can be that dispersity data due to pandemic period end up in a similar approximation between analyzed specific locations.

To conclude, the scenario which predict better COVID-19 propagation with the influence of meteorological variables in pandemic are the second scenario by ABS. It can end up with the same conclusion that in lockdown period, it means that, model to predict the COVID-19 propagation based on meteorological variables in lockdown and pandemic period by locations can be predicted more accurate a mathematical model.

It should be pointed out that slightly difference is between lockdown and pandemic models, it can be due to the fact that other factors are taking into account in pandemic period.

5.3.3 Regression fit based on air pollutants parameters during in lockdown period

With the aim of providing the best performance to predict COVID-19 propagation with the air pollutants parameters, diverse scenarios are evaluated with the multilinear regression model in lockdown period.

The regression coefficients of each air pollutant parameter which form the regression fit equation to predict COVID-19 incidence with air pollution data can be seen in Table 5.3.

Table 5.3. Standardized regression coefficients to predict COVID-19 incidence for air pollutants variables in each scenario during lockdown period.

	1st scenario	**2nd scenario (ABS 403)**	**2nd scenario (ABS 192)**	**3rd scenario**
PM10	0.0137	-0.0873	0.0552	0.0241
PM2.5	-0.0282	0.0344	-0.1201	-0.0779
SO_2	-0.0307	0.2116	-	0.0102
NO	-0.0816	-0.4867	-0.0944	-0.0440
NO_2	0.1747	0.2865	0.1741	0.0809
CO	0.0712	0.2472	-	0.1205
O_3	-0.0225	-0.3309	-0.3617	-0.0375

In reference to first scenario, it can clearly appreciate that nitrogen dioxide is the variable that have more importance in the COVID-19 prediction because of the standard regression coefficient has the highest value in the equation in absolute value. This air pollution component is directly related with the increment of COVID-19 cases, that is, nitrogen dioxide increases and consequently, it will produce an increment of the COVID-19 cases.

Moreover, the nitrogen monoxide is the second variable which have influence in the multilinear regression fit and describes better the COVID-19 propagation. On the contrary, nitrogen monoxide is behaved different to the nitrogen dioxide, at low concentration of nitrogen monoxide, COVID-19 propagation increase. Other air pollutants variables evaluated have low relative importance in the regression equation due to the fact that low standardize regression coefficients are obtained.

Hence, the increasing of the PM10 and carbon oxide concertation is related with the increase of COVID-19 propagation. On the contrary, it is produced an increment of the positive cases when the concentration of PM2.5, sulfur dioxide and ozone are in low values.

Concerning the second scenario, the model to explain the influence of COVID-19 propagation with the main pollution in air by unique ABS are assessed for ABS 403 - Barcelona-8L and ASB 192 – Sabadell-2, during lockdown period.

The air pollutant parameters that describes better the performance of COVID-19 cases in Barcelona-8L ASB is nitrogen monoxide which has a high influence in the dependent variable of the studied model. It is a different approximation that is found in the first scenario taking into account all ABS.

In reference to other atmospheric pollutants, the standardize coefficient regression have a high influence in the COVID-19 propagation, except the suspended particles (PM2.5 and PM10) which present low coefficients.

Following the data analyze, the contrary conclusion of the first scenario is reached with the regression coefficients of the suspended particles in this scenario. Concretely, PM10 in a low concentration influence in the expansion of COVID-19 propagation. On the contrary way, increase COVID-19 cases if the concentration of the air pollutant PM2.5 is high.

Regarding the second scenario with ABS 192 – Sabadell-2, ozone is the air pollutant variable that describes better the influence in COVID-19 propagation, followed by the nitrogen dioxide and suspended particle PM2.5. Hence, an augmentation of the COVID-19 cases is influenced by the low concentration of ozone and PM10 and a high concentration in the air pollutant nitrogen dioxide.

In regards to other atmospheric contaminants, suspended particle PM2.5 and nitrogen monoxide are not enough correlated to confirm that it has influenced the multilinear regression coefficient. It should be noted that sulfur dioxide and carbon monoxide are not correlated in this specific scenario and ABS 192, this fact is due to the Sabadell-2 ABS have not data for these air pollutants parameters.

Concerning the third scenario where it simplified the ABS which municipalities less than 200.000 inhabitants, COVID-19 incidence and air pollution regression fit is reproduced.

As it is observed in the table above, the tendency of standard regression coefficients is similar than in the first scenario. An increment of COVID-19 propagation can be produced by a high concentration values of suspended particles PM10, sulfur dioxide, nitrogen dioxide and carbon monoxide and a low concentration of the atmospheric pollutant PM2.5, nitrogen monoxide and ozone.

In this case, carbon monoxide has the highest value in the model in absolute value, so it is the pollutant in air which has more influence in the prediction of COVID-19 spread.

In order to assess the model to predict COVID-19 propagation and atmospheric pollutants regression fit in Barcelona-8L ABS, Figure 5.21 can be plotted to make the comparison between standardized original COVID-19 cases and the reconstructed COVID-19 data of each scenario.

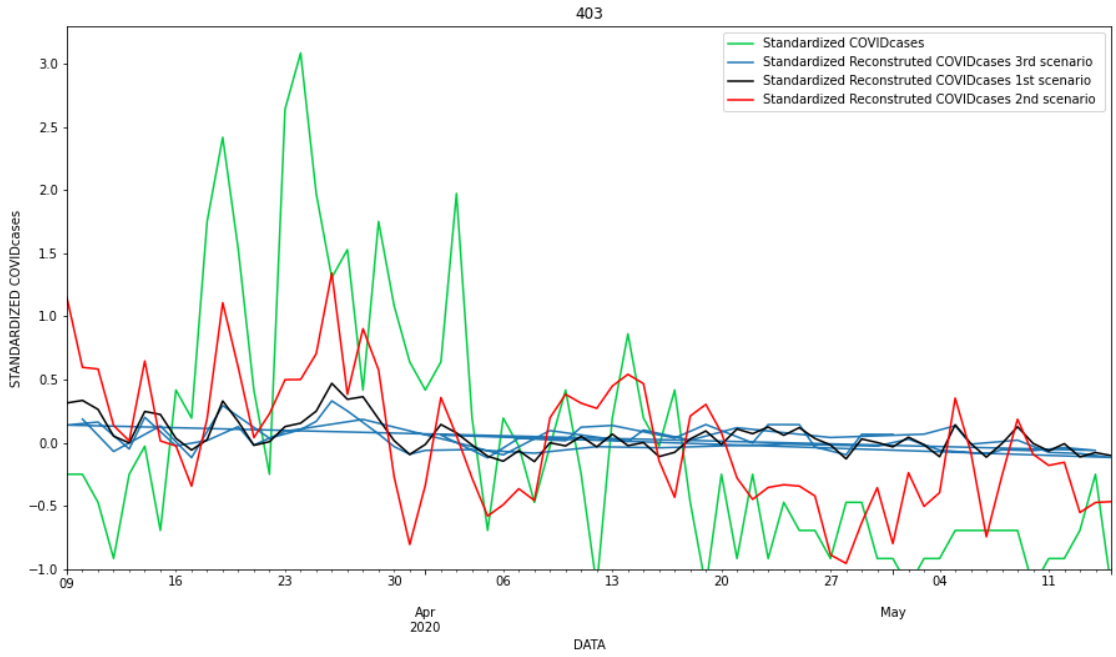

Figure 5.21 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on atmospheric pollutants parameters (standardized values) for the three scenarios in the ABS "403 – Barcelona-8L" during lockdown period.

In Figure 5.21, it should be pointed out reconstructed COVID-19 tendency in second scenario are a similar adjustment in real COVID-19 cases curve. However, a bad performance in this specific case is reached from 18th of March to 4th of April which can be caused to a positive cases upturn. On the other hand, first and third scenario are no do like a proper adaptation in reference to COVID-19 confirmed cases extract from *Dades Obertes de Salut.*

With the objective to quantify the goodness of fit between original cases and reconstructed COVID-19 incidence, a determination coefficient is evaluated.

The model fit in the first scenario taking into account ABS 403 – Barcelona 8L is around 0.4885, so the 49% of the variation in COVID-19 cases is due to the modification of the combination of the air pollutant concentration data. Concerning other scenarios, determination coefficient is 0.6281 in scenario by unique ABS and 0.5024 in case of general regression fit with municipalities' ABS with more than 200.000 inhabitants.

To conclude, the first and third scenario for Barcelona-8L ABS in lockdown are close values due to the fact of model influences in localities with high population. It should be pointed out that model to predict COVID-19 with air pollutant parameters influence are produced a high performance by selecting a unique ABS.

For the purpose of corroborate the predicted model COVID-19 transmission, regression fit adjustment between the model obtained from multiple linear regression and COVID-19 cases extracted from Catàleg de *Dades Obertes de Salut* can be displayed in Figure 5.22 for ABS 192 – Sabadell-2 during lockdown period.

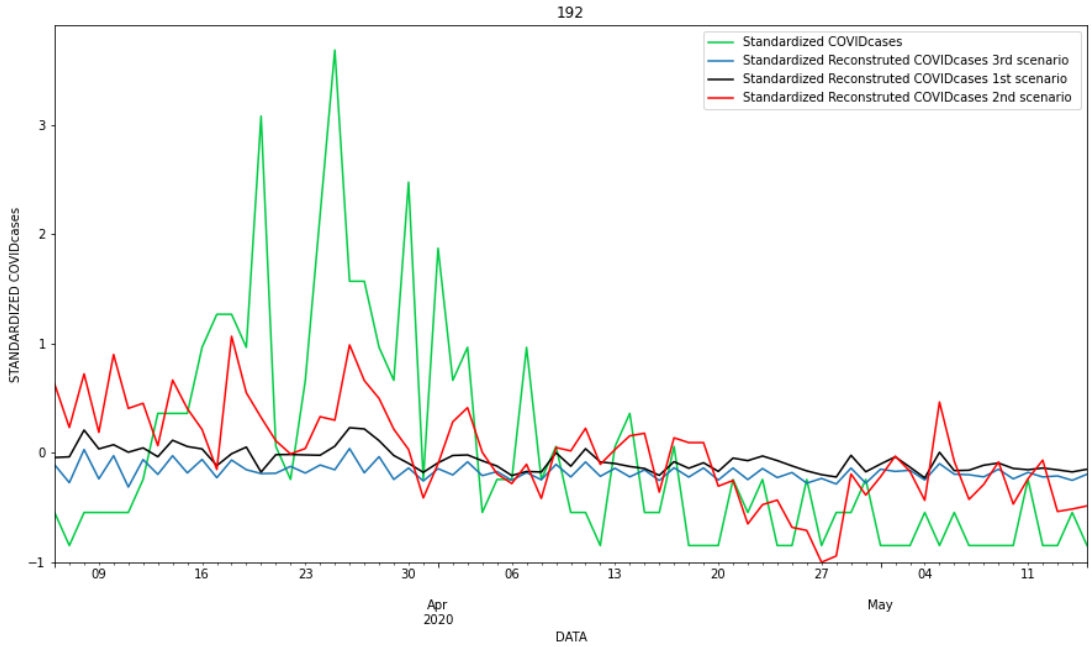

Figure 5.22 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on atmospheric pollutants parameters (standardized values) for the three scenarios in the ABS "192 – Sabadell-2" during lockdown period.

It can observe in Figure above that the disposition of different scenarios are worse than in ABS 403 during lockdown period. The prediction of reconstructed COVID-19 cases can not arrive a real confirmed case, in spite of second scenario are emphasized for the best performance over evaluated scenarios.

Thus, it is relevant to quantify the goodness of fit about standardize reconstructed data of each scenario in comparison with the original COVID-19 cases in Figure 5.22 by a determination coefficient.

In reference to the determination coefficient of ABS 192 – Sabadell-2 in lockdown, it is obtained values of 0.5318, 0.5929 0.5222 for each scenario which has a similar performance in comparison with Barcelona-8L ABS. It is important to highlighted that the scenario that predict better COVID-19 cases is the second by analyzing with a specific location.

Concerning the influences of ABS during lockdown, it can appreciate that a good performance is obtained for Sabadell ABS. However, the best performance is the second scenario which Barcelona-8L is reached a better determination coefficient. These facts can be explained with the specific geographical area influence and the register cases distribution in Sabadell with 10 ABS and Barcelona with 67 ABS.

To conclude the section, the best model to predict COVID-19 incidence based on atmospheric pollution in lockdown is the second scenario assessing by ABS, hence it will expect a better mathematical model by localities.

5.3.4 Regression fit based on air pollutants parameters during in pandemic period

The model to forecast the COVID-19 propagation and atmospheric pollutants regression fit during pandemic are done and, in this section, the results for each scenario will be commented.

Table 5.4 shows the standardize regression coefficients which create the regression fit equation in this case in order to predict COVID-19 cases based on air pollutants data.

Table 5.4 Standardized regression coefficients to predict COVID-19 incidence for air pollutants variables in each scenario during pandemic period.

	1st scenario	2nd scenario (ABS 403)	2nd scenario (ABS 192)	3rd scenario
PM10	0.0266	0.00150	-0.0769	-0.0930
PM2.5	-0.0425	-0.0984	-0.2310	-0.0801
SO_2	-0.0749	-0.2091	-	-0.1038
NO	-0.00711	-0.1095	-0.3269	-0.0510
NO_2	0.1512	-0.00149	-0.5635	0.1390
CO	0.1224	0.2222	-	0.2179
O_3	-0.1055	-0.2441	-0.3467	-0.1457

As it can see in the Table 5.4 concerning the first scenario, the atmospheric pollutant with high influence in the regression model equation is nitrogen dioxide. In lockdown, it is occurred the same, but the other air pollutants variables will remain in a minor influence. However, carbon monoxide and ozone are high correlated in pandemic period, this fact can be due to the fact that other factors are implicated such as the free mobility in the location which is a difference in lockdown period.

An increment in COVID-19 propagation is because of a high concentration on nitrogen oxide and carbon oxide and a decrease in ozone concentration. Other air pollutants parameters in pandemic have low relative weight in the equation of the model to predict COVID-19 spread based on atmospheric pollutants.

In order to assess the influence of ABS in reference to second scenario by unique ABS, it is done COVID-19 propagation and pollutant in air regression fit for ABS 403 - Barcelona-8L and ASB 192 – Sabadell-2, during pandemic period.

Regarding to Barcelona-8L ABS in second scenario, air pollutant parameter which has more influence is ozone, sulfur dioxide and carbon monoxide which presents a high magnitude. In comparison with the results from lockdown period has different tendency, it can be because of pandemic data has more dispersity of air pollutants parameters.

Concerning Sabadell-2 ABS, nitrogen oxides are the atmospheric pollutant that describes the COVID-19 influence variable, followed by ozone and nitrogen monoxide. Thus, low concentration of the pollutant in air will involve an augmentation in positive cases. It should be pointed out that carbon monoxide and sulfur dioxide have an inexistence coefficient due to the fact that the ABS 192 do not have a pollutant station which measured these pollutants, so no data found.

Regarding the third scenario, the results from a general regression fit between COVID-19 incidence and atmospheric pollutants taking into account the municipalities' ABS with more than 200.000 inhabitants in pandemic period will be commented. The important air pollutants parameters in this scenario are carbon monoxide, ozone and nitrogen dioxide that will be more influences variables in the COVID-19 and air pollutant regression fit model. An increase of

COVID-19 confirmed cases are produced by a high nitrogen dioxide and carbon monoxide concentration and a low concentration of ozone.

Hence, it is relevant to visualize Figure 5.23 which is plotted aforementioned scenarios performance in comparison with original COVID-19 cases for 403 – Barcelona-8L ABS.

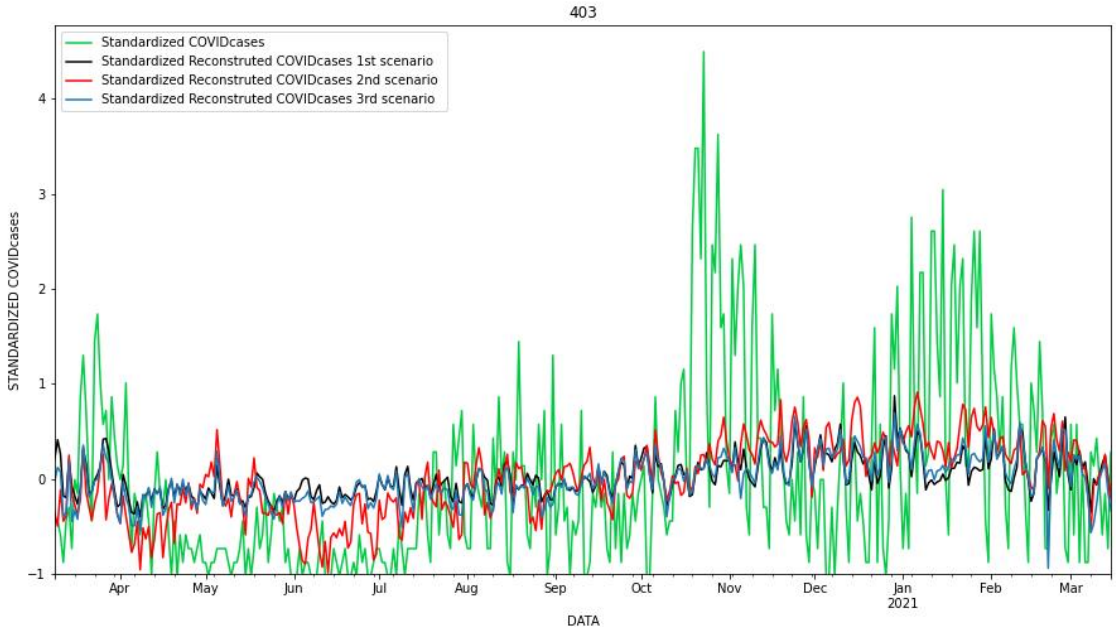

Figure 5.23 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on atmospheric pollutants parameters (standardized values) for the three scenarios in the ABS "403 – Barcelona-8L" during pandemic period.

Overall, the reconstructed COVID-19 incidence with atmospheric pollutants influence of evaluated scenario do not predict sufficiently real COVID-19 cases during pandemic. Among three scenarios, it should be pointed out that second scenario are tried to follow original cases curve in some months, such as between April and May.

In order to quantify the readjust linking the reconstruct COVID-19 cases from the regression model prediction and COVID-19 positive cases from *Dades Obertes de la Salut*, goodness of fit is calculated in this case.

Determination coefficients for model to predict COVID-19 propagation based on atmospheric pollutants influence in ABS 403 – Barcelona-8L during pandemic are 0.4992, 0.5764 and 0.5192 for each scenario, respectively. It is clearly observed a best performance when the model to predict COVID-19 based on atmospheric pollutants are done by a specific location, it means that, *Àrees bàsiques de* salut (ABS).

Following the procedure done in the last sections, it is reproduced model regression fit between COVID-19 propagation and air pollutants in ASB 192 – Sabadell-2 which can show in Figure 5.24.

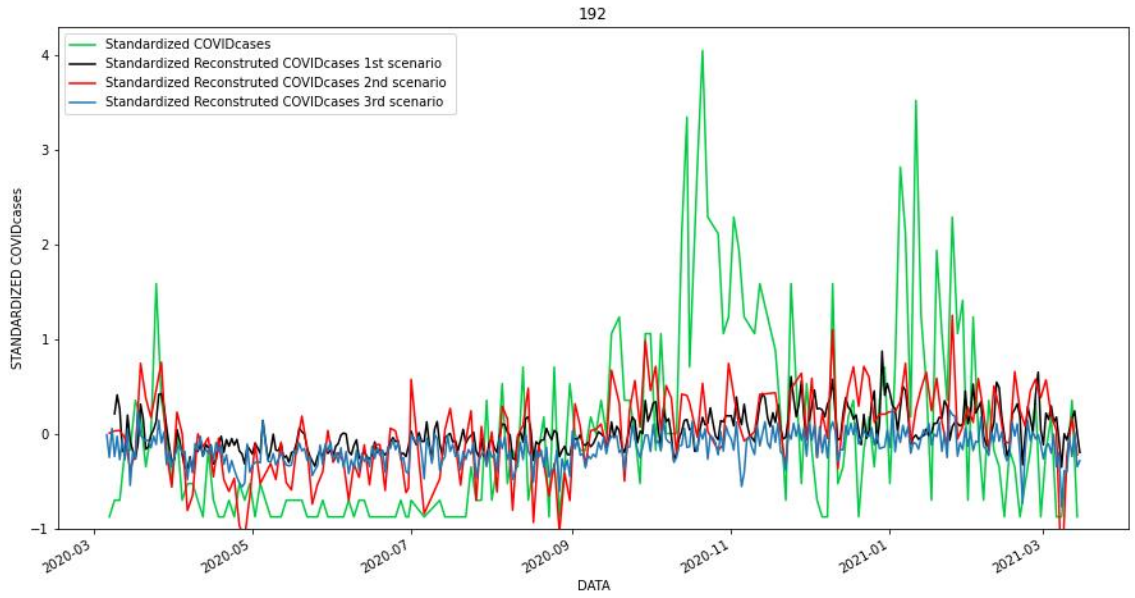

Figure 5.24 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on atmospheric pollutants parameters (standardized values) for the three scenarios in the ABS "192 – Sabadell-2" during pandemic period.

As it can be observed in Figure 5.24 regarding Sabadell-2 ABS, the performance is better than Barcelona-8L ABS because of the adjustment between reconstructed COVID-19 cases and the real one is closed in a specific month, for instance it can see that fact in April 2020, from August to October 2020 and at the end of pandemic period. It cannot appreciate clearly the scenario which has the proper trend, so for that reason, it is crucial to evaluate the goodness of fit quantitatively.

With the aim to clarify the best trend and select the best approximation scenario, goodness of fit is computed by determination coefficient, thus, it can come to the result of 0.542, 0.6038 and 0.5395 in specific scenario in reference to Sabadell-2 ABS.

In comparison with Barcelona ABS, best adjustments between reconstructed COVID-19 spread from regression fit equation and original COVID-19 confirmed cases can be seen during pandemic period. The difference can be produced by the dispersity of data, influence on detected COVID-19 variants, among other factors such as mobility.

It should be noted that there is a slightly performance in predict COVID-19 and atmospheric pollutants regression fit during pandemic period. This fact can be because of large amount of data which will predict a more reliable mathematical model and detected COVID-19 variants so high collected data ration will be produced.

To sum up, model to predict COVID-19 propagation and atmospheric pollutants regression fit in pandemic are achieved superior goodness of fit in the second scenario work with unique ABS. Thus, it can end up with identical conclusion that in meteorological regression fit, so the model to predict COVID-19 propagation with meteorological and atmospheric pollutant will be efficient to model by specific locations.

5.3.5 Regression fit based on atmospheric environmental parameters during in lockdown period

The aim of this section is to discuss a common framework to predict propagation of COVID-19 through the air vector associated with the meteorological and air pollution parameters in the three possible scenarios aforementioned.

The standardize regression coefficients in reference to a general multilinear regression fit of all scenarios and the selected variables during lockdown period can be seen in the Table 5.5.

Table 5.5 Standardized regression coefficients to predict COVID-19 incidence environmental (meteorological and atmospheric pollutants) parameters in each scenario during lockdown period.

	1st scenario	2nd scenario (ABS 403)	2nd scenario (ABS 192)	3rd scenario
Temperature	-0.1057	-0.4150	0.0784	-0.1216
Relative humidity	-0.0135	-0.2845	-0.2847	0.1185
Solar radiation	-0.0629	-0.1952	-0.3132	-0.0344
PM10	0.0123	-0.1038	0.0156	0.00246
PM2.5	-0.000638	-0.0373	-0.1285	-0.0490
SO_2	-0.0433	-0.1796	-	0.0114
NO	-0.0912	-0.0528	-0.0956	-0.0758
NO_2	0.222	0.6754	0.1512	0.1038
CO	0.0695	-0.2909	-	0.1259
O_3	0.000718	-0.1930	-0.1930	-0.0249

As it can appreciate in reference to the first scenario, the model equation for predict propagation COVID-19 with the environmental and air pollution variables are formed by the coefficients in Table 5.5. It is important to highlight those variables with relative importance in predict COVID-19 are the same than obtained in the last section, that is, nitrogen dioxide and temperature are highly influence with the propagation of COVID-19 among other data. Hence, suspended particle PM2.5 and ozone have an extremely low regression, and these are barely influenced in the dependent variable COVID-19 spread.

In reference to the meteorological variables, these are directly correlated with COVID-19 cases, an increment of positive cases is produced by the decrease of temperature, relative humidity and solar radiation.

Therefore, the pollutant in air has the exact disposition than in the first scenario of COVID-19 and air quality regression fit. The high concentration of PM10, NO_2 and CO are related with an increasement in COVID-19 propagation. The other way around, SO_2 and NO with a low concentration present in the air influenced positively in the accumulation of positive COVID-19 cases.

Regarding the second scenario, a multiple linear regression is done for a single ABS, for instance ABS 403 – Barcelona 8L and separately for ABS 192 – Sabadell-2 to assess the influence in the selection of the specific ABS.

The standardize regression coefficient obtained for meteorological and air pollution variables point up a high relation with the prediction of COVID-19 cases, this may be due to the fact input data of a specific ABS adjust more accurately than case of a general ABS data in the first scenario. It should be noted that temperature and nitrogen dioxide are variables which predict better the positive COVID-19 cases, follows by carbon monoxide and relative humidity.

The performance of the variable is changed in comparison to the first scenario, that is, the decrease of the variables involves and increment of COVID-19 propagation, except in nitrogen dioxide case. This fact can be explained in the sense of the second scenario are extremely related with the unique ABS data and in case of first scenario are affected for the tendency of the majority of ABS data carrying the performance.

In regard to Sabadell-2 ABS, the variable with highest influence in order to predict COVID-19 propagation is solar radiation followed by the relative humidity and ozone air pollutant. Sulfur dioxide and carbon dioxide pollutant in air has no relation to COVID-19 prediction because there are no data about these pollutants in this specific ABS. Moreover, temperature, PM10 and nitrogen monoxide have not hardly influence in the prediction of COVID-19.

An augmented COVID-19 positive cases will give to a decrease of relative humidity, solar radiation, low concentration of PM2.5, nitrogen monoxide and ozone and a temperature increase and PM10 and nitrogen dioxide high concentration value.

The evaluated variables in ABS 192 have different tendency in comparison with 403ABS, it can be caused by the extremely influence of population and geographical situation.

As it can appreciate in the table above regard to third scenario, carbon monoxide and relative humidity have the highest values in comparison with the other data, so these are the parameters which describes better the COVID-19 performance in this specific case.

Regarding with meteorological variables, the relative humidity has different tendency than conclusions extract from the study of COVID-19 incidence and meteorological variables, it means that, at high levels of relative humidity can be produced an increase of COVID-19 cases. On the contrary, an expansion of COVID-19 propagation will be produced by a temperature and solar radiation drop.

Concerning the air pollutants parameters, follows the similar approach than results from the COVID-19 propagation and pollutant in air. However, nitrogen dioxide is highly correlated than the nitrogen monoxide.

Results obtained from regression coefficients which formed regression model equation can be found in Figure 5.25 taking into account all scenarios during lockdown.

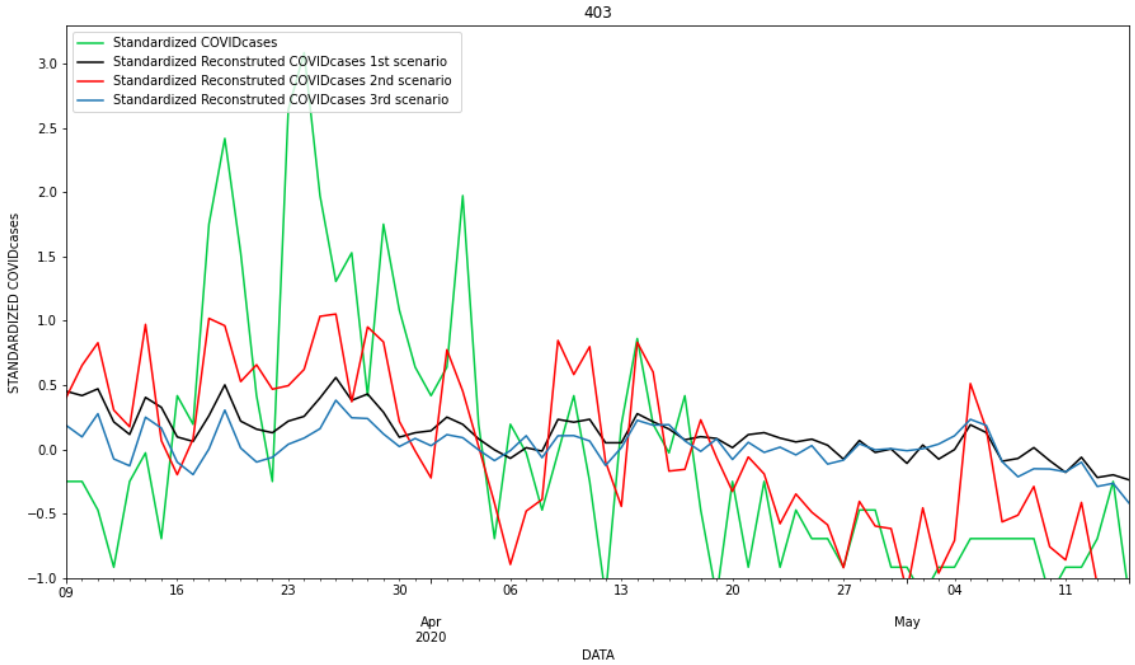

Figure 5.25 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on environmental (meteorological and atmospheric pollutants) parameters (standardized values) for the three scenarios in the ABS "403 – Barcelona-8L" during lockdown period.

It can observe that second scenario which is predicted by a unique ABS is the best adaptation to original COVID-19 curve, nevertheless, it is not adjusted well in the start of lockdown and at the end of studied period. Second and third scenario follow a similar straight trend which is not adjusted properly over lockdown period.

With the goal of seeing clearly the adjustment multilinear regression model between the original COVID-19 cases and the COVID-19 cases predict from the multiple linear regression fit in each scenario during lockdown, a determination coefficient is calculated by ABS 403 – Barcelona-8L.

Thus, the goodness of fit in COVID-19 through the air vector composed by meteorological and air pollution regression fit are 0.5171, 0.6983 and 0.5200, respectively for each evaluated scenario. Hence, it is indicated an enhancement in the propagation COVID-19 prediction considering common framework of meteorological and air pollutant variables than these variables evaluating separately. It should be pointed out that the scenario which reconstructed better COVID-19 transmission are the second by ABS.

In order to check the influence of ABS, it is reproduced the same procedure but regarding to ABS 192 – Sabadell-2 in Figure 5.26.

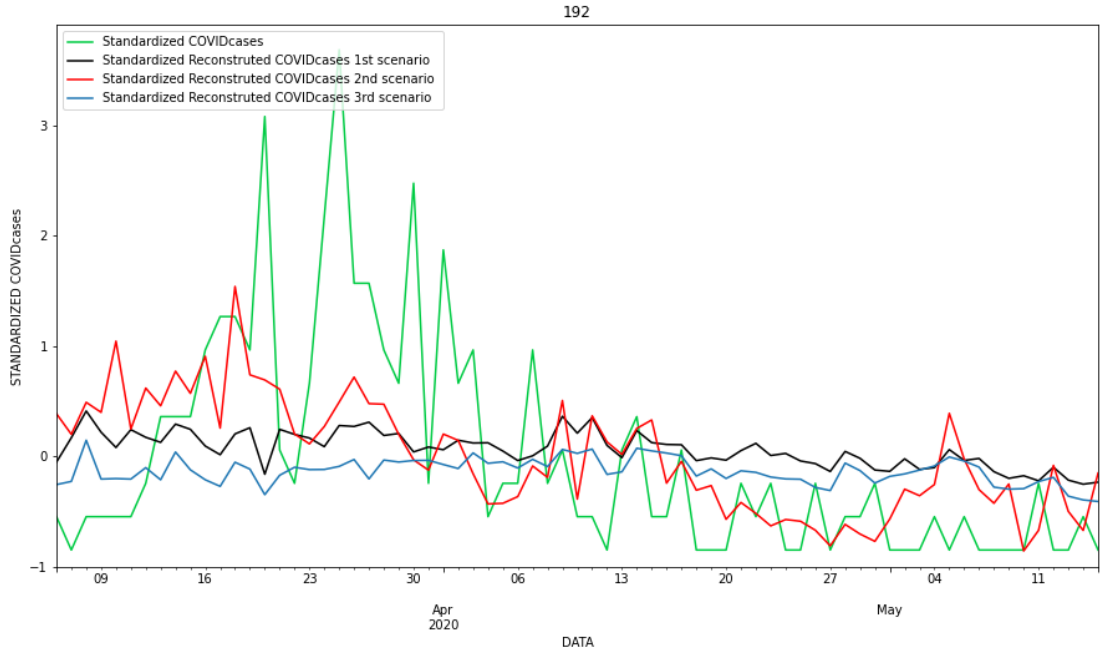

Figure 5.26 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on common framework of meteorological and atmospheric pollutants parameters (standardized values) for the three scenarios in the ABS "192 – Sabadell-2" during lockdown period.

Concerning Figure 5.26, it has reached the same disposition than in Figure 5.25 which is in reference to Barcelona-8L ABS in lockdown. Hence, it should be pointed out the performance in second scenario, but this curve does not predict enough the original one.

In order to verify quantitatively the adjustment between reconstructed data from regression fit and COVID-19 positive cases, determination coefficient is evaluated for each scenario which values are 0.5382, 0.6286 and 0.5034.

Overall, goodness of fit of COVID-19 and environmental regression fit in Sabadell-2 ABS is obtained a low performance compared with the same model regression fit in Barcelona-8L. These differences are clearly related with the geographical area in the specific ABS in Barcelona and Sabadell and the register distribution of positive cases in ABS municipalities.

Moreover, it can conclude that the best performance to predict COVID-19 is the scenario by unique ABS and it is reached better adjustment when it is modelled meteorological and atmospheric pollutants simultaneously than separately.

To summarize, the model to forecast COVID-19 propagation in reference a common framework between meteorological and atmospheric pollutants in lockdown period can achieve better adjustment by locations. Therefore, the improvement is more desirable taking into account the meteorological and atmospheric data simultaneously than independent.

5.3.6 Regression fit based on atmospheric environmental parameters during in pandemic period

In this part, the purpose is to assess the model to predict COVID-19 propagation based on atmospheric environmental and air pollution variables simultaneously during the pandemic period. Thus, model to predict COVID-19 incidence and atmospheric environmental variables

regression fit are composed by the standardize regression coefficients in Table 5.6 which are detailed for each evaluated scenario.

Table 5.6 Standardized regression coefficients to predict COVID-19 incidence environmental (meteorological and atmospheric pollutants) parameters in each scenario during pandemic period.

	1st scenario	**2nd scenario (ABS 403)**	**2nd scenario (ABS 192)**	**3rd scenario**
Temperature	-0.0590	-0.0616	0.1652	0.0177
Relative humidity	-0.0941	-0.2328	-0.2603	-0.1846
Solar radiation	-0.1653	-0.1737	-0.3909	-0.2274
PM10	0.00863	0.0984	-0.0980	-0.0242
PM2.5	-0.0424	-0.1321	-0.00308	-0.1056
SO_2	-0.0778	-0.1992	-	-0.0924
NO	-0.0338	0.0280	-0.1241	-0.0382
NO_2	0.2165	0.07395	0.1523	0.0983
CO	0.07694	-0.0217	-	0.1449
O_3	-0.0285	-0.1778	-0.2919	-0.0920

Regarding the general scenario, solar radiation and nitrogen dioxide are the variables which high relative weight in the model regression equation, it means that, the variables that describes more favorable the COVID-19 incidence in pandemic. The other regression coefficient variables are influenced in minor way in the independent variable of COVID-19.

The meteorological variables disposition in the model to predict COVID-19 are indirectly correlated to COVID-19 cases, it means that, a decrease of meteorological variables will produce an increasement COVID-19 propagation. In case of the air pollutants variables, the majority of the atmospheric pollutants are negative correlated with COVID-19 incidence except the nitrogen dioxide.

Follow the same procedure, a second scenario is computed in order to obtain the forecast of COVID-19 incidence with atmospheric environmental variables in a unique ABS, for instance in ABS 403 -Barcelona-8L and ABS 192 – Sabadell-2.

Concerning the second scenario in Barcelona-8L ABS, the standardize regression coefficient applied a multiple linear correlation model can be seen in Table 5.6.

Relative humidity, solar radiation, sulfur dioxide and ozone are the variables that have more influence in order to predict the COVID-19 propagation. Both of them are inversely correlated with positives cases, that is, an increase in COVID-19 confirmed cases are produced by a low value of solar radiation and relative humidity and a low concentration of ozone and sulfur dioxide air pollutant.

As table above shows, the variables that describes better the performance of model regression fit to predict COVID-19 are the meteorological variables, it means that, temperature, relative humidity and solar radiation. It can be concluded that meteorological variables have

more influence in comparison with atmospheric pollutants in the model to predict COVID-19 spread in the scenario by unique location.

It should be noted that the tendency of the mayor influenced variables are negative correlated with the increasement of confirmed cases, on the contrary than nitrogen dioxide and temperature. As it mentioned above, sulfur dioxide and carbon monoxide are air pollutants variables that in case of Sabadell-2 ABS cannot have data, so it cannot establish a relationship in reference to COVID-19 cases.

In reference to the third scenario, a model to predict COVID-19 through an air vector composed by meteorological and air pollutants regression fit is executed in pandemic period.

In order to show the model regression fit equation in this case, the regression coefficient of each variable taking into account in this model can be seen in Table 5.6.

The meteorological variables are predominated by the air pollutant due to the fact that relative humidity and solar radiation have a high relative weight, so it directly influences in the prediction of COVID-19 cases. In general, COVID-19 increasement is produced by the low meteorological variables and air pollutants, except the temperature, nitrogen dioxide and carbon monoxide.

In order to estimate to the model to predict COVID-19 transmission with the atmospheric environmental variable influence regression fit in different scenarios, Figure 5.27 can be plotted taking into consideration ABS 403 - Barcelona-8L.

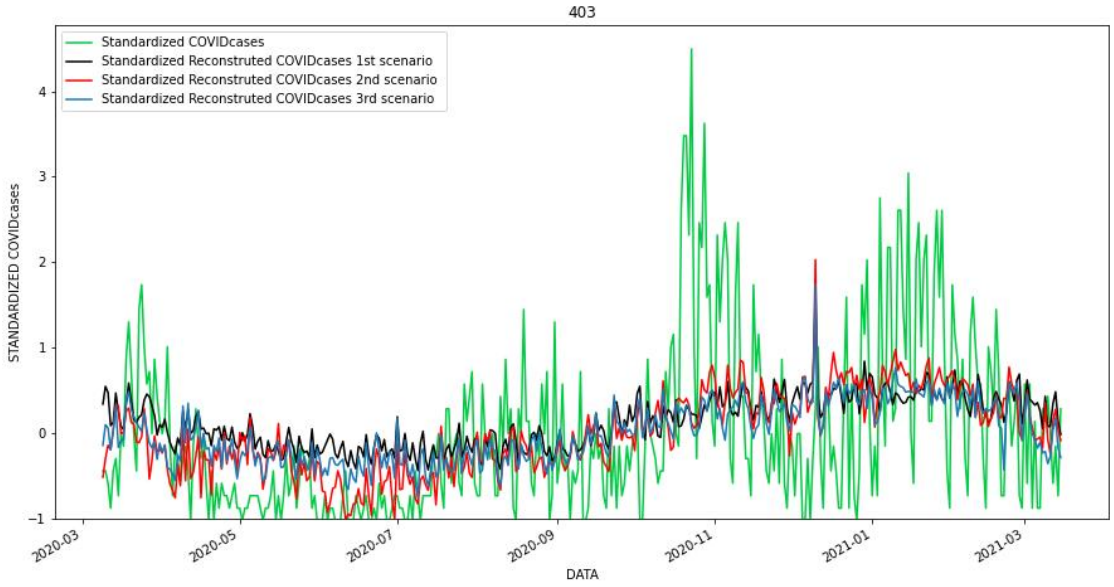

Figure 5.27 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on environmental (meteorological and atmospheric pollutants) parameters (standardized values) for the three scenarios in the ABS "403 – Barcelona-8L" during pandemic period.

Figure 5.27 shows disposition of each scenario in reference to standardized COVID-19 cases during pandemic period. Scenarios have a similar tendency along pandemic date, nevertheless, second scenario seems to be the best approximation with real cases. However, the adjustment of both curves decreases in presence of some upturn in real cases.

With the aim of checking the adaptation between reconstructed COVID-19 cases from the regression equation fit and real positive cases, determination coefficients are calculated for the

different scenarios which result are 0.5351, 0.6085 and 0.5629 during pandemic period. As it can see the best scenario to predict COVID-19 transmission is by specific locations (ABS) and that value is differed to lockdown period. This difference between the goodness of fit can be produced by the COVID-19 variants detection during pandemic which can disrupt model prediction and consequently, give rise to a bad performance.

Furthermore, it clearly observed an improvement in the adjustment in comparison with meteorological and atmospheric pollutants taking into consideration separately.

Regarding to the ABS 192 – Sabadell-2, the representation of model to predict COVID-19 with meteorological and air pollutant simultaneously regression fit can be observed in Figure 5.28.

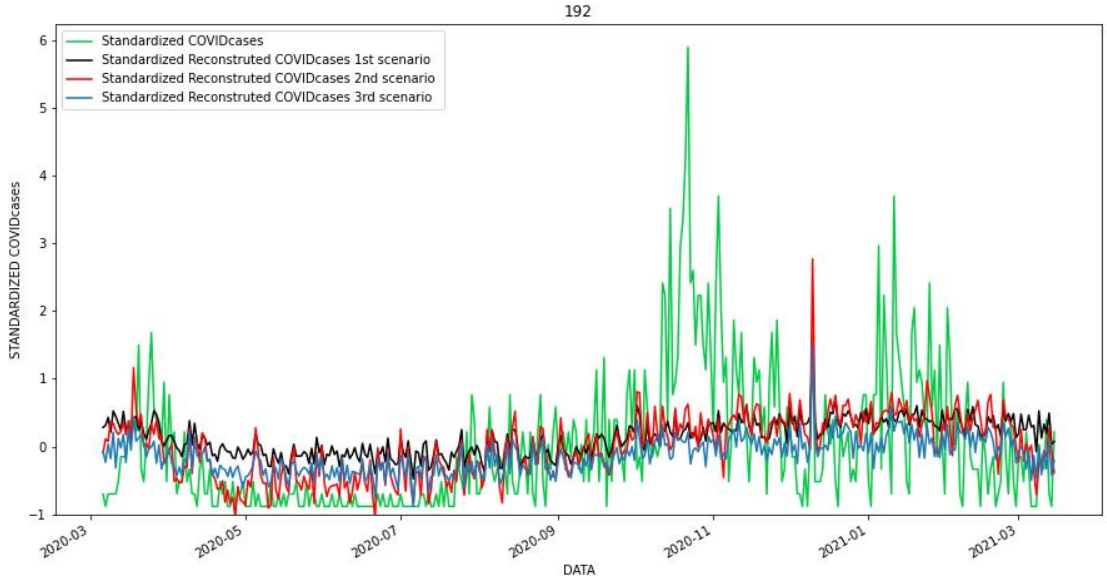

Figure 5.28 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on environmental parameters (meteorological and atmospheric pollutants) parameters (standardized values) for the three scenarios in the ABS "192 – Sabadell-2" during pandemic period.

As Figure above shows, standardized reconstructed COVID-19 cases from analyzed scenarios and original positive cases are quite close along first months in pandemic period. Despite that, the adjustment from October 2020 to March 2021 are considered a bad performance due to the fact that reconstructed COVID-19 cases cannot reach in positive cases uptick when the COVID-19 variants or other influenced factors. It cannot conclude about the best scenario performance over pandemic period; therefore, it will be useful to quantify the goodness of fit.

It is essential to calculate determination coefficient of each scenario in reference to COVID-19 data to end up in the best predicted regression fit. Hence, determination coefficients in reference to ABS 192 – Sabadell-2 are 0.5684, 0.6052 and 0.5769 which are a closed values in comparison with Barcelona-8L ABS. However, it can observe a worst adjustment than lockdown period which is the same extracted conclusion in Barcelona-ABS because of a mobility factor and also, the data are more dispersed.

Moreover, the second scenario is always the best adjustment to predict COVID-19 confirmed cases and it is relevant to mention the high influence taking into account the model with meteorological and air pollutant variables simultaneously and independent.

To end up, the best model to predict COVID-19 through the air vector associated with the meteorological and air pollution in lockdown and pandemic is the second scenario by specific locations. It is relevant to corroborate the fact that the prediction of COVID-19 incidence is better taking into account all atmospheric environmental variables, that is, meteorological and atmospheric pollutant simultaneously.

The lockdown data is adjusted better than in pandemic due to the fact that other factors can be affected, for instance COVID-19 variants, mobility, among other influenced factors during pandemic.

5.3.7 Regression fit based on mobility improved with atmospheric environmental parameters in lockdown period

Mobility is a key factor which will be assess due to the fact that it has a high repercussion in COVID-19 propagation during lockdown and pandemic period according to some scientific publication aforementioned in Section 4.6.4.

Hence, it is aimed to investigate contribution by introducing the effect of mobility data in the model to predict COVID-19 incidence through the air vector associated with the meteorological and air pollution during the lockdown and pandemic.

In order to evaluate the adaptation in the model prediction, a model to predict COVID-19 and mobility regression fit are computed and then, it will be done a model to predict COVID-19 incidence through air vector associated with environmental variables and mobility multilinear regression.

As it mentioned above, this section is evaluated in a unique ABS, it means that, by specific location which is the best case obtained in the lasts sections in order to see the improvement of the adjustment by introducing the effect of mobility data. Thus, the model COVID-19 propagation and atmospheric environmental variables introducing mobility data are done for ABS 403 – Barcelona-8L and ABS 192 – Sabadell-2 separately.

Concerning the Barcelona-8L ABS during lockdown period, standardize regression fit model to predict COVID-19 transmission and mobility factor and reconstructed COVID-19 transmission from COVID-19 based on mobility improved with atmospheric environmental variables with the mobility effect regression fit are compared with real COVID-19 positive cases, hence, it can be seen in Figure 5.29.

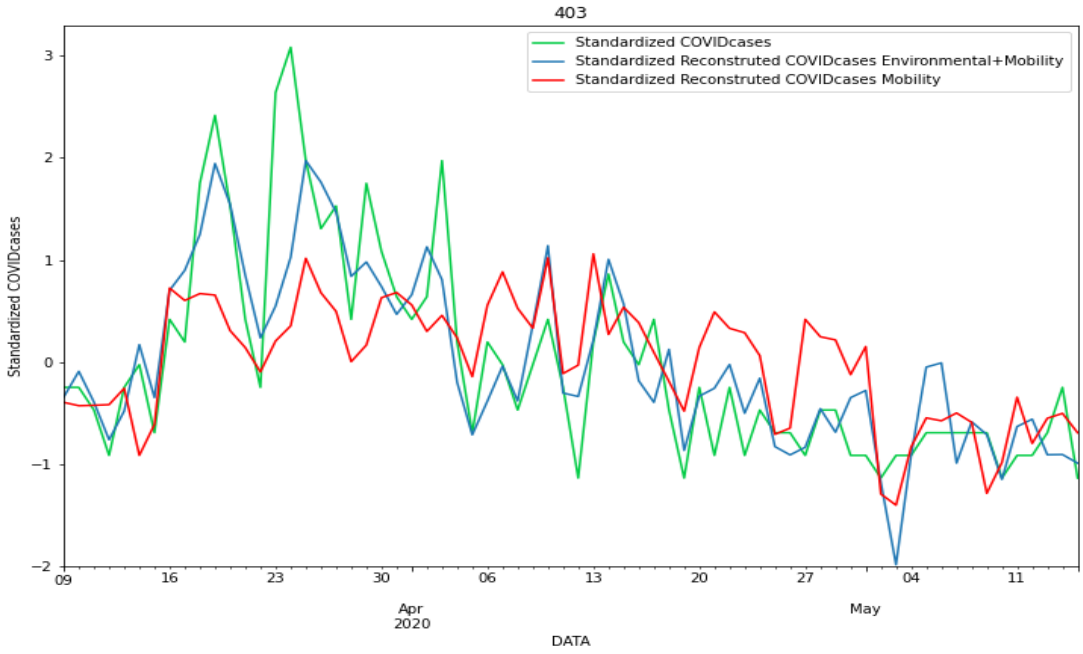

Figure 5.29 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on environmental (meteorological and atmospheric pollutants) parameters (standardized values) for the second scenarios in the ABS "403 – Barcelona-8L" during lockdown period.

As it can appreciate, both reconstructed models to predict COVID-19 incidence has a similar disposition than standardize original cases from *Dades Obertes de Salut.* Hence, COVID-19 spread with the effect of mobility follows the tendency of positive cases, nevertheless, it cannot achieve properly original COVID-19 data curve on some occasion when there are positive cases upticks, for instance these facts are occurred in range from 16th of March to 5th of April 2020. Moreover, it should be noticed that COVID-19 and mobility regression fit model is exceeded in the prediction of real COVID-19 incidence in 9th ,22nd and 29th of April.

Concerning, the standardized reconstructed COVID-19 cases taking into consideration atmospheric environmental and mobility data clearly follow the standardize real COVID-19 cases trend. However, the tendency of pick from 1st to 8th of May would be predicted more accurately taking into account only the mobility variables.

In order to quantify the adjustment between the reconstructed COVID-19 cases, determination coefficient is calculated in both cases taking into consideration original register positive cases.

The determination coefficient in model to predict positive cases considering mobility factor in reference to real confirmed cases during lockdown is 0.6741. Therefore, goodness of fit in model to predict COVID-19 spread with air vector is composed with the meteorological and air pollution with mobility during the lockdown is 0.8482 which is a high adjustment in the prediction of COVID-19 cases.

Thus, the introduction of mobility factor in model to predict COVID-19 cases based on meteorological and atmospheric pollution variables simultaneously has increased the adjustment from 0.6983 to 0.8482 which is a high improvement in the prediction of COVID-19 cases.

With the aim of assessing the mobility data effect, a model to predict COVID-19 based on atmospheric environmental variables with the mobility regression fit and model only taking

into account mobility factor is executed. As a way to visualize the adjustment between reconstructed COVID-19 cases from regression fit models and real COVID-19 data, a Figure 5.30 can be plotted.

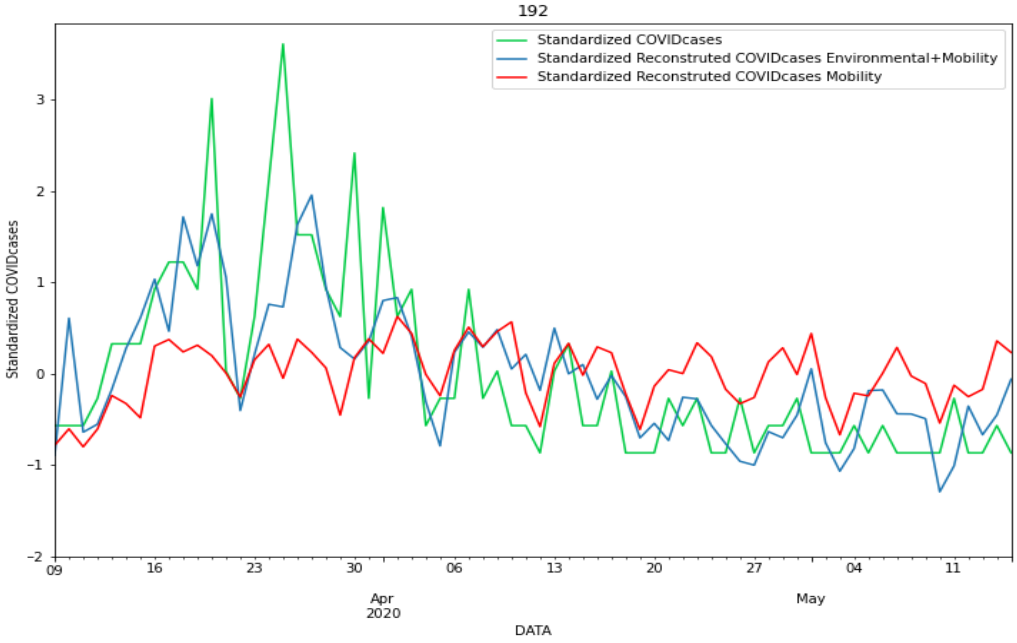

Figure 5.30 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on environmental (meteorological and atmospheric pollutants) parameters (standardized values) for the second scenarios in the ABS "192 – Sabadell-2" during lockdown period.

Figure 5.30 shows the representation of reconstructed COVID-19 cases in regard to original ones in Sabadell-2 ABS which performance as not good as the obtained in Barcelona-8L ABS.

Regarding reconstructed COVID-19 cases with mobility factor, it is not properly predicted the original COVID-19 incidence. In some punctual picks, there are accurate achievement but in overall, it is not predicted better the standardize original COVID-19 cases.

Therefore, common framework of air vector through meteorological and atmospheric pollutants is added in the last-mentioned model giving rise to a high enhancement in the prediction. The first uptick picks in the start lockdown cannot achieve the trend, it can be to the fact that some other factors are contributed in COVID-19 prediction.

With the goal of quantify the goodness of fit of the performance in Figure 5.31, determination coefficient is assessed for ABS 192 – Sabadell-2 in lockdown period. Results from model to predict COVID-19 transmission with mobility effect and model COVID-19 cases through atmospheric environmental variables and mobility contribution are 0.5631 and 0.7664, respectively.

Overall, Table 5.7 shows the goodness of fit of each regression fit evaluated in second scenario (single ABS) to see the high contribution of mobility factor.

Table 5.7 Goodness of fit of each regression fit evaluated in the second scenario by ABS 403 - Barcelona-8L and ABS 192 – Sabadell-2 during lockdown period.

	2nd scenario (ABS 403)	2nd scenario (ABS 192)
Regression fit based on meteorological data	0.5937	0.5689
Regression fit based on air pollutants data	0.6281	0.5929
Regression fit based on atmospheric environmental parameters	0.6983	0.6286
Regression fit based on mobility factor	0.6741	0.5631
Regression fit based on mobility improved with atmospheric environmental parameters	0.8482	0.7664

As aforementioned, Sabadell-2 ABS performance is not good as Barcelona.8L ABS, nevertheless in this specific Sabadell district is obtained an accurate performance in reference to consider mobility factor.

To end up, the introduction of mobility factor in the common framework to predict COVID-19 through the air vector associated with the meteorological and air pollution by localities has produced a huge improvement to predict COVID-19 cases during lockdown.

5.3.8 Regression fit based on mobility improved with atmospheric environmental parameters in pandemic period

A study to evaluate the effect of mobility factor in the model to predict COVID-19 cases through a common framework composed by meteorological and atmospheric pollutant regression fit has been performance by location during pandemic period.

Following the same approach than Section 5.3.7, the model COVID-19 propagation and environmental variables with mobility contribution are assessed in case of ABS 403 – Barcelona-8L and ABS 192 – Sabadell-2.

First, it is relevant to study the mobility effect with COVID-19 propagation during the pandemic in order to see the mobility factor influences, after that, it is relevant the addition of environmental variables in the model COVID-19 transmission and mobility factor.

In order to visualize the mentioned reconstructed regression model fit tendency, Figure 5.31 can be displayed which are composed by original COVID-19 cases, reconstructed confirmed cases from model to predict COVID-19 incidence and mobility factor and with the introduction of meteorological and atmospheric pollutants variables.

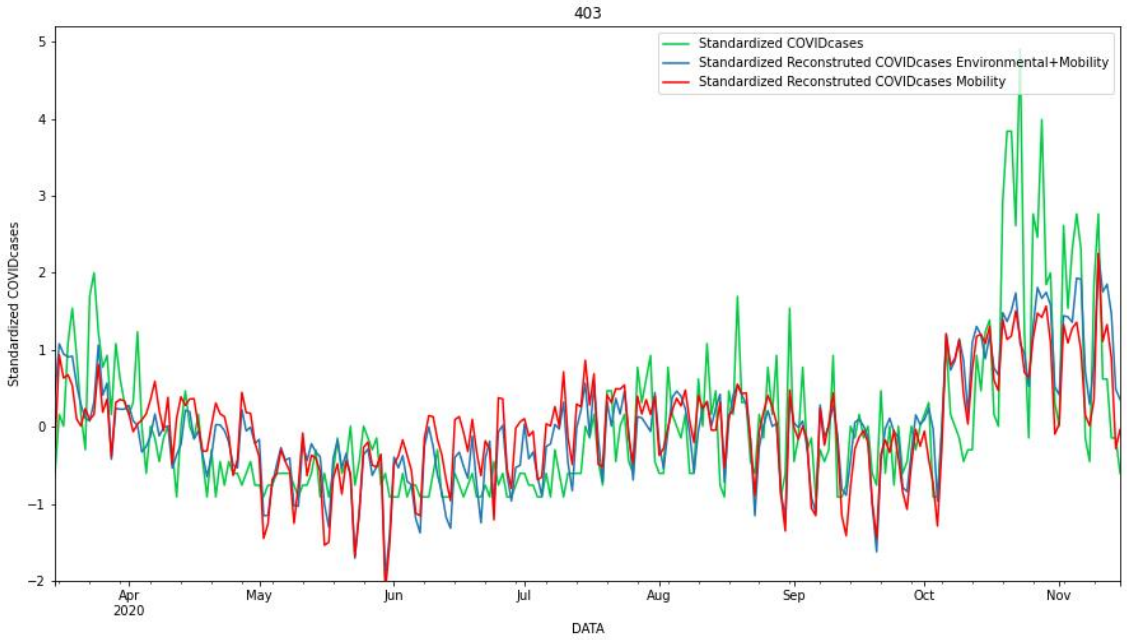

Figure 5.31 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on environmental (meteorological and atmospheric pollutants) parameters (standardized values) with mobility effect for the second scenarios in the ABS "403 – Barcelona-8L" during pandemic period.

Figure 5.31 shows a good adjustment in reconstructed model taking into consideration mobility during pandemic and introducing the environmental variables are improved but not at all, it can be due to the fact a possible limited weight environmental variables in the model.

However, prediction is reached a bad achievement in some uptick COVID-19 cases along pandemic period, these can be influenced for COVID-19 variant detection which there were an augmented COVID-19 confirmed cases.

With the aim to quantify the COVID-19 cases and mobility adjustment for ABS 403 during pandemic period, the goodness of fit is executed and has a value of 0.7324 which indicates a high performance between predicted COVID-19 cases and mobility. The adjustment of ABS 403 during pandemic present a better result than in lockdown (0.6741), it can be due to a large amount of data are done a more realistic and reliable mathematical model. Another reason can be that mobility effect has closed related to pandemic because of mobility are limited by municipalities during lockdown period.

Thus, the determination coefficient in case of model to predict COVID-19 incidence with a common framework between atmospheric environmental and mobility regression fit is 0.7722 which indicate a better adjustment with original COVID-19 cases with the introduction of meteorological and atmospheric pollutants.

In order to see the tendency mentioned in regard with ABS 192 – Sabadell-2, a Figure 5.32 can be displayed between the standardize data of original COVID-19 cases, reconstructed COVID-19 cases from mobility regression fit and reconstructed COVID-19 incidence from the atmospheric environmental and mobility regression fit.

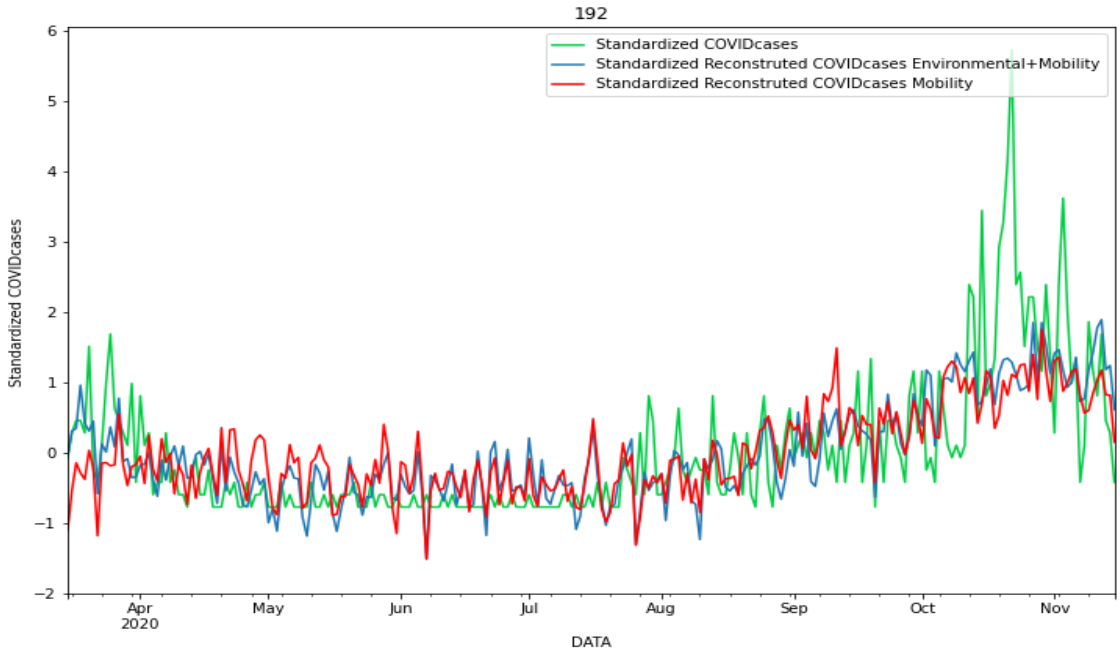

Figure 5.32 Comparison of the real versus reconstructed COVID-19 cases with the regression fit based on environmental (meteorological and atmospheric pollutants) parameters (standardized values) with mobility effect for the second scenarios in the ABS "192 – Sabadell-2" during pandemic period.

In Figure above shows, the adjustment in case of reconstructed cases taking into account atmospheric environmental variables and mobility is slightly better than reconstructed cases with mobility data. In spite of that, it seems that reconstruction of COVID-19 cases from the influence of atmospheric environmental variables and mobility are not properly adapted that in case of lockdown period.

It should be pointed out that a dispersity of COVID-19 data registered and COVID-19 variants which are influenced during pandemic period and for that, the prediction are not proper. This fact, it can be appreciated in Figure 5.32 in period from October to November where a variant is caused positive cases uptick.

Another important fact is the comparison between the evaluated ABS, it can end up in exact conclusion extract in lockdown period, it means that, Sabadell-2 ABS has a low adaptation in COVID-19 prediction.

In order to verify the goodness of fit in Sabadell-2 ABS, a determination coefficient is calculated between the reconstructed COVID-19 cases and mobility with real cases and reconstructed COVID-19 propagation through air vector composed by atmospheric environmental variables and mobility effect in reference original positive cases.

As it can expected, the mobility data has a higher influence in the model to predict COVID-19 incidence composed of meteorological and atmospheric pollutants and mobility factor with a determination coefficient of 0.7409, nevertheless, model COVID-19 and mobility factor has a goodness of fit of 0.6908.

With the aim to see global effect to introduce mobility factor in model regression fit, Table 5.8 shows determination coefficients of specific ABS ("403 – Barcelona-8L" and "192 – Sabadell-2") for each regression fit assessed during pandemic period.

Table 5.8 Goodness of fit of each regression fit evaluated in the second scenario by ABS 403 - Barcelona-8L and ABS 192 – Sabadell-2 during pandemic period.

	2nd scenario (ABS 403)	**2nd scenario (ABS 192)**
Regression fit based on meteorological data	0.5833	0.5748
Regression fit based on air pollutants data	0.5764	0.6038
Regression fit based on atmospheric environmental parameters	0.6085	0.6052
Regression fit based on mobility factor	0.7324	0.6908
Regression fit based on mobility improved with atmospheric environmental parameters	0.7722	0.7409

As table above shows, it is clearly appreciated that Barcelona-8L ABS has a better performance than Sabadell-2 ABS during pandemic period.

Thus, the determination coefficient in pandemic is worse than in lockdown because of the deficit of adjustment in November, it means that, it is clearly appreciated that there are picks which are not possible to create a close adjustment with input data, decreasing the goodness of fit.

To conclude, it is demonstrated that mobility aspect is a key factor in the adjustment of COVID-19 original cases. The goodness of fit is increased in model to predict COVID-19 cases and atmospheric environmental variables regression fit with unique ABS by addition of mobility data. However, the adjustment in reconstructed COVID-19 confirmed cases in lockdown is better than during pandemic due to the fact mobility without municipality restriction and COVID-19 variants detection which are influencing factor in COVID-19 incidence.

6. CONCLUSIONS

The effect of meteorological and air quality data on SARS-CoV-2 transmission has been studied in order to develop a prediction model for the COVID-19 incidence during the lockdown and pandemic periods.

The results demonstrate that certain meteorological and air quality variables affect COVID-19 incidence, and the contribution of this set of variables has been analyzed.

Regarding meteorological variables, temperature, relative humidity and solar radiation are influencing factors in COVID-19 propagation. The correlation is negative for these variables, so an increase of temperature, relative humidity and solar radiation is related to a decrease in the COVID-19 number of new cases. On the contrary, the correlation coefficients for wind velocity and rainfall indicate that these parameters have limited influence on COVID-19, so they can be discarded in the predictive model.

As for the air pollutants, the major pollutants in air, such as NO, NO_2, PM10, PM2.5, CO, SO_2 and O_3, have showed a positive correlation, except for the ozone, on COVID-19 incidence. Nevertheless, it is not possible to clearly appreciate a significant relationship between air pollutants concentration and COVID-19 propagation. This may be due to the fact that the presence of some of them is influenced by other mechanisms different than just emission. It is the case of ozone, whose concentration in the troposphere depends on the presence of natural and polluting chemical precursors and meteorological parameters such as solar radiation.

The correlations of both types of atmospheric environmental variables present a cyclical disposition of seven days, not yet clearly explained by the peak of infection cases that can occurs due to the update and publication of COVID-19 data from *Dades Obertes*. It should be pointed out that the data is published on the web every day, but these are displayed up to 3 days before because of the delay in PCR results, respiratory infection diagnosis and deaths notifications. On the other hand, the correlation obtained during pandemic period is weaker than during the lockdown, as it was expected, due to the weight of the contribution of other factors during the pandemic period, mainly related to mobility, but also with the appearance of different COVID-19 variants, seasonality, among others.

With the correlation analysis performed was not plausible the determination of a specific offset for each atmospheric environmental variable given the lack of patterns in the results. The reason may be found in the different incubation periods of COVID-19 variants during pandemic and the diverse management of the detection and notification of COVID-19 cases within the diverse ABS, among other factors. Thus, an offset of 15 days was selected according to the offset study and the literature review about COVID-19 incubation time and detected variants during the pandemic period.

Concerning the model to predict COVID-19 incidence, it can be concluded that the best scenario to reproduce COVID-19 transmission in the lockdown and pandemic periods is using the regression fit based on the specific *Àrea bàsica de salut* (ABS), so a more accurate predictive model would be derived with regression coefficients obtained at ABS or municipality's level.

A common framework composed by meteorological and atmospheric pollutant regression fit has performed better than taking into account both sets of variables separately.

With regard to the contribution of the mobility factor, it has been demonstrated that the reconstructed COVID-19 cases using the regression fit of atmospheric environmental variables

together with mobility regression clearly follow real COVID-19 cases tendency obtaining a better fit than with mobility parameters alone. Therefore, the implementation of certain atmospheric parameters enhances the prediction and so this type of variables should be considered also in COVID-19 augmented epidemiologic models.

Some next steps to enhance the predicted COVID-19 cases by mathematical modelling will be mentioned. In order to upgrade the model, the correspondence of weather and air quality stations can be specific for each ABS whereas the current model they are still related by municipalities. Another fact is that the current study has been done taking into consideration the number of positive COVID-19 cases, so it will be interesting to introduce and study the severity of the disease particularly for the pollutant concentrations in air and their relation to respiratory diseases than may worsen the COVID-19 evolution in patients. Moreover, an accurate mobility study of each atmospheric environmental variables will be relevant in order to understand better the contribution effect of mobility during lockdown and pandemic periods. Lastly, the implementation of the approach in other regions with different meteorological and air quality conditions is needed to check the validity of the model. For example, precipitation may have some effect in COVID-19 transmission that may have been hidden in the present study given the low levels of precipitation in the case study of Catalonia.

7. REFERENCES

1. **Àrees bàsiques de salut.** [Online] Instituto de Estadística de Cataluña (idescat), 25 de February de 2021. https://www.idescat.cat/codis/?id=50&n=39&lang=es.

2. **Jason Sam Leo Lorenzo, Wilson Wai San Tam, Wei Jie Seow.** Association between air quality, meteorological factors and COVID-19 infection case numbers. 2021.

3. **Channa Zhao Xinyu Fang, Yating Feng, Xuehui Fang Jun He, Haifeng Pan.** Emerging role of air pollution and meteorological parameters in COVID-19. 2021.

4. **Huiying Huang, Xiuji Liang, Jingxiu Huang, Zhaohu Yuan, Handong Ouyang, Yaming Wei, Xiaohui Bai.** Correlations between Meterological indicators, air quality and the COVID-19 pandemic in 12 cities across China. 2020.

5. **Maria A. Zoran, Roxaana S. Savastru, Marina N. Tautan, Laurentiu A. Baschir.** Assesing the impact of air pollution and climate seasonality on COVID-19 multiwaves in Madrid, Spain . 2021.

6. **Montse Marquès, José L. Domingo.** Positive association between outdoor air pollution and the incidence and severity of COVID-19. A review of the recent scientific evidences . 2021.

7. **Eric B. Brandt, PhD, Andrew F. Beck, MD, MPH and Tesfaye B. Mersha, PhD.** Air pollution, racial disparities, and COVID-19 mortality. 2020.

8. **Dades meteorològiques de la XEMA.** [Online] Servei Meteorològic de Catalunya, March 18, 2019. https://analisi.transparenciacatalunya.cat/Medi-Ambient/Dades-meteorol-giques-de-la-XEMA/nzvn-apee.

9. **Metadades estacions meteorològiques automàtiques. [Online]** Servei Meteorològic de Catalunya, March 21, 2019. https://analisi.transparenciacatalunya.cat/Medi-Ambient/Metadades-estacions-meteorol-giques-autom-tiques/yqwd-vj5e.

10. **Metadades variables meteorològiques. [Online]** Servei Meteorològic de Catalunya, March 21, 2019. https://analisi.transparenciacatalunya.cat/Medi-Ambient/Metadades-variables-meteorol-giques/4fb2-n3yi.

11. **Registre de casos de COVID-19 a Catalunya per municipi i sexe. [Online]** Departament de Salut, March 2020, 2020. https://analisi.transparenciacatalunya.cat/Salut/Registre-de-casos-de-COVID-19-a-Catalunya-per-muni/jj6z-iyrp.

12. **Registre de casos de COVID-19 a Catalunya per àrea bàsica de salut (ABS) i sexe. [Online]** Departament de Salut, 31 de March de 2020. https://analisi.transparenciacatalunya.cat/Salut/Registre-de-casos-de-COVID-19-a-Catalunya-per-rea-/xuwf-dxjd.

13. **Qualitat de l'aire als punts de mesurament manuals de la Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica.** [Online] Departament d'Acció Climàtica, Alimentació i Agenda Rural, October 15, 2020. https://analisi.transparenciacatalunya.cat/Medi-Ambient/Qualitat-de-l-aire-als-punts-de-mesurament-manuals/qg74-87s9.

14. **Qualitat de l'aire als punts de mesurament automàtics de la Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica.** [Online] Departament d'Acció Climàtica, Alimentació i Agenda Rural, 18 de March de 2020.

https://analisi.transparenciacatalunya.cat/Medi-Ambient/Qualitat-de-l-aire-als-punts-de-mesurament-autom-t/tasf-thgu.

15. **Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica (XVPCA).** [Online] Medi Ambient i Sostenibilitat. https://mediambient.gencat.cat/ca/05_ambits_dactuacio/atmosfera/qualitat_de_laire/avaluacio/xarxa_de_vigilancia_i_previsio_de_la_contaminacio_atmosferica_xvpca/.

16. **Punts de mesurament i equipament de la Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica.** [Online] https://mediambient.gencat.cat/web/.content/home/ambits_dactuacio/atmosfera/qualitat_de_laire/avaluacio/xarxa_de_vigilancia_i_previsio_de_la_contaminacio_atmosferica__xvpca/Equipament.pdf

17. **Àrees bàsiques de salut.** [Online] Instituto de Estadística de Cataluña (idescat), 25 de February de 2021. https://www.idescat.cat/codis/?id=50&n=39&lang=es.

18. **Distritos de Barcelona.** [Online] Wikipedia. https://es.wikipedia.org/wiki/Distritos_de_Barcelona.

19. **Centres d'Atenció Primària i Comunitària.** [Online] Ajuntament de Sabadell, 2017. https://sabadell.cat/ca/serveis-sanitaris/centres-d-atencio-primaria-i-comunitaria.

20. **Michael Marks, PhD, Pere Millat-Martinez, MD, Dan Ouchi, MSc, Chrissy h Roberts, PhD, Andrea Alemany, BM, Marc Corbacho-Monné.** Transmission of COVID-19 in 282 clusters in Catalonia, Spain: a cohort study. 2021.

21. **Nazar Zakia, and Elfadil A. Mohamed.** The estimations of the COVID-19 incubation period: A scoping reviews of the literature. **2021.**

22. **Yuanyuan Han, Mao Jiang, Da Xia, Lichao, Xin Lv, Xiaohua Liao and Jie Menga.** COVID-19 in a patient with long-term use of glucocorticoids: A study of a familial cluster. 2020.

23. **Wafa Dhouib, Jihen Maatoug, Imen Ayouni, Nawel Zammit, Rim Ghammem, Sihem Ben Fredj and Hassen Ghannem.** The incubation period during the pandemic of COVID-19: a systematic review and meta-analysis. 2021.

8. APPENDICES

8.1 Data extraction automation codes

8.1.1 Meteorological web scraping code

```
####################METEOROLOGICAL DATA####################
import pandas as pd
from sodapy import Socrata

client = Socrata("analisi.transparenciacatalunya.cat", None)
    ###### TIME FILTER: Lockdown period (01.03.20 - 15.05.20) and pandemic period
      # (01.03.20 - 15.03.2021)
###Lockdown period
print('    Meteo data in lockdown period with the desirable variables    ')
# Wind speed
meteoLD_WS_time = client.get("nzvn-apee", where='data_lectura >= "2020-02-01T00:00:00.000" AND data_lectura <= "2020-05-15T00:00:00.000" AND codi_variable ="30"', limit=500000)
# Temperature
meteoLD_T_time = client.get("nzvn-apee", where='data_lectura >= "2020-02-01T00:00:00.000" AND data_lectura <= "2020-05-15T00:00:00.000" AND codi_variable ="32"', limit=5000000)
# Relative humidity
meteoLD_HR_time = client.get("nzvn-apee", where='data_lectura >= "2020-02-01T00:00:00.000" AND data_lectura <= "2020-05-15T00:00:00.000" AND codi_variable ="33"', limit=5000000)
# Precipitation
meteoLD_P_time = client.get("nzvn-apee", where='data_lectura >= "2020-02-01T00:00:00.000" AND data_lectura <= "2020-05-15T00:00:00.000" AND codi_variable ="35"', limit=5000000)
# Solar Radiation
meteoLD_SR_time = client.get("nzvn-apee", where='data_lectura >= "2020-02-01T00:00:00.000" AND data_lectura <= "2020-05-15T00:00:00.000" AND codi_variable ="36"', limit=5000000)
meteoLD_time = meteoLD_WS_time + meteoLD_T_time + meteoLD_HR_time + meteoLD_P_time + meteoLD_SR_time
meteos = pd.DataFrame.from_records(meteoLD_time)
print(meteos)
```

```
print(len(meteos))
##Pandemic period
print('    Meteo data not in pandemic period with the desirable variables    ')
# Temperature
meteo_T_time = client.get("nzvn-apee", where='data_lectura >= "2020-02-01T00:00:00.000" AND data_lectura <= "2021-03-15T00:00:00.000" AND codi_variable ="32"', limit=5000000)
# Relative humidity
meteo_HR_time = client.get("nzvn-apee", where='data_lectura >= "2020-02-01T00:00:00.000" AND data_lectura <= "2021-03-15T00:00:00.000" AND codi_variable ="33"', limit=5000000)
# Solar Radiation
meteo_SR_time = client.get("nzvn-apee", where='data_lectura >= "2020-02-01T00:00:00.000" AND data_lectura <= "2021-03-15T00:00:00.000" AND codi_variable ="36"', limit=5000000)
meteo_time = meteo_T_time + meteo_HR_time + meteo_SR_time
meteos = pd.DataFrame.from_records(meteo_time)
print(meteos)
print(len(meteos))
```

8.1.2 Epidemiological web scraping code

```
#################EPIDEMIOLOGICAL DATA####################
import pandas as pd
from sodapy import Socrata

client = Socrata("analisi.transparenciacatalunya.cat", None)
   ###### TIME FILTER: Lockdown period (01.03.20 - 15.05.20) and pandemic period
      # (01.03.20 - 15.03.2021)
##Lockdown period
print('    Clinics data in lockdown period    ')
clinicsLD_time = client.get("xuwf-dxjd", where='data >= "2020-02-01T00:00:00.000" AND data <= "2020-05-15T00:00:00.000"', limit=50000)
clinics = pd.DataFrame.from_records(clinicsLD_time)
print(clinics)
print(len(clinics))

##Pandemic period
```

```
print('     Clinics data in pandemic period     ')
clinics_time = client.get("xuwf-dxjd", where='data >= "2020-02-01T00:00:00.000" AND data <= "2021-03-15T00:00:00.000"', limit=5000000)
clinics = pd.DataFrame.from_records(clinics_time)
print(clinics)
print(len(clinics))
```

8.1.3 Manual measurement of air pollutants web scrapping code

```
#####MANUAL MEASURAMENTS IN AIR POLLUTANTS #####
import pandas as pd
from sodapy import Socrata

client = Socrata("analisi.transparenciacatalunya.cat", None)
  ###### TIME FILTER: Lockdown period (01.03.20 - 15.05.20) and pandemic period
      # (01.03.20 - 15.03.2021)
###Lockdown period
print('     MANUAL Contaminant data in lockdown period     ')
mancontLD_time = client.get("qg74-87s9", where='ano = "2020" AND mes >= "2" AND mes <= "6"', limit=5000)
cont_manual = pd.DataFrame.from_records(mancontLD_time)
print(cont_manual)
print(len(cont_manual))
###Pandemic period, so limit=50000
print('     MANUAL Contaminant data in pandemic period     ')
mancont1_PM10 = client.get("qg74-87s9",where='ano = "2020" AND mes >= "2" AND mes <= "12" AND nom_contaminant="PM10"', limit=500000)
mancont2_PM10 = client.get("qg74-87s9",where='ano = "2021" AND mes >= "1" AND mes <= "3" AND nom_contaminant="PM10"', limit=500000)
mancont1_PM25 = client.get("qg74-87s9",where='ano = "2020" AND mes >= "2" AND mes <= "12" AND nom_contaminant="PM2.5"', limit=500000)
mancont2_PM25 = client.get("qg74-87s9",where='ano = "2021" AND mes >= "1" AND mes <= "3" AND nom_contaminant="PM2.5"', limit=500000)
mancont_time=mancont1_PM10 +mancont2_PM10 +mancont1_PM25+mancont2_PM25
cont_manual = pd.DataFrame.from_records(mancont_time)
print(cont_manual)
print(len(cont_manual))
```

8.1.4 Automatic measurements of air pollutants web scrapping code

```
#####AUTOMATIC MEASURAMENTS IN AIR POLLUTANTS#######
import pandas as pd
from sodapy import Socrata

client = Socrata("analisi.transparenciacatalunya.cat", None)
   ###### TIME FILTER: Lockdown period (01.03.20 - 15.05.20) and pandemic period
       # (01.03.20 - 15.03.2021)
###Lockdown period
print('     AUTO Contaminant data in lockdown period     ')
autcontLD_time = client.get("tasf-thgu",where='data >= "2020-02-01T00:00:00.000" AND data <= "2020-05-15T00:00:00.000"',limit=50000)
cont_auto = pd.DataFrame.from_records(autcontLD_time)
print(cont_auto)
print(len(cont_auto))
###Pandemic period, so limit=500000
print('     AUTO Contaminant in pandemic period     ')
autcont_O3 = client.get("tasf-thgu",where='data >= "2020-02-01T00:00:00.000" AND data <= "2021-03-15T00:00:00.000" AND contaminant ="O3"',limit=500000)
autcont_NO = client.get("tasf-thgu",where='data >= "2020-02-01T00:00:00.000" AND data <= "2021-03-15T00:00:00.000" AND contaminant ="NO"',limit=500000)
autcont_NOX = client.get("tasf-thgu",where='data >= "2020-02-01T00:00:00.000" AND data <= "2021-03-15T00:00:00.000" AND contaminant ="NOX"',limit=500000)
autcont_NO2 = client.get("tasf-thgu",where='data >= "2020-02-01T00:00:00.000" AND data <= "2021-03-15T00:00:00.000" AND contaminant ="NO2"',limit=500000)
autcont_PM10 = client.get("tasf-thgu",where='data >= "2020-02-01T00:00:00.000" AND data <= "2021-03-15T00:00:00.000" AND contaminant ="PM10"',limit=500000)
autcont_PM25 = client.get("tasf-thgu",where='data >= "2020-02-01T00:00:00.000" AND data <= "2021-03-15T00:00:00.000" AND contaminant ="PM2.5"',limit=500000)
autcont_CO = client.get("tasf-thgu",where='data >= "2020-02-01T00:00:00.000" AND data <= "2021-03-15T00:00:00.000" AND contaminant ="CO"',limit=500000)
autcont_SO2 = client.get("tasf-thgu",where='data >= "2020-02-01T00:00:00.000" AND data <= "2021-03-15T00:00:00.000" AND contaminant ="SO2"',limit=500000)
autcont_time = autcont_O3+autcont_NO+autcont_NOX+autcont_NO2+autcont_PM10+autcont_PM25+autcont_CO+autcont_SO2
cont_auto  = pd.DataFrame.from_records(autcont_time)
```

```
print(cont_auto )
print(len(cont_auto ))
```

8.2 Data processing codes

8.2.1 Meteorological data processing code

```
import pandas as pd
import numpy as np
import datetime as datetime

##### Input data: weather station and ABS
locations = pd.read_excel('Data/Location_ABS_Meteo_stations.xlsx')
ABS_code = locations['CodiABS']
estacio = locations['Estacio']
ABS_name = locations['nomABS']
#### Meteorological data from webscraping
# Select in webscraping meteo the desireable option: Lockdown or Pandemic
from webscrapping_METEO import *
   # Change the format data from webscraping
meteos[['codi_variable', 'valor_lectura']] = meteos[['codi_variable', 'valor_lectura']].apply(pd.to_numeric)
meteos["data_lectura"] = pd.to_datetime(meteos["data_lectura"], dayfirst=True)
meteos_data = pd.to_datetime(meteos['data_lectura'], format='%D:%M:%Y').dt.date
meteos["data_lectura"] = meteos_data
meteos["data_lectura"] = pd.to_datetime(meteos["data_lectura"], dayfirst=True)
dummy = list(meteos.columns)
dummy.append('ABS_Code')
dummy.append('ABS')
meteo_final = pd.DataFrame(columns=dummy)
for m in range(len(ABS_name)):
    print (ABS_name[m], estacio[m])
   meteo_ABS = meteos[meteos['codi_estacio'] == estacio[m]]
   meteo_ABS['ABS_Code'] = ABS_code[m]
   meteo_ABS['ABS'] = ABS_name[m]
   meteo_final = meteo_final.append(meteo_ABS)
```

```
## Output data: meteorological data related to ABS
 meteo_final.to_csv('Meteo_ABS_webscrapping_v1.csv')
#meteo_final.to_csv('Meteo_ABS_webscrapping_PANDEMIC.csv')
```

8.2.2 Air pollutants data processing code

```
import pandas as pd
import numpy as np
import datetime as datetime


##### Input data: weather station and ABS
locations = pd.read_excel('Data/Location_ABS_Contaminants_stations.xlsx')
INE_codi = locations['MunicipiCodi']
ABS_codi = locations['CodiABS']
ABS_name = locations['nomABS']
#########Automatic measuraments of air pollutants
print ('==== AUTOMATIC MEASURAMENTS DATA OF AIR POLLUTANTS ======')
# Select in webscraping cont auto the desireable option: Lockdown or Pandemic
from webscrapping_CONT_AUT import *
   # Change the format data from webscraping
cont_auto[['codi_eoi', 'magnitud', 'codi_ine', 'codi_comarca', 'altitud', 'latitud', 'longitud', 'h01', 'h02', 'h03', 'h04', 'h05', 'h06', 'h07', 'h08', 'h09', 'h10',  'h11', 'h12', 'h13', 'h14', 'h15', 'h16', 'h17', 'h18',  'h19',  'h20',  'h21',  'h22',  'h23',  'h24']]=cont_auto[[ 'codi_eoi',  'magnitud',  'codi_ine', 'codi_comarca', 'altitud', 'latitud', 'longitud','h01', 'h02', 'h03', 'h04', 'h05', 'h06', 'h07', 'h08', 'h09', 'h10',   'h11',  'h12',  'h13',  'h14',  'h15',  'h16',  'h17',  'h18',  'h19',  'h20',  'h21',  'h22',  'h23', 'h24']].apply(pd.to_numeric)
cont_auto["data"] = pd.to_datetime(cont_auto["data"], dayfirst=True)
# Daily values and merge stations from the same ABS
cont_auto ['VALOR'] = cont_auto[['h01', 'h02', 'h03', 'h04', 'h05', 'h06', 'h07', 'h08', 'h09', 'h10', 'h11','h12',  'h13',  'h14',  'h15',  'h16',  'h17',  'h18',  'h19',  'h20',  'h21',  'h22',  'h23', 'h24']].mean(axis=1)
cont_auto = cont_auto.drop(['h01', 'h02', 'h03', 'h04', 'h05', 'h06', 'h07', 'h08', 'h09', 'h10',  'h11', 'h12', 'h13', 'h14', 'h15', 'h16', 'h17', 'h18', 'h19', 'h20', 'h21', 'h22', 'h23', 'h24'], axis =1)
dummyFinal = pd.DataFrame(columns=cont_auto.columns)
for m in range(len(INE_codi)):
   print ('ABS =====> ',ABS_name[m])
   cont_ABS = cont_auto[cont_auto["codi_ine"] == INE_codi.iloc[m]]
   cont_ABS ['ABS_Code'] = ABS_codi[m]
```

```
    cont_ABS['ABS'] = ABS_name[m]
     if len(set(cont_ABS['nom_estacio'])) == 1:
        dummyFinal = dummyFinal.append(cont_ABS)
    if len(set(cont_ABS['nom_estacio'])) > 1:
        estaciones = list(set(cont_ABS['nom_estacio']))
        dates_ABS = list(set(cont_ABS['data']))
        variables_ABS = list(set(cont_ABS['contaminant']))
        for v in range(len(variables_ABS)):
            dummy1 = cont_ABS[cont_ABS["contaminant"] == variables_ABS[v]]
            for d in range(len(dates_ABS)):
                dummy2 = dummy1[dummy1["data"] == dates_ABS[d]]
                dummy3 = dummy2['VALOR']
                dummy4 = dummy3.mean(axis=0)
                dummy5 = dummy2.iloc[0]
                dummy5['nom_estacio'] = dummy5['municipi']+'_Mean'
                dummy5[['VALOR']] = dummy4
                dummyFinal = dummyFinal.append(dummy5)
cont_auto_2 = dummyFinal[dummyFinal.VALOR.notnull()]
cont_auto_2.index = cont_auto_2['data']
## Output data: automatic measuraments data related to ABS
cont_auto_2.to_excel('Contaminacion_atmos_2020_2021_auto_ord.xlsx')
#cont_auto_2.to_excel('Contaminacion_atmos_2020_2021_auto_ord_PANDEMIC.xlsx')
###########Manual measuraments of air pollutants
print('====== MANUAL MEASURAMENTS DATA OF AIR POLLUTANTS ======')
# Select in webscraping cont auto the desireable option: Lockdown or Pandemic
from webscrapping_CONT_MAN import *
    # Sort webscraping data (daily data organized)
cont_manual = cont_manual [['codi_eoi', 'nom_estacio', 'ano', 'mes','magnitud',
'nom_contaminant','unitats', 'tipus_estacio', 'codi_ine', 'nom_municipi', 'd01', 'd02', 'd03', 'd04',
'd05', 'd06', 'd07', 'd08', 'd09', 'd10', 'd11', 'd12', 'd13', 'd14', 'd15', 'd16', 'd17', 'd18', 'd19', 'd20',
'd21', 'd22', 'd23', 'd24', 'd25', 'd26', 'd27', 'd28', 'd29', 'd30', 'd31', 'altitud', 'latitud', 'longitud',]]
    # Change the format data from webscraping
cont_manual[['codi_eoi', 'ano', 'mes','magnitud','codi_ine', 'altitud', 'latitud', 'longitud', 'd01',
'd02', 'd03', 'd04', 'd05', 'd06', 'd07', 'd08', 'd09', 'd10', 'd11', 'd12', 'd13', 'd14', 'd15', 'd16', 'd17',
'd18', 'd19', 'd20', 'd21', 'd22', 'd23', 'd24', 'd25', 'd26', 'd27', 'd28', 'd29', 'd30', 'd31']] =
cont_manual[['codi_eoi', 'ano', 'mes','magnitud','codi_ine', 'altitud', 'latitud', 'longitud', 'd01',
```

```
'd02', 'd03', 'd04', 'd05', 'd06', 'd07', 'd08', 'd09', 'd10', 'd11', 'd12', 'd13', 'd14', 'd15', 'd16', 'd17', 'd18', 'd19', 'd20', 'd21', 'd22', 'd23', 'd24', 'd25', 'd26', 'd27', 'd28', 'd29', 'd30', 'd31']].apply(pd.to_numeric)
Codi_EOI = ([])
Nombre = ([])
nombre_contaminante = ([])
Contaminante = ([])
Codi_INE = ([])
Municipio = ([])
Dates = ([])
contaminacion = ([])
for i in range(len(cont_manual)):
    Y = cont_manual['ano'].iloc[i]
    M = cont_manual['mes'].iloc[i]
    for j in range(31):
        try:
            dummy = datetime.date(Y, M, int(j+1))
            cont_dummy = cont_manual['d'+str(j+1).zfill(2)].iloc[i]
            if not np.isnan(cont_dummy):
                Dates.append(dummy)
                contaminacion.append(cont_dummy)
                Codi_EOI.append(cont_manual['codi_eoi'].iloc[i])
                Nombre.append(cont_manual['nom_estacio'].iloc[i])
                nombre_contaminante.append(cont_manual['nom_contaminant'].iloc[i])
                Contaminante.append(cont_manual['nom_contaminant'].iloc[i])
                Codi_INE.append(cont_manual['codi_ine'].iloc[i])
                Municipio.append(cont_manual['nom_municipi'].iloc[i])
        except:
            pass
cont_manual_2 = pd.DataFrame(index=Dates)
cont_manual_2['CODI EOI'] = Codi_EOI
cont_manual_2['NOM ESTACIO'] = Nombre
cont_manual_2['CONTAMINANT'] = nombre_contaminante
cont_manual_2['VALOR'] = contaminacion
```

```
cont_manual_2['CODI INE'] = Codi_INE
cont_manual_2['MUNICIPI'] = Municipio
cont_manual_2['DATA'] = cont_manual_2.index
cont_manual_2.to_excel('Contaminacion_Atmos_Manual_2020_2021_ORDENADO_v1.xls')
#cont_manual_2.to_excel('Contaminacion_Atmos_Manual_2020_2021_ORDENADO_PAN
DEMIC.xls')
   # Merge data when there is more than one station per ABS
dummyFinal = pd.DataFrame(columns=cont_manual_2.columns)
for m in range(len(INE_codi)):
   print('ABS =====> ', ABS_name[m])
   cont_ABS = cont_manual_2[cont_manual_2["CODI INE"] == INE_codi.iloc[m]]
   cont_ABS ['ABS_Code'] = ABS_codi[m]
   cont_ABS['ABS'] = ABS_name[m]
   if len(set(cont_ABS['NOM ESTACIO'])) == 1:
      dummyFinal = dummyFinal.append(cont_ABS)
   if len(set(cont_ABS['NOM ESTACIO'])) > 1:
      estaciones = list(set(cont_ABS['NOM ESTACIO']))
      dates_ABS = list(set(cont_ABS['DATA']))
      variables_ABS = list(set(cont_ABS['CONTAMINANT']))
      for v in range(len(variables_ABS)):
         dummy1 = cont_ABS[cont_ABS["CONTAMINANT"] == variables_ABS[v]]
         for d in range(len(dates_ABS)):
            dummy2 = dummy1[dummy1["DATA"] == dates_ABS[d]]
            if len(dummy2) > 0:
               dummy3 = dummy2['VALOR']
               dummy4 = dummy3.mean(axis=0)
               dummy5 = dummy2.iloc[0]
               dummy5['NOM ESTACIO'] = dummy5['MUNICIPI']+'_Mean'
               dummy5[['VALOR']] = dummy4
               dummyFinal = dummyFinal.append(dummy5)
cont_manual_3 = dummyFinal[dummyFinal.VALOR.notnull()]
cont_manual_3["DATA"] = pd.to_datetime(cont_manual_3["DATA"], dayfirst=True)
## Output data: automatic measuraments data related to ABS
#Lockdown
```

```
cont_manual_4=cont_manual_3.query("DATA >= '2020-02-01 00:00:00' and DATA <='2020-05-15 00:00:00'")
cont_manual_4.to_excel('Contaminacion_atmos_2020_2021_manual_ord_v1.xlsx')
#Pandemic
#cont_manual_4=cont_manual_3.query("DATA >= '2020-02-01 00:00:00' and DATA <='2021-03-15 00:00:00'")
#cont_manual_4.to_excel('Contaminacion_atmos_2020_2021_manual_ord_PANDEMIC.xlsx')
```

8.2.3 Air pollutants data merge code

```
import pandas as pd
from datetime import datetime
##### Input data: Files from Datapreprocesing code
# Select the data for Lockdown or Pandemic in automatic measuraments of air pollutants
cont_auto_2 = pd.read_excel('Contaminacion_atmos_2020_2021_auto_ord_v1.xlsx')
#cont_auto_2=pd.read_excel('Contaminacion_atmos_2020_2021_auto_ord_PANDEMIC.xlsx')
cont_auto_2.index = cont_auto_2['data']
# Select the data for Lockdown or Pandemic in manual measuraments of air pollutants
cont_manual_4 = pd.read_excel('Contaminacion_atmos_2020_2021_manual_ord_v1.xlsx')
#cont_manual_4=pd.read_excel('Contaminacion_atmos_2020_2021_manual_ord_PANDEMIC.xlsx')
cont_manual_4.index = cont_manual_4['DATA']
# Automatic and manual measuraments of air pollutants merge
DF_Contaminantes_merge=pd.DataFrame(columns('VALOR_merge','CONTAMINANT_merge','CODI_ABS_merge','ABS_merge'))
#List of air pollutants and ABS code
contaminantres=list(set(cont_auto_2['contaminant'].append(cont_manual_4['CONTAMINANT'])))
ABS_codi = list(set(cont_auto_2['ABS_Code'].append(cont_manual_4['ABS_Code'])))
for a in range(len(ABS_codi)):
    print (ABS_codi[a])
    for c in range(len(contaminantes))
        print (contaminantes[c])
        dummy_auto=cont_auto_2[(cont_auto_2['contaminant']==contaminantes[c])&(cont_auto_2['ABS_Code'] == ABS_codi[a]) ]
```

```
        dummy_manual=cont_manual_4[(cont_manual_4['CONTAMINANT']==contaminante
        s[c]) & ( cont_manual_4['ABS_Code'] == ABS_codi[a]) ]
      if len(dummy_auto) > 0 or len(dummy_manual) > 0:
        dummy = pd.concat([dummy_auto, dummy_manual], axis=1)
        valor_merge= dummy['VALOR'].mean(axis=1)
        dummy['VALOR_merge'] =  valor_merge
        dummy['CONTAMINANT_merge'] =  contaminantes[c]
        dummy['CODI_ABS_merge'] =  ABS_codi[a]
        dummy.iloc[:,18] = dummy['ABS'].iloc[:,0].fillna(dummy['ABS'].iloc[:,1])
        dummy['ABS_merge'] = dummy['ABS'].iloc[0,0]
DF_Contaminantes_merge=DF_Contaminantes_merge.append(dummy[['VALOR_merge','C
ONTAMINANT_merge','CODI_ABS_merge','ABS_merge']] )
## Output data: Automatic and manual air pollutants merge and related to ABS
DF_Contaminantes_merge.to_excel('Contaminacion_atmos_2020_2021_MERGED_v1.xlsx')
#DF_Contaminantes_merge.to_excel('Contaminacion_atmos_2020_2021_MERGED_PAND
EMIC.xlsx')
```

8.3 Correlation model codes

8.3.1 Meteorological correlation code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import datetime as datetime
##Input data: parametres and variables definition
f=1
f2=1000
f3=2000
#Analyse periods
#Lockdown
stt = '01-03-2020'
ent = '15-05-2020'
#Pandemic
#stt = '01-03-2020'
#ent = '15-03-2021'
#OFFSETS
```

```
#of = [0,1,2,3,4, 5,6, 7,8,9,10,11,12,13,14, 15]
of = [0,1,2,3,4, 5,6, 7,8,9,10,11,12,13,14, 15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30]
#Create an Excel file to store the data for a selected offset
offset_save =30
file_offsets_out = 'Timeseries_Pand_Offset_Meteo_'+str(offset_save)+'_out.xls'
writer = pd.ExcelWriter(file_offsets_out, engine = 'xlsxwriter')
###### Input data: weather stations, municipalities and ABS
locations = pd.read_excel('Data/Location_ABS_Meteo_stations.xlsx')
INE_codi = locations['MunicipiCodi']
munic = locations['MunicipiDescripcio']
ABS_codi = locations['CodiABS']
ABS_name = locations['nomABS']
estacio = locations['Estacio']
###### Input data: epidemiological webscraping (select lockdown or pandemic)
print ('========== EPIDEMIOLOGICAL DATA ==================')
# Select in webscraping clinics ABS the desireable option: Lockdown or Pandemic
from webscrapping_CLINICAL_ABS import *
# Modify the format in web scraping
clinics[['regiosanitariacodi', 'sectorsanitaricodi', 'abscodi', 'sexecodi', 'numcasos']] = clinics[['regiosanitariacodi','sectorsanitaricodi', 'abscodi', 'sexecodi', 'numcasos']].apply(pd.to_numeric)
clinics["data"] = pd.to_datetime(clinics["data"], dayfirst=True)
###### Input data: meteorological data (select lockdown or pandemic file)
df_meteorologicos = pd.read_csv('Data/Meteo_ABS_webscrapping_v1.csv')
df_meteorologicos = pd.read_csv('Data/Meteo_ABS_webscrapping_PANDEMIC.csv')
df_meteorologicos['DATA'] = pd.to_datetime(df_meteorologicos['data_lectura'], dayfirst=True)
#######Meteorological variables to take into account
var_cont = [30,32,33,35,36]
var_cont_label = ('VV10','T','HR',‘PPT','RS')
colors= plt.cm.tab20(np.linspace(0,1,len(var_cont)))
######## Dataframes to store final results
pcV = pd.DataFrame(columns=['CODI_INE','Location','CODI_ABS','ABS_NAME', 'Offset','N_32','32','N_33','33','N_36','36'])
```

```
pcM = pd.DataFrame(columns=['CODI_INE','Location','CODI_ABS','ABS_NAME',
'Offset','N_32','32','N_33','33','N_36','36'])
pcV['Offset'] = of
###Procesing code
for a in range(len(ABS_codi)):
    flag = False
    flag2 = True
    print ('ABS =====> ',ABS_name[a])
    pcV['Offset'] = of
    pcV['CODI_INE'] = INE_codi[a]
    pcV['Location'] = munic[a]
    pcV['CODI_ABS'] = ABS_codi[a]
    pcV['ABS_NAME'] = ABS_name[a]
    # Extract epidemiological data based on ABS and dates
    allcasos = clinics[clinics["abscodi"] ==ABS_codi[a]]  # datos para una ABS
    # Data from desireable period
    mask = (allcasos['data'] > stt) & (allcasos['data'] < ent)
    casosLocDates = allcasos.loc[mask]
    casosLocDates = casosLocDates.sort_values(by="data")
    if casosLocDates.empty:
        print('No cases between dates ')
        # pcV[:] = '-'
        pcV[:] = np.nan
    else:
        stt1 = casosLocDates.iloc[0, 0]
        ###### Prepare data COVID per ABS
        # Create a subtotal per day
        casos = casosLocDates[["data", "numcasos"]].groupby("data").sum()
        # Fill missed dates in cases with 0 cases
        idx = pd.date_range(stt1, ent)
        casos.index = pd.DatetimeIndex(casos.index, dayfirst=True)
        casos = casos.reindex(idx, fill_value=0)
        # Reset index and rename date column
        icasos = casos.reset_index()
```

```
icasos.pop('index')
icasos = icasos.reset_index()
######## Prepare METEO data per ABS
meteo_ABS = df_meteorologicos[df_meteorologicos["ABS_Code"] == ABS_codi[a]]
if len(meteo_ABS) == 0:
   print('No contamination data for this ABS ', ABS_name[a])
   pcV[:] = np.nan
else:
  # Extract data (30 days before for offset)
   stt_lag = pd.to_datetime(stt1) - datetime.timedelta(days=31)
   mask = (meteo_ABS['DATA'] > stt_lag) & (meteo_ABS['DATA'] < ent)
   meteo_ABS = meteo_ABS.loc[mask]
   for v in range(len(var_cont)):
      print (var_cont[v])
      meteo_ABS_var=meteo_ABS[meteo_ABS["codi_variable"] == var_cont[v]]
      if len(meteo_ABS_var) == 0:
         print ('No data for this parameter')
      else:
         meteo_ABS_var.index = meteo_ABS_var['DATA']
       meteo_ABS_var_1=[meteo_ABS_var["valor_lectura"],meteo_ABS_var["DATA"]]
          meteo_ABS_var_2=pd.concat(meteo_ABS_var_1, axis=1, keys= ['valor_lectura',
          'data'])
          meteo_ABS_var_3=meteo_ABS_var_2[['valor_lectura','data']].groupby('data').m
          ean()
         meteo_ABS_var_3 = meteo_ABS_var_3.sort_index()
         casos = casos.sort_index()
         # Merge both dataframes (epidemiological and meteorological)
         data = pd.merge(meteo_ABS_var_3, casos, left_index=True, right_index=True)
         # Use series that have more than 30 days with covid cases
         if len(data.numcasos[data.numcasos >0]) < 31:
            print ('Not enough COVID cases ')
         else:
            flag = True
            # Procedure with correlation offset=0
```

```
            pc = data.corr(method='pearson')
            n = len(data)
            pcL = [0]*len(of)
            pcL[0] = pc.iloc[0, 1
            for i in range(1,len(of)):
               stt_off = pd.to_datetime(stt1) - datetime.timedelta(days=of[i] + 1)
               #Procesing data for each variable
               im = pd.DataFrame(meteo_ABS_var_3[meteo_ABS_var_3.index > stt_off])
               im['Date_original_contaminante'] = im.index
               im.index = im.index + datetime.timedelta(days=of[i])
               # Procesing data for COVID
               casos['Date_original_COVID'] = casos.index
               #Merge both
               data2 = pd.merge(casos, im, left_index=True, right_index=True)
               if len(data2.numcasos[data2.numcasos >0]) < 31
                  print ('Not enough COVID cases for this offset ')
               else:
                 l['data_'+str(i)] = data2
                  # Correlate data with offset
                  data3 = data2[['numcasos','valor_lectura']]
                  pc = data3.corr(method='pearson')
                  pcL[i] = pc.iloc[0, 1]
                  if of[i] == offset_save
                 data3.to_excel(writer,sheet_name=str(ABS_codi[a])+'_'+str(var_cont[v])
            pcV[str(var_cont[v])] = pcL
            pcV['N_'+str(var_cont[v])] = n
      if flag:
         plt.figure(num=f2)
         ##Create figures
         #plt.savefig(ABS_name[a]+'_TimeSeries.png')
         f2=f2
         pcM = pcM.append(pcV, ignore_index=True)
         print(pcV)
         fig = plt.figure(num=f)
```

```
for v in range(len(var_cont)):
   if not np.isnan(sum(pcV[str(var_cont[v])])):
plt.scatter(pcV.index,pcV[str(var_cont[v])],c=colors[v],labe=str(var_cont_label[v]))
      plt.plot(pcV.index,pcV[str(var_cont[v])],  c= colors[v]
plt.text(0.2,0.8,INE_codi[a],  transform=fig.transFigure)
plt.text(0.2,0.75,'Codi INE: '+str(INE_codi[a]),  transform=fig.transFigure)
plt.text(0.2,0.7,'Codi ABS: '+str(ABS_codi[a]), transform=fig.transFigure)
plt.text(0.2,0.65,'TotalCOVIDcases:'+str(data3['numcasos'].sum()),transform=fig.tr
ansFigure)
plt.legend(loc=4)
plt.xlabel('Offset'+' '+ABS_name[a])
plt.ylabel('corr.')
##Create figures
#plt.savefig(ABS_name[a]+'_Offset_corrPAND.png')
f=f+1
###### Output data: meteorological correlation for each offset(selct Lockdown or Pandemic)
writer.save()
xlsname = 'test_AllABS_meteo_Alloffsets_' + stt + '_' + ent + '.xls'
#xlsname = 'test_AllABS_meteo_Alloffsets_PAND_' + stt + '_' + ent + '.xls'
pcM.to_excel(xlsname)
```

8.4 Offset analysis code

8.4.1 Meteorological offset analysis code

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

##### Input data: meterological correlation data (select the document file lockdown or
pandemic)
corrMeteoData = pd.read_excel('Data/test_AllABS_meteo_alloffsets_01-03-2020_15-03-
2020.xls', dtype={'CODI_ABS': str})
#corrMeteoData = pd.read_excel('Data/test_AllABS_meteo_alloffsets_PAND_01-03-
2020_15-03-2021.xls', dtype={'CODI_ABS': str})
var_cont = [30,32,33,35,36]
var_cont_label = ('VV10','T','HR',‘PPT','RS')
### Maximum offset value
```

```
for v in range(len(var_cont)):
    # In anbsolute value
     ABS_code=list(set(corrMeteoData['CODI_ABS']))
    l['maxcorr_'+var_cont[v]]=corrMeteoData[["CODI_INE","Location","CODI_ABS","ABS
    _NAME",  "Offset", var_cont[v]]].iloc[0:1]
    for a in range(len(ABS_code)):
        dummy = corrMeteoData[["CODI_INE","Location", "CODI_ABS","ABS_NAME",
         "Offset", var_cont[v]]][corrMeteoData['CODI_ABS'] == ABS_code[a]]
        maximo = dummy[var_cont[v]].abs().max()
        dummy2 = dummy[dummy[var_cont[v]].abs() == maximo]
        if np.isnan(maximo):
            dummy.iloc[0, :].replace(0,np.NaN)
             l['maxcorr_'+var_cont[v]] = l['maxcorr_' + var_cont[v]].append(dummy.iloc[0,
            :].replace(0,np.NaN))
        else:
            l['maxcorr_'+var_cont[v]] = l['maxcorr_' + var_cont[v]].append(dummy2)
    l['maxcorr_'+var_cont[v]].drop(l['maxcorr_'+var_cont[v]].index[0],  inplace=True)
    l['maxcorr_'+var_cont[v]]=l['maxcorr_'+var_cont[v]].rename({'Offset':'Offset_'+var_cont[
    v]}, axis=1)
maxcorrMeteoData = pd.merge(l['maxcorr_'+var_cont[0]], l['maxcorr_'+var_cont[1]])
for i in range(2, len(var_cont)):
    maxcorrMeteoData = pd.merge(maxcorrMeteoData, l['maxcorr_'+var_cont[i]])
xlsname = 'Maxoffset_corrMeteodata_ABS_Lockdown.xls'
#xlsname = 'Maxoffset_corrMeteodata_ABS_Pandemic.xls'
maxcorrMeteoData.to_excel(xlsname)
### Offset analysis from the max values of offsets
df_Results = pd.DataFrame(index=('Media','Moda', 'ACC'), columns = var_cont)
## Box plots
f=1
plt.figure(num=f)
maxcorrMeteoData.boxplot(column=['Offset_30,'Offset_32','Offset_33','Offset_35','Offset_36
'])
plt.xticks(ticks= range(1,len(var_cont)+1),labels= var_cont_label)
plt.ylabel('Offset')
plt.xlabel('Variable')
for i in range(len(var_cont))
```

```
    n=np.isnan(maxcorrMeteoData[var_cont[i]])[np.isnan(maxcorrMeteoData[var_cont[i]]) == False].size
    means = maxcorrMeteoData['Offset_'+var_cont[i]].mean()
    df_Results.iloc[0,i] = round(means,1)
    modas = maxcorrMeteoData['Offset_'+var_cont[i]].mode()
    if len(modas)==1:
        df_Results.iloc[1,i] = modas[0]
    if means >0 :
        plt.text(i+1, means  , 'n='+str(n))
#plt.savefig('BoxPlot_corrMeteodata_pandemic.png')
plt.savefig('BoxPlot_corrMeteodata_lockdown.png')
f=f+1
#Frecuency histogram
plt.figure(num=f, figsize=(8,4))
for i in range(len(var_cont)):
    plt.subplot(2,3,i+1)
    maxcorrMeteoData['Offset_'+var_cont[i]].hist(bins =30)
    plt.xlim(0,30)

    plt.title(var_cont_label[i])
    plt.xlabel('Offset')
    plt.ylabel('Freq.')
    plt.xticks((0,5,10,15,20,25,30))
plt.tight_layout()
plt.savefig('Hist_Meteo_lockdown.png')
#plt.savefig('Hist_Meteo_Pandemic.png')
f=f+1
for i in range(len(var_cont)):
offset=df_coeff_acum[var_cont[i]][df_coeff_acum[var_cont[i]]==df_coeff_acum[var_cont[i]].max()].index
    if len(offset) > 0:
        df_Results.iloc[2,i] = offset[0]
df_Results.to_excel('OFFSETS_Meteo_lockdown.xls')
#df_Results.to_excel('OFFSETS_Meteo_Pandemic.xls')
```

8.5 Multiple linear regression codes

8.5.1 Preprocessing multiple linear regression code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
 import datetime as datetime
##### Input data: meterological variables, weather stations, ABS
var_cont = [32,33,36]
var_cont_label = ('T','HR','RS')
locations = pd.read_excel('Data/Location_ABS_Meteo_stations.xlsx')
INE_codi = locations['MunicipiCodi']
munic = locations['MunicipiDescripcio']
ABS_codi = locations['CodiABS']
ABS_name = locations['nomABS']
estacio = locations['Estacio']
###### Input data from meteorological correlation
meteo_input = pd.ExcelFile('Data/Timeseries_Pand_Offset_Meteo_15_out.xls')
#meteo_input = pd.ExcelFile('Data/Timeseries_Pand_Offset_Meteo_15_out_pandemics.xls')
sheets = meteo_input.sheet_names
for s in range(len(sheets)):
    l[sheets[s]] = pd.read_excel(meteo_input, sheets[s])
## Meteorological variables
contaminantes =([])
ABS_code =([])
for i in  range(len(sheets)):
   ABS_code.append(sheets[i].split('_')[0])
   contaminantes.append(sheets[i].split('_')[1])
ABS_code = list(set(ABS_code))
contaminantes = list(set(contaminantes))
for c in range(len(contaminantes)):
#Create a datagrame for each meteorological variable
   l['df_'+contaminantes[c]] = pd.DataFrame(columns= ('VALOR_LECTURA','CODI_ABS'))
for i in  range(len(sheets)):
```

```
    data = pd.read_excel(meteo_input, sheets[i])
    data.index = data.iloc[:,0]
    data2 = data.iloc[:,2]
    data2=data2.to_frame()
    data2['CODI_ABS'] = sheets[i].split('_')[0]
    if len(l['df_'+sheets[i].split('_')[1]]) ==0
       l['df_'+sheets[i].split('_')[1]] = data2
    else:
       l['df_'+sheets[i].split('_')[1]] = l['df_'+sheets[i].split('_')[1]].append(data2)
  ## Fullfill dataframes and sort by date
for c in range(len(contaminantes)):
   l['df_'+contaminantes[c]] = l['df_'+contaminantes[c]].rename(columns={"valor_lectura": var_cont_label[var_cont.index(int(contaminantes[c]))]})
   l['df_'+contaminantes[c]]['DATA'] = l['df_'+contaminantes[c]].index
DF_contaminantes = pd.merge(l['df_'+contaminantes[0]] , l['df_'+contaminantes[1]] , how='outer', on = ('DATA','CODI_ABS'))
for c in range(2,len(contaminantes)):
   DF_contaminantes = pd.merge(DF_contaminantes , l['df_'+contaminantes[c]] , how='outer', on = ('DATA','CODI_ABS'))
DF_contaminantes[['CODI_ABS']] = DF_contaminantes[['CODI_ABS']].apply(pd.to_numeric)
###### Input data: epidemiological webscrapping
# Select in webscraping clinics ABS the desireable option: Lockdown or Pandemic
from webscrapping_CLINICAL_ABS import *
# Modify the format in web scraping
clinics[['regiosanitariacodi', 'sectorsanitaricodi', 'abscodi', 'sexecodi', 'numcasos']] = clinics[['regiosanitariacodi','sectorsanitaricodi', 'abscodi', 'sexecodi', 'numcasos']].apply(pd.to_numeric)
clinics["DATA"] = pd.to_datetime(clinics["data"], dayfirst=True)
df_COVID=clinics[["data","abscodi","numcasos"]].groupby(['abscodi','data']).sum().reset_index()
df_COVIDs = df_COVID.rename(columns={'data' :'DATA' , 'abscodi' :'CODI_ABS' })
df_COVIDs['CODI_ABS'] = df_COVIDs['CODI_ABS'].map(int)
df_COVIDs['CODI_ABS'] = df_COVIDs['CODI_ABS'].map(str)
df_COVIDs['DATA'] = pd.to_datetime(df_COVIDs['DATA'], dayfirst=True)
df_COVIDs[['CODI_ABS']] = df_COVIDs[['CODI_ABS']].apply(pd.to_numeric)
```

```
##Merge both dataframes
DF = pd.merge(DF_contaminantes , df_COVIDs , how='left', on = ('DATA','CODI_ABS'))
###### Data ready for Stadistics with offset already applied to contaminantes series
DF.to_excel('TimeSeries_Meteo_Pand_Offset_15_Regresion.xlsx')
#DF.to_excel('TimeSeries_Meteo_Pand_Offset_15_Regresion_pandemics.xlsx')
```

8.5.2 Meteorological multiple linear regression code

```
from sklearn import linear_model
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
###### Input data: variables, data extract from preprosesing code
# var = [30,32,33,35,36]
var = ('T','HR','RS')
print('Import data')
DF_Data = pd.read_excel('Data/TimeSeries_Meteo_Pand_Offset_15_Regresion.xlsx')
DF_Data = DF_Data.fillna(0)
##Organize the importet dataframe.
DF_Data = DF_Data[['Unnamed: 0','CODI_ABS', 'DATA', 'RS','T','HR', 'numcasos']]
DF_Data = DF_Data.drop('Unnamed: 0', axis=1)
## Extract from the dataframe data and ABS column
datos = DF_Data.iloc[:,2:]
ABS_code = list(set(DF_Data['CODI_ABS']))
Contaminantes = DF_Data.columns[2:-1]
# Standardize data
print(' Standardize data')
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
# transform data
scaled = scaler.fit_transform(datos)
dummy = pd.DataFrame(scaled, columns= datos.columns)
## To obtain standardize COVID data per ABS and date
dummy_1 = pd.concat([DF_Data, dummy], axis=1)
COVID_datos = dummy_1.drop(['RS','T','HR'], axis=1)
```

```
COVID_datos.columns =['CODI_ABS','DATA','numcasos','numcasos_st']
COVID_datos.set_index('DATA', inplace = True)
datos = dummy
DF_Data_standard = DF_Data.iloc[:,0:2]
for v in range(len(var)):
    DF_Data_standard [var[v]] = datos.iloc[:,v]
# Multiple lineal regression
print('Multiple Lineal regression')
reg = linear_model.LinearRegression()
reg.fit(datos.iloc[:,:-1], datos.iloc[:,-1])
a=([])
for i in range(len(datos)):
    dummy=([])
    for j in range(len(reg.coef_)):
        dummy.append(datos[datos.iloc[:,:-1].columns[j]].iloc[i]*reg.coef_[j])
    a.append(sum(dummy)+reg.intercept_)
print ('Regression coef.')
for c in range(len(Contaminantes)):
    print (Contaminantes[c],'   ', str(reg.coef_[c]))
COVID_reconstructed = pd.DataFrame(a,index = DF_Data['DATA'])
#Aqui se relaciona lo anterior con el codigo ABS respectivo
COVID_reconstructed ['CODI_ABS'] = np.array(DF_Data['CODI_ABS'])
f= 100
for m in range(len(ABS_code)):
    plt.figure(f, figsize =(20,5))
    plt.subplot(2,1,1)
    plt.figure(m)
#m=163 ###ABS 192
m=351 ###ABS 403
    # Standardized COVID data
COVID_datos_1 = COVID_datos[COVID_datos['CODI_ABS'] == ABS_code[m]]
COVID_datos_1['numcasos_st'].plot(label='Standardized COVIDcases', c='#00cc44')
plt.legend(loc = 2)
plt.gca().set_ylim(bottom=-1)
```

```
reconstruidos_ABS = COVID_reconstructed[COVID_reconstructed['CODI_ABS'] == ABS_code[m]]
reconstruidos_ABS.iloc[:,0].plot( label='Reconstruted COVIDcases ', c='r')
plt.gca().set_ylim(bottom=-1)
plt.legend(loc = 1)
plt.title (ABS_code[m])
plt.savefig('Reconstruccion_Meteo_'+str(ABS_code[m])+'.png')
##Preparation to calculate goodness fit
COVID_datos_3 = COVID_datos_1.drop(['numcasos'], axis=1)
COVID_datos_st = COVID_datos_3 [['numcasos_st','CODI_ABS']]
   # Datos originales estandarizados
COVID_datos_1 = COVID_datos[COVID_datos['CODI_ABS'] == ABS_code[m]]
##########To calculate in case of ABS Sabadell
m = 163
   # Standarized original COVID data
COVID_datos_1 = COVID_datos[COVID_datos['CODI_ABS'] == ABS_code[m]]
   ## Standarized reconstruted COVID data
reconstruidos_ABS = COVID_reconstructed[COVID_reconstructed['CODI_ABS'] == ABS_code[m]]
COVID_datos_3 = COVID_datos_1.drop(['numcasos'], axis=1)
COVID_datos_st = COVID_datos_3 [['numcasos_st','CODI_ABS']]
#########################
# To calculate goodness fit
from sklearn.metrics import r2_score
r2 = r2_score(COVID_datos_st, reconstruidos_ABS)
print('r2 score for model is', r2)
```

8.5.3 Mobility multiple linear regression code

```
from sklearn import linear_model
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
###### Input data: mobility data, epidemiological data, variables
print('Import mobility data')
df_mobility = pd.read_excel('Data/Mobility_data_v1_ABS403.xlsx')
```

```
df_mobility['DATA'] = df_mobility[['datetime']]
###Analysis period
df_mobility_1=df_mobility.query("DATA >= '2020-03-15 00:00:00' and DATA <='2020-11-15 00:00:00'")
# Variables
var = ('q','m','r','p','delta')
print('Import clinical data')
DF_Data = pd.read_excel('Data/TimeSeries_Meteo_Pand_Offset_15_Regresio_ABS403.xlsx')
DF_Data = DF_Data.fillna(0)
DF_Data=DF_Data.query("DATA >= '2020-03-15 00:00:00' and DATA <='2020-11-15 00:00:00'")
DF_Data = DF_Data.drop(['Unnamed: 0','RS','T', 'HR'],  axis='columns')
# Mergre both input data
DF_Datamobility = pd.merge(df_mobility_1 , DF_Data , how='outer', on = ('DATA','CODI_ABS'))
DF_Datamobility_1= DF_Datamobility[['CODI_ABS', 'DATA','q','m','r','p','delta','numcasos']]
datos = DF_Datamobility_1.iloc[:,2:]
## Crate a list of contaminants and ABS
ABS_code = list(set(DF_Datamobility_1['CODI_ABS']))
Contaminantes = DF_Datamobility_1.columns[2:-1]
print('Standardize data')
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
# transform data
scaled = scaler.fit_transform(datos)
dummy = pd.DataFrame(scaled, columns= datos.columns)
## To obtain standardize COVID data per ABS and date
dummy_1 = pd.concat([DF_Datamobility_1, dummy], axis=1)
COVID_datos = dummy_1.drop(['q','m','r','p','delta',], axis=1)
COVID_datos.columns =['CODI_ABS','DATA','numcasos','numcasos_st']
COVID_datos.set_index('DATA', inplace = True)
datos = dummy
DF_Datamobility_1_standard = DF_Datamobility_1.iloc[:,0:2]
for v in range(len(var)):
```

```
    DF_Datamobility_1_standard [var[v]] = datos.iloc[:,v]
# Mutiple lineal regression
print('Multiple Lineal regression')
reg = linear_model.LinearRegression()
reg.fit(datos.iloc[:,:-1], datos.iloc[:,-1])
print('Mobility Multiple Lineal regression')
a=([])
for i in range(len(datos)):
    dummy=([])
    for j in range(len(reg.coef_)):
        dummy.append(datos[datos.iloc[:,:-1].columns[j]].iloc[i]*reg.coef_[j])
    a.append(sum(dummy)+reg.intercept_)
print ('Regression coef.')
for c in range(len(Contaminantes)):
    print (Contaminantes[c],'   ', str(reg.coef_[c]))
COVID_reconstructed = pd.DataFrame(a,index = DF_Data['DATA'])
COVID_reconstructed ['CODI_ABS'] = np.array(DF_Data['CODI_ABS'])
f= 100
for m in range(len(ABS_code))
    plt.figure(f, figsize =(15,8))
    plt.figure(m)
    # Standardize original cases
COVID_datos_1 = COVID_datos[COVID_datos['CODI_ABS'] == ABS_code[m]]
COVID_datos_1['numcasos_st'].plot(label='Standardized COVIDcases', c='#00cc44')
plt.legend(loc = 2)
plt.gca().set_ylim(bottom=-2)
    #Reconstruct dara per ABS (mobility, meteo and contamination)
from MultipleLinearRegression_Pand_MeteoContMobility_ABS_403 import
reconstruidos_ABS_mobilitymeteocont
reconstruidos_ABS_mobilitymeteocont.iloc[:,0].plot( label='Standardized Reconstruted
COVIDcases Environmental+Mobility')
plt.legend(loc = 2)
plt.gca().set_ylim(bottom=-2)
    #Reconstruted data per ABS (Only mobility)
```

```
reconstruidos_ABS_mobility = COVID_reconstructed[COVID_reconstructed['CODI_ABS'] == ABS_code[m]]
reconstruidos_ABS_mobility.iloc[:,0].plot( label='Standardized Reconstruted COVIDcases Mobility', c='r')
plt.gca().set_ylim(bottom=-2)
plt.legend(loc = 1)
plt.title (ABS_code[m])
plt.ylabel('Standardized COVIDcases')
plt.savefig('Reconstruccion_MOVILITYMETEOCONT_'+str(ABS_code[m])+'.png'
## Preparation to calculate goodness fit
COVID_datos_3 = COVID_datos_1.drop(['numcasos'], axis=1)
COVID_datos_st = COVID_datos_3 [['numcasos_st','CODI_ABS']]
#To calculate goodness fit
from sklearn.metrics import r2_score
r2 = r2_score(COVID_datos_st, reconstruidos_ABS_mobility)
print('Mobility r2 score for model is', r2)
```

8.6 Self-evaluation Questionnaire

a) Evaluate the acquired **competences** according to the **tasks** you have carried out.

Degree Competences		**Task in which you have observed the competence**	**Self evaluation [Rank 1 to 10]**	**Aspects to be improved**
SPECIFIC COMPETENCES				
A1.1	Effectively apply knowledge of basic, scientific and technological materials pertaining to engineering.	All task in general	9	
A1.2	Design, execute and analyze experiments related to engineering	All task in general such as Pythpn codes and their results interpretation	8	
A1.3	Be able to analyze and synthesize the continuous progress of products, processes, systems and services, whilst applying criteria of safety, economic viability, quality and environmental management. (G6)	Analyzing the effect of environmental variables in COVID-19 transmission in reference to literature review.	9	

A1.4	Know how to establish and develop mathematical models by using the appropriate software in order to provide the scientific and technological basis for the design of new products, processes, systems and services and for the optimization of existing ones. (G5)	Each code to analyses effect of environmental variables with COVID-19 transmission.	8	
A2.1	Be able to apply the scientific method and the principles of engineering and economics to formulate and solve complex problems that arise in processes, equipment, installations and services, in which the material undergoes changes to its composition, state or energy content, these changes being characteristic of industrial chemistry and other related sectors such as pharmacology, biotechnology, materials sciences, energy, food and the environment. (G1)	Programming methods and principles in Python codes during all the task of the project.	9	
A2.2	Conceive, project, calculate and design processes, equipment, industrial installations and services in the field of chemical engineering and related industrial sectors in terms of quality, safety, economics, the rational and efficient use of natural resources and the conservation of the environment. (G2)	-	-	
A2.3	Lead and technically and economically manage projects, installations, plants, companies and technological centres in the ambit of chemical engineering and related industrial sectors. (G3)	-	-	
A3.1	Apply knowledge of mathematics, physics, chemistry, biology and other natural sciences by means of study, experience, practice and critical reasoning in order to establish economically viable solutions for technical problems (I1).	In Python codes such as introducing multiple regression fit and interpret results.	9	
A3.2	Design and optimize products, processes, systems and services for the chemical industry on the basis of various areas of chemical engineering, including processes, transport, separation operations, and chemical, nuclear, elctrochemical and biochemical reactions engineering (I2).	To design the model to predict COVID-19 with meteorological and air pollutants variables independently and simultaneously.	9	
A3.3	Conceptualize engineering models and apply innovative problems solving methods and appropriate IT applications to the design, simulation,	All the task in the project such as method applications of modeling the	8	

	optimization and control of processes and systems (I3).	influence of COVID-19 with environmental variables.		
A3.4	Be able to solve unfamiliar and ill-defined problems by taking into account all possible solutions and selecting the most innovative. (I4)	In the study of different scenarios in reference to analyze the influence of environmental variables and COVID-19 confirmed cases.	8	
A3.5	Lead and supervise all types of installation, process, system and service in the different industrial areas related to chemical engineering (I5).	-	-	
A3.6	Design, construct and implement methods, processes and installations for the integrated management of waste, solids, liquids and gases, whilst also taking into account the impacts and risks of these products (I6).	-	-	
A4.1	Lead and organize companies and production and service systems by applying knowledge and abilities regarding industrial organization, commercial strategy, planning and logistics, mercantile and labour legislation, and financial and costs accounting (P1).	-	-	
A4.2	Lead and manage the organization of work and human resources by applying criteria regarding industrial safety, quality management, occupation risk prevention, sustainability and environmental management (P2).	All tasks in general developed in the project	8	
A4.3	Manage research, development and technological innovation whilst ensuring the transfer of technology and taking into account property and patent rights (P3).	-	-	
A4.4	Adapt to structural changes in society caused by economic, energy or natural factors so as to be able to solve any resulting problems and to contribute technological solutions with a high commitment to sustainability (P4).	All the project in general such as the influence of COVID-19 transmission, detected COVID-19 variants and factors that affect in confirmed cases.	9	
A4.5	Lead and monitor the control of installations, processes, products, certification, auditing,	-	-	

	verification, testing and reports (P5).			
A5.1	Carry out, present and defend (once all the curriculum credits have been obtained) an original individually produced piece of work before a university panel. The work will consist of a professional integrated Chemical Engineering project that synthesizes (TFM1)	In the task of writing a master thesis and a future defense in a tribunal.	In execution	
TRANSVERSAL COMPETENCES				
B1.1	Communicate and discuss proposals and conclusions in a clear and unambiguous manner in specialized and non-specialized multilingual forums (G9).	When it is written the master thesis (conclusions, results…)	9	
B1.2	Adapt to changes and be able to apply new and advanced technologies and other important developments with initiative and entrepreneurial spirit. (G10)	Task with introduction of different factors in programming program such as Python.	10	
B2.1	Lead and define multidisciplinary teams that are able to make technical changes and address management needs in national and international contexts. (G8)	-	-	
B3.1	Work in a team with responsibilities shared among multidisciplinary, multilingual and multicultural teams	The project is carried out with PROCEED project where some technological units are implicated.	9	
B4.1	Be able to learn autonomously in order to maintain and improve the competences pertaining to chemical engineering that enable continuous professional development. (G11)	Using the Python program for the first time.	10	
B5.1	Carry out and lead the appropriate research, design and development of engineering solutions in new or little understood areas, whilst applying criteria of creativity, originality, innovation and technology transfer. (G4)	All the task in general studying the effect of environmental variables in refrence of COVID-19 prpagation.	8	
B5.2	Bring together knowledge, make judgements and take decisions on the basis of incomplete or limited knowledge whilst taking into account the social and ethical responsibilities of professional practice. (G7)	All task in general, especially in the offset selection.	9	

NUCLEAR COMPETENCES				
C1.1	Have an intermediate mastery of a foreign language, preferably English	In all task of the project (literature review, writing the project, among other tasks).	9	
C1.2	Be advanced users of the information and communication technologies	All tasks in general with the use of Excel, Word, websites, email, among others.	10	
C1.3	Be able to manage information and knowledge	All the tasks that composed the master thesis.	9	
C1.4	Be able to express themselves correctly both orally and in writing in one of the two official languages of the URV	Each task in the project and meetings every week, writing the master thesis, emails.	10	
C2.1	Be committed to ethics and social responsibility as citizens and professionals	All tasks in general.	10	
C2.2	Be able to define and develop their academic and professional project	All tasks in general.	10	

b) Evaluate the final master project and suggest improvements.

Key steps	**Evaluation [Mark 1 to 10]**	**Improvement proposed**
Selection/assignment of the project (dissemination, communication, assignment requirements...)	9	
Stay (welcome, length, relationship, follow-up made by the company...)	10	
Follow-up made by URV tutor	10	
Other aspects to be considered (which ones...). Reicieved ease to work and doing the project in Eurecat simultaneosly	10	