|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Картинки по запросу UNS logo | Картинки по запросу hospital pasteur |  |  |  | |  |  |

Irina GRIGOREVA – Oualid MANAI – Walid RHAZADI – Sergey GORIANIN

Irina-gry@yandex.ru

MBDS 2017

Project manager : Pr Gabriel Mopolo

Scientific Resp : Pr Serge Miranda, Director of MBDS (www.mbds-fr.org)

Annotation

In this research, correlations between environmental and health data are investigated. The aim of the research was to investigate whether or not the air pollution affects the growth of the number of patients with Dyspnea in Nice. The data used for the research were collected by the AirPaca service and the Hospital Pasteur between January 2014 and December 2016.

A preliminary Study on the Correlation Between the ENVIRONMENT and THE HEALTH DATA IN NICE

**Table of contents**

[Introduction 2](#_Toc478133512)

[1 Method 3](#_Toc478133513)

[1.1 Description of the Health Data 3](#_Toc478133514)

[1.2 Description of the Environmental Data 3](#_Toc478133515)

[1.3 Analysis of the Data 4](#_Toc478133516)

[2. Results 5](#_Toc478133517)

[2.1 The calculation of the linear correlation coefficient of unfiltered data 5](#_Toc478133518)

[2.2 The calculation of the linear correlation coefficient of filtered data by diagnosis 5](#_Toc478133519)

[2.3 The calculation of the linear correlation coefficient of interval data 6](#_Toc478133520)

[2.4 Linear Regression 7](#_Toc478133521)

[3. Charts 8](#_Toc478133522)

[Conclusion 9](#_Toc478133523)

[References 11](#_Toc478133524)

[Annexes 13](#_Toc478133525)

# Introduction

The most important mechanism for management decisions aimed at improving air quality and reducing the negative impact of environmental factors on the human body is to carry out a health risk assessment of the population. Methodology for assessing the health risk is an element of mathematical modeling of the causal relationships between environmental factors and health, under their influence in the specific conditions of time and area.

At the present time, there is a large number of studies on the impact of air pollution on human health. Studies in various geographical areas have shown associations of respiratory symptoms and conditions with long-term exposure to total suspended particulates (TSP) and SO2 ([1-7](#_References)), to particulate matter ([8-10](#_References)), to black smoke ([11](#_References)), and to NO2 ([7](#_References)). Furthermore, studies of hospital admissions and mortality studies point to an association of short- and long-term exposure to air pollution with symptoms that are related both to pulmonary and to cardiac diseases ([12-19](#_References)).

Thus, the objective of this study - to test the hypothesis about the dependence of the presence of the above substances in the air and respiratory disorders of patients of the Central Hospital of Nice.

As a result of work were carried out the calculation and evaluation of the correlation of the environmental and the health data in Nice through mathematical calculations, constructing hypotheses and testing them.

This work is a part of the Big Data project called “Big Bridge – SE : Projet Big Data Santé et Environnement dans la ville de Nice en partenariat avec l’IMREDD" of students of MBDS ([23](#_References)).

# 1 Method

## 1.1 Description of the Health Data

For the health part of this research was used the data about patients with Dyspnea of the Central Hospital of Nice, in the period from January 2014 to December 2016. The data structure is presented in Table 1.1. Data is presented in csv format.

Volume: 7854 rows, 25 columns.

|  |  |
| --- | --- |
| Field | Description |
| Gender | *Gender* |
| Age | *Age* |
| Address | *Address* |
| Postal code | *Postal code* |
| Ville | *Ville* |
| Admission | *The date of the admission* |
| Sortie | *The date of the exit* |
| Examen | *The date of the examination* |
| Categorie de Recours | *The category of remedy – group of medicines that patient got* |
| Libelle de Recours | *Diagnosis and value of DEP (peak expiratory flow) and suspicion and/or theoretical diagnosis* |
| Code de Recours | *Code of ICD-10 (International Statistical Classification of Diseases and Related Health Problems) by World health organization* |
| Libelle gravite | *Level of seriousness* |
| Libelle CCMU | *Wording of Classification Clinique des maladies aux urgences)* |
| Destination Confirmee | *Сonfirmed patient organization* |
| Type de sortie | *Format of the treatment* |
| Diag1 – diag10 | *Final diagnosis - code of ICD-10 (International Statistical Classification of Diseases and Related Health Problems) by World health organization* |

**Table 1.1 – The Structure of The Health Data**

## 1.2 Description of the Environmental Data

For the environmental part of this research was used the data about the air pollution in Nice from the service AirPaca (20), in the period from January 2014 to December 2016. The data structure is presented in Table 1.2. Data is presented in csv format.

Volume: 23016 rows, 6 columns.

|  |  |
| --- | --- |
| Field | Description |
| Station | *Station* |
| Pollutant | *Pollutant* |
| Description | *Full name of the pollutant* |
| Unité | *µg/m3* |
| Date | *Date* |
| Value | Value |

**Table 1.2 – The Structure of The Environmental Data**

## 1.3 Analysis of the Data

The statistical analysis of the data from 7854 patients and 6 pollutants (day-by-day) was processed with the R language. R is an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis ([21](#_References)).

R Language allows to group and filter the data, automatically calculate a correlation coefficients, as well as to build the necessary charts for visualization.

For the calculate measure of dependence between quantities was used the Pearson product-moment correlation coefficient, or "Pearson's correlation coefficient", commonly called simply "the correlation coefficient". It is obtained by dividing the covariance of the two variables by the product of their standard deviations.

The population correlation coefficient ρX,Y between two random variables X and Y with expected values μX and μY and standard deviations σX and σY is defined as:

where E is the expected value operator, cov means covariance, and corr is a widely used alternative notation for the correlation coefficient.

The Pearson correlation is +1 in the case of a perfect direct (increasing) linear relationship (correlation), −1 in the case of a perfect decreasing (inverse) linear relationship (anticorrelation), and some value in the open interval (−1, 1) in all other cases, indicating the degree of linear dependence between the variables. As it approaches zero there is less of a relationship (closer to uncorrelated). The closer the coefficient is to either −1 or 1, the stronger the correlation between the variables (22).

To find the linear correlation were used the calculated values of the deviation from the mean value for the entire set of data:

where x – the value of the data set, xav - the mean value.

The architecture of the system is presented in [Annexes](#_Annexes).

# 2. Results

## 2.1 The calculation of the linear correlation coefficient of unfiltered data

The data sets for analysis in this case:

* a set of average values (for all stations) for each pollutant: NO, NO2, NOX, O3, PM10, PM2,5;
* a set of the total number of patients the day-by-day.

The result is presented in Table 2.1.

|  |  |  |
| --- | --- | --- |
| Data set 1 | Data set 2 | The correlation coefficient |
| ∆NO | ∆Patients | 0.1804087 |
| ∆NO2 | ∆Patients | 0.1012164 |
| ∆NOX | ∆Patients | 0.1616413 |
| ∆O3 | ∆Patients | -0.2199504 |
| ∆PM10 | ∆Patients | 0.05497636 |
| ∆PM2,5 | ∆Patients | 0.0786113 |

**Table 2.1 – Result of the searching correlation. Case 1**

As we can see from the results, the largest correlation coefficient with the set of data with numbers of patients has data of pollutant NOX. But this value can’t be considered sufficient for the approval of the existence correlation.

## 2.2 The calculation of the linear correlation coefficient of filtered data by diagnosis

The data sets for analysis in this case:

* a set of average values (for all stations) for each pollutant: NO, NO2, NOX, O3, PM10, PM2,5;
* a set of the total number of patients the day-by-day, filtered by diagnosis to 2 data sets: patients with lung diseases and patients with heart diseases.

The result is presented in Table 2.2.

|  |  |  |
| --- | --- | --- |
| Data set 1 | Data set 2 | The correlation coefficient |
| ∆NO | ∆Patients(lung) | 0.1719213 |
| ∆NO2 | ∆Patients(lung) | 0.08888615 |
| ∆NOX | ∆Patients(lung) | 0.1535681 |
| ∆O3 | ∆Patients(lung) | -0.2045952 |
| ∆PM10 | ∆Patients(lung) | 0.06846391 |
| ∆PM2,5 | ∆Patients(lung) | 0.0973422 |
| ∆NO | ∆Patients(heart) | 0.1060834 |
| ∆NO2 | ∆Patients(heart) | 0.03138159 |
| ∆NOX | ∆Patients(heart) | 0.08371402 |
| ∆O3 | ∆Patients(heart) | -0.1386231 |
| ∆PM10 | ∆Patients(heart) | -0.005957092 |
| ∆PM2,5 | ∆Patients(heart) | 0.002553244 |

**Table 2.2 – Result of the searching correlation. Case 2**

In the results, there is no such coefficient to affirm the correlation between the data.

## 2.3 The calculation of the linear correlation coefficient of interval data

To search the correlation considering the factors such as the presence of disease incubation period, and other anthropological factors, it was decided to group the data of air pollution and the total number of patients shown in paragraph 2.1, by 5 days. For each pollutant - the average value of 5 days, and for health data - the total number of patients of the 5 days.

The result is presented in Table 2.3.

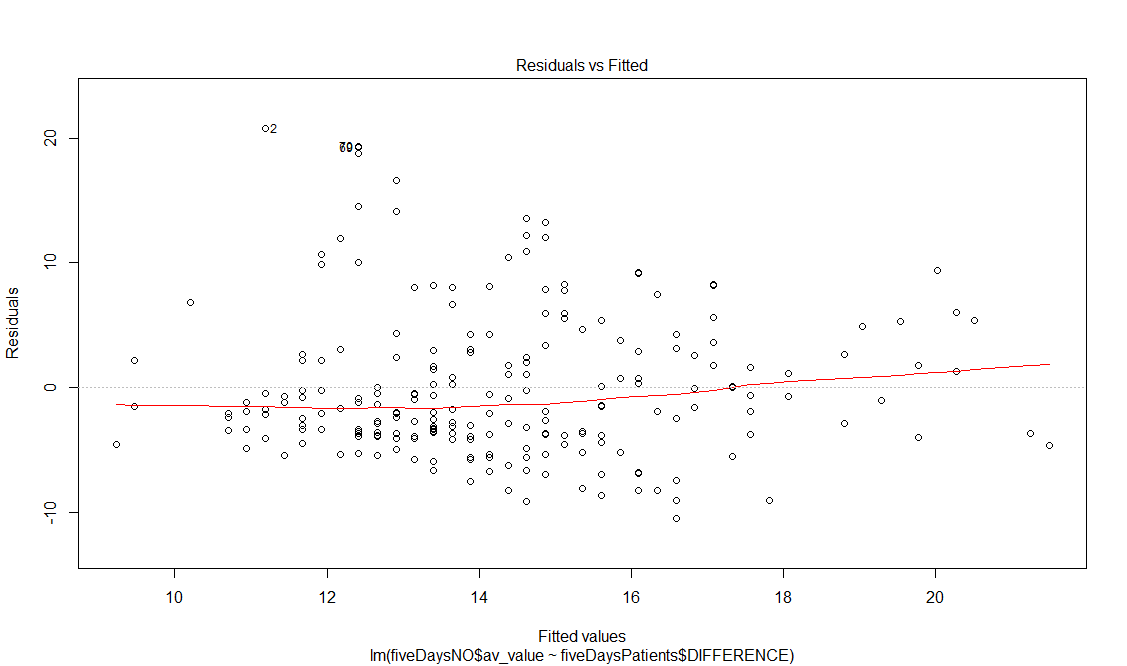
|  |  |  |
| --- | --- | --- |
| Data set 1 | Data set 2 | The correlation coefficient |
| ∆NO | ∆Patients | 0.36395 |
| ∆NO2 | ∆Patients | 0.2697953 |
| ∆NOX | ∆Patients | 0.3613143 |
| ∆O3 | ∆Patients | -0.371573 |
| ∆PM10 | ∆Patients | 0.1243245 |
| ∆PM2,5 | ∆Patients | 0.1249298 |

**Table 2.3 – Result of the searching correlation. Case 3**

These coefficients are not sufficient to confirm the existence of a correlation.

## 2.4 Linear Regression

For the NO pollutant (interval data set), which has the highest correlation coefficient with patient data, was constructed a linear regression model for visual verification of the correlation of data. “Residuals vs Fitted” chart:



# 3. Charts

Figure 1 Number of patients – Value of NO Chart

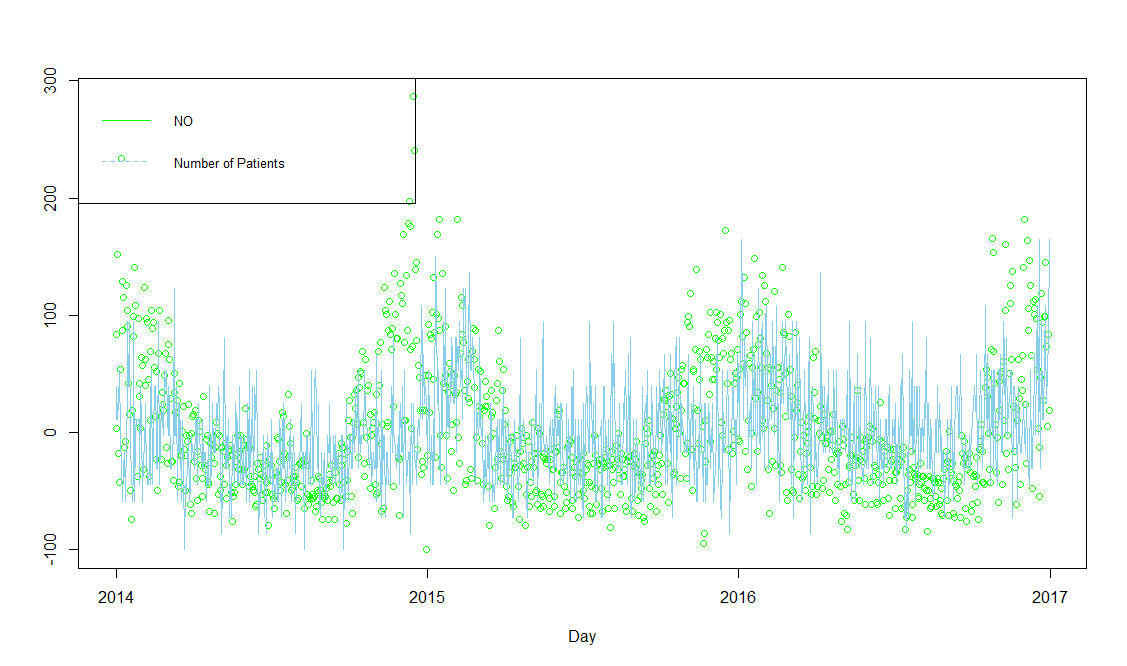


Figure 2 Number of patients – Value of NO2 Chart

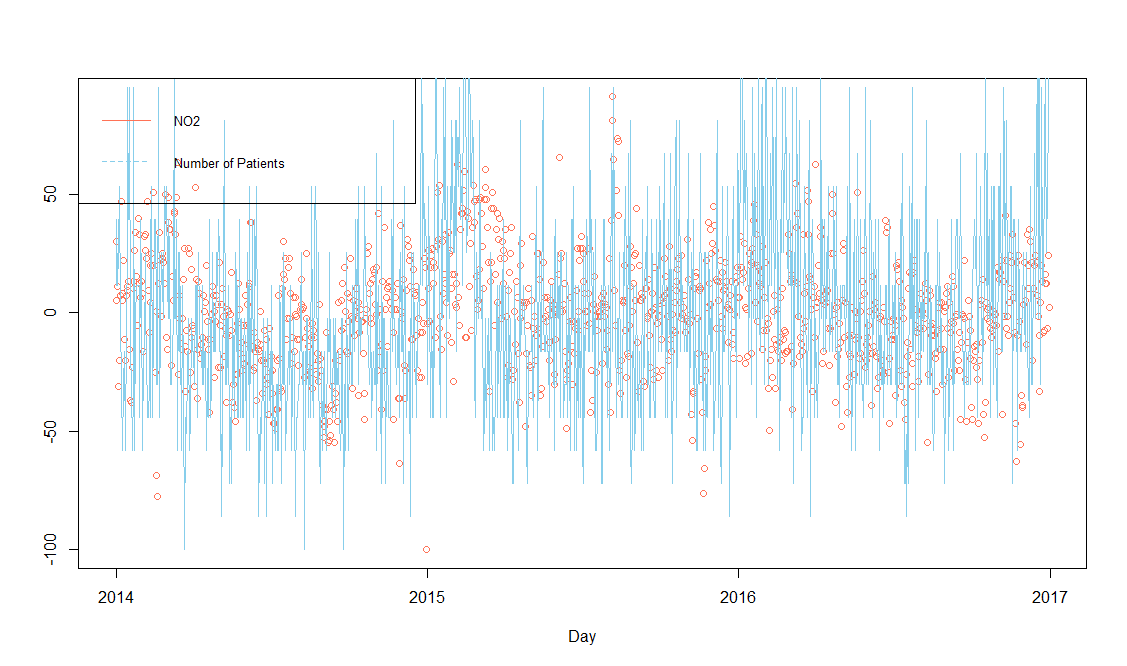
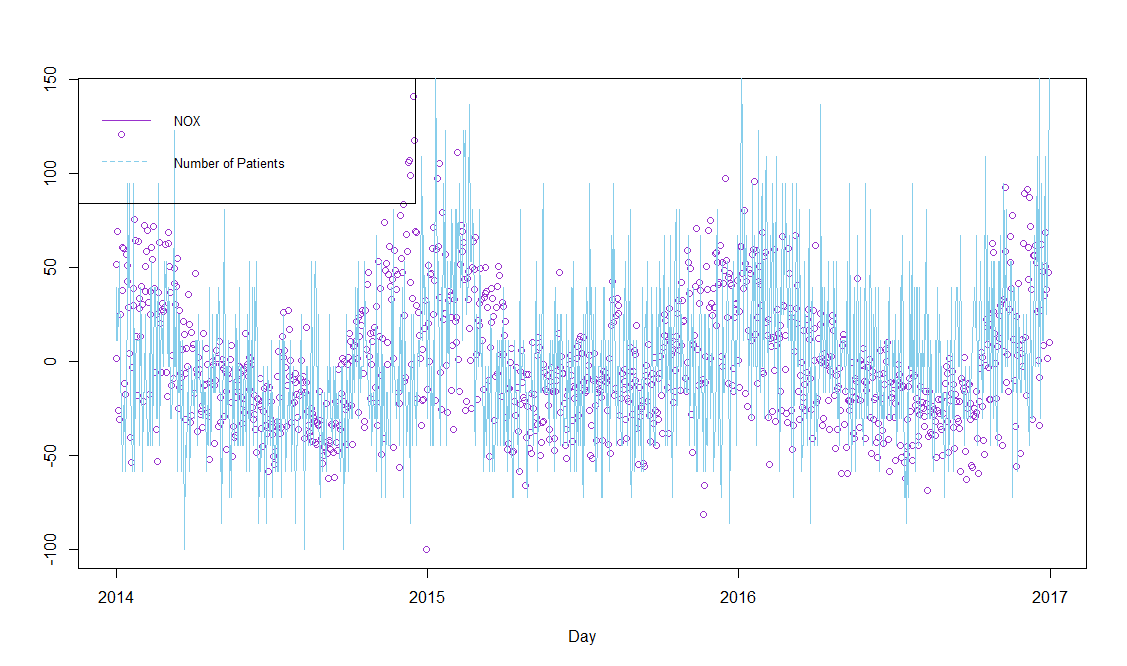


Figure 3 Number of patients – Value of NOX Chart;



# Conclusion

An analysis of the presented data for the existence of a linear correlation was carried out. The linear correlation between data of air pollution and the number of patients with Dyspnea was not established in this study.

As can be seen from the results, there is a sense to continue experiments of finding correlation with such pollutants: NO, NО2 and NОХ, which show the greatest possibility of correlation.

Also, it is planned to enrich the project with unstructured data coming both from social networks and captors in connected watches (monitoring heart beats in real time) to enable real-time cartography for monitored individuals.

For further analysis of the existing data, the following approaches are proposed, which can be implemented with the language R:

1) Clustering: to identify the Risk Factors (such as age, for example, or place of living) for patients with Dyspnea in Nice for existing health data;

2) Classification: a method that allows a person to be assigned to a group defined in point 1, according to his description. It allows to determine the risk of Dyspnea for a person by means of predictive functions. It can be used in the future when including to the system the individual user profiles.

Then - to correlate personal Risk Factors and the environment data for constructing recommendations for the user (for example, caution about visiting heavily polluted areas if a person has high blood pressure).

# References

1. Holland W. W., Reid D. D.The urban factor in chronic bronchitis. Lancet11965445448
2. PAARC: Groupe Cooperative /Lelouche JPollution atmosphérique et affections respiratoires chroniques ou à répétition. Bull. Eur. Physiopathol. Respir.18198287116
3. Schenker M. B., Samet J. M., Speizer F. E., Gruhl J., Batterman S.Health effects of air pollution due to coal combustion in the Chestnut Ridge region of Pennsylvania: results of cross-sectional analysis in adults. Arch. Environ. Health381983325330
4. Euler G. L., Abbey D. E., Magie A. R., Hodlkin J. E.Chronic obstructive pulmonary disease symptom effects of long term cumulative exposure to ambient levels of total suspended particulates and sulfur dioxide in California Seventh-Day Adventist residents. Arch. Environ. Health421987213222
5. Portney P., Mullahy J.Urban air quality and respiratory disease. Reg. Sci. Urban Econ.201990407418
6. Schwartz J.Particulate air pollution and chronic respiratory disease. Environ. Res.621993713
7. Forsberg, B., N. Stjernberg, and S. Wall. 1997. Prevalence of respiratory and hyperreactivity symptoms in relation to levels of criteria air pollutants in Sweden. Eur. J. Public Health 7/3:291–296.
8. Abbey D. E., Lebowitz M. D., Mills P. K., Petersen F. F., Beeson W. L., Burchette R. J.Long-term ambient concentrations of particulates and oxidants and development of chronic disease in a cohort of nonsmoking California residents. Inhal. Toxicol.119952134
9. Abbey, D. E., B. E. Ostro, F. Petersen, and R. J. Burchette. 1995. Chronic respiratory symptoms associated with estimated long-term ambient concentrations of fine particulates less than 2.5 microns in aerodynamic diameter (PM2.5) and other air pollutants. J. Exp. Anal. Environ. Epidemiol. 5/2:137–159.
10. Abbey D. E., Hwang B. L., Burchette R. J.Estimated long term ambient concentrations of PM10 and development of respiratory symptoms in a nonsmoking population. Arch. Environ. Health501995139150
11. Scarlett J. F., Griffiths J. M., Strachan D. P., Anderson H. R.Effect of ambient levels of smoke and sulphur dioxide on the health of a national sample of 23-year-old subjects in 1981. Thorax501995764768
12. Schwartz J., Dockery D. W.Increased mortality in Philadelphia associated with daily air pollution concentrations. Am. Rev. Respir. Dis.1451992600604
13. Spix C., Heinrich J., Dockery D., Schwartz J., Volksch G., Schwinkowski K., Collen C., Wichmann H. E.Air pollution and daily mortality in Erfurt, East Germany, 1980–1989. Environ. Health Perspect.1011993518526
14. Dockery D., Pope A., Xu X., Spengler J. D., Ware J. D., Fay M. E., Ferris B. J., Speizer F. E.An association between air pollution and mortality in six U.S. cities. N. Engl. J. Med.329199317531759
15. Touloumi G., Pocock S. J., Katsouyanni K., Trichopoulos D.Short-term effects of air pollution on daily mortality in Athens—a time-series analysis. Int. J. Epidemiol.231994957967
16. Schwartz J.Air pollution and daily mortality: a review and meta-analysis. Environ. Res.6419943652
17. Pope A., Thun M., Namboodiri M., Dockery H. D. W., Evans J. S., Speizer F. E., Heath C. W.Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. Am. J. Respir. Crit. Care Med.1511995669674
18. Schwartz J., Morris R.Air pollution and hospital admissions for cardiovascular disease in Detroit, Michigan. Am. J. Epidemiol.14219952325
19. Burnett R., Dales R., Krewski D., Vincent R., Dann T., Brook J. Associations between ambient particulate sulfate and admissions to Ontario Hospitals for cardiac and respiratory diseases. Am. J. Epidemiol.14219951522
20. AirPaca. Association de surveillance de la qualité de l'air agréée par le ministère de l'environnement – URL – <http://www.airpaca.org/>
21. Wikipedia. R (programming language) – URL – <https://en.wikipedia.org/wiki/R_(programming_language)>
22. Correlation and dependence – URL – <https://en.wikipedia.org/wiki/Correlation_and_dependence>
23. Grigoreva I., Gorianin S., Manai O., Rhazadi W. Big Bridge – SE : Projet Big Data Santé et Environnement dans la ville de Nice en partenariat avec l’IMREDD

# Annexes

*Architecture of the system:*

