MetOp Satellites Data Processing for Air Pollution Monitoring in Morocco

Mohamed Akram Zaytar, Chaker El Amrani
Faculty of Science and Technology in Tangier, Abdelmalek Essaadi University, Laboratory of Informatics Systems and Telecommunications (LIST), Morocco

Article Info

Article history:

Received Jun 9, 2018
Revised Jul 1, 2018
Accepted Jul 20, 2018

Keyword:

Air pollution
Data aggregation
Data analysis
Data processing
Deep learning

ABSTRACT

This paper presents a data processing system based on an architecture comprised of multiple stacked layers of computational processes that transforms Raw Binary Pollution Data coming directly from Two EUMETSAT MetOp satellites to our servers, into ready to interpret and visualise continuous data stream in near real time using techniques varying from task automation, data preprocessing and data analysis to machine learning using feedforward artificial neural networks. The proposed system handles the acquisition, cleaning, processing, normalizing, and predicting of Pollution Data in our area of interest of Morocco.

Corresponding Author:

Mohamed Akram Zaytar,
Department of Informatics,
Laboratory of Informatics Systems and Telecommunications (LIST),
Abdelmalek Essaadi University, PO. Box 416, Tangier, Morocco.
Email: MedAkramZaytar@gmail.com

1. INTRODUCTION

Over the last decade, Air Pollution environmental threats significantly increased [1]-[4], and Climate change effects became many and wide ranging [5]. There is no doubt that excessive levels of air pollution are causing a lot of damage to human and animal health as well as to the wider environment. For these reasons, careful scientific research and monitoring of air pollutants is a necessity that must be exercised with a great deal of attention and precision.

Nowadays, as much as we want to quickly evaluate and conclude from existing pollution and climate data, most of the problems we face center around preparing, cleaning, processing, and transforming the large amounts of raw environmental data we receive from satellites in near real time. In our case, the raw data takes multiple primitive formats such as BUFR (Binary Universal Form for the Representation of meteorological data), GRIB 2, HRIT/LRIT, HRPT/LRPT. in this paper, we are going to present a system for processing BUFR based binary files coming directly from the satellite's sensors and transform it into a data set that is ready for data analysis specific tasks likes inference and visualisation.

The main source of the data we process is EUMETSAT. EUMETSAT is an intergovernmental operational satellite agency with a total of 30 European Member States. The organization's mission statement is to gather accurate and reliable satellite data on weather, climate and the environment around the clock, and to deliver them to its member and cooperating states, international partners, and to users world-wide [6].

The data we are most interested in comes directly from a type of satellites named Metop. Metop is a series of three polar orbiting meteorological satellites, we currently get data from two of them, Metop-A and Metop-B, they both are in a lower polar orbit, at an altitude of approximately 817 kilometres, they provide detailed observations of the global atmosphere, oceans and continents. The last satellite, Metop-C, is planned to be launched in 2018.

The system transforms the data from its primitive BUFR format, which is a binary data format maintained by the world meteorological organization, to comma separated files (CSV). The BUFR format is a somewhat controversial and a hard-to-work-with data format because of the difficulty of manipulating and experimenting with its encoded values.

Our proposed solution is a software system composed of multiple stacked layers. The first one decompresses and processes the BUFR binary data, decodes it, structures and combines its decoded messages under the CSV (comma separated values) format, and finally normalizes it. Because deep learning models are used in different climate related problems [7]-[9], we trained and measured the performance of an ANN based architecture when filling missing value points and interpolating new ones. The system produces a near continuous data stream on the 2-D surface of our area of interest.

The software solution proposed by this paper is a system that can be directly plugged into the endpoints of the near real time data stream, it will allow for fast experimentation and visualization of already processed raw data points coming directly from the Metop-X satellites series, it will also result in space and time reduction and optimization since it focuses on interest areas, we look forward for our solution to further improve and accelerate the research process done on top of the EUMETCAST data stream pipeline.

2. PROCEDURE

2.1. Data Processing

The following figure demonstrates the procedure taken to pre-process and normalize the data:

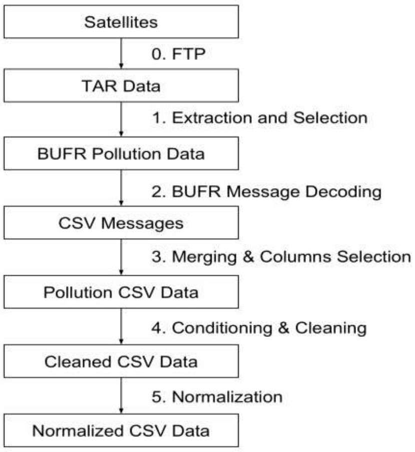


Figure 1. Decoding BUFR Data to Comma separated merged messages

In the first step, the system gets the raw tar files through the FTP protocol, after extracting the compressed files we get multiple Binary BUFR files which follow a strict naming convention in the following form (INSTRUMENT_ID-PRODUCT_TYPE-PROCESSING_LEVEL-SPACECRAFT_ID-SENSING_START-SENSING_END-PROCESSING_MODE-DISPOSITION_MODE-PROCESSING_TIME) that corresponds to multiple important variables such as the instrument identifier, orbit, and time frame, the system filters the data down to get pollution files in the time the satellite is scanning the area of interest using regular expressions on the names of the extracted files (under the pollution code name of ”TRG”). What we finally get are multiple BUFR pollution files corresponding to the area of interest that are ready to be decoded.

In the second step, the system uses a third party software solution named BUFRExtract [10] to decode the BUFR files into bulks of exported messages, each message containing a description of its columns and the values in each one in a text file format. In the third step, the system performs fast merge/selection techniques to combine all of the messages into two comma separated files corresponding to the scanning timeframe, one for the Metop-A satellite and the second for Metop-B. Both CSV files contain the following columns of interest:

Table 1. Extracted Features

No.	Feature	Unit
1	Year	Integer
2	Month	Integer
3	Day	Integer
4	Hour	Integer
5	Minute	Integer
6	Second	Integer
7	latitude	DEGREE
8	Longitude	DEGREE
9	CH4 Density	km.m-2
10	CO2 Density	km.m-2
11	N2O Density	km.m-2

After exporting the necessary values into multiple structured CSV files, the system groups rows by location points and the exact date (Year-Month-Day- Hour-Minute-Second) and applies the mean function on the pollutant values to take the average of possible redundant measurements. In the fourth step, the system deals with cleaning data points that are substantively unreasonable using logical conditions on data points of CH_4, CO_2 and N_2O using Z-scores.Lastly, the system normalizes all pollution points into values in [−1, 1] to accelerate convergence in the training phases, using the following formula for all three numerical variables:

$$X_i^{\ j} \leftarrow \frac{X^j{}_i - Mean(X^j)}{MAX(X^j)\ -\ MIN(X^j)}\ \forall j \in \{1,2,3\}, i \in \{1,\ldots,m\}$$

As a general description of the process, each half an hour, the system receives one compressed tar file through the servers' end points, the system automatically decompresses the file into BUFR BIN, selects files corresponding to the area of interest, and decodes them using a third party library (BUFRextract) to the corresponding messages and turns them into two CSV files containing all of the values of interest in near real time, this results in a considerable reduction in the dimensionality of the data and the space it normally occupies.

The second part of the system fills the missing values in the 2-D surface of interest and also generates new data points using algorithmic search and a neural network architecture to get a near continuous data stream output that is ready for exploration, visualisation, and interpretation.

2.2. Intelligent Interpolation

The prediction of missing values is based on three pre-trained Feed-Forward Fully Connected Neural network models fit to fill the missing values in the 2-D surface of our interest for the three pollutants (CO2, CH4, and N2O), and the general architecture of our ANNs is as shown in Figure 2.

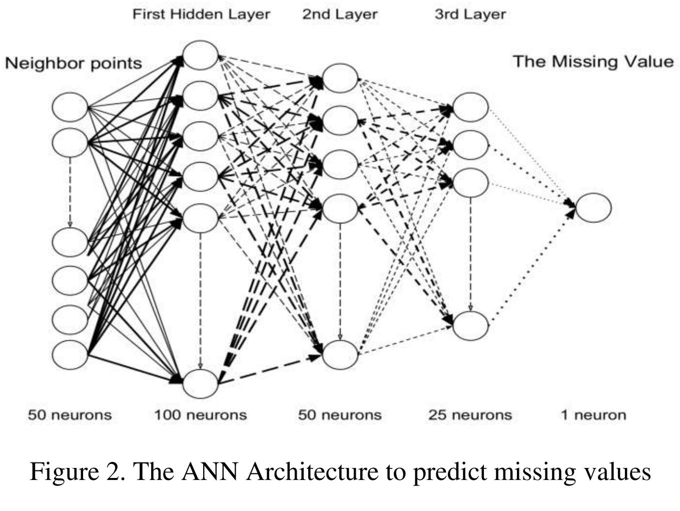


Figure 2. The ANN Architecture to predict missing values

As an activation function for our model, we chose the rectifier function. The general process in which selected missing points are predicted (or not), is shown as Figure 3:

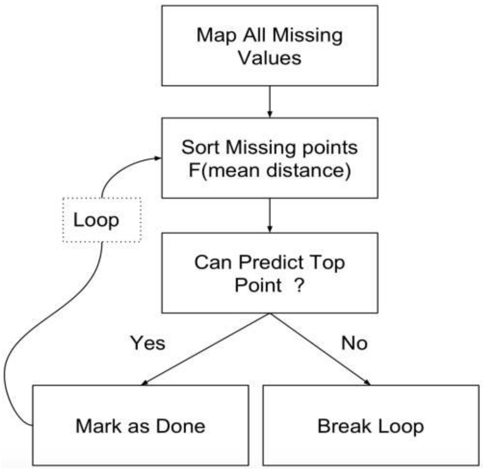


Figure 3. The Filling missing points Procedure

The system predicts missing values following this procedure:

1. Map all missing values with the nearest 100 neighbor values.
2. Sort points in function of the number of neighboring missing points and the average distance, giving a new score for each missing point in the form of : $S = avg * num$
3. If the Top missing point's average distance from all neighbor points is greater than 50 km, or if there is no top missing point, break the loop and finish the process.
4. If the average distance is less than 50km, predict the missing point using the ANNs models and mark the point as done and loop back to step 2.

The system automatically loops over these steps until all missing values are filled (for possible predictions), the system repeats this whole procedure for the three pollutants of interest.

3. RESEARCH METHOD

3.1. Data Description

The first Dataset used in this study was collected in the form of bulks of BUFR message files coming directly from two satellites, Metop-A and Metop-B, and precisely from the Infrared atmospheric sounding interferometer (IASI) sensor, which is composed of a Fourier transform spectrometer and an associated integrated Imaging Subsystem (IIS). The Fourier transform spectrometer provides infrared spectra with high resolution between 645 and 2760cm-1 (3.6m to 15.5m).

The main goal of IASI is to provide atmospheric emission spectra to derive temperature and humidity profiles with high vertical resolution and accuracy. Additionally it is used for the determination of trace gases such as ozone, nitrous oxide, and carbon dioxide, as well as land and sea surface temperature and emissivity and cloud properties.

IASI measures in the infrared part of the electromagnetic spectrum at a horizontal resolution of 12 km over a swath width of about 2, 200km. With 14 orbits in a sun-synchronous mid-morning orbit (9:30 Local Solar Time equator crossing, descending node) global observations can be provided twice a day (every 12 hours), the satellites take around 25 minutes to scan The area of interest, we get pollution data from points approximately 20 km apart from each other.We constructed the second dataset from already preprocessed data points in the goal of training, testing, and validating our neural network models and solve the problem of filling missing data points and interpolating new points in the selected area of interest.

3.2. Intelligent Interpolation

We generated new empty points values in which all of the points in the area of interest are distanced from each other by 5km, the system then intelligently interpolate all empty points.

3.2.1. Data Collection

We collected 150 Gigabytes of preprocessed data or the equivalent of around 800 million data point to build an intelligent model capable of predicting missing pollutant values. After collecting the data set, we ran a general statistic on missing data points and we present the following results based on the sampled dataset as shown in Table 2.

Table 2. Missing Values in the Data Set

No.	Pollutant	% of missing values
1	N2O	0.003 %
2	CO2	72 %
3	CH4	71 %

3.2.2. Training and testing data

The data was transformed into a table where the features are the 50 nearest points and the target variable is the data point used to train the artificial neural network, the distance between the target point and the furthest point set to a maximum and the same conditions we applied when selecting valid missing points were applied when transforming the data. When training the model to predict new point values (Interpolation), the system adds new points (marked missing) so that every point has a point at least 5 km near the next one, after creating new grids of 2-D points, training sets were selected based on availability of the neighboring points.

3.2.3. Training

6 Models consisting of 3 fully connected hidden layers with 100, 50, and 25 neurons respectively were used, the first 3 models constructed to predict missing and corrupted values and the last 3 were trained to interpolate new point values, the training details are given as:

a. All of the neurons parameters were randomly initialized using the uniform distribution between −0.1 and +0.1.
b. The Mini-Batch gradient Descent was used to optimize the parameters.
c. A learning rate of ε=0.001 was chosen.
d. Batches of 1024 samples and 200 epochs were trained.

3.2.4. Validation

For the validation to be efficient, we used 10-fold cross validation technique, splitting the data set into multiple training and testing sets to verify the efficiency of the trained models and to avoid overfitting.

3.2.5. Interpolation Method

The system uses three pre-trained neural network models to predict newly generated points and interpolate the whole surface. The process is similar to the procedure of predicting missing values, however, the system doesn’t set a threshold on the average of distances in order to break the loop of predictions. It predicts and fills all new data points at a fixed neighbouring distance of 5km, the following graph demonstrates the process as shown in Figure 4.

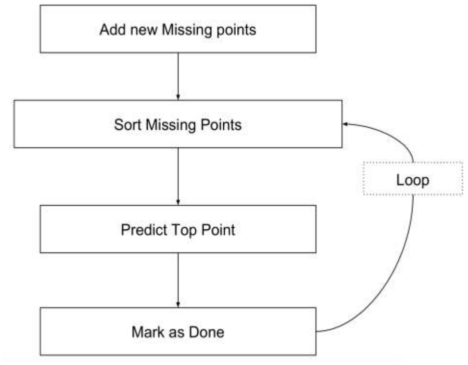


Figure 4. Interpolation by Feed-Forward Neural Networks

The system predicts all points and updates the sorted list of missing points as it goes until filling all of the missing values, the only difference that this model have with the previous one is that it does not have a criteria for whether to predict a missing point or not.

3.3. Benchmarking

To measure the performance of our ANN-based interpolation system, we benchmark its predictions against two state of the art algorithmic methods of spatial interpolation, Kernel smoothing and Kriging.

3.3.1. Kernel Smoothing

A kernel smoother is a statistical technique for estimating a real valued function f(X) (X ∈ Rp) by using its noisy observations, when no parametric model for this function is known. The estimated function is smooth, and the level of smoothness is set by a single parameter. To put it in mathematical terms, the idea of the nearest neighbor smoother is the following. For each point Xi, take N nearest neighbors and estimate the value of F(Xi) by averaging the values of these neighbors. This type of interpolation is most appropriate for low-dimensions (p < 3) (the dimensionality curse [11] is one reason for that). Actually, the kernel smoother represents the set of irregular data points as a smooth line or surface, in our case (2-D surface) this is a perfectly reasonable solution. One way to fill these points would be to use Scipy's [12] (precisely scipy.interpolate.Rbf) implementation of Radial Basis Function interpolation which is intended for the smoothing/interpolation of scattered data.

3.3.2. Gaussian Process Regression or Kriging

Kriging or Gaussian process regression is a method of interpolation in which the interpolated values are modelled by a Gaussian process governed by prior covariances, as opposed to a piecewise-polynomial spline chosen to optimize smoothness of the fitted values. Under suitable assumptions on the priors, kriging gives the best linear unbiased prediction of the intermediate values. Interpolating methods based on other criteria such as smoothness may not yield the most likely intermediate values. The method is widely used in the domain of spatial analysis and computer experiments. The technique is also known as Wiener Kolmogorov prediction. We'll compare the results of Kriging interpolation on the dataset using the Gaussian Process Regression implementation in the Python's scikit-learn library.

3.4. Hardware

A Python implementation of the deep neural network architecture with hidden layers of 100, 50, 25 number of neurons (respectively), Google's TensorFlow [13] library was used to build and train the model. An NVIDIA Tesla K80 single GPU device, with 4992 CUDA cores, 24 GB of GDDR5 memory, and 480 GB/s aggregate memory bandwidth was used to train the neural network models.

4. RESULTS AND DISCUSSION

The resulting solution is a system composed of three layers of processes, the first layer decompress, decode, and normalizes the data. the second layer is a three ANN stack to fill in the missing pollutant values, and lastly the final layer which is composed of another stack of neural network models to interpolate new data points in our area of interest.

4.1. Data Processing

The decompressing, decoding, merging, cleaning and normalizing of Raw BUFR data result in a considerable reduction in resources. Since our algorithm runs in linear time, and considering the volume of data the system processes at each step, a simple computer configuration (4 Gigabyte RAM, 4 cores with no parallelism) result in the following durations as shown in Figure 5.

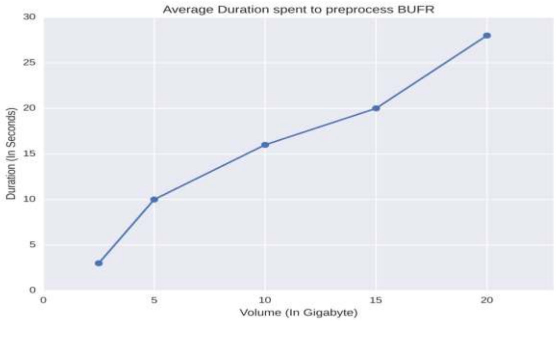


Figure 5. Average duration of the preprocessing stage

These tests were conducted multiple times for each volume category, to ensure high precision. We conclude that the system scales pretty well and can process large volumes of data (up to terabytes per hour) in relatively short duration of time.

4.2. Intelligent Interpolation

The resulting surface of interest is a 1887 by 1776 km2 rectangle, the system predicts a maximum number of 123,568 points, the following figures showcase examples of predictions in a fixed date, using kriging, smoothing and our neural network model as shown in Figures 6, 7, 8.

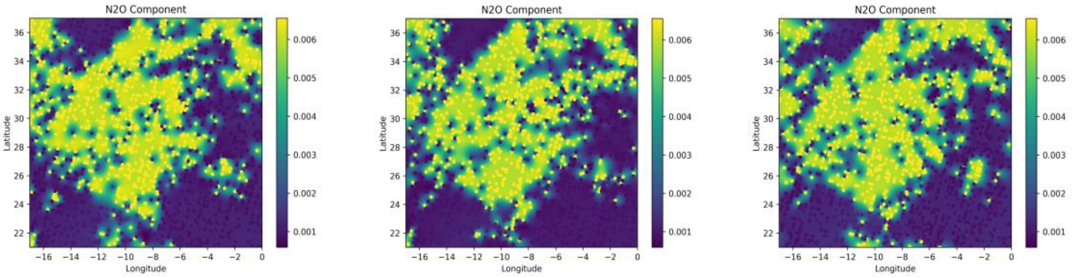


Figure 6. Interpolation visualisations of N2O

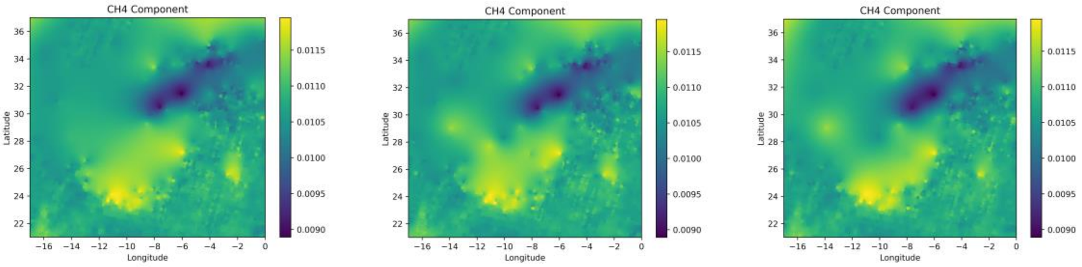


Figure 7. Interpolation visualisations of CH4

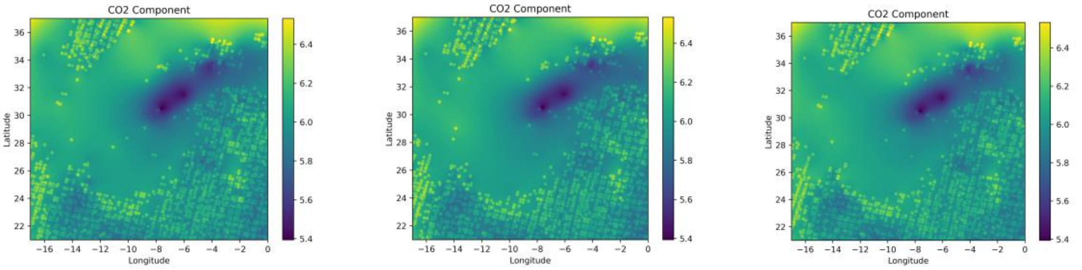


Figure 8. Interpolation visualisations of CO2

In the above figures, the rounded markers represent a known sample of pollution data points, and the interpolated surface represent the the resulting predictions. We got the following training results after cross validating the models as shown in Figures 9 and 10. As expected, the system produces better results when filling missing values, and generally worse results when filling in new data points. but when comparing interpolated data using the 3 methods, we find interesting results, the following graph showcases the results of comparisons.

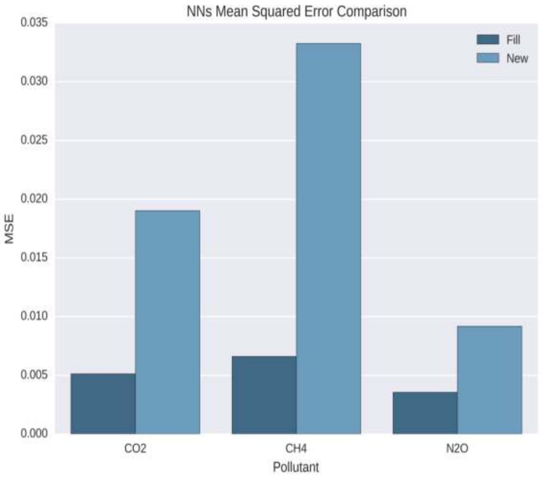


Figure 9. Comparing the accuracy of different models using MSE

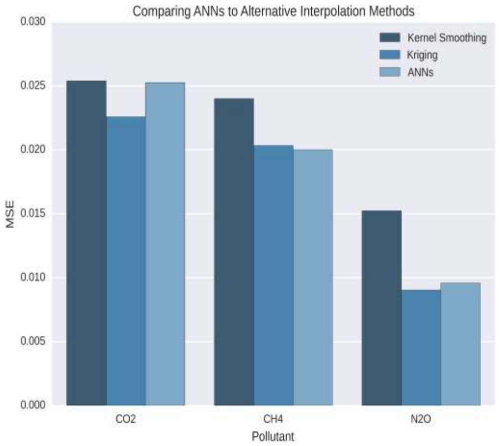


Figure 10. Accuracy of the suggested Interpolation Methods

4.3. Discussion

As we can see from the results, the optimal interpolation technique is generally better than our trained neural network models, however, in the case of N2O and CH4 we can say that our model is competitive with the other two classical 2-D interpolation algorithms, and since we had 70% missing data, that opens the possibility of better performance with greater volumes of data, if trained on larger volumes of data, our system can make better predictions and therefore introduce an optimal solution and a competitor to the kriging or smoothing interpolation algorithms.

5. CONCLUSION

At the present time, the size, variety and complexity of raw data is huge and continues to increase every day. The use of data processing systems to store, process, and analyze data streams has changed how we discover and visualise big data in general. In this paper, we presented a software solution composed of multiple stacked layers of subsystems that transform and process considerable volumes of raw pollution data in near real time, taking the data from its native compressed format to a structured, cleaned, normalized, and continuous data stream that is light and easy to experiment with.

In the future, significant challenges and problems concerning Big Environmental Data must be addressed by the industry and academia, current work on topics ranging from utilizing AI for plant monitoring [14], working on social awareness concerning climate change [15, 16], and the use of biological methods [17] to fight climate change is important. But new challenges to tackle are in the field of environmental data science, future work focused on how to build new environmental data learning paradigms, scientific computing environments, and an all around better infrastructure for pollution monitoring is a necessity for all of us.

ACKNOWLEDGEMENTS

The authors are thankful to the Ministry of Higher Education and Scientific Research, and the National Centre for Scientific and Technical Research (CNRST) for funding this study, under the project: PPR/2015/7.

REFERENCES

[1] Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., & Pozzer, A. “The contribution of outdoor air pollution sources to premature mortality on a global scale”. *Nature*, 2015; 525(7569): 367-371.

[2] Aguscik, A., Ikob, R., & Putra, S. A. “The Level of Malondialdehyde in People Exposed to Air Pollution”. *International Journal of Public Health Science (IJPHS)*, 2017; 6(1): 99-103.

[3] Hajat, A., Allison, M., Diez-Roux, A. V., Jenny, N. S., Jorgensen, N. W., Szpiro, A. A., & Kaufman, J. D.. “Long-term exposure to air pollution and markers of inflammation, coagulation, and endothelial activation: a repeat-measures analysis in the Multi-Ethnic Study of Atherosclerosis (MESA)”. *Epidemiology (Cambridge, Mass.),* 26(3), 310.

[4] Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z. J., Weinmayr, G., Hoffmann, B., ... & Vineis, P. "Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project". *The Lancet*, 2014; 383(9919): 785-795.

[5] Pio, D. V., Engler, R., Linder, H. P., Monadjem, A., Cotterill, F. P., Taylor, P. J., ... & Salamin, N. "Climate change effects on animal and plant phylogenetic diversity in southern Africa". *Global Change Biology,* 2014; 20(5), 1538-1549.

[6] EUMETSAT, http://www.eumetsat.int/website/home/AboutUs/index.html, 13 05 2017.

[7] Mohamed Akram Zaytar, Chaker El Amrani, "Sequence to Sequence Weather Forecasting with Long Short-Term Memory Recurrent Neural Networks." *International Journal of Computer Applications*, 2016; Volume 143 - No.11.

[8] Jyothi, M. N., Dinakar, V., Teja, N. S. R., & Kishore, K. N. "NARX Based Short Term Wind Power Forecasting Model". *Indonesian Journal of Electrical Engineering and Computer Science*, 2015; 15(1): 20-25.

[9] Khalel, S. I., Rahmat, M. F., & Mustafa, M. W. B. "Sensoring Leakage Current to Predict Pollution Levels to Improve Transmission Line Model via ANN". *International Journal of Electrical and Computer Engineering*, 2017; 7(1): 68.

[10] BUFR File Support Software, http://www.elnath.org.uk/, 13 05 2017.

[11] Bach, F. "Breaking the curse of dimensionality with convex neural networks". *Journal of Machine Learning Research*, 2014; 18(19): 1-53.

[12] Jones, Eric, Travis Oliphant, and Pearu Peterson. "SciPy: open source scientific tools for Python.", (2004).

[13] Abadi, Martn, et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems.", arXiv preprint arXiv:1603.04467 (2016).

[14] Ilamathi, P., Selladurai, V., & Balamurugan, K. "Predictive modelling and optimization of nitrogen oxides emission in coal power plant using Artificial Neural Network and Simulated Annealing". *IAES International Journal of Artificial Intelligence*, 2012; 1(1): 11.

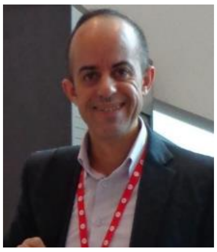
[15] Sulistyawati, S., & Nisa, I. "Climate Change and Health Teenager's Perceptions as a Basis for Interventions". *International Journal of Public Health Science (IJPHS)*, 2016; 5(3): 267-273.

[16] Montanaro, T., Corno, F., Migliore, C., & Castrogiovanni, P. "SmartBike: an IoT Crowd Sensing Platform for Monitoring City Air Pollution". *International Journal of Electrical and Computer Engineering (IJECE)*, 2017; 7(6): 3602-3612.

[17] Ghanemi, A., & Boubertakh, B. "Biological tools to deal with pollution: selected advances and novel perspectives". *International Journal of Public Health Science (IJPHS)*, 2014; 3(1): 57-62.

BIOGRAPHIES OF AUTHORS

Mohamed Akram Zaytar Obtained his Bachelor of Applied Mathematics and Computer Science degree from The Faculty of Science and Technology, Tangier in 2013. He received his master, in Computer Systems and Networking from the FSTT, in 2016. Currently, a PhD student at the FSTT. His primary research interest are in Data Science, Artificial Intelligence, Machine Learning, Cloud Computing, and Big Data.



Dr. Chaker El Amrani is Doctor in Mathematical Modelling and Numerical Simulation from the University of Liege, Belgium (2001). He joined Abdelmalek Essaadi University, Morocco in 2003. He is currently Chair of the Computer Engineering Department at the Faculty of Science and Technology, Tangier. He lectures distributed systems and is promoting High Performance Computing education in the University. His research interests include Cloud Computing, Big Data Mining and Environmental Information Systems. Dr. El Amranis research has been supported by national and international organisms. Dr El Amrani has served as an active volunteer in IEEE Morocco Section. He is currently Vice Chair of IEEE Communication and Computer Societies Morocco Chapter, and advisor of the IEEE Computer Society Student Branch Chapter at Abdelmalek Essaadi University. He is the NATO Partner Country Project Director of a real-time remote sensing initiative for early warning and mitigation of disasters and epidemics in Morocco.