Immissions profiling using AI

Marius Darie, Angela Călămar, Constantin Sorin Burian, Marius Kovacs, Alexandru Simion

National Institute for Research and Development in Mine Safety and Protection to Explosion - INSEMEX, G-ral Vasile Milea Street, no 32-34, Petroșani, România

marius.darie@insemex.ro

Abstract. Industrial activity and not only, generates both emissions and immissions of pollutants into the atmosphere. Thus, their magnitude and dynamics will present a specific footprint for each pollutant. The article aims to identify immission profiles using tools specific to artificial intelligence applied to a wide set of recordings of environmental parameters. The first part briefly presents the issue of environmental protection and specific regulations at national and European level, and the second part showcases the database of environmental parameters and the theoretical model of data processing. The last part of the paper is dedicated to results obtained and their analysis which shows the presence of emissions patterns (profiles) measured in different locations and time periods. Seasonality and its impact on emission profiles were also analysed. On this occasion, the use of the Hurst exponent allowed the segregation of various time series of data based on the resulting memory interpreted as a measure of immissions’ persistence. Jumps in the temporal dimension of values allowed the anonymous association of immissions with different locations. Analysis of the topology of clusters associated with immission profiles highlighted the presence of two types: rare clusters and dense clusters. Rare clusters can be associated with immission having accidental dynamics and dense clusters can be associated with systematic immissions. Use of the framed method allows for a classification of pollutants resulting in increased chances of solving the environmental impact.

1. Introduction
Following the EU’s adherence to the Paris Agreement of 2016 and once the Energy Union Strategy was published, the Union has taken on an important role in combating climate change, through its 5 main dimensions: energy security, decarbonisation, energy efficiency, internal energy market and research, innovation and competitiveness [1].

The European Union has set energy and climate targets for 2030 as follows:
- The target for reducing domestic greenhouse gas emissions by at least 40% compared to 1990, by 2030;
- The target for renewable energy consumption of 32% in 2030;
- The target for improving energy efficiency by 32.5% in 2030;
- The goal of interconnecting the electricity market at a level of 15% by 2030.

Thus, our country needs to provide more reliable and transparent measurements of industrial emissions and immissions in certain areas.
Sources of anthropogenic pollution cover a wide range of physical and chemical activities, which are the main contributors to urban air pollution. Air pollutants result from electricity production, millions of vehicles, materials dumped by population and manufacturing processes of many products necessary for mankind. The first five major classes of pollutants are: particles, sulfur dioxide, nitrogen oxides, volatile organic compounds and carbon monoxide. Hundreds of millions of tons of polluting air are generated annually. In general, each category of pollution source has a different contribution to raising the level of one of the five main classes of pollutants.

Atmosphere is the largest vector in pollutants’ propagation, its effects being felt from local to global level on all biotic and abiotic elements. Thus, air quality monitoring plays an essential role in the integrated monitoring system, being a key element in underlying control strategies. It should be mentioned that the notion of "measurement" must be understood in the sense of knowing and mastering present and future sources of pollution, in order to protect the environment at different spatial scales.

Knowledge of air pollution mechanisms, involves the study of four categories of phenomena, namely:

• Emissions of air pollutants related to natural and anthropogenic sources;
• Transfer of pollutants from source to receiver (transmission);
• Immissions, respectively temporal spatial distribution and level of pollutant concentrations in the atmosphere;
• Effects of pollution on human health, the environment, buildings and materials.

The current paper studies the distribution of dust and gases’ immissions (CO, NO2, SO2, CO, O3) in a known area, to identify whether certain patterns (profiles) of measured emissions in various locations and time periods can be achieved.

Immissions, respectively the temporal-spatial distribution and the level of pollutants’ concentration in the atmosphere, are monitored by measurements (and mathematical modelling) of concentrations, in order to attain knowledge of the state and evolution of air quality. This component is achieved by measurements and mathematical modelling. The objectives of immission monitoring derive from the need to attain knowledge of the state and evolution of air quality and its protection [2].

Air quality is monitored by surveillance networks consisting of stations for impact monitoring and background stations (background monitoring), located at a great distance from pollution sources, which determine the background levels of atmospheric constituents so that the influence of anthropogenic activities can be assessed [2].

Monitoring air quality in a geographical area requires the development of a coherent program to address all pollution issues in the entire area, where pollution sources may or may not be found, which also involves the responsibility of local authorities [3].

It should be noted that limitation of air pollution has evolved and is constantly evolving, depending on better knowledge of the harmful effects of various air pollutants. The main elements that have led and continue to lead to continuous revision of norms are climate change and acid rain associated with air pollutants.

Law 104/2011 on quality of the environment, sets the maximum allowed immission limit values for pollutants [4].

The current paper considers emissions for the following gases: CO, NO2, SO2, CO, O3.
Sulfur oxides (SO2 and SO3) result from combustion processes, the most important being sulfur dioxide (SO2) because a much higher proportion is produced. The amount of SO2 produced is closely related to the sulfur content of the fuel used [5].

Nitrogen oxides (NO and NO2) also result from combustion processes, resulting from the combination of nitrogen in the atmosphere with oxygen in the air. The reaction takes place at high temperatures, proportional to its concentration (at above 1500°C a proportion of about 0.25% nitrogen oxides is formed, and at about 2000°C a proportion of about 1.75% is formed).

Carbon monoxide (CO) is mainly formed by incomplete combustion of fossil fuels. Carbon monoxide can accumulate at dangerous levels, especially during the calm of winter and spring (it is much more chemically stable at low temperatures), when the burning of fossil fuels reaches a maximum [6].

At relatively low concentrations, it affects the central nervous system, weakens the heart rate, thus reducing the volume of blood distributed in the body, reduces visual acuity and physical capacity, causes irritability, migraines, rapid breathing, reduces the ability to concentrate.

Ozone (O3) is a very oxidizing gas, very reactive, with a pungent odour. It concentrates in the stratosphere and provides protection against life-threatening UV radiation. Ozone present soil acts as a component of "photochemical smog". It is formed by a reaction involving, in particular, nitrogen oxides and volatile organic compounds. The concentration of ozone in soil causes respiratory tract irritation and eye irritation.

Suspended particles can be defined as solid matter, whose effective diameter is larger than that of a molecule, but less than 100 μm.

Suspended particles (Particulate Matter = substance in the form of particles – PM10 and PM2.5) are the amount of substance, in the form of particles, present in a given volume.

The usual units are micrograms of particulate matter, per cubic meter of air (μg / m3). Since smaller particles have a greater harmful effect on human health than larger ones, usually only those that have a certain diameter or smaller than that are collected and reported. For example, a high concentration of particles with diameters between 0.1 μm and 1 μm in the air, produces haze. In general, dust in the urban atmosphere comes from car traffic, burning fossil fuels, industrial activities, etc.

Larger particles are generally filtered into nose and throat through cilia and mucus, and particles smaller than 10 micrometers can settle into lungs and cause health problems. The size of 10 micrometers is not a strict boundary between breathable and non-breathable particles, but it was agreed as monitorization limit for airborne particles by most regulators. So, particles below 10 microns, (PM10), can enter bronchioles and alveoli of the lungs and can trigger bronchoconstriction.

Most of the measurements were performed in the yard of INCD INSEMEX Petrosani, to the right from the main gate.

2. The data set
The data set under analysis represents a set of 16,020 records structured in 12 fields as follows: Timestamp, co_gas, o3_gas, p_raw, rh_raw, t_raw, pm1_raw, pm10_raw, pm25_raw, no_gas, no2_gas, so2_gas.

Measurement units for recorded values are: t [°C], p_raw [mbar], p* _raw [g/m^3], all others [ppm].

The time interval covered by the recordings is 2019-12-09 14:01:00 to 2021-01-15 11:00:00.
Distribution of recordings is shown in the diagram in Figure 1.

![Figure 1. Time distribution of recordings](image)

Dispersion of the decimal logarithm of recorded values (interquartile range - box plot diagram) is shown in the diagram in Figure 2.

![Figure 2. Box plot diagram of recordings](image)

3. **Processing and analysis of the data set**

Distribution of the number of missing values in the records, for each measurement channel, is shown in Figure 3.
The channels with most missing records were nitrogen oxides, followed by sulfur dioxide and ozone. The assignment of a location based on the jumps in the time field of recordings is shown in the diagram in Figure 4.

Recordings containing at least one missing field were then filtered. The number of recordings remaining after extracting those that have missing values (NA) is 5595.

To filter the noise in the recordings, the moving average was calculated for 60 successive recordings (a time interval of 180 minutes).
The Hurst exponent was similarly calculated, to highlight the variability and memory of immissions. Thus, the suffixes “.ma” and “.Hurst” were added to the names of resulting fields.

The Ward D2 clustering method applied to the standard set of records was used to identify immission profiles. The diagram in Figure 5 shows the resulting dendrogram.

![Dendrogram of recordings](image1)

**Figure 5.** Dendrogram of recordings

The number of clusters (profiles) was established following the maximum uniformity of population size, distributed on clusters. The physical environmental parameters (pressure, humidity, temperature) and the assigned location based on time jumps of records were omitted from the clustering process. For this case, a recommended number of 5 clusters (profiles) resulted. (Figure 6)

![Optimal number of clusters](image2)

**Figure 6.** Optimal number of clusters
Boxplot diagrams for each of the immission parameters are shown below. The suffix “.ma” shows the values averaged at 3 hours and the suffix “.Hurst” shows the values of the corrected R over S Hurst exponent calculated for 3-hour intervals.

Analysis of the principal components, which this time also included the physical environmental parameters (pressure, temperature and humidity) showed that the evolution of immissions is characterized by preponderance of the first 3 components, based on variations in the specific variances of principal components (Figure 7).

![Figure 7](image)

**Figure 7.** Variations of variants specific to principal components

Distribution of records for the first two principal components with the highlighting of the resulting clusters is shown in Figure 8.

![Figure 8](image)

**Figure 8.** Distribution of records for the first two principal components
Analysis of the rotation matrix resulting from the analysis of the principal components showed that an important role is played by carbon monoxide and dust (weight 0.4) in the first main component (Figure 9).

The second main component is very much influenced by the exponent Hurst for dust (weight 0.5). This underlines the importance of considering the Hurst exponent. (Figure 10)

![Figure 9. Rotation matrix values for the first main component](image1.png)

![Figure 10. Rotation matrix values for the second main component](image2.png)
The third main component is determined by the emissions of nitrogen monoxide and atmospheric pressure (0.4) and ozone (-0.4). (Figure 11)

Characterization of immission profiles obtained based on the clustering process is shown in Table 1.

Profile 0 represents the calculated median for the entire filtered data set.

**Table 1.** Relative values of immissions and environmental parameters relative to profile 0

| Parameter   | Profile [%] | 0 (median) | 1 | 2     | 3     | 4     | 5     |
|-------------|-------------|------------|---|-------|-------|-------|-------|
| co_gas.ma   |             | 373.41     | 19.44 | 291.14 | 141.28 | 153.24 | 85.42 |
| o3_gas.ma   |             | 15.48      | 168.09 | 130.05 | 22.28  | 278.78 | 82.21 |
| p_raw.ma    |             | 942.57     | 104.57 | 100.48 | 100.74 | 93.96  | 100.12|
| rh_raw.ma   |             | 98.85      | 73.31  | 72.60  | 190.02 | 93.87  | 105.32|
| t_raw.ma    |             | 0.37       | 6754.05 | 67.39  | 31.53  | -6.03  | 193.75|
| pm1_raw.ma  |             | 18.86      | 29.22  | 290.38 | 73.44  | 249.45 | 57.90 |
| pm10_raw.ma |             | 21.73      | 49.24  | 204.02 | 68.76  | 211.46 | 57.69 |
| pm25_raw.ma |             | 19.75      | 33.27  | 261.04 | 73.53  | 238.78 | 57.95 |
| no_gas.ma   |             | 6.58       | 2206.23 | 5.29   | 206.77 | 34.63  | 98.73 |
| no2_gas.ma  |             | 14.62      | 536.46 | 9.01   | 203.68 | 97.43  | 98.00 |
| so2_gas.ma  |             | 1.44       | 967.36 | 2.66   | 148.65 | 558.18 | 22.80 |
| co_gas.Hurst|             | 0.98       | 98.98  | 106.19 | 99.03  | 99.02  | 82.18 |
| o3_gas.Hurst|             | 1.01       | 100.00 | 107.92 | 93.58  | 101.96 | 89.42 |
| pm1_raw.Hurst|            | 0.95       | 87.37  | 120.48 | 93.00  | 106.45 | 84.85 |
| pm10_raw.Hurst|           | 0.94       | 89.36  | 119.05 | 91.00  | 108.79 | 84.85 |
| pm25_raw.Hurst|           | 0.95       | 96.84  | 107.61 | 93.94  | 106.45 | 84.85 |
| no_gas.Hurst |             | 0.69       | 178.26 | 53.66  | 119.70 | 84.81  | 95.52 |
| no2_gas.Hurst|             | 0.72       | 168.06 | 59.50  | 123.61 | 75.28  | 86.57 |
| so2_gas.Hurst|             | 0.84       | 148.81 | 65.60  | 95.12  | 125.64 | 84.69 |
The column of profile 0 contains the median of values for the entire filtered data set and the columns of profiles 1 to 5 show a relative value as a percentage of profile 0.

The analysis of the location-profile contingency matrix showed that locations 9 and 10 are well represented in profile 3 and location 11 in profiles 2, 4 and 5. Profile 1, with very few records, is found only in location 5. Locations 0 to 4 were filtered due to missing values in the recordings.

| Location | Profile 1 | Profile 2 | Profile 3 | Profile 4 | Profile 5 |
|----------|----------|----------|----------|----------|----------|
| 5        | 0        | 0        | 0        | 0        | 0        |
| 6        | 0        | 1        | 0        | 0        | 0        |
| 7        | 0        | 1        | 0        | 0        | 0        |
| 8        | 0        | 1        | 0        | 0        | 0        |
| 9        | 0        | 0        | 465      | 5        | 0        |
| 10       | 0        | 0        | 1339     | 51       | 0        |
| 11       | 0        | 2297     | 0        | 897      | 536      |

4. Conclusions
Incomplete recordings led to a significant truncation of the set of immission measurements.

The channels with the most missing values were the nitrogen oxide channels, followed by the sulfur dioxide and ozone channels.

Preponderance of carbon monoxide and dust in the first main component highlighted the presence of proximal sources of emissions from coal combustion.

Use of the Hurst exponent brought a benefit in the clustering process which is also highlighted in the high contribution in the second main component.

Use of maximum uniformity of profile populations as a criterion, allowed the objective choice of the optimal number of profiles.

The data processing approach allowed unsupervised processing of the data set for the purpose of obtaining immission profiles.

Acknowledgements
The data used came from recordings of environmental parameters measured by the auto-laboratory, endowment of INCD INSEMEX, in the time interval 2019-12-09 14:01:00 - 2021-01-15 11:00:00.

For performing the calculus and plotting the diagrams were used the R language and R Studio environment [7], [8] and [9]. Along the R Studio environment were used accelerometry [10], pracma [11], and readxl [12] packages.

References
[1] Integrated National Plan for Energy and Climate Change 2021-2030, April 2020.
[2] “Atmospheric pollution - the major problem of society” - course notes developed by the PROGRAMMING SYSTEMS Laboratory for MODELING and SIMULATION”, 2018.
[3] G. Lăzăroiu, “Modern solutions for clean air”, 2015.
[4] LAW 104 OF 15th June 2011 on quality of environmental air, 2011.
[5] P. Ursu, D. P. Frosin, I. Bergea-Tatu et. al., “Atmospheric air protection”, Practical guide, 1978.
[6] V. Voicu, “Pollution control in industry - technical publishing house”, Bucharest, 2002.
[7] R Development Core Team R: “A language and environment for statistical computing”. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org., 2008.
[8] R Core Team R: “A language and environment for statistical computing”. R Foundation for Statistical Computing, Vienna, Austria. [Online] 2015. Available at: http://www.R-project.org.
[9] L. Henry and H. Wickham."rlang: Functions for Base Types and Core R and 'Tidyverse' Features". R package version 0.4.10. 2020. Available at: https://CRAN.R-project.org/package=rlang.
[10] D. R. Van Domelen “accelerometry: Functions for Processing Accelerometer Data”. R package version 3.1.2, 2018. Available at: https://CRAN.R-project.org/package=accelerometry.
[11] H. W. Borchers “pracma: Practical Numerical Math Functions”. R package version 2.3.3. https://CRAN.R-project.org/package=pracma, 2021.
[12] H. Wickham and J. Bryan “readxl: Read Excel Files”. R package version 1.3.1, 2019. Available at: https://CRAN.R-project.org/package=readxl.