Missing Values Interpolation in PurpleAir Sensor Data based on a Correlation with Neighboring Locations using KNIME Analytics Platform

Samir Omanovic*, Admir Midzic**, Zikrija Avdagic*, Damir Pozderac* and Amel Toroman**
* University of Sarajevo, Faculty of Electrical Engineering, Sarajevo, Bosnia & Herzegovina
** University of Bihac, Technical Faculty, Bihac, Bosnia & Herzegovina
samir.omanovic@etf.unsa.ba, admir.midzic@unbi.ba, zikrija.avdagic@etf.unsa.ba,
damir.pozderac@etf.unsa.ba, amel.toroman@unbi.ba

***Abstract* – Missing values handling in any collected data is one of the first issues that must be resolved to be able to use that data. This paper presents an approach used for missing values interpolation in PurpleAir particle pollution sensor data, based on a correlation of the measurements from the observed locations with the measurements from its neighboring locations, using KNIME Analytics Platform. Results of our experiments with data from five locations in Bosnia & Herzegovina, presented in this paper, show that this approach, which is relatively simple to implement, gives good results. All modeling and experiments were conducted using KNIME Analytics Platform.**

Keywords - interpolation; correlation; sensor data; KNIME; missing values; particle pollution (PM2.5)

I. INTRODUCTION

We are witnessing an increasing presence of data collection from many different sensors in various fields of application. These data should be analyzed and used to find certain relationships, rules, patterns, etc. which is not so easy, bearing in mind the volume and the quality of the data. One of the problems with the collected data is related to missing values. There is no straightforward answer how to deal with missing values in the collected data and various approaches are used [1].

Nowadays, there is more and more awareness of environmental pollution, so there are more and more activities in that field with an emphasis on pollution measurements using sensors like PurpleAir sensors. PurpleAir sensors are low-cost sensors for measurement of air quality monitoring. They use laser particle counters to provide real time measurement of temperature, humidity, PM1.0, PM2.5 and PM10. This paper is only about particle pollution (PM2.5) data. These sensors can be connected to a WiFi network. That way, the measurements from the sensor are accessible on that network. Additionally, users of PurpleAir sensors can register their sensors on the PurpleAir real-time map. The PurpleAir real-time map is a web application that displays a network of community owned PurpleAir sensors. It enables downloading and use of data for various research projects [2,3].

This research is supported by the Ministry of Education and Science of entity Federation of Bosnia & Herzegovina of the state Bosnia & Herzegovina.

One of the major issues related to the use of the collected data is that there are missing values (missing measurements). To be able to use data in a good way, it is important to solve that issue at the beginning of any research. Measurement errors and sensor failure are not so rare in data collection and are one of major reasons for missing measurement values [4].

The data collected from PurpleAir sensors are time-series data. There are many methods to interpolate time-series data and that way to solve missing values issue. In [5] is given good overview of the interpolation methods:

- deterministic like nearest-neighbor, polynomial, based on distance weighting, based on Fourier's theory, and others; and
- stochastic like regression methods, autoregressive methods, machine learning methods, methods based on data dynamics, and others.

This paper is focused on resolving the issue of missing PM2.5 values in data from PurpleAir sensor, using existing data from neighboring locations. Neighboring locations usually share the same climatic and geographical properties which means that they mostly share the same causes of particle pollution. Logically, some causes will always be related to some local properties, but if neighboring locations are close enough then we can conclude that they share many common causes of particle pollution and that they share common climatic and geographical conditions. Based on that, we can say that there should be a significant correlation between PM2.5 data from these locations. Generally, correlation should not be used as a proof of mutual causality, but it can be an indicator of sharing common causes, which is important in this case. So, we can say that common causality is proved by the proximity of locations. In this case we can conclude that the correlation between locations is a good indicator of sharing common causes of particle pollution.

The main idea is to use correlation as a base for calculating weights, instead of distance weighting. This way, missing values handling is not so complicated process, while the obtained results are very good.

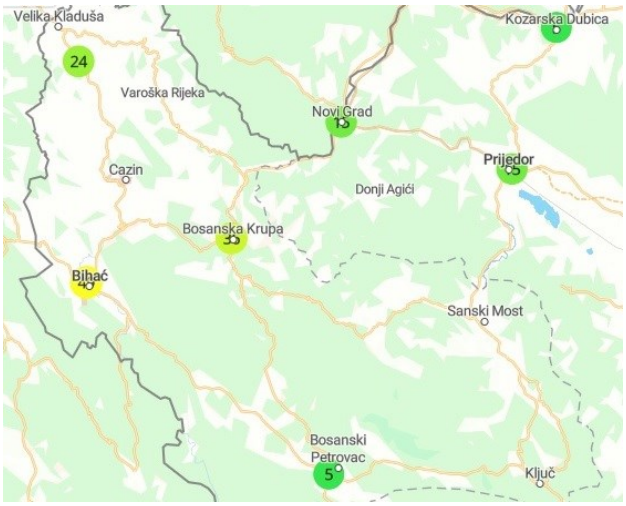


Figure 1. Map of locations with PurpleAir Sensors (image taken from the PurpleAir real-time map on 30.09.2022)

II. DATA

In our project we analyze data publicly available on the PurpuleAir real-time map [6], collected on several PurpleAir sensors located in Bosnia and Herzegovina. This paper uses data from sensors located in cities: Bihac, Bosanski Petrovac, Bosanska Krupa, Prijedor, and Velika Kladusa. Fig. 1 shows these locations on the map. Results and conclusions in this paper are based on the data from listed locations, for the period 01.01.2021 – 29.09.2022, and 60 minutes averages were used. Data was downloaded as CSV files on 30.09.2022.

If we observe listed locations on the map on Fig. 1 we can note that city Bosanska Krupa is surrounded by cities Bihać (air distance around 33 km), Velika Kladuša (air distance around 33 km), Prijedor (air distance around 47 km), and Bosanski Petrovac (air distance around 35 km). Based on the geographical location of these cities we decided to analyze and interpolate missing values for the city Bosanska Krupa, based on data from cities: Velika Kladuša, Bihać, Bosanski Petrovac, and Prijedor.

Table 1. shows locations list with data counts for the observed period 01.01.2021-29.09.2022 (primary, 60 minutes average) It is expected to have 15288 measurements for 636 days x 24 hours/day, but each of the sets do not have all entries. For the observed location there are 13970 entries, and 1318 entries are missing.

There are 787 entries in the data sets of all neighboring locations with entry time (*created_at* column) that do not exists in the data set for the observed location. That means that these values can be interpolated based on the existing data from the neighboring locations. 10060 entries in the data sets of all 5 locations are with the same entry time (created_at column) and they can be used for finding correlation between the observed location and their neighboring locations.

III. MODELING USING KNIME

KNIME Analytics Platform [7] is a very popular tool for various research projects. In this project, related to PurpleAir data, we use it for modeling different kind of workflows including those related to handling missing values in data.

TABLE I. LOCATIONS AND THEIR DATA COUNTS

ID	*Location name*	*Location Type*	*Data Counts*
#1	Velika Kladuša	neighboring	Number of entries: 12519 Missing entries: 2769
#2	Bihać	neighboring	Number of entries: 14050 Missing entries: 1238
#3	Bosanski Petrovac	neighboring	Number of entries: 15275 Missing entries: 13
#4	Prijedor	neighboring	Number of entries: 13992 Missing entries: 1296
#5	Bosanska Krupa	**observed**	Number of entries: 13970 **Missing entries: 1318**

At the beginning of the workflow, we use CSV Reader input nodes to read data from CSV files. Then manipulation nodes: Row Filter, Duplicate Row Filter, and Sorter are used to remove records with missing PM2.5 values, remove duplicated rows and sort data by date and time. We do not need rows with missing PM2.5 values in further processing.

By using Joiner nodes and the inner join setting for these nodes we create a new set composed of joined measurements for all five locations involved (four neighboring and one observed – target). That set contains only existing measured values on all locations for the same date and hour. If measurement is missing for any of five locations, then that date and hour is not included in this set. After that outliers are removed by using Numeric Outliers node. This new set is used to calculate the correlation between the four neighboring locations (Velika Kladuša, Bihać, Bosanski Petrovac, and Prijedor) and the observed location (Bosanska Krupa), using the Linear Correlation node. This node will calculate the correlation for all combinations of input columns, so it is necessary to filter the values of interest for the further processing. The results shown on Fig. 2.a indicate that there is a moderate to high positive correlation of PM2.5 values between the neighboring locations and the observed location. p-value stands for 'probability value' and it indicates how likely is that a result occurred by chance alone. Fig. 2.b shows zero p-values which indicates zero probability that this result is by chance.

S First column name	S Second column name	D Correlation value
PM2.5_ATM_ug/m3 (L #1)	PM2.5_ATM_ug/m3 (L #5)	0.7081996022459...
PM2.5_ATM_ug/m3 (L #2)	PM2.5_ATM_ug/m3 (L #5)	0.663273428193733
PM2.5_ATM_ug/m3 (L #3)	PM2.5_ATM_ug/m3 (L #5)	0.5906837388301...
PM2.5_ATM_ug/m3 (L #4)	PM2.5_ATM_ug/m3 (L #5)	0.6766711805903...

(a) Correlation values

D Correlation value	D p value	I Degrees of freedom
0.7081996022459994	0.0	8153
0.663273428193733	0.0	8153
0.5906837388301037	0.0	8153
0.6766711805903904	0.0	8153

(b) Two-sided p-values

Figure 2. Filtered result of the Linear Correlation node.

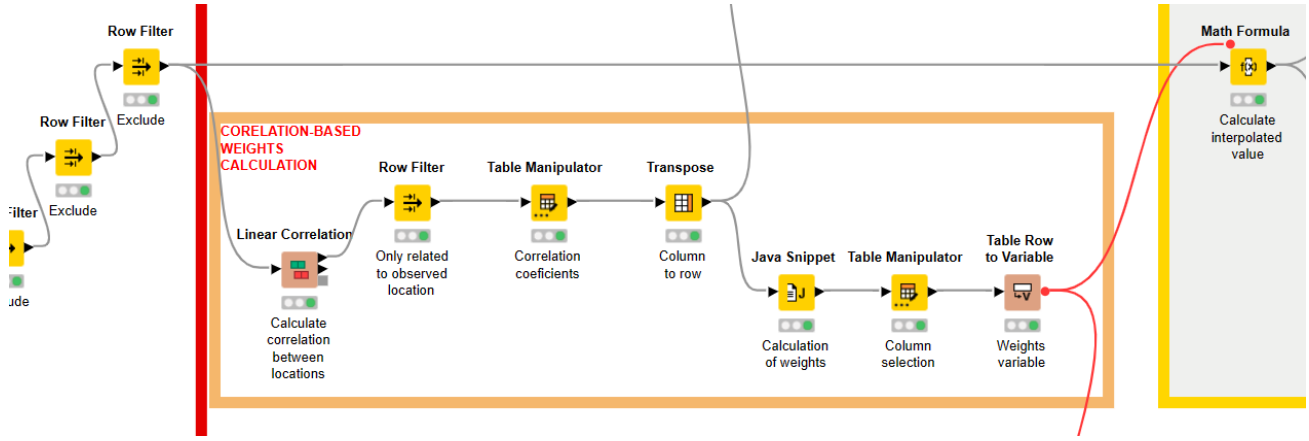


Figure 3. Part of the KNIME workflow where the correlation-based weight coeficients were calculated

In general, the Linear Correlation node calculates, for each pair of selected columns, a correlation coefficient – a measure of the correlation of the two columns in data (two variables). When columns contain numeric values then Pearson's product-moment coefficient (1) is calculated [7].

$$\rho_{X,Y} = cov(X,Y) / (\sigma_X \sigma_Y) \quad (1)$$

$cov(X,Y)$ is a covariance of two columns X and Y. σ_X is a standard deviation for the column X, and σ_Y is a standard deviation for the column Y.

The Fig. 3 shows a snippet of the workflow with nodes related to weights calculation based on the correlation values. In the Java Snippet node are calculated four weight values (w_i values) based on the correlation values (cor_i) using (2).

$$w_i = cor_i / \Sigma cor_i \text{ , } i = 1,2,3,4. \quad (2)$$

Weights are then transformed to flow variables using Table Row to Variable node. These weights variables are then used to calculate interpolated values as a weighted sum of measurements from the neighboring locations. In a further analysis of the results, it is concluded that it would be good to correct the interpolated result by multiplying it with the average ratio between the expected and the calculated value. The calculated multiplication factor has a value of 0.896. Fig. 4 shows quality indicators generated by the Numeric Scorer node.

Calculated weights and the multiplication factor are then used to interpolate missing values for the observed location. For that it is necessary to prepare the dataset that contains all rows with missing PM2.5 values in the dataset for the observed locations, but only for those date and time for which there are entries in datasets on all four neighboring locations. That way we ensure that we can calculate the interpolated value. For joining datasets, we again use Joiner nodes.

Statistics - 0:166 - Numeric Scorer (Quality)

File Edit Hilite Navigation View

Table "Scores" - Rows: 7 | Spec - Column: 1 | Properties | Flow Variables

Row ID	New_calculated_value
R^2	0.574
mean absolute error	9.652
mean squared error	218.999
root mean squared error	14.799
mean signed difference	1.11
mean absolute percentage error	0.407
adjusted R^2	0.574

Figure 4. Numeric Scorer node shows some quality indicators for the results.

In the final phase we joined the interpolated data and the existing data for the observed location to be able to make further analysis of the results. Our observations are presented in the next section of the paper.

So, the interpolation is based on a weighted sum, where the weighting factors are not calculated on the basis of distance but on the basis of correlation between data. The reason for this is that climatic influences, which are a very important part of this type of problem, cannot be connected only to the distance of locations. There are a lot of factors that come into play, from terrain configuration to air flow, and all of these factors are reflected through the correlation factor, which is why the correlation factor was taken into account when calculating weights, instead of distance.

IV. Discussion

Fig. 5 presents comparison of the final calculated

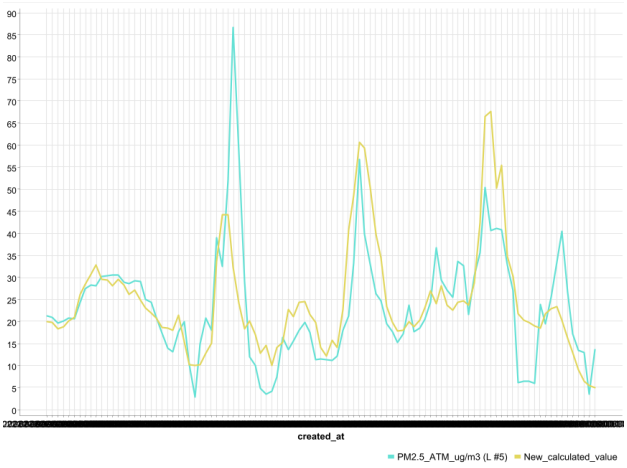


Figure 5. Checking results by comparing calculated value (after applying the correction factor) and known values.

(interpolated) values and the real (expected) values. We can note that the calculated values are near the real values in most of the points (60 minutes averages). It can be noted that there are differences and that they are mostly related to higher peaks in real values. That can be explained as a local influence (local factors related to the observed location) that cannot be calculated (predicted) based on the data from the neighboring locations.

Fig. 6 shows how interpolated values filled a larger gap of missing values. It can be noted that the transition from the interpolated to the existing values doesn't deviate much. For a few hours where there are interpolated values and for a few hours where there are original (real) values it can be noted that numbers are close to each other – around 20. Also, it can be noted that values decrease up to 18.44 and then they rise, which means that the trend is captured too. Also, it can be noted that the average of three marked interpolated values on Fig. 6 is 21,11 which is close to the average of three marked original values that are following them in a sequence, which is 20,18. If we bear in mind the fact that pollution influences between locations are not transmitted instantly but with a certain delay, then these small deviations (10-20%) are expected. All that is an argument that this approach is good enough, especially when there are larger gaps of missing values, like in the presented example.

Fig. 7 shows how interpolated values filled a small gap of missing values. It can be noted that the transition from the original to the interpolated values and back to the original values again is quite smooth.

The presented model can be seen as a simple model that can at least partially fill in the missing values in the data. Since it relies on existing data from neighboring locations, it has limited possibilities because it depends on them. In practice, this means that missing values cannot be interpolated if only a single value from one of the four neighboring sites is missing.

Of course, extensions of the model are possible in such a way that values are predicted based on more than four or less than four neighboring locations. In the case of an increase in the number of locations, it would be logical to expect an improvement in accuracy, but this increase cannot be significant, and it is difficult to verify it, because it would be very complicated to try to model microclimatic influences only for the observed location. Using fewer locations would result in less accurate values. The decision to use the four listed locations was more aimed at roughly covering the four sides (north, south, east, west) and thus covering the effects of climatic factors for those sides. That is why is less aimed at taking the closest locations.

Creating a model for something involves making a number of decisions related to the complexity of the model. The introduction of additional parameters usually leads to greater precision, but this may not be of vital interest for solving a particular problem. Therefore, in the modeling process, it is very important to balance between the practical usefulness of the model and the ideal model. Although the presented model has its shortcomings, it is still good enough for practical application and allows us to make good enough estimations of missing PM2.5 values for the observed location.

Regardless of all the shortcomings, this shows that, due to the fact that the locations share similar climatic characteristics, such a model allows us not only to fill in the missing values but also to roughly predict the values at the observed location, in case we don't have a sensor there, or the sensor doesn't work.

The number of missing values at the neighboring locations does not necessarily mean that the presented model will not be able to produce many interpolated values. The result depends on overlapping missing periods. If for the missing period on the observed location we have data on the neighboring location then the total number of missing values on the neighboring locations is not important. For example, although for the location Velika Kladuša are missing 2769 out of totally possible 15288 values which is 18,11%, the number of missing values on the observed location for which there are values on neighboring locations is 787, out of 1318 that are missing for the observed location. That means that for the observed location there are 8,62% missing values before applying the model and after applying the model number of missing values decreased to 3,47% which is an improvement of 5,15%. It would be possible to decrease this further by repeating the same process with other neighboring locations.

V. CONCLUSION

Our experiments showed that interpolating PM2.5 values in data from PurpleAir sensor, using existing data from neighboring locations and weighted sum based on a historical correlation of data has many advantages including: (1) easy implementation, and (2) possibility to interpolate large gaps in the data.

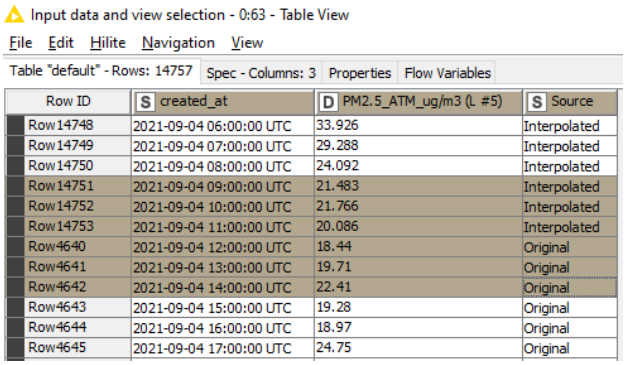
Comparing interpolated values and known (measured) values is possible to conclude that the interpolation results are very good. Discussion about results, and figures presented, argument our conclusions.

KNIME Analytical Platform tool enables very effective work during data analysis and modelling so that research focus is on modeling and model representation in the form of a workflow, and not on programming of a solution.

Our goal is to continue this research by calculating missing data using approaches based on the data for the observed location only, which can be a problem for larger gaps in data, so that we can compare approaches. Our final goal is to work on predictions of PM2.5 values for locations where PurpleAir sensor is temporary not working and not showing results on PurpleAir real-time map.

ACKNOWLEDGMENT

The authors thanks to the Ministry of Education and Science of entity Federation of Bosnia & Herzegovina of the state Bosnia & Herzegovina for the financial support to the project entitled: “Analysis of data from PurpleAir



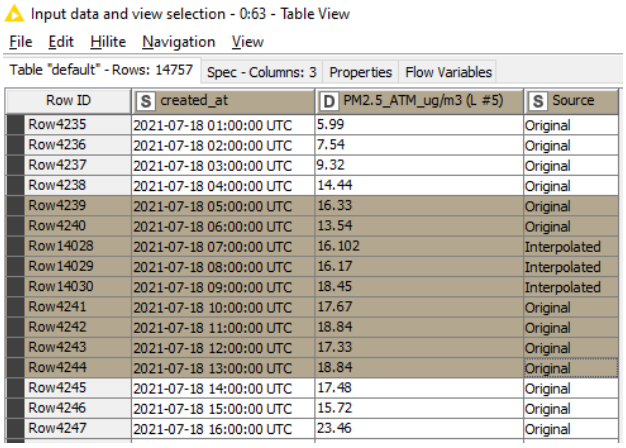
Input data and view selection - 0:63 - Table View

File Edit Hilite Navigation View

Table "default" - Rows: 14757 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	created_at	PM2.5_ATM_ug/m3 (L #5)	Source
Row14748	2021-09-04 06:00:00 UTC	33.926	Interpolated
Row14749	2021-09-04 07:00:00 UTC	29.288	Interpolated
Row14750	2021-09-04 08:00:00 UTC	24.092	Interpolated
Row14751	2021-09-04 09:00:00 UTC	21.483	Interpolated
Row14752	2021-09-04 10:00:00 UTC	21.766	Interpolated
Row14753	2021-09-04 11:00:00 UTC	20.086	Interpolated
Row4640	2021-09-04 12:00:00 UTC	18.44	Original
Row4641	2021-09-04 13:00:00 UTC	19.71	Original
Row4642	2021-09-04 14:00:00 UTC	22.41	Original
Row4643	2021-09-04 15:00:00 UTC	19.28	Original
Row4644	2021-09-04 16:00:00 UTC	18.97	Original
Row4645	2021-09-04 17:00:00 UTC	24.75	Original

Figure 6. Part of time-series where interpolated values filled a larger gap of missing values. Note the transition from the interpolated to the existing values.



Input data and view selection - 0:63 - Table View

File Edit Hilite Navigation View

Table "default" - Rows: 14757 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	created_at	PM2.5_ATM_ug/m3 (L #5)	Source
Row4235	2021-07-18 01:00:00 UTC	5.99	Original
Row4236	2021-07-18 02:00:00 UTC	7.54	Original
Row4237	2021-07-18 03:00:00 UTC	9.32	Original
Row4238	2021-07-18 04:00:00 UTC	14.44	Original
Row4239	2021-07-18 05:00:00 UTC	16.33	Original
Row4240	2021-07-18 06:00:00 UTC	13.54	Original
Row14028	2021-07-18 07:00:00 UTC	16.102	Interpolated
Row14029	2021-07-18 08:00:00 UTC	16.17	Interpolated
Row14030	2021-07-18 09:00:00 UTC	18.45	Interpolated
Row4241	2021-07-18 10:00:00 UTC	17.67	Original
Row4242	2021-07-18 11:00:00 UTC	18.84	Original
Row4243	2021-07-18 12:00:00 UTC	17.33	Original
Row4244	2021-07-18 13:00:00 UTC	18.84	Original
Row4245	2021-07-18 14:00:00 UTC	17.48	Original
Row4246	2021-07-18 15:00:00 UTC	15.72	Original
Row4247	2021-07-18 16:00:00 UTC	23.46	Original

Figure 7. Part of time-series where interpolated values filled a small gap of missing values. Note the transition from the original to the interpolated values and back to the original again.

sensors" approved by the contract no. 05-35-2135-1/22 from 27.10.2022.

REFERENCES

[1] Struminskaya, B., Lugtig, P., Keusch, F., & Höhne, J. K. (2020). Augmenting Surveys With Data From Sensors and Apps: Opportunities and Challenges. Social Science Computer Review, 0(0). https://doi.org/10.1177/0894439320979951

[2] Lu T, Liu Y, Garcia A, Wang M, Li Y, Bravo-Villasenor G, Campos K, Xu J, Han B. Leveraging Citizen Science and Low-Cost Sensors to Characterize Air Pollution Exposure of Disadvantaged Communities in Southern California. Int J Environ Res Public Health. 2022 Jul 19;19(14):8777. doi: 10.3390/ijerph19148777. PMID: 35886628; PMCID: PMC9322770.

[3] A. Caseiro, S. Schmitz, G. Villena, J. V. Jagatha, and E. von Schneidemesser, "Ambient characterisation of PurpleAir particulate matter monitors for measurements to be considered as indicative," *Environ. Sci.: Atmos.*, p., 2022, doi: 10.1039/D2EA00085G.

[4] Kasam, A.A., Lee, B.D. & Paredis, C.J.J. Statistical methods for interpolating missing meteorological data for use in building simulation. Build. Simul. 7, 455–465 (2014). https://doi.org/10.1007/s12273-014-0174-7

[5] Lepot, M.; Aubin, J.-B.; Clemens, F.H.L.R. Interpolation in Time Series: An Introductive Overview of Existing Methods, Their Performance Criteria and Uncertainty Assessment. Water 2017, 9, 796. https://doi.org/10.3390/w9100796

[6] PurpleAir, "Map Start-up Guide," PurpleAir Community. [Online]. Available: https://community.purpleair.com/t/map-start-up-guide/90. [Accessed: Dec. 06, 2022].

[7] KNIME AG: "KNIME Analytics Platform – Allowing anyone to build and upskill on data science". [Online]. Available: https://www.knime.com/knime-analytics-platform. [Accessed: Dec. 07, 2022].

[8] KNIME AG: "Linear Correlation". [Online]. Available: https://hub.knime.com/knime/extensions/org.knime.features.base/latest/org.knime.base.node.preproc.correlation.compute2.CorrelationCompute2NodeFactory [Accessed: Dec. 07, 2022].