A Unified Air Quality Assessment Framework Based on Linear Fuzzy Space Theory

Endre Pap
Singidunum University
Belgrade, Serbia
epap@singidunum.ac.rs

Zora Konjović
Singidunum University
Belgrade, Serbia
zkonjovic@singidunum.ac.rs

Đorđe Obradović
Singidunum University
Belgrade, Serbia
djobradovic@singidunum.ac.rs

Ivan Radosavljević
Singidunum University
Belgrade, Serbia
iradosavljevic@singidunum.ac.rs

Abstract **- Air quality is one of the most critical issues that humankind is facing today. There are diverse types of indices measuring air pollution, which are mostly based on aggregation functions. This paper proposes a model aimed at forecasting aggregated air pollution indices, which enables modelling data uncertainties. The proposed original model consists essentially of two sub models. The first one models Air Quality Index (AQI), while the second one models concentrations of pollutants. Multi-contaminant air quality index is modelled as an aggregation of the Pollutant Standard Index (PSI) obtained via fuzzy linear transformation defined by fuzzy breakpoints. We model concentrations of pollutants by regression (XGBoost, Deep Neural Network, ADA Boost, and Histogram Gradient Boosting Regressor) using fuzzy time series of two groups of data (pollutants' concentrations and meteorological parameters). In doing so, the target variable was modeled in two ways. The first model is a set of independent classes defined by proper fuzzy membership functions, while the second one is a set of classes connected by an ordering relation. Simulation results are presented showing the model performance for each of the target variable models in terms of prediction mean absolute errors. The main result of the paper is a unified model of air quality assessment relying upon a consistent mathematical theory called Linear fuzzy space.**

Keywords - Linear fuzzy space; AQI index; aggregation operator; connected classes.

I. INTRODUCTION

This paper is an extended version of our conference paper "Linear Fuzzy Space Based Framework for Air Quality Assessment" [1] presented at the INTELLI 2022 Conference held in Venice in May 2022.

In relation to the initial work, this work introduces an improvement related to modeling the target variable. While in the original paper the target variable was modeled by independent classes described by appropriate membership functions, in this paper the target variable is modeled as a set of classes with an ordering relation.

The research presented in both the original and extended papers is motivated by the fact that in the last decade, humankind has been facing air pollution as one of the most important issues with adverse effects on human health, but also on the economy of societies. According to WHO's annual World Health Statistics report from 2016, outdoor air pollution causes approximately 4.2 million deaths per year [2]. As reported by the European Environmental Agency (EEA) in 2018, the number of deaths in Europe related to concentrations of the particles PM2.5 was about 379,000 [3]. Therefore, there is a great need for air pollution forecasting models that will express the air pollution as a simple value that is understandable for a wide audience.

Air pollution is an extremely complex spatio-temporally determined dynamic system distinctly characterized by the presence of imprecision and uncertainty. Therefore, it is not easy to give a precise air pollution forecast, which would be of great importance for public health.

To cope with uncertainty and imprecision, we use a fuzzy approach. More precisely, the one based on our previous results presented in [4] [5] [6] [7] [8], where we introduced mathematical models for basic concepts: fuzzy point, fuzzy spatial relation, fuzzy ordering, fuzzy distance, fuzzy measurement and simple geometrical fuzzy objects (line, triangle, circle). For modelling the temporal dimension of air pollution, we use a combination of time series models with techniques supporting the manipulation of imprecise and uncertain data, known under the umbrella term Fuzzy Time Series (FTS). This model enables a more adequate air pollution forecast.

Multi-contaminant Air Quality Index/Common Air Quality Index (AQI/CAQI) manages multiple effects due to the exposure to more pollutants, gives more complete information on the possible impacts of air pollutants and a direction for a more accurate, consistent, and comparable

AQI/CAQI system. Hence, we opt for multi-contaminant AQIs/CAQIs as a model of air pollution estimate.

For that purpose, the ordering relations $\leq^{RF}$ and $\leq^{LF}$ were introduced and it was proved that they agree with the definition of the fuzzy ordering with respect to the t-norm T and the equivalence E ($T-E$ ordering) from the Linear fuzzy space.

Simulations were performed for the two proposed target variable models on the same data set using the same regression models (XGBoost, Deep Neural Network, ADA Boost, and Histogram Gradient Boosting Regressor).

The obtained results showed the superiority of the model of connected classes over the model of independent classes in terms of mean square error.

The rest of the paper is organized into five sections. Section 2 presents related work, Section 3 brings theoretical foundations, while Section 4 presents the model of the proposed framework. Section 5 shows model application and simulation results for the real data set (82457 samples/16 variables/24h measurements). Finally, Section 6 summarizes the research results, identifies deficiencies, and outlines future research.

II. Related work

This section brings analysis of related work and basic underlaying preliminaries of our research.

As already said, air pollution is a complex spatio-temporally determined dynamic system characterized by the presence of imprecision and uncertainty, which makes air pollution modelling and prediction a challenging task. The research field itself is vivid, yearly generating hundreds of publications, which deal with the modelling task. There are also several recent papers that provide more-less inclusive overview of the field like [9] with specific objectives (a) to address current developments that push the boundaries of air quality research forward, (b) to highlight the emerging prominent gaps of knowledge in air quality research, and (c) to make recommendations to guide the direction for future research within the wider community. This research identifies Earth system modelling as offering considerable potential by providing a consistent framework for treating scales and processes.

One important issue of air quality modelling is the quality indicator, which is called Air Quality Index (AQI) in the USA and Common Air Quality Index (CAQI) in Europe. These two terms share the same semantics so we shall use them interchangeably through the rest of this text.

As shown in [10], a multi-contaminant model of AQI in which aggregation functions (aggregation operators) are applied to combine several numerical values into a single representative is predominant by far.

The simplest AQI model calculates a sub-index (AQI_i) for each pollutant i by the following linear interpolation formula:

$$AQI_i = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - C_{low}) + I_{low} \,.$$

Here, C is the monitored ambient average concentration of pollutant i; C_{low} is the breakpoint lower than or equal to C; C_{high} is the breakpoint higher than or equal to C; and I_{low} and I_{high} are the sub-index values corresponding to C_{low} and C_{high}, respectively. The overall AQI is then calculated as a simple max aggregation:

$$AQI = \max_{i=1}^{m}(AQI_i).$$

There is ongoing research for new aggregation functions, which involve the influence of multiple pollutants [11] [12] [13] [14] [15] [16] [17] [18]. Among these AQIs, arithmetic pollutant aggregation integrates pollutants in a linear or nonlinear way, and weighted pollutant aggregation further assigns varied weights from different approaches. The General Air Quality Health Index ($GAQHI$) is proposed as a pollutant-aggregated, local health-based AQI paradigm suitable for representing a complex multi-contaminant situation:

$$Is = \left(\sum_{i=1}^{n}(AQI_i)^{\alpha}\right)^{\frac{1}{\alpha}},$$

where $\alpha \in [1, \infty]$.

An interesting modification of the United States Environmental Protection Agency (EPA) AQI is proposed in [11], giving a new index $RAQI$, which is the product of three terms:

$$RAQI = F_1 * F_2 * F_3$$

were

$$F_1 = \max(I_i), \qquad i = 1{,}5$$

$$F_2 = \frac{\sum_{i=1}^{5} Ave_{daily}\,(I_i)}{Ave_{annual} \cdot \left(\sum_{i=1}^{5} Ave_{daily}\,(I_i)\right)}$$

and the Shannon entropy function is introduced in the third term:

$$F_3 = \frac{Ave_{annual} \cdot Entropy_{daily}\left(\max_{i=1}^{5}(I_i)\right)}{Entropy_{daily}\left(\max_{i=1}^{5}(I_i)\right)}$$

This model strives to avoid ambiguity (indicating a less polluted air as highly polluted) and ellipticity (indicating highly polluted air as less polluted) by introducing entropy.

In addition, there are interesting approaches like in [12] that model the air pollution index via a mixture of distributions based on its structure and descriptive status as well as research doing with development of aggregate air quality index for a specific agglomeration [13] and tools used to inform the public about the status of the ambient air quality [14] in which different AQIs are analyzed to contribute to the sharing of air quality management practices and information to raise public awareness and to help policymakers to act accordingly.

There are also results that utilize fuzzy logic for modelling air quality indices, like those presented in [15] [16] [17] [18].

In the paper [15], Atacak and coauthors present the model in which the input variables are air pollutant criteria (PM10, SO2, CO, NO2, O3), and the output variable is fuzzy AQI. The fuzzification process is defined via the boundary values of the universal sets and the corresponding fuzzy sets (trapezoidal for input, and triangular for output variables). The rule base representing the relationship between input variables and output variables has 243 rules. The max-min inference strategy and centroid method are chosen for the inference and defuzzification process. The paper [16] proposed an index that, in addition to criteria air pollutants (CO, SO2, PM10, O3, NO2), includes benzene, toluene, ethylbenzene, xylene, and 1,3-butadiene due to their considerable health effects. Different weighting factors were then assigned to each pollutant according to its priority. Trapezoidal membership functions were employed for classifications and the final index consisted of 72 inference rules.

Time series is a model that is extensively used in air quality modelling. In [17], the authors present a comparative study of the results obtained from several models for air pollution index forecast, which shows that the fuzzy time series models outperformed the other models in terms of forecasting accuracy and computation time. Finally, [18] utilizes a fuzzy time series-Markov chain model for predicting the daily air pollution index.

Current air quality research relies heavily on machine learning. Another characteristic that could be attributed to them is the still partial observation of the phenomenon, with rare attempts at comprehensive modeling of this extremely complex system. With the pretension to show only a rough picture of the area and to connect it with the earlier claim, we have presented here a few characteristic papers. The first criterion for the selection of the presented papers was that they deal with air quality research, the second criterion was that they applied artificial intelligence techniques for research and, finally, the third criterion was that they were published in the course of 2022. An exception to the last criterion is the paper [19] from 2019, which we presented because it provides an overview of neural network models applied in the field, with the focus on the most frequently studied pollutants (PM10, PM2.5, nitrogen oxides, ozone). In this source, most of the work is devoted to the long-term forecasting of outdoor PM10, PM2.5, oxides of nitrogen, and ozone. Most of the identified works used meteorological and source emissions predictors almost exclusively. Furthermore, ad-hoc approaches are found to be predominantly used for deciding optimal model predictors, proper data subsets with the optimal model structure. Multilayer perceptron and ensemble-type models are predominantly implemented. The paper [20] is a review paper that is based on 128 articles published from 2000 to 2022. The review reveals that input uncertainty was predominantly addressed while less focus was given to structure, parameter, and output uncertainties. Ensemble approaches are used mostly, followed by neuro-fuzzy networks. The use of bootstrapping, Bayesian, and Monte Carlo simulation techniques, which can quantify uncertainty, was also found to be limited. Authors recommend the development and application of approaches that can both handle and quantify uncertainty surrounding the development of ANN models. The source [21] is also a review paper showing recent attempts to use deep learning techniques in air quality forecasts. There are presented deep networks, e.g., convolutional neural networks, recurrent neural networks, long short-term memory neural networks, and spatiotemporal deep networks, and their connection to the nonlinear spatiotemporal features across multiple scales of air pollution. The source presents deep learning techniques for air quality forecasts in diverse aspects, e.g., data gap filling, prediction algorithms, estimations with satellite data, and source estimations for atmospheric dispersion forecasts. The paper [22] is a focused one. This paper develops a hybrid modeling framework that combines the elastic net and multivariate relevance vector machine for interval-valued PM2.5 time series forecasting. Instead of directly modelling linear and nonlinear patterns of time series, there is introduced the multi-factor interval division approach and bivariate empirical mode decomposition algorithm into linear and nonlinear pattern modeling, respectively. The last source presented here [23] developed the LIFE Index-Air tool, where the air pollutant concentrations are predicted by Artificial Neural Networks trained using a set of air quality modeling simulations. Authors argued that the results show that this approach based on ANN, calibrated using a limited number of air quality modeling system simulations, can reproduce the concentration values competently.

Previous analysis of air quality research shows that research in the field of air quality modeling and prediction is active and diverse, with research that deals with inaccuracies and uncertainties in the data applying a fuzzy approach being very common. A significant lack of current research, in the opinion of the authors of this paper, is a consistent theoretical basis for modeling air quality as a temporally and spatially determined system with inaccuracies and uncertainties in the data (including spatial and temporal data).

Our work represents an investigation of the possibility to use the theory of Linear fuzzy space, which we are developing, as a theoretical basis for modeling air quality as a time-space system with inaccuracies and uncertainties in the data. In addition to theoretical results used in this paper, the theory of Linear fuzzy space includes models of imprecise 2D geometric objects (line, triangle, circle), models of imprecise spatial operations (spatial measurements, spatial transformations, spatial relations), and initial results proving the theory extensibility to a fuzzy finite automata model. In this way, a consistent framework is obtained that applies to modeling and simulation of air quality systems, including urban pollution where the space geometry plays an important role.

III. PRELIMINARIES

The theoretical foundations of our model are based on Linear fuzzy space theory, multi-contaminant fuzzy AQIs, fuzzy time series, and Machine Learning (ML)-based classification.

This section presents in brief basic underlaying preliminaries of our research, Linear fuzzy space, fuzzy aggregation operators and fuzzy time series.

A. Linear fuzzy space

In this subsection, we present the fundamental concepts of the Linear fuzzy space theory: fuzzy point, linear fuzzy space, fuzzy space ordering, fuzzy linear combination, and fuzzy space metrics, as defined in [4, 5, 6, 7, 8].

Definition 1 *Fuzzy point* $P \in R^2$, denoted by $\tilde{P}$ is defined by its membership function $\mu_{\tilde{P}} \in \mathcal{F}^2$**,** where the set $\mathcal{F}^2$ contains all membership functions $u: R^2 \to [0,1]$ satisfying the following conditions:

i) $(\forall u \in \mathcal{F}^2)(\exists_1 P \in R^2)\, u(P) = 1$,

ii) $(\forall X_1, X_2 \in R^2)(\lambda \in [0,1])\, u(\lambda X_1 + (1-\lambda)X_2) \geq min\big(u(X_1), u(X_2)\big)$,

iii) function u is upper semi-continuous,

iv) $[u]^\alpha = \{X | X \in R^2, u(X) \geq \alpha\}$ α-cut of function u is convex.

Here, a point from R^2 with a membership function $\mu_{\tilde{P}}(P) = 1$, is denoted by P (P is the core of the fuzzy point $\tilde{P}$), the membership function of point $\tilde{P}$ is denoted by $\mu_{\tilde{P}}$, while $[P]^\alpha$ stands for the α-cut (a set from R^2) of the fuzzy point.

Figure 1 shows geometrical illustration of a fuzzy point membership function $\mu(X)$ and its α-cut.

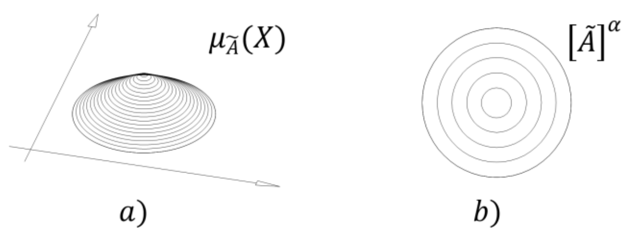

Figure 1. Geometrical illustration of a) a fuzzy point membership function and b) its α-cuts

Definition 2 R^2 *Linear fuzzy space* is the set $\mathcal{H}^2 \subset \mathcal{F}^2$ of all functions, which, in addition to the properties given in Definition 1, are:

i) Symmetric with respect to the core $S \in R^2$

$$(\mu(S) = 1), \mu(V) = \mu(M) \wedge \mu(M) \neq 0 \Rightarrow$$
$$d(S,V) = d(S,M)$$

where $d(S,M)$ is the distance in R^2.

ii) Inverse-linearly decreasing regarding points' distance from the core, i.e.:

$$\text{If } r \neq 0: \quad \mu_{\tilde{S}}(V) = \max\left(0, 1 - \frac{d(S,V)}{|r_S|}\right),$$

$$\text{If } r = 0: \quad \mu_{\tilde{S}}(V) = \begin{cases} 1 & if \quad S = V \\ 0 & if \quad S \neq V \end{cases},$$

where $d(S,V)$ is the distance between point V and the core S ($V, S \in R^2$) and $r \in R$ is a constant.

The elements of that space are represented as ordered pairs $\tilde{S} = (S, r_S)$ where $S \in R^2$ is the core of $\tilde{S}$, and $r_S \in R$ is the distance from the core for which the function value becomes 0.

Definition 3. A t-norm is a function T: $[0,1] \times [0,1] \to [0,1]$ that satisfies the following properties:

i) Commutativity: $T(a,b) = T(b,a)$

ii) Monotonicity: $T(a,b) \leq T(c,d)$ if $a \leq c$ and $b \leq d$

iii) Associativity: $T(a, T(b,c)) = T(T(a,b),c)$

iv) The number 1 acts as identity element: $T(a,1) = a$

Definition 4. Let X be a nonempty set and function T be a t-norm. Fuzzy relation $E: X \times X \to [0,1]$ is called *fuzzy equivalence relation* (T*-equivalence*) with respect to t-norm T if the following axioms for $x, y, z \in X$ hold:

i) Reflexivity $\quad E(x,x) = 1$

ii) Symmetry $\quad E(x,y) = E(y,x)$

iii) T-Transitivity $\quad T(E(x,y), E(y,z)) \leq E(x,z)$

Definition 5. Let $E: X \times X \to [0,1]$ be a T*-equivalence*. Fuzzy relation $L: X \times X \to [0,1]$ is called *fuzzy ordering* with respect to norm T and equivalence E ($T - E$ *ordering*) if for any $x, y, z \in X$ the following hold:

i) E- Reflexivity $\quad E(x,y) \leq L(x,y)$

ii) T-E Anti-symmetry $\quad T\big(L(x,y), L(y,x)\big) \leq E(x,y)$

iii) T- Transitivity $\quad T\big(L(x,y), L(y,z)\big) \leq L(x,z)$

Definition 6 Let $\mathcal{H}^2$ be a linear fuzzy space. Then, a function $f: \mathcal{H}^2 \times \mathcal{H}^2 \times [0,1] \to \mathcal{H}^2$ called a *linear combination* of the fuzzy points $\tilde{A}, \tilde{B} \in \mathcal{H}^2$ is given by:

$$f\big(\tilde{A}, \tilde{B}, u\big) = \tilde{A} + u \cdot \big(\tilde{B} - \tilde{A}\big),$$

where $u \in [0,1]$, the operator $+$ is the sum of fuzzy points, and the operator $\cdot$ is the scalar multiplication of the fuzzy point.

Note: The thesis [4] defines operations sum of fuzzy points ($+$) and scalar multiplication of the fuzzy point ($\cdot$) and proves that an ordered quadruple $(\mathcal{H}^2, +, R^2, \cdot)$ is a vector space.

Figure 2 is a geometrical illustration of the linear combination of the fuzzy points $\tilde{A}, \tilde{B} \in \mathcal{H}^2$.

Measurement in space, especially the distance between plane geometry objects, is defined as a generalization of the concept of physical distance.

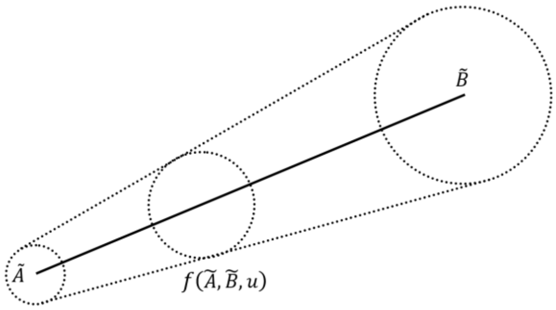

Figure 2. Geometrical ilustration of a linear combination of the fuzzy points $\tilde{A}, \tilde{B} \in \mathcal{H}^2$

Definition 7. Let $\mathcal{H}^2$ be a linear fuzzy space and $\tilde{d}: \mathcal{H}^2 \times \mathcal{H}^2 \to \mathcal{H}^+$, L, R: $[0,1] \times [0,1] \to [0,1]$ be symmetric, associative, and non-decreasing for both arguments, and $\mathrm{L}(0,0) = 0$, $\mathrm{R}(1,1) = 1$. The ordered quadruple $(\mathcal{H}^2, \tilde{d}, \mathrm{L}, \mathrm{R})$ is called fuzzy metric space and the function $\tilde{d}$ is a *fuzzy metric*, if and only if the following conditions hold:

i) $\tilde{d}(\tilde{X}, \tilde{Y}) = \tilde{0} \Leftrightarrow [\tilde{X}]^1 = [\tilde{Y}]^1$

ii) $\tilde{d}(\tilde{X}, \tilde{Y}) = \tilde{d}(\tilde{Y}, \tilde{X}),\ \widetilde{\forall} X, \tilde{Y} \in \mathcal{H}^2$

iii) $\forall \tilde{X}, \tilde{Y} \in \mathcal{H}^2$:

if $s \le \lambda_1(x,z) \wedge t \le \lambda_1(z,y) \wedge s + t \le \lambda_1(x,y)$

$$\tilde{d}(\tilde{X}, \tilde{Y})(s+t) \ge \mathrm{L}\big(d(x,z)(s), d(z,y)(t)\big)$$

if $s \ge \lambda_1(x,z) \wedge t \ge \lambda_1(z,y) \wedge s + t \ge \lambda_1(x,y)$

$$\tilde{d}(\tilde{X}, \tilde{Y})(s+t) \le \mathrm{R}\big(d(x,z)(s), d(z,y)(t)\big)$$

The α-cut of a fuzzy number $\tilde{d}(x,y)$ is given by

$$\left[\tilde{d}(\tilde{X}, \tilde{Y})\right]^\alpha = [\lambda_\alpha(x,y), \rho_\alpha(x,y)]\ (x, y \in R^+, 0 < \alpha \le 1).$$

The fuzzy zero, $\tilde{0}$ is a non-negative fuzzy number with $[\tilde{0}]^1 = 0$.

B. *Fuzzy aggregation operators*

An aggregation operator has natural properties such as monotonicity and boundary conditions. In practice, the data is usually normalized, so the definition of aggregation becomes:

Definition 8. An aggregation function (operator) is a function $A^{(n)}: [0,1]^n \to [0,1]$ that satisfies the following conditions:

1. is nondecreasing (in each variable)
2. $A^{(n)}(0, \dots, 0) = 0$ and $A^{(n)}(1, \dots, 1) = 1$.

Aggregation applies to various fields and takes diverse forms, from the simple to quite sophisticated and complex ones, modelling the interaction between criteria, which are managed by monotone set functions and corresponding integrals [24] [25] [26] [27]. In our research, a fuzzy aggregation operator is the basic instrument for multi-contaminant AQI/CAQI modelling.

C. *Fuzzy time series*

Most of the real-world tasks that utilize time series rely on multivariate time series models [28] [29] [30] [31]. The common multivariate time series model is [28]:

Let $Z_t = [Z_{1,t}, Z_{2,t}, \dots, Z_{m,t}]'$ be an m -dimensional jointly stationary real-valued vector process such that $E(Z_{i,t}) = \mu_i$ is a constant for each $i = 1,2, \dots, m$ and the cross-covariances between $Z_{i,t}$ and $Z_{j,s}$ for all $i = 1,2, \dots, m$ and $j = 1,2, \dots, m$ are functions only of the time difference $(s - t)$.

On the other hand, the original definition of the univariate first fuzzy time series model is [31]:

Definition 9. Let $Y(t)(t = \dots, 0, 1, 2, \dots)$, a subset of R^1 be the universe of discourse on which fuzzy sets $f_i(t)(i = 1, 2, \dots)$ are defined and $F(t)$ is the collection of $f_i(t)(i = 1, 2, \dots)$. Then, $F(t)$ is called a fuzzy time series on $Y(t)(t = \dots, 0, 1, 2, \dots)$.

Our time series model is a combination of the previous two where we apply the same common multivariate model, which is modified to support imprecise values. In our model, we simply replace a crisp point with a linear fuzzy space point [4]:

Definition 10. Let $Y(t)(t = \dots, 0, 1, 2, \dots)$, a subset of R^1 be the universe of discourse. Let $\mathcal{H}^l\ (l = 1,2)$ be a linear fuzzy space. Furthermore, let $f_i(t)(i = 1, 2, \dots)$ be fuzzy sets defined as points on a linear fuzzy space over the given universe of discourse, and $\widetilde{F}_j(t)(j = 1, 2, \dots m)$ be collections of these fuzzy points. Then, $\widetilde{F}_t = [\tilde{F}_{1,t}, \tilde{F}_{2,t}, \dots, \tilde{F}_{m,t}]'$ is called a linear fuzzy space based fuzzy time series on $Y(t)(t = \dots, 0, 1, 2, \dots)$.

This definition enables all features of linear fuzzy space to be used. For example, a process vector can be of a mixed type (some components can be crisp, some can be fuzzy) whilst spatial relations defined on the linear fuzzy space hold.

IV. Fuzzy model of Air pollution indices prediction

In this example, we show how the linear fuzzy space is used for time series-based forecasting. Fuzzy time series defined by the linear fuzzy space, as described in subsection C of Section III, are used to model air quality forecast.

A. *Data model*

The data model used in this paper consists of temporal georeferenced samples. Each sample is a time series covering the earlier 24h in 1h sample rate (total 385 real values). Each time series corresponds to one variable. Variables are divided into two groups: six common air pollutants known as “criteria air pollutants” [32], and ten meteorological parameters, the Global Data Assimilation System (GDAS) [33] shown in Table I and Table II, respectively.

TABLE I. AIR POLLUTANTS PARAMETERS

ID	Description	Unit
PM10	Suspended particles smaller than 10 μm	$\mu g/m^3$
PM25	Suspended particles smaller than 2.5 μm	$\mu g/m^3$
SO2	Sulphur dioxide	ppb
CO	Carbon Monoxide	ppm
NO2	Nitrogen Dioxide	ppb
O3	Ground-level Ozone	ppm

TABLE II. GDAS PARAMETERS

ID	Description	Unit
PRSS	Pressure at Earth surface	hPa
TPP6	Accumulated precipitation (6 h accumulate.)	m
RH2M	Relative Humidity at 2m AGL	%
TO2M	Temperature at 2m AGL	K
TCLD	Total cloud cover (3- or 6-h average)	%
U10M	U-component of wind at 10 m AGL	m/s
V10M	V-component of wind at 10 m AGL	m/s
TMPS	Temperature at surface	K
PBLH	Planetary boundary layer height	m
irradiance	Irradiance/solar power	W/m2

B. *Linear Fuzzy Space-based air pollution index*

Since air pollutants are measured in different physical units and scales, the first step is to transform them into a common domain (0-500). This transformation is usually defined by breakpoint tables and the resulting values are called Pollutant Standard Index (PSI). Instead of using discrete functions, we propose a fuzzy linear transformation defined by fuzzy breakpoints (Figure 3).

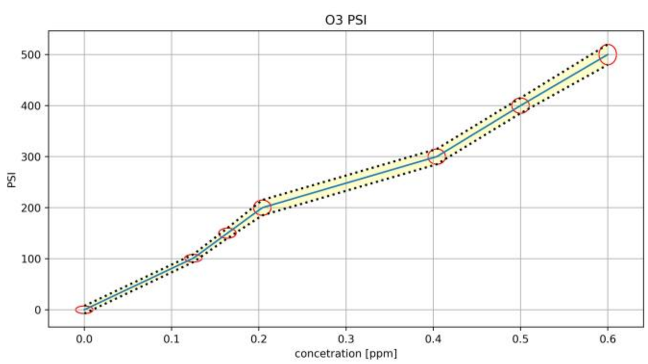

Figure 3. Fuzzy linear transformation for Ground-level Ozone (O3)

A fuzzy linear transformation is defined by an ordered list of 2D Fuzzy points $\tilde{P} = (\tilde{X}, \tilde{Y})$. Each 2D fuzzy point consists of two components $\tilde{X} = (X, r_X)$ and $\tilde{Y} = (Y, r_Y)$, which are 1D fuzzy points.

Then, Fuzzy PSI (FPSI) is defined as:

$$\widetilde{FPSI_\iota} = \widetilde{l\iota nterp}\left(C, \left[\widetilde{P_0}, \dots \widetilde{P_n}\right]\right) = (\,FPSI_i, r_{FPSI})\,,$$

$$FPSI_i = \frac{Y_{high} - Y_{low}}{X_{high} - X_{low}}(C - X_{low}) + Y_{low}$$

$$r_{PSI} = \frac{r_{Yhigh} - r_{Ylow}}{r_{Xhigh} - r_{Xlow}}(C - X_{low}) + r_{Ylow}$$

where $\widetilde{l\iota nterp}$ is a fuzzy linear transformation from concertation fuzzy space into index fuzzy space. Fuzzy points $\widetilde{P_{h\iota gh}}$ and $\widetilde{P_{low}}$ are fuzzy points whose roots of $\tilde{X}$ components are nearest to the concertation C.

$FPSI$ can be further represented by a linguistic variable, or it can be used directly in the aggregation process.

A single fuzzy value $FAQI$ is obtained by applying some fuzzy aggregation operator (aggreg) to all (n) components $FPSI$ indices:

$$FAQI = aggreg(FPSI_i), i = 1, n$$

To simplify the decision-making process and/or facilitate general understanding, a fuzzy linguistic variable defined by corresponding fuzzy sets can be easily introduced in such a model.

C. *Prediction model*

In our model, we opt for multivariate regression to forecast $FAQI$. However, other classification methods can easily be incorporated in the proposed model.

A prediction model in which the target variable consists of unordered classes is shown in Figure 4.

The shortcoming of this approach is that the individual membership functions, that is, the fuzzy classes, are viewed as if they were independent variables, i.e., the fact that the **Very low** and **Low** classes are "closer" than the **Very Low** and **Very High** classes is not considered.

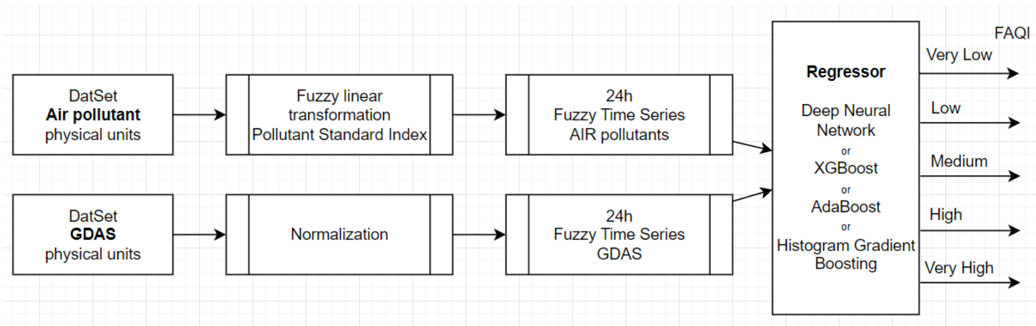

Figure 4. Prediction model with target variable modeled by unordered classes

To address this problem, this paper used the fact that classes can be arranged by the ordering relation:

Very Low ≤ Low ≤ Medium ≤ Hight ≤Very High
$$(\ \tilde{\mathbf{0}} \leq \tilde{\mathbf{1}} \leq \tilde{\mathbf{2}} \leq \tilde{\mathbf{3}} \leq \tilde{\mathbf{4}}\ )$$

If we represent the **Very Low** class with the fuzzy number $\tilde{0}$ and the **Very High** class with the fuzzy number $\tilde{4}$, then we can use the analogy with integers and real numbers, so the ordering could be like an ordering among integer numbers, and the extension of a strict class membership could be a Type 2 fuzzy membership over discourse R. Then, for example, the number 1.2 stands for a partial belonging to classes 1 and 2, which simultaneously reflects the arrangement of classes. This also expresses in a simpler way the fact:

$$F_j(AQI) = 0 \ \rightarrow F_i(AQI) = 0, \forall i, i > j$$

To implement this idea, we have introduced the notion of fuzzy relations $\leq^{RF}$ and $\leq^{LF}$.

Definition 11. Let $\mathcal{H}$ be a linear fuzzy space defined on R^1. Fuzzy relations $\leq^{RF}$ and $\leq^{LF}$ on the set $\mathcal{H}$ are defined by the following membership functions:

$$\mu(\tilde{A} \ \leq^{RF} \ \tilde{B}) = \begin{cases} 1 & if\ A > B \\ \dfrac{B - A}{r_B - r_A} & if\ A \leq B \wedge A - r_A > B - r_B \\ 0 & if\ A \leq B \wedge A - r_A \leq B - r_B \end{cases}$$

$$\mu(\tilde{A} \ \leq^{LF} \ \tilde{B}) = \begin{cases} 1 & if\ A < B \\ \dfrac{B - A}{r_B - r_A} & if\ A \leq B \wedge A - r_A > B - r_B \\ 0 & if\ A \leq B \wedge A - r_A \leq B - r_B \end{cases}$$

where $\tilde{A} = (A, r_A)$ and $B = (\tilde{B}, r_B)$ are fuzzy points from $\mathcal{H}$, A is the core of the point $\tilde{A}$, r_A is the parameter determining the membership function of the point $\tilde{A}$, B is the core of the point $\tilde{B}$, and r_B is the parameter determining the membership function of the point $\tilde{B}$.

Also, we have formulated the following two theorems and proved the first one (the proof of the second is analogue to the first).

Theorem 1. Let $T_M -$ equivalence $E: \mathcal{H} \times \mathcal{H} \rightarrow [0,1]$ is given by

$$E(\tilde{A}, \tilde{B}) = \begin{cases} 1 & if\ A = B \ \wedge \ r_A = r_B \\ 0 & otherwise \end{cases}$$

and a minimum $T_M -$ norm $(T_M(a, b) = min(a, b))$. Then the fuzzy relation $\leq^{RF}$ is an ordering compliant with Definition 4.

Proof.

The following notation will be used in the proof. With A, B, C we shall denote the cores of the fuzzy points $\tilde{A}$, $\tilde{B}$, $\tilde{C}$ and with r_A, r_B, r_C corresponding maximal distances from the cores for which the membership functions are note zeroes.

What we have to prove here is that the relation $\leq^{RF}$ has the following properties:

(i) E Reflexivity $E(\tilde{A}, \tilde{B}) \leq \mu(\tilde{A} \leq^{RF} \tilde{B})$

(ii) T_M E Anti-symmetry
$$T_M\left(\mu(\tilde{A} \leq^{RF} \tilde{B}), \mu(\tilde{B} \leq^{RF} \tilde{A})\right) \leq E(\tilde{A}, \tilde{B})$$

(iii) T_M Transitivity
$$T_M\left(\mu(\tilde{A} \leq^{RF} \tilde{B}), \mu(\tilde{B} \leq^{RF} \tilde{C})\right) \leq \mu(\tilde{A} \leq^{RF} \tilde{C})$$

***E*-Reflexivity**

If $E(\tilde{A}, \tilde{B}) = 0$, the proof is trivial. If not, from $E(\tilde{A}, \tilde{B}) = 1$, follows $A = B \wedge r_A = r_B$ having the consequence $\mu(\tilde{A} \leq^{RF} \tilde{B}) = 1 \Rightarrow 1 \leq 1$.

T_M E-Anti-symmetry

The proof is trivial for $E(\tilde{A}, \tilde{B}) = 1$.

If $E(\tilde{A}, \tilde{B}) = 0$, it should be proved that $\mu(\tilde{A} \leq^{RF} \tilde{B}) = 0 \vee \mu(\tilde{B} \leq^{RF} \tilde{A}) = 0$ is true.

Suppose that $\mu(\tilde{A} \leq^{RF} \tilde{B}) \neq 0 \vee \mu(\tilde{B} \leq^{RF} \tilde{A}) \neq 0$ is true. By substituting $\mu(\tilde{A} \leq^{RF} \tilde{B}) = a$ and $\mu(\tilde{B} \leq^{RF} \tilde{A}) = b$ we distinguish four cases.

1) $\boldsymbol{a = 1 \wedge b = 1}$**:**

$(A \leq B \wedge A + r_A \leq B + r_B) \wedge (B \leq A \wedge B + r_B \leq A + r_A)$ implies $A = B \wedge r_A = r_B \Rightarrow E(\tilde{A}, \tilde{B}) = 1$, which is not possible due to the assumption that $E(\tilde{A}, \tilde{B}) = 0$.

2) $\boldsymbol{a = 1 \wedge b < 1}$**:**

$(A \leq B \wedge A + r_A \leq B + r_B) \wedge (B \leq A \wedge B + r_B \leq A + r_A)$ implies $A = B \wedge r_A \leq r_B \wedge r_B > r_B$, which is not possible.

3) $\boldsymbol{a < 1 \wedge b = 1}$: Analogous to the proof of case 2).

4) $\boldsymbol{a < 1 \wedge b < 1}$**:**

$(A \leq B \wedge A + r_A > B + r_B) \wedge (B \leq A \wedge B + r_B > A + r_A)$ implies $A = B \wedge r_A > r_B \wedge r_B < r_A$, which is not possible.

T_M Transitivity

It should be proven that, for each $\tilde{A}, \tilde{B}, \tilde{C}$ from $\mathcal{H}$ defined over R^1, $min\left(\mu(\tilde{A} \leq^{RF} \tilde{B}), \mu(\tilde{B} \leq^{RF} \tilde{C})\right) \leq \mu(\tilde{A} \leq^{RF} \tilde{C})$ holds. With adopted notation, we can distinguish two cases:

1) $A > B \vee B > C$, and
2) $A \leq B \wedge B \leq C$.

For $A > B \vee B > C$ the proof is trivial because $\mu(\tilde{A} \leq^{RF} \tilde{B}) = 0$ or $\mu(\tilde{B} \leq^{RF} \tilde{C}) = 0$.

For $A \leq B \wedge B \leq C$ we distinguish two cases:

1) $A + r_A \leq C + r_C$ and
2) $A + r_A > C + r_C$.

1) The proof is trivial because of $\mu(\tilde{A} \leq^{RF} \tilde{C}) = 1$.

2) Let $\mu(\tilde{A} \leq^{RF} \tilde{C}) = \frac{C-A}{r_C - r_A} = a < 1$.

Then, the inequality $min\left(\mu(\tilde{A} \leq^{RF} \tilde{B}), \mu(\tilde{B} \leq^{RF} \tilde{C})\right) \leq a$ is true if $\mu(\tilde{A} \leq^{RF} \tilde{B}) \leq a$ is true or $\mu(\tilde{B} \leq^{RF} \tilde{C}) \leq a$ is true and, again, three cases emerge:

(i) $A + r_A \leq B + r_B$;

(ii) $A + r_A > B + r_B \geq C + r_C$;

(iii) $B + r_{\boldsymbol{B}} < C + r_C$.

For the case (i) $\mu(\tilde{A} \leq^{RF} \tilde{B}) = 1$, and for the case (iii) $\mu(\tilde{B} \leq^{RF} \tilde{C}) = 1$ the consequence of the previous statements are the following three cases:

(i) $1 \leq a \vee \frac{C-B}{r_B - r_C} \leq a$;

(ii) $\frac{B-A}{r_A - r_B} \leq a \vee \frac{C-B}{r_B - r_C} \leq a$;

(iii) $\frac{B-A}{r_A - r_B} \leq a \vee 1 \leq a$.

Let's continue with contradiction and suppose the opposite:

(i) $1 > a \vee \frac{C-B}{r_B - r_C} > a$;

(ii) $\frac{B-A}{r_A - r_B} > a \vee \frac{C-B}{r_B - r_C} > a$;

(iii) $\frac{B-A}{r_A - r_B} > a \vee 1 > a$.

Then for (i) holds that $\frac{B-A}{r_A - r_B} \geq 1 > a \wedge \frac{C-B}{r_B - r_C} > a$, while for (iii) holds that $\frac{B-A}{r_A - r_B} > a \wedge \frac{C-B}{r_B - r_C} \geq 1 > a$.

We can represent the cases (i), (ii), and (iii) as

$\frac{B-A}{r_A - r_B} > a \wedge \frac{C-B}{r_B - r_C} > a$,

This leads to the consequence

$B - A > a(r_A - r_B) \wedge (C - B) > a(r_B - r_C) \Rightarrow C - A > a(r_A - r_C)$, i.e., $\frac{C-A}{r_A - r_C} > a$.

However, this is not possible due to the assumption $\frac{C-A}{r_A - r_C} = a$ ∎

Theorem 2. Let $T_M -$ equivalence $E: \mathcal{H} \times \mathcal{H} \to [0,1]$ is given by

$$E(\tilde{A}, \tilde{B}) = \begin{cases} 1 & if\ A = B \ \wedge \ r_A = r_B \\ 0 & otherwise \end{cases}$$

and a minimum $T_M -$ norm $\left(T_M(a, b) = min(a, b)\right)$. Then the fuzzy relation $\leq^{LF}$ is an ordering compliant with Definition 4.

The proof of Theorem 2 is analogous to the proof of Theorem 1.

Figure 5 shows the mapping between the unordered set of fuzzy membership functions and the new, ordered classes.

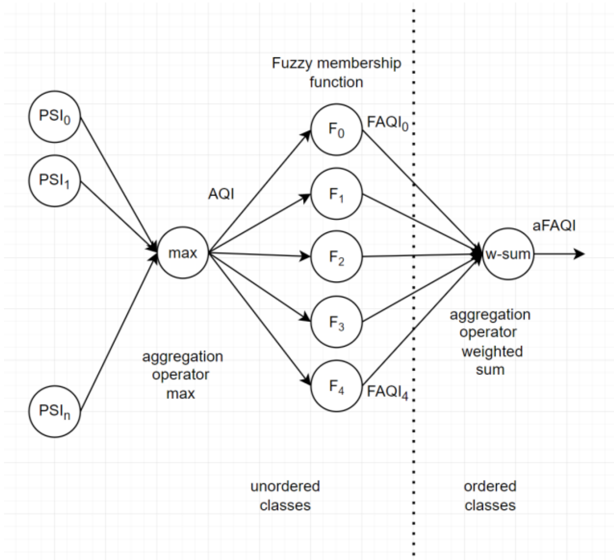

Figure 5. Modeling ordered classes

The mapping is modeled through the weighted sum aggregation operator, which is defined as

$$FAQI = \sum_{i=0}^{4} i \cdot FAQI_i(AQI)$$

where i is the integer number standing for the classes (**0** for **Very Low**, **1** for **Low**, **2** for **Medium**, **3** for **High** and **4** for **Very High**) and $FAQI_i(AQI)$ is the value of the function of the measurement sample belonging to class i .

That way we obtain the prediction model in which, instead of five output values, we have a single value that captures interactions of five independent quantities, while keeping the computational complexity same as that for unordered classes.

The resulting prediction model is shown in Figure 6.

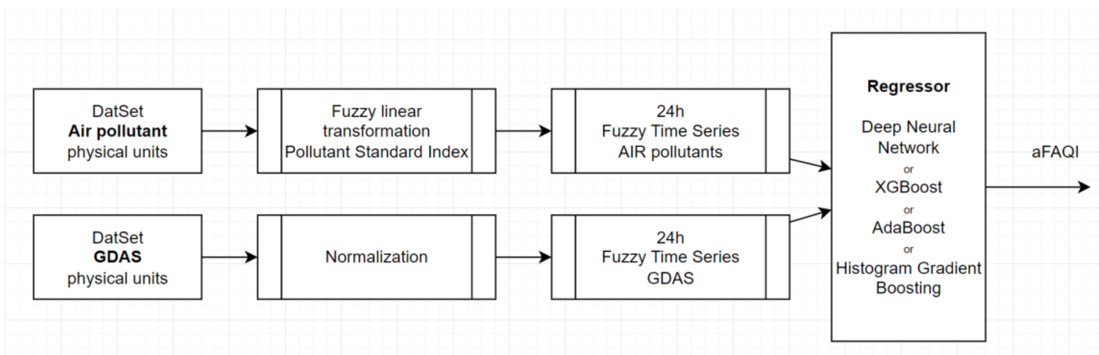

Figure 6. Prediction model with target variable modeled by ordered classes

To be able to compare the ordered and unordered case, it is necessary to calculate the total error as the mean value of the error of the example for all classes, which is calculated as:

$$Err = \frac{1}{n} \sum_{i=0}^{4} Err_i \cdot \sum_{j=0}^{n} FAQI_i(AQI_j)$$

V. MODEL APPLICATION AND SIMULATION RESULTS

To present the proposed model/methodology, we ran an experiment on a large and diverse data set. The used data set contains more than 82000 samples each with 385 real values. GDAS values are interpolated to fit five geo locations and merged with measurements of the concentration of the air pollutants.

For our experiment we used a virtual machine (VM) created as part of PARADOX HP Proliant SL250s with following components: Intel Xeon Processors E5-2670 (Sandy Bridge, 8 Core, 20M Cache, 2.6GHz), 106 nodes with total 1696 CPU cores and 32GB per node. Our VM is configured to use 8 cores and 32GB RAM on Debian GNU/Linux 10, Architecture x86-64.

All experiments were implemented using Python with Tensorflow 2.0, Pandas, XGBoost and Scikit-Learn main packages.

A. Data set

In this experiment, we used five data sets from five distinct locations in the USA, each in the same format. The sources of data are [32] (measurements of the concentration of the air pollutants) and [33] (meteorological data). Samples are indexed by temporal attribute, datetime, ranging from January 1, 2015. to December 31, 2021. All ten meteorological GDAS and six air pollutants are stored in a 24 hours' time slot with 1h sample rate (385 real values in total). Table III presents the data in more detail including sample sizes per location.

TABLE III. DATA SETS

Data set ID	site	Samples
11-001-0043	Washington, DC	27,981
13-089-0002	Near Atlanta, GA	21,468
18-097-0078	Indianapolis, IN	16,774
22-033-0009	Baton Rouge, LA	6,569
32-003-0540	Las Vegas, NV	9,665

The same source supplies data about land use (COMMERCIAL, RESIDENTIAL) and type of location (URBAN, SUBURBAN) as shown in Table IV.

TABLE IV. SITE TYPES

Data set ID	City	Land use	Location
11-001-0043	Washington, DC	COMMERCIAL	URBAN
13-089-0002	Near Atlanta, GA	RESIDENTIAL	SUBURBAN
18-097-0078	Indianapolis, IN	RESIDENTIAL	SUBURBAN
22-033-0009	Baton Rouge, LA	COMMERCIAL	URBAN
32-003-0540	Las Vegas, NV	RESIDENTIAL	URBAN

PSI calculation was done using PSI functions (Table V), which transform the physical value domain into a real value interval [0, 500].

TABLE V. PSI BREAKPOINTS

PSI	PM10 μg/m^3	SO2 ppm	CO ppm	NO2 ppm	O3 ppm
0	0	0	0	0	0
50	50	0.03	4.5	-	0.06
100	150	0.14	9	-	0.12
200	350	0.3	15	0.6	0.2
300	420	0.6	30	1.2	0.4
400	500	0.8	40	1.6	0.5
500	600	1	50	2	0.6

B. Fuzzy air quality index

In this example of framework application, the fuzzy air quality index is modelled via a simple max aggregation function applied to five *FPSI* indices of each criteria air pollutants:

$$FAQI = max\ (FPSI_{CO}, FPSI_{PM10}, FPSI_{NO2}, FPSI_{O3}, FPSI_{SO2})$$

Finally, we introduce a fuzzy linguistic variable (*Very low*, *Low*, *Medium*, *High*, *Very high*) defined by corresponding fuzzy sets, as depicted in Figure 7.

This fuzzy linguistic variable actually corresponds to the definition of CAQI, which describes air quality through these five categories. However, individual countries apply different scales (for example, Canada has a scale with four categories, while the USA has scale with six categories).

It is obvious that in our model this index can easily be adapted to specific needs.

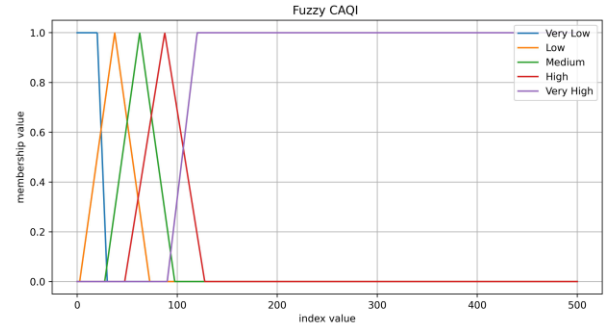

Figure 7. Fuzzy CAQI

C. ML experiments

In our experiments, we applied four multivariate regressors.

The first multivariate predictor regressor in this experiment is XGBoostRegressor with 24*10 GDAS and 24*6 air pollutant variables as input, and 5 real valued outputs, each corresponding to a single fuzzy set (FAQI_*very low* to *very high*), as depicted in Figure 3. Data set is split up into train (80%) and test (20%) subsets and trained with 1000 estimators with *max_depth* 4 and enabled early stopping method to avoid overfitting.

The second multivariate predictor regressor in this experiment is a deep neural network with 24*10 GDAS and 24*6 air pollutant variables as input, and 5 real valued outputs, each corresponding to a single fuzzy set (FAQI_*very low* to *very high*), with one hidden layer consisting of 20 Rectified Linear Units (ReLU) nodes. The activation functions in output layer are Sigmoid. The data set is split up into train (80%) and test (20%) subsets. Two dropout layers with 10% random filters are incepted between active layers to prevent overfitting.

The third multivariate predictor regressor in this experiment is ADA Boost regressor with 100 estimators, learning rate 1 and boosting algorithm SAMME.R.

The fourth multivariate predictor regressor in this experiment is Histogram gradient boosting regressor with squared error loss function, learning rate 0.1, max iterations 100, max leaf nodes 31, min samples leaf numbers 20, without regularization and max bins 255.

D. Simulation results

The mean absolute errors for FAQI prediction with unordered classes are shown in Table VI (XGBoost), Table VII (Deep Neural Network), Table VIII (ADA Boost), and Table IX (Histogram gradient boosting).

The tables show that all regressors behave similarly. Moreover, they are good in prediction for categories *Medium*, *High* and *Very high* and poor in prediction for categories *Very low* and *Low*. Having that in mind and the main purpose of the FAQI to alert of dangerous air pollution (*High* and *Very high*, possibly *Medium*), the results showed that further research was needed and justified.

TABLE VI. XGBOOST

Data set ID	Very low	Low	Medium	High	Very high
11-001-0043	0.229	0.230	0.049	0.003	0.001
13-089-0002	0.239	0.217	0.036	0.004	0.001
18-097-0078	0.213	0.213	0.065	0.006	0.001
22-033-0009	0.224	0.218	0.063	0.009	0.000
32-003-0540	0.083	0.236	0.185	0.045	0.012

TABLE VII. DEEP NEURAL NETWORK

Data set ID	Very low	Low*	Medium	High	Very high
11-001-0043	0.405	0.399	0.065	0.003	0.001
13-089-0002	0.460	0.398	0.045	0.004	0.002
18-097-0078	0.425	0.406	0.074	0.005	0.002
22-033-0009	0.426	0.392	0.056	0.008	0.001
32-003-0540	0.109	0.362	0.291	0.039	0.010

TABLE VIII. ADA BOOST REGRESSOR

Data set ID	Very low	Low*	Medium	High	Very high
11-001-0043	0.3336	0.328	0.1146	0.0195	0.0001
13-089-0002	0.3516	0.3140	0.0684	0.0095	0.0002
18-097-0078	0.339	0.3226	0.1221	0.048	0.001
22-033-0009	0.337	0.3283	0.096	0.046	0.0001
32-003-0540	0.1526	0.3282	0.2879	0.0965	0.035

TABLE IX. HISTOGRAM GRADIENT BOOSTING REGRESSOR

Data set ID	Very low	Low*	Medium	High	Very high
11-001-0043	0.2478	0.2574	0.0425	0.0031	0.001
13-089-0002	0.2608	0.2414	0.0332	0.0067	0.001
18-097-0078	0.2313	0.2492	0.0615	0.0062	0.001
22-033-0009	0.2397	0.2348	0.0549	0.0139	0.001
32-003-0540	0.083	0.2509	0.1794	0.0472	0.018

In the next step, a prediction model with ordered classes was applied. Using the same regressors, the results shown in Table X were obtained.

TABLE X. RESULTS OBTAINED WITH ORDERED CLASSES APPROACH

Data set ID	Deep neural network	XGBoost	AdaBoost	Histogram Gradient Boosting Regressor
11-001-0043	0.15353	0.1583	0.1558	0.1509
13-089-0002	0.11429	0.1209	0.1238	0.1092
18-097-0078	0.16035	0.1521	0.1517	0.1532
22-033-0009	0.15211	0.1539	0.1448	0.1460
32-003-0540	0.29561	0.3416	0.3210	0.3118

Finally, Table XI shows the comparative results obtained using two proposed models of the target variable.

TABLE XI. PERFORMANCES OF THE PREDICTION MODELS WITH UNORDERED AND ORDERED CLASSES

Data set ID	Mean absolute error	
	Unordered classes	Ordered classes
11-001-0043	0.40341	0.15353984
13-089-0002	0.45718	0.11429921
18-097-0078	0.42278	0.16035105
22-033-0009	0.42340	0.15211709
32-003-0540	0.11300	0.29561046

Simulation results for the target variable modelled as independent classes show that this model is characterized by a distinct property, which is a satisfactory performance for higher values of air quality index, and significantly worse (mean absolute errors higher for an order of magnitude) performance for lower values. In that case, the overall mean absolute prediction error was between 0.403 and 0.457 (except for the data set 32-003-0540 with mean absolute error of 0.11). This was a notable deficiency of the model calling for improvement that should ensure at least approximately equal performance for all categories.

Simulation results for the model in which the target variable is modelled as a set of ordered classes showed better performance. In this case, the overall mean absolute prediction error was between 0.114 and 0.160 (except for the data set 32-003-0540 with mean absolute error of 0.295), and the maximum error (again for the low pollution value classes) was of the same order of magnitude as for the high pollution value classes.

VI. CONCLUSION

This paper proposes a framework aimed at forecasting the aggregated air pollution index that is based on our Linear fuzzy space theory and fuzzy aggregation operators. The proposed model consists of two sub models. The first one models the concentration of pollutants, while the second one models multi-contaminant air quality index. We model the concentration of pollutants by regression using fuzzy time series of two groups of data: measured concentrations of pollutants and meteorological parameters with the target variable modeled in two ways. In the first case, the target variable consists of independent classes defined by proper membership functions, while in the second case, it is modeled by a set of classes connected by an ordering relation. The multi-contaminant air quality index is modeled as a fuzzy aggregation of PSI obtained via fuzzy linear transformation defined by fuzzy breakpoints.

Simulation results for the target variable modelled as unordered classes show that this model is characterized by a distinct property, which is a satisfactory performance for higher values of air quality index, and significantly worse (mean absolute errors higher for an order of magnitude) performance for lower values. This was a notable deficiency of the model calling for improvement that should ensure at least approximately equal performance for all categories. To improve the model performance, we model the target class as a set of ordered classes, which gave better performance in terms of mean absolute error.

However, air pollution is a result of an extremely complex and interdependent interaction among multiple factors (air pollutants, environment, time, climate conditions, etc.) additionally burdened with uncertainty and imprecision in data. This makes a single index an extremely rough approximation of the considered pollution situation.

Indeed, there is a potential for improvements in the research topics tackled in this paper, which shapes further research directions. The possible improvements could be further divided into two rough partitions.

The first, which is of fundamental kind, is about rethinking the air pollution index concept (for example, making it contextually dependent, or making it multidimensional). Recent research related to aggregation of sequence of fuzzy measures [34] and distortion functions [35] could provide for further improvement of the air pollution index modelling.

The second one is more about "technical" improvements of the model proposed in this paper: use of additional variables (like those in Table IV), training data balancing (in our experiment clearly indicated by results obtained for the data set 32-003-0540), learning shapes of membership functions from historical data, and alike. Improvements should specifically address creation of precision metrics in linear fuzzy space, enabling estimations of sensitivity of interval partitions selection in time series, aggregation models, and fuzzy sets parameters. Recent theory development (see [36], [37]) gives a method for identification of the optimal solution for convex and non-convex optimization in fuzzy approach that could help to do this.

The two partitions intersect at utilization of Artificial Intelligence (AI) methods, particularly fuzzy approach relying upon Linear fuzzy space and other stuff concerning imprecision and ambiguity management, and machine learning techniques.

ACKNOWLEDGMENT

The authors acknowledge funding provided by the Science Fund of the Republic of Serbia #GRANT No. 6524105, AI - ATLAS.

REFERENCES

[1] E. Pap, Đ. Obradović, Z. Konjović and I. Radosavljević, "Linear Fuzzy Space Based Framework for Air Quality Assessment," in *INTELLI 2022: The Eleventh International Conference on Intelligent Systems and Applications*, Venice, Italy, 2022.

[2] World Health Organization, "World health statistics 2016: monitoring health for the SDGs, sustainable development goals. 1.Health Status Indicators. 2.Global Health. 3.Health Priorities. 4.Mortality. 5.Universal Coverage. 6.Life Expectancy. 7.Statistics. I.World Health," World Health Organization, Geneve, Switzerland, 2016.

[3] C. Guerreiro, A. G. Ortiz, F. d. Leeuw, M. Viana, and A.Colette, "Air quality in Europe — 2018 report," European Environment Agency, Copenhagen, Denmark, 2018.

[4] Đ. Obradović, "Model of the computer simulation system for managing geospace under uncertainty conditions" - PhD Thesis, Novi Sad, Serbia: Faculty of Technical Sciences University of Novi Sad, 2011.

[5] Đ. Obradović, Z. Konjović, E. Pap, and A. Šoštarić, "The linear fuzzy space: Theory and applications," in *Artificial Intelligence: Theory and Applications. Studies in Computational Intelligence, vol 973*, E. Pap, Ed., Springer, Cham, 2021, pp. 227-253.

[6] Đ. Obradović, Z. Konjović, E. Pap, and I. Rudas, "Fuzzy Geometry in Linear Fuzzy Space," in *Intelligent Systems: Models and Applications, Revised and Selected Papers from the 9th IEEE International Symposium on Intelligent Systems and Informatics SISY 2011.* , vol. 3, E. Pap, Ed., Berlin, Heidelberg, Springer, 2013, pp. 137-153.

[7] Đ. Obradović, Z. Konjović, E. Pap, and I. J. Rudas, "Linear Fuzzy Space Based Road Lane Detection," *Knowledge Based Systems,* vol. 38, no. 1, pp. 37-47, January 2013.

[8] Đ. Obradović, Z. Konjović, E. Pap, and N. Ralević, "The maximal distance between imprecise point objects," *Fuzzy Sets and Systems,* pp. 76-94, 2011.

[9] R. S. Sokhi, N. Moussiopoulos, A. Baklanov, J. Bartzis, I. Coll, S. Finardi, R. Friedrich, C. Geels, T. Grönholm, T. Halenka, M. Ketzel, A. Maragkidou, V. Matthias, J. Moldanova, L. Ntziachristos, K. Schäfer, P. Suppan, G. Tsegas, G. Carmichael, and V. Franco, "Advances in air quality research – current and emerging challenges," *Atmospheric Chemistry and Physics,* vol. 22, pp. 4615–4703, Apil 2022.

[10] X. Tan, L. Han, X. Zhang, W. Zhou, W. Li, and Y. Qian, "A review of current quality indexes and improvements under the multi-contaminant air pollution exposure," *Journal of Environmental Management,* vol. 279, pp. 111681, 1 February 2021.

[11] W. Cheng, "Revised air quality index derived from an entropy function," *Atmospheric Environment,* vol. 38, no. 3, pp. 383-391, 2004.

[12] A. AL-Dhurafi, N. Masseran, Z. H. Zamzuri, and M. A. M. Safari, "Modeling the Air Pollution Index based on its structure and descriptive status," *Air Quality, Atmosphere & Health,* pp. 171–179, 2018.

[13] G. Kyrkilis, A. Chaloulakou, and P. Kassomenos, "Development of an aggregate air quality index for an urban mediterranean agglomeration: Relation to potential health effects," *Environment International,* vol. 33, no. 5, pp. 670-676, 2007.

[14] A. Fino, F. Vichi, C. Leonardi, and K. Mukhopadhyay, "An overview of experiences made and tools used to inform the public on ambient air quality," *Atmosphere,* vol. 12, no. 11, pp. 1: 1524, 2021.

[15] I. Atacak, N. Arici, and D. Guner, "Modelling and evaluating air quality with fuzzy logic algorithm-Ankara-Cebeci sample," *International Journal of Intelligant Systems and Applications in Engineering,* vol. 5, no. 4, pp. 263-268, 2017.

[16] M. H. Sowlat, H. Gharibi, M. Yunesian, M. T. Mahmoudi, and S. Lotfi, "A novel, fuzzy-based air quality index (FAQI) for air quality assessment," *Atmospheric Environment,* vol. 45, no. 12, pp. 2050-2059, 2011.

[17] J. Koo, S. Wong, G. Selvachandran, H. V. Long, and L. H. Son, "Prediction of Air Pollution Index in Kuala Lumpur using fuzzy time series and statistical models," *Air Quality, Atmosphere & Health,* no. 13, pp. 77–88, 2020.

[18] Y. Alyousifi, M. Othman, R. Sokkalingam, I. Faye, and P. C. L. Silva, "Predicting Daily Air Pollution Index Based on Fuzzy Time Series Markov Chain Model," *Symmetry,* vol. 12, no. 2, pp. 293, 2020.

[19] S. M. Cabaneros, J. K. Calautit, and B. R. Hughes, "A review of artificial neural network models for ambient air pollution prediction," *Environmental Modelling & Software,* vol. 119, pp. 285-304, 2019.

[20] S. M. Cabaneros and B. Hughes, "Methods used for handling and quantifying model uncertainty of artificial neural network models for air pollution forecasting," *Environmental Modelling & Software,* vol. 158, pp. 105529, 2022.

[21] B. Zhang, Y. Rong, R. Yong, D. Qin, M. Li, G. Zou, and J. Pan, "Deep learning for air pollutant concentration prediction: A review," *Atmospheric Environment,* vol. 290, no. 119347, 2022.

[22] Z. Wang, H. Li, H. Chen, Z. Ding, and J. Zhu, "Linear and nonlinear framework for interval-valued PM2.5 concentration forecasting based on multi-factor interval division strategy and bivariate empirical mode decomposition," *Expert Systems with Applications,* vol. 205, no. 117707, 2022.

[23] H. Relvas, D. Lopes, J. Ferreira, A. Silva, S. Rafael, M. Lopes, S. M. Almeida, V. Martins, E. Diapouli, A. Korhonen, O. Hänninen, M. Lazaridis, and A. I. Miranda, "Scenario analysis of strategies to control air pollution," *Urban Climate,* vol. 44, no. 101201, 2022.

[24] M. Grabisch, J. L. Marichal, R. Mesiar, and E. Pap, Encyclopedia of mathematics and its applications, series number 127 Aggregation functions, Cambridge, UK: Cambridge University Press, 2009.

[25] E. P. Klement, R. Mesiar, and E. Pap, Triangular norms, Springer Dordrecht, 2000.

[26] E. Pap, Ed., Artificial Intelligence: Theory and Applications, Springer Cham, 2021.

[27] E. Pap, Ed., Intelligent Systems: Models and Applications Revised and Selected Papers from the 9th IEEE International Symposium on Intelligent Systems and Informatics SISY 2011, Springer Berlin, Heidelberg, 2013.

[28] W. W. S. Wei, Multivariate Time Series Analysis and Applications, John Wiley & Sons Ltd., 2019.

[29] Q. Song and B. Chissom, "Fuzzy time series and its model," *Fuzzy Sets and Systems,* vol. 3, no. 54, pp. 269–277, 1993.

[30] S. Panigrahi and H. Behera, "Fuzzy Time Series Forecasting: A Survey," in *Computational Intelligence in Data Mining. Advances in Intelligent Systems and Computing*, H. Behera, J. Nayak, B. Naik and D. Pelusi, Eds., Springer, Singapore, 2020.

[31] A. Sahin, M. F. Dodurka, T. Kumbasar, E. Yesil, and S. Siradag, "Review Study on Fuzzy Time Series and Their Applications in the Last Fifteen Years," in *International Fuzzy Systems Symposium (FUZZYSS'15)*, Istanbul, Turkey, 2015.

[32] EPA United States Environment Protection Agency, "Emission & Generation Resource Integrated Database (eGRID2020)," Environmental Protection Agency, 13 October 2022. [Online]. Available: https://www.epa.gov/system/files/documents/2022-01/egrid2020_data.xlsx. [Accessed 1 February 2022].

[33] Air Resources Laboratory, "Global Data Assimilation System (GDAS1) Archive Information," Air Resources Laboratory, 21 January 2021. [Online]. Available: https://www.ready.noaa.gov/gdas1.php. [Accessed 10 January 2022].

[34] L. Nedović and E. Pap, "Aggregation of sequence of fuzzy measures," *Iranian Journal of Fuzzy Systems,* vol. 17, no. 6, pp. 39-55, 2020.

[35] L. Nedović, E. Pap, and Đ. Dragić, "Aggregation of triangle of distortion functions," *Information Sciences,* vol. 563, pp. 401- 417, 2021.

[36] S. Das and D. Chakraborty, An Introduction to Analytical Fuzzy Plane Geometry, Springer Cham, 2019.

[37] S. Das and D. Chakraborty, "Conceptualizing fuzzy line as a collection of fuzzy points," *Information Science,* vol. 598, no. 2, pp. 86-100, 2022.