IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 October 2023, accepted 14 December 2023, date of publication 21 December 2023,
date of current version 27 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3345466

RESEARCH ARTICLE

Understanding Climate Change and Air Quality Over the Last Decade: Evidence From News and Weather Data Processing

ALIN-GABRIEL VĂDUVA[1], MIHAI MUNTEANU[2], SIMONA-VASILICA OPREA[1], ADELA BÂRA[1], AND ANDREEA-MIHAELA NICULAE[1,3]

[1]Department of Economic Informatics and Cybernetics, Bucharest University of Economic Studies, 010374 Bucharest, Romania
[2]Department of Statistics and Econometrics, Bucharest University of Economic Studies, 010374 Bucharest, Romania
[3]Doctoral School of Economic Informatics, Bucharest University of Economic Studies, 010374 Bucharest, Romania

Corresponding author: Alin-Gabriel Văduva (vaduvaalin19@stud.ase.ro)

This work was supported by the Ministry of Research, Innovation and Digitization, Consiliul National al Cercetarii Stiintifice (CNCS)-Unitatea Executivă pentru Finanțarea Învățământului Superior, a Cercetării, Dezvoltării și Inovării (UEFISCDI), under Project PN-III-P4-PCE-2021-0334, within PNCDI III.

ABSTRACT Climate change is a phenomenon that is sometimes denied or trivialized. However, in recent years, we have faced extreme phenomena such as fires, floods, excessive temperatures, etc. which affect our physical and mental condition and the environment, often leading to significant material damage. To understand these problems and highlight the meteorological and phenomenological changes encountered in the last decade, time series were web-scraped and analyzed from several open data sources: weather news broadcast in Romania, air quality, temperature, etc. The extraction and organization of data recorded between 2009 and 2023 are formulated as a framework that can be reproduced and replicated to continue the monitoring. The exploratory analysis of the categorical and numerical data highlights intricate patterns and correlations within meteorological conditions across regions and seasons. From temperature trends to air quality fluctuations, the study underscores the dynamic interplay of weather phenomena, paving the way for informed forecasting and deeper climate research. At the same time, data processing includes Latent Dirichlet Allocation, K-prototype clustering analysis, in addition to K-means clustering with dimensional reduction techniques, all of which are employed to further reveal the extreme phenomena in news and regions with higher occurrence. Therefore, in this paper, we propose a data processing framework for multiple datasets and analytics, extracting valuable information on climate change and identifying the exposed regions to extreme phenomena.

INDEX TERMS Climate change, news, web scraping, NLP, data analysis, data clustering.

I. INTRODUCTION

The context of climate change is global. The 2030 agenda for sustainable development, unanimously adopted by all United Nations Member States in 2015, serves as a collective roadmap for promoting peace and prosperity for both humanity and the planet. This enduring framework is centered on the 17 Sustainable Development Goals (SDG), which represent an immediate and universal call to action for nations across the spectrum of development. Numerous papers have been dedicated to the SDG as they play a crucial role in sustainable development [1], [2]. The SDG underscore the interconnectedness of various challenges, emphasizing that eradicating poverty and addressing other forms of deprivation must be pursued in tandem with efforts to enhance healthcare and education, mitigate inequality, stimulate economic growth, all while combating climate change and preserving the integrity of our oceans, agricultural land, and forests [3], [4], [5]. This holistic approach aims to create a better world for present and future generations.

Tackling climate change is an important objective for the European Commission (EC), as global temperatures

The associate editor coordinating the review of this manuscript and approving it for publication was Hocine Cherifi.

(especially heat waves) have risen dramatically and the last decade from 2011 to 2020 was the warmest on record. Out of the 20 warmest years on record, 19 have occurred since the year 2000. According to data from the Copernicus Climate Change Service, 2022 marked the hottest summer and the second warmest year ever recorded. The prevailing scientific consensus attributes this trend to the increase in Greenhouse Gas (GHG) emissions resulting from human activities. Today, the average global temperature is 0.95 to 1.20°C higher than it was at the close of the 19th century. Scientists consider a 2°C increase compared to pre-industrial levels as a critical threshold with potentially disastrous consequences for the climate and the environment [6], [7]. This is the reason why the international community collectively recognizes the imperative to limit global warming to well below a 2°C increase. European Union (EU) climate action holds significance as climate change is already exerting various effects on Europe, with consequences varying by region. These effects encompass biodiversity loss, reduced crop yields [8], forest fires, and rising temperatures (heatwaves) [9]. Furthermore, climate change can have adverse impacts on human health [10], [11]. As of 2019, the EU ranked as the world's fourth-largest emitter of GHG, following China, the United States and India. The EU's contribution to global GHG emissions decreased from 15.2% in 1990 to 7.3% in 2019 [12]. Electric vehicles and Renewable Energy Sources (RES) are seen as two important pillars in combating GHG emissions [13]. In 2021, the EU legally committed to achieving climate neutrality, which means zero net emissions, by the year 2050 [14]. An interim goal of reducing emissions by 55% by 2030 was also established. This commitment to zero net emissions is codified in climate legislation, with the European Green Deal (EGD) serving as the roadmap for the EU's journey to becoming climate-neutral by 2050. Researchers have investigated the role of photovoltaics for the EGD and the recovery plan [15]. Geopolitics also plays an important role [16]. The specific legislation required to meet the EGD's objectives is outlined in the Fit for 55 packages introduced by the EC in July 2021. This package includes the revision of current laws related to GHG emissions reduction and energy consumption [17]. Additionally, the EU is actively working toward achieving a circular economy by 2050, fostering a sustainable food system and safeguarding biodiversity. To fund the EGD's ambitious goals, the EC unveiled the Sustainable Europe Investment Plan in January 2020, with the aim of attracting a minimum of €1 trillion in public and private investments over the next decade.

Severe climate changes gradually increased until it became evident that they are more likely to intensify and disturb humans' activities in the next few years. First seasonal changes were noticed, then more phenomena that create damage emerged indicating clear changes that we do not know if they would be reversible or not. More and more studies acknowledged that it would be a great challenge to make people aware of the environmental issues [18]. Our motivation is to highlight these phenomena through the analysis of historical data and through the identification of patterns whose knowledge will probably allow us to avoid natural disasters. As the years pass, people have become used to these changes and consider them natural, taking measures to withstand heat waves or trying to find ways to fight the hail that destroys their crops. But the extent of these phenomena has experienced an alarming intensification, the measures to combat being useless in the way of floods or destructive fires. To understand these phenomena and their impact on humans, big data analytics were proposed [19], emphasizing the climate change impact on supply chains. The researchers identified challenges and opportunities regarding these operations and proposed research themes on big data analytics and climate change to ease the transition to a clean environment.

In this paper, we build a dataset from weather news broadcast in Romania and meteorological data including air quality variables and aim to analyze and extract insights that could indicate a climate change trend or regions where some phenomena recently intensified and are more frequent. Exploratory Data Analysis (EDA), clustering, and Latent Dirichlet Allocation are applied to extract insights, patterns, and the relevant topics from the news' text.

Our contribution is mainly to create a data processing framework that can be replicated to reproduce the analysis and obtain more results that allow society to take measures to protect our environment and prevent extreme phenomena such as fire, hail, floods, heat waves, etc. The remainder of this paper is structured as following: in section II, the most relevant similar research papers are depicted highlighting the previous contribution in the field; section III is dedicated to the proposed data processing framework and the methods; section IV provides the EDA, as well as the results obtained from the K-prototypes and PCA or t-SNE combined with K-means methods. In section V, the conclusion is drawn.

II. LITERATURE REVIEW

Information related to air quality, weather, extreme phenomena, and even climate change is readily available given the increasingly high data volumes from around the globe. As of 2022, the volume of generated data was estimated to be around 97 zettabytes [20]. With this wealth of data sources, both offline and online, and the continuous volumes of data generated every second, the topic of climate change can be more efficiently tackled and analyzed in-depth. Thus, analyzing this vast amount of data is essential to address and better understand climate change. Numerous research papers have been written regarding this major concern, and an even higher number of news articles focus on raising awareness about the importance of climate change. The study performed by [21] analyzes climate change thematic academic papers, in contrast with news coverage on the same topic, revealing the different focus of the two seemingly related articles. Academic papers predominantly focus on the quantitative and direct effects of climate change on agriculture [22], [23], [24] or water management [25], [26]. In contrast, most media outlets center on the societal impact [27], [28]. News articles cover a

much wider variety of topics [21], [27], especially regarding climate-related data, making them a valuable, abundant, and challenging source of information.

In the context of data accessibility, the internet has been essential in ensuring an increasing number of individuals are becoming informed about climate-related topics. Moreover, online, raw data found on the internet, often having an unstructured form, is more accessible to analyze in recent times, all due to innovations in the software domain [29]. With updated products, packages and even dedicated software, data is more easily Extracted, Transformed, and Loaded (ETL) into convenient, easy-to-use files or databases. To perform such extraction, newer articles use Web Scraping (WS) as a tool to obtain the needed data, in real-time [30], [31] or one-time-only [32], [33], [34]. WS tools used in academics mainly focus on two directions [34]: 1) data analysis and visualization for predictions and decision-making processes and 2) developing models for obtaining feasible, clean data, unattainable in the past. Regarding the first focus, data analysis and visualization, WS is sporadically utilized to extract meteorological data [34]. The data is used to create a dashboard containing weather forecasts (predictions for meteorological variables such as humidity, precipitations, and temperature) as a support for decisions related to future weather. Article [35] uses WS tools to extract air quality data, such as pollutants (PM_{10}, $PM_{2.5}$, CO, CO_2, SO_2, O_3, NO, H_2S, NO_x) and Air Quality Index values to accurately predict AQI using the collected data. For the second focus of developing data extraction models, some authors [13], [14] use WS to efficiently collect and organize news content. With the increasing volume of news articles, manual inspection becomes impractical, prompting the need for automated solutions. Article [13] extracts news from 15 Senegalese websites, storing date, URL, and body text in separate databases. Article [14] employs bots to extract data from multiple sources, providing users with the latest news based on their preferences through a unified database.

Simply acquiring data from the internet through WS proves insufficient. It often leads to datasets that are hard to manage and even harder to replicate, should one add another source to the scraping tool. In order to properly assess the extracted information, two concepts have predominant uses in this domain: mapping (with its main purpose of process replication) [29] and Natural Language Processing – NLP (used to extract pertinent information from the collected data) [36], [37], [38]. Article [29] presents the importance of mapping the data obtained through WS, to ensure alignment with the scope of the research. On the other hand, NLP techniques are automatic processes used to analyze extensive text corpora. Article [36] uses an automated NLP process from the entire Corpus of the Assessment Reports of the Intergovernmental Panel on Climate Change. Through tokenization, stemming, and determining N-grams the authors obtained the most used climate change-related terms. These texts highlight the frequency of references to developed countries, sea levels, greenhouses, energy and the climate influence altogether. While the application of WS and NLP techniques in climate change research is crucial, a similar approach is also being adopted in Romania, where most research papers [39], [40], [41] use WS tools along with NLP for collecting news articles related to fake news [39], [40] or for trend analysis [41].

In article [41], authors extract and classify news articles using various semantic models, emphasizing weather-related topics. Key terms include: ''grad'' (degree), ''temperature'' (temperature), ''maximă'' (maximum), ''precipitations'' (precipitations), ''ninsoare'' (snowfall), ''slab'' (weak), ''vânt'' (wind), ''minima'' (minimum). In contrast, article [39] presents an automated scraping model using the PHP library Text Language Detect to extract only Romanian articles. With an emphasis on fake news detection, the article consists of several supervised machine-learning techniques, such as classical models (Naïve Bayes, Support Vector Machine), deep learning models (LSTM, GRU and CNN) and transformer models (RoBERT). CNN and NB models achieve the highest accuracy. In another study [40], the authors focus extensively on the BERT (Bidirectional Encoder Representations from Transformers) model, while using collected data using BeautifulSoup, a Python tool. As previously noted, BERT does not offer high accuracy when it comes to identifying fake news. While BERT provides powerful insights for contextual understanding, another NLP technique is lately gaining interest [42], [43] for comprehensive text analysis: Latent Dirichlet Allocation (LDA). Using a probabilistic approach, researchers use LDA to uncover various topics existing in the corpus. LDA analysis is present in papers concerning climate change [44], [45], [46], [47], [48]. One research of interest [44] uses LDA to analyze news media coverage on climate change and discover the increasing media attention on this topic from 2007 onwards.

However, before diving into different analysis techniques, it is vital for air quality and climate-related data to be processed beforehand [49], [50], [51], [52], [53]. Pre-processing serves the essential purpose of data preparation, cleaning and enhancement, ensuring a higher analysis quality [50], [51], [52], [53]. Some works in the literature use pre-defined preprocessing techniques [49], [53], [54], such as fuzzy logic or ANFIS. Other works focus on using their own pre-processing analysis steps [50], [51], [52], including relevant data manipulation: from normalizing and scaling data to principal component analysis and feature extraction. Most works go even further and present a framework for cleaning and preparing data in more clear and concise steps [50], [55], [56], [57], [58]. Building upon the insights gained from LDA's topic modeling, while using pre-processed data, the next analysis step often involves clustering techniques. One such clustering method is K-means, a simple yet useful unsupervised learning technique, frequently addressed in climate [59], [60] or air pollution analysis [61], [62]. K-means main drawback comes from its incapacity to process categorical data. To address this deficiency, several papers have been focusing on using

K-Prototypes [63], [64], [65], [66], a different clustering algorithm that can successfully handle mixed data. However, regarding climate change and air pollution topics, there are certain limitations in the existing literature, as there are not nearly enough analyses performed using the K-Prototypes algorithm.

III. METHODS

A. DATA PREPROCESSING FRAMEWORK

Meteorological datasets offer a compilation of weather and atmospheric data. They encapsulate a diverse array of meteorological indicators, such as temperature, precipitation, humidity, and atmospheric pressure, to name a few. While the timespan of such datasets can vary widely, ranging from immediate real-time data to extensive historical archives covering vast epochs, the dataset in the current investigation covers the period from February 2009 to August 2023. The principal aim of this study is to describe weather patterns, and the ramifications of climate change specific to Romania. By considering the capabilities of advanced NLP methodologies, particularly leveraging the OpenAI API linked with the renowned ChatGPT, alongside with web scraping mechanisms, this study extracts meteorological insights from news articles regarding weather events in Romania. Integrating these with data associated with Romania's primary regions, a unified dataset emerges. To pre-process this data, the following steps are proposed:

1) STEP 1 - SET UP THE SOURCES, PARAMETERS, TIME SPAN AND KEYS

In the context of meteorological research, consider a sequence of data sources $s = \{s_1, s_2, s_3\}$, a set of corresponding parameters p_s representative of attributes, such as air quality, temperature, and humidity, and a time frame T, $\forall t \in T$, representing the time interval of interest. The sequence of parameters p_s can be divided into three environmental datatypes corresponding to three data sources:

- Air quality data ($p_{s1} \in s_1$): describing the quantity of chemical compounds at time t, their origin being a data source s_1 from the sequence of sources denoted by s;
- Meteorological parameters ($p_{s2} \in s_2$): describing the meteorological conditions at time t, also originated from the data source s_2;
- Meteorological phenomena ($p_{s3} \in s_3$): sequence of events observed at time t_i, from time t, originated from the data source s_3;

Ideally, data corresponding to these parameters should be accessible across all geographic divisions within a country. Let $r = \{'Banat', 'Transilvania', 'Moldova', 'Muntenia', 'Oltenia', 'Dobrogea'$ be the set containing the studied geographic regions from Romania. Given that a single geographic region can encompass multiple counties or districts, based on the state's administrative divisions, it is prudent to single out a representative county or administrative territory from each region r. The consolidated dataset, which incorporates the meteorological parameters, should be augmented with two additional columns: 'Region' and 'County/Administrative territory'. In this structured data architecture, the combination of the 'Date' and 'Region' columns corresponding to t and r will act as a composite primary key, a mechanism that will be instrumental for future integrations with secondary datasets.

2) STEP 2 - WEB SCRAPING FOR WEATHER ARTICLES

The additional dataset required for this research is sourced from a news portal featuring meteorological articles, ensuring temporal alignment with the period established in the first dataset. The dataset is structured with the following columns:

- article_url: This signifies the article's web link.
- article_title: This encapsulates the headline of the article.
- article_lead: A concise summary or overview of the article's content.
- article_text: This covers the main body or content of the article.
- article_date: Indicates the date when the article was published.

To obtain this data, web scraping methodologies are employed. Specifically, for the purposes of this study, the Scrapy Python library is employed. Scrapy, an open-source framework, facilitates efficient extraction of web data. To operationalize this, we design a web scraper within Scrapy by establishing a Spider class, a crucial component that facilitates the extraction of website content. After this, the parsing logic is necessary. It is imperative at this point to meticulously define CSS selectors, ensuring that data extraction is both precise and comprehensive.

3) STEP 3 - PREPROCESS THE WEATHER ARTICLES

Upon completing the initial data extraction and populating the columns, the OpenAI API is employed to map the regions cited in the article_text column. Additionally, the associated meteorological phenomena will be identified. The specific model used for this task is the gpt-3.5-turbo-0613, with its temperature parameter set to 0 and a token limit set at 256. A critical aspect of this phase is the creation of the system prompt. The outcomes produced by the Large Language Model (LLM) have a critical effect on the prompt's clarity and precision. This strategic approach is referred to as "Prompt Engineering." Aptly crafted prompts ensure more accurate and relevant results. Multiple prompts are engineered, and the results are compared to check the reliability of the results and their conformity with the expected outputs of the LLM model. Outputs rendered by the LLM are conserved in Python lists of dictionaries. The generic format adopted is:

$$[\{'region'_1 : ['phen'_1, 'phen'_2, \ldots, 'phen'_i, \{'region'_2 : ['phen'_1, 'phen'_2, \ldots, 'phen'_j]\}, \ldots, 'region'_n : ['phen'_1, 'phen'_2, \ldots, 'phen'_k]\}].$$

In this case, each unique region may be associated with an array of different phenomena, or possibly no mentioned

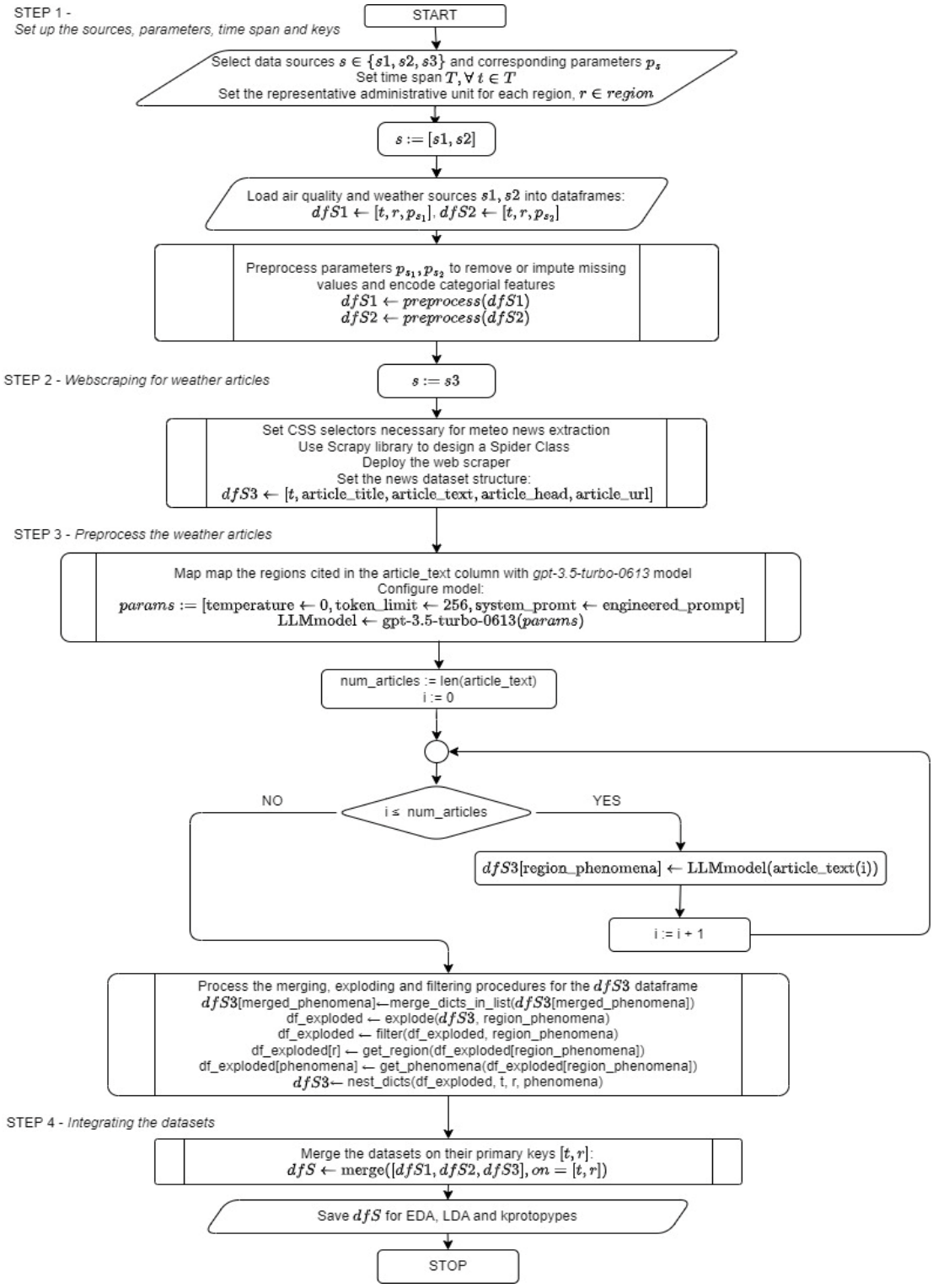


FIGURE 1. Flowchart of the data pre-processing framework.

phenomena at all. In cases where an article contains no relevant data, the LLM is instructed to generate a list encapsulating an empty dictionary. In the process of analyzing the dataset, it has been observed that multiple articles share an identical date in the article_date column. Such occurrences are not uncommon, especially in daily news portals that may publish several meteorological articles on the same day, each focusing on different regions or phenomena. However, these

overlapping dates present a potential challenge, as the ultimate objective is to have a structured dataset where each date corresponds to a unique set of meteorological phenomena across various regions. The procedure can be outlined as follows:

- Identification: Start by identifying all the dates with multiple articles. This can be achieved using Python's Pandas library, particularly by employing functions such as group by and count.
- Merging Dictionaries: For dates with multiple articles, merge the dictionaries to amalgamate regions and their corresponding phenomena. For example, if two entries on the same date cite ''rain'' and ''snow'' for ''Region_A'', the consolidated dictionary should list both phenomena under ''Region_A''.
- Redundancy Removal: Ensure that the merged dictionary for each region does not have duplicate phenomena.
- Creation of a Unified Entry: Once the dictionaries for a particular date have been merged, create a new, unified entry in the dataset. This entry should encapsulate all the regions mentioned across multiple articles for that date and their collective phenomena.
- Assuring the Composite Primary Key: This procedure ensures each date-region combination in the dataset remains unique, facilitating a structured composite primary key.

4) STEP 4 - INTEGRATING THE DATASETS

Upon the completion of the datasets, the first two containing the air quality data (dfS1) and meteorological parameters (dfS2) and the third containing the regions associated with the phenomena (dfS3), the final stage entails their integration. This merging is based on the shared composite primary key - comprising both date (t) and region (r) - present within each dataset. Using Python's Pandas merge function, datasets are consolidated based on the primary key, resulting in a unified dataset (dfS). Post-merging, a rigorous verification step is necessary to eliminate anomalies, ensuring the combined dataset's integrity. The outcome is a cohesive database, necessary for advanced meteorological analysis. The described steps are graphically represented in Figure 1.

B. K-PROTOTYPES ALGORITHM

In data analytics, effectively clustering mixed datasets, comprising both numerical and categorical variables, remains challenging. While K-means excels with numerical data and K-modes with categorical, neither is suited for mixed data. K-prototypes represents a hybrid approach combining both algorithms' characteristics, ensuring a robust solution to cluster data having mixed types. Let X denote the dataset necessary for clustering, $X = \{X_1, X_2, X_3, \ldots, X_n\}$, where X_i is a data object or datapoint, with $1 \leq i \leq n$. We can represent a data object X_i as a vector $X_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$. The K-Prototypes algorithm allows the usability of both numerical and categorical attributes in the clustering process. The aim of the K-prototypes algorithm is to find k groups by minimizing the following defined cost function:

$$J(P, Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} p_{il}\, (x_i, Q_l) \tag{1}$$

where p_{il} is an element from the partition matrix $P_{n\times k}$, $0 \leq p_{il} \leq 1$, Q_l is the center of the cluster denoted by l, and $d(x_i, Q_l)$ represents the dissimilarity measure defined as:

$$d(x_i, Q_l) = d\left(x_{ij}, q_{ij}\right) \tag{2}$$

where:

$$d\left(x_{ij}, q_{ij}\right) = \begin{cases} \left(x_{ij} - q_{ij}\right)^2, & \text{if feature } l \text{ is numerical} \\ \mu_l \delta\left(x_{ij}, q_{ij}\right), & \text{if feature } l \text{ is categorical} \end{cases} \tag{3}$$

and:

$$\delta(p, q) = \begin{cases} 0, & p = q \\ 1, & p \neq q \end{cases} \tag{4}$$

representing the Kronecker symbol.

The term μ_l represents the weight associated for the categorical attribute located in the cluster l. If x_{ij} is a value of a numerical feature, then q_{ij} will be the mean of the j-th numeric feature in the cluster denoted by l. When x_{ij} is a value of a categorical variable, q_{ij} is the mode of the j-th categorical variable in the cluster l [67].

There are four major steps representative for the K-prototypes algorithms:

1. Randomly select K datapoints as initial prototypes/centers from the dataset;
2. For each datapoint in the main dataset X, assign it to the cluster with the nearest prototype to the selected datapoint with respect to (2);
3. After each allocation, update the cluster prototypes/centers.
4. If the updated prototypes are identical to the previous ones, then the algorithm stops. Else, steps 2 and 3 are repeated until convergence.

C. K-MEANS WITH PCA AND T-SNE

Data allows for conducting numerical analysis after encoding categorical variables. K Means, a widely used unsupervised machine learning clustering algorithm that specializes in numerical data, is one part of K-Prototype. On certain datasets, both algorithms can be applied to enhance our ability to gain insights from the data.

Given the dataset $X = X_i$ where $1 \leq i \leq n$, X is partitioned into 'k' clusters $C = C_j$ where $1 \leq j \leq k$. K-Means' main concept revolves around centroids μ_j calculation, data points representing each cluster's mean (center). The aim of this algorithm is to find the k clusters by minimizing the cost function:

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \epsilon c_k} \| x_i - \mu_k \|^2 \tag{5}$$

The steps needed to obtain these clusters and centroids are identical to the steps presented in section IV-B, with the difference in the equation used to calculate the cost function.

For an improved visualization, clustering algorithms often use dimensionality reduction techniques. Two such techniques standout: Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). PCA uses the concepts of Eigenvalues and Eigenvectors for linear dimensionality reduction, while t-SNE uses a non-linear, probabilistic approach that contains the concept of Perplexity.

PCA is a technique that tries to preserve as much information as possible from the original dataset, while also reducing the number of variables. It uses preprocessed data based on which it computes the covariance matrix. Using the newly obtained matrix, the unsupervised algorithm determines eigenvalues (degree of variability) and eigenvectors (PC axes) to create Principal Components that explain the maximum amount of variance of the original data. Given n number of variables, PCA will determine n principal components, each with its own eigenvalue and eigenvector. The optimal number of components can be chosen using various methods, either by specifying a variance threshold, applying Kaiser's Rule (retain only eigenvalues above 1), or by plotting a Scree Plot. In the end, the determined principal components can be used for statistical inference.

On the other hand, while the second technique focuses on dimensionality reduction as well, t-SNE is mostly used for visualization purposes. t-SNE uses high-dimensionality Euclidean distances that are converted into conditional probabilities using probability distributions. These conditional probabilities are often referred to as similarities between two points [68]. The conversion is made after assessing a Gaussian distribution in the original high dimensional space and a Student's t distribution in the desired low-dimensional space, and aiming to map them by matching similarities between the two. Perplexity is used in t-SNE to control the balance between preserving global and local structure in the data. Different perplexity values reveal different aspects of the data structure, as it states how many neighbors the algorithm should take into consideration when reducing the data.

IV. RESULTS

A. EXPLORATORY DATA ANALYSIS

The dataset under examination is meticulously structured, aiming to offer insights into diverse meteorological phenomena and variables recorded over varying dates across distinct regions and counties. The following enumerates the attributes and the nature of the information housed within each:

- **Date ('date'):** This field denotes the specific date on which the recorded weather phenomena were observed, formatted as YYYY-MM-DD. For example, the entry '2009-02-19' represents the 19th of February 2009.
- **Region ('region'):** The 'region' attribute specifies the geographical region of the observation. It is categorical and includes areas such as 'Banat', 'Dobrogea', and 'Moldova', enabling regional analyses of weather patterns and phenomena.
- **Phenomena ('phenomena'):** This attribute holds an array of the observed weather phenomena on the recorded date in the corresponding region and county. Examples of phenomena encompass 'Snow', 'Rain', 'Wind', 'Flood', 'Fog', etc., facilitating an exploration of the diversity and frequency of weather events.
- **County ('county'):** The 'county' field indicates the specific county within the region where the observation was made, represented by abbreviated codes such as 'TM' for Timiş, 'CT' for Constanţa, and 'IS' for Iaşi.
- **PM 10 Quality ('pm10_quality'):** This numeric attribute represents the concentration of particulate matter (PM10) in the air, measured in μg/m^3. It's critical for assessing air quality, with higher concentrations usually indicative of poorer air quality.
- **Air Pressure ('air_pressure'):** Measured in millibars (mbar), this numeric variable reflects the atmospheric pressure recorded at the time of observation, offering insights into the prevailing weather conditions, with lower values often associated with adverse weather.
- **Temperature ('temperature'):** This is a numeric variable capturing the ambient air temperature at the time of observation, measured in degrees Celsius (°C). It serves to elucidate temperature trends and anomalies within the dataset.
- **Humidity ('humidity'):** This field quantifies the relative humidity observed, represented as a percentage (%). It's pivotal for understanding moisture levels in the air and can be associated with the occurrence of specific weather phenomena like rain or snow.

Example Entry:

- Date: 2009-02-19
- Region: Banat
- Phenomena: ['Snow', 'Wind']
- County: TM
- PM10 Quality: 51.63 μg/m^3
- Air Pressure: 995.4 mbar
- Temperature: −4.095 °C
- Humidity: 85.5%

Each entry in the dataset presents a snapshot of the meteorological conditions on a particular day, in a specific county and region, thereby providing a multifaceted view of the weather patterns, atmospheric conditions, and air quality prevalent in the observed areas.

1) DISTRIBUTION OF VARIABLES

The distribution analysis of the variables unveils distinct patterns, providing a glance into the climatic nuances inherent to the dataset (as in Figure 2).

- PM 10 Levels: Displayed some skewness, predominantly concentrating around 20-60 μg/m^3, signaling potential concerns regarding air quality and public health.
- Air Pressure: Exhibited a more symmetrical distribution around 1000 mbar, a typical trait often linked to weather dynamics such as storms.
- Air Temperature: Revealed a bimodal distribution, hinting at the existence of two principal temperature clusters,

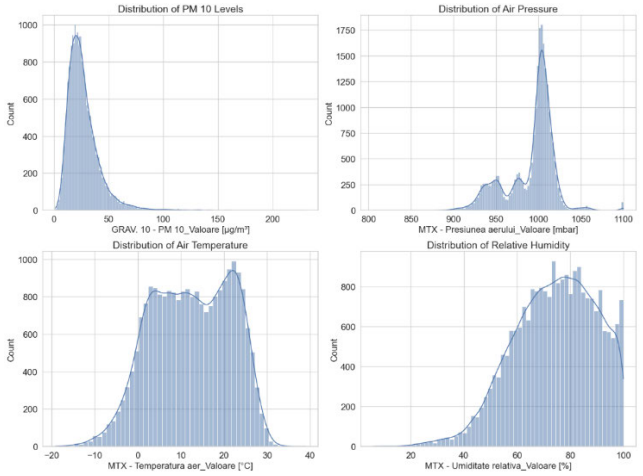


FIGURE 2. Distribution of the variables.

possibly correlating to varied seasons or distinct weather states.

- Relative Humidity: Manifested a slight left skew, with a majority clustering around 70-100%, a range often associated with precipitation phenomena.

2) ANALYSIS OF EXTREME VALUES AND CORRESPONDING PHENOMENA

a: LOW AIR PRESSURE ANALYSIS

Analysis under conditions of low air pressure disclosed frequent occurrences of phenomena such as snow, hail, and blizzards, validating the acknowledged association of low air pressure with severe weather conditions.

b: HIGH PM 10 LEVELS ANALYSIS

During elevated PM 10 levels, the prevalence of phenomena like rain, fog, snowfalls, and frost was observed. The persistence of rain, despite the elevation of PM 10 levels, implies it might not be effective in air purification. Similarly, occurrences of fog, snowfalls, and frost potentially indicate the entrapment of pollutants close to the ground.

c: TEMPERATURE EXTREMES ANALYSIS

High temperatures exhibited a strong association with heatwaves and unexpectedly high frequencies of hail, possibly implying sudden climatic transitions. In contrast, low temperatures showed a strong correlation with frost, snow, and blizzards, indicative of severe winter conditions.

d: HIGH RELATIVE HUMIDITY ANALYSIS

High relative humidity prominently correlated with rain and snow falls and, to a lesser extent, with sleet and wind. This is consistent with the expectation of precipitation during conditions of elevated moisture in the atmosphere.

3) CORRELATION MATRIX ANALYSIS

The correlation matrix is established to discern the degree of linear relationship between the numerical variables (as in Figure 3).

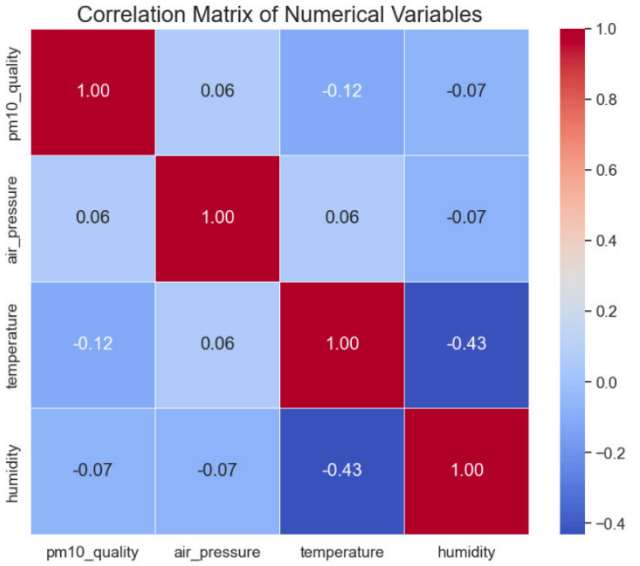


FIGURE 3. Correlation matrix.

- PM 10 Levels & Air Pressure: Exhibited a negligible positive correlation (0.0614), hinting at a marginal increase in PM 10 levels with rising air pressure.
- PM 10 Levels & Temperature: Portrayed a slight negative correlation (−0.1169), suggesting that rising temperatures may marginally decrease PM 10 levels.
- PM 10 Levels & Humidity: A trivial negative correlation (−0.0695) was noted, suggesting a minor decrease in PM 10 levels with increasing humidity.
- Temperature & Humidity: A moderate negative correlation (−0.4326) was observed, affirming the meteorological principle that an increase in temperature is typically accompanied by a decrease in humidity, the most significant correlation observed in this study.

4) REGIONAL TRENDS AND SEASONAL VARIATIONS

a: REGIONAL TRENDS OF NUMERICAL VARIABLES

In-depth regional analysis of numerical variables such as PM 10 levels, air pressure, air temperature, and relative humidity is conducted to discern regional patterns. Box plots are constructed to present regional distribution characteristics for each numerical variable, revealing substantial variations across regions (as in Figure 4).

b: SEASONAL VARIATIONS IN WEATHER PHENOMENA

Seasonal breakdowns of weather phenomena occurrences were generated by mapping each month to its respective meteorological season. The phenomena were then analyzed seasonally to discern the most frequent occurrences in each season over the years, revealing divergent patterns, where different phenomena dominated in varying seasons. In each season, the most frequent phenomena presented varying trends across the years (as in Figure 5).

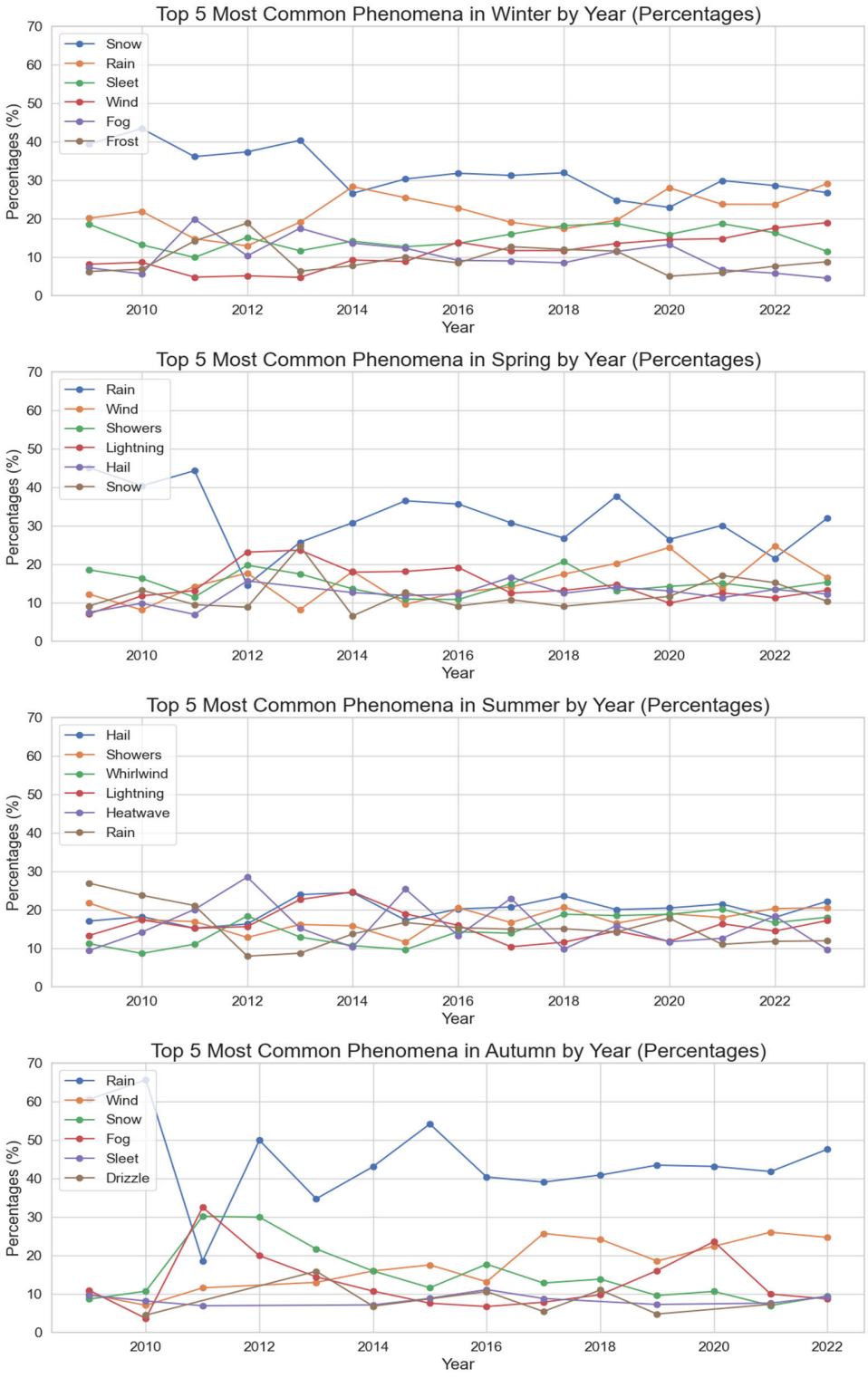


FIGURE 4. Seasonal trends of phenomena.

Winter:

- In the winter season, the most frequent weather phenomenon is "Snow," with notable variations in occurrence over the years.
- "Rain" and "Sleet" are also common occurrences during winter.
- The years 2017 and 2018 exhibit a significant increase in "Snow" occurrences, while 2013 has comparatively lower "Snow" occurrences.

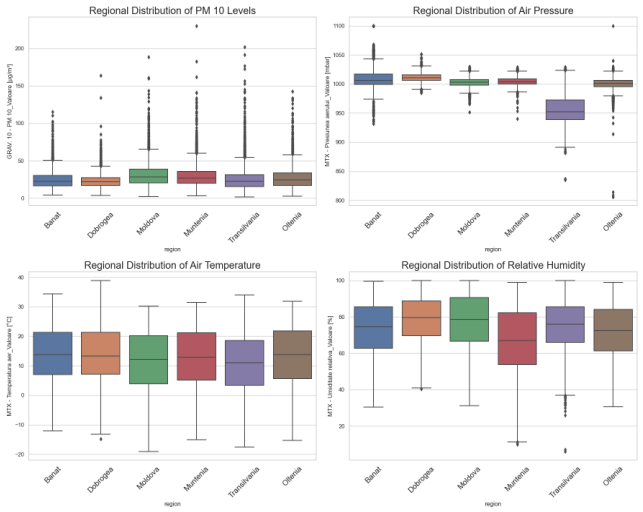


FIGURE 5. Regional box-plots.

- "Blizzard" is a prominent phenomenon in the winter of 2012 and 2017.

Spring:

- Spring exhibits a different set of dominant weather phenomena, with "Rain" being the most frequent.
- "Showers" and "Lightning" are also notable occurrences during spring.
- In 2021, there is a substantial increase in "Snow" occurrences during spring, which is relatively unusual for this season.
- "Hail" is more frequent in the springs of 2014 and 2015.

Summer:

- Summer seasons are characterized by weather phenomena such as "Hail," "Lightning," and "Heatwave."
- "Rain" and "Showers" are also common during summer.
- The years 2014 and 2015 have a higher incidence of "Lightning" and "Hail" during summer.
- "Whirlwind" and "Storm" are notable occurrences in the summer of 2016 and 2017.

Autumn:

- "Rain" dominates in the autumn season, with variable occurrences of other phenomena.
- "Fog" and "Wind" are also observed during autumn.
- In 2015, there is a notable increase in "Snow" occurrences during autumn.
- "Blizzard" is a rare phenomenon but occurred in the autumn of 2017.
- The years 2018 and 2019 exhibit higher occurrences of "Wind" in autumn.

These findings highlight the dynamic nature of weather patterns across different seasons and years. The variations in the prevalence of specific weather phenomena underscore the need for comprehensive seasonal analysis when studying meteorological data. Understanding these patterns can provide valuable insights for weather forecasting and climate research.

5) STATISTICAL TESTS AND TRENDS

A focused analysis is performed to observe the trend in the occurrences of heatwaves over the years. The Mann-Kendall Test is employed to discern whether the occurrences of heatwaves are increasing, decreasing, or remain constant over the years. The Mann-Kendall Test yields a τ value of 0.428 and a p-value of 0.027, suggesting a statistically significant increasing trend in the occurrences of heatwaves over the years in the dataset. To investigate if the occurrences of heatwaves are dependent on the season, a Chi-Square Test for Independence was conducted. The Chi-Square statistic is calculated to be approximately 4051.18, with a p-value of virtually zero. These results indicate a statistically significant relationship between the season and the number of heatwave occurrences, suggesting that heatwaves are not uniformly distributed across different seasons. To explore the relationship between the occurrences of heatwaves and meteorological parameters, a Spearman Rank Correlation test is performed specifically focusing on air temperature. The test yields a Spearman Correlation Coefficient of approximately 0.442 with a p-value of virtually zero. This indicates a statistically significant moderate positive correlation between air temperature and the occurrences of heatwaves, suggesting that rising air temperatures are associated with an increase in the number of heatwave events.

B. LATENT DIRICHLET ALLOCATION

Implementing LDA on the text of news, first the article_text column is translated with Translator from *googletrans*, the punctuation and stop words are removed and the letters are converted to lower case. The translation can be time consuming depending on the text length and the capacity of virtual machine provided by Google Collaboratory. Around 7,000 distinct news were analyzed, and the following three groups of topics emerged. The first major topic is characterized by several normal weather conditions (as in Figure 6). It is focused on temperature measurement units (degrees), consisting of a variety of weather indicators such as wind, sun, rain, clouds, snow, etc.

The second topic focuses on the cold weather, where snow, wind, precipitation, clouds maximum, values are predominant (as in Figure 7).

Whereas the third topic is smaller than the first two and is concentrated on weather with extreme phenomena like hail. It includes rain, snow, shower, clouds, electrical, hail, water, etc. as in Figure 8.

Topic coherence is a measure of how interpretable and coherent the topics produced by an LDA model are. It assesses how well the words within each topic are related. With 3 topics, the coherence score is 0.638. Increasing or decreasing the number of topics led to a lower coherence score. A higher coherence score indicates that the topics generated by the LDA model are more coherent and meaningful.

TABLE 1. Descriptive statistics of the clusters.

	PM 10			Air pressure			Temperature			Humidity		
Cluster	mean	max	min	mean	max	min	mean	max	min	mean	max	Min
C0	22.8	88.0	1.4	946.3	976.5	805.7	11.4	34.0	-17.6	77.4	100.0	28.1
C1	21.2	48.8	2.1	1006.5	1100.0	974.3	10.2	38.9	-16.3	85.6	100.0	51.0
C2	58.9	230.2	35.7	1000.7	1059.3	930.5	5.1	30.1	-19.1	79.4	100.0	24.6
C3	25.8	63.5	2.1	1000.4	1099.8	964.6	19.4	34.3	-14.2	60.3	82.0	6.0

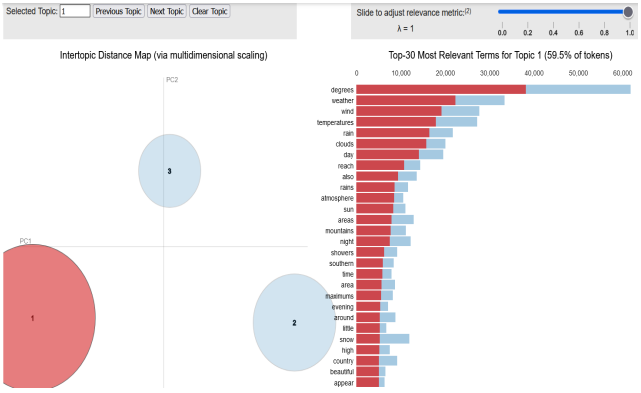


FIGURE 6. The first topic obtained with LDA.

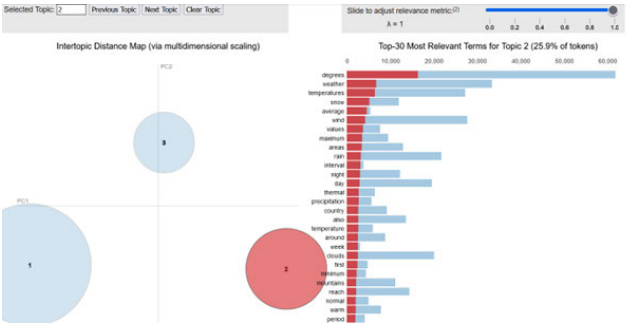


FIGURE 7. The second topic obtained with LDA.

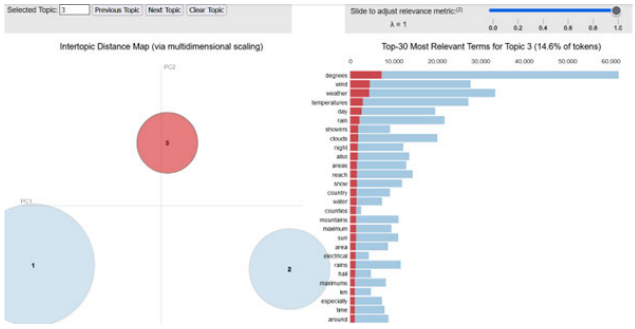


FIGURE 8. The third topic obtained with LDA.

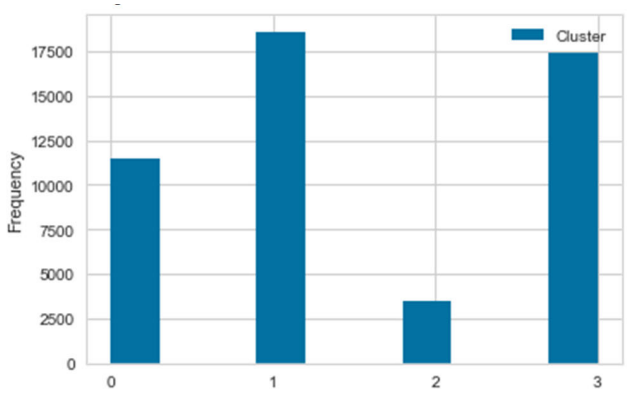


FIGURE 9. Distribution of the clusters.

C. CLUSTERING WITH K-PROTOTYPES

To determine the groups related to the occurrence of the phenomena in the selected regions, we applied K-prototypes model since the data is composed by numerical and categorial features. The following variables are considered for clustering: phenomena, pm10_quality, air_pressure, temperature, humidity. To decide on the number of clusters, we iteratively calculated the cost function using the Elbow method for up to k=11 clusters and finally set k=4. As a result, the records are grouped into 4 clusters, two with more than 17,000 records, one with 11,500 and one with 3,500 records (Figure 9).

Descriptive statistics are centralized in Table 1: As can be noticed, cluster C1 is characterized by low values of PM10, C0 and C3 have moderate values of PM10, while C2 has the highest values. These findings are also depicted in Figure 10. The air pressure is low in C0, while C2 is characterized by the lowest values of temperature. Related to the humidity, C1 has the highest values and C3 has the lowest values.

Further, we investigate the most frequent phenomena in each cluster during the seasons and the following findings are obtained and are depicted in Figure 11:

- C0 has the highest values of precipitations, especially in winter and autumn that includes snow, rain, and sleet. Also, wind has the highest occurrence during winter and autumn.
- C1 has moderate values of precipitations during all seasons, having a high occurrence of snow.
- C2 is characterized by low precipitation (rain, sleet, hail, and snow). In summer, it has the lowest values of precipitations but also has the lowest values of heatwave. During autumn and winter, it has a high occurrence of fog.

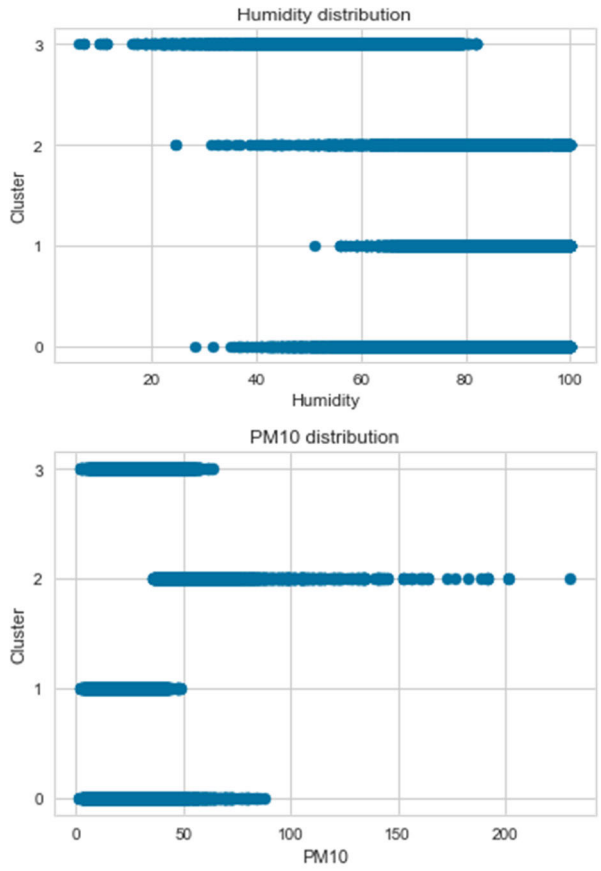


FIGURE 10. PM10 and humidity distribution on each cluster

- C3 is characterized by high precipitation, especially rain during summer, accompanied by high values of lightning, storm, hail and wind. Although it appears to be a precipitation-rich cluster during summer, it records the highest heatwave values that occur in autumn and spring, not just summer.

Additionally, we investigate the extreme or dangerous phenomena and how these extreme weather conditions are related to the clusters. The following extreme phenomena are considered: sleet, flood, blizzard, freezing rain, drizzle, whirlwind, heatwave, storm, hail, drought, cyclone, fires, tornado. An interesting finding is revealed by analyzing

Figure 12. Cluster C2 has the smallest occurrence of these extreme phenomena, C0 and C1 register the most extreme phenomena such as sleet, blizzard, and freezing rain, while C3 has the highest occurrence of hail, storm, whirlwind, heatwave, drought, and fire. It seems that C3 grouped the most dangerous weather phenomena for summer season, while C0 grouped the most dangerous phenomena for winter. For spring and autumn, C0 and C3 have an equal distribution of these extreme weather conditions.

To analyze how the clusters are distributed over the regions, in Figure 13, we plot the records and noticed that there is

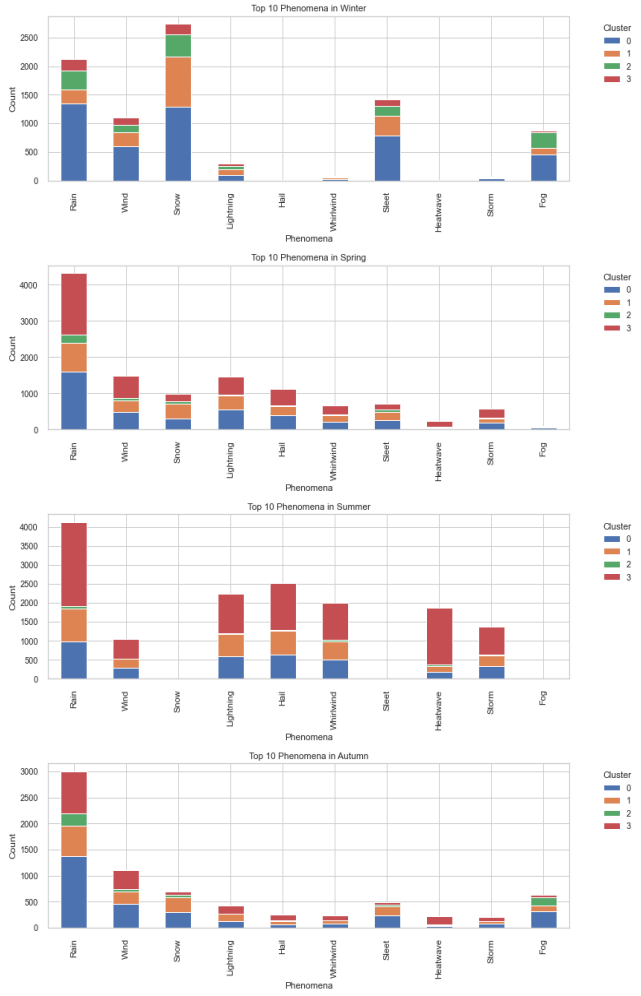


FIGURE 11. Phenomena occurrence during seasons on each cluster.

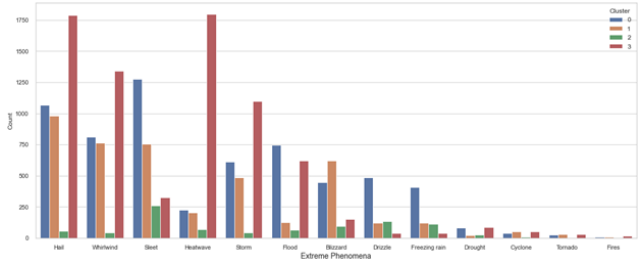


FIGURE 12. Extreme phenomena occurrence on each cluster.

a strong correlation. Thus, Transilvania is representative for C1, Muntenia and Oltenia are predominant in C3 followed by C0, while Moldova, Dobrogea and Banat are characterized by C0 and C3. C2 has the highest occurrence in Muntenia and Moldova.

Therefore, Muntenia and Oltenia are exposed to dangerous phenomena during summer, while Moldova, Dobrogea and Banat are more exposed to extreme weather conditions during winter. Transilvania is predominant in C1, having moderate weather conditions.

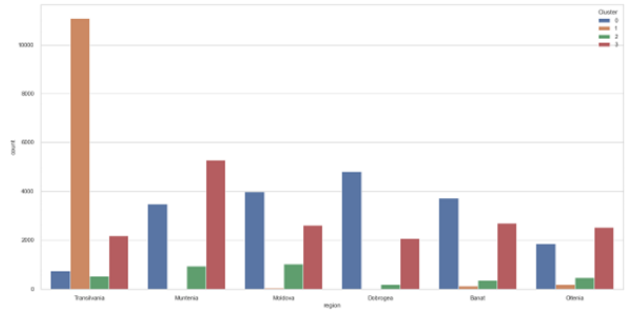


FIGURE 13. Clusters distributions over regions.

The extreme phenomena in each region are displayed in Figure 14. As can be noticed, hail is prevalent in all regions, followed by whirlwind and storm. Sleet is more frequent in Moldova, Muntenia and Transilvania, while heatwave is prevalent in Oltenia and Muntenia, but higher values are also encountered in Dobrogea and Moldova.

As a conclusion, C0 and C3 are identified as the representative clusters for extreme weather conditions in winter and summer respectively, grouping numerous records (11,500 in C0 and 17,500 in C3) and impacting most of the regions in Romania. Thus, in winter three regions, Moldova, Dobrogea and Banat are impacted by severe weather such as sleet, freezing rain, flood. In summer, Muntenia and Oltenia are impacted by extreme phenomena such as heatwaves, fires, storms, and hail.

D. CLUSTERING WITH PCA/T-SNE AND K-MEANS

After comprehensive pre-processing, the final dataset contains 7 mixed variables. Previously, the K-Prototype algorithm is used, as it excels with both numeric and categorical data, and valuable insight was gained. However, due to the limited number of categorical variables, they are encoded to enable the use of the K-means algorithm. In this section, we apply K-means with Principal Component Analysis (PCA) to gain a deeper understanding of the data. Furthermore, a t-SNE visualization is run to compare and validate the relationships within the dataset. After exploring the phenomena variable, we identified 28 different types of phenomena. To maintain interpretability, only extreme or dangerous phenomena are encoded and kept in the data. The variables used in the K-means clustering algorithm are: region, phenomena, pm10_quality, air_pressure, temperature, and humidity.

1) PRINCIPAL COMPONENT ANALYSIS

The initial step in applying the PCA algorithm involves data scaling, especially when using encoded variables, which typically have smaller values than other variables found in our dataset. For visualization purposes, two principal components were determined. The results of the PCA model are summarized in Table 2.

At first glance, the two calculated principal components explain slightly more than 60% of the data's variance, which

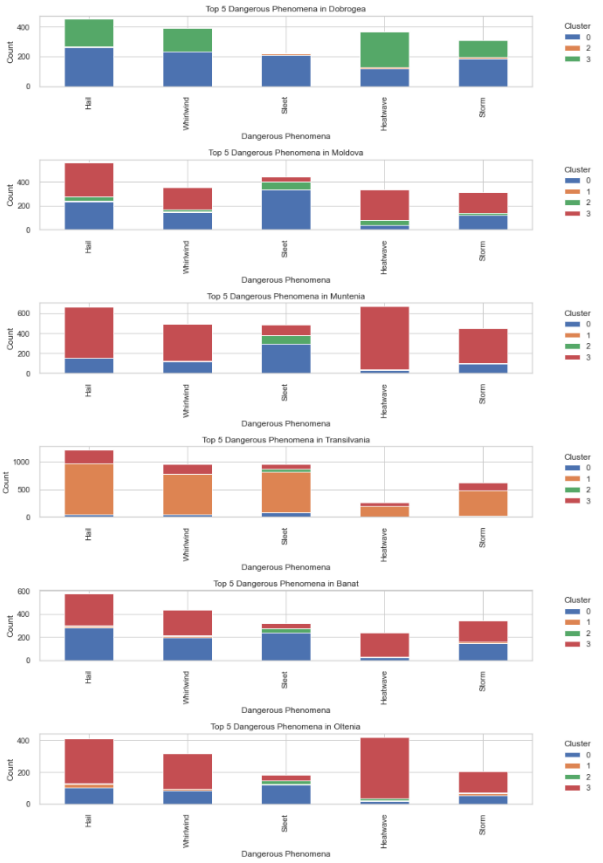


FIGURE 14. Extreme phenomena in each region.

is enough for the set objective. These two PCs were assigned names based on the most influencing variables. PC1 was named Regional Climate Influence, where region and humidity have positive influence and phenomena, pm10_quality, air_pressure and temperature have a negative influence, all of them having a similar impact on the PC. On the other hand, PC2 was labeled Environmental Factors component, where strong negative loadings for region and phenomena, and strong positive loadings for air_pressure and humidity can be found.

Therefore, in Figure 15, PCA is used for data visualization and two different categorical variables are used as the target: first the region, second the extreme phenomena.

In the upper plot, a linear relationship can be seen, clearly distinguished for each of the six regions along a diagonal line. Yellow is represented by Transilvania, contains the most observations and stands out at first sight. Two regions seem to overlap in the upper left corner. The regions, in their encoded order, are: Banat, Dobrogea, Moldova, Muntenia, Oltenia, Transilvania (colored from dark purple to yellow).

In the lower plot, due to the high number of types of extreme phenomena (13: Flood, Blizzard, Sleet, Freezing rain, Drizzle, Whirlwind, Cyclone, Tornado, Storm, Hail, Heatwave, Drought, Fires), a linear relation is harder to observe.

TABLE 2. PCA eigenvectors and eigenvalues.

	region	phenomena	pm10_quality	air_pressure	temperature	humidity	Explained Variance
PC1	0.4554	-0.4185	-0.1122	-0.4908	-0.4575	0.3930	33.99%
PC2	0.5533	-0.4269	0.0266	0.4825	-0.3338	0.4081	27.59%

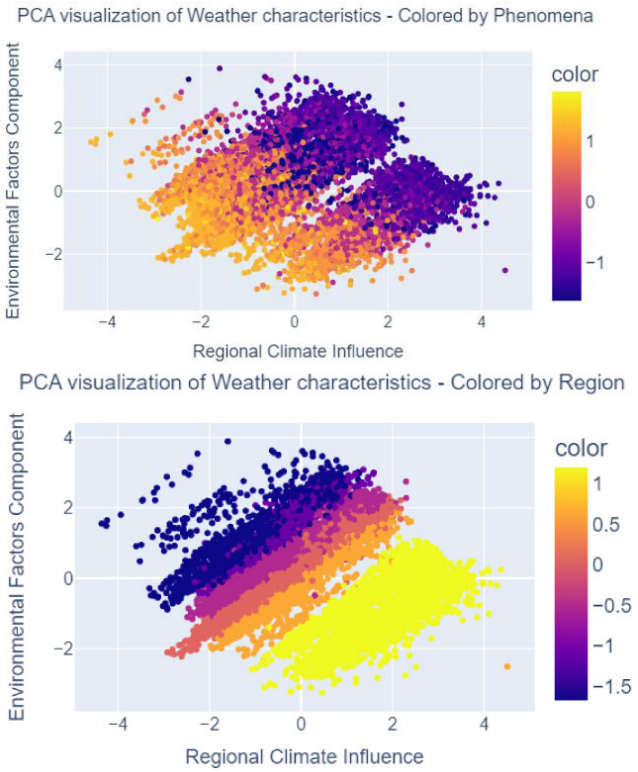


FIGURE 15. PCA visualization with different coloring: upper for Region, lower for Phenomena.

However, in the bottom left side there are predominant Heatwaves, Droughts and Fires, related negatively to both components, Environment and Regional Climate. On the opposite end, in the upper right, there are predominant Floods, Blizzards, Sleets and Freezing rains, related positively to the components.

2) K-MEANS USING PCA

Using the Elbow method for up to k=11 clusters, we finally set k=5 and employed the K-means algorithm using the data obtained from the two previously computed principal components. Figure 16 visually describes the five obtained clusters after applying the algorithm. Cluster 1 tends to have negative values for both components, Cluster 2 has only positive values for Regional Climate Influence component, Cluster 4 tends to have positive values for both components, and Cluster 5 has only negative values for Environmental Factors component. Cluster 3 has an irregular shape, but most of its values are contained in the upper left quadrant.

The five obtained clusters have 14,769 observations in total.

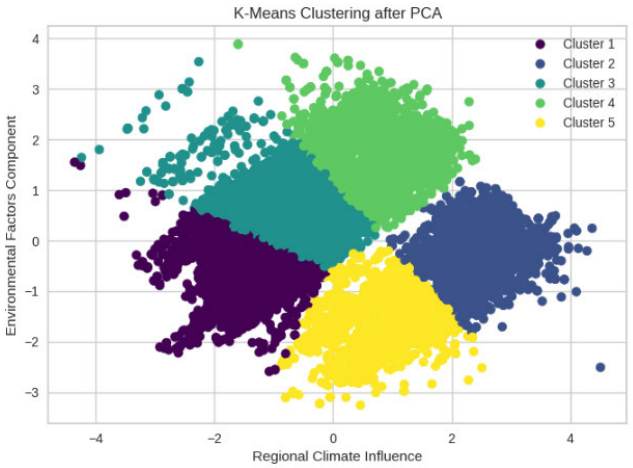


FIGURE 16. K-means clustering visualization after applying PCA algorithm with 2 PCs.

Figure 17 contains their distribution, and one can see that Cluster 3 has the most observations and Cluster 2 the least. Cluster 3 and 4 have a similar number of observations.

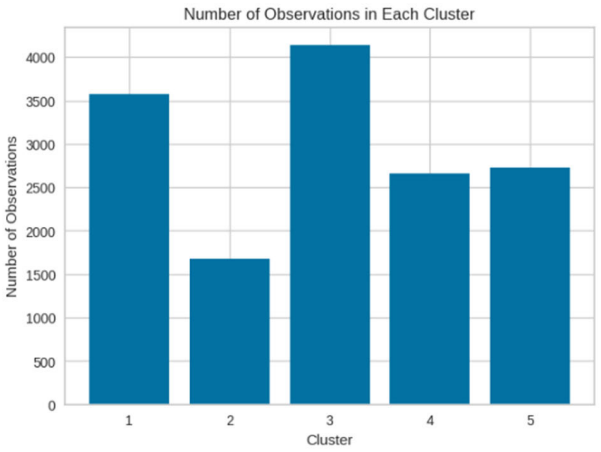


FIGURE 17. Distribution of the K-means clusters.

Descriptive statistics for the five obtained clusters are centralized in Table 3. As can be noticed, K5 has the lowest PM10 values, K1 has average values, while K2, K3 and K4 tend to have a high variability. Cluster K1 is characterized by high values of temperature, always positive, while K4 tends to have lower overall temperatures, but with the highest humidity. These values can be analyzed along with Figure 16 to understand the link between the statistics and the computed PCs.

TABLE 3. Descriptive statistics of the K-means clusters.

	PM 10			Air pressure			Temperature			Humidity		
Cluster	mean	max	min	Mean	max	min	mean	max	min	mean	max	min
K1	27.53	88.28	3.62	1002.14	1099.80	931.59	23.90	32.57	2.92	58.22	86.00	9.99
K2	23.87	133.71	1.44	950.27	993.00	814.15	5.64	23.47	-0.11	83.28	100.00	34.19
K3	22.89	140.86	3.81	1005.92	1099.80	931.59	19.51	31.83	-8.10	74.83	99.75	21.99
K4	24.28	143.55	2.99	1005.61	1099.80	961.84	7.31	26.12	-0.12	86.11	100.00	47.99
K5	20.27	67.35	2.53	956.90	995.75	881.49	19.34	31.12	-4.94	70.11	99.50	5.99

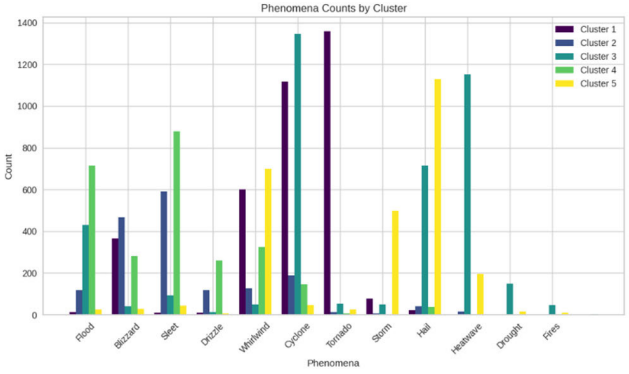


FIGURE 18. Extreme phenomena count by cluster.

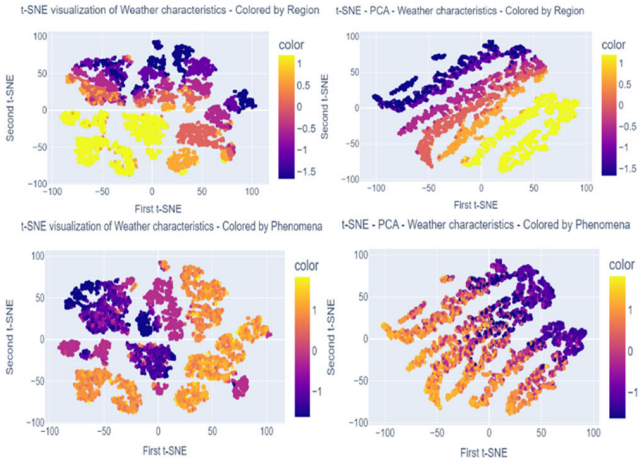


FIGURE 19. Using t-SNE on different sets: left on original data, right on PCA-processed data.

Lastly, in Figure 18 we present for each type of extreme phenomena the number of recorded observations for each cluster.

Following the results from Figure 16 and Table 3, it emphasizes the clusters different characteristics. In cluster K1 extreme phenomena such as whirlwind, cyclone, tornado and blizzard are representative, suggesting that K1 is distinguished by extreme winds. Cluster K2 has cold winters, judging by the higher numbers of blizzard and sleet phenomena. Cluster K5 is on the opposite end with extreme summers, suggested by numerous whirlwinds, storms, hail and heatwaves. K3 and K4 have mixed phenomena, with K3 being characterized by extreme instability: there are numerous floods, cyclones, hails, and even heatwaves and droughts.

3) T-SNE

In the end, we employed the t-SNE algorithm on the original dataset, used by the PCA algorithm, and on the PCA-processed dataset, to enhance the visualization goal and insights gained from it. t-SNE, the non-linear probabilistic algorithm, identifies different kinds of relationships in the data. With a perplexity of 50 and after 1000 iterations, this technique was used to obtain two specific components.

Figure 19 presents a comparison between using t-SNE on the two different datasets (original and reduced). On the left, clusters and sub-clusters can be clearly distinguished among the data, especially the left yellow ones (Transylvania) and dark blue ones (Banat). It is interesting to note how different the two outputs are, as the right plot maintains the linearity obtained and observed after using the PCA algorithm. Without using K-means or K-prototypes, one can clearly see the existence of some type of clusters in the right plot.

Moreover, when looking at the phenomena categorical variable and its encoding, the t-SNE algorithm behaved a little differently than before. On the left plot, extreme phenomena sub-clusters seem to be scattered all over the plot, with some phenomena even having sub-clusters in all the four quadrants. This plot by itself shows the presence of mini clusters, but one would be tempted to say that there are no actual clusters in the data, which is not exactly precise, as this article emphasized so far. The right plot, however, presents in an easier to understand way the presence of grouped data. In a diagonal way, from a dark-colored upper-right corner (Flood and Blizzard phenomena, for instance) to a light-colored bottom-left corner (Drought and Fires), t-SNE presents a correlation between extreme phenomena and weather characteristics.

V. CONCLUSION

In this study, we proposed a data pre-processing framework, and a comprehensive analysis was performed to investigate the trends and characteristics of heatwave occurrences over time and the relationship of these occurrences with various meteorological parameters and geographical factors. Using the Mann-Kendall Test, we identified a statistically significant increasing trend in the occurrences of heatwaves over

the years in the dataset. Moreover, the Chi-Square test for independence highlighted a significant relationship between the season and the number of heatwave occurrences, implying that heatwaves are not uniformly distributed across different seasons. This relationship was further reinforced by a Spearman Rank Correlation test that depicted a moderate positive correlation between rising air temperatures and an increase in the number of heatwave events. The results from the three statistical tests indicate that the trend towards extreme weather phenomena is persistent, not merely occasional. These consistent statistical findings underscore the ongoing and significant shift in weather patterns towards more extreme events.

Latent Dirichlet Allocation (LDA) was applied to a collection of around 7,000 distinct news articles, which identified three primary topics related to weather conditions. The coherence score validated that these topics were coherent and meaningful. Among these, one topic emphasized general weather conditions, another stressed on cold weather indicators, while the third was centered on extreme phenomena like hail.

Clustering techniques, specifically the K-prototypes model, were applied to understand the groupings related to the occurrence of phenomena in selected regions. Four distinct clusters emerged, characterized by various combinations of PM10 quality, air pressure, temperature, and humidity. These clusters provided insights into the predominant weather conditions for different regions in Romania. For instance, Muntenia and Oltenia predominantly experienced extreme weather conditions during the summer, while Moldova, Dobrogea, and Banat were more prone to harsh conditions in the winter. In terms of extreme weather phenomena, two clusters, C0 and C3, emerged as representative of winter and summer conditions, respectively.

Lastly, a combination of Principal Component Analysis (PCA), t-SNE and K-means clustering was employed to further understand the data. This technique further solidified previous findings, with distinct clusters emerging that showcased the influence of regional climates and various environmental factors on weather phenomena. This multi-pronged approach, utilizing statistical tests, topic modeling, and clustering, provides a comprehensive overview of the changing dynamics of heatwaves and extreme weather conditions over time, particularly in Romania.

The convergence of results from the three methods - Latent Dirichlet Allocation (LDA), K-Prototypes, and K-Means clustering - reinforces the findings. Despite their varied approaches, all methods consistently point to similar conclusions, affirming the robustness and reliability of the analysis in understanding the complex weather dynamics in Romania.

The analysis also unveils a complex relationship between air quality and weather events. Notably, during periods with elevated PM 10 levels, it has been observed an increased prevalence of weather phenomena like rain, fog, snowfalls, and frost. This finding suggests intricate interactions between atmospheric pollutants and specific meteorological conditions, warranting further investigation into their combined effects on environmental and public health.

As future work, we aim to introduce anomaly detection algorithms to better identify and predict heatwaves and extreme weather events. To support this, we will enhance our data gathering efforts, focusing on collecting a wider range of meteorological data. This enriched dataset will provide a solid foundation for the effective implementation of these advanced algorithms, enabling a more nuanced understanding of climate dynamics, particularly in Romania.

REFERENCES

[1] R. Bali Swain and F. Yang-Wallentin, "Achieving sustainable development goals: Predicaments and strategies," *Int. J. Sustain. Develop. World Ecology*, vol. 27, no. 2, pp. 96–106, Feb. 2020, doi: 10.1080/13504509.2019.1692316.

[2] M. Ziolo, I. Bak, and K. Cheba, "The role of sustainable finance in achieving sustainable development goals: Does it work?" *Technol. Econ. Develop. Economy*, vol. 27, no. 1, pp. 45–70, Dec. 2020, doi: 10.3846/tede.2020.13863.

[3] M. Bogers, F. Biermann, A. Kalfagianni, R. E. Kim, J. Treep, and M. G. de Vos, "The impact of the Sustainable Development Goals on a network of 276 international organizations," *Global Environ. Change*, vol. 76, Sep. 2022, Art. no. 102567, doi: 10.1016/j.gloenvcha.2022.102567.

[4] N. Subramaniam, S. Akbar, H. Situ, S. Ji, and N. Parikh, "Sustainable development goal reporting: Contrasting effects of institutional and organisational factors," *J. Cleaner Prod.*, vol. 411, Jul. 2023, Art. no. 137339, doi: 10.1016/j.jclepro.2023.137339.

[5] B. F. Giannetti, F. Agostinho, J. J. C. Eras, Z. Yang, and C. M. V. B. Almeida, "Cleaner production for achieving the sustainable development goals," *J. Cleaner Prod.*, vol. 271, Oct. 2020, Art. no. 122127, doi: 10.1016/j.jclepro.2020.122127.

[6] R. Kongboon, S. H. Gheewala, and S. Sampattagul, "Greenhouse gas emissions inventory data acquisition and analytics for low carbon cities," *J. Cleaner Prod.*, vol. 343, Apr. 2022, Art. no. 130711, doi: 10.1016/j.jclepro.2022.130711.

[7] J. P. Romero and C. Gramkow, "Economic complexity and greenhouse gas emissions," *World Develop.*, vol. 139, Mar. 2021, Art. no. 105317, doi: 10.1016/j.worlddev.2020.105317.

[8] H. El Bilali, I. H. N. Bassole, L. Dambo, and S. Berjan, "Climate change and food security," *J. Agricult. Forestry*, vol. 66, no. 3, pp. 197–210, Sep. 2020, doi: 10.17707/AgricultForest.66.3.16.

[9] N. Seddon, A. Smith, P. Smith, I. Key, A. Chausson, C. Girardin, J. House, S. Srivastava, and B. Turner, "Getting the message right on nature-based solutions to climate change," *Global Change Biol.*, vol. 27, no. 8, pp. 1518–1546, Apr. 2021, doi: 10.1111/gcb.15513.

[10] F. Charlson, S. Ali, J. Augustinavicius, T. Benmarhnia, S. Birch, S. Clayton, K. Fielding, L. Jones, D. Juma, L. Snider, V. Ugo, L. Zeitz, D. Jayawardana, A. La Nauze, and A. Massazza, "Global priorities for climate change and mental health research," *Environ. Int.*, vol. 158, Jan. 2022, Art. no. 106984, doi: 10.1016/j.envint.2021.106984.

[11] C. Latkin, L. Dayton, M. Scherkoske, K. Countess, and J. Thrul, "What predicts climate change activism?: An examination of how depressive symptoms, climate change distress, and social norms are associated with climate change activism," *J. Climate Change Health*, vol. 8, Oct. 2022, Art. no. 100146, doi: 10.1016/j.joclim.2022.100146.

[12] W. F. Lamb et al., "A review of trends and drivers of greenhouse gas emissions by sector from 1990 to 2018," *Environ. Res. Lett.*, vol. 16, no. 7, Jul. 2021, Art. no. 073005, doi: 10.1088/1748-9326/abee4e.

[13] Ö. Andersson and P. Börjesson, "The greenhouse gas emissions of an electrified vehicle combined with renewable fuels: Life cycle assessment and policy implications," *Appl. Energy*, vol. 289, May 2021, Art. no. 116621, doi: 10.1016/j.apenergy.2021.116621.

[14] J. Jesic, A. Okanovic, and A. A. Panic, "Net zero 2050 as an EU priroty: Modeling a system for efficient investments in eco innovation for climate change mitigation," *Energy, Sustainability Soc.*, vol. 11, no. 1, pp. 1–16, Dec. 2021, doi: 10.1186/s13705-021-00326-0.

[15] I. Kougias, N. Taylor, G. Kakoulaki, and A. Jäger-Waldau, "The role of photovoltaics for the European green deal and the recovery plan," *Renew. Sustain. Energy Rev.*, vol. 144, Jul. 2021, Art. no. 111017, doi: 10.1016/j.rser.2021.111017.

[16] M. Leonard, J. Pisani-Ferry, J. Shapiro, S. Tagliapietra, G. Wolf, P. Institute, and E. University, "The geopolitics of the European green deal," *Int. Organisations Res. J.*, vol. 16, no. 2, pp. 204–235, Aug. 2021, doi: 10.17323/1996-7845-2021-02-10.

[17] J. Gupta and C. Vegelin, "Sustainable development goals and inclusive development," *Int. Environ. Agreements: Politics, Law Econ.*, vol. 16, no. 3, pp. 433–448, Jun. 2016, doi: 10.1007/s10784-016-9323-z.

[18] L. Steg, "Psychology of climate change," *Annu. Rev. Psychol.*, vol. 74, pp. 391-421, Jan. 2023, doi: 10.1146/annurev-psych-032720-042905.

[19] T. Papadopoulos and M. E. Balta, "Climate change and big data analytics: Challenges and opportunities," *Int. J. Inf. Manage.*, vol. 63, Apr. 2022, Art. no. 102448, doi: 10.1016/j.ijinfomgt.2021.102448.

[20] DOMO. *Data Never Sleeps 10.0*. Accessed: Sep. 11, 2023. [Online]. Available: https://www.domo.com/data-never-sleeps

[21] P. A. Cortés and R. Quiroga, "How academic research and news media cover climate change: A case study from Chile," *Frontiers Commun.*, vol. 8, Aug. 2023, Art. no. 1226432, doi: 10.3389/fcomm.2023.1226432.

[22] M. Habib-ur-Rahman, A. Ahmad, A. Raza, M. U. Hasnain, H. F. Alharby, Y. M. Alzahrani, A. A. Bamagoos, K. R. Hakeem, S. Ahmad, W. Nasim, S. Ali, F. Mansour, and A. EL Sabagh, "Impact of climate change on agricultural production; issues, challenges, and opportunities in Asia," *Frontiers Plant Sci.*, vol. 13, Oct. 2022, Art. no. 925548, doi: 10.3389/fpls.2022.925548.

[23] E. F. Lambin, H. J. Geist, and E. Lepers, "Dynamics of land-use and land-cover change in tropical regions," *Annu. Rev. Environ. Resour.*, vol. 28, no. 1, pp. 205–241, Nov. 2003, doi: 10.1146/annurev.energy.28.050302.105459.

[24] M. Zhang, G. Li, T. He, G. Zhai, A. Guo, H. Chen, and C. Wu, "Reveal the severe spatial and temporal patterns of abandoned cropland in China over the past 30 years," *Sci. Total Environ.*, vol. 857, Jan. 2023, Art. no. 159591, doi: 10.1016/j.scitotenv.2022.159591.

[25] Y. Nan, M. Bao-hui, and L. Chun-kun, "Impact analysis of climate change on water resources," *Proc. Eng.*, vol. 24, pp. 643–648, Dec. 2011, doi: 10.1016/j.proeng.2011.11.2710.

[26] O. Hoegh-Guldberg, P. J. Mumby, A. J. Hooten, R. S. Steneck, P. Greenfield, E. Gomez, C. D. Harvell, P. F. Sale, A. J. Edwards, K. Caldeira, N. Knowlton, C. M. Eakin, R. Iglesias-Prieto, N. Muthiga, R. H. Bradbury, A. Dubi, and M. E. Hatziolos, "Coral reefs under rapid climate change and ocean acidification," *Science*, vol. 318, no. 5857, pp. 1737–1742, Dec. 2007, doi: 10.1126/science.1152509.

[27] V. Hase, D. Mahl, M. S. Schäfer, and T. R. Keller, "Climate change in news media across the globe: An automated analysis of issue attention and themes in climate change coverage in 10 countries (2006–2018)," *Global Environ. Change*, vol. 70, pp. 2006–2018, Sep. 2021, Art. no. 102353, doi: 10.1016/j.gloenvcha.2021.102353.

[28] M. T. Boykoff and J. M. Boykoff, "Climate change and journalistic norms: A case-study of U.S. mass-media coverage," *Geoforum*, vol. 38, no. 6, pp. 1190–1204, Nov. 2007, doi: 10.1016/j.geoforum.2007.01.008.

[29] R. N. Landers, R. C. Brusso, K. J. Cavanaugh, and A. B. Collmus, "A primer on theory-driven Web scraping: Automatic extraction of big data from the Internet for use in psychological research," *Psychol. Methods*, vol. 21, no. 4, pp. 475–492, Dec. 2016, doi: 10.1037/met0000081.

[30] Y. N. Kunang and S. D. Purnamasari, "Web scraping techniques to collect weather data in South Sumatera," in *Proc. Int. Conf. Elect. Eng. Comput. Sci. (ICECOS)*doi: 10.1109/ICECOS.2018.8605202.

[31] J.-C. Bricongne, B. Meunier, and S. Pouget, "Web-scraping housing prices in real-time: The COVID-19 crisis in the U.K.," *J. Housing Econ.*, vol. 59, Mar. 2023, Art. no. 101906, doi: 10.1016/j.jhe.2022.101906.

[32] G. Boeing and P. Waddell, "New insights into rental housing markets across the United States: Web scraping and analyzing craigslist rental listings," *J. Planning Educ. Res.*, vol. 37, no. 4, pp. 457–476, Dec. 2017, doi: 10.1177/0739456X16664789.

[33] B. Batrinca and P. C. Treleaven, "Social media analytics: A survey of techniques, tools and platforms," *AI Soc.*, vol. 30, no. 1, pp. 89–116, Feb. 2015, doi: 10.1007/s00146-014-0549-4.

[34] A. El Mhouti, M. Fahim, A. Soufi, and I. El Alama, "A Web scraping framework for descriptive analysis of meteorological big data for decision-making purposes," *Int. J. Hybrid Innov. Technol.*, vol. 3, no. 1, pp. 47–64, Oct. 2023, doi: 10.21742/ijhit.2653-309X.2022.2.1.04.

[35] G. Kalaivani and S. Kamalakkannan, "Web scraping technique for prediction of air quality through comparative analysis of machine learning and deep learning algorithm," in *Proc. Int. Conf. Augmented Intell. Sustain. Syst. (ICAISS)*, Nov. 2022, pp. 263–273.

[36] M. Moinuddin. *Using Natural Language Processing for Automating the Identification of Climate Action Interlinkages Within the Sustainable Development Goals*. Accessed: Nov. 2022. [Online]. Available: https://www.researchgate.net/publication/365729518

[37] L. Tanguy, N. Tulechki, A. Urieli, E. Hermann, and C. Raynal, "Natural language processing for aviation safety reports: From classification to interactive analysis," *Comput. Ind.*, vol. 78, pp. 80–95, May 2016, doi: 10.1016/j.compind.2015.09.005.

[38] J. Farrell, "The growth of climate change misinformation in U.S. philanthropy: Evidence from natural language processing," *Environ. Res. Lett.*, vol. 14, no. 3, Mar. 2019, Art. no. 034013, doi: 10.1088/1748-9326/aaf939.

[39] M. C. Buzea, S. Trausan-Matu, and T. Rebedea, "Automatic fake news detection for Romanian online news," *Information*, vol. 13, no. 3, p. 151, Mar. 2022, doi: 10.3390/info13030151.

[40] C. Busioc, V. Dumitru, S. Ruseti, S. Terian-Dan, M. Dascalu, and T. Rebedea, "What are the latest fake news in Romanian politics? An automated analysis based on BERT language models," in *Ludic, Co-design and Tools Supporting Smart Learning Ecosystems and Smart Education* (Smart Innovation, Systems and Technologies) Deutschland, Germany: Springer, 2022, pp. 201–212, doi: 10.1007/978-981-16-3930-2_16.

[41] A. Simion, M. Dascalu, and S. Trausan-Matu, "Analysis of trends in online Romanian news using semantic models," in *Proc. 22nd Int. Conf. Control Syst. Comput. Sci. (CSCS)*, May 2019, pp. 410–415, doi: 10.1109/CSCS.2019.00075.

[42] S. E. Uthirapathy and D. Sandanam, "Topic modelling and opinion analysis on climate change Twitter data using LDA and BERT model," *Proc. Comput. Sci.*, vol. 218, pp. 908–917, 2023, doi: 10.1016/j.procs.2023.01.071.

[43] E. Atagün, B. Hartoka, and A. Albayrak, "Topic modeling using LDA and BERT techniques: Teknofest example," in *Proc. 6th Int. Conf. Comput. Sci. Eng. (UBMK)*, Sep. 2021, pp. 660–664, doi: 10.1109/UBMK52708.2021.9558988.

[44] T. R. Keller, V. Hase, J. Thaker, D. Mahl, and M. S. Schäfer, "News media coverage of climate change in India 1997–2016: Using automated content analysis to assess themes and topics," *Environ. Commun.*, vol. 14, no. 2, pp. 219–235, Feb. 2020, doi: 10.1080/17524032.2019.1643383.

[45] L. L. Benites-Lazaro, L. Giatti, and A. Giarolla, "Topic modeling method for analyzing social actor discourses on climate change, energy and food security," *Energy Res. Social Sci.*, vol. 45, pp. 318–330, Nov. 2018, doi: 10.1016/j.erss.2018.07.031.

[46] D. Valle, P. Albuquerque, Q. Zhao, A. Barberan, and R. J. Fletcher, "Extending the latent Dirichlet allocation model to presence/absence data: A case study on North American breeding birds and biogeographical shifts expected from climate change," *Global Change Biol.*, vol. 24, no. 11, pp. 5560–5572, Nov. 2018, doi: 10.1111/gcb.14412.

[47] P. Brzustewicz and A. Singh, "Sustainable consumption in consumer behavior in the time of COVID-19: Topic modeling on Twitter data using LDA," *Energies*, vol. 14, no. 18, p. 5787, Sep. 2021, doi: 10.3390/en14185787.

[48] W. Ejaz, M. Ittefaq, and S. Jamil, "Politics triumphs: A topic modeling approach of analyzing news media coverage of climate change in Pakistan," *J. Sci. Commun.*, vol. 22, no. 1, Jan. 2023, Art. no. A02, doi: 10.22323/2.22010202.

[49] S. Hariri and A. Sill, "Simulation process support for climate data analysis," Sigarch, Assoc. Comput. Machinery, ACM Digit. Library, New York, NY, USA, Tech. Rep., 2012.p. 247, Art. no. 29.

[50] Q. Xie, J. Hao, J. Li, and X. Zheng, "Carbon price prediction considering climate change: A text-based framework," *Econ. Anal. Policy*, vol. 74, pp. 382–401, Jun. 2022, doi: 10.1016/j.eap.2022.02.010.

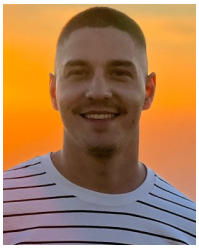
[51] Q.-T. Phan, Y.-K. Wu, and Q.-D. Phan, "An overview of data preprocessing for short-term wind power forecasting," in *Proc. 7th Int. Conf. Appl. Syst. Innov. (ICASI)*, Sep. 2021, pp. 121–125, doi: 10.1109/ICASI52993.2021.9568453.

[52] Q.-T. Phan, Y.-K. Wu, Q.-D. Phan, and H.-Y. Lo, "A novel forecasting model for solar power generation by a deep learning framework with data preprocessing and postprocessing," *IEEE Trans. Ind. Appl.*, vol. 59, no. 1, pp. 220–231, Jan. 2023, doi: 10.1109/TIA.2022.3212999.

[53] K. S. Lei and F. Wan, "Pre-processing for missing data: A hybrid approach to air pollution prediction in Macau," in *Proc. IEEE Int. Conf. Automat. and Logistic*, Aug. 2010, pp. 418–422.

[54] A. Aggarwal and D. Toshniwal, "A hybrid deep learning framework for urban air quality forecasting," *J. Cleaner Prod.*, vol. 329, Dec. 2021, Art. no. 129660, doi: 10.1016/j.jclepro.2021.129660.
[55] S. Bontemps, P. Defourny, C. Brockmann, M.Herold, V. Kalogirou, and O. Arino, "New global land cover mapping exercise in the framework of the esa climate change initiative," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2012, pp. 44–47.
[56] M. El Barachi, M. AlKhatib, S. Mathew, and F. Oroumchian, "A novel sentiment analysis framework for monitoring the evolving public opinion in real-time: Case study on climate change," *J. Cleaner Prod.*, vol. 312, Aug. 2021, Art. no. 127820, doi: 10.1016/j.jclepro.2021.127820.
[57] Y. S. Chang, K.-M. Lin, Y.-T. Tsai, Y.-R. Zeng, and C.-X. Hung, "Big data platform for air quality analysis and prediction," in *Proc. IEEE 27th Wireless Opt. Commun. Conf. (WOCC)*, Apr. 2018, pp. 1–3.
[58] Z. Wang, M. Liang, and D. Delahaye, "Automated data-driven prediction on aircraft estimated time of arrival," *J. Air Transp. Manage.*, vol. 88, Sep. 2020, Art. no. 101840, doi: 10.1016/j.jairtraman.2020.101840.
[59] S. Patra, S. Sahoo, P. Mishra, and S. C. Mahapatra, "Impacts of urbanization on land use /cover changes and its probable implications on local climate and groundwater level," *J. Urban Manage.*, vol. 7, no. 2, pp. 70–84, Sep. 2018, doi: 10.1016/j.jum.2018.04.006.
[60] P. Camus, F. J. Mendez, R. Medina, and A. S. Cofiño, "Analysis of clustering and selection algorithms for the study of multivariate wave climate," *Coastal Eng.*, vol. 58, no. 6, pp. 453–462, Jun. 2011, doi: 10.1016/j.coastaleng.2011.02.003.
[61] P. Govender and V. Sivakumar, "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)," *Atmos. Pollut. Res.*, vol. 11, no. 1, pp. 40–56, Jan. 01, 2020, doi: 10.1016/j.apr.2019.09.009.
[62] J. Corte-Real, B. Qian, and H. Xu, "Regional climate change in Portugal: Precipitation variability associated with large-scale atmospheric circulation," *Int. J. Climatol.*, vol. 18, no. 6, pp. 619–635, May 1998, doi: 10.1002/(SICI)1097-0088(199805)18:6<619:AID-JOC271>3.0.CO.2-T.
[63] J. Hong, Y. Xiang, Y. Liu, J. Liu, R. Li, F. Li, and J. Gou, "Development of EV charging templates: An improved K-prototypes method," *IET Gener., Transmiss. Distrib.*, vol. 12, no. 20, pp. 4361–4367, Nov. 2018, doi: 10.1049/iet-gtd.2017.1911.
[64] A. Wijayanto, Y. K. Suprapto, and D. P. Wulandari, "Clustering on multidimensional poverty data using PAM and K-prototypes algorithm: Case Study: Jambi province 2017," Surabaya, 2019, pp. 210–215, doi: 10.1109/ISITIA.2019.8937130.
[65] C. Li, X. Wu, X. Cheng, C. Fan, Z. Li, H. Fang, and C. Shi, "Identification and analysis of vulnerable populations for malaria based on K-prototypes clustering," *Environ. Res.*, vol. 176, Sep. 2019, Art. no. 108568, doi: 10.1016/j.envres.2019.108568.
[66] S. A.-S. Tahfim and C. Yan, "Analysis of severe injuries in crashes involving large trucks using K-prototypes clustering-based GBDT model," *Safety*, vol. 7, no. 2, p. 32, Apr. 2021, doi: 10.3390/safety7020032.
[67] J. Ji, T. Bai, C. Zhou, C. Ma, and Z. Wang, "An improved k-prototypes clustering algorithm for mixed numeric and categorical data," *Neurocomputing*, vol. 120, pp. 590–596, Nov. 2013, doi: 10.1016/j.neucom.2013.04.011.
[68] H. Perez and J. H. M. Tah, "Improving the accuracy of convolutional neural networks by identifying and removing outlier images in datasets using t-SNE," *Mathematics*, vol. 8, no. 5, p. 662, Apr. 2020, doi: 10.3390/MATH8050662.

ALIN-GABRIEL VĂDUVA was born in June 2000. He received the bachelor's degree in economic informatics. He is currently pursuing the M.Sc. degree in databases—support for business. In the professional realm, he is employed as a Data Scientist in the banking sector, where he has contributed to anti-money laundering projects. His research interests include mathematics, machine learning, data mining, deep learning, and generative AI.



MIHAI MUNTEANU was born in October 1999. He received the B.Sc. degree in statistics and economic forecasting. He is currently pursuing the M.Sc. degree in data science and applied statistics. Specializing in both theoretical and applied data analytics, his proficient in ML techniques, data visualization, and statistical analysis. Currently, he is a Data Scientist and the Team Leader of ML in the banking sector. He is leading projects in customer service automation and anti-money laundering.

SIMONA-VASILICA OPREA was born in July 1978. She received the M.Sc. degree from the Infrastructure Management Program, Yokohama National University, Japan, in 2007, the first Ph.D. degree in power system engineering from Bucharest Polytechnic University, in 2009, and the second Ph.D. degree in economic informatics from the Bucharest University of Economic Studies, in 2017. She currently teaches databases, database management systems, and software packages with the Faculty of Economic Cybernetics, Statistics and Informatics, Bucharest University of Economic Studies.

ADELA BÂRA was born in October 1978. She is currently the Director of the Data Science Excellence Center, Bucharest University of Economic Studies. She is the author of 19 books in the domain of economic informatics and over 60 published scientific papers and articles. She has participated in ten research projects, financed by national and international research programs. Her research interests include databases, data warehousing, big data, data mining, artificial neural networks, machine learning, the IoT, business intelligence, and informatics solutions for energy systems.

ANDREEA-MIHAELA NICULAE was born in January 1998. She received the bachelor's and master's degrees (summa cum laude). She is currently pursuing the Ph.D. degree with the Doctoral School of Economic Informatics, Bucharest University of Economic Studies. She is a Research Assistant with the Bucharest University of Economic Studies. She also teaches courses in databases, software packages, and managerial microeconomics. Her research interests include databases, big data, advanced analytics, and machine learning, particularly in the air quality domain.