Detection of features in the series of electric field strength measurements using data transformation and machine learning

A A Adzhieva1,*, V A Shapovalov1, I Kh Mashukov2, A A Tokbaeva3
1Kabardino-Balkarian State Agricultural University, 1 Tarchokova, Nalchik, Russia
2High-Mountain Geophysical Institute, 2 Lenin avenue, Nalchik, Russia
3Kabardino-Balkarian State University, 173 Chernyshevsky street, Nalchik, Russia
*E-mail: aida-adzhieva@mail.ru

Abstract. Data transformation is a common information processing procedure that can detect features hidden in the data that are not visible in their original form. The article shows some examples of such processing to detect features in the data of long-term observations of the intensity of the surface atmospheric electric field. This uses the data of the EFM 550 field sensor installed in the city of Nalchik. The results are visualized using the Matplotlib Python library to visualize the results of such data processing. The selected data transformation methods are also well suited for machine learning. To demonstrate the possibilities of finding features, several methods of teaching without a teacher are used. It is noted that most of the identified features in the series of measurements of the surface atmospheric electric field are associated with local meteorological phenomena.

1. Introduction
In the atmosphere, there are always mobile electric charges associated with positive and negative ions, as well as with elements of clouds and precipitation.

The volume charges in the atmosphere formed by these mobile charges and their transport play a major role in all phenomena of atmospheric electricity. They are formed as a result of unequal movement under the influence of an electric field of ions of different signs, differing in their characteristics. Volume charges can also occur during various electrification processes (dust, smoke, fragments of ice crystals, splitting of droplets into smaller ones, collision of ice droplets with ice crystals, friction of falling droplets on the air, etc.), when the atmosphere is filled with charged particles of mainly the same sign, which are then spread over considerable distances from the place of their formation [1].

The electric field in the atmosphere, like any electric field, is characterized at any point by the value of the potential and the intensity that are due to all the electric charges available both on the Earth's surface and in the atmosphere. Just like uncharged particles, the ions in the atmosphere are constantly moving. It is due to this that the atmosphere has electrical conductivity, in the lower layers it is small, in the higher layers it is significant. The characteristics of the electric field of the atmosphere also depend on the distribution of the conductivity of the atmosphere, although the electrical conductivity of the air changes slightly during the year [2].

Under undisturbed atmospheric conditions, i.e. in clear weather both over land and over the oceans, the atmosphere is positively charged with respect to the earth's surface, the field strength near the earth's surface is 130 V/m and is directed in the absence of clouds from top to bottom. With altitude, the field strength decreases exponentially: at an altitude of 10 km, it is only about 5 V/m. Above 20 km, the field strength is very low, and the air conductivity in these layers is sufficient to equalize the potential differences. The current strength at different altitudes is usually kept constant, since the electrical conductivity of the atmosphere increases with altitude, and the electric field strength decreases with altitude. Near the earth's surface, the intensity of the electric field of the atmosphere depends on the shape of the terrain – it weakens in the depressions of the terrain, on the streets – «urban canyons» and increases near the protruding elements of the landscape, high-rise masts and buildings [2-4].

Monitoring of the electric field in the surface layer shows extremely high variability of the field depending on various factors of air pollution, radioactivity, including meteorological ones-temperature, humidity, visibility range, wind speed, precipitation, cloud cover, etc. The intensity of the electric field of the atmosphere
experiences changes in the daily and annual course, they can be especially large when significant volume charges appear in the atmosphere, as, for example, occurs during the phenomena of fogs, heavy rains, snowstorms, haze, dry winds, dust storms, volcanic eruptions, etc. [5-9]. The biggest disturbances are associated with the development of clouds, especially thunderstorms. In thunderclouds, there is a strong electrification of cloud elements and a spatial separation of positive and negative charges into separate parts of the cloud. As a result, both inside the clouds and between the clouds and the earth, there are huge potential differences, at which the field strength reaches tens of thousands of volts per meter. The field strength between the cloud and the ground can even change its direction, that is, get an upward direction. When the potentials in the atmosphere reach the breakdown values, spark electric discharges and lightning occur, both in the clouds and between the clouds and the earth [10].

Thus, the electric field can act as an indicator of the quality of the atmosphere and be a predictor of the development of dangerous weather events. All this makes the study of the ground-level electric field extremely relevant and promising.

2. Materials and methods
To detect features in the data of long-term observations of the surface atmospheric electric field intensity conducted in the monitoring center at the Federal State Budgetary Institution «HMGI» [11-13], a program for processing and visualizing data on measuring atmospheric parameters «AEFM-DM-2017» is used. In this case, the data of the characteristics of the electric field strength in the surface layer of the field sensor EFM 550 of the company «Vaisala» installed in the city of Nalchik are used (figure 1). The main component of data preprocessing here is «data transformation».

![Figure 1. The field strength according to the EFM 550 measurement data.](image)

Data transformation refers to the transformation and matching of data sets with each other or with a specific schema. In machine computing, data transformation is the process of converting data from one format or structure to another format or structure. This is a fundamental aspect of most data integration and data management operations. Data transformation can be simple or complex, depending on the required data changes between the source (initial) data and the target (final) data. The tools and technologies used to transform the data can vary greatly depending on the format, structure, complexity, and volume of the data being converted. The transformation procedure consists in the fact that the computer performs, in principle, typical operations on data structures and values (sorting, sampling, arithmetic and logical actions, creating and changing data structures and elements, etc.) in the number and sequence specified by the algorithm for solving a computational problem, which is implemented at the physical level by a sequential set of machine commands (program). As a rule, the data must either be converted to a form corresponding to each other, or to a specific target format. If the data is standardized, i.e. brought to a single form, then their format is defined as the main one. Then all incoming data from other sources will be brought exactly and only to this target view.

From a mathematical point of view, data transformation is the application of a deterministic mathematical function to each point in the data set, that is, each data point \( z_i \) is replaced by a transformed value \( y_i = f(z_i) \), where \( f \) is a function. The function that is used to transform the data is almost always reversible, and is usually
continuous. One-dimensional functions can be applied pointwise to multidimensional data to change their particular distributions. It is also possible to change some properties of multidimensional distributions using appropriately constructed transformations. Such transformations are usually used to make the data more suitable for statistical processing.

Thus, data transformation is important not only for analyzing the data, but also for preserving the relationships between them. The data can also be transformed to facilitate visualization, for example, for a graphical representation. Another significant reason for data conversion may be to improve interpretability, even if no formal statistical analysis or visualization is intended.

In fact, data transformation acts as the main path to the feature description of objects, which is necessary for machine learning, and each object is described by a set of its numeric or non-numeric characteristics, which are called features.

Machine learning algorithms can be divided into several categories: Supervised Learning, Reinforcement Learning, and Unsupervised Learning. The teacher-led learning algorithm takes the labeled data and creates a model that performs predictions by providing new data. It solves both classification and regression problems. Reinforcement learning uses a reward system and a trial-and-error approach to maximize the rewards received over the long term. Unsupervised and unclassified data is used in unsupervised learning, where you need to find patterns and get a data structure to find the desired value. The main forms of unsupervised learning are clustering and dimensionality reduction.

Unsupervised learning (unsupervised learning, self-learning, spontaneous learning) is a class of machine learning methods in which the test system is spontaneously trained to perform a task without interference from the experimenter. In unsupervised learning, the task of machine learning is to derive a function to describe a hidden structure from «unlabeled» data, while the system tries to independently find patterns directly from the example given. From the point of view of cybernetics, this is a type of cybernetic experiment. As a rule, this is only suitable for tasks in which the descriptions of a set of objects (the training sample) are known, and it is necessary to detect internal relationships, dependencies, and patterns that exist between objects.

Learning without a teacher is the cutting edge of data mining technologies and a serious step towards creating a strong artificial intelligence. For example, due to the fact that the vast majority of the accumulated data of mankind is not marked up, it is impossible to apply traditional teaching with a teacher to them. At the same time, learning without a teacher allows you to successfully work with unmarked data sets and identify the patterns inherent in them, which a person cannot detect [14-20].

Data clustering methods, where an algorithm picks up similar data, finds common features, and groups them together into subsets called clusters, are the most popular family of unsupervised machine learning methods. Examples include K-Means algorithms, Dendrogram-based hierarchical clustering, and Density-based spatial clustering for applications with noise (Dendrogram).

Dimensionality reduction methods are used to reduce the number of features so as to completely eliminate correlating features and as a way of grouping similar observations in the data, each object of the high-dimensional feature space is modeled into a two - or three-coordinate point in such a way that similar data elements in the multidimensional space are projected into neighboring points, and dissimilar objects are more likely to be modeled by points far apart from each other. The goal is to represent the data in a smaller-dimensional space, minimizing information loss as much as possible. Here, as an example, we can distinguish the Principal Component Analysis (PCA) method and the t-SNE (t-Distributed Stochastic Neighbor Embedding) method.

3. Results and discussion

The general scheme of using unsupervised learning algorithms includes: data exploration, detection of outliers (anomalies), i.e. those data areas where the object's behavior significantly differs from the characteristic (expected) behavior, identification (recognition) of images (obtaining associations).

For the study, unsupervised learning was implemented based on two Python platforms: Scikit-learn and TensorFlow/Keras. The features were obtained by converting the data of the time series of measurements of the electric field strength of the atmosphere in the python – tsfresh package [21]. In order to demonstrate the ability to search for features by consistently applying one of the dimension reduction methods (PCA) and clustering methods (DBSCAN) for unsupervised learning, the results on feature detection were obtained (figure 2). In unsupervised learning, the input data is split based on the characteristics presented, and the prediction of the properties is based on which cluster the example belongs to.
Figure 2. Example of data processing with transformation and using unsupervised learning methods.

It should be noted that in a particular case, the features identified in the series of measurements of the surface atmospheric electric field correlate well with local meteorological phenomena. All this indicates the complexity of the actual picture of the processes occurring in the clouds and the implementation of weather events associated with them.
4. Conclusion
The search for analytical dependencies on the data of long-term measurements is greatly simplified. Using ready-made tools of packages for working with data, specialists will be able to identify hidden patterns in information arrays, analyze data more deeply [22], detect anomalies, perform automatic feature construction and generate synthetic data sets.

References
[1] Petrova G G et al 2018 Investigation of the role of aerosols in the formation of electric field variations in the surface atmosphere (Bulletin krasec. physical and mathematical sciences) 5 42–54
[2] Anisimov S V et al 2018 Global and regional components of the electricity of the undisturbed lower atmosphere of the middle latitudes Physics of the Earth 5 104–114
[3] Adzhiev A Kh et al 2017 Influence of orography on the dynamics of electrical phenomena in the atmosphere Global electric circuit 80–81
[4] Adzhiev A Kh and Kupovykh G V 2015 Measurements of the electric field of the atmosphere in high-altitude conditions of the Elbrus region Proc. of the Russian Academy of Sciences. Atmospheric and ocean physics 51(6) 710
[5] Boldyrev A S, Boldyrev A K, Kupovykh V G, Pestov D A, Pestova O V and Redin A A 2013 On the issue of monitoring the electric field of the atmosphere according to ground-based observations Modern Problems of science and education 6 875–84
[6] Krasheninnikov A V, Loktev D N and Solovyov S P 2019 Features of the atmospheric electric field in the conditions of aerosol pollution of the atmosphere by technogenic sources Trigger effects in geosystems 399–404
[7] Voschan O N and Spivak A A 2017 Variations of the electric field in the surface atmosphere as a result of the passage of cold atmospheric fronts Dynamic processes in geospheres 9 80–88
[8] Spivak A A, Rybnov Yu S and Kharlamov V A 2017 Variations of geophysical fields during hurricanes and squalls in Moscow Dynamic processes in geospheres 8 88–94
[9] Chernieva N V and Firtov P P 2018 Formation of the local electric field of the atmosphere in Kamchatka under the influence of natural processes (Vladivostok: Dalnauka) p 126
[10] Kuzmin V A 2018 The relationship of thunderstorms in clouds and the electric field strength of the atmosphere at the Earth's surface Systems for ensuring technosphere safety 141–3
[11] Adzhieva A A, Mashukov I Kh and Shapovalov V A 2017 Local monitoring of the surface electric field intensity of the atmosphere in Nalchik. Slope processes. Environmental problems. Environmental pollution (Nalchik: Aeterna) p 90
[12] Adzhiev A Kh et al. 2017 Atmospheric Electric Field Strength Meter Experimental devices and techniques 5 125–129
[13] Mashukov I Kh, Adzhieva A A and Shapovalov V A 2017 Investigation of the characteristics of the electric field strength in the surface layer by measurements of the EFM550 sensor network The third All-Russian conference «Global Electric circuit» (Borok) p 51
[14] Alemasov E P and Zariyeva R S 2020 Prospects of application of machine learning technologies Information technologies in construction, social and economic systems 2 32–34
[15] Kharitonova A E 2020 Statistics and machine learning Reports of the TSHA p 312
[16] Galimov R G 2017 Fundamentals of machine learning algorithms-learning without a teacher Alley of Science 14(1) 807–9
[17] Rashka S 2021 Python Machine Learning (M.: Dialektika) p 656
[18] Kingma D and Adam Ba J 2014 A method for stochastic optimization Computer Science. Mathematics arXiv:1412.6980v9 [cs.LG]
[19] Hastie T, Tibshirani R and Friedman J 2009 Unsupervised learning The elements of statistical learning (New York: Springer) p 485
[20] Sinaga K P and Yang M S 2020 Unsupervised K-means clustering algorithm IEEE Access (2020)2988796
[21] Christ M, Braun N, Neuffer J and Kempa-Liehr A W 2018 Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh - A Python package) Neurocomputing 307 72–77
[22] Kanarachos S et a. 2017 Detecting anomalies in time series data via a deep learning algorithm combining wavelets, neural networks and Hilbert transform Expert Systems with Applications 85 292–304