Machine Learning Models Applied to Weather Series Analysis

Chapter · March 2021

DOI: 10.1007/978-3-030-71903-6_21

Francesca Fallucchi
Università Telematica Guglielmo Marconi

Ernesto William De Luca
Otto-von-Guericke-Universität Magdeburg

Machine learning models applied to weather series analysis

Francesca Fallucchi[1][0000-0002-3288-044X] and Riccardo Scano[2][0000-0002-1937-0736] and Ernesto William De Luca[1][0000-0003-3621-4118]

[1] Guglielmo Marconi University, 00193 Rome, Italy, f.fallucchi@unimarconi.it
[2] Council for Agricultural Research and Economics, 00198 Rome, Italy

Abstract. In recent years the explosion in high-performance computing systems and high-capacity storage has led to an exponential increase in the amount of information, generating the phenomenon of big data and the development of automatic processing models like machine learning analysis. In this paper a machine learning time series analysis was experimentally developed in relation to the paroxysmal meteorological event "cloudburst" characterized by a very intense storm, concentrated in a few hours and highly localized. These extreme phenomena such as hail, overflows and sudden floods are found in both urban and rural areas. The predictability over time of these phenomena is very short and depends on the event considered, therefore it is useful to add data driven methods to the deterministic modeling tools to get the anticipated predictability of the event, also known as nowcasting. The detailed knowledge of these phenomena, together with the development of simulation models for the propagation of cloudbursts, can be a useful tool for monitoring and mitigating risk in civil protection contingency plans.

Keywords: machine learning, nowcasting, cloudburst.

1 Introduction

In this paper regression models were applied to a meteorological time series with the aim of identifying a physical correlation that could possibly be integrated into a nowcasting system [27]. The volume and "speed" of the data used in this study are characteristic of the big data domain [4], whose analysis requires massive processing models provided by machine learning.

In particular the rain data collected at the Manziana weather station north of Rome was analyzed with the aim of identifying a possible correlation between the event of extreme rain and the monitored meteorological quantities (temperature, humidity, pressure) [19]. These meteorological measures have characterized the minutes just before the phenomenon of the rainstorm, assuming that they were the "trigger" of the phenomenon itself. A rainstorm is defined as serious if the intensity is 80-100 millimeters per hour, with durations of about ten minutes. Unfortunately it is very difficult to establish the point where rainstorms occur because they are very concentrated and sometimes there are no detection tools in the points where the intensities are maximum [2]. The rains of the storms instead have longer durations and are more extensive, therefore better measurable.

Current monitoring systems are doppler radars, multispectral satellites and ground stations that measure physical quantities directly. As a result, there are large amounts of meteorological data observed available for forecasting models, including for example the Nation Oceanic and Atmospheric Administration (NOAA) which is reaching 100 TB per day. The main current forecasting approach for these phenomena is based on applications that make use of the information obtained from the weather-radar network which detects the movement of storms in real time [19][20].

Specifically, the weather radar is an instrument designed for the detection of atmospheric precipitation carried out by means of a rotating antenna that sends a pulse signal in the microwave band. The presence of raindrops along the signal path generates a change in reflectivity which is detected by the antenna itself and from which an estimate of the intensity of precipitation can be obtained.

The data thus recorded (on average every 10 minutes) are used to create georeferenced maps of reflectivity and therefore of intensity of precipitation with a resolution of about 1 kilometer and with greater reliability than other detection systems (satellites and ground stations).

Nowday available tools, capable of explicitly simulating the non-hydrostatic dynamics of convective phenomena, are aimed at large-scale and medium-long time atmospheric forecasting (cf. European Center for Medium-Range Weather Forecasts - ECMWF) [19]. The classic numerical models produce long-term forecasts of 1 to 10 days and for large areas of about 5 kilometers.

On the other hand, nowcasting is oriented towards high resolution (1km x 1km) and short-term (max 1 hour) forecasts, therefore usable for immediate emergency decisions in response to extreme phenomena. Due to climate changes and orographic reasons [3], the consequent flash floods can take on a particularly violent character being triggered by rainfall which in a few hours reaches cumulative values above 500 millimeters, thus increasing the level of risk.

In order to improve the tools for forecasting extreme weather events, such as heavy rainfall, ensemble forecasting systems have been developing for some years to be used in parallel with classic systems with limited area [5][9]. These high resolution models are able to provide a probabilistic forecast of the state of the atmosphere on a small scale by simulating the convective phenomena with a horizontal resolution of about 1.5 to 2.2 kilometers. The forecasting ability of these phenomena is very short and depends on the event considered, it is therefore necessary to combine data driven modeling tools with physical methods that allow to simulate the event with greater precision [2][5]. In general the use of different data sources could allow a level of reliability of the warning system, for example the social media analysis with NLP methods and information extraction could improve strategies of disaster management, but in this case the time scale would be of the order of days [10].

2 Related Works

The evolution of high-performance computing systems has enabled the development of machine learning based on big data [Changhyun C. et al., 2018] especially with meteorological data as explanatory variables [Dueben P. et al., 2018][Abrahamsen E. B. et al., 2018]. By learning the models' strengths and weaknesses [17], the climate community is starting to adopt AI algorithms as a way to help improve forecasts, but some researchers don't rely on these 'black boxes' deep-learning systems to forecast imminent weather emergencies such as floods. Nevertheless some AI algorithms are proving useful for weather forecasting, indeed in 2016 researchers reported the first use of a deep-learning system to identify weather fronts showing that the algorithm could replicate human expertise [Liu, Y. et al., 2016]. After 2016 test, computer scientists have incorporated an AI algorithm into the weather service's hail forecasts. [McGovern, A. et al. Bull, 2017]. Machine learning techniques have demonstrated promising results for forecasting chaotic systems purely from past time series measurements of system state variables (training data) without prior knowledge of the system dynamics, filling the gaps in underlying mechanistic knowledge [Pathak J. et al., 2018]. The prediction of extreme weather events depends on the formulation and analysis of complex dynamical systems characterized by high intrinsic dimensionality of the underlying attractor to which are not applicable the classical order-reduction methods through projection of the governing equations. Alternatively, data-driven techniques aim to quantify the dynamics of specific critical modes. [Wan Z. et al., 2018][Sebastian Scher et al., 2016]. The occurrence of natural disasters such as floods and cloudbursts is increasing due to the climate change, moreover the damage is becoming larger and larger due to the rapid urbanization over the world [Gozzini B, 2017]. By using big data and the machine learning model to predict the occurrence of heavy rain damage, it's possible greatly reduce the damage through proactive disaster management [Changhyun C. et al., 2018]. High-resolution nowcasting for extreme weather is an essential tool needed for effective adaptation to climate change and in particular deep learning techniques have shown dramatic promise in high-resolution (1km x 1km) short-term (1 hour) predictions of precipitation [Agrawal S., Hicdkey J., Xingjian S. et al., 2019]. Lastly, to be able to model a trigger of geospatially localized natural adverse events, is also useful to implement an ontology interoperable disaster risk reduction system (DRR) and to organize an Emergency Response System (ERS) [1][8][16][21].

3 Theoretical Background

The machine learning analysis [18] developed in this work makes extensive use of fundamental mathematical models [22][24][26] and the statistical theories supporting data analysis [6][7][13].

In general a machine learning method assigns an x point (feature) of a R_k space to an y point (pattern) of another R_h space. The features are usually numerical vectors while the patterns are labels, sortable or non-sortable, appropriately coded with real numbers. The supervised approach requires that the algorithm receives sample reports (training set) before the test.

Classification is a supervised method for assigning data (x_n, y_n) to predefined classes through the likelihood function with which it is possible to classify the n data by separating them linearly through a hyperplane.

In general in a set of points of the n-dimensional space, the classifier is a subspace of dimension n-1, obtained by applying a projection function P from the space with k dimensions to one k + h dimensions, where the additional dimensions h are the weights that reassign the labels of the training set ($P(x_k)$, y_k) in order to make the points of the plane separable.

A binary classifier is the Naive Bayes classifier able to decide whether the binary hypothesis y = (1,0) is much probable for a vector of features (x) observed by applying the Bayes theorem:

$$p(y|x) = p(x|y)\, p(y) / p(x) \tag{1}$$

where p(y) is the prior probability of hypothesis y, while the likelihood p(x|y) is estimated from a training set.

The data driven algorithms and techniques used in this analysis have the aim of modeling time series problem starting from the data sampling (x,y) of the physical process itself. The inductive model used here is part of the general techniques commonly designated with the term of soft computing that find application in the treatment and processing of uncertain or incomplete information [12][13][25].

In particular the supervised machine learning process was divided into the following analysis phases:

- pre-processing of the categorized (labeled) data has been carried out through normalization, smoothing (rounding and cleaning), reduction of dimensionality (selection of specific main attributes) and finally the choice of the algorithm parameters such as threshold values and cross-validation in which the parts of the training set are recombined for the next training;
- division of the data into two sets with which is performed the training and test algorithm model, verifying its predictive capacity by comparing the output with the real data (see Figure 1).

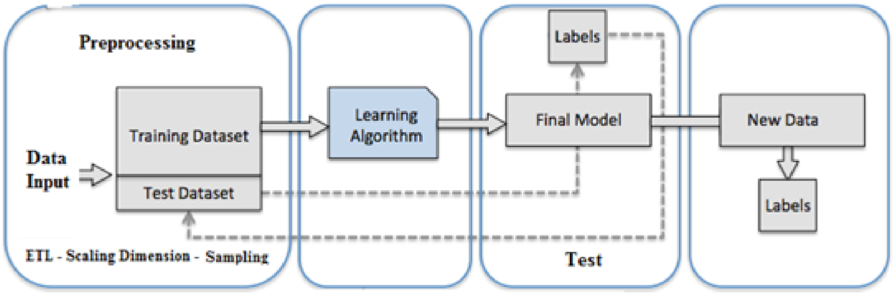


Fig. 1. Supervised machine learning algorithm.

The model's performance was determined with the use of the RSME (root mean square error), the accuracy measurement ((VP + VN) / (P + N)) and with the confusion matrix in which the four classification results are identified in true positives (VP), false positives (FP), true negatives (VN) and false negatives (FN).

On the other hand the adoption of the artificial neural networks (RNA) mathematical model to find relationships between the data, would have been computationally expensive in terms of calculation time and sample size in particular for the training phase [14].

4 Methodology

The use of machine learning models for "physics-free" data driven meteorological forecasts in small-scale areas, have the advantage of being computationally economic and allow short-term forecasts for already trained inference models. Deep learning procedure was not adopted due to the small size of the selected dataset that would produce overfitting, which is a typical problem of low bias and high variance models implemented by tensorflow.

The analysis carried out here is theoretically based on the inferential model of the precipitation ($R_t^{lat,lon}$) conditional probability (P) at time t compared to each physical variable ($V_{t-1}^{lat,lon}$) (same latitude and longitude) measured at soil and with a retrograde temporal lag (t-1), such as to be considered as a physical trigger factor of the rainstorm (r):

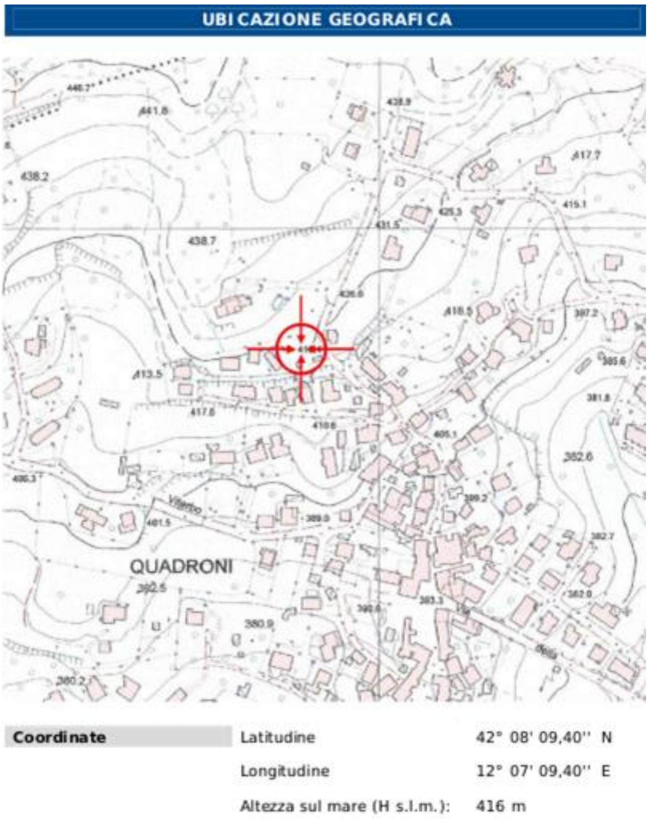
$$P(R_t^{lat,lon} > r \mid V_{t-1}^{lat,lon}) \tag{2}$$

The considerable amount of data would make the Bayesian approach too complex [14][15] compared to the regression and classification models of the time series used here.

The pluviometric datasets refer to a tipping rain gauge with electronic correction of the values and to a thermo-hygrometer for temperature and humidity measurements, provided by the civil protection of the Lazio region, documented and of known information quality comparable over time and space.

The temporal coverage of the dataset used in this work refers to the range of years from 2015 to 2019 with an update frequency of 15 minutes. The 15th minute interval is bounded by the maximum radio transmission frequency of the weather station data.

The spatial coverage is punctual and refers to the “Manziana” weather station (see Figure 3) which provided representative data for the area being analyzed. The choice of a specific weather station was mainly due on the theoretical prerequisite of the physical uniformity (omogeneus boundary conditions) and on the quality of the measurement equipment (99% data availability). Over the period considered, the average annual temperatures are of the order of 15°C while the average annual precipitation is about 998 mm.



UBICAZIONE GEOGRAFICA

Coordinate	Latitudine	42° 08' 09,40'' N
	Longitudine	12° 07' 09,40'' E
	Altezza sul mare (H s.l.m.):	416 m

Fig. 3. Spatial localization of “Manziana” weather station.

The rainiest period is concentrated between August and November when the Italy Tyrrhenian coasts are affected by intense storms. The rainfall dataset allows us to characterize the extreme events recognized as storms with the parameter of the cumulative rainfall over 15-minute intervals. The indicators chosen for the analysis (station location, precipitation, temperature, humidity, pressure and time interval) provide the necessary information on the meteorological phenomenon analyzed [23].

The time interval (2015 - 2019) was chosen to have the maximum continuity of the gapless series and considering the presence of various cloudburst phenomena due to global warming [3][23], indeed the time span considered was marked by several paroxysmal meteorological events, characterized by high quantities of rain often concentrated over a day. The indicators were chosen to obtain a possible data-driven correlation between the parameters characterizing the physical phenomenon.

Conversely the possible limitations of the indicators are due to the limited number of years of the historical series and to the absence of a physical analysis of the same meteorological phenomenon [3][5].

Below (see Figure 4) there is the class diagram that represents the general preprocessing algorithm developed.

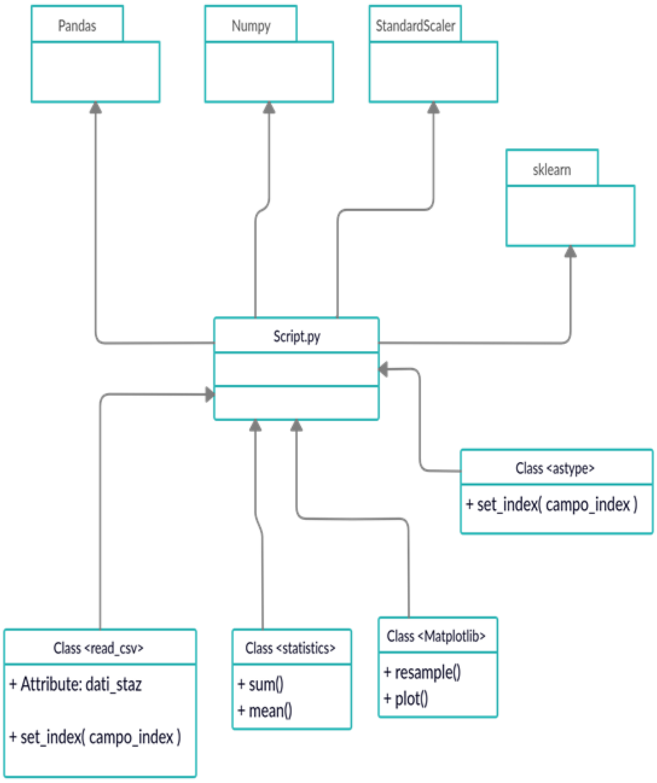


Fig. 4. Preprocessing algorithm class diagram.

The above detailed algorithm consisting of:

- dataset reading and indexing;
- cleaning and normalization with the deletion of missing data (Null), incorrect or meaningless (Nan) for the forecast;
- conversion of variables in numeric (float) and date format;
- computation of average values for all variables over the entire time span;
- daily resample and scaling for plotting and visual verification of the variables trend;
- calculation of the scatter matrix (see Figure 5) which provides the graphic distribution and the variables relationship;

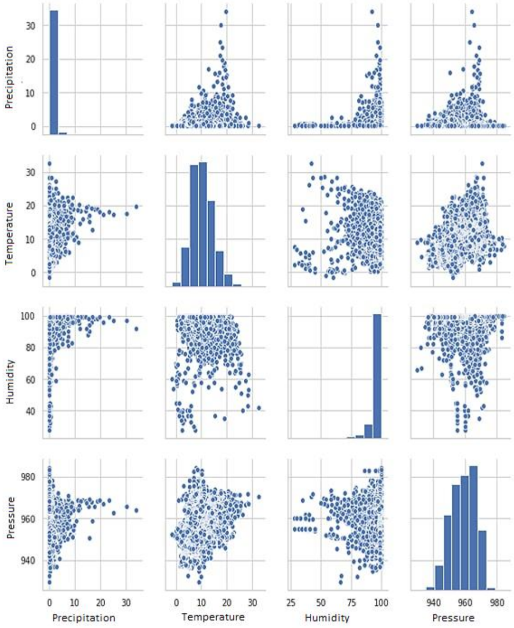


Fig. 5. Scatter matrix.

- computation of the heatmap-type correlation matrix between the dataset variables (see Figure 6), from which it is already clear the greatest correlation between precipitation and temperature.

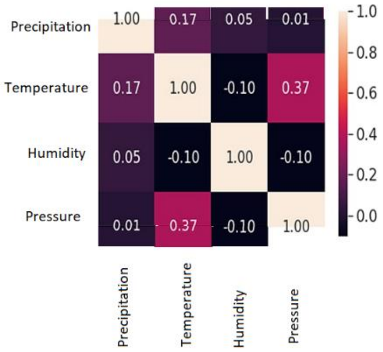


Fig. 6. Correlation matrix.

The following flow diagram (see Figure 7) represents the general algorithm of the entire processing scheme:

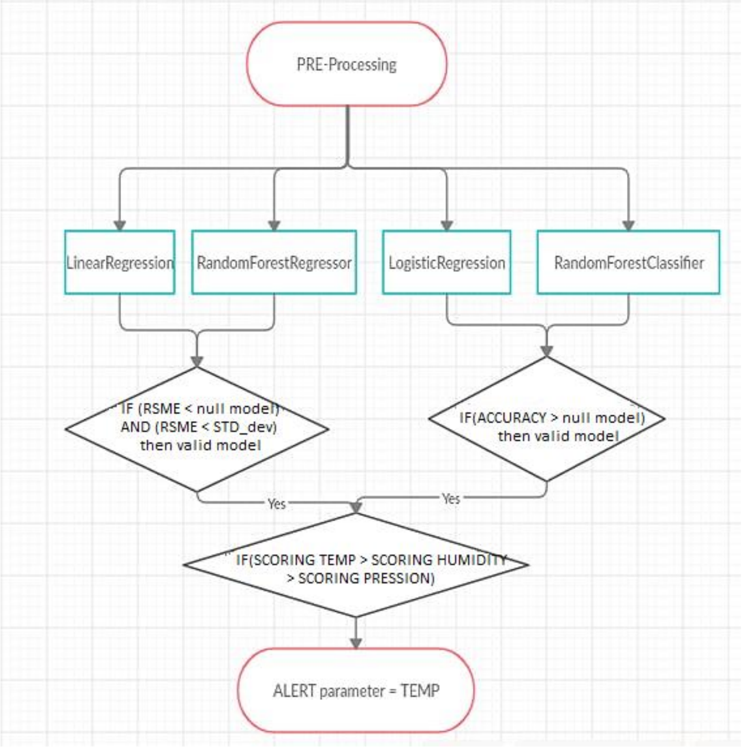

PRE-Processing
LinearRegression
RandomForestRegressor
LogisticRegression
RandomForestClassifier
IF (RSME < null model) AND (RSME < STD_dev) then valid model
IF(ACCURACY > null model) then valid model
Yes
Yes
IF(SCORING TEMP > SCORING HUMIDITY > SCORING PRESSION)
ALERT parameter = TEMP

Fig. 7. Processing algorithm flow diagram.

- backward translation of the series by a time step (15’) for the variables (temperature, pressure, humidity) whose correlation with the precipitation must be verified;
- regression path definition with the application of the linear and random forest cross-validation models (with 2 training sets and 1 test set (cv = 3)) and with the RSME (root mean squared error) as performance metric;
- RMSE computation for the null model: (average number of precipitation minus precipitation values), which must be major of the error obtained with the linear and random forest models;
- definition of the classification path by transforming the precipitation into a categorical variable {1,0} by calculating the percentage variation over a time step (precipitation percentage change greater than average percentage change produces true (1) otherwise false (0));
- Application of the cross-validation logistic and random forest classification model (with 2 training sets and 1 test set (cv = 3)) and with the accuracy as performance metric;
- calculation of the null model accuracy: average number of precipitation values classified as significant (greater than average percentage change) compared to the total (100%), which must be less than accuracy obtained with the logistic and random forest models.

- extraction from the historical series of the record corresponding to the maximum precipitation value, from which is obtained the temperature drop ($\Delta t = t_i - t_{i-15'}$) in the case of extreme precipitation event (mm max).

5 Results and Discussion

Three summary tables are reported (see table 1, 2, 3) which represent the results obtained for each model (classification, regression) applied to the current case and the discriminating value for each corresponding metric (major accuracy for classification and minor RSME for regression).

Table 1. Correlation with precipitation.

Temperature	0.17
Pression	0.01
Humidity	0.05

Table 2. Correlation with precipitation.

	Analysis type (response variable = Precipitation) (characteristic variables = Temperature, Humidity, Pression)						
	CLASSIFICATION			REGRESSION			
	ML Model			ML Model			
	Random forest	Logistic	Null	Random Forest	Linear	Null	Standard deviation
Accuracy	0.755	0.76501188	0.76501189				
RSME Temperature				1.462	**1.395**	1.4114	1.4115
RSME Humidity				1.407			
RSME Pression				1.414			

Table 3. Temperature threshold

(supposed as trigger meteorological value for the cloudburst).

	Precipitation (mm max cumulated in 15’)	Temperature
t_{i-1}	-	23.8
ti (= $t_{i-1\ +15'}$)	34.0	19.6
Threshold (Δt)		-4.2

Therefore cross-validation model (logistic, random forest) of the classification path obtains accuracy values lower than null model. On the other hand, the linear model of the cross-validation regression path obtains RMSE error values lower than random forest model, null model and the simple standard deviation. Then from the comparison of the results obtained for the cross-validation linear regression model applied to each atmospheric parameter (temperature, humidity, pressure) temporally anticipated (15’), it is clear that the temperature is the variable that predicts the trend of the precipitation better than the others (lower RSME) and therefore also possible extreme phenomena such as cloudbursts. Moreover the temperature jump that anticipates the cloudburst event is obviously linked to the geomorphological characteristics of the analyzed area, but the correlation between the two physical variables, temperature jump and severe precipitation, can have a general validity.

6 Conclusions

In this paper meteorological time series analysis has been described, focusing in particular on the paroxysmal cloudburst weather event. Then data driven methods were implemented based on machine learning models to analyze the temporal propagation of the cloudburst. It has been experimentally verified that the data-driven approach, with the use of machine learning models, could improve the results of the forecast analysis of extreme meteorological phenomena, in particular when combined with traditional physical forecasting models [5][9][11].

Data Availability

Data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

Authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Simas F., Barros R., Salvador L., Weber M., Amorim S.: A Data Exchange Tool Based on Ontology for Emergency Response Systems. Metadata and Semantics Research, 74-79 (2017). doi: 10.1007/978-3-319-70863-8_7
2. Li J., Liu L., Duy Le T., Liu J.: Accurate data-driven prediction does not mean high reproducibility. Nature machine intelligence (2020). doi: 10.1038/s42256-019-0140-2
3. Almanacco della Scienza CNR, n. 9 (2017).
4. Rezzani A.: Big Data. Maggioli editore (2013).
5. Wan Z., Vlachas P., Koumoutsakos P., Sapis T.: Dataassisted reduced-order modeling of extreme events in complex dynamical systems. Plos One (2018).
6. Ozdemir S.: Data Science, Apogeo (2018).
7. Virkus S., Garoufallou E.: Data Science from a Perspective of Computer Science. Metadata and Semantics Research, 209‒219 (2019). doi: 10.1007/978-3-030-36599-8_19
8. Zschocke T., Villagrán de León J., Beniest J.: Enriching the Description of Learning Resources on Disaster Risk Reduction in the Agricultural Domain: An Ontological Approach. MTSR, 320-330 (2010). doi: 10.1007/978-3-642-16552-8_29
9. Hosni H., Angelo Vulpiani A.: Forecasting in light of big data. Physics (2017).
10. Grunder-Fahrer S., Schlaf A., and Wustmann S.: How Social Media Text Analysis Can Inform Disaster Management. Springer (2017). doi: 10.1007/978-3-319-73706-5_17
11. Pathak J., Wikner A., Fussell R., Chandra S., Hunt B., Girvan M., Ott E.: Hybrid Forecasting of Chaotic Processes: Using Machine Learning in Conjunction with a Knowledge-Based Model. American Institute of Physics (2018).
12. Mandrioli D., Paola Spoletini P.: Informatica Teorica. CittàStudi edizioni (2011).
13. Melucci M.: Information Retrieval. Franco Angeli (2013).
14. Russell S., Norvig P.: Intelligenza Artificiale, vol. 1-2. Pearson (2010).
15. Sipser M.: Introduzione alla teoria della computazione, Maggioli editore (2016).
16. Fallucchi F., Tarquini M., De Luca E.W.: Knowledge Management for the Support of Logistics During Humanitarian Assistance and Disaster Relief (HADR). Springer, Cham (2016). doi: 10.1007/978-3-319-47093-1_19
17. David S., Hrubes P., Moran S., Shipilka A., Yehudayoff A.: Learnability can be undecidable. Nature machine intelligence (2019).
18. Raschka S.: Machine Learning con Python. Apogeo (2015).
19. Marsili Libelli S.: Modelli matematici per l'ecologia. Pitagora editrice (1989).
20. Banbura M., Giannone D., Modugno M., Reichlin L.: Now-casting and the real-time data flow. European central bank (2013).
21. Santos L., Sicilia M., Padrino S.: Ontologies for Emergency Response: Effect-Based Assessment as the Main Ontological Commitment. MTSR, 93-104 (2011).
22. Comincioli V.: Problemi e modelli matematici nelle scienze applicate. Casa editrice Ambrosiana (1993).
23. Report annuario dei dati ambientali. ISPRA (2017).
24. Cammarata S.: Reti neuronali. Dal Perceptron alle reti caotiche e neurofuzzy. Etas Libri (1997).
25. Cusani R., Inzerilli T.: Teoria dell'Informazione e Codici. Ed. Ingegneria 2000, (2008).
26. M. G. Bergomi, P. Frosini, D. Giorgi, N. Quercioli: Towards a topologicalgeometrical theory of group equivariant non-expansive operators for data analysis and machine learning. Nature Machine Intelligence, vol. 1, n. 9.
27. Hickey J.: Using machine learning to "Nowcast" precipitation in high resolution. Google AI Blog (2020).