A REVIEW OF GEOSPATIAL DATA ANALYSIS AND VISUALIZATION USING MACHINE LEARNING

Floarea-Maria BREBU, Lecturer PhD. eng. jur., Politehnica University Timisoara, Civil Engineering Faculty, Romania, floarea.brebu@upt.ro
Cosmin-Constantin MUȘAT, Associate Professor PhD. eng., Politehnica University Timisoara, Civil Engineering Faculty, Romania, cosmin.musat@upt.ro
Clara-Beatrice VÎLCEANU, Associate Professor habil. PhD. eng. ec., Politehnica University Timisoara, Civil Engineering Faculty, Romania, beatrice.vilceanu@upt.ro
Ioan-Sorin HERBAN, Prof. habil. PhD. eng., Politehnica University Timisoara, Civil Engineering Faculty, Romania, sorin.herban@upt.ro
Carmen GRECEA, Prof. habil. PhD. eng., Politehnica University Timisoara, Civil Engineering Faculty, Romania, carmen.grecea@upt.ro

Abstract*: Currently, machine learning, including artificial neural networks of different architectures and support vector machines, provides extremely important tools for analyzing, processing and visualizing geo and intelligent environmental data. Machine learning represents an important complement to traditional techniques, such as geostatistics.*
In this article, we present a review of several applications from the last period of using machine learning for geospatial data: regional classification of environmental data, continuous mapping, environmental data, and pollution, including the use of automated algorithms, optimization (design / redesign) of monitoring networks.

Keywords: *geospatial data; satellite images; machine learning; prediction models; analysis; visualization*

1. Introduction

Thousands of satellite data sets are freely available online, but scientists need the right tools to efficiently analyze the data and share the results. The roots of machine learning in remote sensing date back to the 1990s. It was originally introduced to automate the knowledge-based building for remote sensing and was thus further developed by experts at that time, how the proposed approach of the system for automatic learning has generated the highest analysis and visualization accuracy compared to conventional methods. After such similar developments, machine learning was soon adopted as an important tool by the remote sensing community.

In recent decades, it has been used in all kinds of projects, to analyse and visualize this geospatial data. Geospatial data analysis using machine learning proves to be an advantage in anticipating the processing and analysis results of this large-dimensional data [1].

There are multiple famous machine learning algorithms in use today, and new algorithms are popping up every other day. Some of the widely known algorithms are Support Vector Machines, Neural Networks, Random Forests, K-Nearest Neighbours, Decision Trees, K-Means, Principal Component Analysis.

It can be said that Machine learning (ML) is very useful for prediction, analysing and visualization geospatial data in many domains. ML tools are mainly founded out a place for filtering, interpretation, and prediction information [2].

In today's era, by using analytical machine learning algorithms that include artificial neuronal networks (ANN), vector support machines (SVMS) specialists can analyse the processing and visualization of geospatial data from satellites.

Thus, this scientific report aims to review some of the recent best practices with ML using geospatial data, considering the learning process and the properties of this data for the full understanding of:

- How much progress has been made in handling spatial properties of data in ML?
- What are some of the best practices to this respect?
- How to store and access spatial data for ML?
- How can we process, access, manipulate, and display geospatial data?
- Where do opportunities exist for future research with and on spatially explicit ML?

2. Machine Learning and methodologhy

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Machine learning is an important component of the growing field of data science. Using statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase, requiring them to assist in the identification of the most relevant business questions and subsequently the data to answer them.

The learning system of an automatic learning algorithm is divided into three main parts [6]:

A. **A Decision Process**: In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labelled or unlabelled, your algorithm will produce an estimate about a pattern in the data.

B. **An Error Function**: An error function serves to evaluate the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model.

C. **A Model Optimization Process**: If the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this evaluate and optimize process, updating weights autonomously until a threshold of accuracy has been met.

There are three subcategories of machine learning [8]:

- **Supervised** machine learning models are trained with labelled data sets, which allow the models to learn and grow more accurate over time. Supervised machine learning is the most common type used today.
- In **unsupervised** machine learning, a program looks for patterns in unlabelled data. Unsupervised machine learning can find patterns or trends that people are not explicitly looking for.
- **Reinforcement** machine learning trains machines through trial and error to take the best action by establishing a reward system. Reinforcement learning can train models to play games or train autonomous vehicles to drive by telling the machine when it made the right decisions, which helps it learn over time what actions it should take.

Bibliographic selection in our article started for a first step from the questions listed in the previous section, thus identifying that machine learning is a field of future in prediction, processing, analysis, and visualization of geospatial data. Many satellites orbit around the earth and provide users with a large amount of data, such as optical and radar images, thus making available free images by accessing the databases of: Copernicus, Sentinel, Landsat, etc.

The bibliographic study conducted in this research report was to inform the scientific sector about the problems that have been resolved so far in the prediction, processing, analysis, and visualization of geospatial data using machine learning. This leads to bibliographic selection by keywords: geospatial data, satellite images, machine learning, prediction models, analysis, visualization, and published in databases as: IEEE, Springer, Elsevier, Scopus, reasarchgate.net, Web of Science, Crossref. Also, for the selection of the bibliographical articles, we also considered that they should be published in journals, be part of research projects in the research area proposed, and to present best practices from the literature.

We have selected articles that comply with the following criteria: Published after 2017, have a minimum of 3 citations, published in journal or major conferences, and international research projects.

Articles based only on title and summary without access to all information have been excluded.

To understand the applicability of machine learning in geospatial data we searched in literature and Moocs.

Based on these criteria we have identified 19 scientific articles that present different problems solved in the field of machine learning for the processing, analysis, and visualization of geo-space data.

3. Results

These bibliographic studies have been classified according to keywords (satellite images, geospatial data, big data, predictive, processing, analysis, visualization in machine learning), depending on the period when they were published (2017-2021 in journals, conferences, and research projects), as well as on what problems are solved in this field (Table 1).

We have studied, from the bibliographic source, in the scientific report, several important applications of machine learning algorithms for geospatial data: regional classification of environmental data, mapping of continuous environmental data including automatic algorithms, optimization (design/redesign) of monitoring networks.

All these studies have as a solution the use of the Python programming language because it has developed libraries to streamline prediction, analysis, visualization of geo-space data and share results.

Tabel 1. Taxonomy and representative publications of Geospatial data using Machine Learning

Category	Publications
representative classification algorithms of machine learning	[1], [2], [10], [12]
practices are introduced into new ways of working with geospatial data using artificial intelligence and automated learning techniques	[4], [7], [12], [16], [18] [19]
evolution from Jupyter to JupyterLab or Pyjeo from JEO-Lab	[3], [9], [11], [13], [14], [17]
analyse a large GeoData, resulting in an interactive map	[1], [2], [10], [12]
Artificial intelligence for Earth monitoring MOOCs	[5], [6], [8], [15]

Articles 1, 2 and 10 of bibliographic research refer to how users can realize a map (topography, historical or general) at different scales using different databases that store geospatial data (OpenStreetMap, Sentinel etc.) [1] [2] [10].

In these studies, the focus is on the elimination and aggregation of the building layer, for which each building in a large scale was classified as “0-eliminated,” “1-retained,” or “2-aggregated.” Machine-learning classification algorithms were then used for classifying the buildings. Furthermore, the buildings were classified using representative classification algorithms of machine learning: decision tree (DT), k-nearest neighbor (k-NN), naive Bayes (NB), and support vector machine (SVM). In addition, geometric, topological, and thematic properties were used as input features.

The result of articles research consists in the classification of buildings using representative classification algorithms of machine learning through decision tree (DT), k-nearest neighbour (k-NN), naive Bayes (NB) and vector support machine (SVM). In addition, geometric, topological, and thematic properties were used as input characteristics.

The authors of article 2 did not target the whole process of generalizing buildings; rather they considered the classification of buildings, which could be used as a preparatory step for the generalization of buildings.

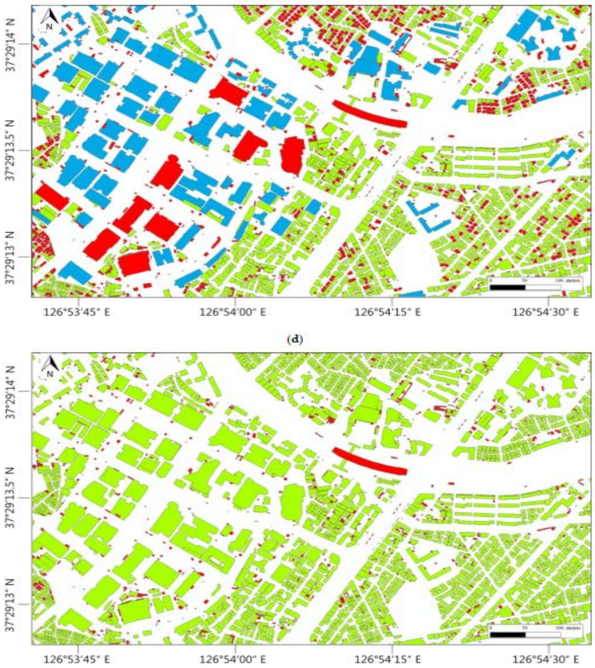
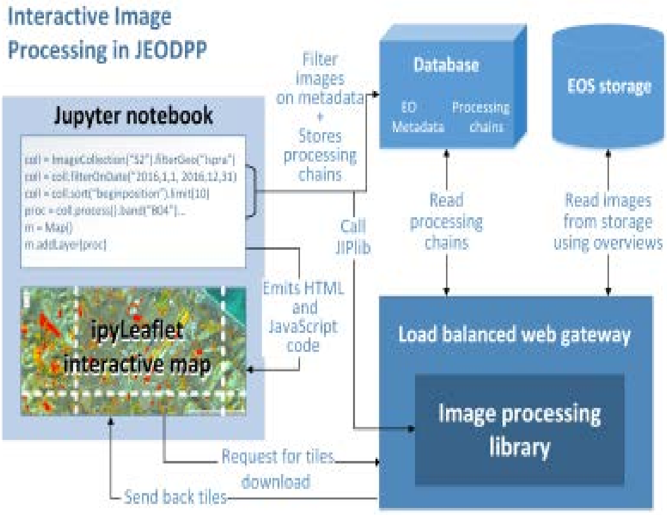


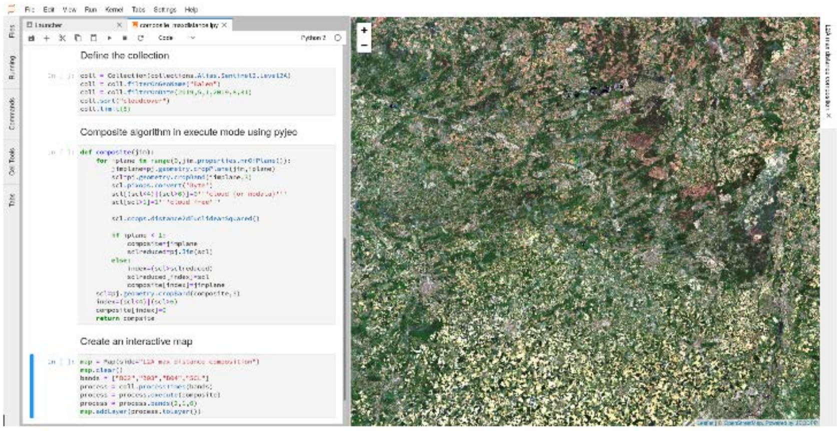
Figure 1. Visualized results [2]

Another field approach set out in Articles 3, 9, 11,13,14, and 17 proposes the analysis and visualization of data using satellite data provides precise procedures to analyse a large GeoData, resulting in an interactive map [3] [9] [11] [13] [14] [17]. These papers concentrate on the advances regarding the interactive analysis and visualization layer. The following aspects are detailed: evolution from Jupyter to JupyterLab or Pyjeo from JEO-Lab, availability of new data collections, the possibility to execute arbitrary Python code, and applications for users without programming capabilities exploiting the temporal dimension of geospatial data cube.


Interactive Image Processing in JEODPP
Jupyter notebook
ipyLeaflet interactive map
Filter images on metadata + Stores processing chains
Call JIPlib
Emits HTML and JavaScript code
Database
EO Metadata
Processing chains
EOS storage
Read processing chains
Read images from storage using overviews
Load balanced web gateway
Image processing library
Request for tiles download
Send back tiles

```
coll = ImageCollection("S2").filterGeo("Ispra")
coll = coll.filterOnDate("2016,1,1, 2016,12,31)
coll = coll.sort("beginposition").limit(10)
proc = coll.process().band("B04")...
m = Map()
m.addLayer(proc)
```

Figure 2. The interactive processing and visualization model [3]


File Edit View Run Kernel Tabs Settings Help
Launcher
composite_maxdistance.ipynb
Code
Python 2
Files
Running
Commands
Cell Tools
Tabs
+
−

Define the collection

```
coll = Collection(collections.Atlas.Sentinel2.Level2A)
coll = coll.filterOnGeoName("Balen")
coll = coll.filterOnDate(2019,6,1,2019,6,31)
coll.sort("cloudcover")
coll.limit(5)
```

Composite algorithm in execute mode using pyjeo

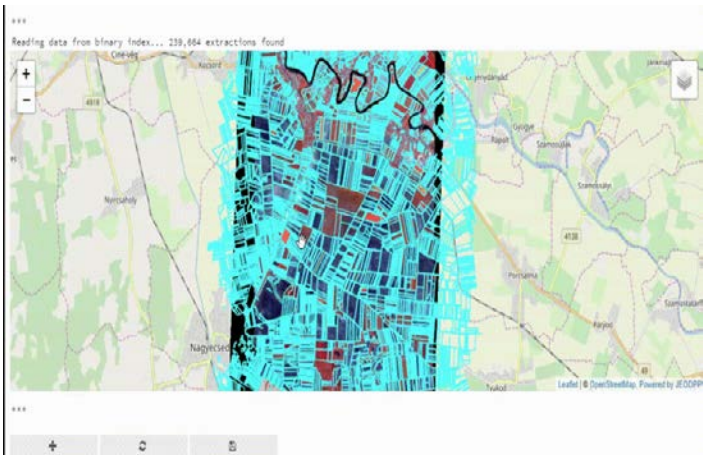
```
def composite(jim):
    for iplane in range(0,jim.properties.nrOfPlane()):
        jimplane=pj.geometry.cropPlane(jim,iplane)
        scl=pj.geometry.cropBand(jimplane,3)
        scl.pixops.convert('Byte')
        scl[(scl<4)|(scl>6)]=0'''cloud (or nodata)'''
        scl[scl>1]=1'''cloud free'''

        scl.ccops.distance2dEuclideanSquared()

        if iplane < 1:
            composite=jimplane
            sclreduced=pj.Jim(scl)
        else:
            index=(scl>sclreduced)
            sclreduced[index]=scl
            composite[index]=jimplane
    scl=pj.geometry.cropBand(composite,3)
    index=(scl<4)|(scl>6)
    composite[index]=0
    return composite
```

Create an interactive map

```
map = Map(title="L2A max distance composition")
map.clear()
bands = ["B02","B03","B04","SCL"]
process = coll.process(bands=bands)
process = process.execute(composite)
process = process.bands(2,1,0)
map.addLayer(process.toLayer())
```

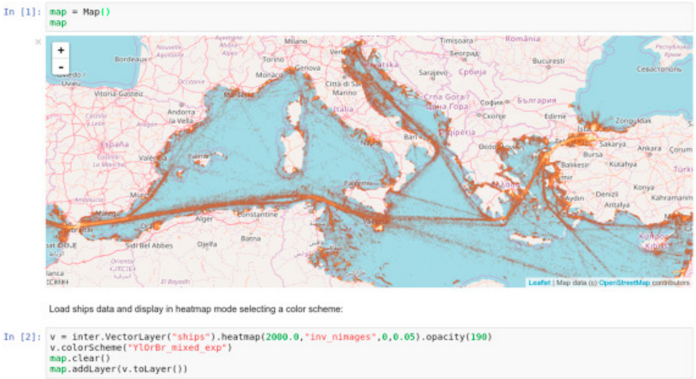
Figure 3. Interactive analysis and visualization in JEO-lab [18]

The results presented in these articles are implemented on the JRC Big Data Platform (JEODPP, JEO-lab). The JEODPP platform and has been characterized by providing users with a wide variety of raster and vector geospatial datasets. This was done for all functions originating from the software suite for processing geospatial data, as well as a series of morphological image analysis functions, including hierarchical image segmentation based on constrained connectivity [3]. The result of the research obtained was the realization of various interactive maps for: agriculture, forests etc.



```
Reading data from binary index... 239,664 extractions found
```

Figure 4. Interactive map for agriculture [3]



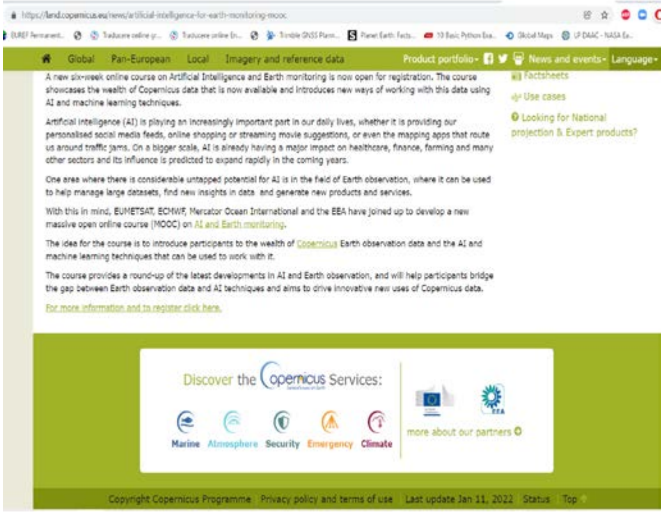
```
In [1]: map = Map()
        map
```

Load ships data and display in heatmap mode selecting a color scheme:

```
In [2]: v = inter.VectorLayer("ships").heatmap(2000.0,"inv_nimages",0,0.05).opacity(190)
        v.colorScheme("YlOrBr_mixed_exp")
        map.clear()
        map.addLayer(v.toLayer())
```

Figure 5. Jupyter notebook with interactive visualization and rendering of the density map of the ships detected from Sentinel-1 images over the Mediterranean sea during the period October 2014 to September 2016 [17]

To predict, analyse, visualize this geospatial data and share results, in Articles 5 and 15, Copernicus is helping users with a MOOC course: Artificial intelligence for Earth monitoring MOOCs. In 6, 7, 8, 12, 15, 16, 18 and 19 the authors give examples of basic methods, applications, and visualizations to processes satellite data sets for Earth science research.



https://land.copernicus.eu/news/artificial-intelligence-for-earth-monitoring-mooc

A new six-week online course on Artificial Intelligence and Earth monitoring is now open for registration. The course showcases the wealth of Copernicus data that is now available and introduces new ways of working with this data using AI and machine learning techniques.

Artificial intelligence (AI) is playing an increasingly important part in our daily lives, whether it is providing our personalised social media feeds, online shopping or streaming movie suggestions, or even the mapping apps that route us around traffic jams. On a bigger scale, AI is already having a major impact on healthcare, finance, farming and many other sectors and its influence is predicted to expand rapidly in the coming years.

One area where there is considerable untapped potential for AI is in the field of Earth observation, where it can be used to help manage large datasets, find new insights in data and generate new products and services.

With this in mind, EUMETSAT, ECMWF, Mercator Ocean International and the EEA have joined up to develop a new massive open online course (MOOC) on AI and Earth monitoring.

The idea for the course is to introduce participants to the wealth of Copernicus Earth observation data and the AI and machine learning techniques that can be used to work with it.

The course provides a round-up of the latest developments in AI and Earth observation, and will help participants bridge the gap between Earth observation data and AI techniques and aims to drive innovative new uses of Copernicus data.

For more information and to register click here.

Figure 6. Artificial Intelligence for Earth Monitoring MOOC [5]

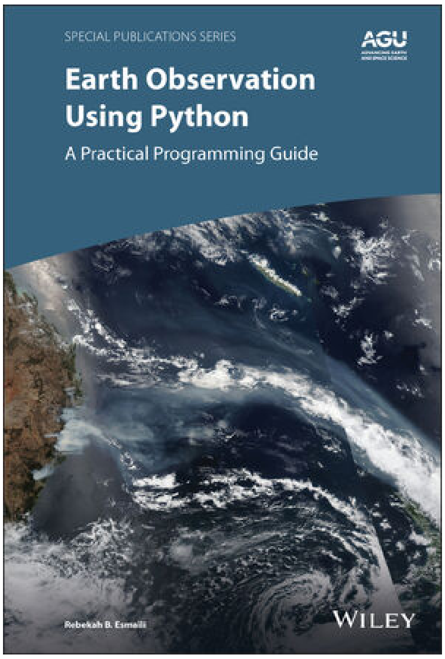

SPECIAL PUBLICATIONS SERIES
AGU
Earth Observation Using Python
A Practical Programming Guide
Rebekah B. Esmaili
WILEY

Figure 7. Earth Observation using Python: A Practical Programming Guide [16]

These practices are introduced into new ways of working with this data using artificial intelligence and automated learning techniques.

4. Discussions

In these all-bibliographic research, three broad classes of machine learning algorithms were used. One class is learning the supervised, the second is learning the unsupervised, and the third is enhanced learning. The difference between supervised and unsupervised is that, while using supervised algorithms, one has a data set that contains the output column; while using unsupervised algorithms, one has only a huge data set, and it is the duty of the algorithm to group the data set into different classes, based on the relationship he identified between different records. Learning consolidation is slightly different. In consolidation learning, the algorithm is provided with an environment, and the algorithm makes decisions in that environment. It continues to improve with each decision based on the feedback it receives for its latest decision. These are the three famous algorithms used to analyse and visualize geospatial data from satellite images.

Several applications of automatic learning algorithms for geospatial data have been used for: regional classification of environmental data, mapping of continuous environmental data, including automated algorithms, and optimization (design / redesigning) of monitoring networks.

The following typical geospatial data analysis issues and appropriate approaches / methods that can be used to solve them can be concluded from the study:

- Spatial predictions/interpolations: deterministic interpolators, geostatistics, machine learning. Spatial predictions in a high-dimensional geo-feature space – machine learning.
- Modelling and spatial predictions with uncertainties (e.g., considering measurement errors): geostatistics, machine learning.

- Multivariate joint predictions of several variables: geostatistics (co-kriging), machine learning (multi-task learning).
- Risk mapping – modelling of local probability density function: geostatistics (indicator kriging, simulations), machine learning (mixture density networks).
- Modelling of spatial variability and uncertainty, conditional simulations (spatial Monte Carlo simulations): geostatistical conditional stochastic simulations (sequential Gaussian simulations, indicator simulations etc.).
- Optimization of monitoring networks (spatial sampling design/redesign): space filling models, geostatistics (kriging, simulations), machine learning (Support Vector Machines). The basic idea of using Support Vector Machines for spatial sampling design is that only support vectors are important measurement points contributing to the solution of mapping problem.
- Data mining in a high-dimensional geo-feature space: machine learning (supervised and unsupervised learning algorithms).

These solved problems are currently used for research in modelling geospatial data, for making an interactive map of the studied area (example a country) such as: topo-climatic modelling, natural hazard assessments (landslides, avalanches), pollution mapping (inside the radon, heavy metals, air and soil pollution), natural resources assessment, remote classification of remote sensing images, analysis and visualization of socio-economic data etc.

An unresolved problem in predicting, analysing, and visualizing geospatial data using Auto Learning is creating a single interactive map for a city that contains all these solutions that have been resolved so far, such as:

- progress made in handling spatial properties of the data in Machine Learning using Big Data platforms;
- best practices in this respect are the use of Python programming language with libraries developed for the research field;
- storing and accessing space data for ML;
- processing, accessing, the manipulation and display of geospatial data.

As stated in the previous paragraph, as an opportunity for future research using ML will be prediction, analysis, and visualization of geospatial data will be the creation of a single interactive map for an administrative territory and, of course, the realization of the transformation from the WGS84/Mercator projection into the 1970 stereographic projection currently used in Romania.

5. Conclusions

Automatic learning algorithms are adaptive, non-linear, extremely powerful universal tools. They have been used successfully in many geo- and environmental applications. In principle, they can be used efficiently in all stages of environmental data extraction: analysis of exploratory spatial data, recognition and modelling of spatial-temporal models, and decision-oriented mapping.

Current trends in ML applications for geo- and environmental sciences refer to: reducing nonlinear dimensionality and visualizing data; analysis and modelling of data in high-dimensional geo-characteristic spaces; rapid modelling of physical processes and other processes in hybrid models; extraction, modelling, and predictions of spatial-temporal models / structures (data exploitation and forecasting).

It should be noted that, being data-based models, they need in-depth knowledge of experts to be able to be applied correctly and efficiently, from preprocessing data to interpreting and justifying results.

6. References

1. *J. Wegner, R. Roscher, M. Volpi, and F. Veronesi, "Foreword to the Special Issue on Machine Learning for Geospatial Data Analysis," ISPRS Int. J. Geo-Inf., vol. 7, no. 4, Art. no. 4, Apr. 2018, doi: 10.3390/ijgi7040147.*
2. *J. Lee, H. Jang, J. Yang, and K. Yu, "Machine Learning Classification of Buildings for Map Generalization," ISPRS Int. J. Geo-Inf., vol. 6, no. 10, Art. no. 10, Oct. 2017, doi: 10.3390/ijgi6100309.*
3. *P. Soille et al., "A versatile data-intensive computing platform for information retrieval from big geospatial data," Future Gener. Comput. Syst., vol. 81, pp. 30–40, Apr. 2018, doi: 10.1016/j.future.2017.11.007.*
4. *N. Tohidi and R. B. Rustamov, "A Review of the Machine Learning in GIS for Megacities Application," in Geographic Information Systems in Geospatial Intelligence, R. B. Rustamov, Ed. IntechOpen, 2020. doi: 10.5772/intechopen.94033.*
5. *"Artificial Intelligence for Earth Monitoring MOOC — Copernicus Land Monitoring Service," Jan. 08, 2022. Accessed: Jan. 08, 2022. [Online]. Available: https://land.copernicus.eu/news/artificial-intelligence-for-earth-monitoring-mooc*
6. *R. B. Esmaili, "A New Practical Guide to Using Python for Earth Observation," Eos, Aug. 06, 2021. http://eos.org/editors-vox/a-new-practical-guide-to-using-python-for-earth-observation (accessed Jan. 08, 2022).*
7. *"An Intro to the Earth Engine Python API | Google Earth Engine," Google Developers, Jan. 08, 2022. https://developers.google.com/earth-engine/tutorials/community/intro-to-python-api-guiattard (accessed Jan. 08, 2023).*
8. *FutureLearn, "Artificial Intelligence (AI) for Earth Monitoring," FutureLearn, Jan. 08, 2022. https://www.futurelearn.com/courses/artificial-intelligence-for-earth-monitoring (accessed Jan. 08, 2023).*
9. *D. H. Hagos et al., "ExtremeEarth Meets Satellite Data from Space," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 14, pp. 9038–9063, 2021, doi: 10.1109/JSTARS.2021.3107982.*
10. *J. Uhl, S. Leyk, Y.-Y. Chiang, W. Duan, and C. Knoblock, "Map Archive Mining: Visual-Analytical Approaches to Explore Large Historical Map Collections," ISPRS Int. J. Geo-Inf., vol. 7, no. 4, Art. no. 4, Apr. 2018, doi: 10.3390/ijgi7040148.*
11. *A. Kovacs-Györi et al., "Opportunities and Challenges of Geospatial Analysis for Promoting Urban Livability in the Era of Big Data and Machine Learning," ISPRS Int. J. Geo-Inf., vol. 9, no. 12, Art. no. 12, Dec. 2020, doi: 10.3390/ijgi9120752.*
12. *M. Gangappa, C. Kiran, and P. Sammulal, "Techniques for Machine Learning based Spatial Data Analysis: Research Directions," Int. J. Comput. Appl., vol. 170, no. 1, Art. no. 1, Jul. 2017, doi: 10.5120/ijca2017914643.*
13. *M. Campos-Taberner et al., "Understanding deep learning in land use classification based on Sentinel-2 time series," Sci. Rep., vol. 10, no. 1, Art. no. 1, Oct. 2020, doi: 10.1038/s41598-020-74215-5.*
14. *"Accelerated Analytics Platform | OmniSci," Jan. 21, 2022. https://www.omnisci.com/ (accessed Jan. 21, 2023).*
15. *V. Syrris, P. Hasenohr, B. Delipetrev, A. Kotsev, P. Kempeneers, and P. Soille, "Evaluation of the Potential of Convolutional Neural Networks and Random Forests for Multi-Class Segmentation of Sentinel-2 Imagery," Remote Sens., vol. 11, no. 8, Art. no. 8, Apr. 2019, doi: 10.3390/rs11080907.*

16. *A. Kovacs-Györi et al., "Opportunities and Challenges of Geospatial Analysis for Promoting Urban Livability in the Era of Big Data and Machine Learning," ISPRS Int. J. Geo-Inf., vol. 9, no. 12, Art. no. 12, Dec. 2020, doi: 10.3390/ijgi9120752.*
17. *P. Kempeneers and P. Soille, "Optimizing Sentinel-2 image selection in a Big Data context," Big Earth Data, vol. 1, no. 1–2, Art. no. 1–2, Dec. 2017, doi: 10.1080/20964471.2017.1407489.*
18. *P. Kempeneers, O. Pesek, D. De Marchi, and P. Soille, "pyjeo: A Python Package for the Analysis of Geospatial Data," ISPRS Int. J. Geo-Inf., vol. 8, no. 10, Art. no. 10, Oct. 2019, doi: 10.3390/ijgi8100461.*
19. *"Working with Geospatial Data in Python," GeeksforGeeks, Aug. 23, 2021. https://www.geeksforgeeks.org/working-with-geospatial-data-in-python/ (accessed Jan. 11, 2023).*