International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I12P201

Volume 70 Issue 12, 1-12, December 2022

Original Article

Metageosystem Analysis Based on a System of Machine Learning and Simulation Algorithms

Stanislav Yamashkin[1], Anatoliy Yamashkin[2], Milan Radovanović[3,4], Marko Petrović[3], Ekaterina Yamashkina[1]

[1]*Institute of Electronics and Lighting Engineering, National Research Mordovia State University, Mordovia, Russia*
[2]*Geography Faculty, National Research Mordovia State University, Mordovia, Russia*
[3]*Geographical Institute "Jovan Cvijić", Serbian Academy of Sciences and Arts, Belgrade, Serbia*
[4]*Institute of Sports, Tourism and Service, South Ural State University, Chelyabinsk Oblast, Russia*

[1]*Corresponding Author : yamashkinsa@mail.ru*

Received: 14 October 2022 Revised: 05 December 2022 Accepted: 16 December 2022 Published: 24 December 2022

Abstract *- The study presented in this article aims to solve the scientific problem of increasing the efficiency of using modeling and machine learning models in solving problems of analysis and classification of metageosystems. The article describes an approach aimed at improving the efficiency of machine learning models in solving the problem of classifying metageosystems, which makes it possible to overcome the limitations imposed on the use of convolutional neural network models. The article is also devoted to solving the scientific problem of multifaceted quantitative analysis of intercomponent links in metageosystems of different hierarchical levels based on simulation modeling. It is proved that the study of the structure and properties of metageosystems should be based on the analysis of complex properties and patterns of interaction of territorial systems distributed in space. A set of requirements for the framework for creating simulation models of spatial processes is formulated, and an algorithm for developing a simulation model that describes the spatio-temporal processes occurring in complex territorial systems is presented. The study also showed that combining models into an ensemble based on the proposed metaclassifier architecture makes it possible to increase the stability of the analyzing system: the accuracy of decisions made by the ensemble tends to tend to the accuracy of the most efficient monoclassifier of the system. Systematic analysis of territory descriptors integrated based on data from different sources significantly increases the accuracy of metageosystem classification.*

Keywords *- Geosystem approach, Metageosystems, Earth remote sensing data, Classification, Machine learning, Simulation.*

1. Introduction

Geosystems act as an object of systemic spatial analysis in modern science in the field of spatial data analysis. The doctrine of geosystems is relevant both in the field of studying natural objects and processes and in the field of analyzing their interaction with social and economic systems. In this case, it is necessary to introduce the concept of "metageosystem", digital models of the territory as the main object of spatial analysis. The study presented in this article aims to solve the scientific problem of increasing the efficiency of using machine learning models in solving the problem of classifying metageosystems based on Earth remote sensing (ERS) data.

The solution to the scientific problem of developing a multifaceted quantitative analysis of intercomponent relationships in metageosystems of different hierarchical levels should be based on the construction of digital models of geosystems at the levels of systems of another class. In the first stage, a selection of properties is made that most accurately reflect the object under study with the required level of abstraction. After that, variables that describe the properties with the greatest completeness are selected, and the parameters on which they are observed are determined (time, space, groups). Each variable is described according to its main characteristics, which are significant from a methodological point of view; at the level of data systems, variables are transformed into a set of states, displayed on a single parametric set; at the level of generating systems, the data obtained are converted into forms, based on the nature of the object under study and the objectives of the study.

Models built in this way allow solving the following key tasks:

1) Assessment of the strength and nature of intercomponent links in metageosystems;
2) Determination of the number of factors describing the territorial change in the properties of metageosystems;
3) Interpretation and justification of the physical meaning of the selected factors.

The solution of the designated problem points makes it possible to build digital statistical models of the spatial structure of geosystems based on continuous and discrete approaches; evaluate the resulting models through a quantitative analysis of the quality of displaying the territorial variability of those landscape properties that were not directly introduced into the original multidimensional model of geosystems [1]. The solution to the scientific problem of analyzing the structure and properties of metageosystems should be based on the analysis of complex properties and patterns of interaction of territorial systems distributed in space [2, 3]. Machine learning algorithms and deep neural networks, in particular, are widely used in solving the problem of space-temporal forecasting [4]. At the same time, using such models is associated with several difficulties and threats [5, 6].

First, because neural network models are structures built and trained based on the "black box" strategy, analysing their behavior to adjust the decision-making features becomes unnecessarily complicated. Secondly, training deep models capable of extracting complex hierarchical features about territorial processes and systems requires the preparation of huge training data banks, which, among other things, involves field research [7, 8]. Thirdly, the ambiguity of solving the problem of assessing the accuracy and error of machine learning models causes legitimate concerns of experts, expressed in high uncertainty when comparing the abstract mathematical assessment of the model error with the requirements of standards and performance criteria [9]. As a result, the solution to the problem of spatio-temporal forecasting of territorially distributed processes should be sought in the field of methods and algorithms that allow analyzing models of real-world systems based on the "white box" strategy, which implies the possibility of analyzing the functioning of each component of a complex system, which is an abstraction of a real territorial system and depicting various internal interactions and relationships in it. Spatio-temporal quantitative analysis of the components of this system makes it possible to understand better-distributed processes and a relatively accurate prediction of the directions of development of natural and natural-technogenic phenomena.

2. Research Method

The structure of studies of metageosystems is based on the principles:

1) Territoriality - the study of the natural resource potential and patterns of natural differentiation that determine the general patterns of economic development and patterns of development of environmental, social and economic processes;
2) Consistency - a conjugated study of all aspects of the functioning of metageosystems, taking into account the relationship of all elements that determine the features of ecological, social and economic development; interdependent association of structural elements and connections of different levels of natural, social and production systems, mutual linkage and consistency of all processes and phenomena;
3) Environmental - optimization of the interaction of natural, social and production systems based on the observance of ecological balance.

Based on the methodology of sequential analysis of geosystems of different hierarchical levels, digital models of geosystems of the following levels can be built:

1) The level of source systems involves the selection of properties that maximally reflect the objects under study based on the purpose of the study and variables that describe the selected properties with the greatest completeness.
2) The level of data systems, focused on transforming initial variables into a set of states, displayed on a single parametric set, characterized by temporal, spatial and attributive coordinates.
3) The level of generating systems involves transforming the obtained data into optimal representation forms based on the nature of the object under study and the study's objectives.
4) The level of structured systems, which implements the identification of relationships (connections) between geosystem variables based on using non-parametric and parametric statistics methods, simulation algorithms and machine learning models.

Structured models of geosystems make it possible to approach the consistent solution of problems: assessment of the strength and nature of intercomponent links in geosystems; determination of factors describing the territorial change (variation) of the properties of geosystems (the so-called assessment of the dimension of space); interpretation and justification of the physical meaning of the selected basic factors. These results, in turn, become the basis for solving the problems of constructing digital spatial models of geosystems based on continuous and discrete approaches with the possibility of evaluating the resulting models through a quantitative analysis of the properties of landscapes and their components based on a multidimensional model of geosystems.

The formed multifactorial hierarchical spatial models become the basis for forming information geoportal resources for solving project-oriented tasks in managing regional metageosystems: analyzing the structure and properties of lands, detecting natural and natural-technogenic objects, and predicting the development of emergencies and natural processes. In organizing research on geosystems, it is advisable to single out the following stages: substantiation of problem situations, preparation of initial data and development of algorithms for their merging, typological systematization of information and compilation of a synthetic map of geosystems, ensemble analysis of the susceptibility of geosystems to external influences, generation of output data and use of the results obtained.

2.1. Justification of Problem Situations

The most important stages in the creation of a synthetic map of geosystems are the identification of problem situations related to the development of geoecological processes, the creation of a system of test polygons at the regional and local levels for diagnosing the structure, functioning, and dynamics and evolution of geosystems, the formation of initial data, the modeling of the spatiotemporal structure of geosystems, the generation of output data, analysis of the results and their use for making managerial decisions. Thus, in the scheme of physical-geographical zoning of the Republic of Mordovia, the system of test polygons focused on revealing the spatial and temporal organization of geosystems in the zone of interaction between the forest-steppe of the Volga Upland and mixed forests of the Oka-Don Lowland. The region's economic development is associated with the relevance of making managerial decisions in the water balance regulation and optimization of water supply problems, minimizing the development of exogeodynamic processes, and maintaining soil fertility and biological diversity.

2.2. Data Preparation

The range of problematic situations determines the database structure that reflects the structure of geosystems at the regional and local levels. Regional GIS "Mordovia" includes a system of electronic maps with information on the main analyzed elements of geosystems, Earth remote sensing data (ERS). The use of test polygons is relevant from the standpoint of solving the problem of consolidating an informative set of training data samples for the operation of automated algorithms, the accuracy of which can be significantly increased by calibrating the classification process based on the method of ground measurements during field studies involving direct observation of objects [10]. Within the framework of test polygons, the morphological parts of natural and anthropogenic landscapes within the boundaries of individual land types are considered. Test polygons must meet a set of quality requirements:

1) Sufficient structural diversity, which implies the presence within the study area of landscapes sufficient to create exhaustive sets of reference samples characterizing landforms, soils, water bodies, vegetation cover, and anthropogenic objects;
2) High quality of recorded spatially associated characteristics that ensure the formation of a sufficient set of stable radiative-reflective characteristics of geosystems;
3) The location and size of the polygon sufficient for a representative sample of spatial data, which determines a statistically sufficient number of presented classes of objects;
4) The correctness of the methodological, regulatory and technological support for the process of consolidating spatial data containing objective and accurate information about the spatial and temporal organization of geosystems;
5) *Information content of the meta-description*, which provides essential information about spatial data, including information about the survey, description of meta-geosystems, and administrative characteristics.

Test polygons that meet the presented quality requirements can be used to optimize automated classification models based on the use of machine learning technologies and can be combined into a single system to improve the quality of design work in the field of landscape mapping, farm management, precision farming, and exploration of natural resource deposits, monitoring of natural and natural-technogenic natural processes.

2.3. Systematization of Data

Based on the widespread use of multizone satellite images, the synthetic map of geosystems acts as the central link of the regional geographic information system. It is the basis for the following:

1) Compilation (revision) of thematic maps that reveal the structure of geosystems and the development of geoecological processes;
2) Identification of zones of influence of geotechnical systems on the state of the environment;
3) Assessment of the negative consequences of the direct and indirect technogenic impact on geosystems;
4) Predicting the dynamics of changes in the geoecological situation;
5) Making managerial decisions to prevent (minimize) the development of destructive geoecological processes.

The space-time structure of geosystems can be represented as a set of processes of movement, exchange and transformation of energy, matter and information between its elements and the surrounding geographical space. The determining factors in the formation of the structure, development, dynamics and functioning of geosystems are macroclimatic factors, tectonic landforms, features of the water and geochemical regime, soil-biological and exogeodynamic processes, and plant communities. Synthetic mapping of geosystems provides the following taxonomic units hierarchy: category, subcategory, class, subclass, type, subtype, genus, and subgenus of landscapes. The use of spatial data of varying degrees of generalization contributes to establishing regularities in the spatio-temporal organization of geosystems, increases the reliability of interpretation, and contributes to a more accurate interpretation of diagnostic features.

The stage "Data analysis for compiling a digital map of geosystems" is aimed at implementing the automated analysis of spatial monitoring data using computational algorithms. Software and mathematical models, grouped into ensembles, function based on identifying statistical patterns and reference features of spatially distributed objects.

As the final step, the stage "Visualization, dissemination and practical use of the results obtained" is singled out. For its implementation, it is necessary to ensure the formation of a multimodel geoportal spatial database. The solution to the spatial data visualization problem is achieved through digital maps that function within the framework of geoportal web interfaces and attain the goal of dissemination and practical use of data on regional metageosystems.

Machine learning methods and algorithms can effectively interpret geospatial data characterized by spatial dependence, spatial heterogeneity and scalability [11]. At the same time, applying methods and algorithms of deep machine learning to geospatial data analysis faces many open problems that require a scientifically based solution. Machine learning models used to solve the geosystem classification problem can have different architectures (artificial neural networks, decision trees, support vector machines) and hyperparameters. Moreover, they are able to successfully train on different interpretive territory datasets, which can be multivariate and multimodel.

Types of geosystems reflect the structure of habitats and plant communities. An analysis of the hierarchical structure of geosystems shows revealing the features of the interaction between the forest-steppe and forest geosystems of the Volga Upland and the Oka-Don Lowland to develop methods and algorithms for analyzing and integrating information in spatial data infrastructures at the Mordovia test site (located between 53°38' and 55°11' N, 42°11' and 46° 45' E), it is advisable to single out 6 polygons for high-precision mapping at the topological level.

1) The Nuya polygon (center coordinates: 54°28' N, 45°54' E) reflects the interaction of geosystems of the elevated erosion-denudation and erosion-accumulation plains of the Pliocene-Pleistocene age. The key aspects of the development of synthetic mapping methods are diagnostics of the state and interaction of forest and meadow-steppe geosystems for landscape planning of agricultural, forestry and mining systems; substantiation of the zone of the ecological balance of republican significance, which ensures the stabilization of the water balance and minimization of the development of planar and linear erosion.
2) The Moksha-Temnikov polygon (center coordinates: 54°42′ N, 43°24′ E) is organized in the geosystems of forest landscapes of water-glacial and ancient alluvial plains. The primary vectors for the development of methods for diagnosing the state of metageosystems to justify a set of measures for the organization of the Sanaksar-Temnikov pilgrimage and tourism cluster while maintaining the sustainable functioning of the feeding area of the Carboniferous-Permian aquifer complex, the unique ecosystems of the "Reserved Mordovia".
3) The Alatyr-Smolny polygon (center coordinates: 54°48′ N, 45°29′ E) was created in an ancient hollow of glacial water runoff with coniferous and coniferous-broad-leaved forests; the main objects of research are the mapping of exogeodynamic processes in the zones of linear tectonic faults; planning of recreational systems.
4) The Saransk polygon (center coordinates: 54°11′ N, 45°11′ E) was organized for a comprehensive study of metageosystems to diagnose the development of geoecological processes and landscape planning for urban development of geosystems.
5) The Moksha polygon (center coordinates: 54°13′ N, 44°02′ E) was created for a comprehensive study of slope and floodplain exogeodynamic processes in the zones of geotechnical systems.
6) The Inerka Polygon (center coordinates: 54°03′ N, 45°53′ E) reflects the interaction of paragenetic systems of forest-steppe geosystems of the erosion-denudation plain and intrazonal forest landscapes of the valley of the river. Sura. The priority geoecological problem is the optimization of tourist and recreational development of the natural monument of republican significance, "Lake Inerka".

The database "System of thematic maps for test sites of the Republic of Mordovia" is based on thematic layers, the register of which forms the following data categories:

- geology (primary sections, Quaternary deposits, lithological composition of surface deposits),
- groundwater (depth of groundwater, aquifers, water class and type, content of fluorides, iron chlorides, sulfates),
- relief (digital elevation model, lineaments and ring structures),
- soils (spatial distribution of soils, ecological and geochemical stability),
- metageosystems and landscapes (landscape map, vegetation distribution, geobotanical map, specialization in agriculture).

The generated database "System of thematic maps for test sites of the Republic of Mordovia" is a digital characteristic of the system of test sites in the zone of contact between the forest-steppe of the layer-tier Volga Upland and forest landscapes of the Oka-Don Lowland.

Consolidated data can be used to develop and test automatic and automated methods for mapping metageosystems, modeling geoecological processes, and monitoring the state of development of geoecological situations using remote sensing data.

3. Results and Discussion

Spatio-temporal data models should simulate spatially distributed processes in three-dimensional or four-dimensional space. In the first case, each spatial object is described by such coordinates as latitude, longitude and time; relative or absolute height can also be considered in the second case. Both natural (hydrological, geological, plant) and technogenic (economic, social) systems can be subjected to simulation [12]. The organization of a methodological, architectural and software framework that allows transforming fuzzy boundaries and multifactorial characteristics of natural phenomena into a strictly formalized framework of spatio-temporal abstractions is of current importance [13]. The system properties of natural and natural-technogenic phenomena dynamically change when moving in space and time; accordingly, the display of these internal structural, including hierarchical, characteristics should be included in the spatio-temporal structure of the model at a given level of assumptions and abstraction.

3.1. Designing a Hierarchical Model of Geosystems

The development of models that represent and characterize complex dynamic distributed systems is based on object-oriented, domain-oriented, event-based, multi-agent, and graph approaches [14, 15]. Graph-based modeling strategies are used for the spatiotemporal representation of processes and phenomena in the form of events. One of the advantages of graph modeling is the possibility of flexible integration of the semantic constraints of modeling and the organization of a formalized approach to quantitative assessment, taking into account both the spatial and temporal organization of territorial systems.

Let us formulate a set of requirements for the framework for creating simulation models of spatial processes.

3.1.1. Versatility

The modeling framework should allow the formation of simulation models of a wide class of natural and natural-technogenic territorial systems with sufficient abstraction and representativeness.

3.1.2. Interpretability

Models formed based on the framework should allow for a transparent assessment of the features of the functioning of its individual components and identify cause-and-effect relationships between changes in the system characteristics of the model and its specific constituent elements.

3.1.3. Reliability

Models deployed based on the proposed framework should ensure sufficient minimization of discrepancies between the actual characteristics of the state of the simulated territorial system and the parameters of its abstract representation, calculated based on the algorithms introduced into the system.

3.1.4. Configurability and Extensibility

The modeling framework should allow the formation of flexible models of complex territorial systems that are easily refined and modified when new knowledge about the analyzed objects and processes appears.

3.1.5. Optimization

The simulation framework should allow the processing of large data on territorial systems with minimal computing power.

From the point of view of the set-theoretic approach, the system of simulation modeling of spatial processes is described by a tuple of sets: $MODEL = \langle GRAPH, AGENTS, SIMULATION \rangle$, where $GRAPH$ is a graph model of territorial systems, described by a set of spatially distributed components; $AGENTS$ - a set of model agents that simulate the movement of matter and energy between the components of territorial systems; $SIMULATION$ is a set that defines the parametric and algorithmic features of the simulation subsystem. The $GRAPH$ describes the structural organization of a territorial system with a given level of abstraction (Figure 1).

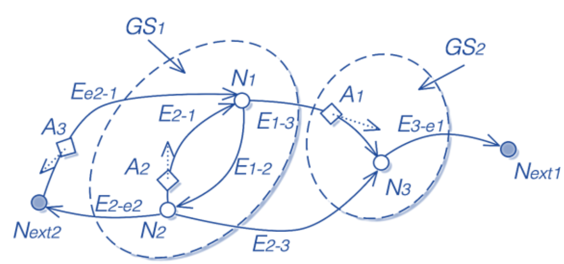


Fig. 1 Generalized graph structure of the model

The set of graph vertices (N) determines the selected territorial components of the system. Each vertex is characterized by a coordinate reference in Cartesian (relative) and geographic (absolute) coordinates. Refining the model leads to adding associated attributes and methods to each vertex. Graph vertices can be internal (accepting flows of objects from other model vertices) and boundary (modeling external connections of the considered geosystem).

The set of graph arcs (E) describes the existing connections between territorial objects and channels within which the transfer of matter and energy is carried out. It is advisable to use elements of the theory of queuing systems To describe the capacitive characteristics of connections (queues). Each arc E_i and vertex N_j of the $GRAPH$ graph is described by a set of static (invariant) and dynamic parameters. The routeing model RT_k is described by the characteristic of length (length_k), initial (source_k) and final (target_k) vertices, sets of vertices (N_k) and arcs (E_k) included in it.

$$\begin{cases} E_i = \langle EP_{i,static}, EP_{i,dynamic} \rangle, i = \overline{1,e} \\ N_j = \langle EP_{j,static}, EP_{j,dynamic} \rangle, j = \overline{1,n} \\ RT_k = \langle \mathrm{length}_k, \mathrm{source}_k, \mathrm{target}_k, \mathrm{N}_k, \mathrm{E}_k \rangle, k = \overline{1,r} \end{cases} \quad (1)$$

The $GRAPH$ graph model is divided into many GS segments to increase the flexibility of the territorial system model. Through segments, patterns of centralized cascading control over the properties of a set of model vertices can be defined. For example, the membership of a set of vertices in N to segment GS_i can determine the appearance of a set of attributes with given values and methods for these vertices that describe specific behaviors. Finally, graph segments can be organized into a multilevel structure to model the hierarchical organization of subordinate geosystems.

Model agents (A). The energy exchange between territorial components is the most important component of functioning natural and natural-technogenic spatial systems. Connections between territorial systems are also realized through the transfer of matter (flows of air, water, solid masses, and living organisms). This aspect can be implemented through agent-based modeling, in which decentralized abstract system objects described by a set of attributes and methods are moved and transformed within the model based on their state, as well as the state of the system as a whole and its individual components. Thanks to the agent component, the modeling of the system's uplinks are implemented.

A separate agent of the model A_i within each discrete stage of simulation is transformed and moved within the graph structure of the model based on the following provisions (aspects):

1) A set of admissible solutions is embedded in a set of algorithms and strategies for choosing behavior. This aspect determines the degree of freedom in the agent's behaviour. It may include such actions as changing the route of movement, stopping, starting the movement, and changing its own state parameters.
2) The current and predicted state of the agent. By means of a set of state properties (current and historical), an abstraction level is achieved to describe spatial objects of different natures: aggregate state of substances, dynamic characteristics of objects, and motivation of traffic participants.
3) The current and predicted state of the observed components of the model or the system as a whole. This group of decision-making parameters determines the possibility of responding to changes in external conditions, including the presence of external influences (merger and separation of hydrological flows, interaction of road traffic systems, and development of exogeodynamic processes).
4) Experience in interacting with agents and elements of the model. This group of provisions is transformed over time based on the accumulated experience, allowing you to model the interactions of smart learning systems. It can be implemented based on reinforcement learning through the transformation of an agent based on the assessment of the influence of the environment in response to certain decisions.

The $SIMULATION$ simulation subsystem is defined by a set of control algorithms $CONTROLERS$ and system parameters ($PARAMETERS$). Within the framework of the register of control algorithms, there is an algorithm for generating model agents, calculating routes for moving system agents, modeling and analyzing the laws of model transformation, and visualizing simulation results. Static parameters also represent system parameters MP_{static} (time-invariant characteristics that have spatial and attribute properties) and dynamic $MP_{dynamic}$ characteristics (values that change over time).

Simulation modeling makes it possible to analyze complex traffic processes with the necessary approximation, to evaluate the effectiveness and consequences of some regions of transformation of the infrastructure of technogenic metageosystems.

The algorithm for developing a simulation model that describes the spatio-temporal processes occurring in complex territorial systems should include the following sequence of steps:

1) Definition of the task and scenarios for using the system. During this project stage, a list of problems is determined that the developed simulation model aims to solve, and a register of key precedents is determined to form the designed solution's functionality space.
2) Designing and detailing the model. At this stage, the decomposition and refinement of the basic framework of the simulation model for solving specific problems are carried out; numerical data are collected, the statistical characteristics of the system are evaluated, and hypotheses are put forward to determine the level of abstraction of the model, which ensures acceptable modeling accuracy and achievement of target effects.
3) Algorithmic concretization of the model. As part of this stage, the necessary list of algorithms is being developed

that determines the functioning of a specific implementation of a system for simulating spatial processes aimed at solving particular problems of analyzing natural and natural-technogenic processes.

4) Implementation of the graphical interface and visualization subsystem. At this stage, the interface component of the system is proposed and developed, which should determine the ways of interaction with the model, methods and tools for analyzing its functioning, optimization and refinement.
5) Iterative approbation of the simulation system allows not only to model specific spatio-temporal processes but also to calibrate the model, refine the register of data used, and optimize the resulting quality.

Two development directions of the framework for simulation modeling spatial processes should be singled out. On the one hand, it is necessary to adhere to a deductive strategy, in which the model-building system should be initially designed to achieve the possibility of covering solutions to the maximum number of problems.

On the other hand, it is necessary to implement the inductive strategy as efficiently as possible, within which the positively proven improvements of particular specific models become the basis for the development of the simulation framework. The combined use of deductive and inductive strategies will provide an evolutionary improvement in the modeling framework and optimization of particular solutions developed on its basis.

3.2. Analysis of Spatial Data with the Construction of a Synthetic Map of Geosystems

In the past few years, the concept of deep machine learning has taken an important place in the field of spatial data analysis [16]. Still, the use of capacious models encounters a number of obstacles that significantly complicate their implementation: they can be effectively trained only on large sets of labeled spatial data, be subject to the problem of overfitting, poor generalization of information and poor interpretability, the process of training deep models places high demands on hardware [17]. The solution to the indicated problems is possible with the use of ensembles of classifiers built on the Ensemble Learning methodology, combining various models into a system and making it likely to increase the accuracy and stability of machine learning models [18].

Data on territorial systems can be multidimensional, including information about the spectral properties of the earth's surface, the features of its spatial organization, and other information of attributive, spatial and temporal nature. A set of features of a local territorial object (including spatial, spectral, and radiometric characteristics) can be formed based on satellite imagery or other sources. In contrast, a spatial area can be classified based on a pixel-by-pixel analysis or by extracting features from fragments of an area of different sizes. The set of features of a local object, which themselves can be represented in the form of tensors, determines the base level of the created geospatial terrain model.

Under the territory's geosystem model, we mean the data characterizing the enclosing hierarchical geosystems. In accordance with the geosystem approach, the enclosing geosystem significantly impacts the properties of hierarchically subordinate formations. Earth remote sensing data are a useful source of information about it [28]. The process of obtaining data at various levels of the hierarchy is potentially subject to automation through integration with spatial data infrastructure services and third-party providers of spatial information.

The optimization algorithm for constructing a metageosystem classification model includes the following sequence of steps:

1) Development of a system of requirements for a data analysis model, definition of a register of analyzed data, qualitative and quantitative requirements for the result of functioning;
2) Designing the basic architecture of the model with the decomposition of top-level components into linear and branching structures, determining the form of input and output signals, and acceptable system quality indicators;
3) Iterative optimization of the model in accordance with the rule "less is better than more" to achieve the effect of reducing the requirements for computing resources and creating prerequisites for solving the problem of overfitting;
4) Heuristic configuration of the hyperparameters of the model with the design of accuracy metrics, optimization algorithms, loss functions and the number of training epochs;
5) Analysis of the probabilistically conditioned learning process of the model by monitoring the dependence of the mathematical expectation and the standard deviation of the classification efficiency measure on the learning epoch;
6) Assessing the quality of the model based on experimentally built error matrices and calculating accuracy and error metrics to conclude that the result meets the requirements for the first stage. The proposed chain of actions leads to obtaining an instance of the model. The entire process of searching for the most effective result is formalized in the form of a tree, the terminal nodes of which correspond to the generated spatial data analysis models.

A systematic analysis of territorial descriptors gives a significant increase in the accuracy of metageosystem classification (Table 1): taking into account the descriptors

Table 1. Comparison of classification accuracy and training time of models within the experiment

Model and relative size of the training sample	Relative classification accuracy, %											Training time, c	
	Annual crops	Forests and parks	Herbaceous vegetation	Highways and roads	Industrial buildings	Pastures	Perennial crops	Residential development	Rivers	Lakes and reservoirs	Accuracy	GPU	на CPU
CNN	88.7	97.0	77.7	65.2	91.2	89.5	76.8	98.3	76.8	99.0	88.5	3.6×102	8.3×103
CNN, with a lack of data	78.9	93.2	65.5	44.0	86.7	86.2	67.5	91.7	59.9	95.2	77.5	1.8×102	4.8×103
SNN (D_{ERS})	81.6	94.5	63.9	39.7	90.7	67.0	64.4	94.2	58.6	94.0	76.1	2.7×102	1.6×102
SNN (D_{ERS} + D_{DEM})	84.4	94.8	80.5	42.0	90.8	73.1	67.7	93.9	58.9	96.5	79.4	2.7×102	1.6×102
SNN (D_{ERS} + D_{LK})	90.2	83.1	83.1	75.5	94.0	80.6	81.2	96.6	74.5	97.1	87.6	2.7×102	1.6×102
SNN (D_{ERS} + D_{DEM} + D_{LM})	90.5	96.7	88.7	77.6	93.8	86.5	83.6	96.9	77.8	97.8	89.2	2.7×102	1.6×102

calculated based on satellite imagery data of the territory (DERS group) made it possible to achieve an accuracy of 76%. Involvement of terrain descriptors (DDEM group) increases the accuracy by 3%, and metrics calculated based on landscape maps (DLM group) by 11%. Simultaneous consideration of descriptors of all categories leads to an increase in classification accuracy by almost 12%.

Thus, when solving specific problems, low-capacity shallow neural network models (SNN), provided that territorial descriptors are taken into account, show accuracy characteristics comparable to those of convolutional models trained on multidimensional spatial data.

It should also be noted that using deeper neural network architectures will improve classification accuracy due to the possibility of extracting complex hierarchical features. However, this will increase the requirements for the hardware on which the calculations are carried out, or it will take a lot of time to conduct experimental studies [21, 22].

Thus, with all the advantages, using deep convolutional neural network models leads to contradictions that need to be resolved. First, their sustainable training requires expert labeling of important training data, which is time-consuming and resource-consuming [23, 24]. Secondly, deep convolutional models are very demanding on computational resources, which are not always available [25, 26].

The combination of neural networks into an ensemble according to the described method, followed by the calculation of the F1 metric and other performance parameters, showed that an increase in the capacity of the models does not lead to a clear improvement in the result, since more powerful models may be more unstable to overfitting, and also require more labeled data for training. When the models were combined into an ensemble, the resulting hypothesis began to be applied based on weighted voting based on the measure of efficiency, which made it possible to avoid gross errors in the classification inherent in each classifier separately. At the same time, the ensemble only slightly loses accuracy to individual classifiers of the system while maintaining overall resistance to errors when determining objects of a particular class of territory.

While a classification system trained from satellite imagery fragments does not allow accurate pixel-by-pixel segmentation, it does provide the ability to automate land cover analysis and land-use systems (including change detection). It can contribute to the updating of digital geographic maps. Under the condition of fine-tuning the algorithm for calculating territorial descriptors and training the classifier based on remote monitoring data of a specific satellite scanning the earth's surface for a long time, the generated models can be used to solve the problem of automated monitoring of changes in the structure of land use and the geophysical envelope (including those caused by technogenic transformation), exogeodynamic processes, deforestation processes, as well as natural processes: fires and floods). To classify land and spatial data on the study area (RS data, digital elevation models, digital maps) on the study area, it is necessary to impose a spatial grid with a step Δ_{grid}, which determines the detail and degree of generalization of the result

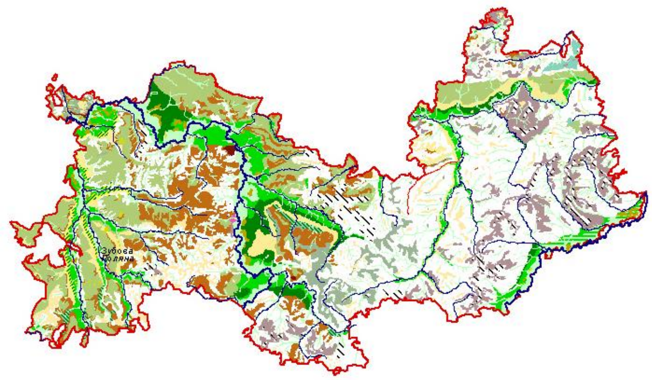
The conducted discretization compares the analyzed territory with the matrix:

$$MAP = (cell_{i,j})_{i=1,j=1}^{m,n} = (cell_{i,j})_{i=1,j=1}^{\frac{height}{\Delta_{grid}},\frac{width}{\Delta_{grid}}}, \qquad (2)$$

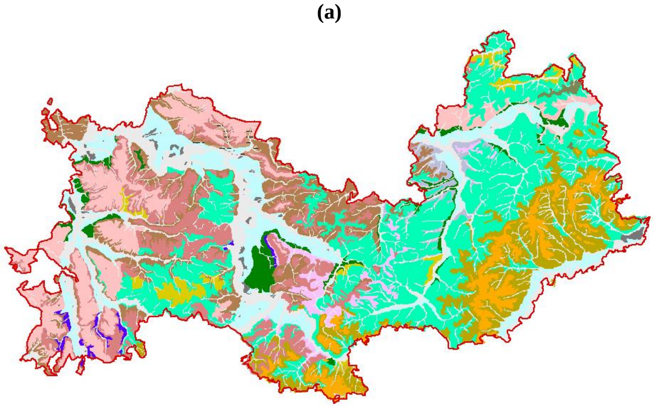
where MAP – test polygon matrix; m, n – number of rows and columns of the matrix MAP; $cell_{i,j}$ – cell of the MAP matrix located at the intersection of the i-th row and j-th column and containing data on a fragment of the territory; $height, width$ – test polygon height and width; Δ_{grid} – grid step size.

Each cell $cell_{i,j}$ of the MAP matrix can be associated with a set of territorial descriptors, the characteristics and calculation algorithm of which are presented in this section of the article. The classification system under consideration can be used for automated validation of digital maps of significant territorial coverage based on the search and visualization of territories that require manual expert verification for the relevance of the displayed information and to identify the need for detailed segmentation.

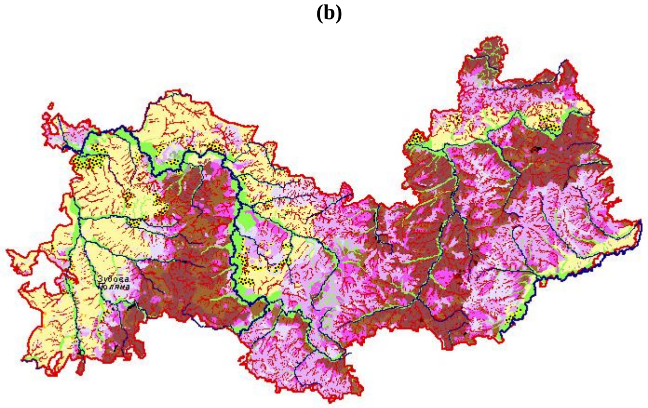
Figure 2 shows images that characterize the thematic layers of the spatial database. Solving the problem of choosing informative territorial features is essential for analyzing the state of metageosystems from the standpoint of identifying and predicting specific properties of spatial objects and requires the selection of an optimal set of thematic layers of digital maps.



(a)



(b)



(c)

Fig. 2 Digital maps of regional GIS: a) Quaternary deposits, b) aquifers, c) spatial distribution of soils

The algorithm for selecting thematic layers for analysis should make it possible to effectively eliminate redundant features and select subsets of parameters most suitable for building data analysis models when solving design problems aimed at functional zoning of territories. The efficient design of sets of factors makes it possible to reduce the complexity of the analyzed data system, increase the stability of algorithms, and reduce the time of information processing.

$$w_i = P\big((f_i^{\mathrm{Miss}} \neq f_i)|\mathrm{Miss}\big) - P((f_i^{\mathrm{Hit}} \neq f_i)|\mathrm{Hit})$$

In the case of solving a binary classification problem, within the framework of the Relief algorithm, an instance of the object x is selected at each iteration. Then two nearest neighbors are found, one from the same class (the closest hit Hit) and the other from a different class (the closest miss Miss). The search for nearest neighbors, characterized by a set of N features, can be carried out based on the calculation of the Euclidean distance for the values of features of objects, followed by taking instances for which the maximum and minimum values of this distance are characteristic:

$$\mathrm{d}(x_p, x_q) = \sqrt{\sum_{f=1}^{N}(x_p^f - x_q^f)^2},$$

The search for nearest neighbors can be based on other principles, including calculating the Manhattan (L1) distance or the Chebyshev distance. After this stage, it is possible to update the weights of each i-th factor by the value Δ_i:

$$w_i^{iter+1} = w_i^{iter} + \Delta_i$$

The modification parameter Δ_i can be increased as the difference in the value of the attribute for objects of different classes increases and decreases by the difference in the value of the attribute for objects of the same class:

$$\Delta_i = \Delta_i^{miss} - \Delta_i^{hit} = \left|x^i - \mathrm{Miss}^i(x)\right| - \left|x^i - \mathrm{Hit}^i(x)\right|$$

In the event that categorical values characterize the i-th factor, the weight coefficients can be modified by the following principle:

$$\Delta_i = \Delta_i^{miss} - \Delta_i^{hit} = \Big(\Big(x^i = \mathrm{Miss}^i(x)\Big)\,?\,0:1\Big) - \Big(\Big(x^i = \mathrm{Hit}^i(x)\Big)\,?\,0:1\Big)$$

The indicated weight transformation technique can be extended for the case of multiclass analysis in which the task is to distribute objects within the class space $Y = \{1, \dots, C\}$, which characterizes by the probability distribution p_c, which determines the frequency of occurrence of objects of each class. At the same time, based on the Euclidean distance, the nearest neighbors $\mathrm{Miss}_c^i(x)$ and $\mathrm{Hit}_c^i(x)$ are searched from the point of view of belonging to the object class C.

$$\Delta_i = \sum\nolimits_{c \in Y, c \neq y(x)} \frac{p_c}{1 - p_c} \left(\left| x^i - \mathrm{Miss}_c^i(x) \right| - \left| x^i - \mathrm{Hit}_c^i(x) \right| \right),$$

In regression problems, the predicted value of y_{pred} is continuous, which makes it impossible to select the nearest neighboring objects (Near) belonging to different classes. In this context, it is advisable to introduce a criterion for distinguishing the predicted continuous values of two instances. This value can be modeled based on calculating the relative distance between the predicted values of the two instances.

Based on the Bayes theorem, the calculation of weight coefficients can be represented as:

$$w_i = \frac{P_{\mathrm{Near}}^{y|f_i} P_{\mathrm{Near}}^{f_i}}{P_{\mathrm{Near}}^{y}} - \frac{(1 - P_{\mathrm{Near}}^{y|f_i}) P_{\mathrm{Near}}^{f_i}}{1 - P_{\mathrm{Near}}^{y}}$$

where $P_{\mathrm{Near}}^{y|f_i}$ is the probability of a different predicted value for neighboring objects with different values of the i-th factor; $P_{\mathrm{Near}}^{f_i}$ – the probability of different values of the i-th factor for neighboring objects; P_{Near}^{y} is the probability of a different predicted value for neighboring features. The presented probabilistic parameters can be estimated based on a series of statistical measurements, and the estimate of the distance between objects can be determined as follows:

$$\mathrm{d}(x_p, x_q) = \frac{\mathrm{f}(y_p, y_q, \alpha)}{\sum_{i=1}^{N} \mathrm{f}(y_p, y_i, \alpha)}$$

where f(·) is a function for determining the distance of the discrete distance between the predicted values based on the scaling parameter α and a set of linear or non-linear transformations; N is an adjustable coefficient that determines the power of the analyzed objects. Before performing the stage of estimating the weights of the analyzed factors, their values should be normalized to solve the problem of excessive data scatter. This problem can be approached based on Z-normalization, which involves the transformation of the factor values based on the calculation of the mathematical expectation $\bar{x}$ and the variance σ_x.

$$\widetilde{x_\iota} = \frac{x_i - \bar{x}}{\sigma_x}$$

It is possible to develop the presented algorithm and ensure this stability when working with unbalanced and noisy data by adding iteration. The correctness of this provision is confirmed by the fact that the algorithm's performance for calculating the weight coefficients deteriorates as the number of irrelevant features becomes significant.

The recursive approach to the successive elimination of irrelevant features allows each iteration to cut off the territorial features with the lowest estimates from consideration so that they do not participate in searching for nearest neighbors and updating the weight coefficients. It is important to note that the choice of the number of iterations and the number of features discarded at each stage is a parameterizable side of the iterative algorithm. At the same time, these characteristics may not be strictly specified: at each stage, it is possible to cut off any number of features with certain statistical characteristics.

On the other hand, the algorithm for estimating the importance of spatial features, having a quadratic time complexity, may not be optimal in terms of performance when working with high-dimensional training datasets. The principle that is the key to optimization when working with large datasets is that evaluating the importance of features is faster within a smaller data dimension. The modified algorithm then applies the assessment of the importance of territorial features to a certain number of random or expertly determined subsets of features. After that, the partial results can be integrated by setting the global feature weight to equal the maximum local weight of that feature across all analyzed subsets. The effectiveness of the presented approach depends to a large extent on the representativeness of correlated features within the selected subsets.

Thus, the presented algorithm makes it possible to assess the influence of data from specific thematic layers of a digital map on the target indicator (discrete or continuous) to solve the problem of determining the optimal set of spatially associated information used in decision-making. Improving the algorithm's stability when working with unbalanced and noisy data is possible based on a recursive approach, which involves the sequential application of the algorithm for assessing the importance of data by cutting off a reasonable number of layers at each stage of the algorithm execution. The solution to the problem of increasing the algorithm's speed when working with big data is possible based on assessing the importance of territorial features for subsets of thematic layers, followed by integrating the results into a single set.

4. Conclusion

The solution to the scientific problem of quantitative analysis of intercomponent links in metageosystems of different hierarchical levels is possible based on simulation modeling. The article formulates a set of requirements for the framework of simulation models of spatial processes. It presents an algorithm for developing a model that describes the spatio-temporal processes occurring in territorial systems.

The implemented software package for simulating the traffic flows of urban metageosystems can be used to create dynamic transport models with the ability to predict the nature of the traffic flow depending on the current traffic situation. It integrates the components of the system into a single model, which subsequently serves as a tool for making strategic decisions regarding the development of the transport structure of a city or region. Provided that the problems of analyzing space-time systems are formalized within the framework of the presented model, the proposed methodology can be used to analyze other natural and natural-technogenic processes, including natural phenomena.

The developed system consists of eight modules: a map editor, a component for working with a database, a subsystem for visualizing interactive graphical web interfaces, an auxiliary module for performing the necessary mathematical calculations, a subsystem for modeling urban traffic, and a recommender subsystem for generating recommendations to improve the structure of the road network. When interacting with each other, the system modules ensure the reliable and uninterrupted operation of the simulation system as a whole.

Two development directions of the framework for simulation modeling spatial processes should be singled out. On the one hand, it is necessary to adhere to a deductive strategy. The model building system should be initially designed to achieve the possibility of covering solutions to the maximum number of problems. On the other hand, it is necessary to implement the inductive strategy as efficiently as possible, within which the positively proven improvements of particular specific models become the basis for developing the simulation framework. The combined use of deductive and inductive strategies will provide an evolutionary improvement in the modeling framework and optimization of particular solutions developed on its basis.

The hierarchical structuring of geosystems optimizes the diagnostics of the leading factors of the interaction of physical and geographical factors, the regularities of the spatio-temporal change of their states, the direction of the development of metabolic processes and the transformation of matter and energy. The general scheme for compiling a digital map of metageosystems in a regional GIS is implemented based on solving the problems of integrating spatial data and their systematization based on the construction of a hierarchy of geosystems and the project-oriented use of automated computing.

The cumulative analysis of territory descriptors integrated based on data from different sources significantly increases the accuracy of metageosystem classification. In the framework of the experiment presented in the article, taking into account the proposed system of descriptors calculated based on satellite imagery data, a digital elevation model and an electronic landscape map, made it possible to achieve an accuracy of 89%, which is much more than this parameter for a convolutional neural network model. At the same time, the analysis of terrain descriptors increases the accuracy by 3% and the metrics calculated based on landscape maps - by 11%. Specialists will interpret the cartograms of the presented descriptors in the field of data analysis in geosciences.

The result of the operation of the algorithm for the automated selection of thematic layers allows the formation of reasonable structured models of geosystems that reflect the strength and nature of intercomponent connections, will enable you to determine the factors that describe the territorial variation of properties, interpret and justify their physical meaning.

Funding Statement

The study was supported by the Russian Science Foundation (grant № 22-27-00651), https://rscf.ru/en/project/22-27-00651/.

References

[1] O. Trofymchuk, V. Okhariev, and V. Trysnyuk, “Environmental Security Management of Geosystems,” *In 18th International Conference on Geoinformatics-Theoretical and Applied Aspects,* vol. 2019, no. 1, pp. 1-5, 2020. *Crossref,* https://doi.org/10.3997/2214-4609.201902083

[2] Jacinta Holloway, and Kerrie Mengersen, “Statistical Machine Learning Methods and Remote Sensing for Sustainable Development Goals: A Review,” *Remote Sensing*, vol. 10, no. 9, pp. 1365, 2018. *Crossref,* https://doi.org/10.3390/rs10091365

[3] Miao Li et al., “A Review of Remote Sensing Image Classification Techniques: The Role of Spatio-Contextual Information,” *European Journal of Remote Sensing*, vol. 47, no. 1, pp. 389-411, 2014. *Crossref,* https://doi.org/10.5721/EuJRS20144723

[4] Supakarn Prajam, Chitapong Wechtaisong, and Arfat Ahmad Khan, “Applying Machine Learning Approaches for Network Traffic Forecasting,” *Indian Journal of Computer Science and Engineering*, vol. 13, no. 2, pp. 324-335, 2022. *Crossref,* https://doi.org/10.21817/indjcse/2022/v13i2/221302188

[5] Ying Li et al., “Deep Learning for Remote Sensing Image Classification: A Survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 6, p. e1264, 2018. *Crossref,* https://doi.org/10.1002/widm.1264

[6] Qiangqiang Yuan et al., “Deep Learning in Environmental Remote Sensing: Achievements and Challenges,” *Remote Sensing of Environment,* vol. 241, p. 111716, 2020. *Crossref,* https://doi.org/10.1016/j.rse.2020.111716

[7] Israr Ullah et al., “An Integrated Approach of Machine Learning, Remote Sensing, and GIS Data for the Landslide Susceptibility Mapping,” *Land,* vol. 11, no. 8, p. 1265, 2022. *Crossref,* https://doi.org/10.3390/land11081265

[8] M. Weiss, F. Jacob, and G. Duveiller, “Remote Sensing for Agricultural Applications: A Meta-Review,” *Remote Sensing of Environment,* vol. 236, p. 111402, 2020. *Crossref,* https://doi.org/10.1016/j.rse.2019.111402

[9] Aaron E. Maxwell, Timothy A. Warner, and Fang Fang, “Implementation of Machine-Learning Classification in Remote Sensing: An Applied Review,” *International Journal of Remote Sensing,* vol. 39, no. 9, pp. 2784-2817, 2018. *Crossref,* https://doi.org/10.1080/01431161.2018.1433343

[10] Stanislav A. Yamashkin et al., “Improving the Efficiency of Deep Learning Methods in Remote Sensing Data Analysis: Geosystem Approach,” *IEEE Access,* vol. 8, pp. 179516-179529, 2020. *Crossref,* https://doi.org/10.1109/ACCESS.2020.3028030

[11] Ming Zhu, Yongning He, and Qingyu He, “A Review of Researches on Deep Learning in Remote Sensing Application,” *International Journal of Geosciences,* vol. 10, no. 1, pp. 1-11, 2019. *Crossref,* https://doi.org/10.4236/ijg.2019.101001

[12] Miaojia Lu et al., “Multiagent Spatial Simulation of Autonomous Taxis for Urban Commute: Travel Economics and Environmental Impacts,” *Journal of Urban Planning and Development,* vol. 144, no. 4, p. 04018033, 2018. *Crossref,* https://doi.org/10.1061/(ASCE)UP.1943-5444.0000469

[13] Mario Valiante et al., “A Spatiotemporal Object-Oriented Data Model for Landslides (LOOM),” *Landslides,* vol. 18, no. 4, pp. 1231-1244, 2021. *Crossref,* https://doi.org/10.1007/s10346-020-01591-4

[14] Andrew Crooks, Christian Castle, and Michael Batty, “Key Challenges in Agent-Based Modelling for Geo-Spatial Simulation,” *Computers, Environment and Urban Systems,* vol. 32, no. 6, pp. 417-430, 2008. *Crossref,* https://doi.org/10.1016/j.compenvurbsys.2008.09.004

[15] Jingwei Hou et al., “Spatial Simulation of the Ecological Processes of Stormwater for Sponge Cities,” *Journal of Environmental Management,* vol. 232, pp. 574-583, 2019. *Crossref,* https://doi.org/10.1016/j.jenvman.2018.11.111

[16] Teja Kattenborn et al., “Review on Convolutional Neural Networks (CNN) in Vegetation Remote Sensing,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 24-49, 2021. *Crossref,* https://doi.org/10.1016/j.isprsjprs.2020.12.010

[17] Jia Song et al., “A Survey of Remote Sensing Image Classification Based on CNNs,” *Big Earth Data,* vol. 3, no. 3, pp. 232-254, 2019. *Crossref,* https://doi.org/10.1080/20964471.2019.1657720

[18] Zhou Huang et al., “An Ensemble Learning Approach for Urban Land Use Mapping Based on Remote Sensing Imagery and Social Sensing Data,” *Remote Sensing*, vol. 12, no. 19, p. 3254, 2020. *Crossref,* https://doi.org/10.3390/rs12193254

[19] Addis Seid, and T. Suryanarayana, “Identification of Lithology and Structures in Serdo, Afar, Ethiopia Using Remote Sensing and GIS Techniques,” *SSRG International Journal of Geoinformatics and Geological Science,* vol. 8, no. 1, pp. 27-41, 2021. *Crossref,* https://doi.org/10.14445/23939206/IJGGS-V8I1P104

[20] Bahareh Kalantar et al., “Landslide Susceptibility Mapping: Machine and Ensemble Learning Based on Remote Sensing Big Data,” *Remote Sensing*, vol. 12, no. 11, p. 1737, 2020. *Crossref,* https://doi.org/10.3390/rs12111737

[21] M. F. Goodchild, “Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0,” *International Journal of Spatial Data Infrastructures Research*, vol. 2, no. 2, pp. 24-32, 2007.

[22] Xiao Xiang Zhu et al., “Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8-36, 2017. *Crossref,* https://doi.org/10.1109/MGRS.2017.2762307

[23] Liangpei Zhang, Lefei Zhang, and Bo Du, “Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22-40, 2016. *Crossref,* https://doi.org/10.1109/MGRS.2016.2540798

[24] Chao Tao et al., “Unsupervised Spectral–Spatial Feature Learning with Stacked Sparse Autoencoder for Hyperspectral Imagery Classification,” *IEEE Geoscience and Remote Sensing Letters,* vol. 12, no. 12, pp. 2438-2442, 2015. *Crossref,* https://doi.org/10.1109/LGRS.2015.2482520

[25] Yann Le Cun, Yoshua Bengio, and Geoffrey Hinton, “Deep Learning,” *Nature*, vol. 521, no. 7553, pp. 436-444, 2015. *Crossref,* https://doi.org/10.1038/nature14539

[26] László Miklós et al., “*Landscape as a Geosystem*,” Cham, Switzerland: Springer, pp. 11-42, 2019. *Crossref,* https://doi.org/10.1007/978-3-319-94024-3

[27] Marko Robnik-Šikonja, and Igor Kononenko, “Theoretical and Empirical Analysis of Relieff and Rrelieff,” *Machine learning*, vol. 53, no. 1, pp. 23-69, 2003. *Crossref,* https://doi.org/10.1023/A:1025667309714

[28] RanWanga, and John A.Gamon, “Remote Sensing of Terrestrial Plant Biodiversity,” *Remote Sensing of Environment,* vol. 231, p. 111218, 2019. *Crossref,* https://doi.org/10.1016/j.rse.2019.111218