Land and Architecture. 2025; 4:183
doi: 10.56294/la2025183

ORIGINAL

Big Data Processing Methods in GIS

Métodos de procesamiento de big data en SIG

Irada Seyidova[1] ✉, Elgun Gamzaev[1]

[1]Azerbaijan University of Oil and Industry, Department of Computer Engineering. Azerbaijan.

Cite as: Seyidova I, Gamzaev E. Big Data Processing Methods in GIS. Land and Architecture. 2025; 4:183. https://doi.org/10.56294/la2025183

Submitted: 20-05-2024 **Revised:** 22-10-2024 **Accepted:** 14-05-2025 **Published:** 15-05-2025

Editor: Prof. Dr. William Castillo-González

Corresponding author: Irada Seyidova ✉

ABSTRACT

The article discusses methods for processing big data in geographic information systems (GIS) with an emphasis on the use of recurrent neural networks (RNN) for forecasting geospatial processes. Modern approaches are described, including distributed computing on clusters (Hadoop, Spark) and cloud platforms (Google Earth Engine), providing efficient processing of spatial data. Particular attention is paid to RNN architectures, such as LSTM, their application in temporal forecasting problems (weather, transport, land use) and comparison with traditional methods. The article provides a numerical example illustrating the use of RNN for time series forecasting, with an accuracy analysis and visualization of the results.

Keywords: Big Data; Geographic Information Systems; Recurrent Neural Networks; Forecasting; Machine Learning.

RESUMEN

En este artículo se examinan métodos para procesar macrodatos en sistemas de información geográfica (SIG), haciendo hincapié en el uso de redes neuronales recurrentes (RNN) para predecir procesos geoespaciales. Se describen enfoques modernos, incluida la computación distribuida en clústeres (Hadoop, Spark) y plataformas en la nube (Google Earth Engine), que proporcionan un procesamiento eficiente de los datos espaciales. Se presta especial atención a las arquitecturas RNN, como LSTM, su aplicación en problemas de previsión temporal (meteorología, transporte, uso del suelo) y su comparación con los métodos tradicionales. El artículo proporciona un ejemplo numérico que ilustra el uso de RNN para la previsión de series temporales, con un análisis de precisión y visualización de los resultados.

Palabras clave: Big Data, Sistemas de Información Geográfica, Redes Neuronales Recurrentes, Predicción, Aprendizaje Automático.

INTRODUCTION

Modern geographic information systems (GIS) are faced with an unprecedented growth in data volumes. Spatial "big data" comes from satellite images, Internet of Things sensors, crowdsourced geodata, etc. These data are characterized by the classic "4Vs" of big data - large volumes, diversity, high velocity, and potential unreliability. Traditional methods of processing geodata often fail to cope with such volumes: calculations become too slow, and conventional desktop software cannot effectively use all available hardware resources. Thus, working with spatial big data requires special approaches, including distributed computing and cloud technologies.[1,2,3]

One of the key approaches to processing large geodata is the use of distributed computing. The idea is to parallelize data fragments on a cluster of several nodes, which speeds up computations proportionally to the number of nodes. Companies that stood at the origins of big data, such as Google and Facebook, popularized the MapReduce paradigm and related technologies for distributed processing of large data arrays. In geoinformatics, extensions of this paradigm were developed for the specifics of spatial data. An example is the SpatialHadoop platform, which built support for spatial indices (R-trees, grids) into Hadoop, allowing MapReduce tasks to be performed directly on geodata.[4,5,6]

Another direction is the use of Apache Spark, which provides distributed computing in RAM (in-memory). Spark has demonstrated high efficiency in geoanalytics tasks compared to Hadoop due to fast in-memory computing and a rich library of algorithms. For example, the Esri ecosystem has created a GeoAnalytics Engine module for running spatial analyses on Spark clusters.[7]

However, the application of distributed systems to GIS data is associated with difficulties. Spatial data is more complex than standard string or tabular data, which are operated by classic big data systems.[6] It is necessary to take into account the spatial locality of data - objects close to each other on the map should preferably be processed on one node in order to reduce the overhead of data transfer over the network. Spatial analysis algorithms (for example, buffer calculation, overlay analysis) often require data exchange between nodes, which does not fit well into the MapReduce model. Nevertheless, the development of parallel GIS algorithms continues. Researchers note that new approaches are needed for efficient parallel processing of geodata, going beyond the classic MapReduce.[8,9]

Another revolutionary trend has been the use of cloud technologies for storing and processing geospatial big data. Cloud GIS allows you to offload heavy calculations and storage of redundant petabytes of data to remote servers, providing users with web interfaces or APIs to access the analysis. A striking example is the Google Earth Engine (GEE) platform, officially launched in 2010. GEE stores a multi-petabyte catalog of satellite images and geospatial datasets in the cloud, and also provides tools for their planetary scaling analysis. The user can write a relatively small script, and the platform will execute it in parallel on thousands of Google nodes, automatically applying operations to all the necessary image tiles or territorial units . This eliminates the need to download huge amounts of raw data to a local computer; the analysis "fits" the data where it is stored. Cloud platforms also provide ready-made data from open sources (e.g., Landsat, Sentinel, MODIS collections), which simplifies pre-processing.[4,11,12,13]

Similar capabilities are offered by other cloud services: Amazon Web Services (AWS SageMaker and EMR service for Spark/Hadoop clusters), Microsoft Azure (Azure Synapse, Azure Maps), as well as specialized geo-platforms such as Microsoft's Planetary Computer or Sentinel Hub.

Commercial GIS vendors also integrate cloud solutions. For example, ArcGIS GeoAnalytics Server from Esri is part of ArcGIS Enterprise and is designed specifically for distributed processing of big geodata. It includes a cluster of servers that perform spatial analysis (aggregation, clustering, anomaly detection, regression analysis) on large volumes of data, using an architecture similar to Hadoop/Spark. GeoAnalytics Server can work with both traditional formats (shapefiles, CSV) and directly connect to big data storage (HDFS, cloud storage), which simplifies integration with existing infrastructure.[1,14,15,16,17]

METHOD

Problem setting

Machine learning methods, and especially deep neural networks, are increasingly being applied to geodata analysis tasks (a field known as GeoAI). For time-dependent data, recurrent neural networks are the most popular. A recurrent neural network (RNN) is a deep network architecture adapted to sequential information, such as time series. In classical fully connected networks, inputs are considered independent, while an RNN operates on an ordered sequence of inputs, with cyclic connections, thanks to which the "memory" of previous steps influences subsequent calculations.

This allows RNN to capture temporal dependencies: trends, seasonal fluctuations, periodic patterns, delays in the influence of some factors on others. Basic RNN suffers from the gradient decay problem, which makes it difficult to remember long-term dependencies (events that are far in time). To solve this problem, improved architectures were developed, primarily Long Short-Term Memory (LSTM) and the closely related GRU. LSTM implements special mechanisms (the so-called GATEs - input, output, and forgetting) allow you to store information in your memory for as long as you like, "forgetting" unimportant information if necessary. Thanks to these improvements, thanks to these improvements, LSTM has become the de facto standard for time series forecasting tasks. In the geospatial field, RNNs have found application in a variety of tasks. For example, in meteorology and climatology, recurrent networks are used to forecast weather parameters over time.[18,19,20,21]

RNNs (including LSTMs) successfully model dependencies in sequences of temperature, precipitation, pressure, etc. observations, which complements traditional numerical atmospheric models. One study proposed a modification of RNNs (the so-called SwiftRNN) for forecasting air visibility, which surpassed in accuracy even complex ConvLSTM models that specifically take into account spatial relationships. Another area is forecasting anthropogenic processes: RNNs are used to forecast road traffic and transport network congestion, predict demand for taxis or car sharing in different areas of the city based on hourly data. In environmental problems, RNNs were used to assess the dynamics of water quality indicators, monitor changes in groundwater levels, and predict river floods. RNNs have also proven effective in forecasting changes in land use and vegetation cover. Van Duynhoven et al. (2021) showed that LSTMs are able to predict changes in land cover classes over time with high accuracy, outperforming classical algorithms such as random forest.(10,22,23,24,25,26)

RESULTS

Problem solving

It is important to note that purely temporal RNN models take into account the sequence of observations over time, but do not always directly account for spatial relationships (e.g., the proximity of measuring stations or pixels in an image). Therefore, in complex geospatial forecasting problems, RNNs are often combined with other neural network components. There are convolutional RNNs (ConvLSTM), which take into account the two-dimensional structure of the input data (e.g., a sequence of raster precipitation fields) and are able to simultaneously track the temporal dynamics and the movement of weather fronts in space. Another approach is graph neural networks (GNNs) combined with RNNs, when the spatial structure (e.g., a road graph or a network of monitoring stations) is modeled by a graph, and a recurrent model for the time series at this point runs at each graph node. However, even basic LSTMs, working with each time series separately or with adjusted features, have already shown high efficiency and have become an important tool for geoforecasting.(3,27,28)

After training, the RNN model generates a forecast for 10 months ahead, which is compared with actual observations. Figure 1 shows a fragment of such a forecast for the test segment. The green line shows the actual change in the indicator over time, and the red line shows the value predicted by the RNN model (the predicted time series). It is clear that the recurrent network was able to correctly capture the general trend of changes and closely repeats the fluctuations of the time series.

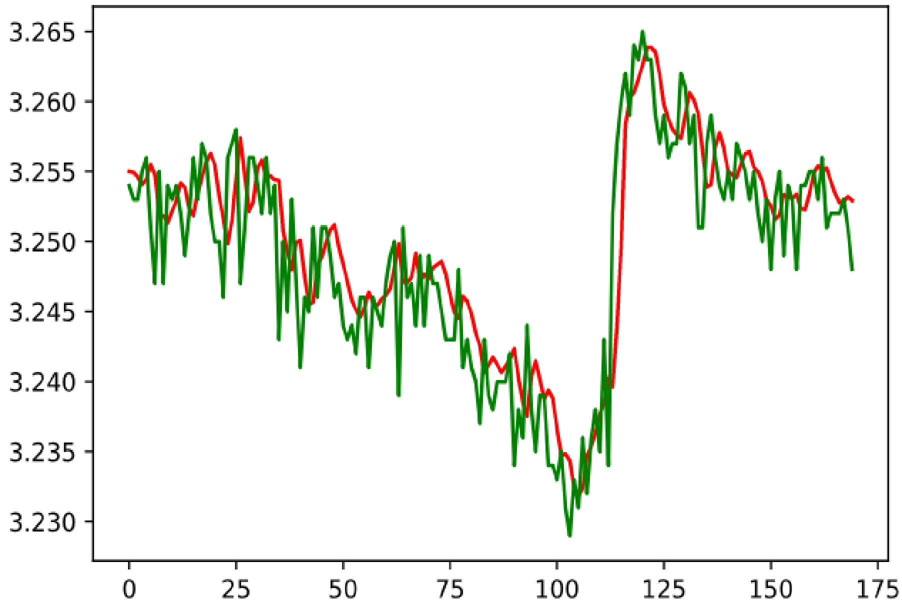

Figure 1. Example of time series forecasting using RNN: actual data (green line) and model forecast (red line) on a test plot

DISCUSSION

The model reproduces the dynamics of the indicator almost exactly, including sharp changes.(5) For example, in the interval of steps 125-135, there is a sharp jump in the value upward, and the model successfully predicts this growth (the forecast curve almost coincides with the real one) - this means that the RNN has recognized the pattern preceding the jump. In quantitative terms, the accuracy of the forecast is high: the root-mean-square error is about 0,003 (in относительном выражении менее 0,1 % from the average level of the series), and the correlation coefficient between the forecast and the fact on the test segment exceeds 0,98. In other words, the deviations of the forecast from reality are minimal. This example illustrates the potential of recurrent neural networks in geodata forecasting tasks. Classical statistical models (for example, ARIMA) often have difficulty in the presence of complex nonlinear trends or variable seasonality, while RNNs are able to learn them automatically. Of course, achieving such high accuracy may require careful selection of model parameters, a sufficient volume of training data and correct adjustment of the training process (e.g. number of epochs, loss function, optimization algorithm). It is also important to avoid overtraining – control the quality on the test data. In our example, the model showed good quality precisely on the data that was not used in training, which confirms its ability to generalize patterns.(29,30,31,32)

CONCLUSIONS

In conclusion, the integration of big data processing methods with deep learning models is a powerful tandem for geoinformatics. Distributed and cloud computing provide the ability to work with planetary-scale data, and recurrent neural networks allow extracting valuable predictive information from this data.

Future research in this area aims to combine spatial and temporal analysis in single models (e.g., advanced spatio-temporal networks), improve the interpretability of neural network forecasts, and optimize the computational costs of training models on truly large volumes of geospatial data. Solving these challenges will contribute to the creation of smarter and more efficient GIS systems capable of analyzing global processes in real time and supporting data-driven decision making.

REFERENCES

1. ArcGIS GeoAnalytics Server – What is ArcGIS GeoAnalytics Server? (Doc. ArcGIS Enterprise 11.2).

2. Eldawy A., Mokbel M.F. (2015). SpatialHadoop: A MapReduce Framework for Spatial Data. 31st IEEE International Conference on Data Engineering.

3. Fan, J., Bai, J., Li, Z., Ortiz-Bobea, A., & Gomes, C. P. (2022). A GNN-RNN approach for harnessing geospatial and temporal information: Application to crop yield prediction. arXiv. https://arxiv.org/abs/2111.08900

4. Google Earth Engine – Overview (2017). Google Developers.

5.https://discuss.pytorch.org/t/time-series-the-prediction-result-of-lstm-is-approximately-straight-line/98896

6. Olasz A., Nguyen Thai B. (2016). Geospatial Big Data processing in an open source distributed computing environment. PeerJ Preprints, 4:e2226v1.

7. Sun, Z., Zhang, H., Liu, Z., Xu, C., & Wang, L. (2016). Migrating GIS big data computing from Hadoop to Spark: An exemplary study using Twitter. 2016 IEEE 9th International Conference on Cloud Computing (CLOUD). IEEE. https://ieeexplore.ieee.org/document/7820291

8. Van Duynhoven A. et al. (2021). Exploring the Sensitivity of RNN Models for Forecasting Land Cover Change. Land, 10(3), 282.

9. Werner M. (2019). Parallel Processing Strategies for Big Geospatial Data. Frontiers in Big Data, 2:44.

10. Zang Z. et al. (2023). A Modified RNN-Based Deep Learning Method for Prediction of Atmospheric Visibility. Remote Sensing, 15(3), 553.

11. Chen Y. Smart Urban and Rural Planning Decision Support System Based on GIS and Geographic Information Big Data. 2024 International Conference on Power, Electrical Engineering, Electronics and Control (PEEEC), 2024, p. 678-82. https://doi.org/10.1109/PEEEC63877.2024.00128.

12. Zhang L, He W, Guo Y, Teng X. A Smart Application Frame of Remote Sensing in Non-grain Production Data Governance. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 2024;XLVIII-1-2024:849-54. https://doi.org/10.5194/isprs-archives-XLVIII-1-2024-849-2024.

13. Yang J, Fricker P, Jung A. From intangible to tangible : The role of big data and machine learning in walkability studies. Computers, Environment and Urban Systems 2024;109:1-20. https://doi.org/10.1016/j.compenvurbsys.2024.102087.

14. Wong ATT. Opportunities and Challenges of Big Data Analytics in Crime Investigation. International Annals of Criminology 2025:1-15. https://doi.org/10.1017/cri.2025.3.

15. Ju C, Huang H. Rural Ecological Environment Monitoring and VR Visualization Analysis of Jilin Province Supported By Big Data. Procedia Computer Science 2024;243:558-66. https://doi.org/10.1016/j.procs.2024.09.068.

16. Chang L, Zhi Y, Binbin Z, Sihang Z. Wildfire risk assessment using multi-source remote sensing data and GIS. 2024 IEEE 4th International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), vol. 4, 2024, p. 420-7. https://doi.org/10.1109/ICIBA62489.2024.10868308.

17. Rajeev A, Shah R, Shah P, Shah M, Nanavaty R. The Potential of Big Data and Machine Learning for Ground Water Quality Assessment and Prediction. Arch Computat Methods Eng 2025;32:927-41. https://doi.org/10.1007/s11831-024-10156-w.

18. Wu J, Gan W, Chao H-C, Yu PS. Geospatial Big Data: Survey and Challenges. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2024;17:17007-20. https://doi.org/10.1109/JSTARS.2024.3438376.

19. Yang S. Strengthening Accounting Information Systems with Advanced Big Data Mining Algorithms: Innovative Exploration of Data Cleaning and Conversion Automation. Informatica 2025;49. https://doi.org/10.31449/inf.v49i11.7302.

20. Xiong R, Song F, Fan Y. Strategies Research on Comprehensive Energy Service Based on Electric Power Big Data. En: Hung JC, Yen N, Chang J-W, editores. Frontier Computing: Vol 3, Singapore: Springer Nature; 2025, p. 347-55. https://doi.org/10.1007/978-981-96-2798-1_38.

21. Djokić V, Djordjević A, Milovanović A. Big data and urban form: a systematic review. J Big Data 2025;12:17. https://doi.org/10.1186/s40537-025-01084-y.

22. Genale AS. Big Data Analytics for Geospatial Application Using Python. Ethics, Machine Learning, and Python in Geospatial Analysis, IGI Global Scientific Publishing; 2024, p. 254-78. https://doi.org/10.4018/979-8-3693-6381-2.ch011.

23. Liu D, Qian X, Yang H. The Application of Big Data Technology in Monitoring and Analyzing the Operation of Economic Policies. En: Gupta R, Bartolucci F, Katsikis VN, Patnaik S, editores. Recent Advancements in Computational Finance and Business Analytics, Cham: Springer Nature Switzerland; 2024, p. 472-82. https://doi.org/10.1007/978-3-031-70598-4_43.

24. Singh S, Reddy KS, Bhowmick MK, Srivastava AK, Kumar S, Peramaiyan P. Accelerating Climate Adaptation with Big Data Analytics and ICTs. En: Pathak H, Lakra WS, Gopalakrishnan A, Bansal KC, editores. Advances in Agri-Food Systems: Volume I, Singapore: Springer Nature; 2025, p. 179-96. https://doi.org/10.1007/978-981-96-0759-4_10.

25. Zheng N, Chen W. The Environmental Status Assessment Modelof Artificial Wetlands Basedon Big Data Technology. Pol J Environ Stud 2025. https://doi.org/10.15244/pjoes/202575.

26. Li L, Jia L. Complex Event Information Mining and Processing for Massive Aerospace Big Data. Scalable Computing: Practice and Experience 2024;25:2540-7. https://doi.org/10.12694/scpe.v25i4.2832.

27. Xue S. Building Material Defect Detection and Diagnosis Method Based on Big Data and Deep Learning. Informatica 2024;48. https://doi.org/10.31449/inf.v48i16.6433.

28. Selmy SAH, Kucher DE, Yang Y, García-Navarro FJ, Selmy SAH, Kucher DE, et al. Geospatial Data: Acquisition, Applications, and Challenges. Exploring Remote Sensing - Methods and Applications, IntechOpen; 2024. https://doi.org/10.5772/intechopen.1006635.

29. Adhikari BK, Mahajan R. Leveraging Big Data Analytics to Enhance Water, Sanitation, and Hygiene (WASH) Systems. Amrit Research Journal 2024;5:98-106. https://doi.org/10.3126/arj.v5i1.73556.

30. Yang L, Ye H. Application and Research of Key Technologies of Big Data for Agriculture. IJISSCM 2024;17:1-20. https://doi.org/10.4018/IJISSCM.344038.

31. Dritsas E, Trigka M. Remote Sensing and Geospatial Analysis in the Big Data Era: A Survey. Remote Sensing 2025;17:550. https://doi.org/10.3390/rs17030550.

32. Zhang J, Lin J, Wu T. An Interval Intuitionistic Fuzzy Characterization Method Based on Heterogeneous Big Data and Its Application in Forest Land Quality Assessment. Int J Fuzzy Syst 2025;27:558-81. https://doi.org/10.1007/s40815-024-01765-5.

FUNDING

The authors did not receive funding for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTION

Conceptualization: Irada Seyidova, Elgun Gamzaev.
Data curation: Irada Seyidova, Elgun Gamzaev.
Formal analysis: Irada Seyidova, Elgun Gamzaev.
Research: Irada Seyidova, Elgun Gamzaev.
Methodology: Irada Seyidova, Elgun Gamzaev.
Project management: Irada Seyidova, Elgun Gamzaev.
Supervision: Irada Seyidova, Elgun Gamzaev.
Validation: Irada Seyidova, Elgun Gamzaev.
Visualization: Irada Seyidova, Elgun Gamzaev.
Writing - original draft: Irada Seyidova, Elgun Gamzaev.
Writing - review and editing: Irada Seyidova, Elgun Gamzaev.