[PD-01-009] Modern Programming Libraries and Infrastructures for Raster Data Analysis

Abstract

This topic entry outlines the recent advances and the transformative integration of artificial intelligence with geospatial science, focusing on modern programming libraries and infrastructures for raster data analysis. We explain how the convergence of AI and geospatial technologies has revolutionized Earth observation analysis through advanced machine learning algorithms, computer vision techniques, and modern infrastructures. The entry highlights the evolution of raster data analysis, from traditional processing methods to sophisticated AI-driven approaches that enable automated feature extraction and object detection at an unprecedented scale. We discuss key developments in programming infrastructure, including Python-based frameworks, GPU-accelerated computing solutions, and cloud-native platforms that have emerged to address the challenges of big Earth data. Special attention is given to dataframe-based analysis approaches and distributed computing frameworks that have enhanced the processing capabilities of large-scale geospatial datasets. We also introduce the cloud computing platforms such as Google Earth Engine, AWS, and Microsoft's Planetary Computer, which have democratized access to Earth observation analysis and empowered researchers to conduct impactful research using geospatial data. By presenting this comprehensive overview, we provide insights into current capabilities and future directions in the field, emphasizing the continuing evolution of GeoAI technologies and their impact on environmental monitoring, urban planning, resource management, etc.

Keywords: cloud computing, GeoAI, machine learning, raster data

Author & citation

Yang, H. L. and Wohlgemuth, J. (2025). Modern Programming Libraries and Infrastructures for Raster Data Analysis. The Geographic Information Science & Technology Body of Knowledge (Issue 1, 2025 Edition), John P. Wilson (ed.). DOI: 10.22224/gistbok/2025.1.20.

Explanation

1. Introduction

The integration of artificial intelligence with geospatial science marks a revolutionary transformation in how we understand and analyze geographic phenomena. Over the past

decade, this convergence has fundamentally reshaped traditional geographic information systems through the incorporation of advanced machine learning algorithms, computer vision techniques, and deep learning methodologies. This integration represents more than a mere technological advancement; it constitutes a new paradigm in geographical information science that enables unprecedented capabilities in spatial analysis and understanding. The emergence of GeoAI has proven particularly timely, coinciding with the exponential growth in Earth observation data from various sensors and platforms, creating new opportunities for environmental monitoring (Chen et al., 2021), urban planning (Rosentreter et al., 2020), and resource management at global scales.

2. The AI Revolution in Geospatial Analysis

The impact of AI on traditional GIS analysis methods has proven profound and multifaceted. Where GIS operations once relied heavily on rule-based processing and manual interpretation, AI-driven approaches for raster data analysis now enable scalable automated feature extraction (Yang et al., 2018), object detection (Zhang et al., 2023), modeling and predictive analytics (Ma et al., 2019) at unprecedented the scale. The emergence of new data-driven methods, particularly Convolutional Neural Networks (CNNs) and other deep learning architectures, have significantly improved semantic segmentation, land use land cover classification and scene classification accuracies (exceeding 90%) as compared to traditional Support Vector Machine or Random Forest (Ma et al., 2021). The transformation extends beyond mere accuracy improvements. Advancements in software and hardware have reduced processing times dramatically from days to hours or minutes for large-scale applications, enabling rapid scientific discovery and operational implementation. This efficiency gain has particular significance in time-critical applications such as disaster response and environmental monitoring, where rapid analysis can directly impact decision-making and outcomes. Deep learning models have demonstrated remarkable capabilities in handling complex spatial patterns and relationships i.e., non-linear systems. These models can now effectively process and analyze multiple data sources simultaneously, learning intricate features and patterns that would be impossible to define through traditional rule-based approaches. Significant advancements have allowed researchers to integrate temporal dimensions into analysis, enabling better understanding of dynamic Earth processes.

3. Evolution of Raster Data in Modern Geospatial Analysis

Raster data serves as the fundamental Earth observational data structure in modern geospatial analysis, where spectral information of ground objects is stored in a matrix form, with each element, known as a pixel, containing specific values representing spectral characteristics of geographic phenomena. The complexity and variety of raster data have expanded significantly with advances in sensing technology. More advanced and various sensors (satellite constellations, or UAV-based platforms) now drive diverse remote sensing data products, including several key categories: 1) Optical imagery (e.g. Landsat, Sentinel-2) with multiple spectral bands ranging from visible to infrared wavelengths, 2) Synthetic Aperture Radar (SAR) data (such as Sentinel-1) that provides all-weather monitoring capabilities, 3) hyperspectral imagery capturing hundreds of narrow spectral bands for detailed surface material analysis and 4) other derived raster data products such as Digital Elevation Models (DEMs) derived from lidar, radar interferometry, or stereo photogrammetry (Vicens-Miquel et al., 2024). These diverse raster types are defined by their resolution properties across multiple dimensions. Spatial resolution determines the

ground area represented by each pixel, ranging from sub-meter to kilometers for satellite imagery. Temporal resolution defines the frequency of data acquisition over the same area, critical for monitoring dynamic Earth processes. Spectral resolution describes the spectral bandwidth sensors provided for distinguishing spectral responses from different objects, while radiometric resolution represents the precision of measurements through bit depth. The evolution of raster data has also brought new challenges in data management and analysis. The increase in resolution across some or all dimensions has led to exponential growth in data volume, requiring new approaches to storage, processing, and analysis. Modern geospatial platforms must balance the richness of high-resolution data against computational efficiency and analytical effectiveness. This has driven many programming libraries and innovations in cloud computing, distributed processing, and artificial intelligence applications for effective raster data analysis at scale.

4. Modern Programming Infrastructure for Big Earth Data

The exponential growth in Earth observation data has driven significant advances in programming tools and frameworks specifically designed for largescale geospatial analysis. The Python programming language has emerged as the de-facto foundation for modern raster data analysis, offering an expansive ecosystem that continues to expand in response to growing demands for solutions that scale and sophisticated algorithm development. Multi-dimensional arrays have become the fundamental data type, providing a straightforward representation of remote sensing data across spatial and spectral dimensions. The matured libraries such as PyTorch and TensorFlow over the last decade have led to a fast developing landscape of remote sensing raster data analysis. With the support of engaged communities and strong use cases across domains, the balanced and optimized use of CPU and GPU in these two libraries for large data processing became easier to researchers, which drives active algorithm and model developments. TorchGeo, a PyTorch domain library, supports geospatial data handling with built-in dataset loading, extending PyTorch’s Dataset class with geospatial awareness, enabling efficient loading and management of large satellite image collections. Although not an exhaustive listing of modern options, these frameworks provide essential capabilities for developing and deploying sophisticated AI models for Earth observation analysis. As an important GPU-accelerated Python machine learning library, NVIDIA RAPIDS cuML (Raschka et al, 2020) enables breakthrough performance in geospatial analysis workflows through efficient parallel computing architectures. Typically paired with NVIDIA RAPIDS cuDF (a dataframe library) and Dask adapter packages, cuML provides CUDA GPU-accelerated implementations of traditional machine learning algorithms such as random forests, clustering, and key dimensionality reduction techniques, cuML enables significant processing speeds-up of large-scale Earth observation data as compared to CPU-based implementations (Arndt et al., 2024). This acceleration is particularly valuable for time-critical applications using large collection EO data.

The handling of multi-dimensional data has been transformed by Xarray (Hoyer and Hamman, 2017), which has revolutionized the analysis of labeled multi-dimensional arrays, particularly crucial for time series satellite imagery analysis. Its extension, Xarray-Spatial, provides specialized functionality for spatial vector data analysis, while Rasterio provides a ”Pythonic” I/O interface to raster data through GDAL. Recent developments in these libraries have improved integration with cloud storage systems and support for streaming operations and distributed processing.

5. Dataframe-based Raster Analysis

The evolution of dataframe-based approaches has also significantly enhanced raster data analysis workflows, particularly through the integration of pandaslike operations with geospatial capabilities. GeoPandas, combined with rasterio, enables seamless transitions between vector and raster data operations, while maintaining the familiar dataframe interface that facilitates complex analytical operations. Recent developments in libraries like xarray-spatial have further extended these capabilities by introducing dataframe-like operations specifically optimized for large-scale raster datasets. While pandas and GeoPandas established the foundation for geospatial data handling, the emergence of Polars and its geospatial extension GeoPolars has introduced new performance paradigms for raster analysis. GeoPolars leverages Rust's memory efficiency and parallel processing capabilities to provide significantly faster operations compared to traditional pandas-based workflows, particularly beneficial for large-scale geospatial raster operation. Polars list of capabilities continues to improve, including functionality to leverage NVIDIA RAPIDS cuDF GPU as a backend for accelerated computing. The integration of dataframe operations with cloud-native geospatial formats has further streamlined big data workflows. Platforms like Microsoft's Planetary Computer and Earth Engine have adopted dataframe-like APIs that abstract away the complexity of distributed storage while maintaining the intuitive nature of dataframe operations.

Distributed computing frameworks have evolved to handle the massive scale of remote sensing data, particularly in high-performance computing environments. Among these solutions, Dask stands out as a powerful tool that extends Python's parallel computing capabilities for geospatial data either in array or dataframes. It enables scientists and analysts to process datasets larger than memory while maintaining a familiar Python interface, making it especially valuable for earth observation data processing. Recent work combining Xarray and Dask has demonstrated the potential for processing vast amounts of observational data efficiently, particularly in applications such as non-linear Trust Region Reflectance solving (Raml et al., 2024). A popular ecosystem PanGeo further leverages several of these technologies to enable large scale reproducible geoscience analysis (Hamman et al., 2018; Abernathey et al., 2021).

6. Cloud-Native Computing and Infrastructure Evolution

Another important advancement addressing challenges related to downloading, processing, and analyzing massive amounts of Earth Observation (EO) data is the increased availability of cloud computing resources (Xu et al., 2022). Major technology companies have developed specialized platforms that have transformed how researchers and practitioners interact with geospatial data. Google Earth Engine (GEE) (Gorelick et al., 2017) launched in 2010, pioneered this approach by providing free access to petabytes of satellite imagery and scientific datasets, alongside a JavaScript and Python API for large-scale geospatial analysis. One of the most utilized capabilities provided by GEE is its mature EO data pre-processing support, which enables complex image correction with registration and calibration across petabyte-scale raster collections as shown in (Daams et al., 2023). Amazon Web Services (AWS) has taken a different but complementary approach through its Earth on AWS initiative, hosting major satellite datasets through the Registry of Open Data. AWS's Simple Storage Service (S3) has become a de facto standard for storing Cloud-Optimized GeoTIFFs (COGs), while services like SageMaker and Lambda enable end-to-end scalable machine learning geospatial analysis workflows (Shi et al., 2020). Planetary

Computer represents a newer but rapidly evolving platform that combines data hosting, computing resources, and specialized APIs for environmental monitoring. The platform emphasizes STAC-compliant EO data catalogs (satellite, weather, etc) and provides Python-based tools for interactive analysis, while distinguishing itself through its focus on environmental applications and integration with popular open-source geospatial libraries. These cloud platforms have fundamentally changed how we approach Earth observation analysis by eliminating the need to download data to local machines and providing scalable computing resources. This shift has enabled new forms of scientific collaboration and democratized access to advanced geospatial analysis capabilities, particularly benefiting academic researchers with limited access to computational infrastructure. The evolution of these platforms continues to emphasize machine learning integration, improved interoperability through efforts such as the SpatioTemporal Asset Catalogs (STAC) and the Open Geospatial Consortium (OGC), and enhanced support for real-time monitoring applications.

7. Future Directions

The continued integration of AI with geospatial science represents not just a technological advancement but a fundamental shift in how we observe, analyze, and understand our planet. The trajectory of GeoAI development suggests continued innovation and expanding capabilities, particularly as new satellite sensors are launched and computing technologies advance. Future developments in shared standards, geospatial integration into common technologies, API unification, software portability, edge processing for geospatial raster data analysis, and automated machine learning are likely to further accelerate the evolution of this field, opening new possibilities for Earth observation and analysis.

References

Abernathey, R., Augspurger, T., Banihirwe, A., Blackmon-Luca, C., Crone, T., Gentemann, C., Hamman, J., Henderson, N., Lepore, C., McCaie, T., et al. (2021). Cloud-native repositories for big scientific data. Computing in Science & Engineering, 23(2):26–35.

Arndt, J., Wohlgemuth, J., H. Yang, H.L., Bowman, J., Lunga, D., and King, D. (2024). A science gateway for the repeatable analysis of machine learning predicted gravity anomalies. IEEE Geoscience and Remote Sensing Letters, 21:1–5.

Chen, H., Qi, Z., and Shi, Z. (2021). Remote Sensing Image Change Detection With Transformers. IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-14.

Daams, M. N., Banquet, A., Delbouve, P., and Veneri, P. (2023). Consistent metropolitan boundaries for the remote sensing of urban land. Remote Sensing of Environment, 297:113789,

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202:18–27,

Hamman, J., Rocklin, M., and Abernathey, R. (2018). Pangeo: A bigdata ecosystem for

scalable earth system science. In American Geophysical Union, Fall Meeting 2018. American Geophysical Union.

Hoyer, S. and Hamman, J. (2017). xarray: N-D labeled arrays and datasets in Python. Journal of Open Research Software, 5(1).

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., and Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. ISPRS Journal of Photogrammetry and Remote Sensing, 152, 166-177.

Ma, Y., Zhang, Z., Yang, H. L., and Yang, Z. (2021). An adaptive adversarial domain adaptation approach for corn yield prediction. Computers and Electronics in Agriculture, 187:106314,

Raml, B., Quast, R., Schobben, M., Reimer, C., & Wagner, W. (2024). Unleashing the power of Dask with a high-throughput Trust Region Reflectance solver for raster datacubes. In EGU General Assembly 2024. EGU General Assembly 2024, Wien, Austria. EGU.

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. Information, 11(4), 193.

Rosentreter, J., Hagensieker, R., and Waske, B. (2020). Towards largescale mapping of local climate zones using multitemporal Sentinel 2 data and convolutional neural networks. Remote Sensing of Environment, 237:111472.

Shi, Y., Chen, X., and Zhang, T. (2020). Cloud-Based Deep Learning on AWS Open Data Registry: Automatic Building and Road Extraction from Satellite and LiDAR. In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Geospatial Data Access and Processing APIs (SpatialAPI'20). Association for Computing Machinery, New York, NY, USA, Article 1, 1–2.

Vicens-Miquel, M., Tissot, P. E., Colburn, K. F. A., Williams, D. D., Starek, M. J., Pilartes-Congo, J., Kastl, M., Stephenson, S., Medrano, F. A. (2024). Machine-Learning Predictions for Total Water Levels on a Sandy Beach," Journal of Coastal Research, 41(1), 57-72

Xu, C., Du, X., Fan, X., Giuliani, G., Hu, Z., Wang, W., ... Guo, H. (2022). Cloud-based storage and computing for remote sensing big data: a technical review. International Journal of Digital Earth, 15(1), 1417–1445.

Yang, H. L., Yuan, J., Lunga, D., Laverdiere, M., Rose, A., & Bhaduri, B. (2018). Building Extraction at Scale Using Convolutional Neural Network: Mapping of the United States. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 11, no. 8, pp. 2600-2614.

Zhang, J., Lei, J., Xie, W., Fang, Z., Li, Y., and Du, Q. (2023). SuperYOLO: Super Resolution Assisted Object Detection in Multimodal Remote Sensing Imagery. in IEEE

Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-15.