DOI: 10.12700/AIS.2025.008

First Experiences with the AlphaEarth Foundations Model: A Cosine Distance-Based Evaluation

Balázs Böröcz
Alba Regia Faculty
Óbuda University
Székesfehérvár, Hungary
borocz.balazs@amk.uni-obuda.hu

Gábor Molnár
Alba Regia Faculty
Óbuda University
Székesfehérvár, Hungary
molnar.gabor@amk.uni-obuda.hu

The recently released AlphaEarth Foundations Model was generated using multi-source satellite data (optical, radar, LiDAR, etc.) and consists of a 64-dimensional normalized vector for each 10m by 10m pixel on the whole globe for each year since 2017.
This representation enables the use of cosine distance to quantify similarity between pixels in the parameter space. The model has been made available in Google Earth Engine, providing new opportunities for environmental monitoring. In this study, we revisited previous analyses of agricultural areas and bee pastures to evaluate the applicability of the new methodology. We compared its performance with earlier approaches based on machine learning classifiers and spectral indices calculations. Our results show that, in most cases, computations were significantly faster and required less code, while accuracy improved slightly compared to traditional methods. We also examined the spatial and temporal transferability of the approach, emphasizing its potential for broader applications in land monitoring.

Keywords: AlphaEarth Foundations Model, Google Earth Engine, cosine distance, bee pastures, satellite data, land cover analysis

I. Introduction and Methodology

Until the appearance of AlphaEarth Foundations Model, there were relatively few success stories of AI in practical use for the Earth science community [1]. One of the research directions was establishing generative models, like DiffusionSat, that applied generative Diffusion model that performed well on many modalities (including images, speech, and video) and was tailored to support remote sensing data [2].

AlphaEarth Foundations primarily trains for distribution uniformity, and this will allow for more general-purpose tasks after training the learner for the special use case.

A. AlphaEarth Foundations Embedding model

The AlphaEarth Foundations (AEF) embedding model, developed by researchers at Google DeepMind and Google, was designed to address the challenge of creating accurate global maps from sparse and unevenly distributed ground truth data [3]. Traditional mapping efforts often rely on handcrafted features or domain-specific methods, which can be limited by noise, sensor dependence, or regional biases. AEF introduces a general-purpose geospatial representation that integrates multi-source and multi-model Earth observation data in a unified framework, enabling robust extrapolation of sparse labels into large-scale, high-resolution mapping products.

An embedding model in this context is a way of transforming diverse Earth observation inputs into a compact, universal representation. In AEF, each embedding is a 64-dimensional vector that summarizes signals from multiple modalities, including optical satellite imagery, radar, LiDAR, meteorological records, and other environmental data. These embeddings capture both spatial and temporal dynamics, making it possible to generate consistent and accurate maps even when annotated training data are scarce. By providing a common feature space for heterogeneous data sources, the model supports a wide range of applications, such as land use and land cover mapping, crop monitoring, biodiversity studies, and climate analysis.

The embeddings produced by AEF can be imagined as points distributed on the surface of a 64-dimensional unit sphere (radius = 1). Because the vectors are normalized in this way, their relative position in the embedding space encodes meaningful similarities and differences between objects or locations. This structure enables the use of cosine distance (or cosine similarity) as a metric for separating or clustering different objects: vectors that point in similar directions represent similar geospatial characteristics, while vectors that diverge more strongly correspond to distinct environmental conditions or land cover types.

B. Cosine Distance Computation

In the embedding model, each location or object is represented by a 64-dimensional vector positioned on the surface of a unit sphere. To determine how similar or different two such representations are, we can compare the angle between their vectors. This comparison is performed using cosine distance, which measures how close the orientation of two vectors is to one another. As these vectors are normalized, the cosine of their angle can be easily calculated by the dot product of the vectors, what is simply the sum of the product of the respective vector components.

If two vectors point in almost the same direction, their cosine distance is small, meaning they represent very similar geospatial characteristics. Conversely, if two vectors diverge strongly, the cosine distance is larger, indicating that the underlying environmental conditions or land cover types are different. This makes cosine distance particularly suitable for distinguishing subtle differences in complex, high-dimensional datasets like those produced by the AEF model.

For our task, implementing cosine distance means that once the embeddings are generated for the areas of interest, we can systematically compare them to evaluate their similarity. This allows us to group similar areas together, detect boundaries between distinct land cover types, or identify unusual regions that stand out from their surroundings. In practice, cosine distance becomes a powerful tool for applying the embeddings to classification, clustering, or change detection problems without needing large amounts of labeled training data.

To make the cosine distance operational in our task, we defined a set of training areas that represent the target class or phenomenon of interest. For each training area, embeddings were extracted and then averaged to form a mean vector. This mean vector acts as a reference point in the 64-dimensional space, summarizing the typical characteristics of the training data.

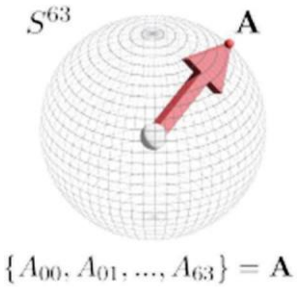


Figure 1 A stack of 64 rasterized AEF layers forms an embedding field, and each individual vector maps to a coordinate on the unit sphere S^{63}. [3]

Next, we calculated the maximum deviation from this mean vector across all training samples. This deviation serves as a threshold value in the model. In practice, when embeddings from unknown areas are compared to the mean vector, their cosine distance must remain below this threshold in order to be considered part of the same class. If the cosine distance exceeds the threshold, the area is classified as different.

A) $$A(A00m, A01m, \ldots, A62m, A63m)$$

B) $$B(A00, A01, \ldots, A62, A63)$$

C) $$cos(\Theta) = \frac{A * B}{|A| * |B|}$$

D) $$D = 1 - cos(\Theta)$$

Equation 1: Study area mean vector components (A). Components of the vector under investigation (B). Cosine similarity (C). Cosine distance (D).

This procedure ensures that the model is not overly sensitive to minor variations within the training data, while still being able to detect meaningful differences. In other words, the threshold defines how much divergence from the training examples we allow before labeling a new area as belonging to another category.

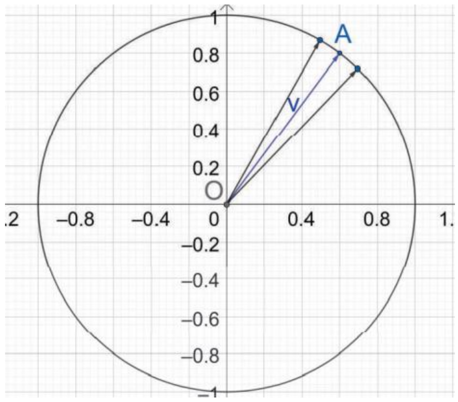


Figure 2 Mean vector of the training samples (blue) and the maximum threshold vector (black).

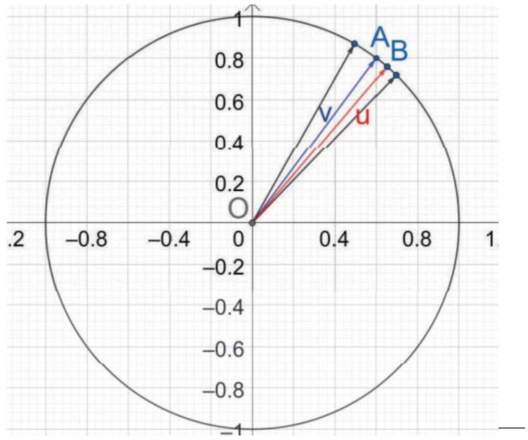


Figure 3 Example of a corresponding vector that falls within the specified threshold value (red).

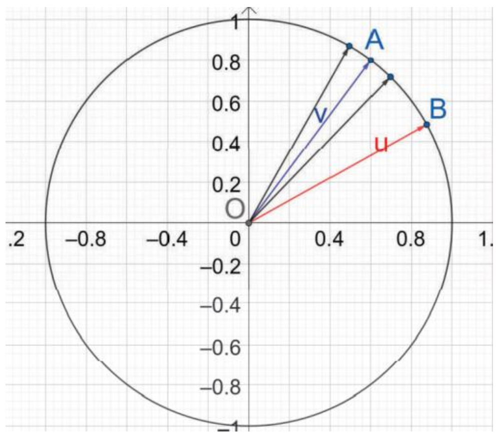


Figure 4 Example of an inappropriate vector that falls outside the specified threshold value (red).

II. Results: Investigated Plant Cultures

The focus of our research was on the detection of various plant cultures, of which we primarily dealt with the detection of bee pastures [4].

In the framework of bee pastures, three flowering plant species of major importance for pollinators were investigated. Rapeseed (*Brassica napus*) is one of the most widely cultivated oilseed crops in Europe, producing large fields that bloom with abundant yellow flowers. Its high nectar and pollen production make it a crucial early spring food source for honeybees, often contributing significantly to colony buildup after winter [5].

Figure 5 Rapeseed. source: https://flybyjing.com/blogs/news/rapeseed-oil

A)

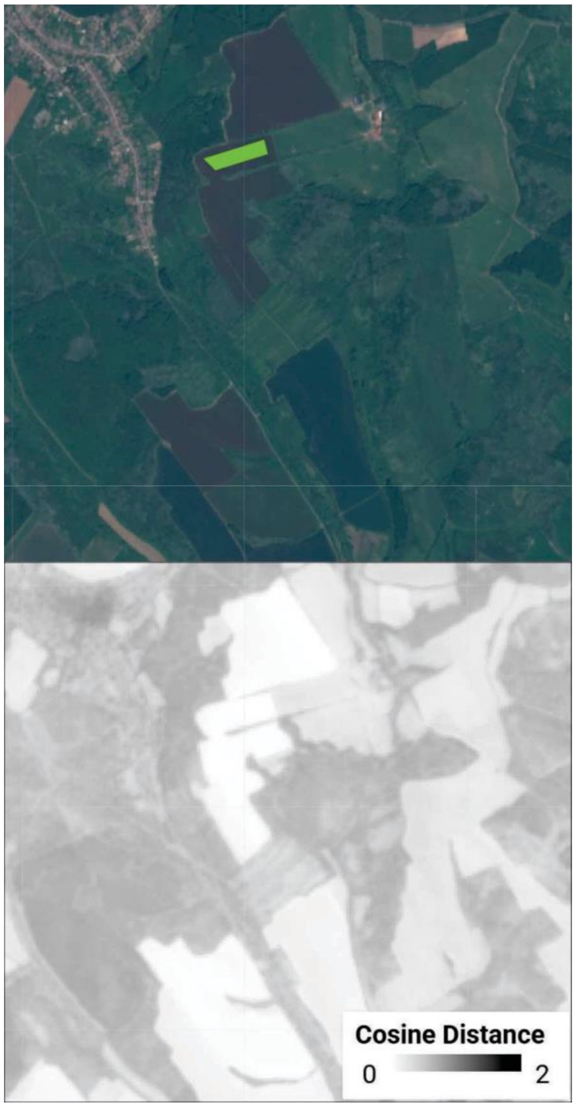
B)

C)

Figure 6 Detail of the rapeseed fields under investigation. The study area is shown with a red polygon (A). The cosine distance of pixels from the mean vector of the study area (B). Rapeseed fields detected after masking (C).

Crimson clover (*Trifolium incarnatum*), a leguminous forage crop, is valued both for soil enrichment through nitrogen fixation and for its dense stands of red flowers that provide substantial nectar resources. It typically blooms later in the season than rapeseed, thereby extending the availability of bee pasture resources and supporting continuous foraging opportunities for pollinators. [6]

Figure 7 Crimson clover. source: https://www.pennington.com/all-products/wildlife/resources/crimson-clover



A)

B)

C)

Figure 8 Detail of the crimson clover fields under investigation. The study area is shown with a green polygon (A). The cosine distance of pixels from the mean vector of the study area (B). Crimson clover fields detected after masking (C).

The detection of crimson clover (*Figure 8C*) also proved to be highly effective. The applied method successfully identified entire crimson clover fields, including areas where the satellite imagery already showed signs of post-flowering vegetation. This indicates that the embedding-based cosine distance approach can reliably detect the crop even when its phenological stage is past peak flowering, highlighting its robustness compared to methods that rely strictly on spectral signals captured at a single time point.

Black locust (*Robinia pseudoacacia*), a fast-growing deciduous tree, is highly regarded in beekeeping due to its profusion of fragrant white flowers. Its nectar is the source of one of the most prized unifloral honeys in Central Europe, commonly known as acacia honey. The species contributes greatly to bee forage, particularly in late spring to early summer, when large tracts of black locust forests bloom simultaneously.

Although black locust plays a central role as a bee pasture species, its detection using only cosine distance in the embedding space was not successful. The spectral and structural characteristics of black locust stands did not produce a sufficiently distinct signature compared to other forest types to allow reliable separation based solely on their distance from the mean training vector. As a result, the expected clear discrimination between black locust forests and other wooded areas could not be achieved through this method.

Due to these limitations, the focus was shifted from species-level detection of black locust to the broader detection and characterization of forested areas. Forests represent a structurally and ecologically coherent land cover type that is more readily distinguishable from other vegetation categories in the embedding space. By treating forests as a single class rather than attempting to separate individual tree species, the methodology could achieve more robust and consistent classification results, making it better suited for large-scale mapping tasks.

A)

B)

C)

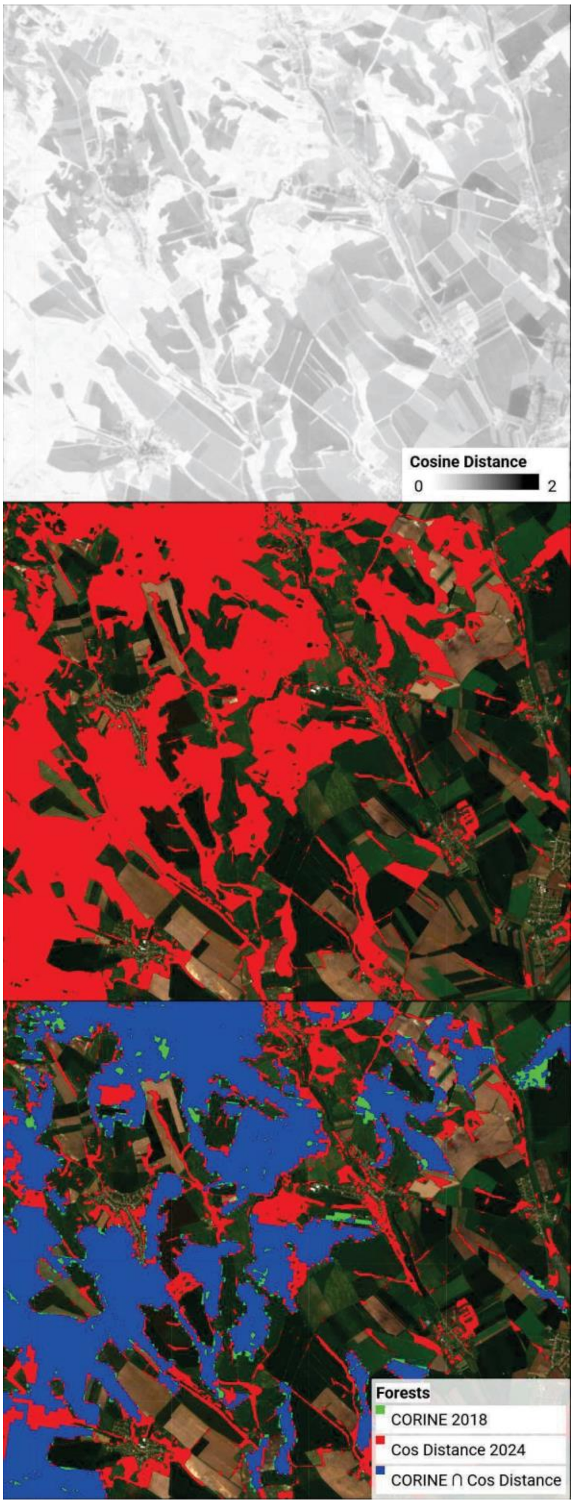


Figure 9 Northern Hungary region, around Szirák. Cosine distance calculated from the study area (A). Detected forest areas after masking (B). Comparison of detected forest areas with the CORINE 2018 land cover database (C).

As shown in Figure 10C, there is a substantial spatial overlap between the forest areas delineated by the CORINE Land Cover database and those detected using the embedding-based methodology (blue). However, the forest areas identified exclusively by our approach (red) reveal additional fine-scale details. This is primarily due to the higher spatial resolution of the embedding data (10 m pixel size), which enables the detection of narrow forest strips, such as tree belts along roads or small wooded patches that are not represented in the CORINE dataset.

Furthermore, part of the discrepancy arises from temporal differences between the datasets. While the CORINE Land Cover database reflects land cover conditions from 2018 (in 100 m pixel size), our analysis represents the state of 2024, capturing land cover changes that occurred in the intervening period. This includes newly established or expanded forested areas that have developed since the CORINE dataset was compiled. As a result, the observed differences reflect both improvements in spatial resolution and genuine temporal dynamics in forest cover.

III. Comparison with Other Methods

When compared to other commonly used classification and change detection techniques, the cosine distance–based analysis of the embedding model demonstrates both notable strengths and clear limitations. The primary advantage of this approach lies in its conceptual simplicity and efficiency: by relying on the geometric relationships between normalized embedding vectors, it enables the identification of areas with similar environmental and spectral characteristics without the need for large annotated training datasets or complex model retraining. This makes the method particularly attractive for rapid assessments and for applications where labeled reference data are limited.

At the same time, the cosine distance approach is inherently sensitive to intra-class variability and subtle differences between land cover types, which can limit its discriminative power in more complex scenarios, such as distinguishing between tree species or different crop types with overlapping phenological patterns. Compared to supervised machine learning classifiers or advanced segmentation algorithms, cosine distance tends to perform best at broader class levels, where the within-class variability is relatively low (e.g., forest vs. non-forest), while more specialized methods may outperform it at finer thematic resolutions.

Overall, the comparison indicates that cosine distance analysis represents a robust and scalable baseline method, which can be effectively applied for large-scale mapping tasks or as a complementary step prior to more sophisticated classification approaches.

A. *Comparison of Rapeseed Detection: Random Forest vs. Cosine Distance*

In the detection of rapeseed, a comparison was made between a Random Forest classifier and the cosine distance–based embedding approach. Random Forest represents a machine learning–based method that requires the algorithm to learn both the characteristics of the target class (rapeseed) and the non-target classes (non-rapeseed) [7]. To achieve reliable performance, the training dataset must contain a large and diverse set of samples representing non-rapeseed surfaces, capturing the full range of spectral and temporal variability present in the study area. This typically necessitates the delineation of multiple, well-distributed training areas. In addition, the implementation of the Random Forest algorithm involves a more complex workflow, including data preprocessing, feature selection, hyperparameter tuning, and model validation.

In contrast, the cosine distance method leverages the properties of the embedding space to separate rapeseed fields from other land cover types using only a few representative training areas. By calculating a mean embedding vector for known rapeseed fields and applying a cosine distance threshold, the method can identify rapeseed areas without explicitly providing the model with examples of non-rapeseed. This results in a simpler algorithmic setup, significantly shorter processing times, and high reliability in detecting rapeseed when clear phenological signals are present in the embedding space.

Furthermore, it is important to note that Random Forest performs best when applied during the flowering period, when rapeseed fields display distinctive spectral characteristics that enable the algorithm to differentiate them from other land cover types more easily. This requires the precise selection of imagery corresponding to the flowering stage, which can vary from year to year. By contrast, the embedding model does not depend on the exact date of data acquisition. It only requires the specification of the target year, as the temporal information is inherently encoded in the embedding representation. This makes the cosine distance approach less sensitive to timing and phenological variability, further simplifying its application in practice.

B. *Comparison of Rapeseed Detection: Index-Based Method vs. Cosine Distance*

In addition to machine learning and embedding-based approaches, index-based methods were also evaluated for detecting rapeseed (Brassica napus). This approach relies on the use of spectral indices derived from satellite imagery, specifically the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Yellow Index (NDYI). NDVI is primarily used to mask out clouds and non-vegetated surfaces, ensuring that the analysis focuses only on vegetated areas. NDYI is then applied to identify rapeseed fields during their flowering stage, leveraging the strong yellow spectral signal characteristic of blooming rapeseed.

$$\mathrm{NDYI} = \frac{\mathrm{B3 - B2}}{\mathrm{B3 + B2}} \qquad \mathrm{NDVI} = \frac{\mathrm{B8 - B4}}{\mathrm{B8 + B4}}$$

Equation 2 Normalized Difference Yellow Index (NDYI) (left), Normalized Difference Vegetation Index (NDVI) (right) [8]

A key advantage of the index-based method is that it does not require any training data, which significantly simplifies its application. When the flowering stage is well-defined and the imagery is acquired at the appropriate time, the method can achieve high detection accuracy. Additionally, the algorithm itself is computationally straightforward, even simpler than the cosine distance–based approach.

However, the limitations of this method are closely tied to timing and flowering intensity. It can only be applied during the rapeseed flowering period, as outside this window the yellow spectral signal is absent and NDYI cannot reliably distinguish rapeseed from other crops. In cases of weaker or heterogeneous flowering, rapeseed fields may be partially or fully masked out, leading to underestimation. Another drawback is the need for manual threshold selection for NDYI values, which introduces subjectivity and may affect consistency across different regions or years.

IV. Conclusions

Overall, while index-based methods are attractive due to their simplicity and lack of training requirements, their strong dependence on precise timing and manual parameterization make them less flexible than the embedding model–based cosine distance approach for large-scale or multi-temporal analyses as summarized in Table 1.

TABLE I. Summary of the main characteristics, advantages, and limitations of the three rapeseed detection methods

Method	Training Data	Temporal Sensitivity	Complexity	Accuracy	Main Advantages	Main Limitations
Random Forest	High – target & non–target required	High – flowering period needed	High	High	Robust, detailed classification	Data– and time–intensive; precise timing needed
Cosine Distance (Embedding)	Low – few rapeseed samples	Low – only target year needed	Medium	High	Few samples, fast, temporally flexible	Less effective for subtle class differences
Index-Based (NDVI+NDYI)	None	Very high – peak flowering required	Low	High*	Very simple, fast, no training data	Manual thresholding; weak flowering may cause omission; only works during flowering

*High accuracy when applied at peak flowering with suitable threshold

V. References

[1] Z. Sun, T. t. Brink, W. Carande, G. Koren, N. Cristea, C. Jorgenson, B. Janga, G. P. Asamani, S. Achan, M. Mahoney, Q. Huang, A. Mehrabian, T. Munasinghe, Z. Liu és A. Margolis, „Towards practical artificial intelligence in Earth sciences," *Computational Geosciences*, 2024.

[2] S. Khanna, P. Liu, L. Zhou, C. Meng, R. Rombach, M. B. D. B. Lobell és S. Ermon, „DIFFUSIONSAT: A GENERATIVE FOUNDATION MODEL," in *ICLR Conference*, Vienna, 2024.

[3] C. F. Brown, M. R. Kazmierski, V. J. Pasquarella, W. J. Rucklidge, M. Samsikova, C. Zhang, E. Shelhamer, E. Lahera, O. Wiles, S. Ilyushchenko, N. Gorelick, L. L. Zhang, S. Alj, E. Schechter, S. Askay, O. Guinan, R. Moore, A. Boukouvalas és P. Kohli, „AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data.," Google DeepMind, London, 2025.

[4] B. Böröcz, „Estimation of the location of bee pastures in Hungary based on SENTINEL satellite images," in *AIS 2023 - 18th International Symposium on Applied Informatics and Related Areas - Proceedings*, Székesfehérvár, 2023.

[5] N. Raboanatahiry, H. Li, L. Yu és M. Li, „Rapeseed (Brassica napus): Processing, Utilization, and Genetic Improvement," *Agronomy*, 2021.

[6] T. J. Butler, G. W. Evers, M. A. Hussey és L. J. Ringer, „Flowering in Crimson Clover as Affected by Planting Date," *Crop Science*, pp. 242-247, 2002.

[7] A. Liaw és M. Wiener, „Classification and Regression by randomForest„" *R News*, 2002.

[8] N. Zamani-Noor és D. Feistkorn, „Monitoring Growth Status of Winter Oilseed," *Agronomy*, 2022.