Aitchison-Loss Training with Geospatial Embeddings Sharpens Compositional Land-Cover Maps

Ayato Kanno[1], Narumasa Tsutsumida[1,2]

[1] Graduate School of Science and Engineering, Saitama University, Japan (kanno.a.972@ms.saitama-u.ac.jp)
[2] Space Data Frontiers Research Center, Fujitsu Research, Fujitsu Limited, Japan (rsnaru.jp@gmail.com)

Keywords: Compositional land-cover classification, Mixed-pixel problem, Sentinel-2, OpenEarthMap, Foundation model embeddings, Aitchison distance

Abstract

Accurate land-cover maps are essential, but medium-resolution imagery (e.g., Sentinel-2 at 10 m) often contains mixed pixels that include multiple land-cover types. Standard "hard" classification assigns one class per pixel, hiding minority classes and reducing map usefulness. Compositional classification instead estimates the proportion of each class within a pixel, preserving sub-pixel detail, but requires outputs that are non-negative and sum to one, constraints not naturally handled by typical ML/DL losses. This study proposed and evaluated a deep-learning framework for compositional land-cover estimation at 10 m resolution. It compared two input feature sets: (1) reflectance from 10 Sentinel-2 multispectral bands (B2–B8, B8A, B11, B12) and (2) Embedding V1, a 64-dimensional representation from the AlphaEarth Foundations model that integrates multi-source, multi-temporal Earth observation signals. Ground-truth composition vectors were derived from OpenEarthMap by aggregating 0.25–0.5 m labels to 10 m pixels for eight classes. Three architectures (MLP, 2D-CNN, 3D-CNN) used Softmax outputs to enforce the constant-sum constraint, and two losses (MAE vs Aitchison distance) were tested. Embedding V1 improved estimation accuracy across all model architectures compared to Sentinel-2 spectral bands alone. While 3D-CNN achieved the best performance with Sentinel-2 input (MAE: 0.1126), MLP outperformed all other architectures when Embedding V1 was used (MAE: 0.0989). Comparison of fraction maps revealed that MAE produced spatially smoothed outputs, whereas Aitchison distance yielded sharper and more realistic compositions. The combination of Embedding V1, MLP, and Aitchison distance loss achieved the best overall performance, suggesting that foundation model embeddings combined with compositional losses can improve sub-pixel land-cover estimation.

1. Introduction

Land-cover information is an essential geospatial resource for understanding surface conditions and human activities, and is widely used in various fields such as urban planning, resource management, disaster prevention, and climate change studies (Wang et al., 2018; Holmberg et al., 2023; Zhang et al., 2025). Satellite remote sensing has become a primary approach for large-scale land-cover mapping, as multispectral satellite imagery such as Sentinel-2 is globally available and enables continuous monitoring over large areas (Chaves et al., 2020; Wang et al., 2023). However, the nature of the spatial resolution of satellite imagery often prevents it from adequately capturing the complexity of surface structures. In conventional land cover mapping, a hard classification approach is used: each pixel is assigned to a single land-cover class, often by a model that assigns it to the class with the highest probability. Consequently, mixed pixels containing multiple land-cover types cannot be adequately represented (Wang et al., 2014; Wan et al., 2021). Minority classes within a pixel may be ignored, resulting in information loss and a reduced ability to represent heterogeneous landscapes (Shanmugam et al., 2006).

To address this problem, compositional land-cover classification has been proposed as an approach to estimate the proportions of multiple land-cover classes within a pixel (Wan et al., 2021; Zhang et al., 2023). Rather than assigning a single label, compositional classification estimates fractional land-cover compositions, providing a more detailed representation of heterogeneous surface conditions. Previous studies have estimated land-cover fractions from mixed pixels using raw satellite observations and spectral indices derived from multispectral imagery as input features (Heremans et al., 2016; Hu et al., 2018; Sierra et al., 2025). To infer sub-pixel land-cover compositions from these data, various approaches have been proposed (Wu et al., 2023; Pacheco and McNairn, 2010; Heremans et al., 2016; Zhang et al., 2018; Hu et al., 2018). Linear and machine-learning approaches have demonstrated the feasibility of estimating land-cover fractions from remotely sensed imagery, while MLP-based models have been shown to effectively capture nonlinear relationships between spectral features and land-cover compositions (Heremans et al., 2016). More recently, CNN-based methods have achieved improved estimation accuracy by exploiting spatial context from neighboring pixels (Zhang et al., 2018; Hu et al., 2018). Furthermore, three-dimensional CNNs (3D-CNNs) have been proposed to jointly learn spatial and spectral information, enabling the extraction of local correlations among adjacent wavelength bands and further improving sub-pixel land-cover estimation performance (Perikamana et al., 2021; Sierra et al., 2025; Pattathal V. et al., 2022). Although these MLP- and CNN-based methods can quantify compositional land cover, their architectures are not well-suited to compositional analysis. Specifically, compositional outputs are constrained ratio data characterized by non-negativity and a constant-sum constraint, yet previous studies on compositional classification have relied on conventional regression losses such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which treat each class fraction independently, failing to capture the relative relationships among land-cover classes (Perikamana et al., 2021; Wan et al., 2021; Heremans et al., 2016; Hadi et al., 2025). Therefore, it is necessary to explore loss functions specifically designed for compositional data in fractional land-cover estimation.

Most existing approaches for compositional land-cover classification estimate land-cover fractions primarily from spectral reflectance values and spectral indices derived from multispectral imagery (Heremans et al., 2016; Hu et al., 2018; Sierra et al., 2025). Previous studies have suggested that single-date spectral imagery may be insufficient to fully capture the complex spatial, temporal, and semantic relationships of land surfaces (Chen et al., 2024; Rußwurm and Körner, 2018). To overcome these limitations, recent foundation models trained on large-scale multi-source satellite observations have leveraged self-supervised learning to learn generalized feature representations from massive amounts of unlabeled remote sensing data (Xue et al., 2025; Khan and Ahmad, 2025; Feng et al., 2025). These representations are typically provided as embedding vectors, which encode spectral, spatial, temporal, and contextual information in a high-dimensional feature space (Feng et al., 2025; Brown et al., 2025).

While previous studies have improved land-cover fraction estimation from multispectral satellite imagery by developing sophisticated deep learning models, several research gaps remain to be addressed. First, although land-cover fractions are compositional data characterized by non-negativity and a constant-sum constraint, most existing studies rely on conventional regression losses (MAE and RMSE), potentially limiting their suitability for compositional land-cover estimation. Second, most existing fraction-estimation approaches primarily utilize spectral reflectance values and spectral indices derived from multispectral imagery. While recent foundation-model-derived embedding vectors have demonstrated strong performance in categorical land-cover classification by encoding spatial, temporal, and semantic information beyond conventional spectral features, their effectiveness for compositional land-cover classification and mixed-pixel fraction estimation remains largely unexplored. To address these gaps, this study proposes a compositional land cover classification model with high-dimensional foundation model embedding vectors and a novel compositional loss function to mitigate the mixed pixel problem in land cover classification. To demonstrate our approach, we estimate the proportions of eight land-cover classes within each 10-m pixel, leveraging Sentinel-2 observations and AlphaEarth embeddings (Brown et al., 2025) as inputs. OpenEarthMap (OEM)-based class fractions were prepared as reference data (Xia et al., 2023). To examine the selection of model architectures and loss functions, we compared three model architectures (MLPs, 2D-CNNs, and 3D-CNNs) and introduced the Aitchison distance as the loss function.

2. Methods

2.1 Data

2.1.1 Reference Data OEM was used as high-resolution reference data for compositional land-cover classification. This global dataset was constructed by (Xia et al., 2023) through manual annotation of land-cover classes from aerial and satellite imagery sources acquired at different times. Since the underlying source datasets were collected primarily around 2020, the reference labels were assumed to represent land-cover conditions in 2020. The dataset comprises approximately 2,300 labeled images at a spatial resolution of 0.25–0.5 m, enabling detailed representation of complex surface structures with the definition of eight land-cover classes: Bareland, Rangeland, Developed space, Road, Tree, Water, Agriculture land, and Building. With coverage across 97 regions in 44 countries on six continents, the dataset provides substantial geographic diversity suitable for large-scale land-cover fraction estimation (Table 1).

Class	Number of pixels (M)	Proportion (%)
Bareland	74	1.5
Rangeland	1130	22.9
Developed space	798	16.1
Road	331	6.7
Tree	996	20.2
Water	161	3.3
Agriculture land	680	13.7
Building	770	15.6

Table 1. Number of pixels used in this study from OpenEarthMap by land cover class.

2.1.2 Input Data Two types of input features were prepared for land-cover fraction estimation: Sentinel-2 multispectral imagery and Google Satellite Embedding V1. Sentinel-2 multispectral imagery was obtained from the Harmonized Sentinel-2 MSI Level-2A product (COPERNICUS/S2_SR_HARMONIZED) available on Google Earth Engine at every OEM patch. Sentinel-2 delivers multispectral observations spanning visible to shortwave infrared wavelengths at spatial resolutions of 10 to 20 m and is widely used for land-cover analysis (Chaves et al., 2020). In this study, 10 bands covering the visible, red-edge, near-infrared, and shortwave-infrared regions were selected based on previous studies (Chaves et al., 2020) (Table 2). Since these bands have native spatial resolutions of 10 or 20 m, all bands were prepared at a 10 m resolution, with the 20 m bands resampled accordingly.

Band	Band Name	Center Wavelength (nm)	Spatial Resolution (m)
B2	Blue	492	10
B3	Green	560	10
B4	Red	665	10
B5	Red-edge 1	705	20
B6	Red-edge 2	740	20
B7	Red-edge 3	783	20
B8	NIR	833	10
B8A	Narrow NIR	865	20
B11	SWIR 1	1610	20
B12	SWIR 2	2190	20

Table 2. Sentinel-2 spectral bands used in this study.

We also used Google Satellite Embedding V1 (Embedding V1), generated by the AlphaEarth Foundations model and available in Google Earth Engine (GOOGLE/SATELLITE_EMBEDDING/V1/ANNUAL) (Brown et al., 2025). Embedding V1 provides a 64-dimensional embedding vector for each 10 m grid cell. Unlike conventional spectral bands, these dimensions do not correspond to direct physical measurements such as reflectance values; instead, they represent coordinates in a learned feature space derived from large-scale Earth observation data. The AlphaEarth Foundation model is trained on diverse data sources, including optical imagery, SAR observations, topographic data, climate variables, and geospatially referenced text, enabling the embeddings to encode spectral, spatial, temporal, and semantic characteristics of surface conditions (Brown et al., 2025). Each embedding summarizes observations over one year, thereby capturing seasonal signals and surrounding spatial context in addition to local spectral properties.

2.1.3 Dataset Construction

To construct the training dataset, Sentinel-2 imagery and Embedding V1 data corresponding to each OEM patch were extracted. To reduce cloud contamination and seasonal variability, cloud and shadow masking was applied to Sentinel-2 imagery using the Sentinel-2 Cloud Probability dataset (Schwehr, 2020). Cloud probability–based masking has been shown to effectively remove cloud-contaminated observations and improve land-cover mapping accuracy from Sentinel-2 imagery (Gao et al., 2024). Cloud pixels were identified from the cloud probability band using a predefined threshold, while shadow pixels were detected by combining dark near-infrared pixels with cloud projections along the solar azimuth direction. Water pixels were excluded using the Scene Classification Layer (SCL). Following cloud and shadow removal, pixel-wise median composites were generated from all available Sentinel-2 observations acquired between April 2020 and April 2021, producing a single representative image for each region. For Embedding V1, annual embedding representations corresponding to the same period were obtained. Sentinel-2 imagery, Embedding V1 data, and OEM labels were reprojected into a common coordinate reference system and spatially aligned to ensure positional consistency among all datasets.

Figure 1 illustrates the generation process of the fraction labels used as ground-truth data. The fine-resolution OEM land-cover labels were aggregated within each 10 m Sentinel-2 grid cell, and the fractional coverage of the eight land-cover classes was calculated. The resulting land-cover fractions were used as ground-truth targets for supervised learning.

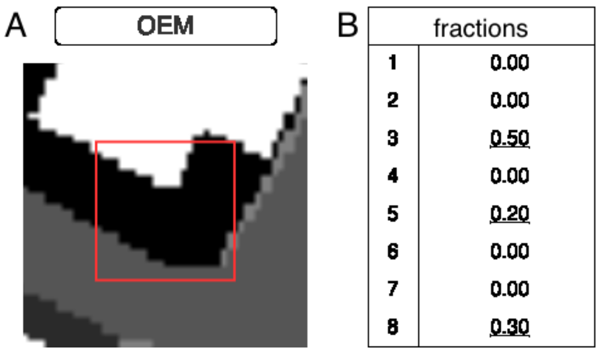

Figure 1. Generation process of fraction labels. (A) High-resolution OpenEarthMap (OEM) land-cover labels. The red square indicates a single 10 m Sentinel-2 pixel. (B) Fraction labels generated by aggregating the OEM labels within the corresponding Sentinel-2 pixel and calculating the fractional coverage of each land-cover class.

2.2 Models

To investigate the effectiveness of different deep-learning architectures for compositional land-cover estimation, we implemented MLP, 2D-CNN, and 3D-CNN. The MLP is a neural network composed of fully connected stacked layers (FC) that learns nonlinear relationships between input features and output land-cover fractions through linear transformations and activation functions. CNN-based models were also used to leverage the spatial context of neighboring pixels. Because many land-cover classes exhibit spatial continuity, convolutional operations are expected to improve fraction estimation by capturing local spatial structures and class boundaries. This study implemented both 2D-CNN and 3D-CNN architectures to investigate differences in how spatial and spectral information are utilized. The 2D-CNN performs convolution operations exclusively in spatial dimensions. Multispectral images with spatial dimensions of $H \times W$ and spectral dimension C are input to the model, and convolution kernels of size $k \times k$ extract spatial features from neighboring pixels. Although spectral bands are treated as independent channels, the model effectively learns local spatial structures and land-cover boundaries. In contrast, the 3D-CNN performs convolution operations in both spatial and spectral dimensions, enabling simultaneous learning of spatial structures and spectral relationships among adjacent wavelength bands. For example, Sentinel-2 multispectral imagery can be represented as a three-dimensional tensor with spatial dimensions ($H \times W$) and spectral dimension (B). By applying convolution kernels of size $k \times k \times k_\lambda$, the 3D-CNN jointly captures spatial context and local spectral correlations.

2.3 Loss Function

The target outputs are land-cover fractions represented as the compositional vector $\mathbf{y} = (y_1, \ldots, y_D)$, where $D = 8$ corresponds to the eight OEM land-cover classes (Table 1). Each component satisfies the non-negativity constraint ($y_i \geq 0$), and the components sum to unity ($\sum_{i=1}^{D} y_i = 1$). Consequently, an increase in one class proportion necessarily alters the relative proportions of the remaining classes. Such data, governed by ratio relationships rather than absolute magnitudes, are referred to as compositional data and require treatment distinct from standard regression problems (Aitchison et al., 2000). MAE is defined as in Equation 1, which evaluates prediction errors for each component independently:

$$\mathrm{MAE} = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{1}{D} \sum_{i=1}^{D} |y_{n,i} - \hat{y}_{n,i}| \right), \tag{1}$$

where N is the number of pixels, $y_{n,i}$ is the reference fraction of class i for pixel n, and $\hat{y}_{n,i}$ is the corresponding predicted fraction.

Because MAE treats class proportions as separate scalar quantities, it does not account for the relative relationships inherent in compositional data. To address this limitation, we employed the Aitchison distance (Aitchison et al., 2000) as a loss function. As defined in Equations 2 and 3, the Aitchison distance measures dissimilarity between two compositions through log-ratio transformation with respect to their geometric means. By employing the Aitchison distance, we preserve the geometric properties and constraints of land-cover compositions in mixed-pixel fraction estimation.

$$d_A(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\sum_{i=1}^{D} \left(\ln \frac{y_i}{g(\mathbf{y})} - \ln \frac{\hat{y}_i}{g(\hat{\mathbf{y}})} \right)^2} \tag{2}$$

$$g(\mathbf{y}) = \left(\prod_{i=1}^{D} y_i \right)^{1/D}, \tag{3}$$

where $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_D)$ is the predicted composition vector, and $g(\cdot)$ represents the geometric mean.

2.4 Experimental Design

To evaluate the effectiveness of different input features for compositional land-cover estimation, two input configurations were tested: Sentinel-2 spectral bands and Embedding V1. The dataset comprised 2,300 patch pairs, randomly partitioned into 70% for training, 10% for validation, and 20% for testing. Four patches within the test set were reserved for visualization comparisons and kept fixed across all experiments. Three deep-learning architectures were implemented: MLP, 2D-CNN, and 3D-CNN. Models were trained using two loss functions: MAE and Aitchison distance.

Model performance was assessed using three metrics: MAE, RMSE, and Aitchison distance. MAE and RMSE values were computed to evaluate the prediction error across all land-cover classes, while the Aitchison distance assesses the geometric structure and relative proportion relationships characteristic of compositional data. Class fractions were mapped from the predicted outputs and visually compared with the reference data. This assessment examined the spatial continuity of fraction distributions and the reproduction of extreme values (proportions near 0 or 1), which are physically meaningful in land-cover composition. Additionally, residual maps were created by computing the difference between predicted and reference proportions at each pixel for each class to detailed investigation of systematic prediction biases and class-specific error trends.

2.5 Training Strategy

Sentinel-2 spectral bands were normalized to the range $[0, 1]$ using min-max normalization to prevent training instability caused by differences in value scales across bands. Embedding V1 features were used without normalization, as the embedding vectors are already unit-normalized by construction. All models were optimized using Adam, and dropout was applied to prevent overfitting. Early stopping was employed to terminate training when the validation loss did not improve for 20 consecutive epochs. The complete training conditions and hyperparameters are summarized in Table 3.

The MLP consists of three hidden layers, each with 256 units and ReLU activation, with dropout applied after each hidden layer, followed by a fully connected output layer with Softmax activation. For CNN-based models, Group Normalization and ReLU activation were applied after each convolutional layer, and the output was produced by two fully connected layers followed by Softmax. In both cases, Softmax ensures that the predicted fractions sum to unity.

We preliminarily investigated the parameters of patch size, number of channels, and kernel size across CNN-based models. Since no substantial differences in estimation performance were observed across the tested configurations, the smallest values(patch size 5, 32 channels, kernel size 3) were adopted to reduce computational cost. The explored and adopted settings are summarized in Table 4.

Item	Setting
Optimizer	Adam
Learning rate	1.0×10^{-3}
Maximum epochs	200
Normalization	GroupNorm
Activation function	ReLU
Dropout rate	0.2
Number of output classes	8
Output layers	FC + Softmax

Table 3. Training conditions and hyperparameters

Item	Explored values	Adopted value
Patch size	5, 7, 9	5
Number of channels	32, 64	32
Kernel size	3, 5, 7	3

Table 4. Hyperparameter search settings for CNN-based models

3. Results

3.1 Performance by Input Configuration and Model

We conducted experiments for each combination of model architecture and input data type by using the Aitchison distance as the loss function in all experiments. These experiments revealed contrasting performance patterns depending on the input data type. Table 5 summarizes the results for different input configurations and model architectures. When using Sentinel-2 spectral bands, 3D-CNN achieved the best performance, followed by 2D-CNN, suggesting that CNN architectures capable of jointly extracting spatial and spectral information are effective for multispectral fraction estimation. In contrast, MLP outperformed CNN-based models when using Embedding V1 as input.

Input	Model	MAE	RMSE	Aitchison
Sentinel-2	MLP	0.1185	0.2456	9.7737
	2D-CNN	0.1159	0.1978	9.6670
	3D-CNN	**0.1126**	**0.1871**	**9.4980**
Embedding V1	MLP	**0.0989**	0.1895	**8.9905**
	2D-CNN	0.1055	**0.1734**	9.2114
	3D-CNN	0.1094	0.1769	9.4658

Table 5. Evaluation of MLP, 2D-CNN, and 3D-CNN using the Aitchison distance loss function with Sentinel-2 or Embedding V1 inputs. Bold values represent the best performance for each metric within each input type.

3.2 Comparison of Loss Functions

To evaluate the effectiveness of compositional loss functions, we trained an MLP with Embedding V1 as input using MAE and Aitchison distance as loss functions, and compared the resulting estimation performance (Table 6). Training with MAE resulted in lower component-wise error metrics (MAE: 0.0923, RMSE: 0.1354), whereas training with Aitchison distance resulted in lower compositional error (Aitchison distance: 8.9905). Figures 2 and 3 demonstrated class-specific maps of example patches in Dar es Salaam and Tokyo, respectively, and revealed visual differences in estimation results depending on the choice of loss function. When using Aitchison distance loss, spatial distributions of each land-cover class were reproduced more sharply, with class boundaries and proportion changes clearly visible. In contrast, when using MAE, fraction distributions exhibited overall smoothing, suggesting that minimizing absolute error tends to produce spatially averaged predictions.

Model/Input	Loss	MAE	RMSE	Aitchison
MLP/Embedding V1	Aitchison	0.0989	0.1895	**8.9905**
	MAE	**0.0923**	**0.1354**	9.6439

Table 6. Evaluation of MLP using MAE or Aitchison distance loss with Embedding V1 inputs. Bold values represent the best performance for each metric.

https://doi.org/10.5194/isprs-archives-L-4-W1-2026-95-2026 | © Author(s) 2026. CC BY 4.0 License.

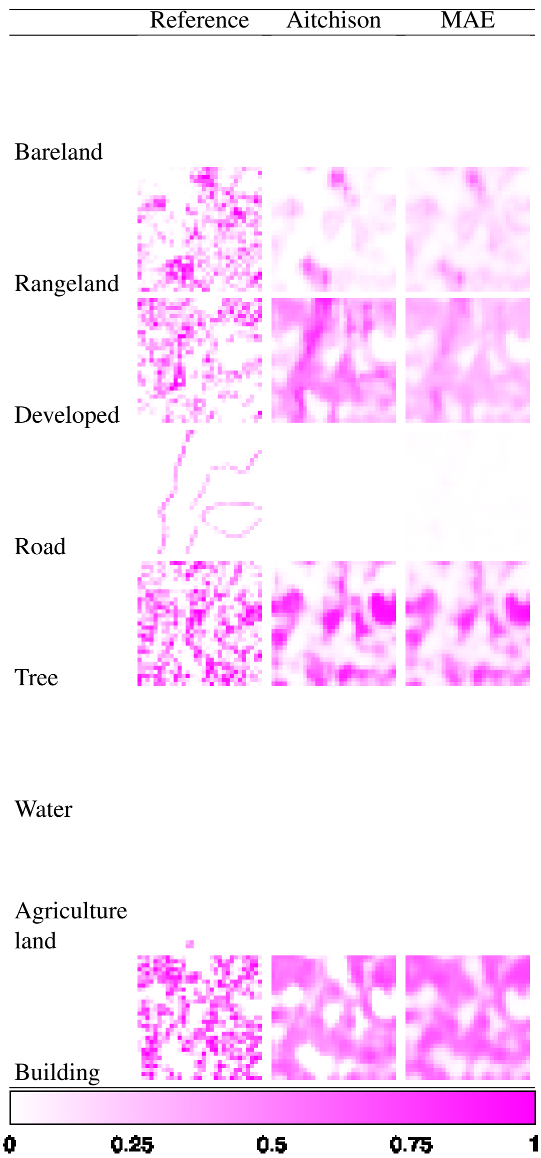

Figure 2. Spatial distribution of predicted land cover fractions by MLP using MAE and Aitchison distance loss functions with Embedding V1 inputs, shown for an example patch in Dar es Salaam.

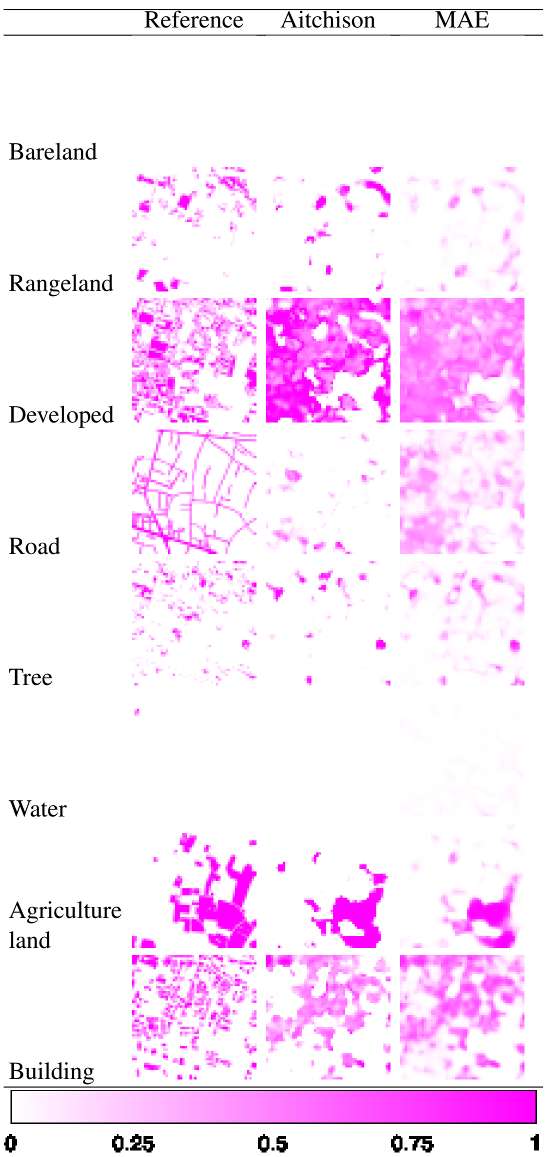

Figure 3. Spatial distribution of predicted land cover fractions by MLP using MAE and Aitchison distance loss functions with Embedding V1 inputs, shown for an example patch in Tokyo.

3.3 Error Distribution Analysis

We generated error maps for the best-performing model configuration (MLP with Embedding V1 input and Aitchison distance loss) at two test sites: Dar es Salaam and Tokyo, as illustrative examples (Figure 4). The maps revealed distinct patterns of over- and underestimation across land-cover classes. Rangeland and Agriculture land classes were predominantly underestimated at both sites. The Road class exhibited linear underestimation patterns that closely followed the road network,suggesting that the model had difficulty accurately estimating road fractions. In contrast, Developed space class was often overestimated across both sites. Tree and Building classes showed moderate errors of comparable magnitude in both directions. Bareland and Water classes showed little to no estimation error.

4. Discussion

This study investigated model architectures, loss functions, and hyperparameters for deep learning-based compositional land-cover classification. Specifically, we used Sentinel-2 spectral bands and Embedding V1 as input features, with land-cover fractions derived from OEM as reference data for MLP and CNN-based model architectures with MAE and Aitchison distance losses. We found the Aitchison distance was an effective loss function for compositional data in terms of class-wise fraction maps. We observed that inputting Embedding V1 improved model accuracy across all architectures, due to the rich information encoded in its 64-dimensional feature vectors, which integrate spectral, spatial, temporal, and semantic information, providing substantially more informative input features than conventional spectral bands in both quality and quantity. Since spatial and temporal information is encoded during the generation of Embedding V1, convolutional feature extraction using 2D- or 3D-CNNs may not yield better performance than MLP. MLP directly maps high-dimensional abstract feature vectors

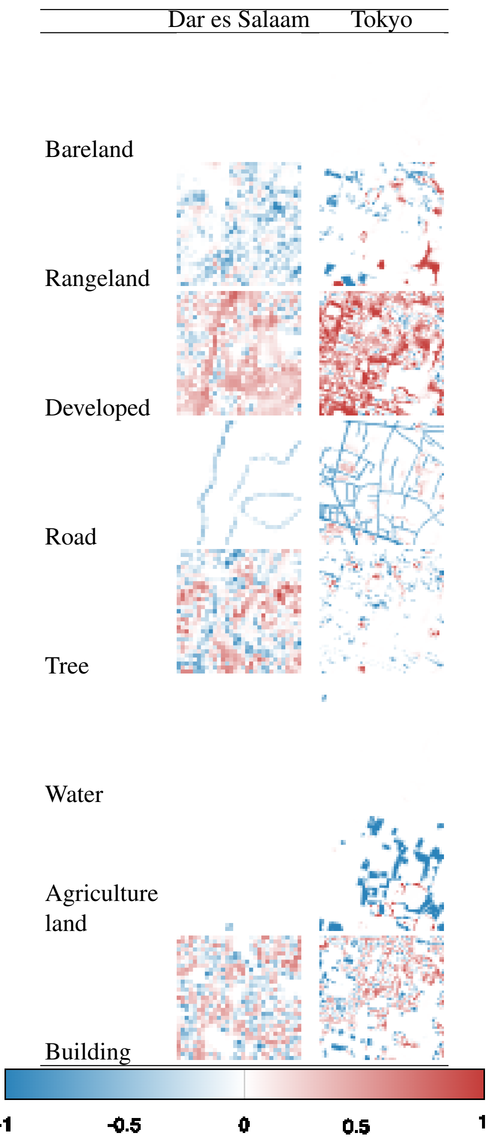

Figure 4. Error maps of predicted land cover fractions relative to the reference, generated by MLP with Aitchison distance loss using Embedding V1 inputs. Examples are shown for selected patches in Dar es Salaam and Tokyo. Blue and red indicate underestimation and overestimation, respectively.

to the output through fully connected layers and may therefore more efficiently leverage the pre-encoded information in Embedding V1. This finding is consistent with prior work reporting that simpler model structures tend to yield greater performance gains with embedding-based inputs (Brown et al., 2025), although that work did not address compositional land-cover analysis. The insight that increasing model complexity does not necessarily improve performance is expected to serve as design guidelines for compositional classification models using embedding vectors.

The selection of the loss function was important: the Aitchison distance yielded lower compositional error, while MAE produced lower component-wise error metrics, indicating that each loss function optimizes for a different aspect of estimation accuracy. The difference in performance between the two loss functions can be explained by the nature of land-cover compositions at 10-m resolution. In practice, a 10 m $\times$ 10 m pixel is often dominated by a small number of classes, meaning that most classes have a fraction near zero for any given pixel. MAE minimizes absolute errors per class independently, which tends to underweight errors near zero and, consequently, produces spatially smoothed outputs that fail to reproduce dominant-class compositions. In contrast, the Aitchison distance evaluates differences via a log-ratio transformation, which assigns relatively large penalties to errors near zero because of the nature of logarithms. This property enables the model to more accurately reproduce fraction compositions in which most classes are absent, thereby suppressing proportional error across the overall compositional structure.

The error distribution patterns can be partly attributed to class imbalance in the training data. Bareland and Water classes together account for only 4.8% of total pixels, yet showed negligible estimation errors, likely because these classes have spectrally distinct characteristics that are relatively easy to distinguish from other classes, even with limited training samples. In contrast, Rangeland, Agriculture land, and Road classes were consistently underestimated due to their spectral similarity with other vegetated or impervious surfaces, making them difficult to distinguish at the sub-pixel level. In particular, Road class accounts for only 6.7% of pixels and is frequently mixed with Developed space and Building in urban environments, likely contributing to its systematic underestimation, while the overestimation of Developed space may reflect a partial mis-attribution of road and building fractions to this class.

While our approach is promising, we should discuss a limitation that should be addressed in future work. Since the training data are derived from OEM that provides land-cover labels at a global scale, each class inherently encompasses diverse spectral and structural characteristics across different geographic regions and climatic zones. This intra-class diversity may introduce inconsistencies in class representations and limit the model's ability to estimate fractions uniformly across all regions. In addition, the inherent class imbalance in the OEM dataset may systematically bias the model toward dominant classes, suppressing accurate estimation of minority classes. Future work should focus on identifying more suitable model architectures for Embedding V1-based inputs and on reducing the influence of class imbalance, aiming to improve generalizability by testing relevant datasets.

5. Conclusions

We proposed a deep learning framework for compositional land-cover fraction estimation at 10-m spatial resolution to mitigate the mixed pixel problem. Our experiment showed that Embedding V1 consistently improved estimation accuracy across all model architectures compared to using Sentinel-2, demonstrating that foundation model embeddings provide richer and more informative features than conventional spectral bands for compositional classification. The optimal model architecture depended on the input type, suggesting that architecture selection should reflect the abstraction level of the input features rather than defaulting to more complex models. The Aitchison distance loss produced sharper and more realistic fraction maps than MAE, confirming that the loss function selection tailored to compositional data properties is more suitable for land-cover fraction estimation than conventional regression losses. However, the error maps revealed that class imbalance and inter-class spectral similarity remain key sources of estimation error. Reducing these errors through class-weighted loss functions or data augmentation strategies represents an important direction for future work.

The proposed framework enables detailed land-cover composition mapping at the sub-pixel level, addressing the mixed-pixel problem not for specific regions but towards global-scale land surface analyses. This approach has the potential to substantially improve the accuracy of land-cover statistics and spatial analysis in applications such as urban planning, green infrastructure assessment, and environmental monitoring, where fine-grained surface information beyond the capability of conventional hard classification is required.

Acknowledgements

This study was funded by JSPS KAKENHI (25K00624).

References

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., Pawlowsky-Glahn, V., 2000. Logratio Analysis and Compositional Distance. *Mathematical Geology*, 32(3), 271–275.

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Zhang, L. L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., Kohli, P., 2025. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data.

Chaves, E. D. M., Picoli, C. A. M., Sanches, I. D., 2020. Recent Applications of Landsat 8/OLI and Sentinel-2/MSI for Land Use and Land Cover Mapping: A Systematic Review. *Remote Sensing*, 12(18), 3062.

Chen, C., Yuan, X., Gan, S., Kang, X., Luo, W., Li, R., Bi, R., Gao, S., 2024. A new strategy based on multi-source remote sensing data for improving the accuracy of land use/cover change classification. *Scientific Reports*, 14, 26855.

Feng, Z., Atzberger, C., Jaffer, S., Knezevic, J., Sormunen, S., Young, R., Lisaius, M. C., Immitzer, M., Jackson, T., Ball, J., Coomes, D. A., Madhavapeddy, A., Blake, A., Keshav, S., 2025. TESSERA: Temporal embeddings of surface spectra for earth representation and analysis.

Gao, X., Chi, H., Huang, J., Han, Y., Li, Y., Ling, F., 2024. Comparison of Cloud-Mask Algorithms and Machine-Learning Methods Using Sentinel-2 Imagery for Mapping Paddy Rice in Jianghan Plain. *Remote Sensing*, 16(7), 1305.

Hadi, F., Farooque, G., Shao, Y., Yang, J., Xiao, L., 2025. DSSFT: Dual Branch Spectral-Spatial Feature Fusion Transformer Network for Hyperspectral Image Unmixing. *Earth Science Informatics*, 18, 352.

Heremans, S., Suykens, J. A., Van Orshoven, J., 2016. The effect of imposing ‘fractional abundance constraints’ onto the multilayer perceptron for sub-pixel land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, 44, 226-238.

Holmberg, M., Junttila, V., Schulz, T., Grönroos, J., Paunu, V.-V., Savolahti, M., Minunno, F., Ojanen, P., Akujärvi, A., Karvosenoja, N., Kortelainen, P., Mäkelä, A., Peltoniemi, M., Petäjä, J., Vanhala, P., Forsius, M., 2023. Role of land cover in Finland’s greenhouse gas emissions. *Ambio*, 52(11), 1697–1715.

Hu, Y., Zhang, Q., Zhang, Y., Yan, H., 2018. A Deep Convolution Neural Network Method for Land Cover Mapping: A Case Study of Qinhuangdao, China. *Remote Sensing*, 10(12).

Khan, H., Ahmad, A., 2025. Evaluating AlphaEarth Foundation Embeddings for Pixel- and Object-Based Land Cover Classification in Google Earth Engine. *Preprints*.

Pacheco, A., McNairn, H., 2010. Evaluating multispectral remote sensing and spectral unmixing analysis for crop residue mapping. *Remote Sensing of Environment*, 114(10), 2219-2228.

Pattathal V., A., Sahoo, M. M., Porwal, A., Karnieli, A., 2022. Deep-learning-based latent space encoding for spectral unmixing of geological materials. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183, 307-320.

Perikamana, K. K., Balakrishnan, K., Tripathy, P., 2021. A CNN based method for Sub-pixel Urban Land Cover Classification using Landsat-5 TM and Resourcesat-1 LISS-IV Imagery. *CoRR*. https://arxiv.org/abs/2112.08841.

Rußwurm, M., Körner, M., 2018. Multi-Temporal Land Cover Classification with Sequential Recurrent Encoders. *ISPRS International Journal of Geo-Information*, 7(4), 129.

Schwehr, K., 2020. Sentinel-2: Cloud Probability in Earth Engine (Version 1).

Shanmugam, P., Ahn, Y.-H., Sanjeevi, S., 2006. A comparison of the classification of wetland characteristics by linear spectral mixture modelling and traditional hard classifiers on multispectral remotely sensed imagery in southern India. *Ecological Modelling*, 194(4), 379-394.

Sierra, S., Ramo, R., Padilla, M., Quirós, L., Cobo, A., 2025. Estimating Fractional Land Cover Using Sentinel-2 and Multi-Source Data with Traditional Machine Learning and Deep Learning Approaches. *Remote Sensing*, 17(19).

Wan, L., Chen, T., Plaza, A., Cai, H., 2021. Hyperspectral Unmixing Based on Spectral and Sparse Deep Convolutional Neural Networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 11669–11682.

Wang, C., Wang, Y., Wang, R., Zheng, P., 2018. Modeling and evaluating land-use/land-cover change for urban planning and sustainability: A case study of Dongying city, China. *Journal of Cleaner Production*, 172, 1529-1534.

Wang, Q., Shi, W., Atkinson, P. M., 2014. Sub-pixel mapping of remote sensing images based on radial basis function interpolation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 92, 1-15.

Wang, Y., Sun, Y., Cao, X., Wang, Y., Zhang, W., Cheng, X., 2023. A review of regional and Global scale Land Use/Land Cover (LULC) mapping products generated from satellite remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 206, 311-334.

Wu, J., Li, Y., Zhong, B., Liu, Q., Wu, S., Ji, C., Zhao, J., Li, L., Shi, X., Yang, A., 2023. Integrated vegetation cover of typical steppe in China based on mixed decomposing derived from high resolution remote sensing data. *Science of The Total Environment*, 904, 166738.

Xia, J., Yokoya, N., Adriano, B., Broni-Bediako, C., 2023. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6254–6264.

Xue, Z., Yang, G., Yu, X., Yu, A., Guo, Y., Liu, B., Zhou, J., 2025. Multimodal self-supervised learning for remote sensing data land cover classification. *Pattern Recognition*, 157, 110959.

Zhang, B., Wang, C., Shen, Y., Liu, Y., 2018. Fully Connected Conditional Random Fields for High-Resolution Remote Sensing Land Use/Land Cover Classification with Convolutional Neural Networks. *Remote Sensing*, 10(12).

Zhang, C., Wang, Q., Atkinson, P. M., 2025. Unsupervised object-based spectral unmixing for subpixel mapping. *Remote Sensing of Environment*, 318, 114514.

Zhang, X., Ge, Y., Chen, J., Ling, F., Wang, Q., Du, D., Xiang, R., 2023. High-quality super-resolution mapping using spatial deep learning. *iScience*, 26(6), 106875.