### 

### **Generative AI holds transformative potential for sustainability, particularly through its application in environmental monitoring and resource management. A poignant example of this is the use of a Variational Autoencoder (VAE) to estimate soil organic carbon content—an essential factor in assessing soil health and promoting sustainable agricultural practices. The VAE model excels in processing and synthesizing complex data, enabling it to capture the intricate relationships within satellite imagery and other geospatial data types. By employing a VAE, we can generate a latent space—a compressed representation of the original input data—that retains critical information pertinent to organic carbon levels. This latent space is then used to regress the organic carbon content accurately, providing a novel and efficient method to monitor and manage soil health on a large scale. Such applications not only enhance our understanding of environmental conditions but also pave the way for predictive and prescriptive analytics in sustainable land management.**

### 

### **The Challenge of Measuring Soil Organic Carbon**

Soil organic carbon is a dynamic and complex component of the soil, influenced by various factors including climate, land use, and farming practices. Accurate measurement and monitoring are vital for making informed decisions regarding land management and environmental policies. Traditional methods, while effective, are not scalable to the granularity required for large-scale global monitoring.

### **AI to the Rescue: Convolutional Variational Autoencoders**

Enter the Variational Autoencoder (VAE), a class of machine learning models that are excellent at handling complex, high-dimensional data. VAEs are particularly adept at data generation and reconstruction, making them suitable for tasks where understanding the underlying distribution of data is crucial.

For our application, we employ a VAE with convolutional layers. Convolutional Neural Networks (CNNs) are highly effective in image processing due to their ability to capture spatial hierarchies in data. By combining CNNs with VAEs, we can effectively model the spatial distribution of SOC using satellite imagery.

### **Methodology: Using Raster Bands for SOC Estimation**

Our approach utilizes multiple raster bands from satellite images, which represent different segments of the electromagnetic spectrum. These bands can capture various soil properties, which correlate with SOC levels. The VAE architecture is designed to encode these high-dimensional inputs into a lower-dimensional latent space, from which the SOC content can be predicted.

1. Data Collection: We gather multispectral satellite images covering different geographic and climatic regions to ensure the model's versatility.
2. Preprocessing: The images are processed to align with the raster bands, and noise reduction techniques are applied to ensure data quality.
3. Model Training: The convolutional VAE is trained on a dataset of images labeled with SOC measurements. This phase involves tuning hyperparameters to optimize the reconstruction and prediction accuracy.
4. SOC Estimation: The trained model predicts SOC by decoding the latent representations of new, unlabeled satellite images.

### **Significant Results and Implications**

The convolutional VAE model showed significant ability to predict SOC from unseen satellite images. The high correlation between the predicted SOC values and the ground truth data indicates the model's effectiveness. This proof of concept opens several avenues:

* Scalability: This method can be scaled to monitor SOC across vast areas, providing critical data for environmental and agricultural management.
* Cost Efficiency: AI-driven approaches reduce the need for extensive physical sampling, lowering the costs and labor involved in SOC monitoring.
* Policy Making: Reliable and timely SOC data can inform better policies for land use and climate change mitigation.

A critical component of our dataset was the 2015 LUCAS (Land Use/Cover Area frame Statistical Survey) soil survey, which provided a robust foundation for training and validating our convolutional VAE model. Conducted between May and November of 2015, the survey collected soil samples from approximately 16,000 sites across Europe, offering a comprehensive snapshot of various soil properties, including organic carbon content.

By integrating these precise ground-truth measurements into our model, we significantly enhanced the accuracy and reliability of our SOC predictions, grounding our AI-driven insights in rigorously collected empirical data. This integration not only improved the model's performance but also demonstrated the practical applicability of combining traditional soil survey methods with cutting-edge AI techniques for environmental monitoring and analysis.

The process of using satellite raster data for soil organic carbon (SOC) prediction through a Convolutional Variational Autoencoder (VAE) involves several steps, from data acquisition and preprocessing to model training and evaluation. Here, we outline a comprehensive workflow using geemap, a Python library that facilitates the use of Google Earth Engine for geospatial data manipulation and visualization. This workflow includes transforming raster data into tensors, creating and fitting a CNN VAE, and finally regressing SOC from the latent space embeddings.

### **Step 1: Data Acquisition with geemap**

The first step involves acquiring satellite imagery raster data. geemap can be utilized to extract multispectral raster datasets from Google Earth Engine, which hosts a vast array of satellite image collections. For this project, we focus on the spectral bands most relevant to predicting SOC characteristics.

### **Step 2: Preprocessing and Transformation into Tensors**

Once the raster data is acquired, the next step is preprocessing. This typically involves normalizing the data, handling missing values, and potentially augmenting the dataset to improve model robustness. The preprocessed raster images are then transformed into tensors, which are the required input format for neural networks. This transformation involves converting the multidimensional raster arrays into tensor data structures suitable for processing by machine learning models.

### **Step 3: Creating and Training the CNN VAE**

With the data in the correct format, we proceed to define our Convolutional VAE architecture. This architecture combines convolutional layers for feature extraction from spatial data, with the encoding-decoding mechanism of a VAE. The encoder compresses the input tensors into a lower-dimensional latent space, while the decoder attempts to reconstruct the input data from this compressed representation. During training, the model learns to minimize the reconstruction loss, adjusted by the latent space regularization to encourage efficient data encoding. We employ Root Mean Square Error (RMSE) as our loss metric, which provides a direct measure of reconstruction accuracy.

We aimed for a RMSE of about 0.08 on an epoch for our data, and finetuned our VAEs until the manifold of the data was learned accordingly for compression.

### **Step 4: Using the Latent Space for SOC Prediction**

After training, the model encodes new, unlabeled data into the latent space. This latent representation, which ideally captures the essential information about the SOC characteristics from the input tensors, is then used as features for regression. Here, we concatenate the latent space vectors with location data derived from the LUCAS 2015 dataset to incorporate geographical context.

### **Step 5: Regression and Analysis**

In the final step, we use a tobit regression model to predict the S content from the concatenated features of latent vectors and geographical data. The regression model is trained to minimize the RMSE between the predicted and actual Organic Carbon values, providing a direct assessment of prediction accuracy. We everage the deep features learned by the convolutional VAEs in order to regress a value of the data collected from the LUCAS survey.

### **Evaluation and Results**

Upon evaluation, the model demonstrates significant predictive capability, as evidenced by low RMSE scores on a validation set. This indicates that the convolutional VAE effectively captures and utilizes the complexities of both the spatial and spectral data to estimate SOC. These results underscore the potential of integrating advanced machine learning models with traditional soil survey data to revolutionize SOC monitoring and analysis. This approach not only improves the scalability and efficiency of SOC estimation but also enhances the accuracy and applicability of the results in real-world scenarios.

### 

### **Implementing Perceptual Losses for Enhanced Model Training**

In addition to the conventional loss metrics such as RMSE, our approach integrates perceptual losses to train the Convolutional VAE on the raster data. Perceptual losses compare the high-level features of the predicted outputs against the targets, essentially focusing on the perceptual similarity rather than pixel-wise accuracy. This method aligns more closely with how humans perceive images and helps in preserving textural and contextual integrity during the reconstruction phase of the VAE.

**The Generative Nature of CNN VAEs in Denoising Applications**

By learning to generate and reconstruct data, the model inherently learns to filter out noise and irrelevant information, focusing instead on the most salient features necessary for accurate Organic Content predictions. This capability makes the convolutional VAE an excellent tool for remote sensing applications where data often contain substantial amounts of environmental and sensor-related noise.

### **Enhancing Model Insights: Statistical Significance and Proof of Concept**

These variables, derived from the model's encoder, are lower manifold representation of the input data and are integral to understanding and predicting soil organic carbon (SOC) content. Statistical tests confirm that these latent variables significantly contribute to the regression model’s predictive accuracy, demonstrating their relevance and utility in capturing meaningful information from the raster data.

One of the notable capabilities of the Convolutional VAE is its effectiveness in denoising and extracting useful information from noisy data sources, such as satellite imagery, especially when averaged out over a year. This attribute is particularly valuable in remote sensing, where data often contain various types of noise due to atmospheric conditions, sensor inaccuracies, or other environmental factors.

### **Implications for Remote Sensing and Environmental Monitoring**

The successful implementation and significant results of this project underscore the potential of advanced machine learning techniques to transform remote sensing applications. By leveraging the power of Convolutional VAEs, researchers and practitioners can enhance their analytical capabilities, leading to more accurate environmental assessments and better-informed policy-making. We can use Generative models such as VAEs for denoising tasks, when estimating soil samples.

### **Conclusion**

The convolutional VAE using raster bands for Organic Carbon estimation is a promising tool in environmental science. This proof of concept not only showcases the feasibility of using AI for soil analysis but also sets the stage for more advanced applications in remote sensing and ecological monitoring. As we continue to refine these models, the potential to revolutionize how we interact with and manage our natural resources becomes increasingly tangible.