UNSUPERVISED LEARNING TECHNIQUES FOR MICROSEISMIC AND CROSWELL GEOPHYSICAL DATA

A Thesis

by

KEYLA GRACIELA GONZALEZ ABAD

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Siddharth Misra
Co-Chair of Committee,	Eduardo Gildin
Committee Member,	Mark Everett
Head of Department,	Jeff Spath

December 2021

Major Subject: Petroleum Engineering

ABSTRACT

Machine learning has served to develop and explore a wide range of applications for geoscientists and petroleum engineers. Fundamental limitations of conventional methodologies include mathematical formulations of physical systems, multi-scale heterogeneity, processing of large datasets, and computational time. The impact of these new technologies has brought the interest of multiple energy industries such as renewables, oil and gas, carbon sequestration, and geothermal. The acquisition of subsurface measurements has been a key factor to characterize reservoir properties. Hence, the integration of machine learning could provide essential information and new knowledge of subsurface monitoring signals. In this work, we focus on the use of unsupervised learning to determine new insights into geophysical tools and subsurface physical properties. We propose three methodologies using microseismic, distributed acoustic sensing (DAS), seismic and electrical resistivity tomography.

A critical aspect of monitoring tools is the high computational power of big data. We applied unsupervised dimensionality reduction to compress, denoise and retrieve vital information of microseismic and DAS data. To achieve this, we implemented high-order SVD for high-dimensional arrays of 3D and 4D space. For the 3D microseismic, we achieved a compression of approximately 75% and a reduction of samples from 1,728,000 to 431,303. We also tested the model to the 3D DAS data where we obtained a compression of 70.2% for a data size of 3.5 GB. Lastly, a 4D HOSVD model was established using a synthetic microseismic tensor, accomplishing a reduction of 83%.

Another major application of unsupervised learning is the clustering algorithms to group observations of similar characteristics. We applied spatial-temporal clustering to identify hidden patterns of subsurface mapping for a geological carbon storage field. The studies were divided according to the geophysical method (crosswell seismic and ERT) and temporal component (single time or time-series). Using crosswell seismic, we developed a multi-level clustering approach to visualize the CO_2 plume behavior. For the first level, we obtained a silhouette score of 0.85, a Calinski-Harabasz of 160666.50, and a Davies-Bouldin value of 0.43. The second level achieved a silhouette, Calinski-Harabasz, and Davies-Bouldin score of 0.74, 59656.01, and 0.32 respectively. We established a total of four clusters of non, low, medium, and high SCO_2.

Finally, we elaborated a spatial-temporal clustering using derived-SCO_2 from daily ERT images. A novel feature extraction methodology was designed to retrieve the spatial and temporal changes of the moving CO_2. Four clusters were determined and linked to the saturation levels. The interval validation of clusters was 0.58 for the DTW-silhouette score, 262791.45 for Calinski-Harabasz, and 0.71 for the Davies-Bouldin index. To evaluate the dynamics of CO_2 flow regimes, we performed a second clustering where 6 distinctive plume patterns were observed. Therefore, machine learning and in particular unsupervised learning can be used to describe complex systems and optimize data processing.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Misra, my committee co-chair, Dr. Gildin, and my committee member, Dr. Everett, for their guidance and support throughout the course of this research.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

Finally, thanks to my mother, sister, and brother-in-law for their encouragement and to my friends for all their love and support.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis (or) dissertation committee consisting of Professors Dr. Siddharth Misra and Dr. Eduardo Gildin of the Harold Vance Department of Petroleum Engineering and Professor Dr. Mark Everett of the Department of Geology & Geophysics.

Funding Sources

Graduate study was supported by a research assistantship from the Texas A&M Energy Institute funded through Convergence Research Incubator at Texas A&M University.

The contents and results of this work are solely the responsibility of the student and his advisory committee, and do not necessarily represent the official views of the funding sources cited above.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

Machine learning has become one of the most important techniques in a wide range of disciplines. These algorithms have the ability to extract patterns in the data without any prior programming. Among the different industries, energy systems (e.g., oil, gas, geothermal, and carbon sequestration) have been in ongoing progress. Some of these machine learning applications are increasingly used for solving big data memory and interpreting complex dynamic processes.

Developments in the oil and gas industry have been extensively observed. Nikravesh (2007) presented the application of artificial intelligence for reservoir characterization using multiple datasets such as geophysical, geological, and production measurements. In carbon capture and sequestration (CCS), machine learning is used to improve and optimize existing technologies and processes of carbon capture design (Rahimi et al., 2021). Nevertheless, some of these techniques can be applied to any subsurface earth resource. For instance, data dimensionality has been used for multiple geophysical tools to reduce their computational size. Consequently, the application of machine learning provides a variety of advantages on subsurface characterization.

Overview of machine learning

Machine learning (ML) is a branch of artificial intelligence that provides the ability of learning data relationships without being explicitly programmed (Samuel, 1959). The impact of ML algorithms can be observed under multiple circumstances such as detecting

hidden patterns, complex approaches, predicting future responses, discovering new models, and intensive traditional coding (Géron, 2019). Extensive applications have been observed to tackle key problems in different domains: climate change, medical diagnosis, speech recognition, fraud detection, and image recognition.

Supervised versus unsupervised learning

Two major machine learning approaches correspond to supervised and unsupervised learning. The main difference between them is linked to the associated response of each measurement. Supervised learning relates each observation to a specific response where we aim to build a model that can map the features measurements and response (labels). Multiple algorithms are being established to obtain an accurate relationship in the supervised domain (James et al., 2021).

Conversely, unsupervised learning lacks the response of the features observations. In other words, there are no labels in the learning process. To retrieve insights into the structure of the observations, we draw inferences between the features or observations. One widely unsupervised technique from unlabeled datasets is clustering. Unsupervised cluster analysis seeks to assign single observations of similar behavior into distinct groups. Lastly, a hybrid approach is sometimes required for an imbalanced dataset of more observations and fewer responses. This is called semi-supervised learning where we aim to incorporate all the observations with and without labels.

Unsupervised learning algorithms

Some of the most important unsupervised learning approaches can be classified as (Géron, 2019):

- Clustering or cluster analysis, to identify groups of similar data structures.
- Anomaly and novelty detection, to remove or retrieve the outliers from the dataset.
- Dimensionality reduction, to reduce the data feature space without losing the most valuable information.
- Association rule learning, to reveal uncover relations from large datasets.

Unsupervised clustering

As mentioned above, clustering splits the data into groups of strong resemblance. Unsupervised clustering assigns a unique class to each observation to represent a well-separated set of clusters. Numerous approaches have been developed which can be divided into partitioning, density, hierarchical, and grid-based algorithms.

K-means, which corresponds to a partitioning-based method, designates a data point to a region of the data where no observation belongs to another region. This algorithm aims to minimize the within-cluster-variation by iteratively updating the regions (cluster centers) and reassigning the observations. Another commonly used method is agglomerative clustering. This hierarchical-based algorithm merges data points of their hierarchy using a specified similarity criterion. The clustering can display a hierarchical path of all the intermediate steps for each data point clustering. Both K-means and agglomerative requires to previously choose the number of clusters, being this a hyperparameter to tune. For the density-based method, techniques such as DBSCAN and Meanshift can be used. Meanshift cluster data points on regions of maxima density and a predefined radius. On the other hand, DBSCAN groups data on their trajectory nearby

space taking into account the shape of the dataset and similar data responses. The use of these density methods does not require the number of clusters.

Dimensionality reduction

In unsupervised transformation, a dataset can be represented in a structure of lower features by only capturing the essential values. Some of the most popular reduction techniques are principal component analysis (PCA), linear discernment analysis (LDA), and non-negative matrix factorization (NMF). PCA transforms a correlated set of features into uncorrelated ones by performing an orthogonal rotation of the dataset. This rotation captures a subset of features of the most impactful features that maximizes the variance of the data (Muller & Guido, 2016). Similarly, LDA projects the high-dimensional dataset to a lower-dimension that maximize the separation of classes (Reddy et al., 2020). On the other hand, NMF extracts components by decomposing the data into a non-negative sum of components; hence, features can only be greater or equal to zero.

For data visualization, t-distributed stochastic neighbor embedding (t-SNE) is widely used to visualize the high-dimensional dataset. It creates a lower-dimensional projection of two dimensions.

Machine learning application in subsurface characterization

A vast number of machine learning models have been developed in subsurface characterization. Datasets comprise a wide range of measurement tools such as well logs, seismic, electromagnetic, electric, remote sensing, and reservoir modeling. This section briefly introduces studies conducted in reservoir engineering.

Li et al. (2018) proposed the use of neural networks to generate synthetic NMR logs. They performed two machine learning models called long short-term memory (LSTM) and variational autoencoder with a convolutional layer (VAEc), where they manage to achieve an accuracy of R^2 of 0,78 and 0.75 respectively.

Wu et al. (2019) developed an SEM (scanning electron microscope) image segmentation model to visualize four rock components: pores, rock matrix, pyrite, and organic components. The workflow involved the use of feature extraction techniques and the random forest classifier. The segmentation was validated, obtaining an F1-score greater than 0.9.

Furthermore, Chakravarty et al. (2021) applied unsupervised learning to map the geomechanical alterations of hydraulic fracturing operations. Ultrasonic transmission waveforms were used to retrieve fracture damage using clustering analysis. The physical significance of each group was evaluated to validate the geomechanical change.

Bao et al. (2020) proved the use of recurrent neural networks for reservoir modeling. The objective was to build a model that can relate control parameters and production outputs. The method shows an accuracy improvement, a reduced computational cost, and a reservoir simulation proxy model.

Thesis objectives and contributions

In this research, we explore the use of machine learning to discover new understandings of spatial-temporal subsurface monitoring. To achieve this, three studies were conducted using real-field and synthetic datasets for unconventional oil and gas, and

carbon sequestration. The imaging tools used were microseismic, distributed acoustic sensing (DAS), crosswell seismic, and electrical resistivity tomography.

The study will provide a set of detailed methodologies to assess data compression, CO_2 mitigation strategies, and the discovery of new knowledge in dynamic systems.

Thesis organization

This work is organized on three major study cases, where we implemented ML to:

- First, reduce microseismic/DAS dimensionality.
- Second, to retrieve unsupervised CO_2 plume visualizations.
- Third, to identify fluid mechanic properties using unsupervised clustering.

This framework aims to explore the use of unsupervised learning on key aspects of subsurface engineering such as big data processing and dynamic reservoir properties.

CHAPTER II

MICROSEISMIC DENOISING AND COMPRESSION USING HIGH-ORDER SINGULAR VALUE DECOMPOSITION [*]

Introduction

Unconventional reservoirs have been one of the primary sources in U.S. for oil and gas production. According to the EIA annual energy outlook (U.S. Energy Information Administration, 2020), unconventionals constitute the main driving force in the growth of oil and gas (O&G) production, representing a key factor in the global energy supply. Subsurface technologies are being continuously applied to gather critical information on the O&G extraction process. Hydraulic fracturing is one of these vital components. Knowledge of hydraulic fracturing is necessary to optimize its design and fracture propagation. Geophysical tools are being used to provide a better understanding of the reservoir and its interaction with hydraulic fractures. Different technologies have been applied for fracturing monitoring such as DAS, DTS, microseismic, VSP, and electromagnetic sensing. Among the various geophysical measurements, our research focuses on the application of microseismic.

Microseismic can be defined as the passive monitoring of small ground displacements associated with seismic waves radiated by small scale fracturing events

---

[*] Reprinted with permission from "Improving Microseismic Denoising Using 4D (Temporal) Tensors and High-Order Singular Value Decomposition" by Gonzalez, K., Gildin, E., and Gibson, R. L, 2021. SPE/AAPG/SEG Unconventional Resources Technology Conference, Copyright 2021 by Gonzalez, K., Gildin, E., and Gibson, R. L.

during hydraulic fracturing. Observations can lead to the estimation of fracture properties (e.g., height, length, azimuth, asymmetry, dip, and complexity) and insights into the fracture propagation (Warpinski, 2009). This indirect measurement has been widely used in the hydraulic fracturing process, due to the direct link between fracture development and the small energy detection. One of the main challenges during microseismic processing is the quality of seismic data, as the Signal-to-Noise ratio (S/N) hinders subsequent analyses. The S/N provides a ratio of the desirable and undesirable seismic energy. Conventional denoising algorithms can be affected by the complexity and assumptions of regularly sampled data (Mandelli, Lipari, Bestagini, and Tubaro, 2019). Another major concern is the computational time and memory requirements. Microseismic is considered to be a big data technology due to the increasing volumes of data. Hence, alternative methodologies need to be applied to manage the data size and quality, since real-time tools are being used for fracturing treatment data interpretation.

Recently, the use of machine learning techniques has increased in seismic processing and data interpretation. Common techniques include deep learning, regressors, classifiers, clustering, and time-series methods. A growing application has been the use of unsupervised learning for dimensionality reduction. Some of the most important algorithms are matrix and tensor decomposition. Multiple studies have been developed using subsurface measurements. Bekara and van der Baan (2007) demonstrated the application of local singular value decomposition (SVD) to enhance seismic signals. Freire and Ulrych (1988) also implemented SVD on VSP processing to separate upgoing and downgoing waves. Furthermore, this application can be extended and applied to higher-

order dimensions. Kreimer and Sacchi (2012) introduced tensor higher-order SVD (HOSVD) for a 4D pre-stack seismic tensor. Microseismic data has also been used as input for denoising and compression (Yatsenko et al., 2019). However, this methodology can be implemented in other sets of data. Afra and Gildin (2016) extended the use of HOSVD for reservoir model parameterization, proving the performance and computational reduction of it. On the other hand, other decomposition techniques have been established on other geophysical measurements such as distributed acoustic sensing (DAS). Brankovic et al. (2021) developed a shifted-matrix decomposition for application to DAS data.

In this research, we propose a novel tensor decomposition workflow to generate the compressed and denoised tensor. We developed a methodology based on the Tucker and high-order singular value decomposition (HOSVD). This framework has the capabilities of working with high-dimensional datasets, that can be compressed, denoised, and used to discover hidden patterns in the data. We first applied HOSVD on a synthetic 3D microseismic tensor and tested it on a field 3D DAS tensor, providing evidence of real-world datasets. This would support the applicability on large datasets with fine spatial and temporal sampling. We also exploit the temporal variation compression by using HOSVD for a 4D microseismic array. The results provided a solution for high-dimensional data processing, reducing computational demands and data size. In particular, it was able to denoise and compressed different sets of data with different measurement principles. Besides, it displayed stability in the 4D HOSVD model by denoising and compressing the tensor without the inclusion of artificial signatures.

Tensor decomposition and unsupervised learning

Tensors are multidimensional arrays that can be defined as a generalization of vectors and matrices to higher dimensions. The tensor can display different orders or numbers of dimensions. Scalars are interpreted as zero-order, first-order as vectors, and second-order as matrices (Rabanser, Shchur, and Günnemann, 2017). A third-order or higher dimension represents tensors of more than two dimensions. A representation of their structure can be observed in figure 1 for vectors, matrices, and higher dimension tensors.

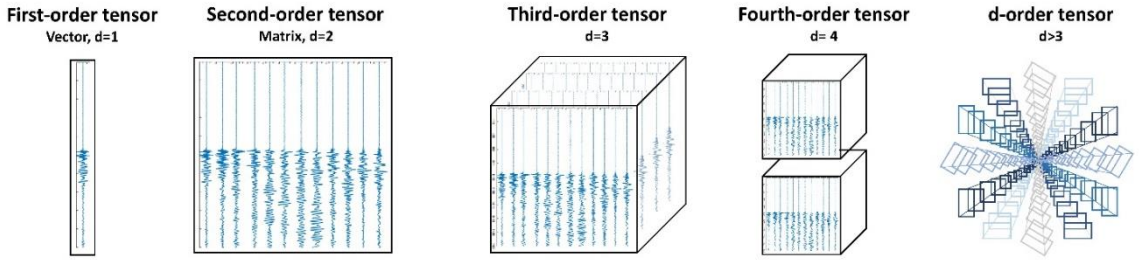


Figure 1. Graphical representation of tensors structure for vectors (d=1), matrices (d=2), third-order tensor (d=3), fourth-order (d=4), and higher-order tensor (d>3)

Tucker decomposition (TD) and higher-order SVD (HOSVD)

Tensor decomposition represents the decomposition of a tensor as a sum of its meaningful parts (Kolda, 2018). This technique has been used for a vast number of applications such as signal processing, neuroscience, and pattern recognition. Different tensor decomposition algorithms have been established, with TD being one of the most popular ones. It can be considered as a generalization of higher-order SVD (singular value decomposition), retaining the main properties of matrix SVD. Specifically, the TD

approach decomposes a tensor into multiple factor matrices and a so-called dense core tensor. In a three-dimension case, the TD can be express as:

$$X_{(N_1 \times N_2 \times N_3)} \approx g \times_1 U \times_2 V \times_3 W \tag{1}$$

$$= \sum_{r_1=1}^{R1} \sum_{r_2=1}^{R2} \sum_{r_3=1}^{R3} g_{r_1 r_2 r_3} \circ \mathrm{u}_{r_1} \circ \mathrm{v}_{r_2} \circ \mathrm{w}_{r_3} = [\![g; U, V, W]\!]$$

Where X represents the original tensor, $g \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ the core tensor and $U \in \mathbb{R}^{N_1 \times r_1}, V \in \mathbb{R}^{N_2 \times r_2}, W \in \mathbb{R}^{N_3 \times r_3}$ the factor matrices (Kolda and Bader, 2007). On the other hand, $\times_1, \times_2, \times_3$ represents the tensor unfoldings for each mode. These unfoldings rearrange the tensors 1-d structures to matrices form which allow us to mathematically obtain a more convenient expression of the tensor (Kolda, 2016).

Different methodologies can be used for computing a TD. They are mainly differentiated on their optimization technique which focus on the minimization and reconstruction of our decomposed tensor. Among the different algorithms, we are going to use HOSVD to reconstruct our original signal. Figure 2 displays the relationship between HOSVD and SVD. The indices of them correspond to the nth element of the tensor on each N vector space.

The optimal solution that we wish to solve is defined as

$$\min_{\hat{X}} \left\| X - \hat{X} \right\|_F^2 \tag{2}$$

being $\hat{X} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$ the compressed sized of the tensor. Moreover, given a relative error (ϵ) we can choose the projection ranks of the core tensor to satisfy the required error.

$$\|X - (g \times_1 U \times_2 V \times_3 W)\| \leq \epsilon \|X\| \quad (3)$$

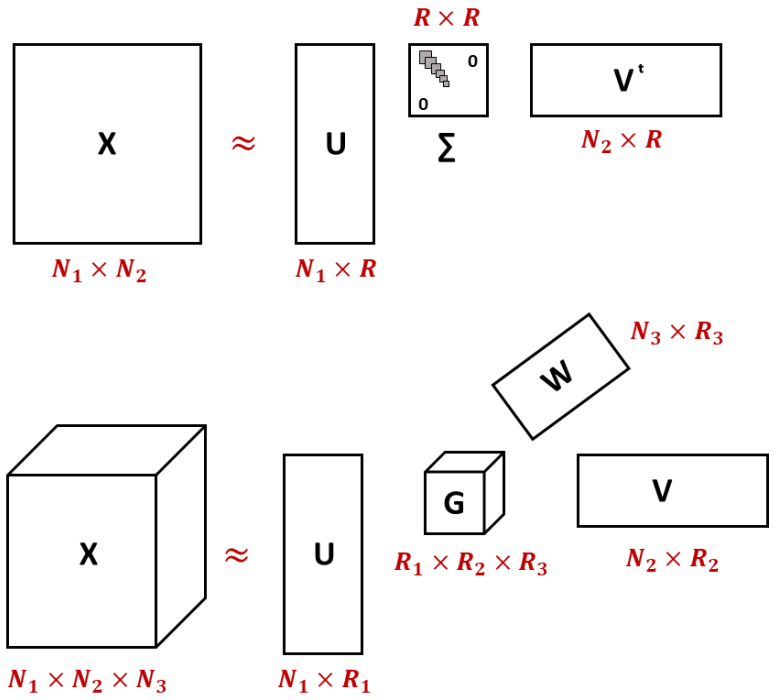


Figure 2. Schematic of the singular value decomposition (SVD) and higher-order SVD (HOSVD). The top image corresponds to the SVD structure which is defined as the three factor matrices (U, V, and ∑). The bottom represents the HOSVD scheme of the three factor matrices and (U, V, W) and one dense core tensor (G).

An example of the use of HOSVD is displayed in figure 3 for a reservoir model parameterization. The input data set is a 3D tensor of stacked state snapshots. These snapshots correspond to the pressures and saturations from the reservoir model simulation. The application of HOSVD allows to maintain and reparametrize the most important state features without losing essential information. The resulting tensor provides a low-dimensional representation of the state snapshots (Afra and Gildin, 2016).

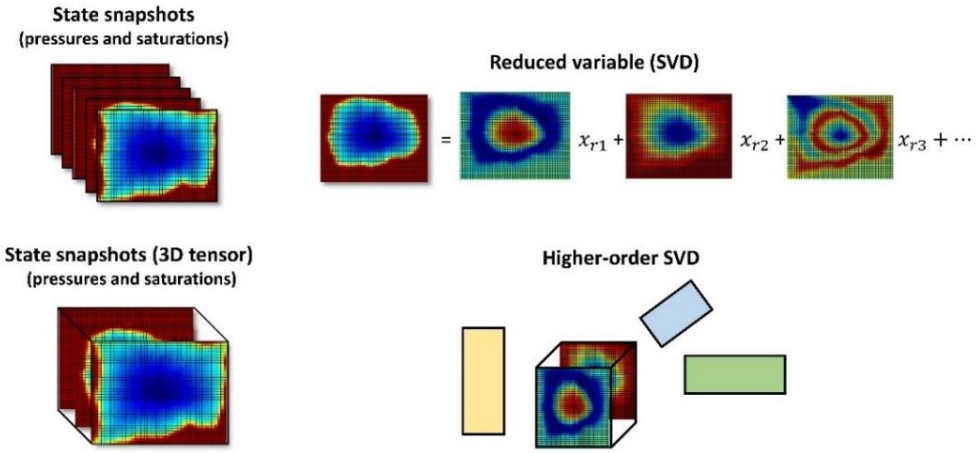


Figure 3. Representation of state snapshot SVD with their respective reduced variables, and HOSVD implementation for 3D tensor form.

HOSVD on microseismic data

In signal processing, the implementation of HOSVD has been used as a compression and denoising algorithm (Yatsenko et al., 2019; Iqbal et al., 2018; Kreimer & Sacchi, 2012). Microseismic data can be arranged into multiple tensor forms due to the different features that could be extracted. For instance, the dimensions of a higher order data tensor could include event location, magnitude, geophone location, or recording time.

Denoising is a major concern during seismic acquisition, as large amplitude noise makes it difficult to detect and process seismic signals. The HOSVD algorithm can act as a noise filter by retaining only the most impactful values, those representing desired waveforms. This approach will increase the S/N ratio and remove the undesired noise (Figure 4). Furthermore, microseismic data sets from DAS are big data; hence, HOSVD is attractive for its ability to compress large, multidimensional data volumes.

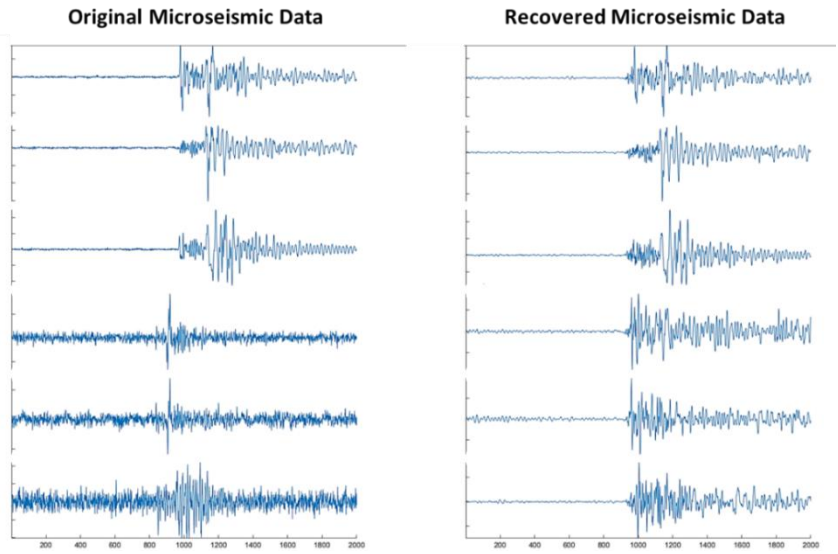


Figure 4. Microseismic traces of original and compressed tensor of 2000 milliseconds. Left: Original microseismic with noisy signals at the last three traces. Right: Recovered microseismic with noise-free signals.

Methodology

Data description

This work investigates the use of HOSVD for microseismic denoising and compression. We also evaluate the interpretability of it by analyzing the reduced tensor components. According to Vesselinov et. al (2019), the decomposed tensors may provide insights into hidden patterns in the data. This would introduce the ability to discover fracture propagation signatures and their importance.

As shown in figure 5, the dataset is comprised of a third-order synthetic microseismic tensor. We used synthetic microseismic signals with different levels S/N ratio. This complexity helped us to incorporate data complexity and noise levels. The tensor modes correspond to 2000 milliseconds recording time, 12 geophones, and 24

location variations. The events location variation can be separated into x-, y- and z-components, creating a final tensor of shape 2000x12x72.

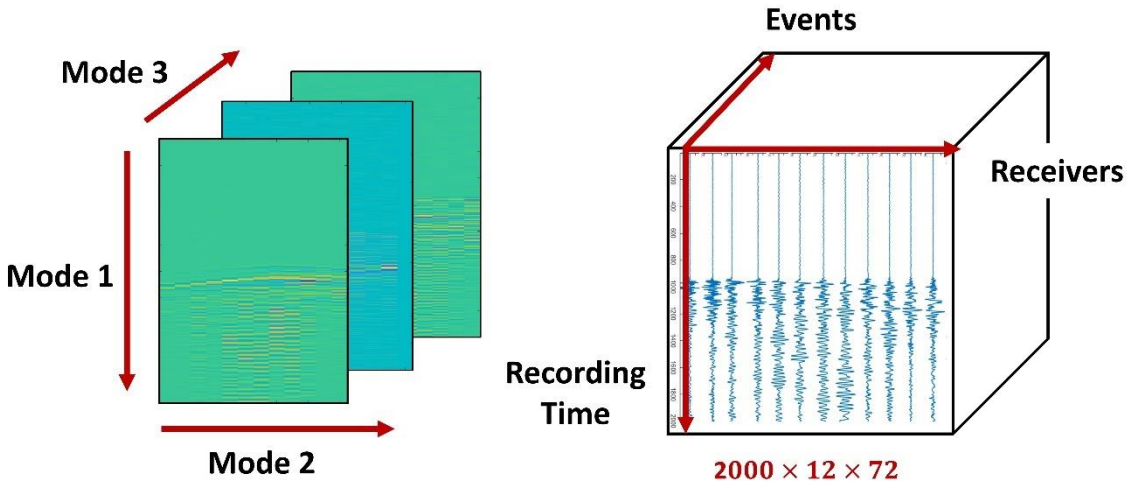


Figure 5. Third-order microseismic tensor with dimensions of 2000x12x72 where mode 1 represents the milliseconds, mode 2 the receivers, and mode 3 the event locations. This tensor will serve as the input data on the HOSVD algorithm.

Furthermore, we tested the proposed workflow to field DAS data. The application to real data allowed us to validate the use of HOSVD for compression and denoising. For this research, we used field DAS from the Marcellus Shale to provide subsurface information during well stimulation. The complete DAS data comprises the 28 stimulation stages of the horizontal well MIP-3H (Figure 6).

The tensor was formed by using 43 SEGY files from stage 9, which corresponds to 20,000 samples in time and 493 traces. Figure 7 shows the data structure with their corresponding dimensions.

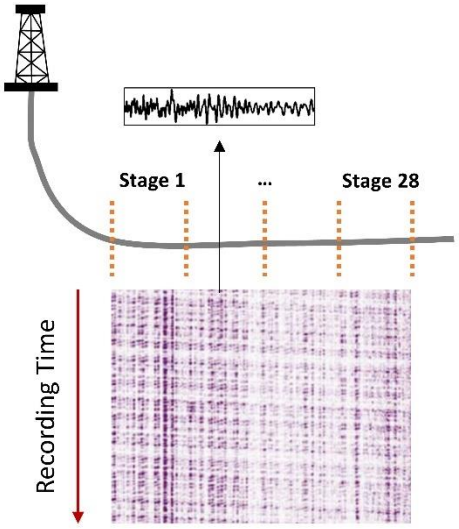


Figure 6. DAS measurements from the horizontal well MIP-3H. These recordings correspond to the different stimulation stages, covering a total of 28.

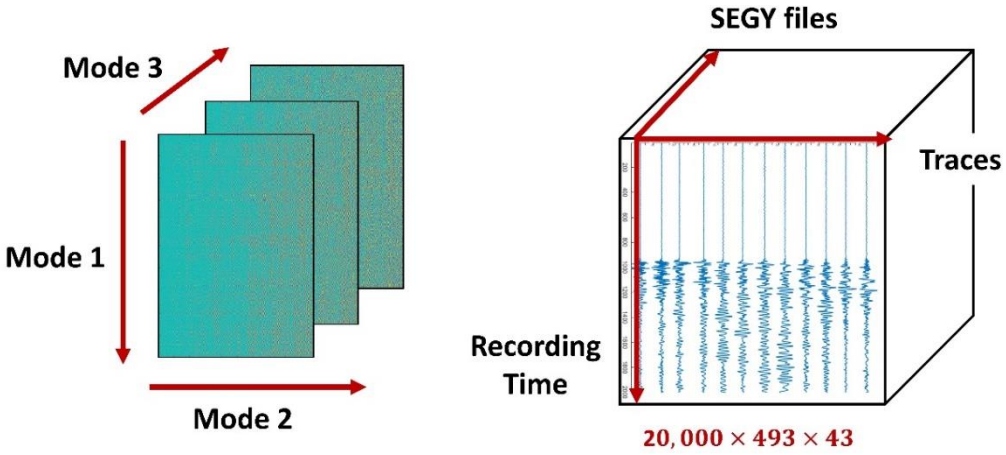


Figure 7. Third-order DAS tensor of 20,000x493x43 dimensions. Mode 1 represents the milliseconds, mode 2 the traces and mode 3 the SEGY files.

Workflow for tensor decomposition

In this section, we provide the proposed workflow for HOSVD implementation. Moreover, we analyze the decomposed tensor using the factor matrices and core tensor. The main steps are displayed in figure 8 and can be divided into five key stages.

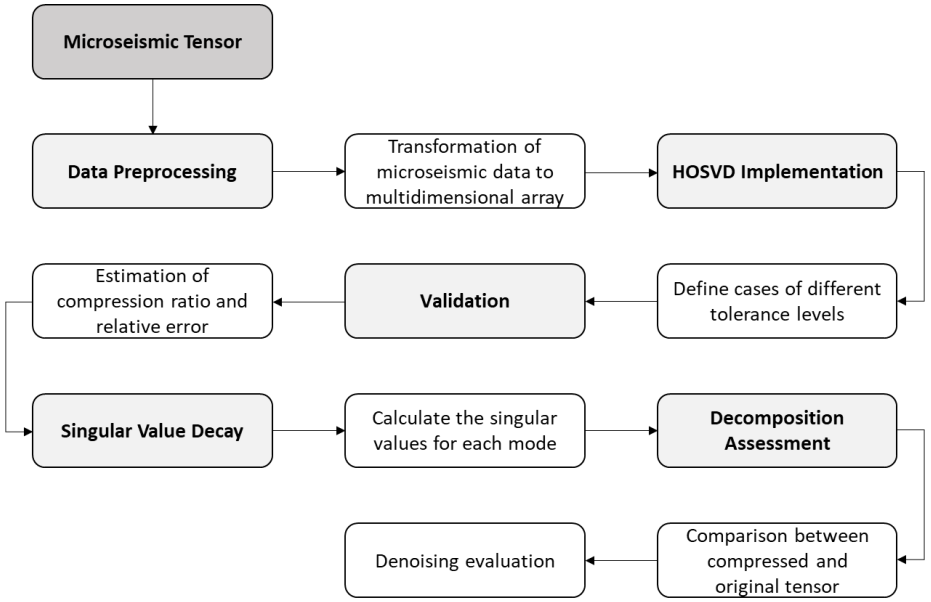


Figure 8. The proposed workflow of HOSVD for a microseismic tensor. The methodology was divided into five stages: data preprocessing, algorithm implementation, validation, estimation of singular values, and decomposition assessment.

Stage 1. Perform data preprocessing: The preprocessing corresponds to the tensorization of the microseismic data. The dataset was constructed to include a combination of noisy signals and noise-free data. The 3D array was later converted into a tensor format for the HOSVD application.

Stage 2. Implement higher-order SVD: We use ST-HOSVD (sequentially truncated-HOSVD) from the SNL tensor toolbox (Bader et al., 2021) to construct the reduced tensor. The algorithm establishes the four orthogonal factor matrices and core tensor according to a tolerance level or requested relative error. In addition, it displays the elapsed time being this a feature for the subsequent analysis of the results.

Stage 3. Estimate compression ratio and relative error: To validate the HOSVD results, we consider the estimation of the compression ratio, relative error, and computing time. The normalized relative error can be expressed as

$$\text{Relative error } = \frac{\left\|X - \hat{X}\right\|}{\|X\|} \tag{4}$$

Where X is the original tensor and X^ the compressed one. On the other hand, the compression ratio (CR) can be computed as

$$\text{CR } = \frac{\#\ total\ bits\ to\ store\ \hat{X}}{\#\ total\ bits\ to\ store\ \ X} \tag{5}$$

Stage 4. Calculate the singular values of each mode: We estimated the normalized singular values to obtain the variation and decay of each tensor mode. The normalized singular values are generated by applying the Frobenius norm.

Stage 5. Compare the reduced tensor with the original: For the analysis of model results, we developed a comparison between the compressed tensor ($\hat{X}$) and the original one (X). We also examined the noisy seismic traces to verify the denoising trade-off.

4D (Temporal) Tensors and HOSVD

Microseismic tensors can be constructed for the analysis of event temporal variations. During the fracturing process, the monitoring of the growth of hydraulic fractures is essential to determine their fracture propagation. In addition, it may reveal the interaction between them and their relationship with natural fractures.

The temporal microseismic decomposition requires an extension of the 3D HOSVD methodology. Under this scenario, we developed a synthetic case of a 4D microseismic tensor with 1000 milliseconds recording time, 10 geophones, 3 event

locations, and 5 the time variations of different reference times. This latter dimension was created to represent the different elapsed times of a single microseismic event. For instance, this could represent the elapse time from 1 hour to 5 hours. The structure of the extended tensor of size (1000x10x3x5) is display in figure 9, being this the input data for the HOSVD algorithm. Here we focus on the stability of the algorithm and its ability to denoise 4D tensors without introducing artificial seismic artifacts.

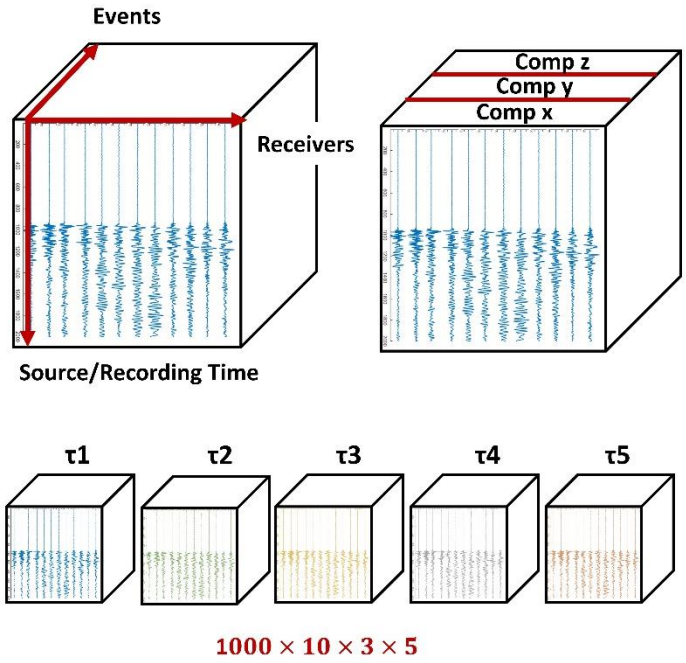


Figure 9. 4D microseismic tensor of dimensions 1000x10x3x5. The dimensions represent the source/recording time in milliseconds (1000), receivers' number (10), components of event location (3), and temporal variation or elapse time (5).

The temporal HOSVD design is illustrated in figure 10. The decomposition is going to provide four factor matrices and a 4D core tensor. This extra factor matrix will be directly related to the time variation of each slice in the source/recording time. The 4D HOSVD is computed as:

$$T_{(N_1 \times N_2 \times N_3 \times N_4)} \approx g \times_1 U \times_2 V \times_3 W \times_4 F = [\![g; U, V, W, F]\!] \quad (6)$$

Where T corresponds to the 4D tensor, $g \in \mathbb{R}^{r_1 \times r_2 \times r_3 \times r_4}$ core tensor, and $U \in \mathbb{R}^{N_1 \times r_1}$, $V \in \mathbb{R}^{N_2 \times r_2}$, $W \in \mathbb{R}^{N_3 \times r_3}$, $F \in \mathbb{R}^{N_4 \times r_4}$ the factor matrices.

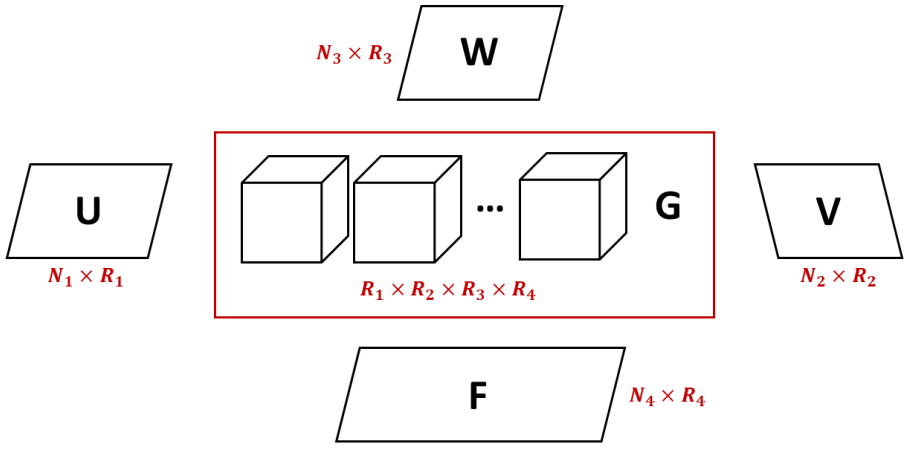


Figure 10. Illustration of 4D HOSVD design which displays the decomposition of a tensor into four factor matrices (U, V, W, and F) and a core tensor of four dimensions.

To investigate the applicability of 4D HOSVD, we tested multiple cases for the compression and denoising of the synthetic microseismic array. In addition, we developed a time variation analysis by subtracting the first (e.g. 1 hour) and last (e.g. 5 hours) elapse times of the compressed tensor. This would represent the variation of the data in time. The difference of it can potentially reveal physical insights into the time fracture propagation.

Results and discussion

Implementation of 3D-HOSVD

Synthetic microseismic data

To evaluate the most appropriate HOSVD compression, we tested four distinctive cases at different tolerance levels (0.2, 0.02, 0.002, and 0.0002). Each one of them is

associated with a specific compression ratio and execution time. A key aspect under any compression technique is the relation between the compression and the removal of significant data. As illustrated in figure 11, we can distinguish in cases one and two the removal of important parts of the synthetic waveforms. The compression ratios were higher (98.8% and 96%) due to the established tolerance level.

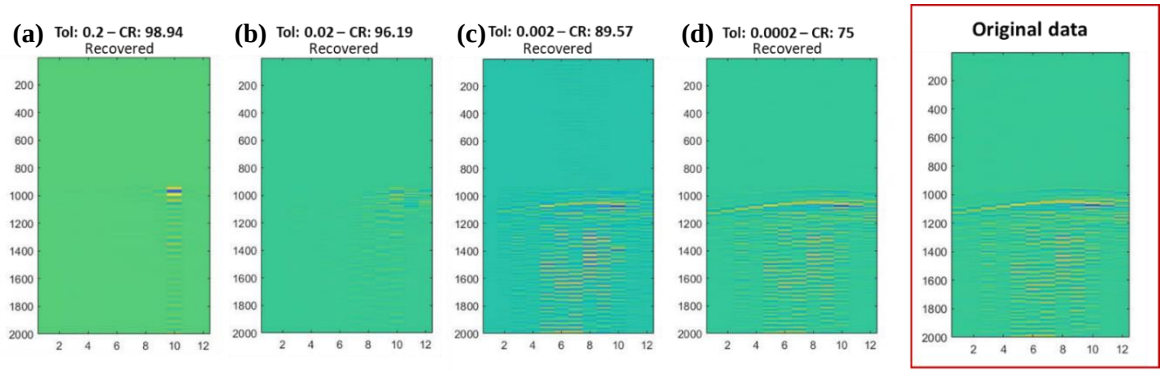


Figure 11. Compressed HOSVD results for four different cases in a single microseismic event. Case 1 (a): tolerance level (TL)=0.2 and compression ratio (CR)=98.95%, case 2 (b): TL=0.02 and CR=96.2%, case 3 (c): TL=0.002 and CR=89.57%, and case 4 (d): TL=0.0002 and CR=75%. The best compressed tensor was achieved on case 4 by prevailing the most significant signatures.

Based on these four cases, the most accurate compression resulted given a tolerance level of 0.0002 (Figure 11. d). The number of samples was reduced from 1,728,000 to 431,303 while achieving a compression ratio of 75% (Figure 12. b). This compression supports the tolerance level and HOSVD implementation by prevailing the noise-free microseismic data. Furthermore, we estimated the singular values for each mode (Figure 12. a). This represents the decay of strength in the recording time, receiver, and event location dimensions. This mode variation reveals the most impactful values that contributed to the original microseismic tensor. For instance, it can be seen that for mode 1 the significant components are arranged up to 50 observed at the inflection point.

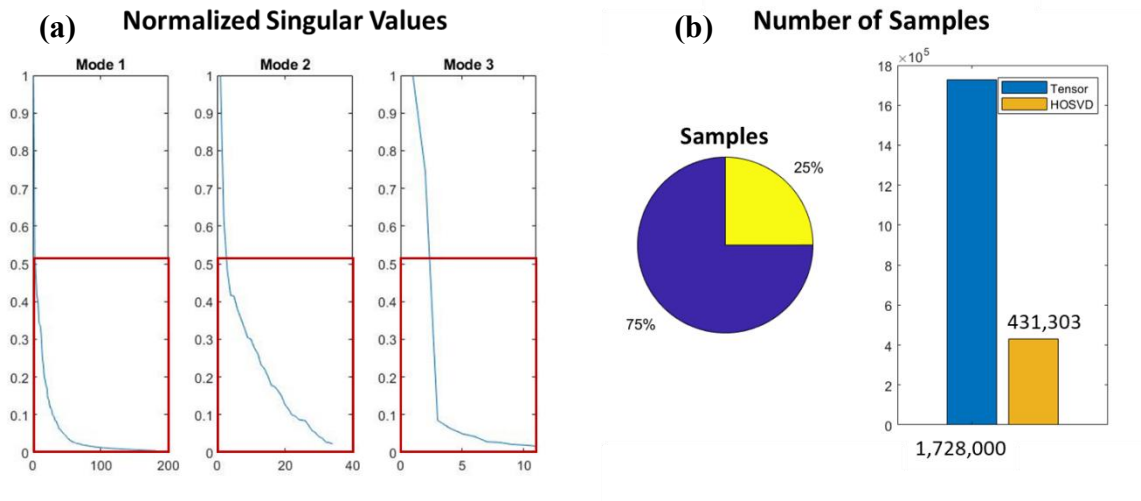


Figure 12. Left: Normalized singular values which were used to analyze the singular values variation and decay. Right: The number of samples reduction from original (1,728,000) to compressed tensor (431,303) under a compression ratio of 75%.

We also conducted a comparative study with the core tensor dimensions, relative error, and compression ratio (Table 1). For all four cases, the values are displayed in the pair plots of figure 13. In plot number one (Figure 13. a), we mapped the relationship between core tensor dimensions and compression ratio. For the higher compression ratio (96%) we obtained a core dimension of 9x3x1. On the other hand, for the lowest compression (75%) we reach a 181x34x11 core tensor. It should be noted that the core tensor of dimensions equal to the original tensor describes the tensor without any reduction. In the second plot, we validated the correlation between the compression ratio and relative error (Figure 13. b). A relative error of 0.0002 displays a compression ratio of 75%, and an error of 0.2 a compression of almost 100%. Thus, a large value compression ratio correlates with a high relative error and a low dimensions core tensor.

Table 1. Relative error, compression ratio, and core tensor dimensions for the four tested cases.

	Relative error	Compression ratio (%)	Core tensor dimensions
Case 1	0.2	98.94	9x3x1
Case 2	0.02	96.19	32x18x2
Case 3	0.002	89.57	84x28x5
Case 4	0.0002	75	181x34x11

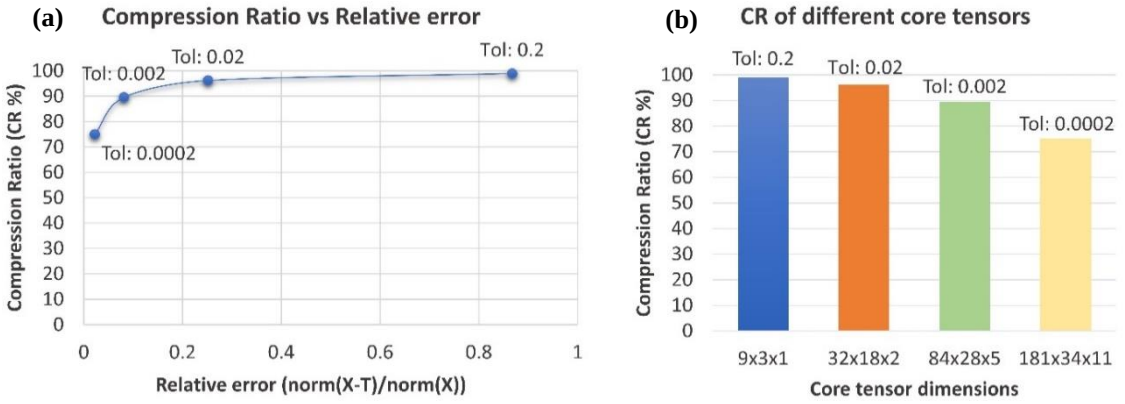


Figure 13. Pair plot of core tensor dimensions and compression ratio (a): A relationship can be established between a high CR and low core dimensions. Pair plot of relative error and compression ratio (b): A high CR corresponds to a high relative error.

In terms of the execution time, the corresponding computational times can be found in table 2. The time increases with the decrease of compression ratio, obtaining a value of 4.910 seconds for the selected compression of 75%.

Table 2. Execution time with their respective compression ratios and relative error. A lower CR will indicate a lower elapse time.

	Compression Ratio (%)	**Computational Time (s)**	**Relative Error**
Case 1	98.9	4.620	0.850
Case 2	96.2	4.639	0.224
Case 3	89.57	4.876	0.082
Case 4	75	4.910	0.023

Field DAS data

With our developed workflow, we also tested the algorithm on a field 3D DAS data. The DAS data are comprised into recorded SEGY files. Each file contains 493 traces with 20,000 samples per trace. We tested three tolerance levels (0.01, 0.02, and 0.05) and analyzed the compression and denoising of each case. As shown in figure 14, we captured a compression ratio of 44.77% for case 1, 70.2% for case 2 and 93.35% for case 3. The results indicate case 2 as the most appropriate one due to the stability of both compression and denoising. Case 1 also manages to accurately display a compression and denoising but in a lower degree while case 3 implemented excessive compression, eliminating most of the significant information.

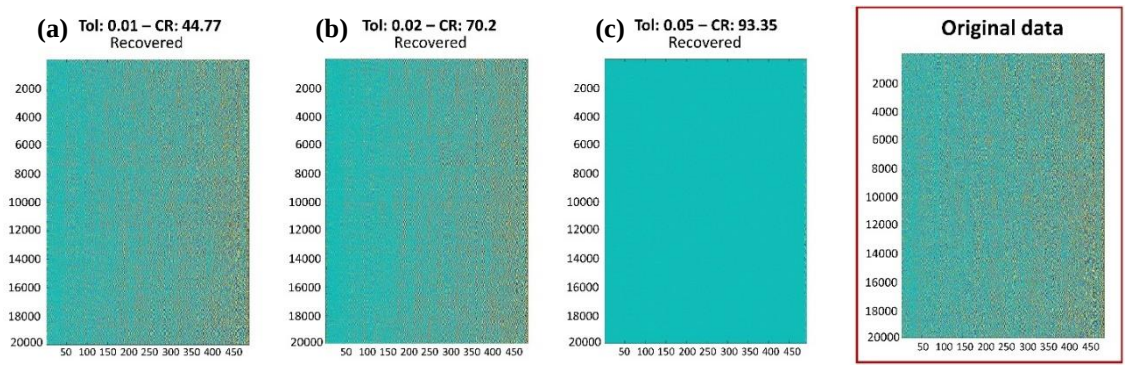


Figure 14. HOSVD results for field DAS data under three tolerance levels. Case 1 (a): tolerance level (TL)=0.01 and compression ratio (CR)=44.77%, case 2 (b): TL=0.02 and CR=70.2%, and case 3 (c): TL=0.05 and CR=93.35%. Original data is display in the red rectangle to understand the impact of HOSVD.

This behavior can also be observed in the reduction of the number of samples. We acknowledge the relationship between the compression ratio and number of samples. A shallow compression will display a lower number of reduced variables; conversely, a deep compression is going to retain a few original samples. In figure 15, we compared this relationship, confirming a strong reduction of samples at increased compression ratio. The best model (70.2%) achieved a reduction from 423,980,000 to 126,337,269 samples.

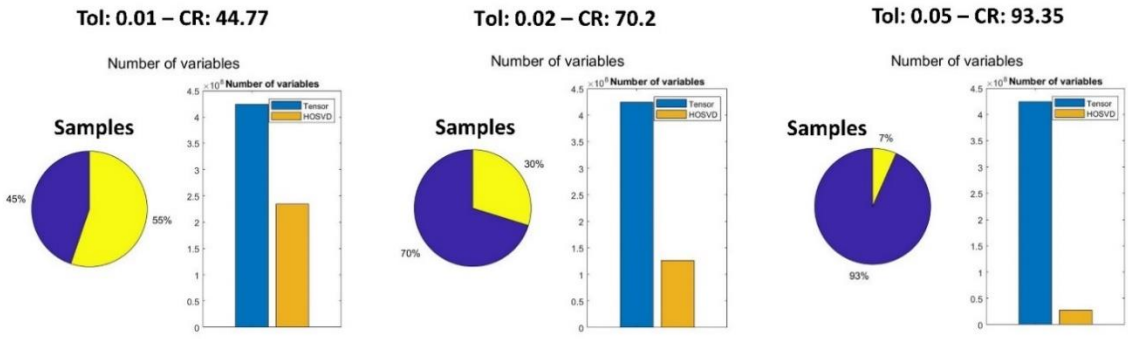


Figure 15. Number of reduced variables at three compression ratios. Case 1 seems to be compressed 45% of samples, case 2 a 70% and case 3 a 93%.

In addition, we determined the singular values on each mode to analyze their decay. The decay can be observed in figure 16. a. which corresponds to the singular values at a compression ratio of 70.2%. This variation may reveal a strong decay with low number of samples. Furthermore, in figure 16. b. we validated the compression and denoising of the DAS traces by comparing the original and recovered results. We can infer that the compression did not include artificial seismic artifacts while denoising the data.

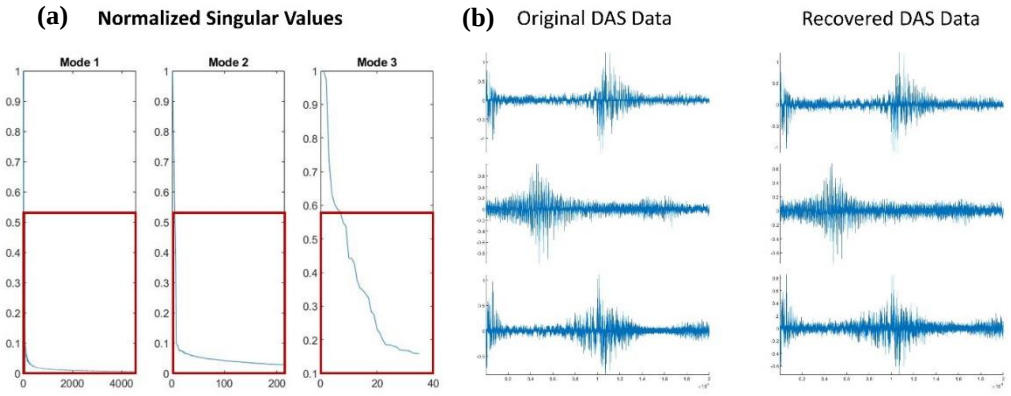


Figure 16. Left: Normalized singular values for best HOSVD compression. Right: Original and recover DAS traces after HOSVD, which allow us to confirm the stability of results.

Temporal variation HOSVD on a synthetic microseismic tensor (4D-HOSVD)

With the proposed workflow, we extended the application to a synthetic 4D temporal tensor. The tensor comprises recording time (1000), receivers (10), event location (3), and elapsed time (5). The P-wave and S-wave arrival are shown in figure 17 in respect of the milliseconds recording/source time and receivers' location. For the HOSVD implementation, we had to fix the core dimensions due to the sensitivity of the algorithm. This provided compression and denoising without any introduction of seismic artifacts. The final compression was achieved at 82.8% with a tolerance level of 0.079.

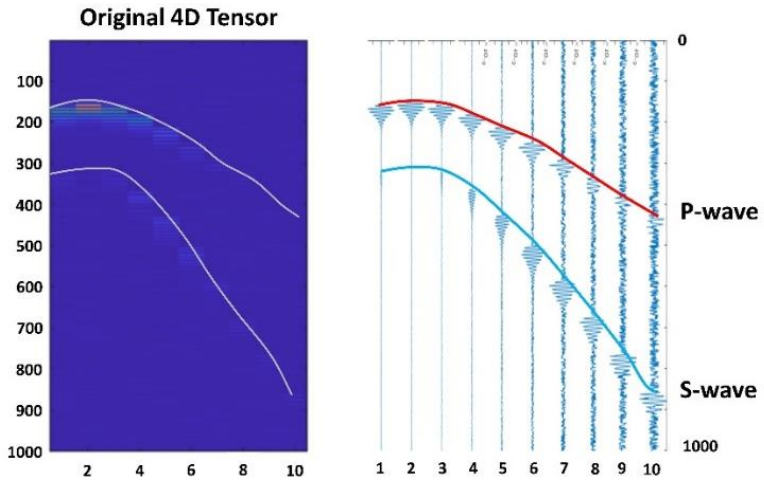


Figure 17. Left: temporal microseismic tensor of fourth-order at ten receivers' location. Right: microseismic traces and their respective p- (red) and s-wave (blue) arrivals.

Figure 18 shows the multidimensional decomposition of temporal variation τ=1 and τ=5. At τ=1 (Figure 18. a) we can acknowledge the compression while maintaining the original seismic information. For τ=5 (Figure 18. b), we validated the compression and denoising by retrieving a noise-free signature. Furthermore, in the recovered seismic image the algorithm was able to provide the p- and s-wave arrivals at almost all receivers.

To obtain insights into fracture propagation, we perform the reference time difference between τ=1 and τ=5. A preprocessed step had to be performed in order to compare them. For this, we aligned/shifted the event signals of τ1 and τ5 at the same p- and s-wave arrivals. The estimated difference was later compressed using the suggested HOSVD algorithm. The results are displayed in figure 19 with the τ=5 compressed tensor. The reduced [τ5- τ1] tensor did not display a substantial difference from the recovered τ5. However, the decomposition maintained the seismic content and denoising.

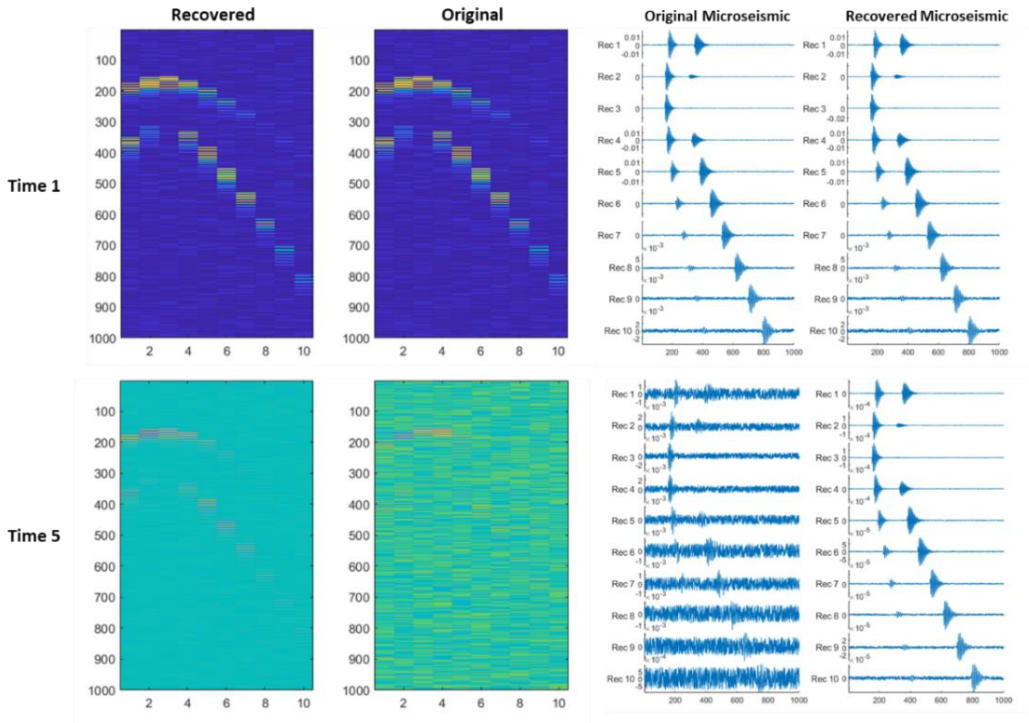


Figure 18. HOVD results on the reference times 1 and 5. Time 1 (a) illustrates the accuracy of compressed tensor by performing the compression without altering the seismic signals. Time 5 (b) displays the capability of HOSVD for denoising seismic traces on a 4D tensor.

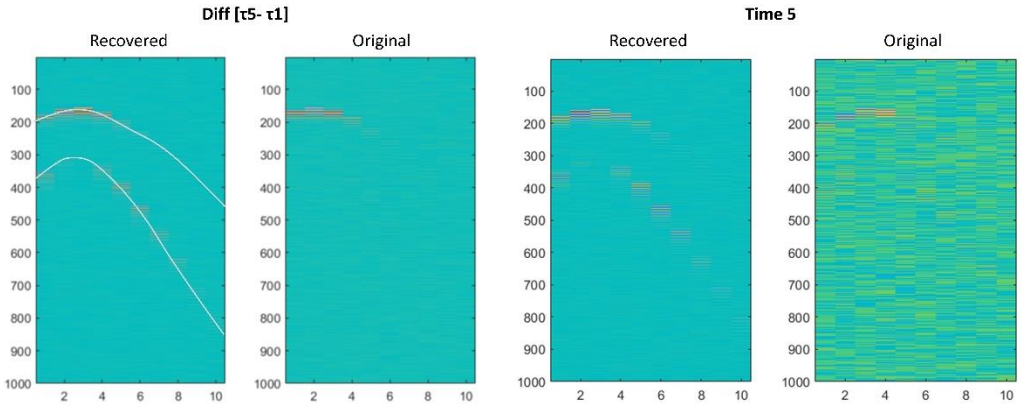


Figure 19. Time variation analysis of the subtraction of first and last time, and time 5 results. The difference aims to display the fracture propagation and behavior of temporal datasets. The solid white lines depict the arrivals of p- and s-wave.

We also carry out an analysis of the seismic traces to assess the HOSVD results (Figure 20). A reduction of seismic amplitudes was observed due to the amplitude differences between τ1 and τ5. Lastly, we recognized the lack of fracture propagation information by noticing the same P- and S-wave arrivals for both times.

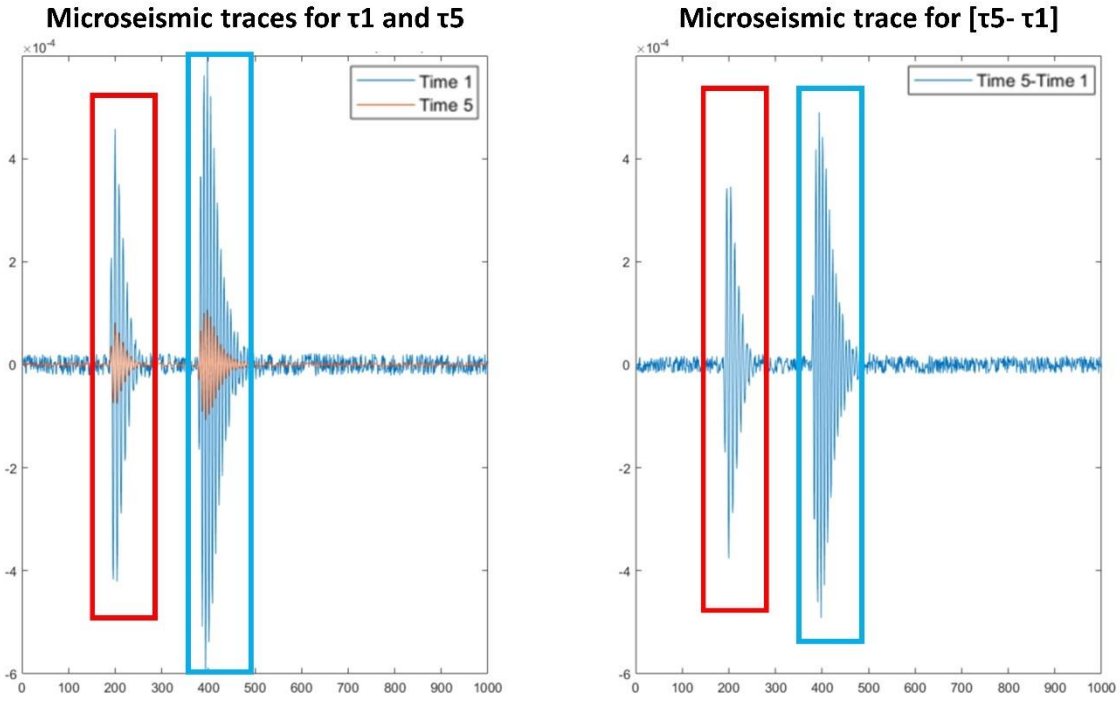


Figure 20. Left: Microseismic traces of time 1 (τ1) and 5 (τ5), being the blue line τ1 and orange τ5. Right: Microseismic trace of τ5 and τ1 subtraction which illustrates the reduction of seismic amplitudes. Red represents p- and blue s-wave arrivals on both plots.

The compression for all conditions revealed the accurate application of HOSVD. This served as an indicator of the model stability and demonstrated the use of 4D-HOSVD for denoising and compression.

Assumptions and Limitations

This work is based on the following assumptions:

- The core tensor is assumed to contain the most impactful signals of microseismic and DAS compression. The architecture for this needs to be investigated for the retrieval of them.
- The synthetic 4D microseismic tensor was constructed to represent the temporal behavior of a single microseismic event at constant geological conditions.
- Synthetic cases for 3D and 4D arrays incorporate a combination of noisy and noise-free signals at different receivers' locations.
- The signals of the 4D tensor are aligned in time for all reference times at both p- and s-wave first arrivals. The key idea was to compare the variation of each temporal change.

Final Remarks

In this work, three multilinear tensor decompositions were generated for 3D microseismic, 3D DAS and 4D temporal microseismic tensors. We developed a novel decomposition workflow for the denoising and compression of seismic traces that enable us to improve microseismic and DAS detection. The following conclusions can be drawn from this research:

1. Based on the HOSVD results, the developed model can be safely applied to multidimensional arrays without adding artificial artifacts. This model displayed stability by preserving the most significant signatures.
2. The HOSVD implementation allowed a reduction of the computational memory and time. The compressed model demonstrated the improvement of big data analysis by cutting down the execution time from months to seconds. In addition, it

provided a fast approximation of conventional seismic filters which involve a high computational cost.

3. The model provided applicability on other datasets for unconventional subsurface measurements, handling data sizes of 4 to 5GB.

4. Mathematical operations can be applied on multidimensional tensors to obtain insights into the data. The proposed 4D tensor only displayed an amplitude decrease which corresponds to a postseismic relaxation along the fracture.

The HOSVD model established the possibility of using tensor decomposition as an unsupervised learning method through the analysis of their decomposed parts. It holds great promise for feature extraction, unsupervised clustering, and feature importance.

CHAPTER III

UNSUPERVISED LEARNING VISUALIZATION OF CO_2 PLUME CONTENT DURING CARBON SEQUESTRATION

Introduction

Geological carbon storage represents one of the main technologies for the reduction of anthropogenic CO_2 emissions. A large amount of carbon dioxide has led to an increase in global temperatures at an average of 0.87 °C. Concerns for the environment have impacted energy production, demanding rapid solutions at low carbon emissions. Numerous industries are developing technologies and mitigation strategies to tackle their high concentration of CO_2. And geosequestration could be part of the solution.

The role of carbon storage represents a potential solution for net-zero emissions. It involves the injection of CO_2, at supercritical conditions, into an underground geological formation. Various carbon storage sites are suitable, being the most important: oil and gas reservoirs, CO_2 enhanced oil recovery (EOR), unused saline aquifers, coal seams, and coalbed methane. In particular, EOR and saline formations have been displayed as promising solutions due to their global storage and economic potential. EOR is considered a key methodology for CO_2 reuse, while saline aquifers the potential to store one Mt of CO_2 per year.

According to the IPCC (2018), over 700 Gt of CO_2 needs to be removed from the air to stabilize the carbon content. This is equivalent to emissions from 143 million vehicles for one year. Hence, new CO_2 sequestration projects are needed to be developed

to address this goal. In addition, investments in carbon capture and storage have increased, expecting to worth $3.5 billion by 2025 (Markets and Markets, 2020). This could establish the geosequestration and carbon capture industry as one of the most attractive ones.

Different aspects of CO_2 storage are needed to consider. The overall process involves the selection of a suitable geological site, transport, injection, surveillance, verification, and assessment of long-term CO_2 storage. Nevertheless, due to the complexity of each process, large uncertainties are encounter. Major initiatives include the analysis of risk management for the development of leakage detection systems, work for effective CO_2 monitoring to understand the behavior of injected carbon, ground motion to predict the extent of pressure change and potential induced seismicity and focus on geochemical research to analyze the chemical interaction between CO_2 and in-site fluids.

To investigate the lifecycle of CO_2 injection and storage, geophysical surveys are acquired. Geophysical signatures are essential to establish a proper reservoir characterization, safe long-term storage, and monitoring of evolving CO_2 plume. Moreover, it can provide a detailed reconstruction of the subsurface at different resolution levels. Integration of diverse monitoring techniques would also enhance the accuracy of results. This includes experimental measurements from other areas such as core, geochemical and geomechanical analyses.

Over the last few years, artificial intelligence has been recognized as a powerful technology to address climate change. Machine learning has served as a tool to assist the ongoing subsurface monitoring and injection process. This is also a well suitable technique to address the high uncertainty for the long-term process and their spatial-temporal

evolution. Areas with potential growth involve computer vision and unsupervised learning approaches. The necessity for obtaining insights into CO_2 processes are essential to the feasibility of carbon storage, and this could be solved by the application of machine learning techniques. In addition, methodologies have been proposed to manage the urgency of rapid CO_2 knowledge, such as real-time visualizations, real-time forecasts, and rapid prediction models (National Energy Technology Laboratory, 2020).

In this work, we proposed an unsupervised clustering model to identify signatures of CO_2 content using field surveillance data. For this, we developed a novel and reliable machine learning methodology where we implemented a multi-level clustering design using unsupervised learning and computer vision techniques. This is a major difference between current models that so far have focused on training models with synthetic data and a known target. The approach is also free of assumptions since this is a data-driven model that does not require human intervention. The model is developed to deal with the unknown features, unknown response, and unbalanced dataset. Its accuracy was evaluated through statistical analysis and the implementation of different clustering algorithms to validate the consistency of CO_2 saturation levels.

CO_2 plume location and movement

CO_2 injection and plume migration

The injection of CO_2 is performed under CO_2 supercritical conditions. Carbon dioxide has the property to behave as both liquid and gas at critical pressures (1070 psi) and temperatures (87.8 °F). As the CO_2 encounters the supercritical setting, the CO_2 starts

behaving like gas with liquid density. This increases the CO_2 storage capacity at low toxicity and environmental impact.

The process of injection occurs through non-corrosive injection wells where materials need to be carefully picked to maintain the well integrity. Subsequently, the injected CO_2 is continuously monitored to assess the reservoir response and regular pressure changes. Pressure and temperature gauges are usually installed to surveille the injection progress and identify any potential well problems (Rackley, 2010).

The CO_2 plume is defined as the volume of carbon dispersed in the reservoir. CO_2 can be immiscible or miscible in presence of other fluids. For instance, water and CO_2 are immiscible while CO_2 and natural gas are miscible. Under immiscible fluids, CO_2 needs to be injected at a higher-pressure rate to displace the in-situ reservoir content. Once the injection stops, it migrates to the base of the caprock since density is lower than reservoir fluids.

Over a longer period, CO_2 can be trapped by capillary forces preventing the movement of the carbon molecules. Furthermore, it starts dissolving the CO_2 due to their chemical interaction with formation water, allowing more storage space in the rock; however, the dissolution could be slow depending on the CO_2 and water ratio. These chemical reactions can also modify the porosity and permeability of the formation. Finally, mineral reactions may occur decreasing the size and connectivity of the pores. In short, CO_2 could be trapped into four mechanisms: structural, capillary, solubility and mineral.

Need for CO_2 plume monitoring

Monitoring of CO_2 is vital to track the movement and plume behavior, and it aims to confirm the injection process and CO_2 prevalence. This is a vital element for risk assessment and mitigation strategies, being a key factor and main challenge. For instance, the identification of early leakages could lead to the prevention of groundwater contamination. It can also assess the effect of geomechanical changes and induced seismicity since the injection of CO_2 (at high-pressure rates) can enhance the movement of the subsurface, increasing the possibility of leakage.

Another important aspect is the prediction of the CO_2 plume. Plume visualization can provide useful information on movement and location. More specifically, the surveillance data permits the validation and modeling of CO_2 growth, allowing to predict the behavior of long-term CO_2 storage. Besides, carbon storage becomes riskier over time. Data-driven frameworks could be established to reduce this uncertainty to more appropriate results. In particular, it can help to select appropriate methodologies for plume interpretation.

Monitoring techniques

Geophysical data analysis has a crucial role in carbon storage processes. A diverse number of geophysical technologies can be used in the lifecycle of CO_2 monitoring. A wealth of knowledge already exists in this area due to their current application in the oil and gas industry, which was rapidly expanded to geosequestration.

Monitoring and verification of CO_2 movement are classified according to their specific surveillance goal. The main commercial measurements involve injection well

monitoring, plume location and movement, ground displacement survey, and leakage detection (Rackley, 2010). For CO_2 plume, the measurement techniques are summarized in table 3, to understand the variability of techniques. Time-lapse seismic is considered the most effective tool due to the high contrast of CO_2 acoustic impedances. Pre- and post-injection seismic are commonly acquired to provide an image of the change of fluids over time.

Table 3. Geophysical methods for plume location and migration according to their physical principles.

Plume location and migration	
Type	Monitoring technique
Seismic	Time-lapsed seismic
	Crosswell seismic
	Vertical seismic profile
	Microseismic
Gravimetry	Time-lapsed gravimetry
Electric and electromagnetic	Electric resistance tomography
	Crosswell resistivity
	Electric spontaneous potential
Remote sensing	Satellite interferometry
	Airborne electromagnetic

Crosswell seismic imaging

Crosswell seismic is another effective tool to monitor the supercritical CO_2 movement. This technique involves the use of downhole seismic sources and receivers' array (figure 21). Both are placed in adjacent wells to transmit and capture high-frequency soundwaves. As the source and receivers move, the process is repeated multiple times to obtain an image of the subsurface properties. The high frequency of the data provides detailed information of thin reservoirs, from 3 to 33 feet, at interwell distances of 33 to 330 feet.

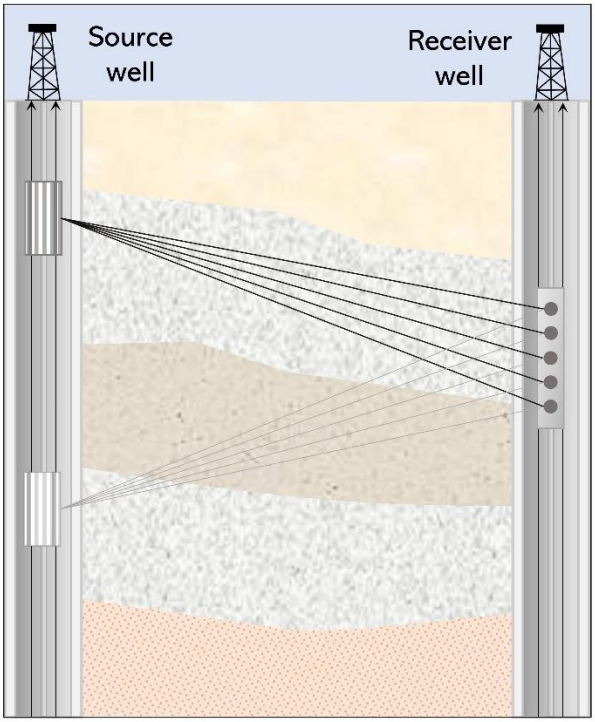


Figure 21. Crosswell survey scheme on a source-receiver profile where the transmissions of sound waves are captured from source to receiver well. This process is repeated as the seismic source and receivers move.

The data can be processed into two fundamental measurements: tomography and reflection imaging. Reflections are used to gather a detailed reflection image of the

formation while for crosswell tomography a velocity map of properties. And velocities are often the best indicators of fluids movement. As the CO_2 is injected, the velocities decrease aiding to identify the injected CO_2. The difference of time-lapsed tomography may provide a direct image of the CO_2 plume where only the areas of velocity change are going to be displayed. It also captures the degree of velocity change which can be associated with the concentration levels of CO_2.

Methodology

SECARB Cranfield project

The Southeast Partnerships (SECARB) Cranfield Project was a commercial CO_2-EOR program located at Cranfield field in Mississippi. This project was designed to establish the feasibility of long-term CO_2 storage at low risk. It also serves to set up strategies for stacked storage, where EOR infrastructure can be used to inject CO_2 above and below EOR operations (Hovorka, 2013). The project started with the CO_2 injection in the fluvial sandstones of the Tuscaloosa formation on an unused saline aquifer.

The research was divided into five stages called: 1) phase-two, 2) high volume injection test, 3) detailed area of study (DAS), and 4) near-surface observatory area (Hovorka et al., 2011). The SECARB project conducted effective subsurface monitoring to evaluate the behavior and permanence of carbon dioxide. Different monitoring techniques were used according to their specific research goal. The Cranfield project focused their analyses on three main goals: risk management, CO_2 plume prediction, and pressure impact.

Time-lapsed seismic, electromagnetic, and tracer chromatography measurements were techniques used for CO_2 plume prediction. For environmental assurance, pressure temperature, and groundwater and soil gas analysis were frequently applied (Hovorka et al., 2011).

Dataset description

The dataset consists of two time-lapsed crosswell tomographies. Three wells were used at an interwell distance of 229 feet for the first profile and 98 feet for the second profile. Figure 22 displays the schematic representation of the well's location and acquisition design. The depth of interest corresponds to a range of 10,400 and 10,510 feet where the supercritical CO_2 condition can be met.

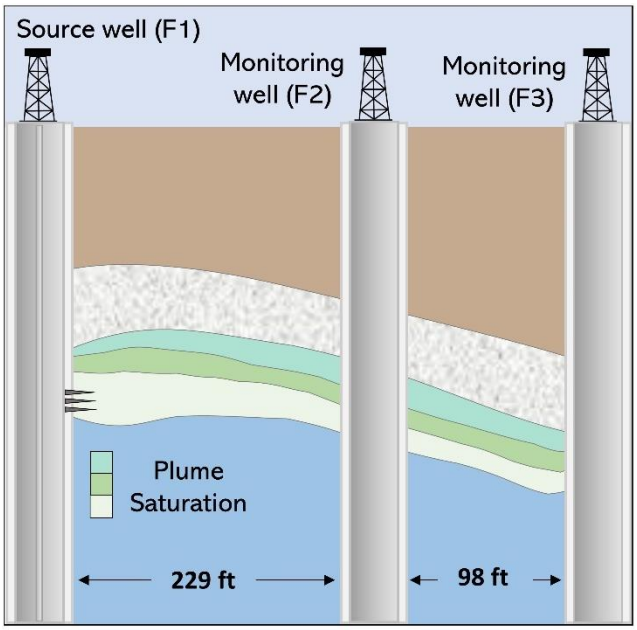


Figure 22. Schematic representation of the study area showing the side view of the crosswell survey.

These profiles were acquired before and after the injection stage using a 10-level hydrophone array. The data was recorded on both profiles at a time difference of 10

months. During the seismic acquisition, the source and receivers were switched to capture the high frequency of soundwaves. The data acquired was processed into two components, reflection imaging, and tomography. The latter provided a seismic velocity map of the subsurface properties and in-situ fluids.

Due to the high correlation of seismic velocity and CO_2 content, the difference between the pre- and post-injection tomography was used. This resulting difference revealed the change of velocity under CO_2 injection and provided an image of the CO_2 plume. Figure 23 displays the percentage change of velocity from the crosswell tomography difference. Values close to zero represent non to low CO_2 content while values close to 14 a higher CO_2 concentration.

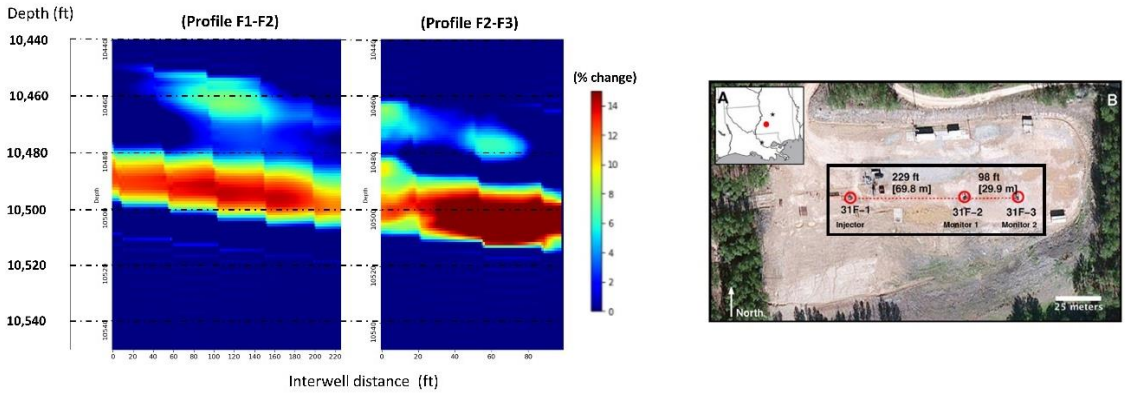


Figure 23. Left: Crosswell tomography image obtained after the data processing of pre- and post-injection profile. Right: Study site of Cranfield field using one injection well (F1) and two monitoring wells (F2 and F3).

Workflow for CO_2 plume visualization

A novel workflow approach was proposed for the visualization of CO_2 content. The workflow is presented in figure 24 where we aim to discover signatures of CO_2 using

unsupervised clustering algorithms. The clustered results reflected the levels of saturation ranging from low to high concentrations.

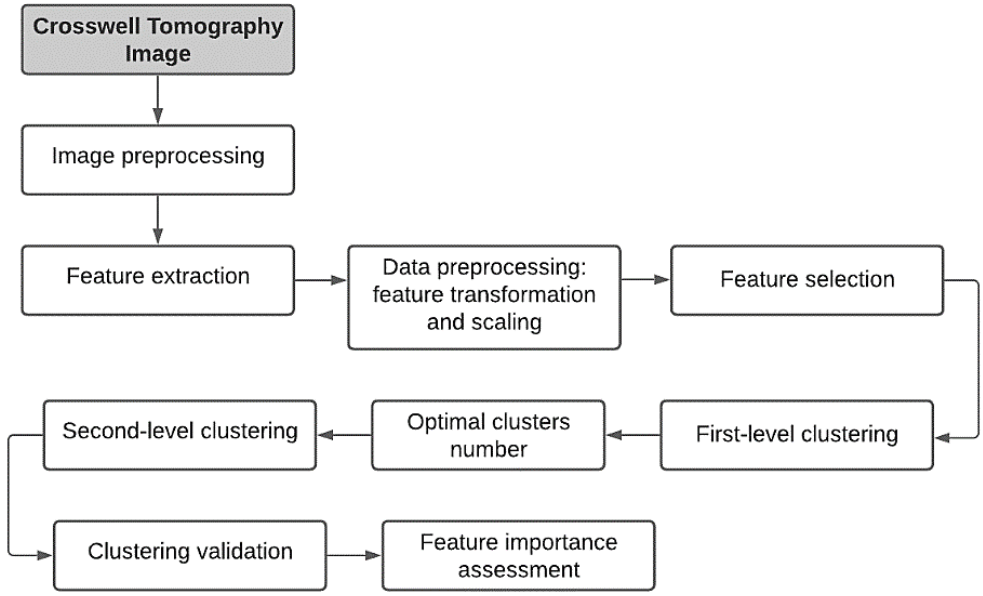


Figure 24. Flowchart of the unsupervised learning clustering for CO_2 plume visualization.

The first step was to process and convert the time-lapsed tomography data into a change of velocity image. The image covers a depth between 10,440 and 10,550 feet, and an interwell distance of 229 feet and 98 feet. The quantification of velocity changes was represented with pixel values to quantify the CO_2 content (figure 25). Pixels values close to zero were linked to low-velocity changes while values between 50 and 255 to medium and high-velocity changes. The images comprise a size of 42749 pixels for profile 1 and 54776 for profile 2.

Due to the nature of the data and unknown features, pixel intensity served as the input to extract relevant features. The number of features was a major factor for the success and accuracy of the machine learning models. Fourteen features were extracted where

each one of them represents a specific image characteristic of the velocity change. Table 4 compiles the extracted features and their description.

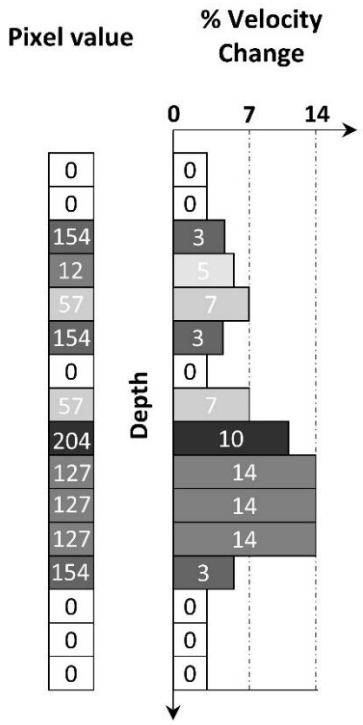


Figure 25. Representation of crosswell seismic tomography using pixel values. The percentage velocity change was determined by the difference between pre and post CO_2 injection. Higher values correspond to higher CO_2 concentrations while values of zero of no-CO_2.

With this new higher-dimensional dataset, the features had to be preprocessed before applying any clustering algorithm. First, robust scaler and power transformation were implemented to achieve a Gaussian distribution and to standardize features in a unique range. This was a vital step to obtain accurate results using similar features weight. Histograms and scatter plots were used to assess this processing stage. Moreover, features needed to be analyzed based on their impact. Statistical tests were performed to select the appropriate set of features, selecting those that displayed low multi-collinearity and high statistical importance. For this study, only nine features were selected:

- Pixel intensity
- GLCM ASM
- GLCM correlation
- GLCM dissimilarity
- Local binary patterns
- Wavelet transform
- Fast-Fourier transform
- Edges
- Hxx (Hessian matrix)

Table 4. Brief description of extracted features from pixel intensity input.

Feature	**Description**
Gray-Level Co-Occurrence Matrix (GLCM)	Statistical analysis of spatial relations between pixels. Statistical methods include contrast, dissimilarity, homogeneity, energy, correlation, and ASM.
Fast-Fourier transform	Transformation of the image from spatial to the frequency domain. Low and high pass filters permit to pass certain image frequencies.
Linear binary pattern (LBP)	Texture operator which labels pixels based on the intensity of the central point
Sobel (Edges)	Gradient of pixels intensity for edge detection. It captures changes of intensity.
Wavelet transform	Time-frequency analysis for selection of suitable frequency band. It is commonly used to remove noisy signals.

Hessian matrix	Second-order derivative of the Gaussian kernel for region detector. It is applied in the Hxx, Hxy, and Hyy direction.

Unsupervised clustering

The lack of ground truth represents a major difficulty for the assurance of a safe, long-term storage site. For the development of real-time visualization tools and leakage system, the discovery of signatures becomes an essential part. In this study, unsupervised clustering was implemented to discover hidden patterns of geophysical data. These measurements serve as indicators of CO_2 content and plume visualization.

The clustering process begins by analyzing the preprocessed dataset. Imbalance in the data was found to be a critical problem for clustering algorithms. Therefore, a novel approach was designed to handle the disparity of class samples. This procedure involved a multi-level clustering to first distinguish CO_2 from Non-CO_2 and to later obtain levels of saturation. Algorithms such as K-means, agglomerative and meanshift were deployed to group the carbon content according to their similitude. Each of them is based on different clustering assumptions, allowing us to evaluate the consistency of results.

Methods for evaluating the clusters

A key aspect during the application of clustering algorithms is the number of clusters. The optimal number was calculated using techniques such as elbow plot, silhouette score, Davies-Bouldin, and Calinski-Harabasz index. These algorithms were applied on each clustering level to evaluate the performance of the different number of clusters. The best one is defined according to each metric scoring range with the purpose to obtain a denser and well-separated number of clusters.

The silhouette score (Rousseeuw, 1987) is given as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad -1 \leq s(i) \leq 1 \tag{7}$$

Where $a(i)$ represents the average distance of each point on the same cluster and $b(i)$ the average distance to the nearest other cluster. The best performance, a score equals to 1, is achieved under lower distances of $a(i)$ and higher distances of $b(i)$. This implies a lower dissimilarity within clusters and a higher dissimilarity between them.

The Davies-Bouldin index (Davies & Bouldin, 1979) computes the average similarity within clusters and between. The mathematical formulation is defined as:

$$R_{ij} = \frac{S_i + S_j}{M_{ij}}, \qquad \bar{R} = \frac{1}{N} * \sum_{i=1}^{N} \max(R_{ij}) \tag{8}$$

S_i and S_j corresponds to the average distance between each point and their respective centroid cluster, and M_{ij} the distance between the cluster's centroids. The optimal number will be the one that minimizes the similarity between clusters ($\bar{R}$).

The Calinski-Harabasz index (Calinski & Harabasz, 1974) is characterized as the following equation:

$$s = \frac{\left[\frac{\sum_{k=1}^{K} n_k \left||c_k - c\right||^2}{K-1}\right]}{\left[\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left||d_i - c_k\right||^2}{N-K}\right]} = \frac{\left[\frac{BGSS}{K-1}\right]}{\left[\frac{WGSS}{N-K}\right]} \tag{9}$$

It represents the ratio of within-cluster (WGSS) and between-cluster sum of square (BGSS). n_k is the number of points per cluster, and c_k and c the cluster and global

centroids. A higher value indicates a denser and well-separated cluster, being this the most optimal one.

Clusters were also validated through the similarity measure of different clustering methods. K-means, agglomerative and meanshift clustering were compared using the adjusted rand index and homogeneity metric. Both of these techniques measure the clustering results by comparing the assigned labels of each cluster. In this case, a value of one indicates a perfect match.

Finally, clusters were analyzed to evaluate the CO_2 distribution and features importance. A frequency histogram was carried out to investigate clusters distribution and levels of CO_2. For the feature importance, different statistical tests were applied. The first test is the ANOVA or analysis of variance F-test which calculates the ratio between the variance of the group means and within-group variances. A large value of it is going to be an indicator that their distribution is unique. A second test was performed called mutual information. This analysis estimates the statistical dependence or joint probability of two variables. In addition, Kendall's Tau correlation coefficients were estimated to measure the relationship between clusters and features. This method relies on the association between variables with the non-Gaussian distribution.

We also conducted post-hoc tests to estimate the features' statistical difference between groups or CO_2 levels. The test we implemented was Tukey's honestly significant difference (Tukey's HSD). The evaluation consists of examining the means clusters difference of the most impactful features, where a high difference between them indicates how different the clusters are.

Two-level clustering

Design of multi-level clustering

In this work, a novel methodology was developed to handle the imbalanced nature of the dataset. An analysis of features was performed where signatures were linked to a specific range of values. Inconsistency can be observed on clustering algorithms since they cannot find a clear pattern in the data. The main issue relies on the fact that an imbalanced dataset introduces bias to favor the majority of class samples. Hence, clusters tend to be not well-separated or dense enough at different cluster levels.

This study adopts a multi-level clustering approach by mapping the features to a lower-dimensional space. The essential idea was to develop a two-level clustering model to obtain levels of CO_2 concentration at a higher clustering efficiency. Various clustering algorithms were applied; being K-means the algorithm used for final clustering results. Other cluster techniques such as agglomerative and meanshift were implemented to validate the clustering results from the K-means approach. It's necessary to mention that all of these algorithms display different mathematical assumptions. Hence, the consistency of these methods provides a good representation of a similar data structure.

The first level of clustering was initially performed to differentiate CO_2 content from non-CO_2 (figure 26). Clustering scores were estimated to assess the performance of the results. This CO_2 cluster data served as the input for the second level clustering. A mask was developed to extract and map the features from the cluster of interest. Then, the second clustering was implemented to this new feature space to generate the final clustering of CO_2 levels (figure 26). This to improve the learning and disparity of cluster

samples. The selection of the optimal number of clusters was determined using the second level clustering scores. Some of the methods used were silhouette, Davies-Bouldin, and Calinski-Harabasz index where these scores represent the measurements of the dissimilarity between and within clusters.

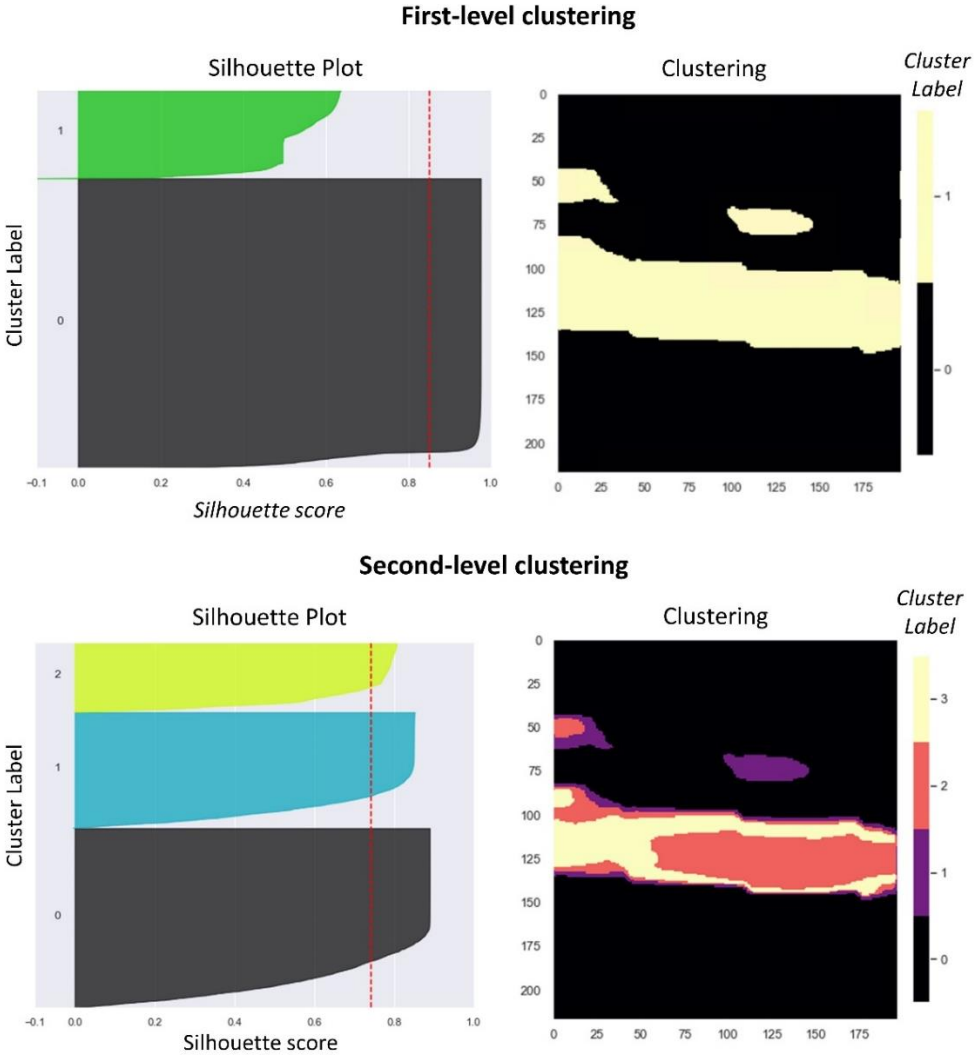


Figure 26. Top: Silhouette plot for two clusters and K-means first-level clustering. Clustered results correspond to no-CO_2 (cluster 0) and CO_2 content (cluster 1). Bottom: Silhouette plot for three clusters and K-means second-level clustering for first-level cluster 1

Traditional K-means clustering was performed to analyze the impact of an imbalanced dataset. Internal clustering scores were also applied to determine the most

suitable number of clusters. The improvement between both of them can be observed through the distribution of clustered data.

Results and discussion

Validation of clustering models

We validate the clustering results by evaluating the density and separation of clusters. To assess the clustering performance, silhouette scores, Davies-Bouldin and Calinski-Harabasz index were computed at the two clustering levels. The coefficients indicate the degree of similarity between and within clusters based on mathematical relationships. This provides a useful measure for the prediction of CO_2 saturation levels. A smaller value of Davies-Bouldin indicates a better performance while for Calinski-Harabasz and silhouette a high value.

At the first level, two clusters were established to differentiate CO_2 from non-CO_2. As shown in table 5, silhouette scores were close to one with a value of 0.85. The Davies-Bouldin index also confirmed the partition by obtaining a value close to zero, indicating accurate segregation of the CO_2 content. These results represent the performance of the clustering, obtaining dense and well-separated clusters of the CO_2 and non-CO_2.

With the first level of CO_2 content, a second clustering was performed. We utilize the same performance algorithms to evaluate the model efficiency and the optimal number of clusters. The scores were computed at 10 clustering levels to choose the most appropriate number. The results displayed a better performance for three clusters. The silhouette and Davies-Bouldin scores achieved values of 0.74 and 0.32, confirming the

need for a second-level approach (table 5). Calinski-Harabasz was also estimated to show a good agreement with the other two clustering scores.

Table 5. Silhouette scores, Davies-Bouldin score, and Calinski-Harabasz index for first and second-level clustering.

Profile	**Clustering level**	**Score**		
		Silhouette	*Davies-Bouldin*	*Calinski-Harabasz*
F2-F3	First-level clustering	0.85	0.43	160666
	Second-level clustering	0.74	0.32	59656
F1-F2	First-level clustering	0.79	0.68	450364
	Second-level clustering	0.68	0.30	72783

For this work, the three clusters represent the three levels of CO_2 saturation ranging from low to high CO_2 concentrations. A total of four clusters were established, providing an image of the CO_2 plume after nearly 10 months of injection.

Traditional clustering vs. two-level clustering

In this section, we compared the traditional clustering and proposed multi-level clustering. The intent of it was to highlight the impact and improvement of this novel approach. Traditional clustering consists of a one-level partitioning of the extracted data. Four clusters were predefined to analyze the clustering behavior and compare it with the proposed methodology.

Figure 27 shows the clustering results for the traditional and two-level k-means. The results are reasonable for the Non-CO_2 cluster; however, the levels of CO_2 displayed

high discrepancies. Differences between them can be attributed to the majority of the Non-CO_2 class. For the CO_2 levels, the traditional clustering presented clusters with low separation and density, resulting in a higher uncertainty of the CO_2 location. Hence, the multi-level clustering display a better approach for regions with high CO_2 content.

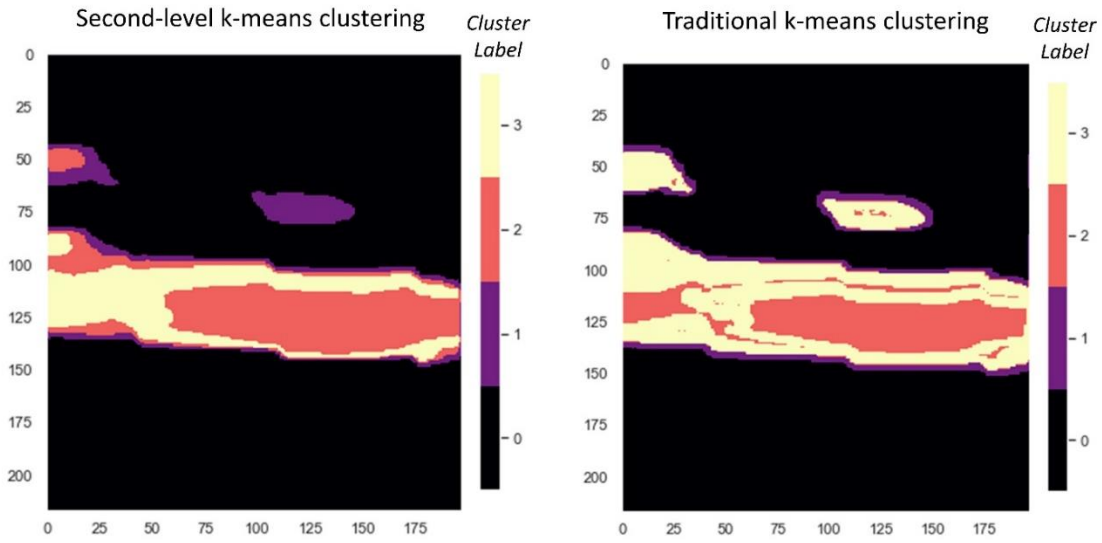


Figure 27. Left: k-means second-level clustering. Right: "traditional" k-means clustering where clusters were established with no previous cluster data.

Comparison of two-level clustering using various clustering methods

A comparative analysis was conducted to confirm the consistency of cluster labels using k-means. As shown in figure 28, three clustering algorithms were used on the two-level approach. The study consists of implementing similarity scores between different clustering methods. To evaluate the k-means CO_2 labels, agglomerative and meanshift clustering were applied. Mean-shift aims to cluster data points based on the discovering of the modes in a data distribution. Meanwhile agglomerative intends to group the samples on their similarity, and recursive clusters merge

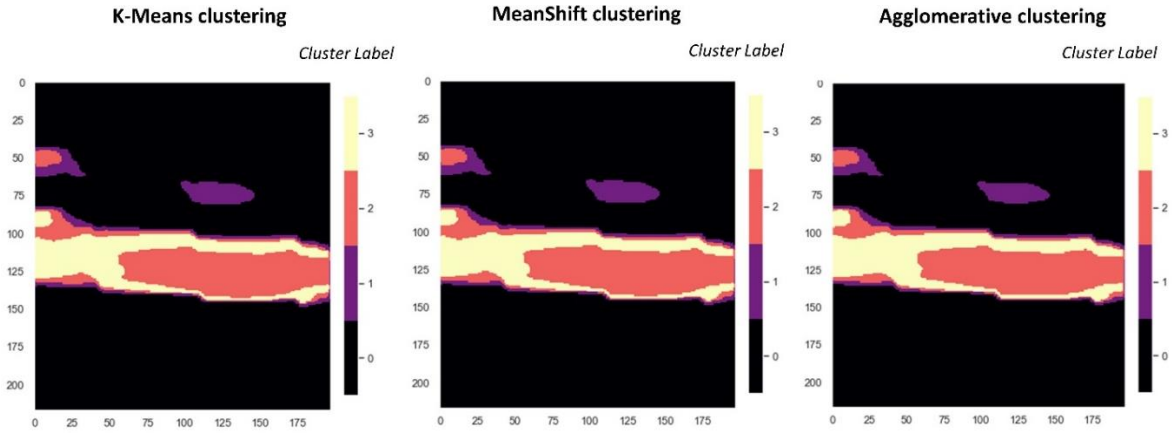


Figure 28. Spatial clustering using the two-level k-means, mean-shift, and agglomerative clustering. The similarity of each clustering algorithm reinforces the consistency and robustness of the proposed workflow.

For the quantitative analysis, we used two pair-wise clustered scores. Table 6 summarizes the results of both adjusted random score and homogeneity. Adjusted random estimates the similarity between two clustering results while ignoring permutations. The homogeneity approach evaluates the clusters labeling based on the principle of clusters containing only a single class. The values range from 0.95 to 0.99, being one a perfect label match. The consistency among these cluster labels confirms the reliability of the model by displaying a high consistency of distinct clustering principles.

Table 6. Comparison of two-level clustering using K-means, meanshift, and agglomerative clustering. Scores close to one indicates a high similitude between two clustering results.

Clustering comparison methods	**Adjusted random score**	**Homogeneity score**
K-means and agglomerative	0.989	0.956
K-means and mean-shift	0.999	0.994
Mean-shift and agglomerative	0.989	0.954

Statistical analysis of spatial clustering

In this section, we perform multiple statistical tests to explore the physical relationship of CO_2 and each cluster label. This will serve as an indicator to physically meaningful labels. Figure 29 shows the histogram distribution of the cluster labels. Based on their frequency, the Non-CO_2 (Cluster 0) constitutes the majority of the cluster samples, accounting for accounts for approximately 76.5% of the data. Clusters 1, 2, and 3 represent regions containing low, medium, and high CO_2 content. Regions containing low, medium, and high are equivalent to 4.5%, 11.5%, and 7.5% of the data, respectively. This behavior can also be observed in the silhouette plots of figure 26, where the thickness of each cluster represents the number of data points belonging to a particular cluster.

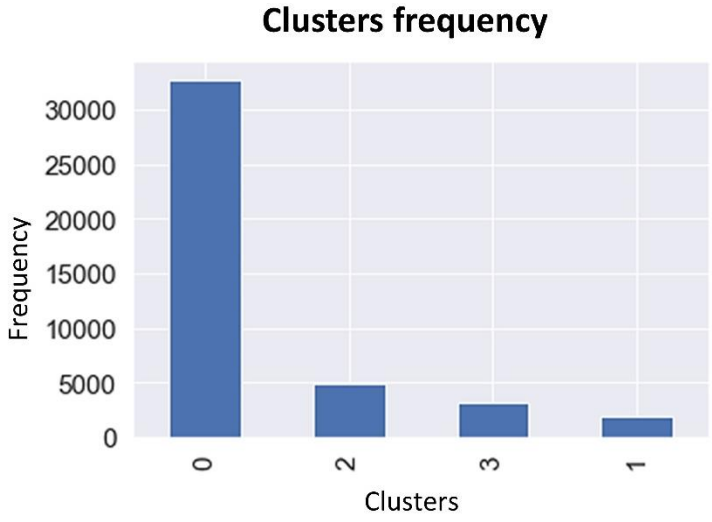


Figure 29. Clusters frequency results using the two-level k-means clustering of the nine extracted features. Cluster 0 is associated with no-CO_2 whereas clusters 1, 2, and 3 with various levels of CO_2 content.

To further investigate the uniqueness of clusters, we estimated the Euclidean distances between each cluster center. The estimation is presented in table 7, where larger distances correspond to larger dissimilarities. According to the results, cluster 3 (high

CO_2) is farthest from cluster 0 (non-CO_2) and cluster 1 (low CO_2). Thus, distinctive characteristics can be observed from each other.

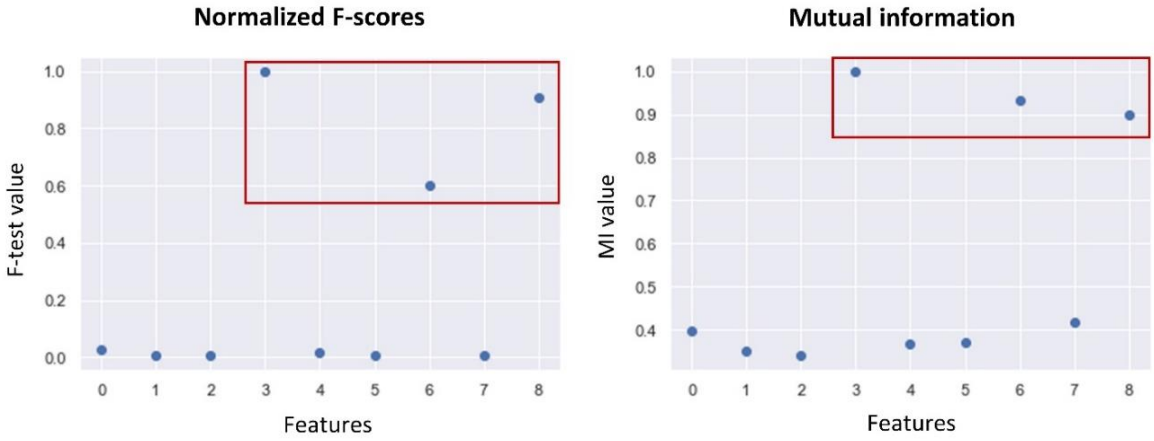
Table 7. Euclidian distances between cluster centers. Cluster "0" represents regions of Non-CO_2, while Clusters "1", "2", and "3" indicate the regions that contain low, medium, and high CO_2 content.

Distances between cluster centers				
Clusters	0	1	2	3
0	0.00			
1	1462.63	0.00		
2	188.97	1273.93	0.00	
3	806.43	656.22	617.74	0.00

Analysis of most impactful and discriminative features

With the extracted features, we aim to analyze their contribution in the clustered data. We first implemented two statistical analyses named ANOVA (analysis of variance) F-test and mutual information values to evaluate the strength of the association between a feature and the clusters (Figure 30). Mutual information quantifies the mutual dependence between a feature and a cluster. In other words, it measures the amount of information obtained about the clusters when a specific feature is implemented in the clustering. Mutual information for a discrete target variable was used in this study, which is based on the entropy estimation of features and target. Meanwhile, ANOVA F-test compares the variances between groups and within groups. This is a specific statistical test that allows the analysis of multiple clusters to determine the features that exhibit significant variation across the clusters. High values of ANOVA F-Test and mutual information indicate that

pixel intensity, fast-Fourier transform coefficients, and wavelet transform coefficients are the most discriminative, informative, and relevant features.



Features number	Feature
0	GLCM ASM
1	GLCM Correlation
2	GLCM Dissimilarity
3	Fast-Fourier transform
4	LBP
5	Edges
6	Wavelet transform
7	Hxx
8	Pixels

Figure 30. Normalized F-test values and mutual information results to determine the most impactful features. Fast-Fourier transform, wavelet transform, and pixels are the signatures that provide most of the clustered information.

Kendall's τ correlation were also estimated to assess the statistical association of ranked data. This non-parametric method was designed for a categorical target, such as classes and clusters. Moreover, this correlation method does not require assumptions of the underlying distributions in data which can be used for non-gaussian distributions. A strong association displays values close to 1 or -1 whereas values close to zero a weaker

relation. It reveals the concordance or discordance of two paired variables. As shown in table 8 the strongest correlations are linked to pixel, fast-Fourier, and wavelet transform.

Table 8. Correlation scores of clusters and features using Kendall's tau for F2-F3 profile. The correlation score displayed wavelet transform and pixels intensity as the most impactful features.

Feature	**Kendall's tau score**
GLCM ASM	0.64
GLCM Correlation	0.62
GLCM Dissimilarity	0.63
Fast-Fourier Transform	0.63
Linear Binary Pattern	0.48
Sobel (Edges)	0.64
Wavelet Transform	**0.96**
Hxx (Hessian matrix)	0.14
Pixels Intensity	**0.95**

For all statistical analyses, it was concluded that pixel, fast-Fourier, and wavelet transform are the features that best describe the spatial CO_2 distribution in the reservoir. In addition to previous tests, Post-hoc "Tukey HSD" was implemented to identify the mean difference between clusters on the most significant features. Table 9 summarizes the statistical mean difference for pixel, fast-Fourier, and wavelet transform. Among the three, Fast Fourier transform is the most significant one. Cluster 3 (high CO_2 content) is the most distinct from both clusters 0 and 1, while clusters 0 and 1 are the most similar.

Table 9. Tukey HSD for post hoc analysis of the significance of the feature for fast-Fourier transform, wavelet transform, and pixels. Mean differences between clusters indicate the significance among them. Cluster "0" indicates Non-CO_2 content, while Clusters "1", "2", and "3" indicate the low, medium, and high CO_2 regions.

Feature	**Clusters being compared**		**Mean difference**
Fast-Fourier Transform	Cluster #	Cluster #	
	0	1	177.02
	0	2	751.82
	0	3	1364.24
	1	2	574.80
	1	3	1187.22
	2	3	612.42
Wavelet Transform	Cluster #	Cluster #	
	0	1	61.25
	0	2	260.83
	0	3	468.20
	1	2	199.57
	1	3	406.94
	2	3	207.36
Pixels	Cluster #	Cluster #	
	0	1	25.72
	0	2	130.41
	0	3	241.26
	1	2	104.70
	1	3	215.54
	2	3	110.85

In addition, we generated boxplots to examine the signature responses per cluster on the high- impact features (figure 31). Among the three features, fast Fourier transform

has the most distinctive values for each cluster. Wavelet transform has a large overlap between clusters 2 and 3 and clusters 1 and 2. On the other hand, pixel intensity only displays large overlap between clusters 2 and 3. This confirms that signatures of impactful features can clearly differentiate levels of CO_2 content.

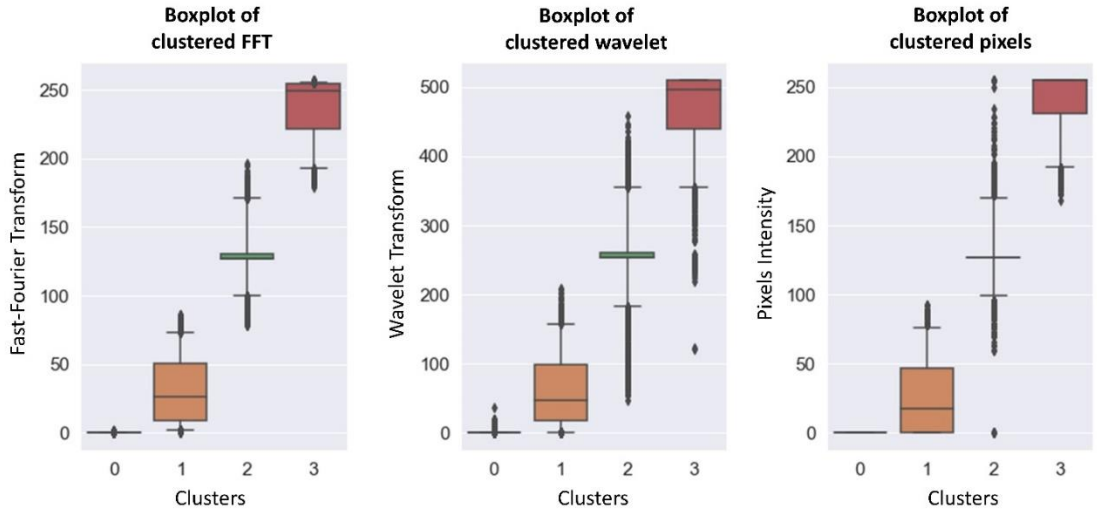


Figure 31. Boxplot of clustered fast-Fourier transform, wavelet transform, and pixels intensity. Boxplots were defined for a low 5th percentile and a high 95th percentile. For fast-Fourier transform, values of 0 were associated with non-CO_2, ~1-80 to low CO_2, ~81-180 medium CO_2, and ~181-257 to high CO_2. For wavelet transform, values of 0 were associated with non-CO_2, ~1-170 to low CO_2, ~171-360 medium CO_2, and ~361-510 to high CO_2. For pixels, values of 0 were associated with non-CO_2, ~1-75 to low CO_2, ~76-175 medium CO_2, and ~176-255 to high CO_2.

Assumptions and Limitations

This study is based on the following assumptions:

- The available dataset corresponds to the derived CO_2 from the crosswell seismic inversion. An accurate approach will involve the use of raw seismic velocities before and after CO_2 injection.

- The model involves the use of only one time-lapse image of CO_2 distributions. A larger dataset will improve our results while incorporating other subsurface scenarios.
- Due to the nature of our data, preprocessing steps were already performed by the seismic processing geophysicist. Therefore, new datasets of raw seismic data will need to include seismic preprocessing.
- The integration of other subsurface measurements would provide a better description of the clustered CO_2 content. Pressure, temperature, and well logs are some of the measurements that could be included in the clustering model.

Final Remarks

In this work, a novel unsupervised learning methodology was implemented to discover patterns of CO_2 levels from field monitoring data. We generated a rapid and reliable plume visualization using geophysical data sets for CO_2 migration assessment. The following conclusions have been drawn from the results:

1. A new workflow was proposed for CO_2 visualization which incorporates the first-time feature extraction, feature selection, and two-level clustering design. This accounts for the unknown features and unbalanced data. In addition, it served to distinguish the variables that best described the CO_2 content, being this a vital step for further analysis.
2. The use of machine learning provided a fast approximation to substitute rock-physics modeling, free of assumptions implemented in physics-driven models.

Thus, the application could be extended to many CO_2 geo-sequestration scenarios at different conditions.

3. Machine learning with unsupervised methods allowed visualizing CO_2 content on regions with similar rock alterations and providing a rapid image of the CO_2 plume. The developed model was intended to represent the growth of plume movement at different time scales.

4. Improved computational performance of this approach was demonstrated using real data from a saline aquifer and existing CO_2-EOR field, reducing modeling/computing time from months to minutes. Real-time visualizations could be implemented to assess the safe long-term storage.

CHAPTER IV

IDENTIFICATION OF FLUID TRANSPORT MECHANISMS USING SPATIAL-TEMPORAL CLUSTERING

Introduction

Subsurface monitoring of geological CO_2 sequestration is critical to ensure storage integrity and efficiency. To identify and map potential pathways of CO_2, appropriate imagining technologies are required. According to Davis et al. (2019), these can be divided into three main categories: surface, near-surface, and subsurface techniques. For subsurface mapping, a set of monitoring tools can be used to establish the CO_2 location by providing a spatial image of the migrated CO_2.

A critical aspect of any CCS project is the real-time location of the injected CO_2 and its migration. Numerous processes can affect the CO_2 plume evolution such as geological heterogeneity, interfacial tension, geological structures, leaky pathways, and gravity forces. For instance, geological heterogeneity can significantly reduce the CO_2 injection capacity due to regions of low permeability. On the other hand, geological structures such as dips can affect the migration of CO_2 by obtaining a condensed supercritical fluid at the bottom of the seal. Hence, the understanding of these dynamic systems is vital to establish safe long-term storage and reducing potential CO_2 leakage.

Machine learning (ML) algorithms are ideal for uncovering hidden relationships of monitoring changes in complex physical behaviors. Haghighat et al. (2013) demonstrated the use of ML using modeled CO_2 leakage and real-time pressure data. Ni

and Benson (2020) developed an unsupervised clustering model to identify capillary flow regimes on five CO_2 coreflooding datasets. Pires de Lima et al. (2019) used convolutional neural networks to model the amount of leaked CO_2 using synthetic pressure and seismic data. Nevertheless, studies of CO_2 systems using ML are often comprised of synthetic datasets and limited subsurface measurements. Moreover, subsurface imaging has to be acquired for an extensive period of time to incorporate the respective spatial plume change. These problems could be solved by using spatial-temporal ML for time-series datasets.

In this chapter, we perform unsupervised clustering to investigate the dynamic fluid properties of subsurface mapping. We aim to identify spatial-temporal patterns of plume migration to assess the feasibility of carbon storage and the processes affecting the trapping. The model is developed using electrical resistivity tomography (ERT) from the field SECARB project.

Crosswell electrical resistance tomography (ERT)

ERT is an electrical geophysical method that involves the use of a direct electric current to measure the electric potential difference in the subsurface. Crosswell ERT configuration is displayed in figure 32 where electrodes are placed on two monitoring wells at a prespecified electrode spacing. An electrode from the first well is going to transmit the electric current while the other ones from the second well will measure the voltage gradient. The acquisition is repeated multiple times to obtain the different current pathways and produce a map of spatial resistivity. The supercritical CO_2 content can be

visualized due to the high resistivity response on brine reservoirs, providing an image of the injected CO_2 plume.

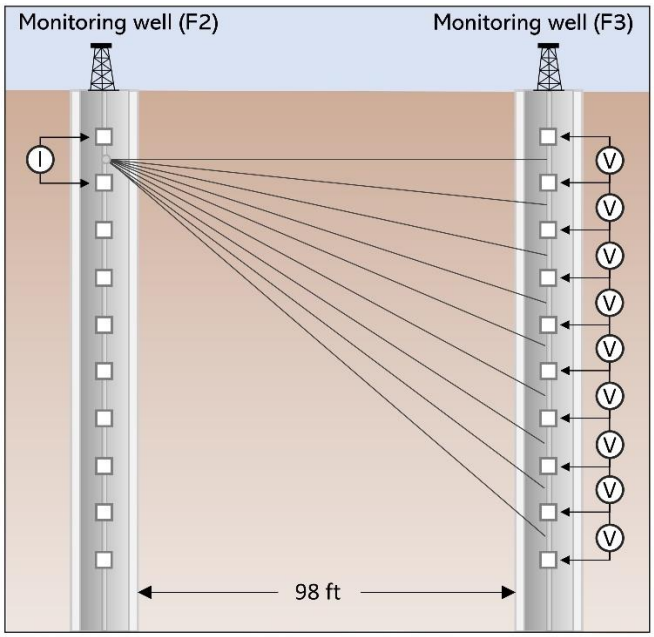


Figure 32. Schematic representation of the crosswell ERT acquisition where the electric potential is measure at the receivers electrodes. I = electric current, V= measured voltage gradient.

Methodology

Dataset description

The dataset utilized contains 91 daily ERT measurements collected from 09 December of 2009 to 12 March of 2010. The ERT acquisition was configured in two monitoring with the placement of 14 electrodes in well F-2 and 7 electrodes in well F-3. The injection well is set to a distance of approximately 229 feet from the F-2 and 327 feet from the F-3. The reservoir depth is set at a range of 10449.5 to 10521.5 feet at a thickness of 72 feet, being this the focalized zone of research (figure 33).

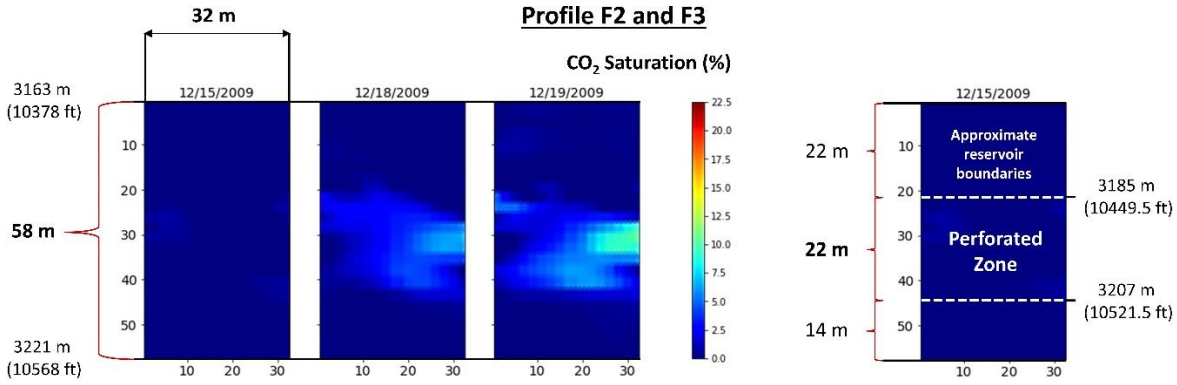


Figure 33. Time-lapse CO_2 saturation images of profile F2 and F3. The injected reservoir corresponds to the perforated zone, displaying a thickness of ~22 m and a length of ~32 m. The saturation ranges from 0 to 22.5 % of CO_2 content.

Processing of the data was previously carried out by Carrigan et al. (2013) where three major steps were performed:

1) Removal of noisy data points using multiple thresholds
2) Modeling baseline measurements to construct a reference dataset
3) Time-lapse inversion obtains CO_2 saturations from the resistivity changes. The processing approach was based on the ratio inversion scheme and Archie's law.

To investigate the different flow regimes, the CO_2 saturations were used as input data due to the lack of resistivity information. The CO_2 is measured as percentage being zero the lowest response and 25 the highest saturation increase. The ERT spatial coordinates correspond to a distance of 33 meters on the x-axis and 57 on the y-axis. These coordinates represent the distance between well F-2 and F-3 (x-axis), and the first and last electrode (y-axis).

Workflow for spatial-temporal clustering

In this study, the purpose is to identify hidden patterns of spatial-temporal plume behavior and uncover potential processes affecting the efficiency of CO_2 migration. The

proposed workflow is shown in figure 34 with the main steps of the spatial-temporal clustering model.

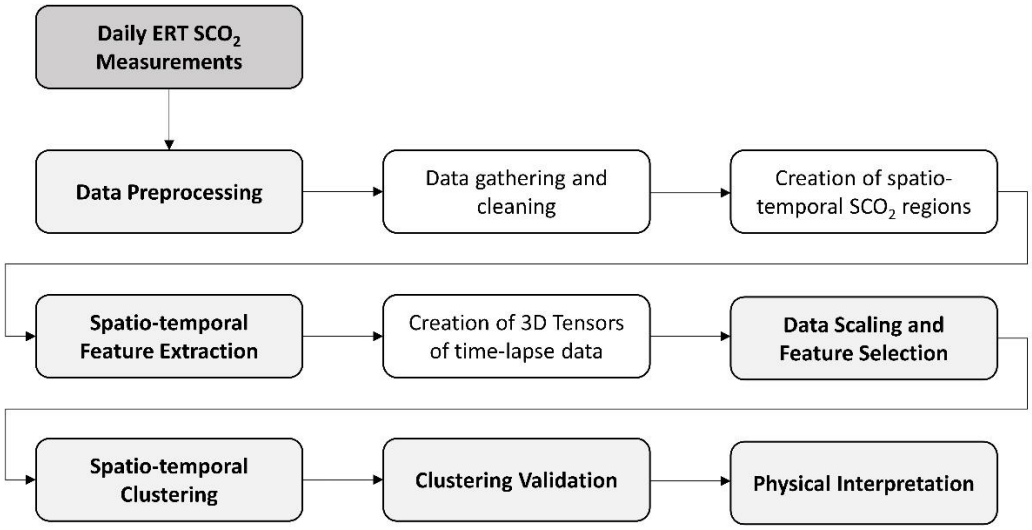


Figure 34. Workflow used for the spatial-temporal clustering of SCO_2 ERT. Six main stages are needed to process, implement, and evaluate time-lapse CO_2 migration.

The steps are summarized as:

Step 1: Data preprocessing. The dataset was gathered and cleaned by identifying the missing and irrelevant data points to increase model quality. The creation of SCO_2 regions was performed as a previous step of feature extraction to retrieve the spatial-temporal changes of the input data.

Step 2: Spatial-temporal Feature Extraction. A novel design was developed to incorporate the previous, current and subsequent stages of daily SCO_2 measurements. We extracted 12 features from a 3D tensor centered in the second temporal dimension and combined them through statistical methods.

Step 3: Data scaling and feature selection. We applied MinMax scaling to transform the features to a compared scale for the unsupervised clustering algorithm. MinMax scaler estimation is presented as:

$$X_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{10}$$

Where "min" and "max" represent the feature range and "x" the feature vector. This standardization will transform the samples into a range of 0 and 1.

From the extracted features, we selected the most impactful ones based on their pairwise correlations and the appropriateness for moving spatial clusters.

Step 4: Spatial-temporal Clustering. With the extracted features, we implemented k-means clustering using the time series distance metric: dynamic time warping (DTW). The optimal number of clusters was assessed by the metric scores of Davie-Bouldin, Calinski-Harabasz, and DTW silhouette.

Step 5: Clustering Validation. We employed statistical analysis to validate the clustering results and retrieve the features that best describe the data response. For this, we applied ANOVA (analysis of variance) and post-hoc Tuckey's test to capture the statistical difference of selected features.

In addition, to assess the DTW k-means approach, we applied conventional clustering algorithms such as agglomerative, mean-shift and Euclidean k-means. We established their performance and comparison with the metric scores of Davie-Bouldin and Calinski-Harabasz.

Step 6: Physical interpretation. Using the wellbore measurements (temperature, pressure, and flow rate) of the injected well, we compared the daily clusters count with F-

1 injection phases. We aim to link the behavior of each class to the wellbore responses by examining their temporal changes. Subsequently, we implemented a second clustering model to group the time frame of clustered results. The data is reshaped to a daily format to discover CO_2 fluid regimes of a similar response.

With this workflow, we expect to discover the predominant CO_2 fluid mechanisms, the evolution of CO_2 plume and the systems affecting the migration and storage efficiency.

Feature extraction design

To incorporate the temporal and spatial change of moving CO_2 content, we developed a novel feature extraction approach. The extraction procedure begins by conditioning the original SCO_2 ERT data. We first created regions of 5 by 5 dimensions and focalized the data to the reservoir and injected zone (figure 35).

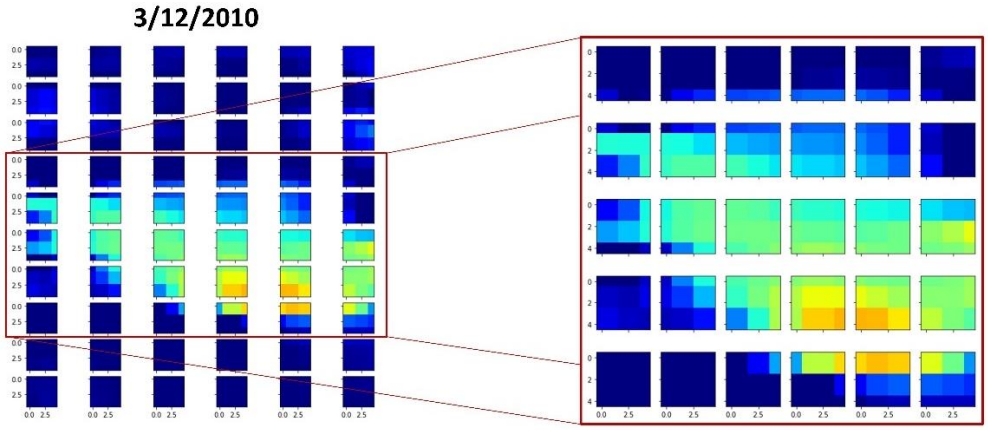


Figure 35. Creation of SCO_2 regions of 5x5 dimensions of the time-lapse images. Each CO_2 image will contain 5x6 regions of 5x5 observations. This was used to account for the local spatial information for the tensor-based feature extraction.

With the daily regions, we designed a tensor-based feature extraction methodology. As illustrated in figure 36, regions were arranged into 3D tensors of 3x5x5 shape where we aim to incorporate the temporal response of the previous, current, and

following day. Subsequently, we performed feature extraction on each slice of the 3D array and combined their responses to obtain a final representation of the feature. The 3D structure will provide the spatial-temporal flow dynamics by adding the change of a moving system. The exact procedure was conducted on all regions for different extracted features.

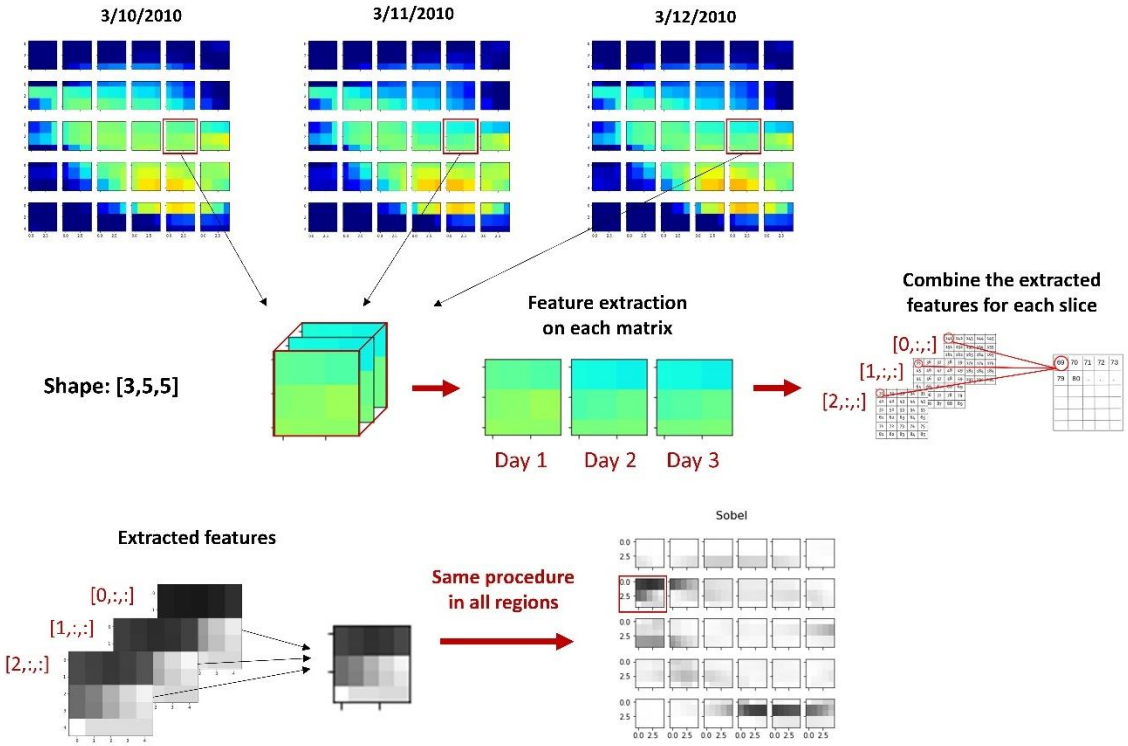


Figure 36. Tensor-based feature extraction design for 3D arrays of 3x5x5 dimensions. To account for the temporal and spatial components of ERT measurements, four steps were implemented: 1) regions are arranged in 3D tensors, 2) feature extraction on each tensor slice, 3) transformation of 3D to a 2D array, and 4) repetition of the process to all tensor regions.

Using the previously described methodology, the extraction techniques applied were: sobel, gray-level co-occurrence matrix, fast-Fourier transform, linear binary pattern, hessian matrix, difference of Gaussians, structure tensor, entropy, height bellow, height above, contrast stretching, and SCO_2 difference. Nevertheless, only 5 features were

selected based on their statistical correlation and impact to describe moving objects. The features used are displayed in table 10 with a brief explanation of the extraction technique.

Table 10. Brief description of extracted features from ERT SCO_2.

Feature	**Description**
Fast-Fourier transform	Transformation of the image from spatial to the frequency domain. Low and high pass filters permit to pass off certain image frequencies.
Structure tensor	Estimation of the weighted sum of squared differences in a centered pixel window.
Height bellow	Difference between the highest elevation point and current point.
Contrast stretching	Stretching or shrinking of pixels intensity levels.
SCO_2 difference	Difference between the previous date and current date observations.

Spatial-temporal clustering

Dynamic time wrapping and k-means

In this study, we used Dynamic time wrapping (DTW) as the distance metric for spatial-temporal clustering. DTW evaluates the optimal aligning between two time-dependent signals (Muller, 2007). This is a technique suitable for time-series datasets due to the ability to capture the similarity/dissimilarity of temporal distances that do not have an equal temporal length. In other words, invariant to time shifts. These pairwise distances are warped in a nonlinear regime to approximate both sequences. The use of k-means clustering and DTW can be divided into two major steps where the algorithm is going to:

- First, it arranges the time-series to similar shapes by using DTW

- And second, it computes the barycenters or clusters centers with the use of DTW. This will provide an average shape of clusters while taking into account temporal shits.

Physical interpretation of spatial-temporal clusters

To provide the physical meaning of clustering results, we examined the association between each cluster class and the wellbore measurements of the injection well. These measurements were acquired daily to monitor the response of injected CO_2. We used the pressure, temperature, and flow rate to analyze the relationship of clusters variability. With the clustering classes, we aggregated their values by the daily appearance, counting their classes each day. Based on both datasets, we aim to correlate the changes in pressure, temperature, and flow rate with the changes of clusters occurrence (Figure 37).

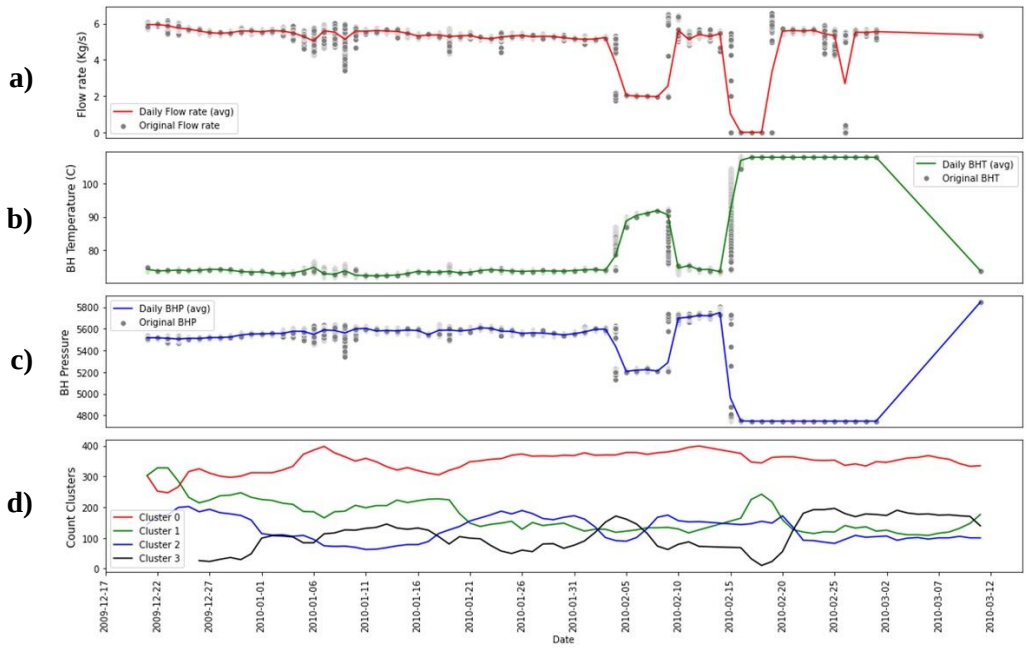


Figure 37. Wellbore measurements: a) flow rate, b) temperature and 3) pressure of injection well; and d) daily clusters occurrences. Clusters count changes were used to correlate the injection phases with the clustering results.

In addition, we established a second clustering model to uncover the temporal patterns from the previous clustering results. As illustrated in figure 38, the clusters were rearranged to group the dates of similar clustering responses. Based on this, we applied k-means to group the dates of each individual clustered ERT image. These clustered dates were compared with the wellbore measurements to investigate the relationship between injection phases.

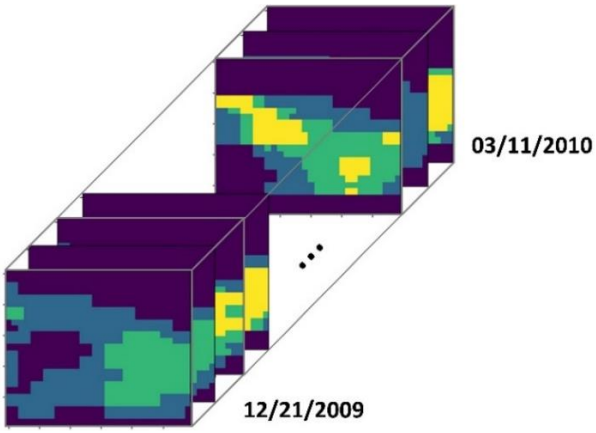


Figure 38. Representation of second clustering input data for the temporal analysis of flow regimes.

Results and discussion

Validation of clustering

To evaluate the performance of the spatial-temporal clustering, Davies-Bouldin, Calinski-Harabasz, and DTW-Silhouette scores were estimated. These clustering metrics, called internal validation measures, assess the goodness of the partition without the use of any external information. The criteria used are commonly based on the compactness (similarity) and separation (distinction) of clusters (Liu et al., 2010). A higher value of Calinski-Harabasz and DTW-Silhouette will indicate a better partition while Davies-Bouldin a lower value.

In the DTW K-means implementation, four clusters were established using the Davies-Bouldin and Calinski-Harabasz scores. Figure 39 presents the original ERT SCO_2 and clustering results for the time-lapse dataset. Clusters classes were linked to levels of SCO_2 content, ranged from zero to high CO_2. The validation measures confirmed the clustering with a Davies-Bouldin index of 0.71, a Calinski-Harabasz of 262791.45, and a DTW-silhouette score of 0.58. These values indicate the measures of the highest clustering performance.

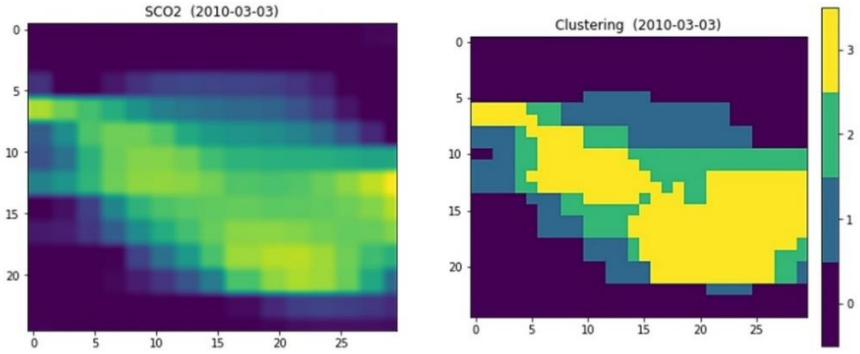


Figure 39. Left: SCO_2 ERT dataset for the 03/03/2010 acquisition**.** Right: DTW K-means clustering. Clustered results correspond to no-CO_2 (cluster 0), low CO_2 (cluster 1), medium CO_2 (cluster 2), and high CO_2 (cluster 3).

Comparison of multiple clustering methods and DTW K-means

A comparative analysis was performed, to validate the use of DTW K-means for spatial-temporal datasets. Three clustering algorithms were implemented to characterize the moving CO_2. K-means with Euclidean distance, agglomerative, and mean-shift were the applied methods. These approaches are often applied to sequences of high spatial components using ordinary distances such as Euclidean. As shown in figure 40, differences between them are mainly associated with: 1) the low characterization of the moving CO_2 content (cluster 3) 2) low separation and compactness of meanshift clusters.

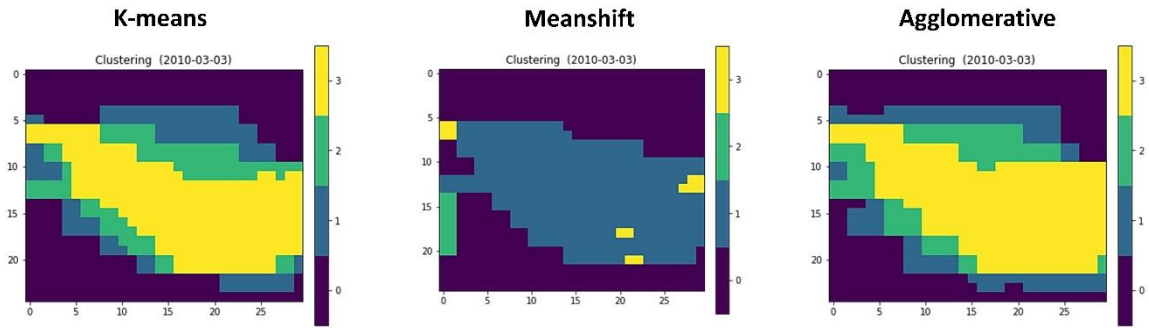


Figure 40. Spatial-temporal clustering using a) Euclidean k-means, b) meanshift and c) agglomerative clustering. Qualitatively k-means and agglomerative display a similar behavior of low migration movements, while meanshift display a poorly clustering performance.

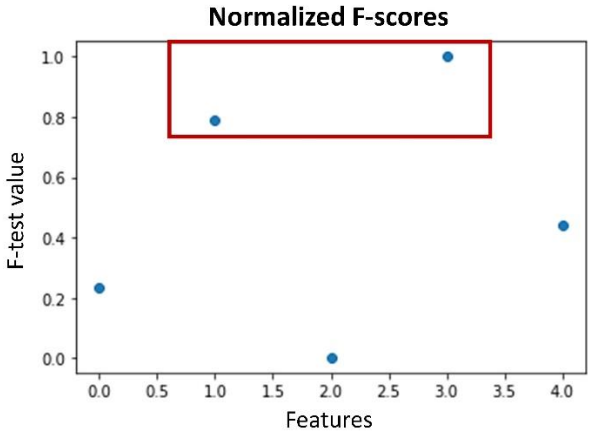
Moreover, Davies-Bouldin and Calinski-Harabasz scores were calculated to compare the goodness of all the clustering results. Table 11 presents the scores of each method, being DTW K-mean the one with a better data partition. Hence, temporal distances could enhance spatial-temporal clustering by incorporating the time-series components.

Table 11. Internal metric scores of clustering results for the algorithms of DTW k-means, Euclidean k-means, meanshift, and agglomerative. A lower Davies-Bouldin value indicates a better performance while for Calinski-Harabasz it displays a higher score. DTW k-means has the lowest Davies-Bouldin and highest Calinski-Harabasz score.

Clustering algorithms	**Score**	
	Davies-Bouldin	*Calinski-Harabasz*
DTW K-means	0.71	262791.45
Euclidean K-means	0.83	157866.86
Agglomerative	0.95	131593.57
Meanshift	1.01	69438.35

Statistical analysis

In this workflow stage, we examined the statistical difference of the extracted features among the spatial-temporal clusters. The analysis of these differences determines the most impactful features for ERT SCO_2 measurements. To quantitatively evaluate the statistical significance, we applied one-way ANOVA (analysis of variance). The results from the ANOVA tests are displayed in figure 41, being contrast stretching and fast-Fourier transform the features that best describe the CO_2 content.



Features number	Feature
0	Structure tensor
1	Fast-Fourier transform
2	SCO_2 difference
3	Contrast stretching
4	Height bellow

Figure 41. Normalized F-test values to establish the most impactful features. Contrast stretching and fast-Fourier transform are the signatures that provide significant clustered information to describe SCO_2.

Subsequently, we conducted the post-hoc Tukey's test to further analyzed the pairwise clusters difference of selected features. Using the ANOVA results, contrast stretching and fast-Fourier transform were assessed (table 12). From the features mean difference, clusters "0" and "3" display the highest difference between each clusters class. Conversely, clusters "2" and "3" reveal the lowest difference or highest clusters similarity.

We also established contrast stretching, as the strongest feature due to their high clusters difference.

Table 12. Tukey HSD for post hoc analysis of the features contrast stretching and fast-Fourier transform. Mean differences between clusters indicate the significance among them. Cluster "0" indicates non-CO_2 content, and clusters "1", "2", and "3" their respective level of saturation (low, medium, and high).

Feature	**Clusters being compared**		**Mean difference**	**Normalized difference**
Contrast stretching	Cluster #	Cluster #		
	0	1	834395206.29	0.45
	0	2	1588207701.76	0.85
	0	3	1867909636.58	1.00
	1	2	753812495.47	0.40
	1	3	1033514430.28	0.55
	2	3	279701934.81	0.15
Fast Fourier Transform	Cluster #	Cluster #		
	0	1	5.89	0.41
	0	2	11.09	0.76
	0	3	14.51	1.00
	1	2	5.20	0.36
	1	3	8.62	0.59
	2	3	3.42	0.24

To validate this difference, we draw boxplots of clustering results for the most impactful features. As shown in figure 42, contrast stretching provides a better representation of the spatial-temporal clusters. We also confirm the similarity between clusters "2" and "3", and the statistical difference of clusters "0" and "3". For fast-Fourier-

transform clusters overlap on all classes. A combination of both features could potentially differentiate them to a higher degree.

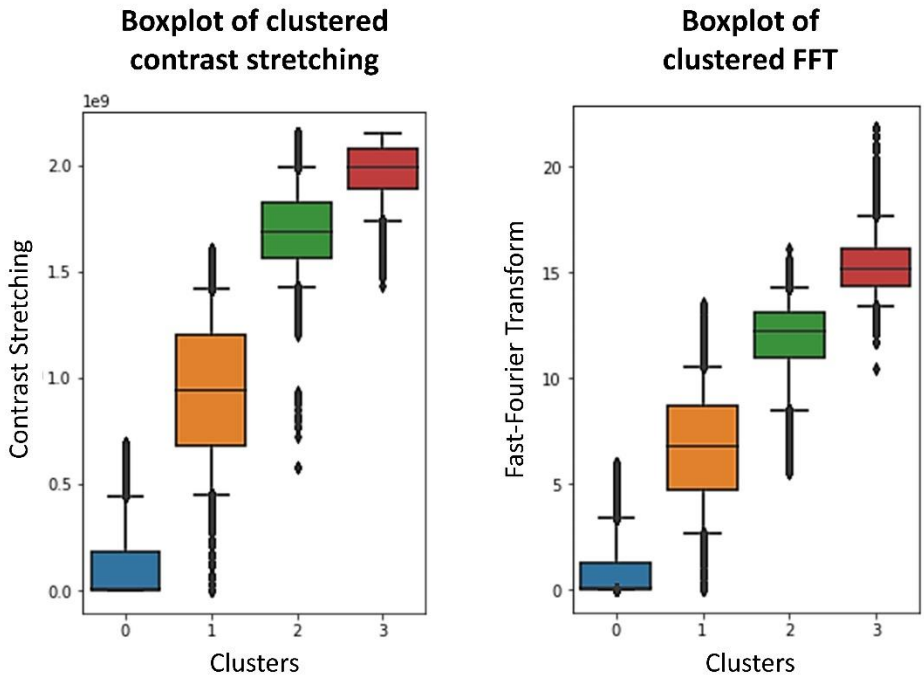


Figure 42. Boxplot of clustered contrast stretching and fast-Fourier transform. Boxplots were defined for a low 5th percentile and a high 95th percentile.

Physical meaning using wellbore measurements and second temporal clustering

The physical interpretation of clusters is investigated with the wellbore measurements of the injection well. Figure 43 displays the daily flow rate, temperature, pressure, and the respective daily count of cluster "3" (High CO_2). We can distinguish from the dates of "2010-01-30" to "2010-03-09" a decrease/increase of both cluster occurrence and flow rate. These changes are attributed to the injection phases which directly affect the CO_2 plume migration.

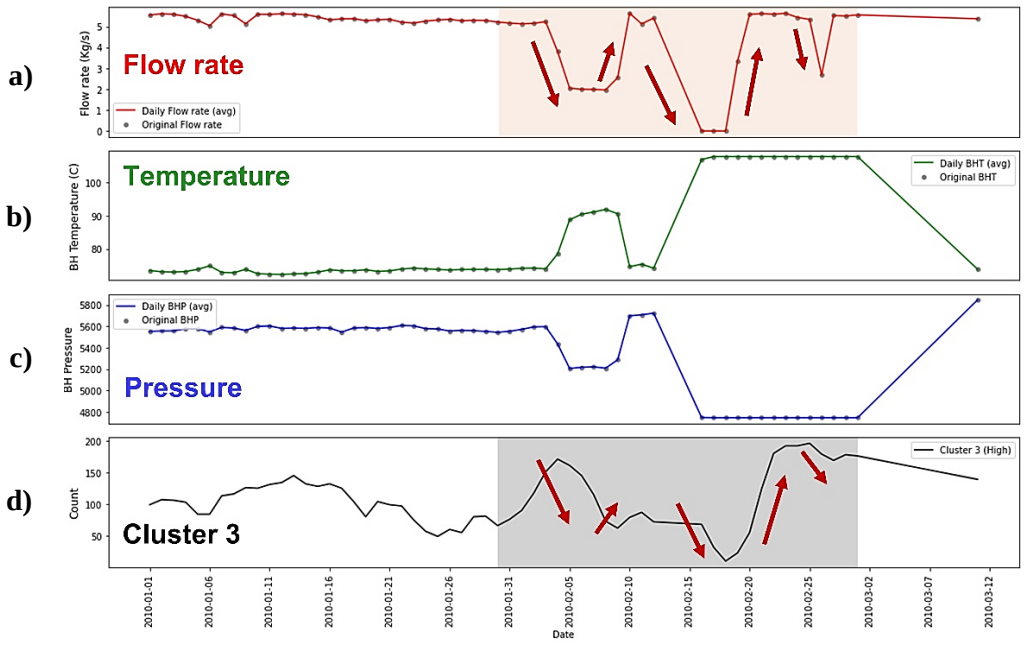


Figure 43. Wellbore measurements: a) flow rate, b) temperature and c) pressure of injection well; and d) daily count of cluster 3 (high SCO_2). The gray background corresponds to the dates from "2010-01-30" to "2010-03-09" where a change of all measurements is observed.

To further evaluate the clustering results, we applied a second clustering from the previous clusters. Dates were clustered to discover flow regimes of similar behavior. Figure 44 shows the second clustering results for the daily domain where we can observe the distinctive plume behaviors of all clusters. The clustering analysis is established in table 13 where we can distinguish different stages of injected CO_2. Lastly, the temporal clusters also display a concordance with the flow rate (figure 45) by identifying temporal patterns in the clustered ERT images. From these results, unique temporal behaviors were uncovered. They can be linked to the phases of plume development and their respective CO_2 flow mechanisms. The movement of clusters is observed at the bottom and top of the reservoir boundaries. In addition, we can distinguish slow access of certain reservoir regions from the clustered plume shape.

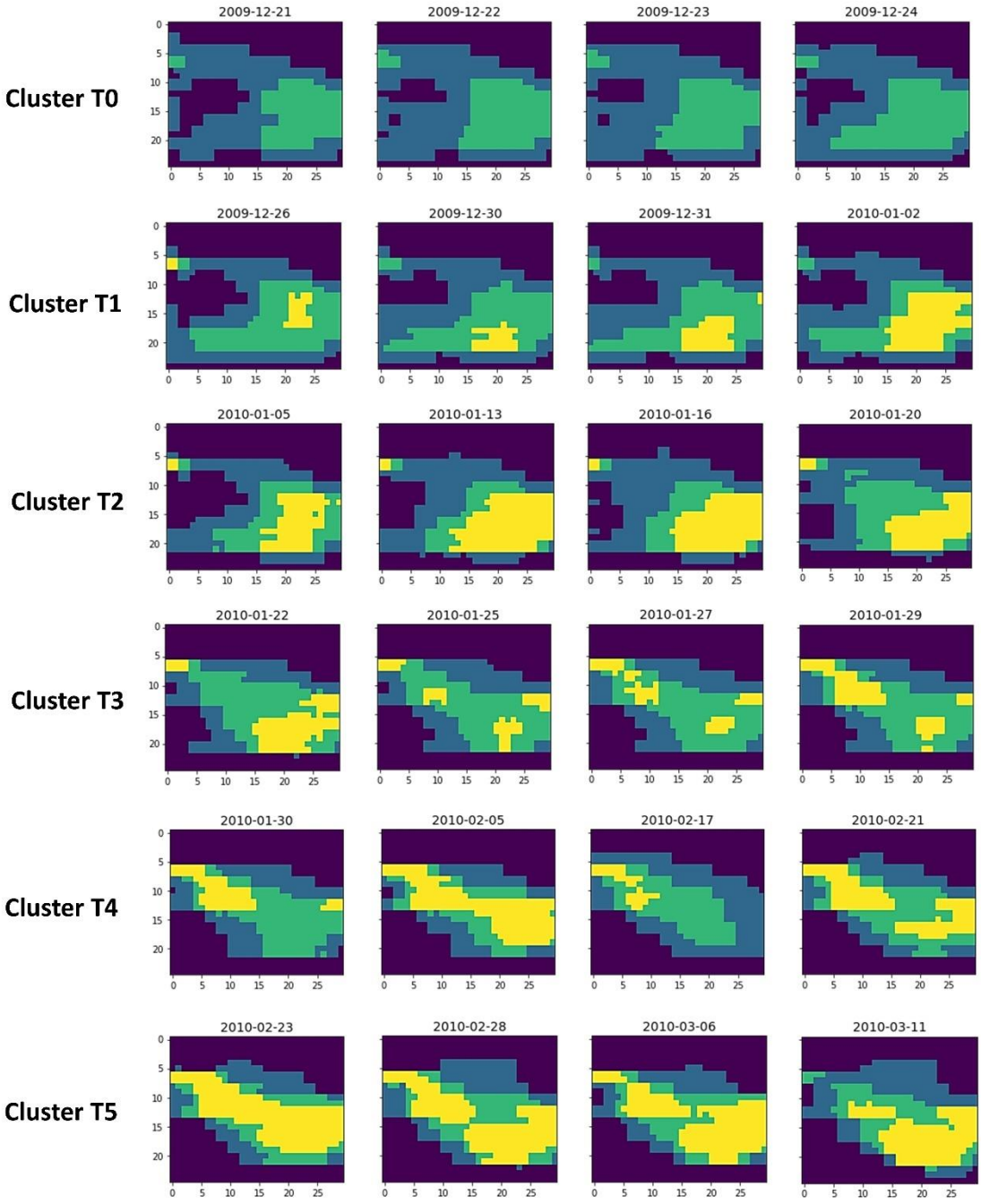


Figure 44. Daily clustered images from the resulted temporal clustering. Six clusters were determined to retrieve CO_2 flow regimes using the first clustering results. "T0" contains 4 images "T1" = 11, "T2" = 16, "T4" = 9, "T3" = 20, and T5 =18.

Table 13. Analysis of daily cluster dates from the derived spatial-temporal clustering.

Temporal clusters	**Clustering analysis**
Cluster T0	Occurrence of clusters 1 and 2 (low and medium SCO_2)
Cluster T1	Occurrence and increment of cluster 3 (high SCO_2)
Cluster T2	No significant clustering changes
Cluster T3	Spatial movement of cluster 3 (high SCO_2) and change of clusters shape
Cluster T4	Decrease and increase of cluster 3 (high SCO_2)
Cluster T5	No significant clustering changes. Cluster 3 (high SCO_2) is arranged into subregions.

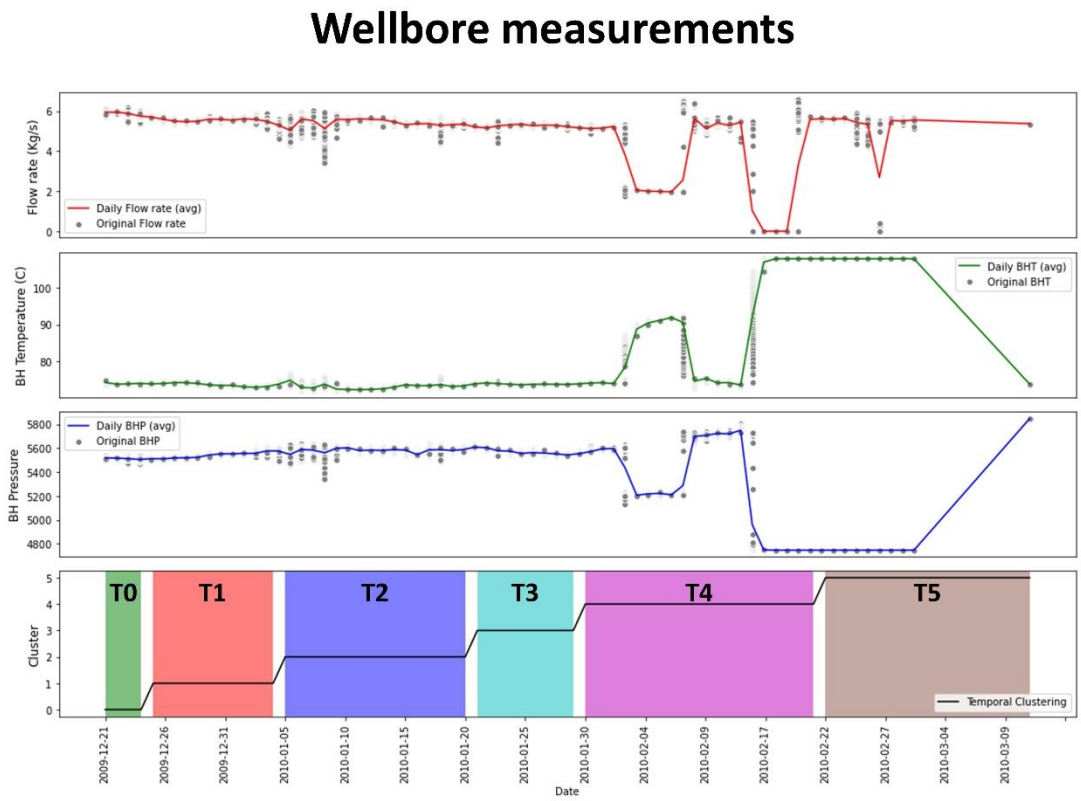


Figure 45. Wellbore measurements and temporal clustering of the daily dates from spatial-temporal results. Changes in plume shape are linked to the temporal clusters by uncovering the drastic changes in flow rate, temperature, and pressure.

Assumptions and Limitations

In this work the following assumptions can be drawn:

- The time-lapse dataset was previously preprocessed by geophysicists, being the input data the derived CO_2 saturations from the ERT images using a ratio inversion scheme.
- A rock physics model was implemented (Archie's model) where assumptions such as the saturation exponent and non-conductive gas phase are incorporated. Resistivity measurements would remove these assumptions from the spatial-temporal model.
- Multiple processes can affect the migration of injected CO_2; hence, other subsurface measurements should be incorporated such as geomechanical and geochemical analyses.
- The proposed pipeline can be used under different geological conditions for multiple mapping tools. This would include subsurface imagining tools on different geo-sequestration fields.

Final Remarks

A spatial-temporal clustering model was developed and validated to uncover hidden patterns of derived-SCO_2 from daily ERT images. We proposed and implemented a novel feature extraction design to include the spatial and temporal neighboring regions. From the clustering work, we can make de following conclusions:

1. The inclusion of previous and subsequent measures enhances the feature extraction and behavior of moving objects. Furthermore, the local windows or

regions, used in the tensor-based extraction, provide a better representation of spatial changes.

2. A growth of CO_2 plume is observed through the changes of plume shape and CO_2 levels. This reveals the dynamics of a CO_2 plume where processes affecting the migration and trapping can be determined (e.g. increase/decrease of the injection flow rate or the geological heterogeneity of the reservoir).

3. The daily second clustering discloses the stages of plume evolution such as early development, CO_2 equilibrium, and multiple saturation changes. In addition, several fluid flow forces are distinguished at the top (buoyancy), bottom (gravity), and regions of partially filled CO_2 (capillary).

CHAPTER V

CONCLUSIONS AND FUTURE WORK

The recent advances and success of machine learning on subsurface geophysical tools have led to the improvement of data exploration in reservoir characterization. In particular, a growth of unsupervised learning applications has been observed due to the high limitations of traditional modeling and the necessity of understanding complex physical systems. To examine the spatial-temporal reservoir patterns, we conducted three independent studies using three distinctive geophysical methods.

In chapter 2, we explore the use of dimensionality techniques to compress, denoise and discover relevant information of microseismic and DAS datasets. Moreover, a CO_2 visualization model was developed (chapter 3) to investigate the signatures of CO_2 content from derived-SCO_2 of crosswell seismic. From chapter 4, we presented a spatial-temporal clustering approach to identify the predominant fluid-flow mechanisms of carbon storage and CO_2 plume migration. The key findings from the studies brought the following conclusions:

- Unsupervised learning for dimensionality reduction can be used to substantially reduce the computational time and memory of large datasets.
- Tensor decomposition provides the possibility of retrieving the insights into the most impactful features for high-dimensional datasets. This also performs parallel processing by denoising and compressing the input data.

- Using unsupervised clustering, we can visualize the CO_2 levels of time-lapse measurements. The new data-driven model gives a better understanding of the CO_2 movement by taking an evidence-based approach of low human intervention.
- The multi-level clustering method outperforms the traditional clustering techniques for datasets of unbalanced nature. This allows discretizing the important information and further analysis of subsurface mapping tools.
- Unsupervised clustering for spatial-temporal datasets is crucial to establish the dynamics of moving systems. Both temporal and spatial components need to be incorporated at different workflow stages to achieve an efficient clustering and impactful set of features.
- Major characteristics of injected fluid can be observed from the temporal clustering of spatial-based measurements. These hidden patterns are tied to the local changes of both domains, revealing the evolution of reservoir fluid flow systems.

Lastly, future work is needed to be addressed for improving the proposed models and retrieving insights from large volumes of data. The main recommendations for future analysis are:

- Investigate the core tensor properties to extract the most impactful signals of high-dimensional datasets. The HOSVD structure allows the examine the components of the tensor decomposition, being the core tensor the one that contains the main information of the system.

- Use other subsurface tools to integrate and analyze the CO_2 plume behavior and evolution. Datasets available from the SECARB project involve measurements of VSP, bottom hole gravity, well logs, and core analysis.
- Combine field with simulated data to incorporate scenarios of CO_2 leakage or any processes affecting the CO_2 migration and trapping.

REFERENCES

Afra, S. & Gildin, E. (2016). Tensor based geology preserving reservoir parameterization with Higher Order Singular Value Decomposition (HOSVD). Comput. Geosci. 94, C (September 2016), 110–120. https://doi.org/10.1016/j.cageo.2016.05.010.

Bao, A., Gildin, E., Huang, J., & Coutinho, E. J. (2020, July). Data-driven end-to-end production prediction of oil reservoirs by EnKF-enhanced Recurrent Neural Networks. In SPE Latin American and Caribbean Petroleum Engineering Conference. OnePetro.

Bekara. M., & van der Baan, M. (2007). Local singular value decomposition for signal enhancement of seismic data. Geophysics, 72(2), V59–V65. https://doi.org/10.1190/1.2435967

Brankovic, M., Gildin, E., Gibson, R., & Everett, M. (2021). A Machine Learning-Based Seismic Data Compression and Interpretation Using a Novel Shifted-Matrix Decomposition Algorithm. Appl. Sci. 11, no. 11: 4874. https://doi.org/10.3390/app11114874.

Caliński, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1–27. https://doi.org/10.1080/03610927408827101

Chakravarty, A., & Misra, S. (2021). Hydraulic fracture mapping using wavelet-based fusion of wave transmission and emission measurements. Journal of Natural Gas Science and Engineering, 104274.

Davies, D., and Bouldin, D. (1979). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Davis, T., Landrø, M., and Wilson, M. (Eds.). (2019). Geophysics and Geosequestration. Cambridge: Cambridge University Press. doi:10.1017/9781316480724

Freire, S., & Ulrych, T. (1988). Application of singular value decomposition to vertical seismic profiling. Geophysics, 53(6), 778–785. https://doi.org/10.1190/1.1442513.

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

Haghighat, S., Mohaghegh, S., Gholami, V., Shahkarami, A., and Moreno, D. (2013). Using Big Data and Smart Field Technology for Detecting Leakage in a CO2 Storage Project. Paper presented at the SPE Annual Technical Conference and Exhibition, New Orleans, Louisiana, USA. doi: https://doi.org/10.2118/166137-MS.

Hovorka, S.D. (2013). Three-Million-Metric-Ton-Monitored Injection at the SECARB Cranfield Project—Project Update. Energy Procedia, 37, 6412-6423.

Hovorka, S.D., Meckel, T.A., Trevino, R.H., Lu, J., Nicot, J., Choi, J., Freeman, D., Cook, P., Daley, T.M., Ajo-Franklin, J., Freifeild, B.M., Doughty, C., Carrigan, C.R., La Brecque, D., Kharaka Y.K., Thordsen, J.J., Phelps, T.J., Yang, C., Romanak, K.D., Zhang, T., Holt, R.M., Lindler, J.S., Butsch, R.J. (2011). Monitoring a Large Volume CO 2 Injection: Year Two Results from SECARB Project at Denbury's Cranfield, Mississippi, USA. Energy Procedia 4: 3478 – 3485.

IPCC. (2018). In: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy,T. Maycock, M. Tignor, and T. Waterfield (eds.)]. World Meteorological Organization, Geneva, Switzerland, 32 pp

Iqbal, Naveed, Liu, Entao, McClellan, James H, Al-Shuhail, Abdullatif, Kaka, Sanlinn I, & Zerguine, Azzedine. (2018). Detection and Denoising of Microseismic Events Using Time-Frequency Representation and Tensor Decomposition. IEEE Access, 6, 22993–23006. https://doi.org/10.1109/ACCESS.2018.2830975

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Kolda, T. (2016, November 16). Parallel Multiway Methods for Compression of Massive Data and Other Applications [Conference presentation]. International Conference for High Performance Computing, Networking, Storage and Analysis, Salt Lake City, CA, United States.

Kolda, T. (2018, January 11). Tensor Decomposition [Conference presentation]. JMM Joint Mathematics Meetings, San Diego, CA, United States.

Kolda, T., & Bader, B. (2007). Tensor Decompositions and Applications. SIAM Review, Vol 51, No. 3. doi. 10.1137/07070111X

Kreimer, N., & Sacchi, M. (2012). A tensor higher-order singular value decomposition for prestack seismic data noise reduction and interpolation. Geophysics, 77(3), V113–V122. https://doi.org/10.1190/geo2011-0399.1

Li, H., & Misra, S. (2018). Long short-term memory and variational autoencoder with convolutional neural networks for generating NMR T2 distributions. IEEE Geoscience and Remote Sensing Letters, 16(2), 192-195.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010, December). Understanding of internal clustering validation measures. In 2010 IEEE international conference on data mining (pp. 911-916). IEEE.

Mandelli, S., Lipari, V., Bestagini, P., & Tubaro, S. (2019). Interpolation and Denoising of Seismic Data using Convolutional Neural Networks. ArXiv, abs/1901.07927.

Markets and Markets. (2020). Carbon Capture, Utilization, and Storage Market by Service (Capture, Transportation, Utilization, Storage), End-Use Industry (Oil & Gas, Iron & Steel, Cement, Chemical & Petrochemical, Power Generation), and Region - Global Forecast to 2025. Retrieved from https://www.marketsandmarkets.com/Market-Reports/carbon-capture-utilization-storage-market-151234843.html

Müller, A.C. and Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media, Inc.

Müller, M. (2007). Dtw-based motion comparison and retrieval. Information Retrieval for Music and Motion, 211-226.

National Energy Technology Laboratory. (2020). About SMART Initiative. Retrieved from https://edx.netl.doe.gov/smart/about-smart/

Ni, H., & Benson, S. M. (2020). Using unsupervised machine learning to characterize capillary flow and residual trapping. Water Resources Research, 56(8), e2020WR027473.

Nikravesh, M. (2007). Computational intelligence for geosciences and oil exploration. In Forging New Frontiers: Fuzzy Pioneers I (pp. 267-332). Springer, Berlin, Heidelberg.

Pires de Lima, R., and Lin, Y. (2019). Geophysical data integration and machine learning for multi-target leakage estimation in geologic carbon sequestration. SEG Technical Program Expanded Abstracts: 2333-2337.

Rabanser, S., Shchur, O., & Günnemann, S. (2017). Introduction to Tensor Decompositions and their Applications in Machine Learning. ArXiv, abs/1711.10781.

Rackley, S. A. (2010). Carbon capture and storage. Butterworth-Heinemann/Elsevier. First Edition. ISBN: 9781856176361

Rahimi, M., Moosavi, S. M., Smit, B., and Hatton, T.A. (2021). Toward smart carbon capture with machine learning. Cell Reports Physical Science, Volume 2, Issue 4. doi: https://doi.org/10.1016/j.xcrp.2021.100396.

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. IEEE Access, 8, 54776-54788.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

U.S. Energy Information Administration. (2021). Annual Energy Outlook 2021. Retrieved from: https://www.eia.gov/outlooks/aeo/pdf/AEO_Narrative_2021.pdf

Vesselinov, V., Mudunuru, M., Karra, S., O'Malley, D., & Alexandrov, B. Unsupervised machine learning based on non-negative tensor factorization for analyzing reactive-mixing. Journal of Computational Physics. doi.org/10.1016/j.jcp.2019.05.039.

Warpinski, N. (2009). Microseismic Monitoring: Inside and Out. J Pet Technol 61: 80–85. doi: 10.2118/118537-JPT

Wu, Y., Misra, S., Sondergeld, C., Curtis, M., & Jernigen, J. (2019). Machine learning for locating organic matter and pores in scanning electron microscopy images of organic-rich shales. Fuel, 253, 662-676.

Yatsenko, M., Brankovic, M., Gildin, E., & Gibson, R. (2019). A Novel Approach to Discovery of Hidden Structures in Microseismic Data Using Machine Learning Techniques. Paper presented at the SPE Europec featured at 81st EAGE Conference and Exhibition, London, England, UK. doi: https://doi.org/10.2118/195522-MS