UNIVERSIDADE FEDERAL FLUMINENSE
INSTITUTO DE GEOCIÊNCIAS
DEPARTAMENTO DE GEOLOGIA E GEOFÍSICA

BERNARDO SARTORI CHEDE ROSAURO DE ALMEIDA

INTEGRATING TOC DATA WITH 3D GEOLOGICAL MODEL FOR ENHANCED SPATIAL SOURCE ROCK CHARACTERIZATION IN THE SANTOS BASIN

DOCTORATE THESIS

POSTGRADUATE PROGRAM IN OCEAN AND EARTH DYNAMICS (DOT)

Niterói
February/2025

BERNARDO SARTORI CHEDE ROSAURO DE ALMEIDA

INTEGRATING TOC DATA WITH 3D GEOLOGICAL MODEL FOR ENHANCED SPATIAL SOURCE ROCK CHARACTERIZATION IN THE SANTOS BASIN

Thesis presented to the Postgraduate Program in Ocean and Earth Dynamics of the Fluminense Federal University, as partial requirement for obtaining the Doctoral degree

Concentration Area: Geology and Geophysics

Advisor
Prof. Ph.D. André L. Belém

Co-Advisor
Prof. Ph.D. Ana Luiza Spadano Albuquerque

Niterói
February/2025

BERNARDO SARTORI CHEDE ROSAURO DE ALMEIDA

INTEGRATING TOC DATA WITH 3D GEOLOGICAL MODEL FOR ENHANCED SPATIAL SOURCE ROCK CHARACTERIZATION IN THE SANTOS BASIN

Tese apresentada ao Programa de Pós-Graduação em Dinâmica dos Oceanos e Terra, da Universidade Federal Fluminense, como requisito parcial para obtenção do grau de Doutor

Área de Concentração: Geologia e Geofísica

Aprovado em 17/03/2024

BANCA EXAMINADORA

_____
Prof. André L. Belém, Dr.
UFF

_____
Prof. Ana Luiza Spadano Albuquerque, Dra.
UFF

_____
Prof. Igor Martins Venancio P. de Oliveira, Dr.
UFF

_____
Prof. Leonardo Guimarães Miquelutti, Dr.
UFF

_____
Rodrigo de Lima Sobrinho, Dr.
UFF

_____
Ismael Humberto Ferreira dos Santos,
CENPES-Petrobras

_____
Alexandre Lopes,
INPETUS

Ficha catalográfica automática - SDC/BIG
Gerada com informações fornecidas pelo autor

A447i Almeida, Bernardo Sartori Chede Rosauro de
Integrating TOC Data With 3D Geological Model For Enhanced Spatial Source Rock Characterization in the Santos Basin / Bernardo Sartori Chede Rosauro de Almeida. - 2025.
94 f.: il.

Orientador: André L. Belém.
Coorientador: Ana Luiza Spadano Albuquerque.
Tese (doutorado)-Universidade Federal Fluminense, Instituto de Geociências, Niterói, 2025.

1. Carbono Orgânico Total. 2. Distribuição 3D. 3. Bacia de Santos. 4. Aprendizado de Máquina. 5. Produção intelectual. I. Belém, André L., orientador. II. Albuquerque, Ana Luiza Spadano, coorientadora. III. Universidade Federal Fluminense. Instituto de Geociências. IV. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Dedication

I dedicate this work to both of my grandmothers Nininha and Elsinha, my mother Ana Paula, my father João Luiz (in memoriam), my grandfather Alberto (in memoriam), and my wife Michele, with all my love and gratitude.

Acknowledgements

I would like to thank my wife, Michele, for her patience, understanding, and affection throughout this journey, especially during the most challenging moments of my doctoral studies.

To my mother Ana Paula and my grandmother Maria Ronilte, for their unconditional support throughout my life and for always believing in my potential.

To my advisor Prof. Dr. André Belém, for all the guidance along this journey, for the careful revisions of the article, for the support during my visiting doctoral period, and also for the numerous "I told you so" that proved valuable.

To my co-advisor Prof. Dr. Ana Luisa, for all her dedication and support from my master's degree through the completion of this doctorate.

To the entire PR4 team for all the discussions and shared knowledge. I especially thank Victor Carreira, Kristoffer Hallam, Luiz Cordeiro, Igor Venancio, Rodrigo Sobrinho, Antonio, and João Ballalai and Gabriela Menezes from the SEAP project, who directly or indirectly contributed to the development of this thesis.

"After all this time?"
"Always."
Harry Potter and the Deathly Hallows

Resumo

CHEDE, S. Bernardo. **Integrating TOC Data With 3D Geological Model For Enhanced Spatial Source Rock Characterization in the Santos Basin.** Thesis (Doctor of Science), Fluminense Federal University, Niterói, p. 94, 2025.

Esta pesquisa desenvolve metodologias para a predição e modelagem espacial do Carbono Orgânico Total (COT) na Bacia de Santos, a maior província petrolífera offshore do Brasil. Utilizando uma abordagem dupla, primeiramente desenvolvemos modelos de aprendizado de máquina para a predição de COT a partir de perfis de poços e, posteriormente, implementamos um fluxo de trabalho integrado para modelagem tridimensional da distribuição de COT usando atributos sísmicos e técnicas geoestatísticas. Na primeira parte, comparamos três algoritmos de aprendizado de máquina—Gradient Boosting Decision Tree (GBDT), Extreme Gradient Boosting (XGBoost) e Multi-Layer Perceptron (MLP)—para prever COT a partir de dados de perfis de poços na Bacia de Santos. Todos os modelos de aprendizado de máquina superaram o método tradicional ΔlogR, com GBDT alcançando a maior precisão de predição. Notavelmente, a redução do conjunto de dados de cinco para três poços mais homogêneos melhorou significativamente o desempenho do modelo, destacando a importância da qualidade dos dados e consistência geológica sobre a quantidade. Na segunda parte, desenvolvemos e implementamos um fluxo de trabalho integrado para modelagem tridimensional de COT na seção do pré-sal. Combinando interpretação sísmica, dados sintéticos de poços e técnicas geoestatísticas avançadas, estabelecemos uma metodologia para criar distribuições de COT geologicamente plausíveis. Três abordagens de modelagem—Krigagem, Geração de Campo Aleatório usando Simulação Gaussiana Sequencial e aprendizado de máquina XGBoost—foram comparadas, cada uma oferecendo vantagens distintas para caracterizar a heterogeneidade espacial e quantificar a incerteza de predição. Os modelos 3D de COT resultantes revelam padrões consistentes de riqueza orgânica na Formação Itapema, com concentrações mais elevadas nas seções média a superior. Esta pesquisa contribui para melhorar as metodologias de estimativa de COT e técnicas de caracterização espacial, aprimorando a avaliação do potencial de hidrocarbonetos em configurações de bacias complexas. Os fluxos de trabalho integrados desenvolvidos aqui fornecem modelos que podem ser adaptados a outras bacias, avançando o campo mais amplo de caracterização quantitativa de rochas geradoras.

Palavras-chave: Carbono Orgânico Total, Distribuição 3D, Bacia de Santos, Aprendizado de Máquina

Abstract

CHEDE, S. Bernardo. **Integrating TOC Data With 3D Geological Model For Enhanced Spatial Source Rock Characterization in the Santos Basin.** Thesis (Doctor of Science), Fluminense Federal University, Niterói, p. 94, 2025.

This research advances methodologies for the prediction and spatial modeling of Total Organic Carbon (TOC) in the Santos Basin, Brazil's largest offshore petroleum province. Employing a two-pronged approach, we first develop machine learning models for TOC prediction from well logs and subsequently implement an integrated workflow for three-dimensional TOC distribution modeling using seismic attributes and geostatistical techniques. In the first component, we compare three machine learning algorithms—Gradient Boosting Decision Tree (GBDT), Extreme Gradient Boosting (XGBoost), and Multi-Layer Perceptron (MLP)—for predicting TOC from well-log data in the Santos Basin. All machine learning models outperform the traditional ΔlogR method, with GBDT achieving the highest prediction accuracy. Notably, reducing the dataset from five wells to three more homogeneous wells significantly improved model performance, highlighting the importance of data quality and geological consistency over quantity. In the second component, we develop and implement an integrated workflow for three-dimensional TOC modeling in the pre-salt section. By combining seismic interpretation, synthetic well data, and advanced geostatistical techniques, we establish a methodology for creating geologically plausible TOC distributions. Three modeling approaches—Kriging, Random Field Generation using Sequential Gaussian Simulation, and XGBoost machine learning—were compared, each offering distinct advantages for characterizing spatial heterogeneity and quantifying prediction uncertainty. The resulting 3D TOC models reveal consistent patterns of organic richness within the Itapema Formation, with higher concentrations in the middle to upper sections. This research contributes to improving TOC estimation methodologies and spatial characterization techniques, enhancing hydrocarbon potential assessment in complex basin settings. The integrated workflows developed here provide templates that can be adapted to other basins, advancing the broader field of quantitative source rock characterization.

Keywords: Total Organic Carbon, 3D Distribution, Santos Basin, Machine Learning

List of Figures

Supplementary Figures

List of Tables

Supplementary Tables

Summary

1. Introduction

The accurate prediction and spatial modeling of Total Organic Carbon (TOC) represents a paramount challenge in petroleum geoscience, with significant implications for resource exploration and development. As a key indicator of organic matter richness, TOC directly influences hydrocarbon generation potential, making its quantification and distribution modeling critical for exploration success (Tissot & Welte, 1984; Peters & Cassa, 1994). Traditional approaches to TOC estimation have relied primarily on well log-based methods or geochemical sampling at discrete points, which provide limited spatial coverage and may inadequately represent basin-scale heterogeneity (Passey et al., 2010; Rodrigues et al., 2021).

Recent advancements in machine learning techniques and geostatistical modeling have opened new avenues for improving TOC prediction accuracy and characterizing its spatial distribution. Machine learning models have demonstrated considerable success in capturing complex, multi-dimensional relationships in well-log data that are otherwise challenging to express using empirical equations (Goodfellow et al., 2016; Zhu et al., 2019). Similarly, three-dimensional geostatistical modeling techniques have evolved to incorporate multiple data sources, enabling more comprehensive representations of subsurface property distributions (Deutsch & Pyrcz, 2014; Caers, 2011). These technological advancements present opportunities to overcome limitations in traditional TOC estimation methods, particularly in geologically complex settings like the Santos Basin.

The Santos Basin presents an ideal geological context for exploring innovative approaches to TOC prediction and modeling. The basin's lacustrine depositional environment during the Early Cretaceous created favorable conditions for organic matter accumulation and preservation, particularly within the Itapema Formation (Moreira et al., 2007; Fernandes, 2017). These organic-rich intervals serve as the primary source rocks for the prolific pre-salt petroleum system. However, the complex depositional and diagenetic history of these formations, coupled with limited well data availability, poses significant challenges for accurately characterizing TOC distribution across the basin.

This research addresses these challenges through a two-pronged approach: first, by developing and comparing advanced machine learning models for TOC prediction from well logs; and second, by implementing an integrated workflow for three-dimensional TOC modeling that incorporates seismic attributes and geostatistical techniques. By leveraging methodologies from both data science and geospatial modeling, this study aims to enhance our understanding of source rock distribution in the Santos Basin pre-salt section, with broader implications for similar geological settings worldwide.

The findings from this research contribute to improving TOC estimation methodologies and spatial characterization techniques, ultimately enhancing hydrocarbon potential assessment in complex basin settings. Furthermore, the integrated workflows developed here provide templates that can be adapted to other basins and geological contexts, advancing the broader field of quantitative source rock characterization. This thesis is structured into two main results and discussion chapters, each corresponding to a scientific article. Chapter 3.1 presents research submitted to Marine and Petroleum Geology, while Chapter 3.2 contains

work in preparation for imminent submission. Together, these chapters offer comprehensive methodologies for TOC prediction and spatial modeling that address fundamental challenges in petroleum geoscience.

2. Material and Methods

This research employed a comprehensive methodological framework combining machine learning techniques for TOC prediction and geostatistical approaches for spatial distribution modeling. The study utilized data from the Santos Basin, offshore Brazil, focusing on the pre-salt section and particularly the Itapema Formation as the primary source rock interval. This section provides a synthesis of the methodological approach, while detailed descriptions of specific methods, algorithms, and workflows are presented in the respective results chapters (Sections 3.1 and 3.2), where each component of the research is thoroughly discussed

2.1. Data Sources and Preprocessing

For the machine learning component (Chapter One), the dataset comprised measurements from five wells in the Santos Basin, including gamma-ray (GR), bulk density (RHOB), deep resistivity (RT), neutron porosity (NPHI), sonic (DT), and TOC values. These wells were selected based on having the most comprehensive TOC data. Geographic coordinates (latitude and longitude) were incorporated to help detect spatial patterns between wells. Prior to model training, extensive data preprocessing was performed, including outlier detection using Density-Based Spatial Clustering of Applications with Noise (DBSCAN), an unsupervised method effective at identifying anomalous values (Ester et al., 1996; Bzdok et al., 2018). Missing values were imputed using k-nearest neighbors (KNN) imputation, selected for its effectiveness in handling complex, non-linear relationships (Troyanskaya et al., 2001; Zhang, 2012).

For the geostatistical modeling component (Chapter Two), we utilized 3D seismic data from the "R0258_3D_IARA_RTM_PSDM" survey of the pre-salt section in the Iara Complex, Sururu field. The seismic data had undergone depth migration processing and was received in SEG-Y format. Since only one well with TOC data was available within the 3D seismic survey area—insufficient for validating the workflow methodology—we generated synthetic TOC data for six wells using a custom Python script. This approach created geologically plausible TOC measurements with values constrained between 4% and 15% for the target formation, implementing sequential dependency to ensure realistic vertical profiles.

2.2. Analytical Approaches

2.2.1 Machine Learning for TOC Prediction

Three machine learning models were implemented and compared for TOC prediction:

1. Gradient Boosting Decision Tree (GBDT**)**: An ensemble learning algorithm constructing multiple decision trees sequentially, refining the model with each iteration to improve predictive performance (Friedman, 2001; 2002). The algorithm approximates the negative gradient of the loss function, allowing iterative improvement of predictions.
2. Extreme Gradient Boosting (XGBoost): An enhanced version of GBDT optimized for efficiency, accuracy, and scalability (Chen & Guestrin, 2016). XGBoost introduces additional optimizations such as regularization and second-order gradient information, making it more robust to overfitting.
3. Multi-Layer Perceptron (MLP): A neural network architecture consisting of at least three fully connected layers, driven by backpropagation to adjust network weights and minimize error between predicted and actual values (Rumelhart et al., 1986; Goodfellow et al., 2016).

Machine learning techniques, particularly ensemble methods and neural networks, have emerged as powerful tools for addressing complex non-linear relationships in geoscientific data (Karpatne et al., 2017). These methods offer significant advantages over traditional empirical equations by automatically extracting patterns from multi-dimensional data without requiring explicit model specification (Bergen et al., 2019). Ensemble methods like GBDT and XGBoost build multiple decision trees sequentially, with each new tree focusing on correcting the errors of previous trees, ultimately producing a robust composite model that captures complex relationships while mitigating overfitting (Hastie et al., 2009). This sequential error correction makes ensemble methods particularly suitable for geochemical data, where relationships between well-log responses and organic content are influenced by multiple factors including mineralogy, porosity, and thermal maturity (Zhao et al., 2017).

Principal Component Analysis (PCA), as utilized in this study, serves as a crucial dimensionality reduction technique that identifies the principal axes of variation in multivariate datasets (Abdi & Williams, 2010). Unlike simple correlation matrices that examine pairwise relationships, PCA captures the covariance structure of the entire dataset simultaneously, revealing latent patterns that might otherwise remain obscured (Jolliffe, 2002). In geoscience applications, PCA is particularly valuable for handling the multicollinearity commonly present in well-log data, where different measurements may respond to the same underlying physical properties (Szabo & Horvath, 1997). By transforming potentially correlated variables into orthogonal components, PCA enables more effective feature selection and improves the stability of subsequent machine learning models, ultimately leading to more robust TOC predictions (Zhang et al., 2023).

Hyperparameter tuning was conducted using Bayesian optimization, systematically exploring parameter combinations within predefined ranges for each model. Model performance was evaluated using multiple metrics including coefficient of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE).

As a comparative baseline, the traditional ΔlogR method (Passey et al., 1990) was implemented for estimating TOC content, providing a benchmark against which the machine learning models were compared.

2.2.2 Geostatistical Approaches for TOC Distribution Modeling

The geostatistical modeling component employed an integrated approach combining seismic interpretation, geostatistical analysis, and property modeling:

1. Seismic Interpretation and Geo-structural Modeling: Key sequence boundaries were interpreted from 2D depth-migrated seismic data to establish the structural framework of the model. Using GemPy, an open-source Python library implementing implicit geological modeling techniques (de la Varga et al., 2019), a pseudo-3D structural sequence model of the Santos Basin pre-salt section was created.
2. Geostatistical Analysis: Vertical variogram analysis was performed on synthetic well TOC data, while lateral correlation length was derived from seismic amplitude data. The variogram analysis revealed a correlation range of 79.8 meters in the vertical direction with a mean semivariogram value of -0.003. For horizontal directions, the directional variogram analysis identified the angle of maximum continuity at 0° with a correlation range of 44.2 meters. The resulting anisotropy ratio of approximately 1:1.8 for horizontal-to-vertical correlation was derived from these synthetic data. It's important to note that this ratio is used primarily to validate the workflow methodology rather than to draw definitive conclusions about the actual depositional patterns of the Itapema Formation. With real data, such analysis could potentially provide insights into whether the formation exhibits compartmentalized depositional patterns or more laterally continuous facies as commonly described in the literature.
3. TOC Distribution Simulation: Three property modeling methods were implemented and compared:
    - Kriging: Applied as a comparative baseline, providing smooth interpolation results with minimized estimation variance.
    - Random Field Generation (RFG): Implemented using Sequential Gaussian Simulation, incorporating seismic-derived spatial correlation structure.
    - XGBoost Regression: Applied as a machine learning approach to predict TOC distribution, incorporating spatial coordinates and contextual variables.

All modeling implementations were performed using open-source Python libraries including GSTools (Müller & Schüler, 2021) and scikit-gstat (Mälicke, 2021), ensuring reproducibility and transparency in the workflow.

This integrated methodological framework combines the strengths of both data science and geostatistical approaches, enabling comprehensive characterization of TOC distribution in the Santos Basin pre-salt section despite data limitations.

3. Results and Discussion

3.1. Chapter One: TOC Prediction from Well Logs Using Gradient Boosting and Neural Network in the Santos Basin, SE Brazil (Submitted to Marine and Petroleum Geology)

Title: **TOC Prediction from Well Logs Using Gradient Boosting and Neural Network in the Santos Basin, SE Brazil**

Authors

Bernardo S. Chede [a,*], Andre L. Belem [a,c], Victor Carreira [a], Ulrich G. Wortmann [d], Luiz H. Cordeiro [a], Kristoffer A.T. Hallam [a], Igor M. Venancio [e], Pedro V.A. Affonso [a], Rodrigo L. Sobrinho [e], Lara P.C. Herculano [a], Andre L.D. Spigolon [f], Ana Luiza S. Albuquerque [a,b]

a Programa de Pós Graduação em Dinâmica dos Oceanos e da Terra, Universidade Federal Fluminense, Niterói, 24210346, Brazil
b Departamento de Geologia e Geofísica, Universidade Federal Fluminense, Niterói, 24210346, Brazil
c Observatório Oceanográfico, Universidade Federal Fluminense, Niterói, RJ, Brazil
d Department of Earth Science, University of Toronto, Canada
e Programa de Geociências (Geoquímica), Universidade Federal Fluminense, Niterói 24020141, Brazil
f Centro de Pesquisas da PETROBRAS, Rio de Janeiro, RJ, Brazil

*Corresponding author. Programa de Pós Graduação em Dinâmica dos Oceanos e da Terra, Universidade Federal Fluminense, Niterói, 24210346, Brazil. E-mail address: bechede@id.uff.br (B.S. Chede).

Abstract

Accurate prediction of total organic carbon in subsurface formations is crucial for evaluating source rock quality and optimizing exploration strategies in hydrocarbon prolific basins. Traditional methods like the ΔlogR technique often require local calibration and may fail to capture the non-linear relationships between well-log parameters and TOC, leading to inaccuracies. This study applies three machine learning (ML) models—Gradient Boosting Decision Trees (GBDT), Extreme Gradient Boosting (XGBoost), and Multi-Layer Perceptron (MLP)—to predict TOC from well-log data in the Santos Basin, Brazil's largest offshore basin. We employed robust data preprocessing techniques, including outlier detection using Density-Based Spatial Clustering (DBSCAN) and feature reduction through Principal Component Analysis (PCA). Bayesian optimization was utilized for hyperparameter tuning to enhance model performance. The results indicate that all ML models outperformed the traditional ΔlogR method, with GBDT achieving the highest prediction accuracy. Reducing the dataset from five wells to three more homogeneous wells significantly improved the GBDT model's performance, underscoring the importance of data quality and relevance. This study demonstrates the potential of ML models in capturing complex, non-linear relationships in geophysical data and highlights the challenges of generalizing these models across diverse geological settings. The findings contribute to improved TOC estimation and can enhance exploration strategies in similar geological contexts.

Keywords

Total organic carbon (TOC), Gradient Boosting, Neural Network, Supervised Machine learning, Unsupervised Machine learning

3.1.1 Introduction

The accurate prediction of total organic carbon (TOC) in subsurface formations is fundamental for evaluating source rock quality and optimizing exploration strategies in hydrocarbon basins (Tissot & Welte, 1984; Peters & Cassa, 1994). As a key indicator of organic matter richness, TOC directly influences hydrocarbon generation, reservoir characteristics, and exploration success (Hood et al., 1975; Peters et al., 2005). Traditionally, TOC estimation relies on well log-based methods, such as the ΔlogR technique (Passey et al., 1990), which combines resistivity (RT) and sonic (DT) logs to identify organic-rich zones. Despite their widespread use, these empirical methods require local calibration, rely on resistivity baselines, and may yield inaccuracies due to variable geological contexts and the inherent limitations in capturing non-linear relationships between well-log parameters and TOC (Schmoker & Hester, 1983; Mahmoud et al., 2017). Recent advancements, such as modifications to ΔlogR (Wang et al., 2016), have been proposed to address these limitations, but they still demand external calibration through thermal maturity indicators like vitrinite reflectance (Ro) or Tmax (Hood et al., 1975; Hunt, 1995).

Given the increasing availability of high-resolution geophysical data and advancements in computational methods, machine learning (ML) models have emerged as a promising alternative for predicting TOC. Unlike traditional approaches, ML techniques can capture complex, multi-dimensional relationships in well-log data that are otherwise challenging to express using empirical equations (Goodfellow et al., 2016; Zhu et al., 2019). Among these, ensemble learning algorithms such as Gradient Boosting Decision Trees (GBDT) and its variant Extreme Gradient Boosting (XGBoost) have demonstrated considerable success in geophysical applications (Chen & Guestrin, 2016; Sun et al., 2023). These models iteratively improve prediction accuracy by minimizing residual errors in successive stages, making them particularly suited for datasets with heterogeneous and non-linear characteristics (Friedman, 2001; Hasan & Karim, 2021; Pan et al., 2022). Similarly, neural networks like Multi-Layer Perceptrons (MLP) have shown strong potential in generalizing complex relationships between well logs and subsurface properties, but their efficacy often depends on dataset size and quality (Hinton et al., 2006; Zhang et al., 2016).

However, the application of ML models for TOC prediction is not without challenges. Noisy well-log data, missing measurements, and spatial heterogeneity across geological formations can hinder model performance (Kamali & Mirshady, 2004; Zhao et al., 2020). Addressing these issues requires robust data preprocessing techniques, including outlier detection using Density-Based Spatial Clustering (DBSCAN) and feature reduction through Principal Component Analysis (PCA) to improve data quality and reduce dimensionality (Ester et al., 1996; Jolliffe & Cadima, 2016). Advanced hyperparameter optimization methods, such as Bayesian optimization, have proven essential in fine-tuning models to balance prediction accuracy and generalization (Snoek et al., 2012; Bergstra & Bengio, 2012).

The Santos Basin, the largest offshore basin in Brazil, presents an ideal geological context for exploring innovative approaches to TOC prediction (Wright & Barnett, 2015; ANP, 2020). The basin's complexity, characterized by diverse lithologies and depositional settings, presents both opportunities and challenges for accurately predicting TOC from well logs (Moreira et al., 2007; Buckley et al., 2015). While traditional methods such as ΔlogR have been successfully applied in some regions, recent studies indicate their limitations in capturing the basin's geological variability (Fernandes, 2017; Venancio et al., 2022a). Thus, applying advanced ML models offers a promising avenue for improving TOC prediction accuracy in such complex environments.

This study investigates the application of three ML models—GBDT, XGBoost, and MLP—to predict TOC from well-log data in the Santos Basin. The study aims to evaluate the models' performance in capturing non-linear relationships between well-log features and TOC, compare them with the conventional ΔlogR approach, and assess the impact of data preprocessing and feature engineering on model accuracy. By leveraging techniques such as PCA for dimensionality reduction and Bayesian optimization for hyperparameter tuning, this research seeks to establish a robust framework for ML-based TOC prediction. The findings could enhance TOC estimation in similar geological contexts and contribute to the growing field of geochemical prediction using data-driven models.

3.1.2 Geological setting of the Santos Basin

The eastern Brazilian basins are classified as continental rift basins that were formed during the Lower Cretaceous, in association with the progressive breakup of the Gondwana Supercontinent and the opening of the South Atlantic Ocean (Moulin et al, 2005). These basins record the deposition of thick continental, fluvial, and lacustrine sediments, intercalated with volcanic rocks (Chaboureau, 2013). Following the rifting phase, the basins underwent a thermal subsidence phase, characterized by gravitational slumping and further sedimentary accumulation, which shaped their ongoing evolution (Estrella et al., 1984).

The Santos Basin, located along the southeastern Brazilian margin, is the largest offshore basin in Brazil, covering more than 350,000 square kilometers (Fig. 1). It spans the coastal regions of Rio de Janeiro, São Paulo, Paraná, and Santa Catarina, and is bounded by the Cabo Frio structural high to the north and the Pelotas structural high to the south (Moreira et al., 2007). This basin, formed during the Lower Cretaceous, is significant not only due to its size but also because it is the primary oil and gas-producing basin in Brazil (ANP, 2020). The extensional stresses that occurred during the rifting of Gondwana led to the deposition of sediments in saline and freshwater lake environments, comprising both siliciclastic and carbonate materials (Buckley et al., 2015; Wright & Barnett, 2015).

The presalt section of the Santos Basin has drawn significant attention due to its prolific hydrocarbon reservoirs, which are associated with two primary depositional phases: the rift and sag phases (Moreira et al., 2007). During the Late Barremian to Early Aptian, the Itapema Formation was developed. characterized by interbedded carbonates and organic-rich shales. This formation serves as the main source rock with total organic carbon (TOC) values reaching as high as 16%, making it a major contributor to hydrocarbon generation in the basin (Fernandes, 2017). The reservoirs themselves are found primarily in

the Barra Velha Formation, which consists predominantly of carbonates deposited during the Aptian sag-phase, enhancing the basin's petroleum migration. Overlying these reservoirs, the Ariri Formation, with evaporite thickness exceeding 2,000 meters, provides an effective seal, trapping hydrocarbons within structural highs and faults generated during the initial stages of rifting, thus ensuring their preservation (Freitas et al., 2022).

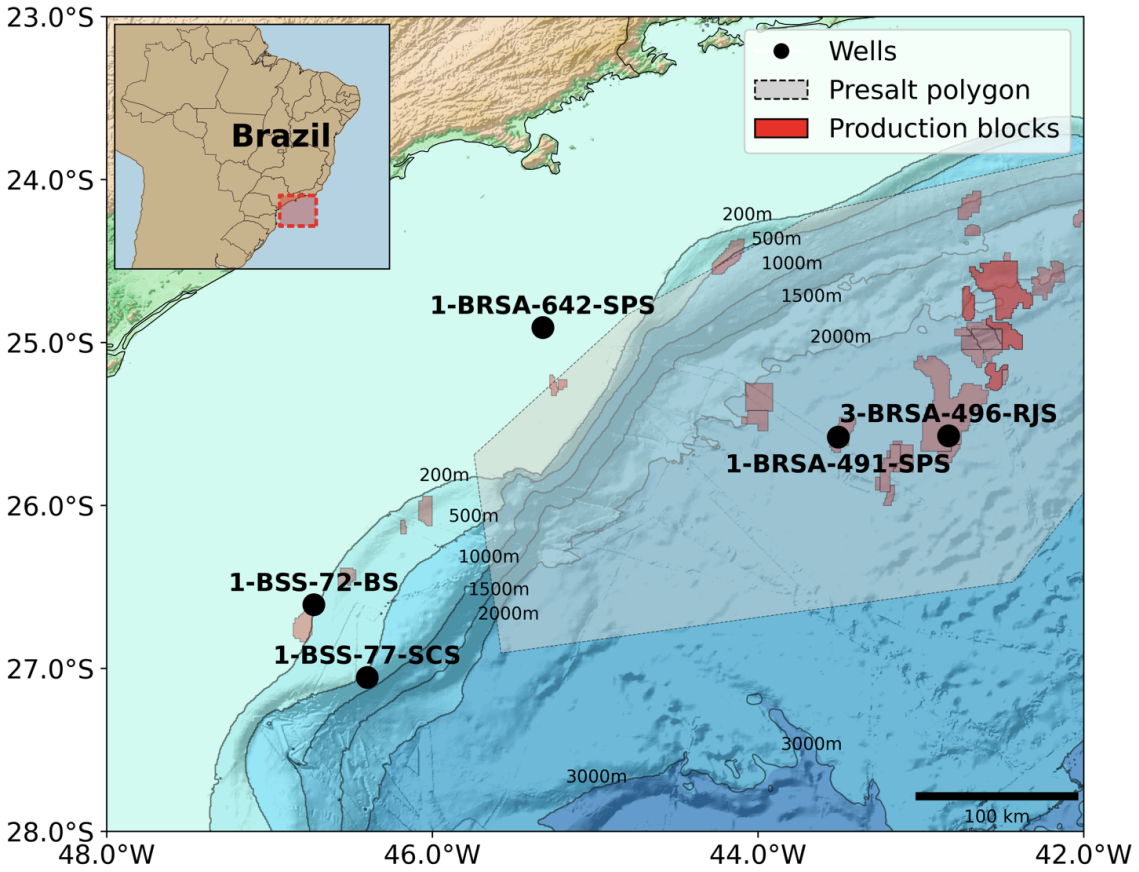


Fig. 1. Location map of the Santos Basin offshore Brazil, highlighting the wells analyzed in this study (black dots). The gray polygon represents the Pre-salt geological section, and the red areas indicate the production blocks.

3.1.3 Methods and data

The dataset consists of five wells (Fig. 1) with measurements of TOC, gamma-ray (GR - gAPI), bulk density (RHOB - g/cm³), deep resistivity (RT - ohm/m), neutron porosity (NPHI - %), and sonic (DT - µs/ft). These wells were selected based on having the most comprehensive TOC data. Latitude and longitude were incorporated to help the model detect spatial patterns between the wells and capture the geographic distribution (Cressie, 1993). However, including spatial coordinates poses the risk of spatial overfitting, where the model may focus on the specific locations rather than identifying the broader geological trends, leading to strong performance on the training data but poor generalization to unseen areas (Pebesma & Bivand, 2023). Prior to model training, feature engineering was applied to the input data, and Fig. 2 shows the workflow.

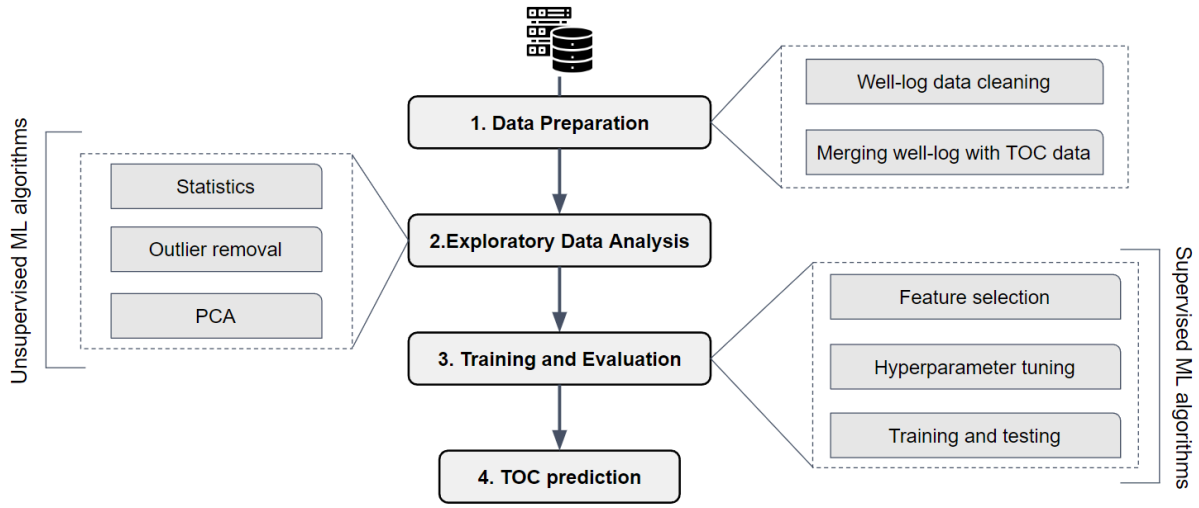


Fig. 2. Workflow chart of this work.

Data preparation is a crucial initial step in ensuring high-quality input for ML models (Damasceno et al., 2022), especially given that well log data often contain irregularities such as washouts and other outlier measurements (Bhattacharya et al. 2022; Reis et al., 2023). The data were preprocessed using the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm, an unsupervised method that effectively detects and isolates anomalous values deviating from the expected distribution (Bzdok et al., 2018). DBSCAN's capacity to identify outliers with high precision without assumptions about the data distribution makes it particularly suited for addressing the complexities and inconsistencies commonly found in well log measurements (Ali et al. 2023). This approach enhances the dataset's integrity by highlighting patterns within the data while isolating values that do not conform to the expected behavior (Ester et al., 1996; Hahsler et al., 2019).

DBSCAN was specifically chosen over other outlier detection methods such as Isolation Forest, Local Outlier Factor, or simple statistical thresholds due to its unique ability to identify outliers based on local density considerations rather than global distribution parameters (Schubert et al., 2017). This characteristic is particularly advantageous for well-log data, where anomalies may manifest differently across varying lithologies and depth intervals (Ali et al., 2023). Unlike methods that assume specific distribution shapes or rely on distance to decision boundaries, DBSCAN's density-based approach can identify irregular clusters of valid measurements while flagging isolated points as potential outliers, regardless of their absolute values (Hahsler et al., 2019). This flexibility allows DBSCAN to effectively handle the heteroscedasticity common in well-log measurements, where the variance of measurements often changes with depth or across different formations (Liu et al., 2012). Furthermore, DBSCAN does not require a predefined number of clusters, making it more suitable for exploratory outlier detection in geoscientific datasets where the underlying distribution structure may not be known a priori (Ester et al., 1996).

After outlier removal, missing values were imputed using the k-nearest neighbors (KNN) imputation method (Troyanskaya et al., 2001). KNN was selected for its simplicity and effectiveness in handling complex, non-linear relationships by leveraging information from nearby data points with similar characteristics (Zhang, 2012). This approach can provide more accurate imputations compared to other methods (Jadhav et al., 2019). KNN can be computationally intensive when applied to large datasets, and the choice of the k-value plays

a critical role in the quality of the imputations, as an inappropriate k can lead to either overfitting or underfitting (Kim & Cho, 2024).

To assess the relationships among the feature variables and their contribution to TOC variability, we employed Principal Component Analysis (PCA), an unsupervised dimensionality reduction technique that transforms potentially correlated variables into a new set of orthogonal components that maximize variance (Jolliffe, 2002). One of the main advantages of using PCA over a simple correlation matrix lies in its ability to provide a comprehensive understanding of complex datasets by revealing patterns and structures not evident through pairwise correlations alone. Correlation analysis, while valuable, can be misleading in the presence of multicollinearity or non-linear relationships, which are common challenges in geophysical datasets (Jolliffe & Cadima, 2016). By transforming the data into uncorrelated components, PCA overcomes these limitations and uncovers latent structures that inform feature selection and model development. This approach is particularly advantageous for evaluating features in geophysical datasets, where multicollinearity and complex relationships are common (Szabo & Horvath, 1997; Abdi & Williams, 2010; Lever et al., 2017; Khan et al., 2019; Skrjanc et al., 2020; Maciejowska et al., 2020; Zhang et al., 2023). By analyzing the structure of the dataset using PCA, we aimed to gain insights into the influence of each variable and guide feature selection for the supervised models.

The final dataset comprises eight features, with descriptive statistics summarized in Table 1, grouped into three categories: (1) well logs commonly used in TOC modeling (GR, DT, RT, RHOB, NPHI); (2) geographic coordinates (LAT, LON); and (3) transformed variables, where each well was assigned a numerical identifier using the LabelEncoder from the sklearn library (WELL_ID) (Pedregosa et al., 2011). The first group of features, consisting of well logs, corresponds to physical rock properties that have been shown to correlate with TOC content (Schmoker, 1979; Passey et al., 1990; Carpentier et al., 1991). The second group includes geographic positioning variables, which may reflect spatial variations in organic matter production, degradation, or maturation processes (Venancio et al., 2022a; Wen et al., 2024). The final group contains encoded discrete information designed to capture well-specific characteristics. This dataset was used to train Gradient Boosting Decision Tree (GBDT), XGBoost, and Multi-Layer Perceptron (MLP) models, with the traditional ΔlogR method (Passey et al., 1990) applied as an alternative approach to TOC modeling.

Table 1. Descriptive statistics of the dataset used for model training and testing, with imputed values shown in parentheses.

	COUNT	**MEAN**	**STD**	**MIN**	**MAX**
WELL_ID	1386	-	-	1	5
LAT (◦)	1386	-26.03	0.71	-27.06	-24.91
LON (◦)	1386	-45.17	1.57	-46.73	-42.83
DEPTH (m)	1386	4330.05	1307.82	549	5718
COT (%)	1386	0.69	0.9	0.06	13.83
GR (gAPI)	1363 (1386)	43.69 (43.43)	24.03 (23.82)	5.74 (5.74)	125.86 (125.86)
RHOB (g/cm3)	1197 (1386)	2.59 (2.59)	0.14 (0.1)	0.12 (2.12)	3.0 (3.0)

DT (µs/ft)	1378 (1386)	70.57 (70.35)	22.68 (22.41)	42.3 (42.3)	179.5 (179.5)
RT (ohm/m)	1300 (1386)	302.0 (190.99)	518.72 (307.1)	0.19 (0.21)	1950.0 (1484.96)
NPHI (%)	925 (1386)	12.47 (14.11)	7.99 (7.66)	0.0 (0.0)	55.32 (40.01)

3.1.3.1 Gradient Boosting Decision Tree

Gradient Boosting Decision Tree (GBDT) is an ensemble learning algorithm that constructs multiple decision trees sequentially, refining the model with each iteration to improve predictive performance (Friedman, 2001; 2002). In every iteration, GBDT fits a new decision tree to the negative gradient of the loss function concerning the current model's predictions, effectively minimizing the residual errors between the true and predicted values. However, GBDT models can be susceptible to overfitting, especially when trained on complex datasets. To mitigate overfitting, several regularization techniques are employed, including shrinkage (learning rate), complexity penalties (such as limiting tree depth or the number of leaves), and subsampling (stochastic gradient boosting), which improves model generalization by introducing randomness in the data sampling process (Zhang & Zung, 2020).

The core principle of GBDT is to approximate the negative gradient of the loss function, which simplifies the optimization process by converting a complex problem into a series of simpler steps (Delcroix et al., 2021). This approximation is crucial as it allows the model to improve predictions iteratively without requiring complex second-order derivatives. During training, when a sample passes through a decision tree, the predicted result may differ from the actual value, and these residuals are key to guiding the construction of subsequent trees (Xia et al., 2024). To further prevent overfitting, early stopping is often employed, which halts the training process when the validation error stops improving. Early stopping is configured by monitoring the performance on a validation set and terminating the training once the model reaches an optimal balance between training accuracy and generalization performance.

The core principle of GBDT is to approximate the negative gradient of the loss function, simplifying the optimization process by converting a complex problem into a series of simpler steps (Delcroix et al., 2021). This gradient approximation is crucial as it allows the model to iteratively refine predictions without requiring second-order derivatives. During training, when a sample passes through a decision tree, its predicted result may differ from the actual value, and these residuals guide the construction of subsequent trees (Xia et al., 2024). To further prevent overfitting, early stopping is often employed, halting training when the validation error stops improving. Early stopping is configured by monitoring performance on a validation set and terminating training once the model reaches an optimal balance between accuracy and generalization performance.

Mathematically, GBDT can be expressed as:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (1)$$

Where:

- $F_m(x)$ is the boosted model after m iterations;
- $F_{m-1}(x)$ is the model from the previous iteration;
- γ_m is the learning rate (shrinkage factor);
- $h_m(x)$ is the new tree added at iteration m, trained on the residuals of the previous model.

The full objective function that GBDT aims to minimize can be represented as:

$$L\left(y_i, F\left(x_i\right)\right) = \sum_{i=1}^{N} \ell\left(y_i, F_{m-1}\left(x_i\right)\right) + \Omega\left(h_m\right) \quad (2)$$

Where:

- $\ell\left(y_i, F_{m-1}\left(x_i\right)\right)$ is the loss function measuring the difference between true values y_i and predictions $F(x_i)$;
- $\Omega\left(h_m\right)$ is the regularization term penalizing model complexity to prevent overfitting.

Each part of this formula contributes to improving model performance and addressing overfitting in the following ways:

- Learning Rate (Shrinkage - γ_m): Shrinkage scales the contribution of each tree by a factor (learning rate), preventing any single tree from overly dominating the model's predictions. Lower learning rates help to avoid overfitting by making the optimization more gradual.

- Complexity Penalties ($\Omega\left(h_m\right)$): Limiting the depth of the trees or the number of leaves in each tree ensures that the individual trees do not become too complex, reducing the risk of overfitting to noise in the training data.
- Stochastic Gradient Boosting: By introducing randomness into the data subsampling (such as using a random fraction of the data for each tree), the model can further enhance its generalization capabilities.
- Limiting the Ensemble: The total number of trees (M) can be restricted to avoid unnecessary model complexity. More trees can improve accuracy, but excessive trees without proper regularization can increase the risk of overfitting.
- Early Stopping and Validation: Early stopping halts the training process when the performance on a separate validation set begins to degrade, ensuring that the model does not overfit the training data.

3.1.3.2 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an enhanced version of the GBDT algorithm, optimized for greater efficiency, accuracy, and scalability. It effectively handles both classification and regression tasks by leveraging Classification and Regression Trees (CART), which can process both continuous and categorical data. Due to these enhancements, XGBoost has become one of the most widely used machine learning

algorithms for TOC prediction (Liu et al., 2021; Sun et al., 2023; Khan et al., 2023; Bione et al., 2024). Like GBDT, XGBoost iteratively fits new trees to the residuals of previous models, progressively improving the overall prediction accuracy. However, XGBoost introduces additional optimizations, such as regularization and second-order gradient information, which make it more robust to overfitting and capable of handling large datasets with higher computational efficiency (Chen et al., 2015).

A key innovation in XGBoost is its use of second-order derivatives (Hessians) in the loss function optimization. While GBDT uses only first-order gradients, XGBoost's second-order approximation through Taylor expansion enables more accurate estimation of the optimal step size for each iteration. For MSE loss function used in TOC prediction, the Hessian equals 2 for all observations, but its incorporation still improves convergence by providing information about the loss function's curvature (Nielsen, 2016). This second-order approach is particularly valuable for modeling the complex, non-linear relationships between well-log parameters and TOC content.

XGBoost, as an ensemble learning method, combines the predictions of multiple weak learners (decision trees) to produce a strong model. The core principle is based on gradient descent, where new trees are built by learning from the residuals of the previous trees, ultimately minimizing the objective function (Chen and Guestrin, 2016). The objective function in XGBoost consists of two main components: the training loss function, which measures the model's performance on the training data, and a regularization term, which controls the complexity of the model and helps prevent overfitting. This balance between the loss function and regularization is essential in ensuring that the model generalizes well to unseen data (Zhang & Gong, 2020).

The regularization term in XGBoost plays a crucial role in controlling overfitting by penalizing both the number of leaves in each tree and the magnitude of the scores assigned to the leaves. The objective function can be expressed as:

$$\mathcal{L}(\omega) = \sum_{i=1}^{n} \ell\left(y_i, \widehat{y}_i\right) + \sum_{k=1}^{T}\left(\frac{1}{2}H_k\omega_k^2 + \lambda\sum_{k=1}^{T}\omega_k + \gamma T\right) \quad (3)$$

Where:

- T is the number of leaves in the tree,
- $\sum_{i=1}^{n} \ell\left(y_i, \widehat{y}_i\right)$ is the sum of the gradients (first-order derivative of the loss function),
- H is the sum of the Hessians (second-order derivative of the loss function),
- ω_k is the vector of scores on leaves,
- λ is the regularization parameter controlling the magnitude of the leaf scores,
- γ penalizes the number of leaves to control tree complexity.

This formulation highlights the balance between model accuracy and complexity control. The term $\frac{1}{2}H_k\omega_k^2$ controls the penalization of leaf scores (preventing overfitting by reducing the impact of outliers or overly confident predictions), while λ and γ regulate the complexity of

the model by penalizing large trees. By carefully managing these regularization components, XGBoost effectively mitigates overfitting, ensuring better generalization on new data.

The key distinction between GBDT and XGBoost lies in their regularization approaches. XGBoost implements a more sophisticated regularization term that combines two components: one that controls the number of leaves in the tree and another that penalizes large leaf weights through L2 regularization (Zhang & Gong, 2020). This dual regularization strategy enables XGBoost to simultaneously control both structural complexity (number of leaves) and prediction magnitude (leaf weights), while traditional GBDT primarily limits complexity through constraints like maximum depth or minimum samples per node. Additionally, XGBoost employs split pruning techniques that evaluate potential splits against regularization penalties, offering superior prevention of overfitting in data-limited settings like TOC prediction from sparse well logs (Ke et al., 2017).

3.1.3.3 Multi-Layer Perceptron

Similarly to XGboost, the MLP employ regularization techniques to prevent overfitting and improve generalization, though their approaches differ. While XGBoost uses tree-specific penalties, such as constraints on the number of leaves and leaf scores through parameters like γ and λ, MLP applies a more general L2 regularization, also known as weight decay, which controls the magnitude of the weights across the network (Ng, 2004). Despite these differences, both models share the goal of balancing model complexity and accuracy, with XGBoost specializing in decision trees and MLP in neural networks.

The MLP, initially inspired by biological neural networks (Rosenblatt, 1958), consists of at least three fully connected layers: an input layer, one or more hidden layers, and an output layer. These layers form the basis of a feedforward neural network, where data flows sequentially from input to output, making MLP suitable for a wide range of regression and classification tasks. The learning process is driven by backpropagation, which adjusts the network's weights to minimize the error between predicted and actual values (Rumelhart, Hinton, & Williams, 1986). This optimization is carried out through gradient descent, where the weights are updated iteratively to minimize the loss function (Goodfellow, Bengio, & Courville, 2016)

The objective function of the MLP consists of two main components: the loss function and a regularization term. The loss function measures how well the model fits the training data, while the regularization term penalizes large weights, helping to avoid overfitting by discouraging excessively large weight values (Bishop, 1995). This balance between model complexity and accuracy is crucial for ensuring that the network generalizes effectively to unseen data.

The objective function of the MLP can be formulated as:

$$L = \frac{1}{n}\sum_{i=1}^{n}\left(y_i, \widehat{y}_i\right)^2 + \lambda\sum_{k}\left(W_k^2\right) \quad (4)$$

Where:

- n is the number of training examples,

- $L\left(y_i, \widehat{y}_i\right)$ represents the loss function (e.g., MSE),
- y_i is the true value, and $\widehat{y}_i$ is the predicted value,
- W_k^2 represents the weights of the k-th layer, and
- λ is the regularization parameter.

This objective function highlights how MLP combines the minimization of prediction errors with regularization to control complexity. While XGBoost regularizes based on the structure of decision trees, MLP's L2 regularization focuses on constraining the network's weights. These differing approaches reflect the distinct characteristics of each model, yet both aim to enhance generalization and prevent overfitting (Hastie, Tibshirani, & Friedman, 2009).

3.1.3.4 Hyperparameter Tuning and Model Evaluation for Performance Optimization

While the GBDT, XGBoost, and MLP algorithms are highly effective ML models for regression and classification tasks, their predictive performance can vary significantly depending on the chosen hyperparameters (Friedman, 2001; Chen & Guestrin, 2016). Hyperparameter tuning is essential for optimizing these models, enabling the identification of settings that minimize both bias and variance, thus enhancing accuracy and generalization (Geman et al., 1992). Equally important is evaluating these models' performance using reliable metrics and validation techniques to ensure they generalize well to unseen data (Kohavi, 1995).

In this study, we employed Bayesian optimization for hyperparameter tuning and stratified K-fold cross-validation for model evaluation. Bayesian optimization was chosen due to its ability to construct a probabilistic model of the objective function, efficiently navigating the hyperparameter space by balancing exploration and exploitation to identify optimal configurations (Snoek et al., 2012; Snoek et al., 2014). This approach contrasts with traditional methods such as grid search (Liashchynskyi & Liashchynskyi, 2019) and random search (Bergstra & Bengio, 2012), which can be less efficient in high-dimensional spaces. The optimization process was implemented using the BayesianOptimization library (Nogueira, 2014).

Cross-validation is critical in ML workflows, as it assesses a model's generalization capability to independent datasets, ensuring robust and unbiased estimates of the models performance (Goodfellow et al., 2016). We employed a 10-fold split with a random state to ensure reproducibility, implemented via the StratifiedKFold function from scikit-learn (Pedregosa et al., 2011). The choice of k = 10 folds were motivated by a balance between bias and variance in model evaluation, as this number offers an optimal trade-off, providing low variance without significantly increasing computational cost (Kohavi, 1995). Also, this k of folds is widely adopted, particularly when the dataset size is moderate, allowing the models to be trained on 90% of the data and validated on the remaining 10%, and ensuring sufficient data for both training and validation while maintaining computational efficiency (Refaeilzadeh et al., 2009).

Our target variable, TOC, is treated as continuous within the context of our dataset, exhibiting a wide range of values. However, stratified K-fold cross-validation requires stratification based on class labels. To accommodate this requirement, we discretized the TOC values into ten intervals (bins) using equal-frequency binning. Forman and Scholz (2010) suggest that discretizing continuous variables can enhance stratification in cross-validation by ensuring that each fold contains a representative distribution of the target variable. By discretizing TOC, we maintained the integrity of the stratification process, thereby enhancing the reliability of performance estimates.

The hyperparameter tuning process involved running each model for 500 iterations. This number was chosen to balance computational efficiency with the need for sufficient exploration of the hyperparameter space, as a larger number of iterations increases the probability of finding near-optimal hyperparameters but also demands more computational resources (Bergstra & Bengio, 2012). For each selected hyperparameter, we explored combinations of values within a specified window range (Supplementary Tables 1-3). These ranges were determined based on a combination of literature recommendations, prior domain knowledge, and initial experimentation. For instance, typical values known to influence model performance significantly were included to ensure the optimization process effectively searched the most promising regions of the hyperparameter space (Snoek et al., 2012).

The objective of the optimization was to maximize the coefficient of determination (R^2), as it quantifies the proportion of variance in the dependent variable that is predictable from the independent variables (Cameron & Windmeijer, 1997). After training, we evaluated the models' performance using several metrics: R^2, mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). These metrics are widely recognized in regression analysis for assessing predictive accuracy and model reliability (Chai & Draxler, 2014; Hyndman & Koehler, 2006). Models achieving higher R^2 and lower MAE, RMSE, and MAPE values are considered to exhibit better predictive performance, indicating a closer alignment between predicted and actual values (Willmott & Matsuura, 2005). The equations for these metrics are presented in Equations 5 to 8 below.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \quad (5)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right| \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2} \quad (7)$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \quad (8)$$

3.1.3.5 ΔlogR method

To provide a comparative baseline and complement the ML models in this study, we applied the ΔlogR method for estimating TOC content, widely used in the industry for evaluating source rock potential and quantifying TOC from well logs (Passey et al., 1990). This method

has been successfully applied in various geological settings and is recognized for its effectiveness in regions with organic-rich source rocks (Kamali & Mirshady, 2004; Khoshnoodkia et al., 2011; Rui et al., 2019)

The ΔlogR method involves overlaying an appropriately scaled porosity log-typically the sonic transit time (DT) curve - onto a resistivity (RT) curve using a specific overlap coefficient. In water-saturated, organic-lean rocks, these two curves run parallel, establishing a baseline. Deviations from this baseline occur in organic-rich intervals due to increased RT and DT caused by the presence of kerogen and hydrocarbons. The distance between the two logs in these intervals is proportional to the TOC content. The key measure in this technique is the ΔlogR, calculated using the following equations 9 and 10 (Passey et al., 1990):

$$\Delta logR = log_{10}\left(\frac{R}{R_{baseline}}\right) + 0.02 \times \left(\Delta t - \Delta t_{baseline}\right) \quad (9)$$

$$TOC = \Delta logR \times 10^{(2.297 - 0.1688 \times LOM)} \quad (10)$$

Where:

- $\Delta logR$ is the difference in log resistivity,
- R and $R_{baseline}$ are the measured and baseline RT values,
- Δt and $\Delta t_{\text{baseline}}$ are the measured and baseline DT values,
- LOM is the Level of Organic Maturity (Hood et al., 1975)
- 0.02 is a constant overlap coefficient specific to the method.

The LOM accounts for the thermal maturity of the organic matter and is critical for accurate TOC estimation. The ΔlogR method has been validated and calibrated in numerous studies, demonstrating its reliability in estimating TOC from well logs (Peters et al., 2005; Abbaszadeh et al., 2016; El Diasty & Ragab, 2019).

In this study, we implemented the ΔlogR method using the script provided by Bione et al. (2024), adapting it to the specific characteristics of each well in our dataset. This involved calibrating baseline values for RT and DT logs and determining appropriate LOM values based on regional geological data. By applying this method, we aimed to establish a benchmark for TOC estimates against which the performance of our ML models could be compared.

3.1.4 Results and discussion

3.1.4.1 Data points, imputation, and feature analysis

After data collection, preprocessing, and feature engineering, the final dataset comprises 1,386 entries, each representing a TOC measurement. Before imputing missing data, the dataset completeness was 66.74%, with DT having the fewest missing values at 0.58%, while the NPHI had the most at 33.26% (Table 1). The TOC values ranged from a minimum of 0.06% to a maximum of 13.83%. The distribution of TOC, however, was skewed towards lower values, with a mean of 0.69%, a median of 0.34%, and a standard deviation of 0.90%, indicating that while most TOC measurements are relatively low, some higher values contribute to increased variability.

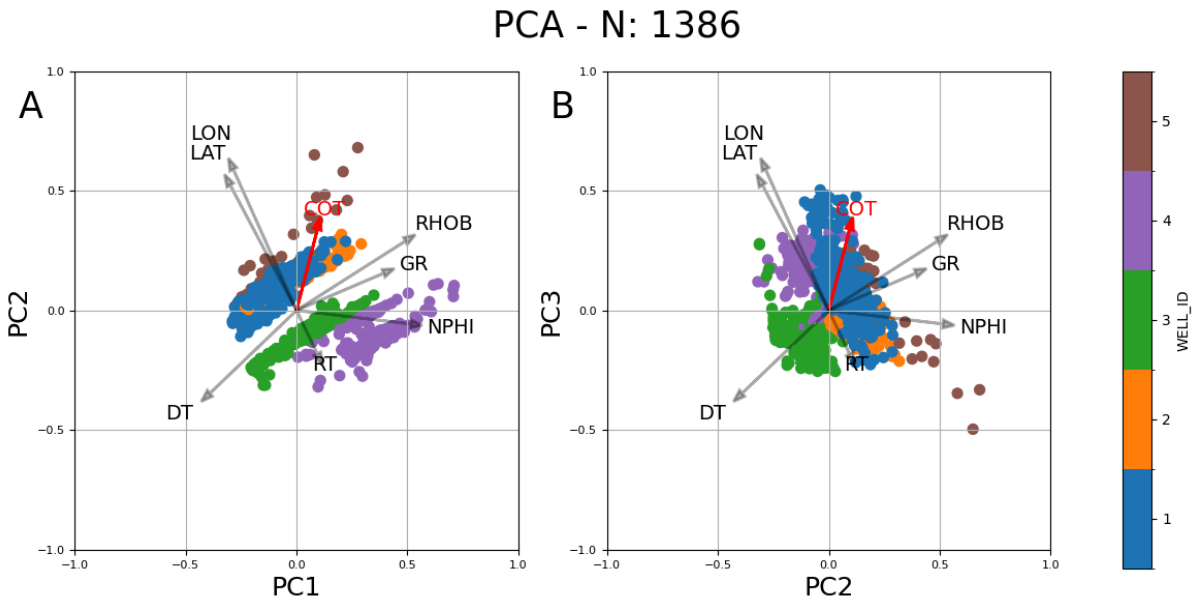


Fig. 3. Principal Component Analysis (PCA) and correlation analysis of the dataset (N = 1386). (A) Plot of PC1 vs PC2. (B) Plot of PC2 vs PC3, with explained variance of 36.8%, 20.9% and 17%, respectively. Gray arrows indicate the direction and magnitude of feature contributions, while the red arrow represents the target variable, COT. Dot colors correspond to individual wells, as indicated by the legend on the right. (C) Spearman's rank correlation coefficients (ρ) between TOC and well-log parameters, showing the strength and direction of relationships.

The PCA results indicated that the first three principal components (PCs) accounted for 75% of the total variance (Fig. 3). In Fig. 3A, the plot of PC1 against PC2 highlights the contributions of various features to the primary dimensions. The variable loadings show that GR, DT, RHOB, and NPHI significantly contributed to the variance captured by these components. This multivariate relationship aligns with established petrophysical principles, where organic-rich intervals typically exhibit higher GR values due to uranium association with organic matter (Schmoker, 1981), extended DT readings from lower matrix velocity (Passey et al., 2010), reduced RHOB measurements reflecting lower density organic material (Vernik and Milovac, 2011), and elevated NPHI values stemming from hydrogen concentration in kerogen (Sondergeld et al., 2010). Additionally, Fig. 3B presents the relationship between PC2 and PC3, revealing that TOC also exhibits a strong association with GR, DT, RHOB, and NPHI. This suggests that these features are influential in explaining TOC variability and could serve as important predictors in ML models aimed at TOC estimation.

The application of PCA as an unsupervised machine learning technique in this study served multiple purposes beyond simple dimensionality reduction (Jolliffe & Cadima, 2016). Unlike supervised learning methods that require labeled data for training, PCA operates without prior knowledge of class labels, making it ideal for exploratory data analysis of well-log parameters (Abdi & Williams, 2010). By examining the structure of variance in the multivariate dataset, PCA reveals intrinsic patterns that might not be immediately apparent through pairwise correlations or visual inspection of individual logs (Lever et al., 2017). This unsupervised approach is particularly valuable in geoscientific applications where the

underlying physical relationships between variables may be complex and non-linear (Maciejowska et al., 2020).

In the context of TOC prediction, PCA serves as a crucial preliminary step before applying supervised machine learning algorithms (Khan et al., 2019). The transformation of potentially correlated well-log parameters into orthogonal components addresses multicollinearity issues that could otherwise compromise the stability and interpretability of regression models (Saporta & Niang, 2009). Furthermore, the visualization of data in PC space (Fig. 3) reveals the absence of clear linear relationships between input features and TOC values, justifying the application of more sophisticated machine learning techniques capable of capturing non-linear patterns. The dispersion of data points in the PC projection, with substantial overlap between different TOC ranges, indicates that simple linear models would be insufficient for accurate prediction, thereby validating our approach of employing ensemble methods and neural networks (Zhao et al., 2020).

The loading vectors in the PCA biplots (Fig. 3) represent the contribution of each original variable to the principal components, with longer vectors indicating stronger influence. The angles between these vectors provide insight into the relationships between variables - small angles suggest positive correlation, perpendicular vectors indicate independence, and opposing vectors imply negative correlation. In our analysis, the clustering of GR, DT, and NPHI vectors in similar directions indicates their positive correlation, while the opposing RHOB vector confirms its expected negative relationship with these parameters in organic-rich intervals. The positions of data points along these loading vectors reflect their relative values for each parameter, allowing identification of samples with particularly high or low values. Notably, the TOC vector (shown in red) demonstrates a moderate alignment with GR and DT, supporting their known relationships with organic content, while showing some orthogonality to other parameters, suggesting additional factors influence organic matter preservation beyond these conventional logs (Zhang et al., 2023).

To complement the qualitative insights from PCA with quantitative measures, Spearman's rank correlation coefficients were calculated between TOC and each well-log parameter (Fig. 3C). This non-parametric correlation metric was selected over Pearson's coefficient due to its robustness against non-linear relationships and outliers, both common in geoscientific datasets (Schober et al., 2018). The analysis revealed moderate positive correlations between TOC and GR ($\rho = 0.50$) and DT ($\rho = 0.42$), a weak negative correlation with RHOB ($\rho = -0.21$) and RT ($\rho = -0.27$), and a moderate positive correlation with NPHI ($\rho = 0.32$). These quantitative relationships align with the qualitative observations from the PCA biplots and confirm the expected petrophysical relationships: higher gamma ray values in organic-rich intervals due to uranium association, extended sonic transit times reflecting lower matrix velocity, reduced bulk density from low-density organic material, and elevated neutron porosity from hydrogen in kerogen (Passey et al., 2010). The moderate strength of these correlations ($|\rho| < 0.6$) further supports the need for multivariate machine learning approaches rather than simple univariate models, as no single log parameter demonstrates a sufficiently strong relationship to predict TOC independently (Mahmoud et al., 2017).

3.1.4.2 Hyperparameter Tuning and Model Evaluation

Hyperparameter tuning was conducted using Bayesian optimization, systematically exploring parameter combinations within predefined ranges for each model to enhance performance. The top ten configurations and their optimized hyperparameters, evaluated based on average RMSE from 10-fold cross-validation, are presented in Supplementary Tables 4–6.

Model convergence was assessed by tracking RMSE values over 500 optimization iterations (Fig. 4). The XGB and GBDT models exhibited negligible improvement beyond approximately 80 and 200 iterations, respectively (Fig. 4A,B). In contrast, the MLP model showed minor enhancements up to 500 iterations (Fig. 4C). However, the incremental RMSE reductions were minimal (XGB: 0.031%, GBDT: 0.024%, MLP: 0.0032%) relative to the computational time required for 500 iterations (XGB: 17 min, GBDT: 52 min, MLP: 230 min). Consequently, the optimization process for all models can be truncated to 200 iterations without significantly compromising performance.

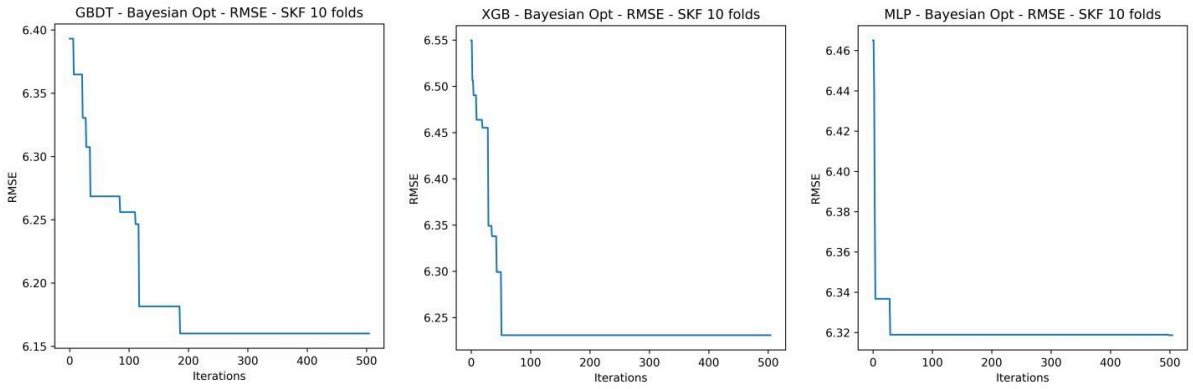


Fig. 4. RMSE Convergence of Bayesian Optimization for GBDT, XGB, and MLP Models Using 10-Fold Stratified Cross-Validation.The X-axis represents the number of optimization iterations, while the Y-axis shows the RMSE values for each model.

3.1.4.3 Comparison of Models

The performance comparison of the models in predicting TOC is illustrated in Fig. 5A, D, and G. Both XGB and GBDT models exhibit relatively similar predictive capabilities; however, GBDT demonstrated a slightly higher maximum R^2 value, indicating that it occasionally achieved better fits to the observed data. Although the average R^2 values for XGB and GBDT were comparable, the MLP model showed notably lower R^2 values, indicating a reduced ability to capture the variability in TOC compared to the ensemble methods. All detailed performance metrics for each model can be found in Table 2. The RMSE values further substantiate these observations, with GBDT achieving consistently lower average and maximum RMSE values compared to XGB and MLP, indicating a more stable error distribution across validation folds. This trend is reflected in the MAE values, where GBDT consistently outperformed XGB and MLP in average and minimum deviations from the measured values. Despite XGB achieving slightly better R^2 metrics in some cases, GBDT's consistently lower RMSE and MAE values suggest that it produced more accurate and reliable predictions overall. The analysis of MAPE values reveals that GBDT generally provided more accurate percentage predictions, although both XGB and GBDT exhibited some degree of heteroscedasticity in the prediction errors.

Table 2. Comparative Performance Metrics of XGB, GBDT, and MLP Models in Predicting TOC Content.

Metrics	**XGB**	**GBDT**	**MLP**
Avg RMSE	0.635	0.641	0.689
Max RMSE	1.093	1.032	1.119
Min RMSE	0.414	0.392	0.492
Avg MAPE	0.824	0.871	0.991
Max MAPE	1.075	1.149	1.209
Min MAPE	0.666	0.638	0.768
Avg R2	0.487	0.470	0.396

Max R2	0.676	0.734	0.517
Min R2	0.264	0.306	0.229
Avg MAE	0.325	0.330	0.381
Max MAE	0.366	0.392	0.424
Min MAE	0.276	0.258	0.324

This is evident in the absolute residual error plots (Fig. 5B, E, and H), which show variability increasing with higher TOC levels, suggesting that model performance may fluctuate at higher concentrations. To further examine the models' reliability, residual error analysis was conducted. Fig. 5B, E, and H display the absolute residuals against measured TOC, revealing patterns that suggest heteroscedasticity in all models, with errors tending to increase with higher TOC values. The presence of such error patterns indicates that the models may not fully capture the variability inherent in high TOC values. Nevertheless, GBDT and XGB models appear to demonstrate slightly better control over this variability than MLP. The residual error ratio plots (Fig. 5C, F, and I) depict the proportion of data points within a ±25% error margin. GBDT encompassed 37.6% of the data within this range, outperforming XGB (32.68%) and MLP (18.62%). This suggests that GBDT not only captures non-linearities effectively but also maintains a more stable predictive accuracy within this critical threshold.

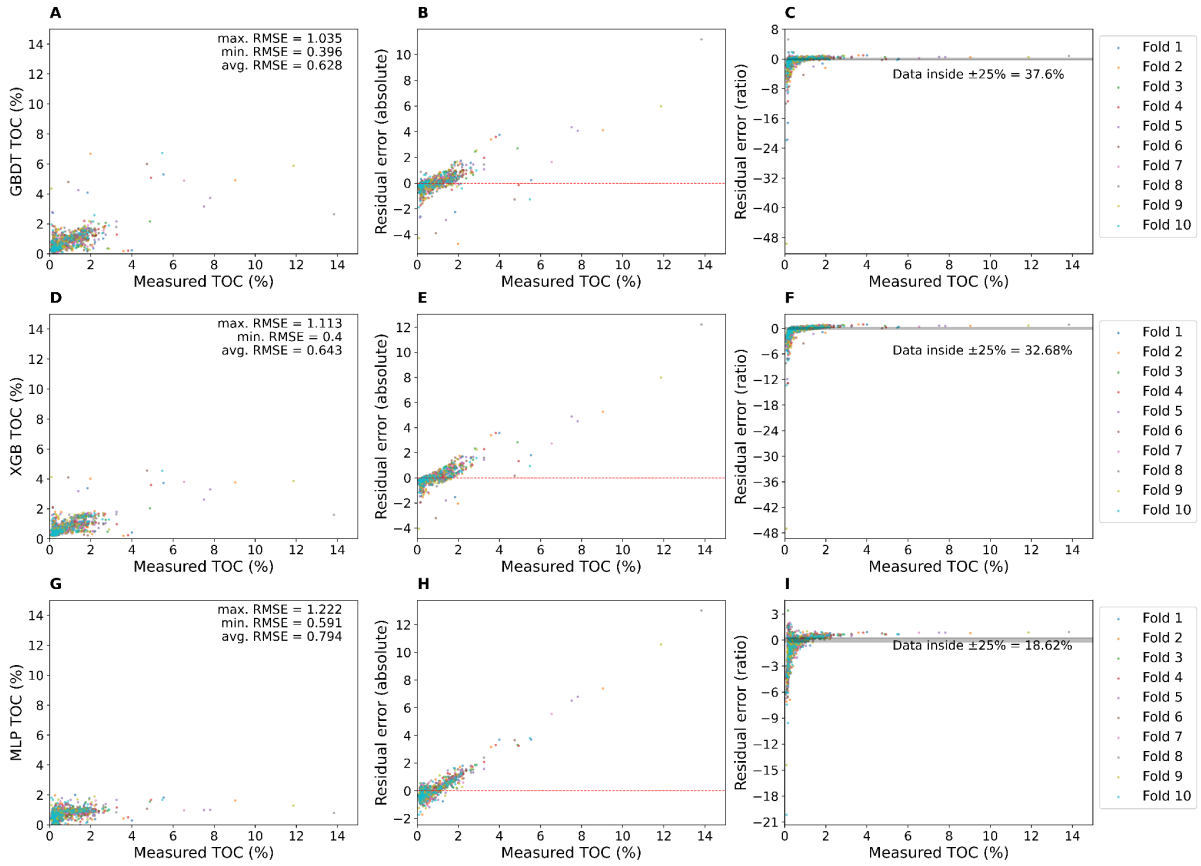


Fig. 5. Evaluation of Prediction Accuracy and Residual Error Distribution for GBDT, XGB, and MLP Models in Predicting TOC Content. (A, D, G) show the scatter plots comparing predicted TOC (%) against measured TOC (%) for the GBDT, XGB, and MLP models, respectively, with corresponding R^2

metrics (max, min, and average). (B, E, H) display the absolute residual error (absolute difference between predicted and measured TOC) as a function of measured TOC for each model. (C, F, I) illustrate the ratio-based residual error (ratio of the residual error to the measured TOC) as a function of measured TOC, indicating the proportion of data points within a ±25% error band for each model. The data points are color-coded by fold number across the 10 cross-validation folds.

The ability of ensemble methods like GBDT and XGB to aggregate predictions from multiple decision trees enhances their robustness against overfitting and improves generalization on unseen data, as discussed in previous works (Friedman, 2001; Chen & Guestrin, 2016; Hastie et al., 2009). The enhanced robustness and consistency make them more adept at managing the complexity and non-linear relationships in geochemical data, such as TOC. The relatively lower performance of the MLP model can likely be attributed to its sensitivity to the dataset's size and noise, as neural networks generally require larger datasets to establish robust internal representations (Zhang et al., 2016; Srivastava et al., 2014). Given the limited dataset and inherent noise in geochemical measurements, the MLP model may have faced challenges in generalizing well across all TOC intervals, leading to a broader error distribution and lower proportion of data within the ±25% error margin. These findings suggest that GBDT, with its ensemble approach, is better suited for this dataset, providing more accurate and consistent predictions across a range of TOC values. Despite XGB's marginally higher average R^2, the broader error analysis highlights GBDT’s advantage in maintaining stability and reducing errors across the measured TOC spectrum. Thus, for future geochemical predictions, ensemble methods like GBDT offer a promising solution to managing variability and complex non-linear relationships in the data.

3.1.4.4 Model Generalizability

To further compare the generalizability of the models, their validation results are shown in Fig. 6. Compared to the GBDT algorithm, XGB and MLP performed poorly in predicting TOC increase at depths of 3900m and 4600m. Overall, the predicted TOC values from the models follow the tendencies of the measured TOC, but the accuracy of the XGB and MLP models fluctuates more. GBDT appears to capture the changes in TOC more effectively, especially at critical depths, which is crucial for accurate subsurface characterization (Friedman, 2001). Additionally, when examining each well individually (Supplementary Figs. 2 to 4), it becomes evident that GBDT consistently captures the variations in TOC better than the other models. This suggests that GBDT not only performs well in a general sense but also adapts effectively to the specific conditions of each well, highlighting its robustness and flexibility. Such adaptability is essential in geoscientific applications where heterogeneity is common (Hastie et al, 2009). However, even though GBDT outperformed the other models, it still didn’t capture the higher variations of TOC within well 3BRSA496RJS and the high variability of well 1BRSA491SPS.

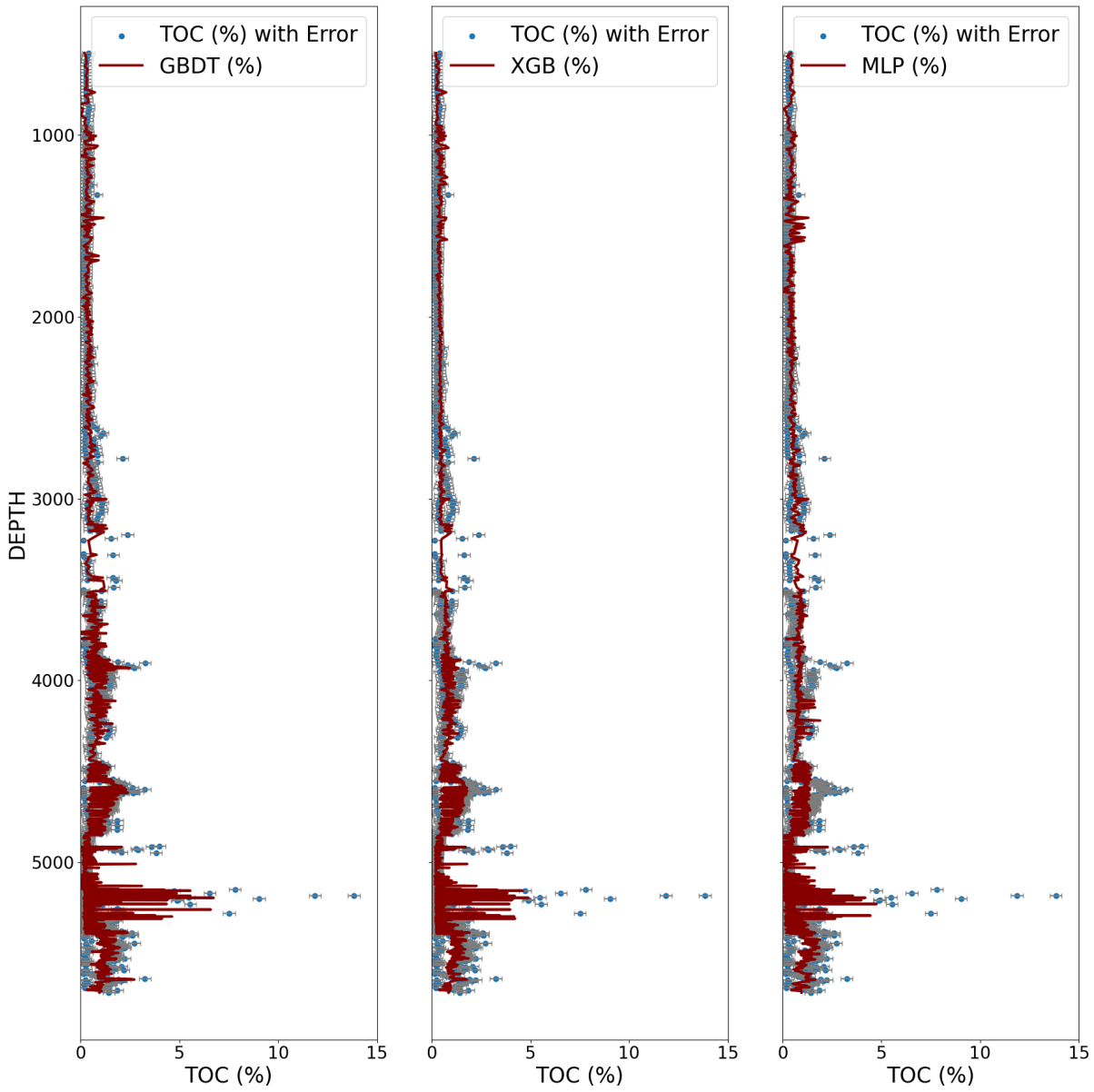

TOC (%) with Error
GBDT (%)
TOC (%) with Error
XGB (%)
TOC (%) with Error
MLP (%)
DEPTH
1000
2000
3000
4000
5000
0
5
10
15
TOC (%)
TOC (%)
TOC (%)

Fig. 6. Measured TOC (%) vs. Depth with Model Predictions for GBDT, XGB, and MLP Models. Each subplot shows the measured TOC values (blue points) with error bars indicating measurement uncertainty. The solid red lines represent the respective model's TOC predictions across the depth interval.

The superior performance of GBDT can be linked to its iterative boosting process, which sequentially reduces the residuals of previous models, thereby enhancing accuracy and reducing overfitting (Liao et al, 2016). In contrast, XGB, while also a boosting algorithm, might not be as fine-tuned for this specific dataset, leading to more pronounced fluctuations. On the other hand, the MLP model may require a larger dataset to generalize effectively and mitigate noise, which is not sufficiently available in this case (Goodfellow, Bengio, & Courville, 2016).

3.1.4.5 Data Imbalance

A critical factor contributing to the decreased performance of all models in predicting TOC values exceeding 3% is the significant data imbalance within the dataset. Specifically, 98.77% of the samples represent TOC values below 3%, while only 1.23% exceed this threshold. This imbalance introduces a bias towards the majority class during training, as algorithms often prioritize overall accuracy—easily achieved by predominantly predicting the majority class (Japkowicz & Stephen, 2002; He & Garcia, 2009). Consequently, the models exhibit reduced sensitivity to the minority class, leading to poorer predictive accuracy in that region. Furthermore, variations in depositional processes due to differences in tectonic settings, paleogeography, paleoenvironment, and paleoclimatic conditions can significantly impact well-log parameters (Chan et al., 2022), resulting in worse prediction due to the inherent heterogeneity between the wells

To address these challenges, several strategies can be considered. Oversampling the minority class increases its representation in the training data, potentially reducing bias but with the risk of overfitting (Chawla et al., 2002). Conversely, undersampling the majority class can balance the distribution but may result in the loss of valuable information (Drummond & Holte, 2003). Expanding the dataset by increasing the number of wells and samples to better represent the basin can enhance model generality (Nguyen et al., 2023), although this approach may not always be feasible. Adjusting model parameters—such as modifying cost functions or decision boundaries—can explicitly penalize misclassification of the minority class, thereby improving model sensitivity (Elkan, 2001). Implementing ensemble methods focused on minority class performance or employing cost-sensitive learning techniques can further enhance model performance in imbalanced datasets (Sun et al., 2009). Additionally, Monte Carlo methods can be utilized to extrapolate data, effectively augmenting the dataset and addressing the imbalance issue. By generating synthetic data points through random sampling based on the statistical distribution of the minority class, the model is exposed to a broader range of scenarios, which can improve predictive performance (Metropolis & Ulam, 1949; Rubinstein & Kroese, 2016). Finally, reducing the number of wells and samples to achieve a more homogeneous dataset can enhance model performance, as discussed in section 4.7.

3.1.4.6 The Traditional Passey Method

In addition to the ML approaches, an automated ΔlogR method was applied to separately estimate TOC values for each well. Incorporating this conventional method allows for a direct comparison between established techniques and the proposed ML models, thereby highlighting the potential improvements offered by newer data-driven methodologies.

The application of the ΔlogR method yielded significantly lower predictive performance compared to the ML models (Fig. 7). Specifically, it resulted in a negative R^2 (-2.125), along with higher RMSE = 4.537 and MAPE = 5.993. These metrics indicate that the ΔlogR method is less effective in predicting TOC values within the studied context. In contrast, ML offer several advantages over the conventional method as follows (Wang et al., 2019b):

(1) ML models can capture complex nonlinear correlations between the input variables and TOC. They are capable of modeling multi-level and multi-scale features, leading to more accurate TOC predictions than the ΔlogR

(2) ML models can incorporate a broader range of logging information to predict TOC content more accurately. The ΔlogR method relies solely on DT and RT logs, which limits its ability to fully exploit the information contained in all available well logs.

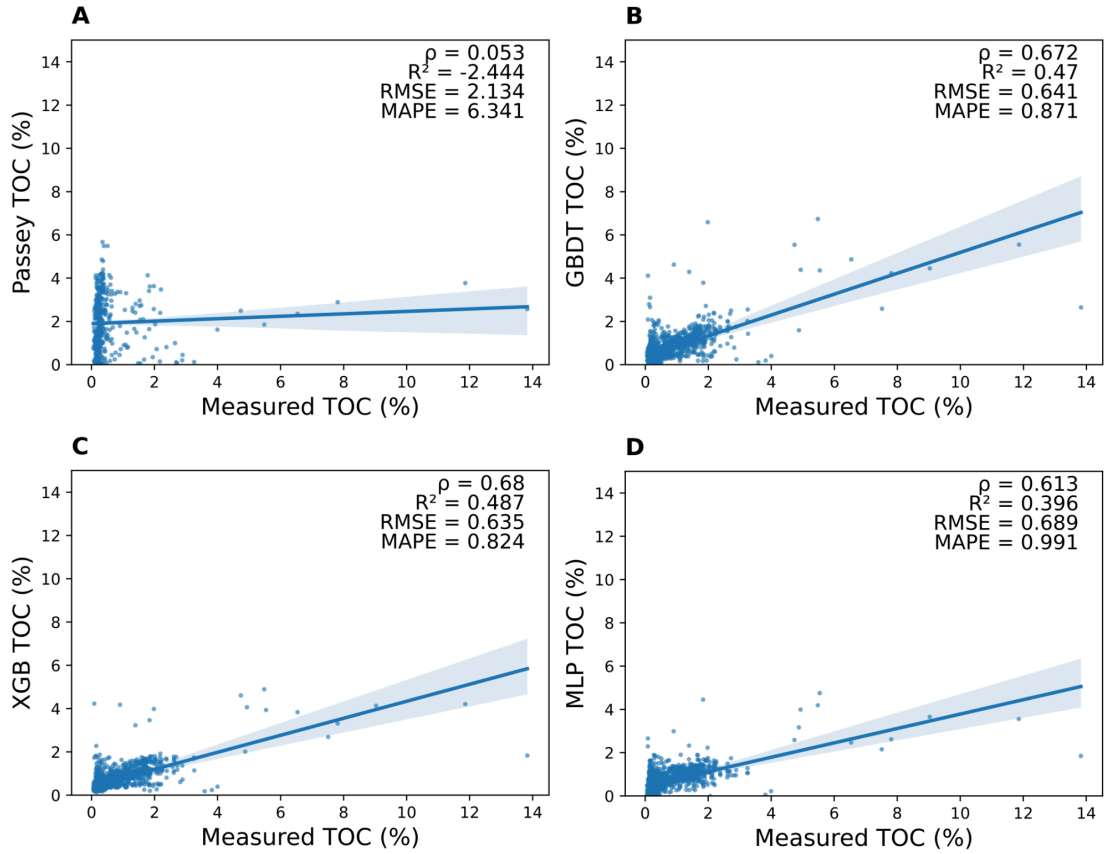


Fig. 7. Comparison of Predicted vs. Measured TOC (%) for Passey (A), GBDT (B), XGB (C), and MLP (D) Models with Metrics. Each subplot shows the linear regression fit with a confidence interval (shaded region). The metrics displayed in each plot include Spearman's rank correlation coefficient (ρ), R^2, RMSE, and MAPE.

3.1.4.7 Better Results By Removing Two Wells

This study demonstrates that ML models may struggle to generalize in highly diverse settings, necessitating more representative data. We reduced the size of our original dataset from 1386 to 860 points, a reduction of 37.95%, by focusing on wells located closer together (1BSS72BS, 1BSS77BS, and 1BRSA642SPS) with more homogeneous characteristics. This resulted in a significant improvement in the performance of the GBDT model.

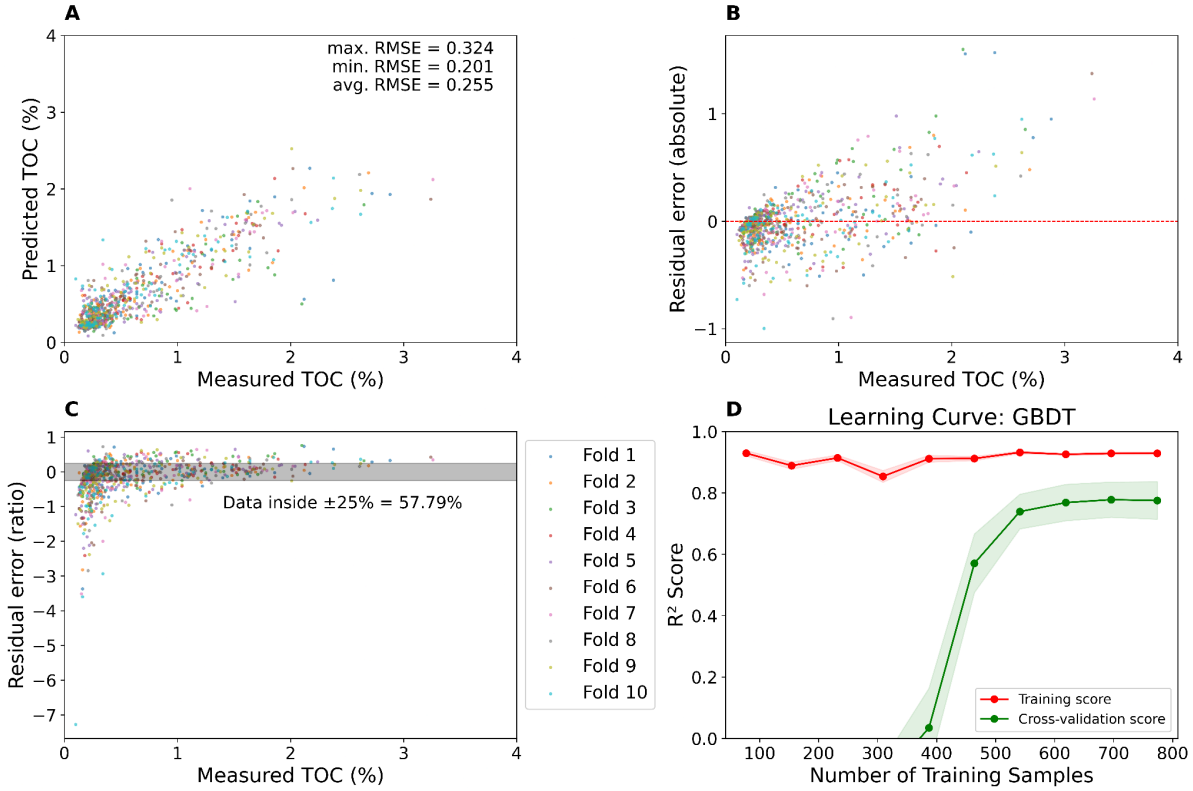


Fig. 8. Results of the GBDT Model Performance for Three Wells: Predicted TOC, Residual Error, and Cross-Validation Scores. (A) Plot of Predicted TOC (%) vs Measured TOC (%) RMSE metrics (maximum, minimum, and average). (B) Plot of Residual Error (absolute) vs Measured TOC (%). (C) Plot of Residual Error (ratio) vs Measured TOC (%) across ten cross-validation folds, with a shaded region indicating ±25% error and the percentage of data falling within this range. (D) Learning curve showing training and cross-validation scores (R^2) as a function of the number of training samples, with shaded regions representing the standard deviation.

As seen in Fig. 8A, the predicted TOC values correlate strongly with the measured TOC, achieving an average RMSE of 0.255. Fig. 8B shows a notable reduction in residual error, while Fig. 8C highlights that 57.79% of the data points fall within the ±25% error margin, an improvement of 59.39% in RMSE and 53.73% in data within the error margin, respectively. These improvements can be attributed to the increased homogeneity of the data used to train the model. As discussed by Hastie et al. (2009), using more homogenous data can lead to better model generalizability.

Moreover, as illustrated by Fig. 8D, the learning curve for the GBDT model shows significant improvement. The training score stabilizes around 0.8 similarly to the previous model trained on a larger dataset. The cv score, however, rises sharply, ultimately converging close to the training score at approximately 0.75, representing an increase of about 50% from the

previous model's cv score of around 0.50. This reduced gap between training and cv scores, accompanied by narrower confidence intervals, indicates enhanced model generalization and reduced overfitting (Friedman, 2001). These findings emphasize the critical importance of optimizing dataset quality and relevance for achieving reliable ML models.

3.1.4.8 Advantage and Limitations of ML models

One advantage of using machine learning methods for estimating TOC is their applicability to stratigraphically unconstrained datasets, although constraining specific intervals could improve model performance and reduce heteroscedasticity. Another advantage is the ability to quickly assess large datasets without the need for well-dependent calibrations. However, these methods do not allow for sensitivity tests to infer paleoenvironmental conditions, unlike process-based modeling techniques (e.g., Mann and Zweigel, 2008; Venancio et al., 2022b).

Generalization to other wells and basins presents another challenge. The observed promising results require cautious extrapolation due to limitations in the available dataset. Robustness testing on diverse datasets representing wider geological variations and TOC concentrations is crucial for ensuring model reliability. While accurate and potentially useful, the model's applicability might be restricted by specific data collection contexts, local environmental conditions, and exploration regions. The diversity in the settings can potentially lead to a range of variations in mineralogy, organic richness, and other factors influencing TOC content, consequently hindering the training of models applicable to diverse data settings. Developing robust predictive models for TOC among such data heterogeneity poses a significant challenge. Increasing the training set size to represent a wider range of well-log characteristics could be a potential strategy, but practical limitations and the difficulty of acquiring sufficiently diverse wells with the selected features need to be considered. Nevertheless, the results presented in this study are an example of a generalized and reproducible regional model of TOC prediction for the Santos Basin that could be adapted to other sedimentary basins at the exploration-level of the oil and gas industry.

3.1.5 Conclusion

In this study, we investigated the use of various ML models for predicting TOC content from well-log data in the Santos Basin. Our conclusions are as follows:

1 - All ML models provided higher prediction accuracy for TOC content than the conventional ΔLogR method, capturing better the complex, non-linear relationships between well-log features and TOC, which the traditional method could not.

2 - Among the ML models, the GBDT model outperformed both the XGB and MLP models. This can be attributed to GBDT's robust ensemble learning technique, which sequentially builds decision trees to correct the errors of previous ones, thereby enhancing overall predictive accuracy and reducing overfitting.

3 - Reducing the dataset from five wells to three wells, resulting in a 37.95% reduction in data points, significantly improved the performance of the GBDT model. This reduction led to a more homogeneous dataset, which improved the model's prediction accuracy, reduced overfitting, and enhanced generalizability. This finding underscores the importance of optimizing dataset quality and relevance to improve model performance.

4 - This study also highlighted the challenges of generalizing ML models to diverse geological settings. The observed improvements with a more homogeneous dataset suggest that ML models may struggle to generalize across highly varied settings. Future work should focus on testing the models on diverse datasets representing wider geological variations and TOC concentrations to ensure robust and reliable model performance. Additionally, exploring advanced techniques for handling data imbalance and heterogeneity, such as oversampling, undersampling, and cost-sensitive learning, could further enhance model accuracy and generalizability.

3.1.6 Declaration of competing interest

Upon submitting this manuscript, I confirm that it is an original work that has not been published, is not under consideration for publication elsewhere, and is not currently being reviewed by any other journal. All authors have approved this submission and declare that there are no conflicts of interest.

3.1.7 Data availability

Data and codes will be available on an online repository, or upon request.

3.1.8 Acknowledgements

Bernardo S. Chede acknowledges the financial support from FEC 479118803/2022-8 and CAPES PDSE scholarship award (CAPES-PRINT - 88887. 935155/2024-00). Andre L. Belem is a CNPq researcher (grant 315004/2020–7). Ana Luiza S. Albuquerque is a senior CNPq researcher (grant 307658/2021–0) and acknowledges the support from FAPERJ (Grant E-26/210.081/2023 and Grant E-26/201.008/2022). Ulrich G. Wortmann was supported through an NSERC Discovery Grant. I.M. Venancio acknowledges the support from FAPERJ (SEI-260003/000677/2023) (JCNE grant 200.120/2023–281226). The authors thank the National Petroleum Agency (ANP/BDEP) for providing the geochemical and well data through the policy of free transfer of public data for academic purposes (Processes: 48610.208058/2023-42, 48610.208057/2023-06).

3.1.9 References

Abdi, H., & Williams, L.J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459.

Ali, M., Zhu, P., Huolin, M., et al. (2023). A Novel Machine Learning Approach for Detecting Outliers, Rebuilding Well Logs, and Enhancing Reservoir Characterization. Natural Resources Research, 32(3), 1047–1066.

ANP (National Agency of Petroleum, Natural Gas and Biofuels), 2020. Boletim de Recursos e Reservas. Available online: https://www.gov.br/anp/pt-br/centrais-deconteudo/dados-estatisticos/arquivos-reservas-nacionais-de-petroleo-e-gas-natural/boletim_reservas_2020.pdf

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281-305.

Bhattacharya, S., Ambrose, W., Ko, L. T., & Casey, B. (2022). Funny-looking things: Interesting features seen on seismic data. Integrated detection and investigation of bad borehole section in the Wolfcamp Formation in the Midland Basin using machine learning, petrophysics, and core characterization. 10(3), 19–27.

Bione, F. R. A., Venancio, I. M., Santos, T. P., Belem, A. L., Rangel, B. R., Souza, I. V. A. F., Spigolon, A. L. D., & Albuquerque, A. L. S. (2024). Estimating total organic carbon of potential source rocks in the Espírito Santo Basin, SE Brazil, using XGBoost. Marine and Petroleum Geology, 162.

Bishop, C. M. (1995). Neural Networks for Pattern Recognition. Oxford University Press.

Buckley, J. D., Bosence, D. W., & Elders, C. F. (2015). Tectonic setting and stratigraphic architecture of an Early Cretaceous lacustrine carbonate platform, Sugar Loaf High, Santos Basin, Brazil. Geological Society, London, Special Publications, 418, p. 1–17.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. Nature Methods, 15(4), 233–234.

Cameron, A. C., & Windmeijer, F. A. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. Journal of Econometrics, 77(2), 329–342.

Carpentier, B., Huc, A. Y., & Bessereau, G. (1991). Wireline logging and source rocks estimation of organic carbon content by the CARBOLOG method. Log Analyst, 32(3), 279–297.

Chan, S. A., Hassan, A. M., Usman, M., Humphrey, J. D., Alzayer, Y., & Duque, F. (2022). Total organic carbon (TOC) quantification using artificial neural networks: Improved prediction by leveraging XRF data. Journal of Petroleum Science and Engineering, 208, 109302.

Charsky, A., & Herron, S. (2013). Accurate, direct total organic carbon (TOC) log from a new advanced geochemical spectroscopy tool: Comparison with conventional approaches for TOC estimation. Society of Petrophysicists and Well-Log Analysts.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, pp. 785–794.

Chen, T., Singh, S., Taskar, B., & Guestrin, C. (2015). Efficient second-order gradient boosting for conditional random fields. In Proceeding of 18th Artificial Intelligence and Statistics Conference (AISTATS'15), volume 1.

Chowdhury, A. A., Das, A., Shahjalal Hoque, K. K., & Karmaker, D. (2022). A Comparative Study of Hyperparameter Optimization Techniques for Deep Learning. In International Joint Conference on Advances in Computational Intelligence (IJCACI 2021).

Cressie, N. (1993). Statistics for Spatial Data. Chapman & Hall.

Damasceno, A.C., Korenchendler, A.L., Da Silva, A.M., Da Silva Praxedes, E., De Almeida Dos Reis, M.A.A., & Silva, V.G. (2022). Source rock evaluation from rock to seismic: Integrated machine learning-based workflow. IMAGE, Houston, TX, USA, 28 August–1 September.

Delcroix, M., & Golea, M. (2021). Gradient Boosting for Regression and Classification. Journal of Machine Learning Research.

Drummond, C., & Holte, R. C. (2003). C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling. Workshop on Learning from Imbalanced Datasets II, 1-8.

Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. IJCAI, 17, 973-978.

Estrella, G., Mello, M. R., Gaglianone, P. C., Azevedo, R. L. M., Tsubone, K., Rossetti, E., Concha, J., & Brüning, I. M. R. A. (1984). The Espirito Santo Basin (Brazil) Source Rock Characterization and Petroleum Habitat. In G. Demaison & R. J. Murris (Eds.), Petroleum Geochemistry and Basin Evaluation (Vol. 35, p. 0). American Association of Petroleum Geologists.

Fernandes, E. (2017). Bacia de Santos: Sumário Geológico e área em oferta. ANP – Agência Nacional de Petróleo, Gás Natural e Biocombustíveis – Seminário Técnico.

Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. ACM SIGKDD Explorations Newsletter, 12(1), 49–57.

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Friedman, J.H. (2002). Stochastic gradient boosting. Computational Statistics and Data Analysis, 38(4), 367–378.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. Neural Computation, 4(1), 1-58.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast Density-Based Clustering with R. Journal of Statistical Software, 91(1), 1-30.

Hasan, S., & Karim, M. (2021). Performance Analysis of Gradient Boosting Methods. International Journal of Data Science and Analytics.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.

Hinton, G.E., Osindero, S., & Teh, Y.W. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18(7), 1527–1554.

Hood, A., Gutjahr, C.C.M., & Heacock, R.L. (1975). Organic metamorphism and the generation of petroleum. AAPG Bulletin, 59, 989–996.

Hunt, J.M. (1995). Petroleum Geochemistry and Geology (2nd ed.). W.H. Freeman and Company, New York.

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence, 33(10), 913-933.

Japkowicz, N., & Stephen, S. (2002). The Class Imbalance Problem: A Systematic Study. Intelligent Data Analysis, 6(5), 429–449.

Jolliffe, I.T. (2002). Principal Component Analysis (2nd ed.). Springer-Verlag, New York, Inc.

Jolliffe, I.T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

Kamali, M.R., & Mirshady, A.A. (2004). Total organic carbon content determined from well logs using DlogR and neuro-fuzzy techniques. Journal of Petroleum Science and Engineering, 45, 141–148.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, 3146-3154.

Khan, M., Liu, T.Q., & Ullah, F. (2019). A new hybrid approach to forecast wind power for large-scale wind turbine data using deep learning with TensorFlow framework and principal component analysis. Energies, 12(12), 2229–2249.

Khan, M. R., Kalam, S., Asad, A., & Abu-khamsin, S. (2023, March). Development of a Deterministic Total Organic Carbon (TOC) Predictor for Shale Reservoirs. In SPE Middle East Oil and Gas Show and Conference. SPE.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI '95), 1137–1143.

Lever, J., Krzywinski, M., & Altman, N. (2017). Principal component analysis. Nature Methods, 14(7), 641–642.

Liashchynskyi, P., & Liashchynskyi, P. (2019). Grid search, random search, genetic algorithm: a big comparison for NAS.

Liao, Z., Huang, Y., & Yue, X. (2016). In silico prediction of gamma-aminobutyric acid type-a receptors using novel machine learning-based SVM and GBDT approaches. BioMed Research International, 1–12.

Liu, X., Tian, Z., & Chen, C. (2021). Total organic carbon content prediction in lacustrine shale using extreme gradient boosting machine learning based on Bayesian optimization. Geofluids, 2021(1), 6155663.

Maciejowska, K., Uniejewski, B., & Serafin, T. (2020). PCA forecast averaging-predicting day-ahead and intraday electricity prices. Energies, 12(14), 3530–3548.

Mann, U., & Zweigel, J. (2008). Modelling source-rock distribution and quality variations: The organic facies modelling approach. In Analogue and Numerical Modelling of Sedimentary Systems: From Understanding to Prediction. John Wiley & Sons, Ltd, pp. 239–274.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo Method. Journal of the American Statistical Association, 44(247), 335-341.

Moreira, J. L. P., Madeira, C. V., Gil, J. A., & Machado, M. A. P. (2007). Bacia de Santos.

Moulin, M., Aslanian, D., Olivet, J., Contrucci, I., Matias, L., Geli, L., Klingelhoefer, F., Nouze, H., Rabineau, M., Labails, C., Rehault, J., & Unternehr, P. (2005). Geological constraints on the evolution of the Angolan margin based on reflection and refraction seismic data (ZaiAngo Project). Geophysical Journal International, 162, 793–810.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7, 21.

Ng, A. Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. Proceedings of the 21st International Conference on Machine Learning.

Nguyen, H., Larsen, E., Oikonomou, D., Alaei, B., Stefos, G., Kvalheim, A., Alaei, A., Evans, D., & Stoddard, D. (2023). A multi-model AI workflow - integrating from rock samples to basin-scale seismic-based rock property prediction. 84th EAGE Annual Conference & Exhibition, 1-5. European Association of Geoscientists & Engineers.

Nielsen, D. (2016). Tree boosting with XGBoost: Why does XGBoost win every machine learning competition? Master's thesis, Norwegian University of Science and Technology.

Nogueira, F. (2014). BayesianOptimization. GitHub repository. https://github.com/fmfn/BayesianOptimization

Pan, S., Zheng, Z., Guo, Z., & Luo, H. (2022). An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. Journal of Petroleum Science and Engineering, 208, 109520.

Passey, Q. R., Creaney, S., Kulla, J. B., Morretti, F. J., & Stroud, J. D. (1990). A Practical Model for Organic Richness from Porosity and Resistivity Logs.

Passey, Q. R., Bohacs, K. M., Esch, W.L., Klimentidis, R., & Sinha, S. (2010). From oil-prone source rock to gas-producing shale reservoir—Geologic and petrophysical characterization of unconventional shale-gas reservoirs. Society of Petroleum Engineers, SPE-131350.

Pebesma, E., & Bivand, R. (2023). Spatial data science: With applications in R. Chapman & Hall/CRC.

Peters, K.E., & Cassa, M.R. (1994). Applied source rock geochemistry. In: Magoon, L.B., Dow, D.G. (Eds.), The Petroleum System—From Source to Trap. AAPG Memoir, Tulsa, pp. 93–120.

Reis, M.A.A.d.A.d., Damasceno, A.C., Roriz, C.E.D., Korenchendler, A.L., & Silva, A.M.d. (2023). Source Rock Evaluation from Rock to Seismic Data: An Integrated Machine-Learning-Based Workflow and Application in the Brazilian Presalt (Santos Basin). Minerals, 13, 1179.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review, 65(6), 386–408.

Rubinstein, R. Y., & Kroese, D. P. (2016). Simulation and the Monte Carlo Method (3rd ed.). John Wiley & Sons.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536.

Saporta, G., & Niang, N. (2009). Principal component analysis: application to statistical process control. In Data Analysis (pp. 1-23). John Wiley & Sons.

Schmoker, J., & Hester, T. (1983). Organic carbon in Bakken Formation, United States portion of Williston Basin. AAPG Bulletin, 67, 2165–2174.

Schmoker, J. (1979). Determination of organic content of Appalachian Devonian shales from formation-density logs. AAPG Bulletin, 63, 1504–1537.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms.

Snoek, J., Swersky, K., Zemel, R. S., & Adams, R. P. (2014). Input warping for Bayesian optimization of non-stationary functions. In ICML.

Sondergeld, C.H., Newsham, K.E., Comisky, J.T., Rice, M.C., & Rai, C.S. (2010). Petrophysical considerations in evaluating and producing shale gas resources. Society of Petroleum Engineers, SPE-131768.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(56), 1929–1958.

Sun, J., Dang, W., Wang, F., Nie, H., Wei, X., Li, P., & Li, F. (2023). Prediction of TOC content in organic-rich shale using machine learning algorithms: Comparative study of random forest, support vector machine, and XGBoost. Energies, 16(10), 4159.

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of Imbalanced Data: A Review. International Journal of Pattern Recognition and Artificial Intelligence, 23(4), 687–719.

Szabo, T., & HorvERL, G. (1997). Roundoff Error Analysis of the PCA Networks. IEEE.

Tissot, B.P., & Welte, D.H. (1984). Petroleum Formation and Occurrence (2nd ed.). Springer-Verlag, Berlin.

Venancio, I.M., Belem, A.L., Santos, T.P., Lessa, D.O., Leonardo, N.F., Bione, F.R.A., Díaz, R., Moreira, M., Bernardes, M.C., Souza, I.V.A.F., Coutinho, L.F.C., & Albuquerque, A.L.S. (2022a). Temporal and spatial differences between predicted and measured organic carbon in South Atlantic sediments: Constraints to organic facies modelling. Marine Petroleum Geology, 138, 105524.

Venancio, I.M., Santos, T.P., Bione, F.R.A., Belem, A.L., Bernardes, M.C., Díaz, R.A., Moreira, M., Carreira, V., Spigolon, A., Souza, I.V., & Albuquerque, A.L.S. (2022b). Preservation factors during cretaceous oceanic anoxic events in the Espírito Santo Basin, southeast Brazil. Geosciences, 12, 351.

Wang, P., Chen, Z., Pang, X., Hu, K., Sun, M., & Chen, X. (2016). Revised models for determining TOC in shale play: Example from Devonian Duvernay shale, Western Canada sedimentary basin. Petroleum Geology, 70, 304–319.

Wen, Z., Han, J., Shang, Y., Tao, H., Fang, C., Lyu, L., Li, S., Hou, J., Liu, G., & Song, K. (2023). Spatial variations and molecular composition of dissolved organic matter in lakes across different frozen ground zones in China. Science of The Total Environment, 871, 162089.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1), 79–82.

Wright, V. P., & Barnett, A. J. (2015). An abiotic model for the development of textures in some South Atlantic early Cretaceous lacustrine carbonates. Geological Society, London, Special Publications, 418(1), 209–219.

Xia, Y., Jiang, S., Meng, L., & Ju, X. (2024). XGBoost-B-GHM: An Ensemble Model with Feature Selection and GHM Loss Function Optimization for Credit Scoring. Systems, 12, 254.

Zhang, D., Gong, Y. (2020). The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure. IEEE Access, 8, 220990–221003.

Zhang, Y., Wang, G., Wang, X., Fan, H., Shen, B., & Sun, K. (2023). TOC estimation from logging data using principal component analysis. Energy Geoscience, 4, 100197.

Zhang, C., Bengio, Y., Morpholio, M., & Larson, J. (2016). Understanding deep learning requires rethinking generalization. In International Conference on Learning Representations, 1-10.

Zhao, P.; Mao, Z.; Huang, Z.; Zhang, C.A. New method for estimating total organic carbon content from well logs. AAPG Bull. 2016, 100, 1311–1327.

Zhao, W., Gao, H., Yan, G., & Guo, T. (2020). TOC prediction technology based on optimization estimation and Bayesian statistics. Lithology and Reservoir, 32(1), 86–93.

Zhu, L., Zhang, C., Zhang, C., Zhang, Z., Nie, X., Zhou, X., Liu, W., & Wang, X. (2019). Forming a new small sample deep learning model to predict total organic carbon content by combining unsupervised learning with semi-supervised learning. Applied Soft Computing, 83, 105596.

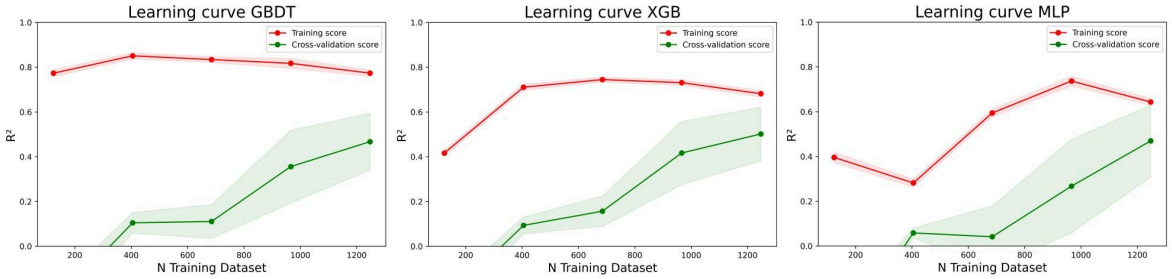
3.1.10 Appendix A.Supplementary data

3.1.10.1 Learning Curve of the Models

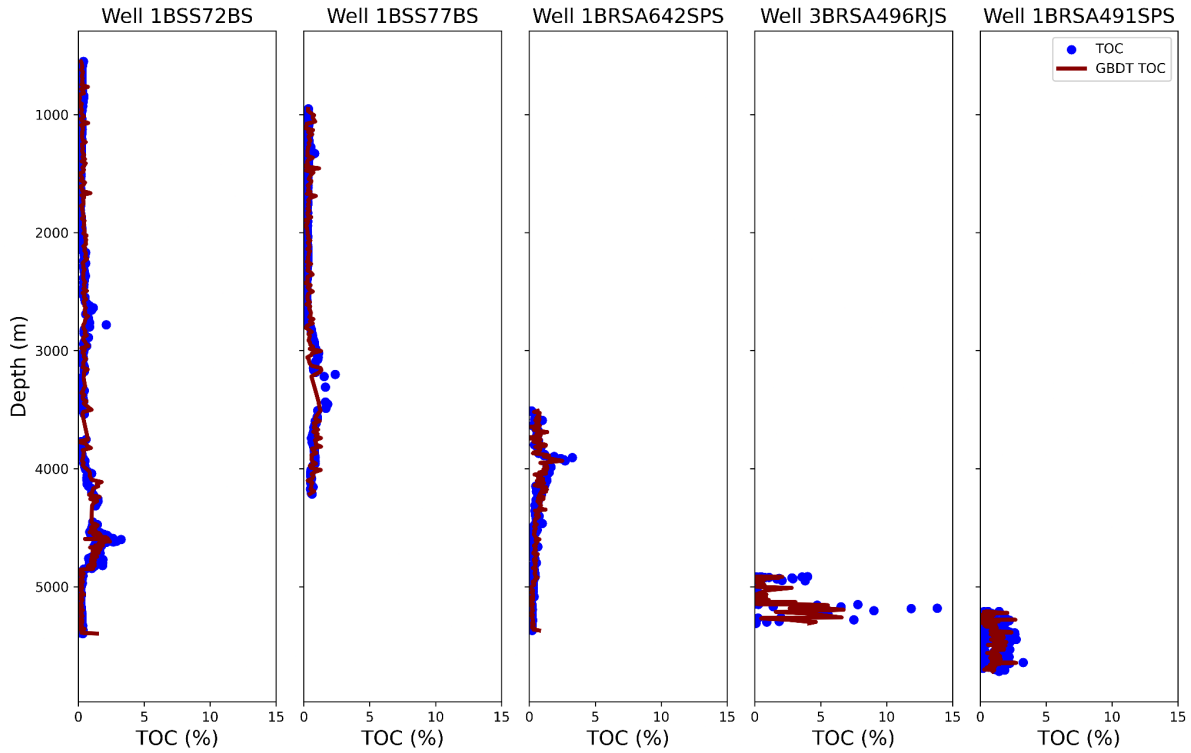
The learning curves for the MLP, XGB, and GBDT models, depicted in Supplementary Fig. 1, provide insight into the training and cross-validation (cv) performance as a function of the number of training examples. For all three models, the training scores start relatively high and show varying degrees of stability as more data is added. The cv scores, however, start much lower and increase steadily with the number of training examples. This trend indicates that all models are benefiting from additional data, as evidenced by the improving R^2 scores in cv and approximation of the training score.

The MLP model shows a notable increase in cv performance, suggesting it requires more data to generalize effectively, aligning with the characteristics of neural networks that typically need larger datasets to perform well (Goodfellow, Bengio, & Courville, 2016). The XGB model also demonstrates significant improvement, but with more fluctuation, potentially due to its sensitivity to the hyperparameters and the nature of the boosting process (Chen & Guestrin, 2016). The GBDT model shows a consistent upward trend, with cv scores gradually approaching the training scores, indicating a reduction in overfitting and enhanced generalization as more data is introduced (Natekin & Knoll, 2013). However, the gap between training and cv scores suggests overfitting and low generalization for all models, indicating that more data is required to improve model performance further.

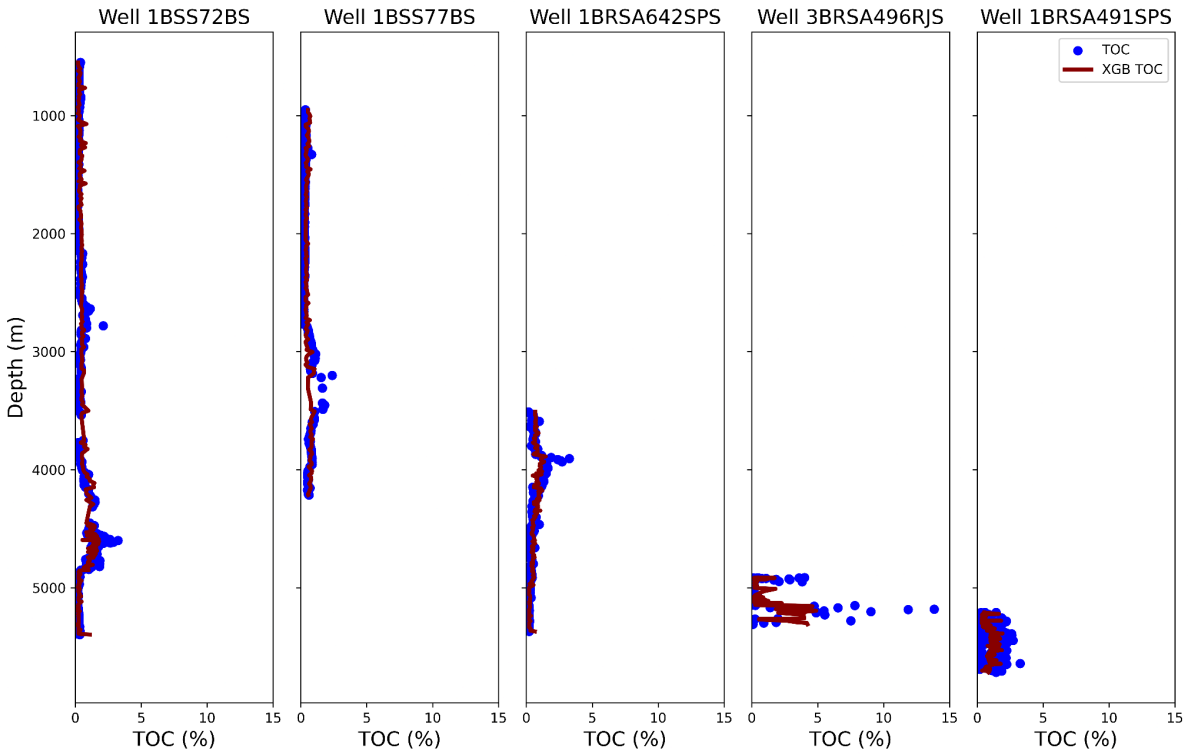
3.1.10.2 Supplementary Figures and Tables



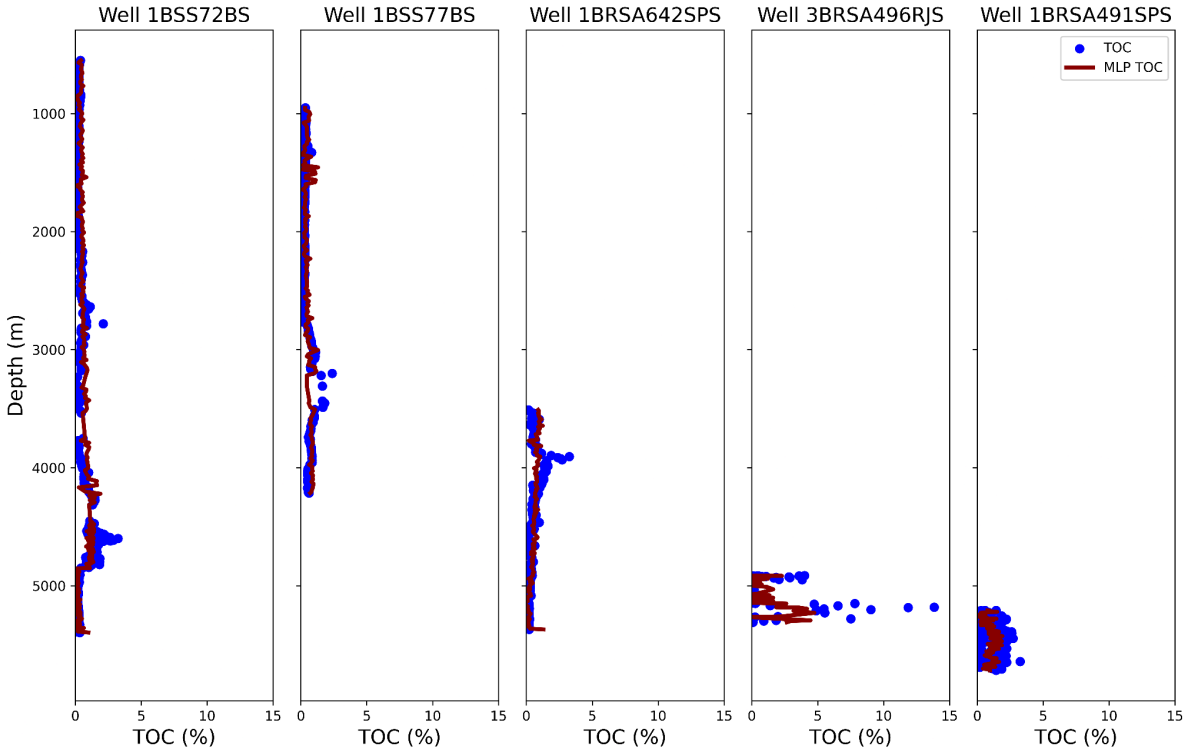
Supplementary Fig. 1. Learning Curves for GBDT, XGB, and MLP Models Showing Training and Cross-Validation Performance. Each plot represents the evolution of the training score (red line) and cross-validation score (green line) in terms of R^2 as the number of training samples increases. Shaded regions indicate the standard deviation of the cross-validation scores, demonstrating the models' generalization ability with varying dataset sizes.



Supplementary Fig. 2. Comparison of Measured and GBDT-Predicted TOC (%) Profiles Across Wells (1BSS72BS, 1BSS77BS, 1BRSA642SPS, 3BRSA496RJS, and 1BRSA491SPS.). The blue dots represent the measured TOC values, while the solid red line represents the predictions made using the GBDT model. Each subplot shows the TOC predictions compared with the actual measured values at varying depths (m) for each well.



Supplementary Fig. 3. Comparison of Measured and XGB-Predicted TOC (%) Profiles Across Wells (1BSS72BS, 1BSS77BS, 1BRSA642SPS, 3BRSA496RJS, and 1BRSA491SPS.). The blue dots represent the measured TOC values, while the solid red line represents the predictions made using the XGB model. Each subplot shows the TOC predictions compared with the actual measured values at varying depths (m) for each well.



Supplementary Fig. 4. Comparison of Measured and MLP-Predicted TOC (%) Profiles Across Wells (1BSS72BS, 1BSS77BS, 1BRSA642SPS, 3BRSA496RJS, and 1BRSA491SPS.). The blue dots

represent the measured TOC values, while the solid red line represents the predictions made using the MLPmodel. Each subplot shows the TOC predictions compared with the actual measured values at varying depths (m) for each well.

Supplementary Table 1. Hyperparameters of the XGB Regressor Algorithm with Corresponding Search Ranges Used for Bayesian Optimization and Their Descriptions.

Hyperparameters	**Range**	**Description**
n_estimator	50 - 1000	Number of gradient boosted trees.
max_depth	1 - 10	Maximum tree depth for base learners
min_child_weight	1 - 20	Minimum sum of instance weight needed in a child
eta	0.001 - 0.05	Learning rate shrinks the contribution of each tree.
subsample	0.5 - 1	Subsample ratio of the training instances
colsample_bytree	0.5 - 1	Subsample ratio of columns when constructing each tree
alpha	0 - 10	L1 regularization term on weights
lambda	1 - 10	L2 regularization term on weights

Supplementary Table 2. Hyperparameters of the GBDT Regressor Algorithm with Corresponding Search Ranges Used for Bayesian Optimization and Their Descriptions.

Hyperparameters	**Range**	**Description**
n_estimators	50 - 1000	Number of boosting stages to perform
max_depth	1 - 10	Maximum depth of the individual regression estimators
min_samples_split	10 - 30	Minimum number of samples required to split an internal node
min_samples_leaf	10 - 30	Minimum number of samples required to be at a leaf node
max_features	0.1 - 0.8	Fraction of features to consider when looking for the best split
learning_rate	0.001 - 0.05	Learning rate shrinks contribution of each tree
tol	0.0001 - 0.1	Tolerance for the optimization

Supplementary Table 3. Hyperparameters of the MLP Regressor Algorithm with Corresponding Search Ranges Used for Bayesian Optimization and Their Descriptions.

Hyperparameters	**Range**	**Description**
hidden_layers_sizes	(25,), (100,) or (100, 5)	Number of neurons in the hidden layers
activation	identity, logistic, tanh or relu	Activation function for the hidden layer
learning_rate	constant, invscaling or adaptive	Learning rate schedule for weight updates
max_iter	500 - 2000	Maximum number of iterations
alpha	0.001 - 1	L2 penalty (regularization term) parameter
tol	00001 - 0.01	Tolerance for optimization

Supplementary Table 4. Top 5 Hyperparameter Results for XGB Based on RMSE During Hyperparameter Tuning.

RMSE	**colsample_ bytree**	**learning _rate**	**max_ depth**	**min_child _weight**	**n_ estimators**	**alpha**	**lambda**	**subsample**
6.231	0.902	0.019	9	6	225	8.732	9.590	0.679
6.236	0.923	0.011	8	2	388	8.217	6.469	0.521
6.245	0.923	0.009	7	7	358	4.112	7.655	0.983
6.247	0.940	0.036	8	6	221	9.753	5.236	0.522
6.272	0.847	0.026	8	8	209	7.862	6.197	0.654

Supplementary Table 5. Top 5 Hyperparameter Results for GBDT Based on RMSE During Hyperparameter Tuning.

RMSE	**learning_ rate**	**max_ depth**	**max_ features**	**min_ samples_ leaf**	**min_ samples_ split**	**n_ estimators**	**tol**
6.160	0.026	8	0.177	10	19	198	0.072
6.170	0.030	5	0.110	11	16	194	0.008
6.172	0.010	8	0.692	11	21	208	0.060
6.176	0.012	10	0.547	11	20	209	0.076
6.182	0.032	7	0.206	11	18	204	0.086

Supplementary Table 6. Top 5 Hyperparameter Results for GBDT Based on RMSE During Hyperparameter Tuning.

RMSE	activation	alpha	hidden_layers_ sizes	learning_rate	max_iter	tol
6.319	tanh	0.878	(25,)	invscaling	1988	0.0041
6.319	relu	0.858	(25,)	invscaling	1607	0.0040
6.325	tanh	0.905	(25,)	invscaling	1607	0.0001
6.325	tanh	0.845	(25,)	invscaling	1606	0.0001
6.325	tanh	0.910	(25,)	invscaling	1606	0.0005

3.2. Integrating TOC Data with a 3D Geological Model for Spatial TOC Distribution

3.2.1 Introduction

The prediction and spatial distribution of Total Organic Carbon (TOC) is paramount for understanding petroleum systems and evaluating source rock potential (Tissot & Welte, 1984; Peters & Cassa, 1994). As a key indicator of organic matter richness, TOC directly influences hydrocarbon generation potential, making its accurate quantification critical for exploration success (Hood et al., 1975; Peters et al., 2005). In this context, the integration of TOC data with three-dimensional geological models represents a significant advancement in characterizing the spatial heterogeneity of source rock properties.

Recent developments in sedimentary basin modeling have emphasized the importance of incorporating geochemical parameters into geological frameworks to improve hydrocarbon potential assessments (Hantschel & Kauerauf, 2009; Baur et al., 2019). The Santos Basin, with its extensive pre-salt petroleum system, represents an ideal geological setting for implementing such integrated approaches. Characterized by complex lacustrine depositional environments and subsequent structural evolution, this basin contains organic-rich intervals within the Itapema Formation that serve as primary source rocks for the prolific pre-salt oil accumulations (Moreira et al., 2007; Gonçalves et al., 2020).

Traditional approaches to TOC estimation often rely on well log-based methods or geochemical sampling at discrete locations, which provide limited spatial coverage and may inadequately represent basin-scale heterogeneity (Passey et al., 2010; Rodrigues et al., 2021). Advancements in geostatistical techniques and machine learning algorithms now enable more sophisticated integration of sparse point data with seismic attributes and geological constraints to generate comprehensive 3D TOC models (Deutsch & Pyrcz, 2014; Karpatne et al., 2017). These models can capture spatial variations in source rock properties across multiple scales, from fine-scale vertical heterogeneity to basin-wide lateral trends.

Despite these advancements, significant challenges remain in modeling TOC distribution in complex geological settings like the Santos Basin pre-salt section. The limited availability of direct TOC measurements at well locations, coupled with the challenges in correlating

seismic attributes with geochemical properties, demands innovative methodological approaches (Thompson et al., 2015; Antonello et al., 2021). Furthermore, the inherent uncertainty in subsurface characterization requires robust uncertainty quantification techniques to assess the reliability of spatial predictions (Caers, 2011; Scheidt et al., 2018).

This study addresses these challenges by presenting an integrated workflow for three-dimensional TOC modeling in the Santos Basin pre-salt section. By combining seismic interpretation, synthetic well data generation, and advanced geostatistical techniques, we establish a methodology for creating geologically plausible TOC distributions that honor both local measurements and regional stratigraphic trends. We explore multiple modeling approaches—including kriging, random field generation, and machine learning—to evaluate their relative strengths in capturing spatial heterogeneity and quantifying prediction uncertainty. The resulting 3D TOC models provide valuable insights into source rock distribution patterns, with implications for petroleum system modeling and exploration strategies in similar geological settings.

3.2.2 Geological setting

The Santos Basin, formed during the Early Cretaceous (Fig. 9) breakup of the Gondwana supercontinent, extends over 350,000 km² along the southeastern margin of Brazil. It is bounded by the Cabo Frio High to the north and the Pelotas High to the south. The basin evolution follows four distinct tectonic-depositional stages: pre-rift, rift, sag (post-rift), and drift (Contreras et al., 2010; Garcia et al., 2012; Neves et al., 2019; Farias et al., 2019; De Oliveira Nardi Leite et al., 2020).

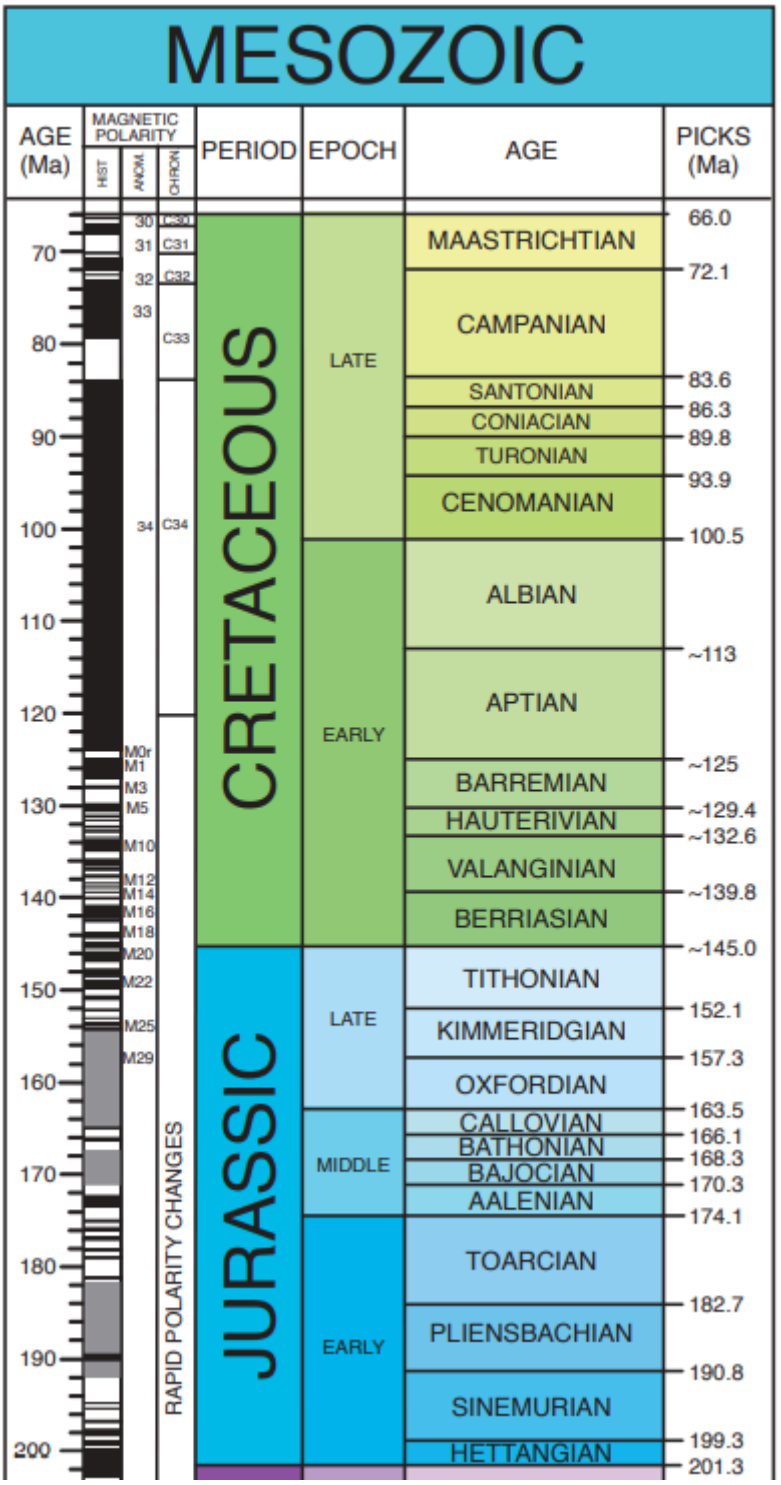


Fig. 9 - Distribution of the main paleogeographic domains through geological time within the evolutionary context of the South Atlantic rift. (Adapted from Walker et al., 2013)

Initial crustal stretching occurred during the Late Jurassic to Early Cretaceous (~161-145 Ma) as Africa and South America began to separate, forming localized depocenters receiving continental and fluvial sediments (Heine et al., 2013; Lovecchio et al., 2013). This pre-rift stage was characterized by predominantly desertic and fluvial environments under continental domain (Moulin et al., 2005). The onset of rifting was accompanied by extensive basaltic volcanism in the Paraná Basin and adjacent areas of the continental margin

between the Pelotas and Espírito Santo basins under conditions of higher temperatures and more active volcanism than today (Gust et al., 1985; Zalán, 2007; Lovecchio et al., 2024).

During the rift stage (Hauterivian–Early Aptian, ~132.6-121.4 Ma), intensive crustal extension created a system of half-grabens and grabens filled by syn-rift siliciclastic and lacustrine sequences (Riccomini et al., 2012). Lacustrine systems developed between the Hauterivian and late Barremian received sedimentary inputs predominantly from continental sources carried by rivers, with occasional marine incursions driven by eustatic variations (Vail et al., 1977a,b; Santos et al., 2023). These lacustrine environments were characterized by increasing salinity that promoted water column stratification and reduced oxygen solubility. These conditions facilitated the proliferation of anaerobic bacteria, contributing to anoxic conditions in the deeper parts of the lakes (Talbot, 1988; Trindade et al., 1995). This created ideal preservation conditions for organic matter and deposition of organic-rich shales that, under appropriate pressure, temperature, and burial conditions, transformed into source rocks with high hydrocarbon potential (Katz, 1995; Thompson et al., 2015). In the Santos Basin, these shales are interbedded with carbonates, reaching thicknesses of 100-300 m with total organic carbon (TOC) contents of 2% to 6%. The generation and expulsion of hydrocarbons began approximately 100 million years ago, peaking between 90 and 70 million years ago (Chang et al., 2008; Lourenço et al., 2014).

Following the main phase of active rifting, the basin entered the sag (post-rift) stage (Late Aptian–Early Albian, ~121.4-113 Ma) (Fig. 10a) and underwent thermal subsidence characterized by periodic marine incursions from the south (Farias et al., 2019). A topographic high, likely composed of basaltic rocks, acted as a partial barrier, regulating marine water influx and creating oscillating conditions favorable for carbonate and evaporite deposition (Carroll and Bohacs, 1999; Lime and Ros, 2019). The paleogeographic scenario resembled a narrow, elongated gulf similar to today's Red Sea between northeastern Africa and the Arabian Peninsula.

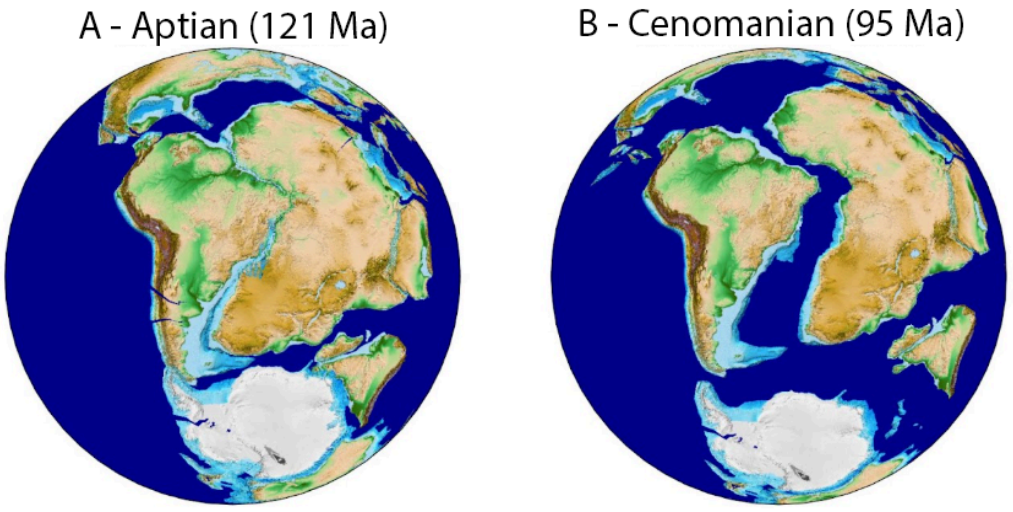


Fig. 10: Paleogeographic reconstruction of Southern Gondwana during evaporite formation and the beginning of the drift phase (Scotese, 2013). (a) Paleoreconstruction during the Aptian (121.4-113 Ma); (b) Paleoreconstruction during the Cenomanian age (100-93.5 Ma).

Increased subsidence rates during this stage caused continuous basin floor deepening. Combined with the warm climate and occasional marine incursions, this led to decreased calcium carbonate ($CaCO_3$) solubility in water, resulting in supersaturation and precipitation of calcite crystals. In hypersaline lake conditions, the greater availability of magnesium allowed for dolomite ($CaMg(CO_3)_2$) formation through Ca-Mg substitution (Tucker, 1990a,b; Bialik et al., 2018). These conditions, together with high evaporation rates, facilitated the formation of a thick evaporite succession up to 2,500 m thick (Chang et al., 1990; Ford and Vergés, 2020), recognized as "salt walls." These deposits are composed primarily of halite (NaCl) with intercalations of anhydrite, carnallite, and tachyhydrite (Gamboa et al., 2008), deposited over a timespan of 400,000-600,000 years (Freitas, 2006) during a short interval between 119-112 Ma at the Aptian-Albian boundary.

With the onset of the drift stage (Cenomanian–Present, ~100 Ma-Present) (Fig. 10b), complete separation between the South American and African continents and the formation of the South Atlantic Ocean led to widespread marine sedimentation (Faugères et al., 1993). This stage started approximately 112-111 million years ago and continues to the present day. Overlying the evaporites from the previous phase, marine to transitional sediments were deposited, primarily platform carbonates and microbialites (between 112-98 Ma and 45-3 Ma) (Valença et al., 2003), deep-water shales (from 96 Ma onward, with clear predominance from 45 Ma to present) (Rodrigues et al., 2021), and shallow-water sandstones and turbidites (from 105 Ma, with greater development between 85-45 Ma) (Pereira and Feijó, 1994; Wen et al., 2019).

As the basin continued to subside thermally during the open marine phase, organic matter distributed across a vast geographic region extending along today's eastern and equatorial margins of Brazil reached maturation (Hood et al., 1975; Kelts, 1989; Mello and Maxwell, 1990). The lacustrine origin of these deposits was particularly significant for petroleum systems in Brazil. The preservation of organic matter under anoxic conditions in these ancient Cretaceous lakes was influenced by orbital forces that affected evaporation cycles and sediment input, changing water chemistry and facilitating the accumulation of organic matter on lake bottoms (Olsen, 1990; Carroll and Bohacs, 2001; Hinnov, 2013). Subsequent events, such as the first marine incursions and the formation of "salt walls," were equally important for retaining this organic matter, while active volcanism in this extensive rift system contributed to its maturation (Williams, 1993; Zalán, 2007; De Mahiques et al., 2017). This thermal subsidence created extensive accommodation space for Tertiary to Quaternary sediment accumulation, while the structural complexity arising from multiple tectonic phases and significant thickness variations in the evaporite layers created considerable heterogeneity in fluid flow and reservoir connectivity (Mohriak et al., 1990; Jackson et al., 2015; Buckley et al., 2015), underscoring the importance of robust 3D geological models for understanding the Santos Basin petroleum systems.

3.2.3 Methods and data

This section describes our integrated approach to generate a comprehensive 3D TOC model using synthetic data. First, we perform seismic interpretation of key sequence boundaries to establish the structural framework of the model. Second, we conduct geostatistical analysis using synthetic well data and seismic attribute data where vertical variogram properties are derived from well data, and directional variogram properties from seismic attributes. Finally,

we implement a simulation workflow to distribute TOC values within the deterministic stratigraphic framework. This multi-step integration of structural interpretation, geostatistical characterization, and property modeling enables us to generate a spatial representation of TOC distribution. The complete workflow is summarized in Fig 11.

The integration of seismic attributes with geostatistical modeling represents a significant advancement in characterizing TOC distribution beyond discrete well locations (Thompson et al., 2015). While conventional approaches rely heavily on well measurements interpolated across large distances, our methodology leverages the spatial continuity information encoded in seismic data to constrain the property models (Doyen, 2007). This multi-scale integration approach overcomes the fundamental limitation of sparse sampling by incorporating the lateral continuity patterns observed in seismic attributes as proxies for geological heterogeneity (Neto et al., 2016). The seismic-derived spatial correlation structure provides critical constraints on the anisotropy and correlation lengths used in subsequent geostatistical simulations, resulting in TOC models that honor both the local measurements at wells and the broader stratigraphic architecture (Dubrule, 2003).

Geostatistical modeling in this study employs three complementary approaches—Kriging, Random Field Generation (RFG), and machine learning—each offering distinct advantages for TOC characterization (Deutsch & Pyrcz, 2014). Kriging provides minimum-variance estimates that honor well measurements exactly while minimizing estimation variance between control points, making it suitable for obtaining smoothed, conservative property distributions (Goovaerts, 1997). RFG, implemented through Sequential Gaussian Simulation, preserves the statistical variability of the input data while honoring spatial correlation structures, thereby capturing the potential heterogeneity expected in lacustrine source rocks (Journel, 2002). The machine learning approach using XGBoost integrates multiple spatial and contextual variables to identify complex, potentially non-linear relationships between location and TOC values, offering a data-driven alternative to traditional variogram-based methods (Karimpouli et al., 2020).

The comparative implementation of these methods enables comprehensive uncertainty quantification, a critical component for robust petroleum system modeling (Scheidt et al., 2018). By generating multiple equiprobable realizations through RFG, quantifying kriging variance, and producing prediction intervals through quantile regression in XGBoost, our workflow provides a probabilistic assessment of TOC distribution that can inform risk analysis in exploration scenarios (Mishra & Datta-Gupta, 2017). This multi-method approach acknowledges that no single technique can fully capture the complexities of subsurface heterogeneity, particularly in data-constrained environments like the pre-salt Santos Basin (Caers, 2011).

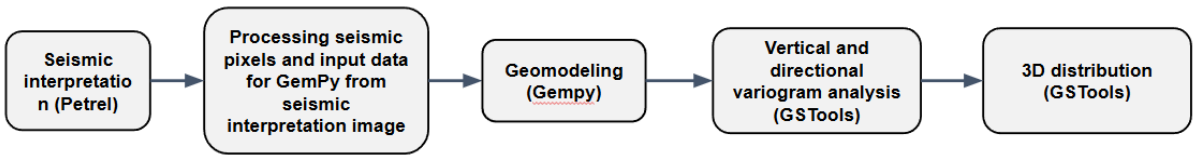


Fig. 11 - Integrated workflow for TOC modeling, showing the sequential process from seismic interpretation and well data integration through geostatistical analysis to final 3D TOC distribution simulation.

3.2.3.1 Input data

The TOC data for this model was generated using a custom Python script that simulates values for 6 synthetic wells (Table 3). The script creates TOC measurements with geologically reasonable variability, constraining values between 4% and 15% for the target formation. To maintain geological plausibility, the algorithm implements a sequential dependency where each new TOC point is generated within ±1.5% of the preceding value, with only the initial point being completely random within the specified range. This approach creates vertical TOC profiles that exhibit realistic transitions while avoiding unrealistic jumps in values. The synthetic dataset comprises a total of 430 TOC measurements across all wells, with a mean TOC value of 6.83% and standard deviation of 2.17% within the Itapema Formation, which is the primary source rock target.

Table 3. Descriptive statistics of the dataset used for TOC modeling, including counts, mean values, standard deviations, minimums, and maximums for the different model variables.

	COUNT	**MEAN**	**STD**	**MIN**	**MAX**
WELL_ID	430	-	-	1	6
DEPTH (m)	430	-6039.97	400.83	-6892.86	--5314.29
TOC (%)	430	2,28	2.82	0.01	11.77
PIXEL AMPLITUDE	430	0.49	0.27	0.00	1.00

This synthetic approach was adopted due to insufficient data in the Santos Basin pre-salt section, where currently only one well with TOC data is available within the 3D seismic survey area—insufficient for validating our workflow methodology. The synthetic wells were strategically positioned throughout the model to achieve optimal spatial coverage and test the algorithm's performance under various conditions.

The 3D seismic data were obtained from ANP, specifically the "R0258_3D_IARA_RTM_PSDM" survey from the pre-salt section of the Iara Complex, Sururu field in the Santos Basin. The data had already undergone depth migration processing and were received in SEG-Y format. While the complete 3D survey covers a larger area, this study focuses on a delimited polygon of approximately 54 km² centered on the central crossline, with a maximum depth of 10 km. The original seismic data uses WGS84 UTM Zone 23S datum with coordinate ranges of 733133.64-744025.19 (easting) and 7229467.38-7239460.50 (northing) (Fig 12). The vertical sampling extends from the seafloor (approximately 2200m depth) to 8000m. The well 1-BRSA-618-RJS, located east of the study area with a total depth of 6098m (TVD), was used for well-to-seismic tie. This well provided robust conventional log data including caliper, density, neutron, gamma ray, resistivity, and sonic logs, as well as lithological markers (well tops), checkshot tables, and composite logs. The well penetrates through the Itapema Formation rocks that comprise the Rift interval of the pre-salt section in the Santos Basin (Moreira et al., 2007). All seismic quality control, well-to-seismic tie processing, and horizon interpretation tasks were

performed using Petrel software (Schlumberger, 2010). Well-to-seismic tie was performed using checkshot data and a 20Hz Ricker wavelet with sonic (DT) and density (DENS) logs, ensuring reliable time-depth correlation. The seismic data exhibits a good signal-to-noise ratio throughout the target interval, with vertical resolution estimated at 30-35m based on dominant frequency analysis.

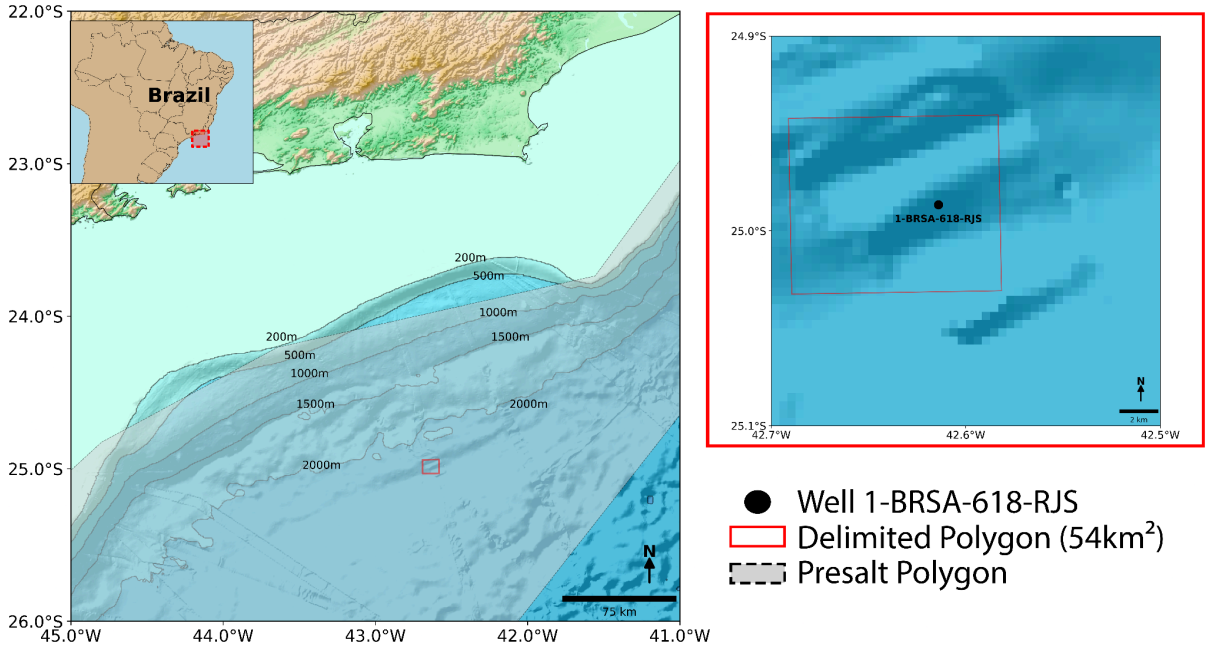


Fig. 12. - Location and bathymetric map of the study area in the Santos Basin, offshore Brazil. Left panel: Regional bathymetric map showing the continental shelf and slope with depth contours (200-2000m). Inset shows the location within Brazil, with the red dashed box indicating the Santos Basin region. Right panel: Magnified view of the study area (red polygon, 54km²) within the Iara Complex, Sururu field, with the location of well 1-BRSA-618-RJS (black circle) used for well-to-seismic tie. The blue shading represents bathymetry, with darker blue indicating deeper waters. The legend identifies the well location, delimited polygon of the study area, and the broader pre-salt polygon extent. Coordinate system: WGS84.

3.2.3.2 Seismic interpretation and geo-structural modeling

Using Petrel software, we conducted a sequence stratigraphic interpretation of the Santos Basin pre-salt section, identifying key formation boundaries. For computational efficiency and enhanced visualization, we developed a Python-based application (Screator) to process the seismic data. The application transforms coordinates to a relative local grid (0-10892m for easting and 0-10000m for northing) (Tearpock and Bischke (2003); Groshong (2006); Wang et al., 2022), and generates the interpreted surface points of the formations in a csv format compatible for Gempy, with columns X, Y, Z and formation names. Additionally, we constrained the vertical sampling to depths ranging from 4750m to 7250m, focusing specifically on the interval spanning from the base of Ariri formation to the base of Piçarras formation.

Screator enables users to interactively replicate the Petrel interpretation while recording coordinates of points along formation interfaces and extracting corresponding seismic attribute values. The seismic image was converted from RGB to grayscale using the luminosity method (Kanan and Cottrell, 2012; Gonzalez and Woods, 2018), defined as:

Grayscale = 0.299×R + 0.587×G + 0.114×B (Fig. 13a). This transformation provides a standardized numerical representation of seismic amplitude that preserves perceptual luminance relationships while reducing dimensionality (Hubral et akl., 1996; Sun et al., 2022). The extraction of amplitude values follows methodologies similar to those employed by Dorn and Shimeld (2020) and Wang et al. (2021), where pixel intensities from grayscale-converted seismic data serve as proxies for relative acoustic impedance. This approach enables quantitative analysis of formation characteristics without requiring full seismic inversion, similar to techniques described by Hami-Eddine et al. (2015) and Chopra and Marfurt (2007). For each interpreted horizon point, we extract and store amplitude values with their corresponding spatial coordinates, creating a dataset that links structural position with seismic response attributes (Fig. 13b).

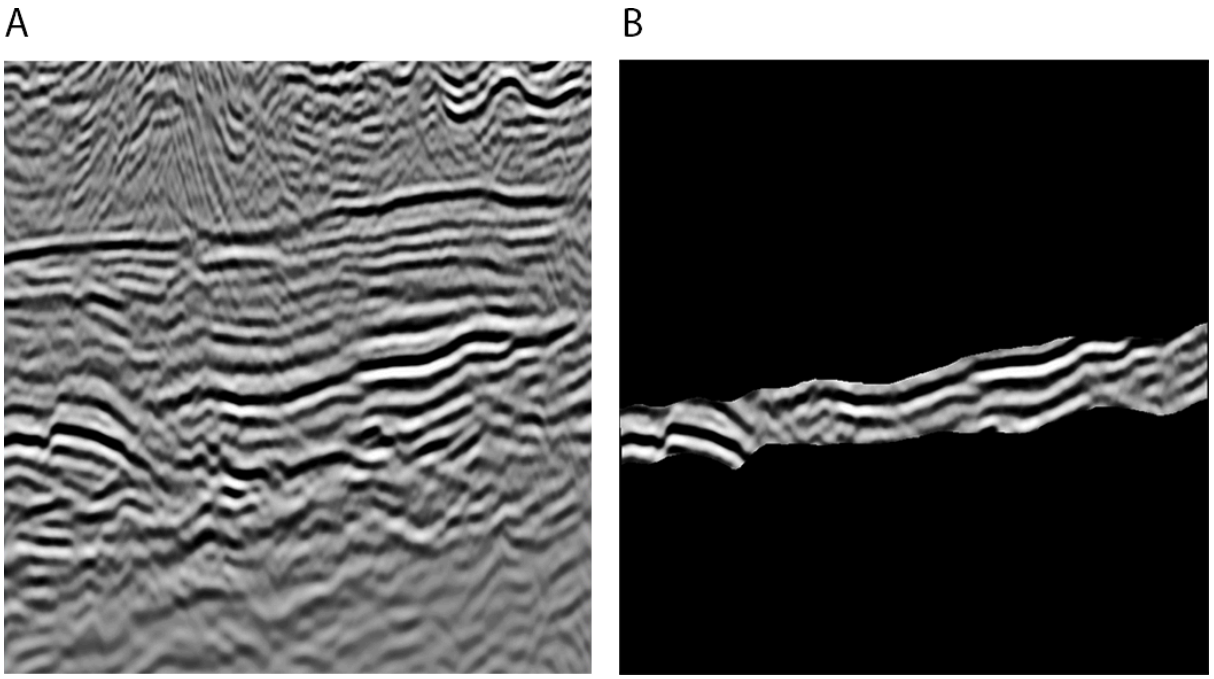


Fig. 13 - Seismic already transformed to gray-scale used for interpretation (A) and section extracted for the Itapema formation (B).

The three-dimensional geo-structural modeling was performe using GemPy (version 2.3.1), an open-source Python library that implements implicit geological modeling techniques. GemPy utilizes a mathematical approach based on potential field interpolation (de la Varga et al., 2019), where geological interfaces are represented as isosurfaces of a scalar field. In this study, surface points from Screator—exported in a CSV format with columns X, Y, Z, and formation (Fig. 14a) —were used to build a pseudo-3D structural sequence model of the Santos Basin pre-salt section.

Prior to initializing the GemPy model, a 5% buffer was applied around the minimum and maximum X, Y, and Z values from the interpreted surface points, ensuring that edges and boundaries would not adversely affect the interpolation results. Four key stratigraphic units (Ariri, Barra Velha, Itapema, Piçarras) were mapped to their respective surfaces through GemPy's stratigraphic stacking functionality, while the Basement boundary was automatically generated by the software.

A crucial aspect of the workflow was the automatic generation of orientation points using a k-nearest neighbors (KNN) algorithm with 15 neighbors (Li et al., 2024). This step analyzed the spatial relationships among the formation points to estimate local dip and azimuth for each surface, thereby minimizing subjectivity and promoting reproducibility. Once the orientation points were defined, the geological surfaces were interpolated through a Theano-optimized solver (Theano Development Team, 2016), allowing for efficient potential field computation. This approach produced a coherent 3D model of the Santos Basin pre-salt section that honors the stratigraphic relationships observed in both the seismic interpretation and well data, while ensuring a quantitative and reproducible workflow for structural interpolation (Fig. 14b and c).

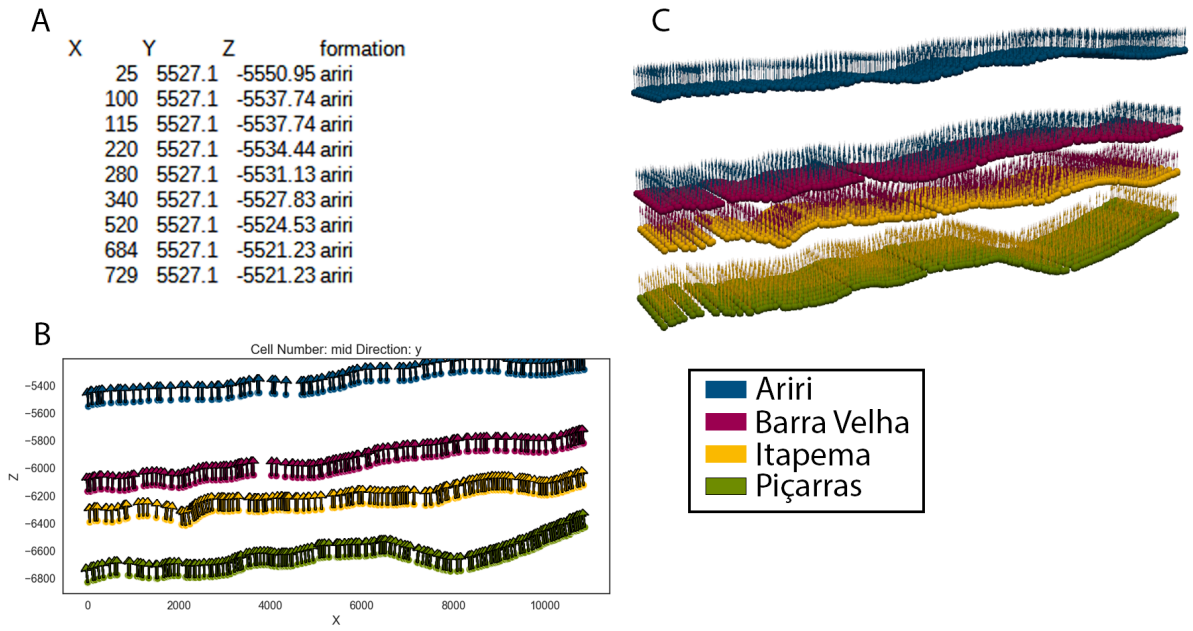


Fig. 14 - GemPy model construction workflow. A) Input data in CSV format showing the X, Y, Z coordinates and corresponding formation labels extracted from seismic interpretation. B) Distribution of surface points (colored dots) and automatically generated orientation points (arrows) for each formation, showing the spatial arrangement used for implicit interpolation. C) 3D visualization of the surface and orientations points.

3.2.3.3 Geostatistical modeling

The geostatistical analysis presented in this study employs an integrated approach using two main information sources: (a) synthetic well TOC data and (b) the 3D structural model derived from seismic interpretation. The TOC synthetic data serves as the basis for vertical variogram modeling and as conditioning points for subsequent geostatistical simulation. As borehole data alone is insufficient to estimate horizontal correlation patterns, we leverage the spatial correlation structures observed in the 3D structural model to inform lateral variability. This multi-source approach incorporates prior knowledge of the Santos Basin pre-salt stratigraphic setting, which provides soft constraints on the vertical and lateral TOC distribution patterns.

For all geostatistical analyses and simulations, we utilized the open-source Python libraries GSTools (Müller and Schüler, 2021) and scikit-gstat (Mälicke, 2021). These libraries provide robust implementations of variogram analysis, kriging, and random field generation methods, allowing for reproducible geostatistical modeling within our Python-based workflow.

3.2.3.3.1 Data preparation

Prior to geostatistical analysis, several preprocessing steps were applied to the TOC data to ensure statistical validity for the subsequent modeling procedures:

1. Trend analysis and detrending: An exponential trend model was fitted to the TOC values with respect to depth using the equation: TOC(z) = a x exp(b x z) where z is the depth, and a and b are parameters determined through curve fitting (Watson, 1986; Wu et al., 2007). This model captures the general decreasing trend of TOC with increasing depth, which is consistent with thermal maturation processes in organic-rich source rocks (Passey et al., 2010; Romero-Sarmiento et al., 2013). The exponential relationship between TOC and depth has been documented in similar basin settings by Piper and Calvert (2009), and Emmel et al. (2018) who demonstrated that organic matter degradation typically follows first-order kinetics with depth. The residuals between the observed TOC values and this trend model were computed to obtain detrended data, following methodologies established by Wang et al. (2015) and Zhao et al. (2017).

2. Normal-score transformation: The detrended residuals were transformed to follow a standard normal distribution using a quantile-quantile transformation (Dunn and Smyth, 1996; Bogner et al., 2011). This transformation was implemented using the probability integral transform method. This is necessary to satisfy the Gaussian assumption required for variogram modeling and random field generation (Deutsch and Pyrcz, 2014). The transformation preserves the rank order of the original data while ensuring that the transformed values follow a normal distribution with zero mean and unit variance.

3. Domain extraction: The spatial domain for TOC modeling was extracted from the 3D geological model by identifying the grid cells associated with the Itapema Formation, which is the primary source rock target in the pre-salt section. This extraction was performed using the GemPy model's lithology identification system, ensuring that the geostatistical analysis focuses exclusively on the formation of interest.

3.2.3.3.2 Vertical variogram analysis of well data

An experimental variogram for the vertical direction was calculated using the normalized TOC residuals from the synthetic wells (Webster and McBratney, 1989). The maximum lag distance was set to one-third of the total vertical domain extent (approximately 250 m), as recommended by geostatistical best practices (Deutsch and Pyrcz, 2014). The bin size was set equal to the vertical sampling interval of the synthetic data (approximately 3 m) to ensure sufficient point pairs for reliable statistics (Marchant and Lark, 2007).

Three theoretical variogram models—Gaussian, Spherical, and Exponential—were fitted to the experimental variogram (Groenigen, 2000), each with a nested nugget effect component (Bostanabad et al., 2018). The best-fitting model was selected based on the coefficient of determination (r^2) criterion, following the methodology proposed by Webster and Oliver (2007) and Olea (2018). The variogram parameters (nugget, range, and sill) from this best-fitting model were subsequently used to define the vertical correlation structure in the random field generation step, consistent with approaches documented by Pyrcz and Deutsch (2014) and Ringrose and Bentley (2015) for heterogeneous reservoir characterization.

The fitted variogram revealed a vertical correlation range of approximately 324.8 meters, indicating the maximum vertical distance over which TOC values exhibit spatial correlation in the Itapema Formation. This correlation length is consistent with the expected vertical heterogeneity of organic matter distribution in lacustrine carbonate source rocks of the pre-salt section (Creaney and Passey, 1993; Zuo et al., 2022).

3.2.3.3.3 Lateral correlation length derived from seismic data

The calculation of directional variograms from the primary TOC data in this study is constrained by the limited borehole data. We overcame this limitation by establishing a correlation between seismic amplitude values and TOC measurements at the synthetic well locations. Based on this correlation, we make the assumption that lateral changes in seismic amplitude reflect variations in TOC distribution. Therefore, correlation lengths derived from variogram analysis of the seismic amplitude can be used to inform the lateral continuity patterns in the TOC model (Murphy and O'Brien, 1977; Doyen, 2007; Neto et al., 2016).

This approach allows us to determine a vertical-to-horizontal anisotropy ratio for the correlation structure in the Itapema Formation. Anisotropy in geostatistical modeling refers to the directional dependence of spatial correlation, which is common in sedimentary environments due to depositional processes creating preferential alignment of geological features (Isaaks and Srivastava, 1989; Gringarten and Deutsch, 2001; Bhandari et al., 2015; Sinan et al., 2020). The horizontal variogram analysis was performed along the available crossline seismic section, which provides information about lateral continuity in one principal direction within the Santos Basin pre-salt section.

We calculated the Spearman's rank correlation coefficient (Hartmann et al., 2018) between TOC measurements and the seismic amplitude values extracted at the synthetic well locations. This approach has been successfully applied in source rock characterization studies by Løseth et al. (2011) and Carcione et al. (2015), who demonstrated significant correlations between seismic attributes and TOC content in organic-rich formations. The Spearman coefficient was chosen over Pearson correlation because it assesses monotonic relationships without assuming linearity, making it more robust for geophysical attribute correlation (Journel and Alabert, 1988; Wang and Li, 2016; Dupont et al., 2018).

The variogram analysis of the seismic amplitude data was performed on the normalized grayscale values, using the GSTools Python library (Müller and Schüler, 2021) which

implements robust algorithms for spatial statistics and random field generation. We randomly sampled 25% of the pixels from the available seismic section for variogram calculation to ensure computational efficiency while maintaining statistical robustness (Leuangthong et al., 2004; Manchuk and Deutsch, 2012). The experimental variogram was calculated using a maximum lag distance of half the section length—a commonly accepted rule-of-thumb in geostatistics to ensure reliable statistics at larger separations (Journel and Huijbregts, 1978; Goovaerts, 1997)—a tolerance angle of 22.5° and a bin size of 20 m, following established geostatistical best practices (Chiles and Delfiner, 2012; Kapageridis, 2015).

To capture the geometric characteristics of the pre-salt sequences, we determined the optimal direction of maximum continuity by iteratively computing directional variograms at different angles (Boisvert et al., 2009; Mariethoz and Caers, 2014). This approach identifies structural anisotropy, which reflects the preferential orientation of geological features due to depositional and tectonic processes (Kupfersberger and Deutsch, 1999; Boisvert and Deutsch, 2011). We examined angles between 0° and 180° in 15° increments, representing the typical gentle dip of the Santos Basin pre-salt section as documented in previous structural studies (Moreira et al., 2007; Gomes et al., 2009).

The experimental variogram was fitted with the same theoretical model type selected for the vertical direction to maintain consistency in the geostatistical approach (Gringarten and Deutsch, 2001; Emery, 2010). This practice ensures mathematical compatibility between vertical and horizontal correlation structures when implementing geometric anisotropy in 3D geostatistical simulations (Deutsch, 2002; Eltom et al., 2020). The horizontal correlation length would subsequently be used to define the anisotropy ratio in the random field generation, accounting for the difference between vertical and lateral heterogeneity typically observed in lacustrine carbonate deposits (Frykman, 2001; Corbett et al., 2012).

3.2.3.3.4 Random field generation

Conditioned Gaussian random fields were generated for the Itapema Formation using the variogram models and correlation lengths established in the previous steps. The random field generation was implemented using the GSTools library's Randomization method (Heße et al., 2014), which provides an efficient algorithms for provides efficient algorithms for Sequential Gaussian Simulation (SGS) on regular grids, consistent with approaches described by Lantuéjoul (2013) and Nussbaumer et al. (2018) for geological property modeling.

The key parameters and steps in the random field generation process were:

1. Variogram model: The theoretical variogram model selected in the vertical variogram analysis, extended to 3D with the anisotropy ratios derived from the lateral correlation length estimation.
2. Geometric anisotropy: Anisotropy ratios were calculated as the ratios between the horizontal correlation lengths and the vertical correlation length. For this synthetic dataset, the calculated ratio was approximately 1:1.8 (horizontal-to-vertical), which differs from what literature suggests for lacustrine carbonate depositional systems. Studies by Di et al. (2016) and Chitale et al. (2015) have documented that in natural lacustrine carbonate settings, lateral continuity often exceeds vertical heterogeneity

by one to two orders of magnitude (typically ranging from 20:1 to 100:1) due to episodic depositional processes and subsequent diagenetic modifications. This difference highlights a limitation of our synthetic approach, as the generated data doesn't fully capture the pronounced anisotropy that would be expected in actual pre-salt formations. However, the methodology demonstrated here could be applied to real data in future studies to better characterize the true spatial heterogeneity of the Itapema Formation.

3. Conditioning data: The normal-score transformed TOC residuals from the synthetic wells were used as conditioning points, ensuring that the generated random fields honor the well data. This conditioned simulation was implemented through ordinary kriging as an intermediate step for SGS, following established geostatistical protocols (Pyrcz and Deutsch, 2014).
4. Grid resolution: The random fields were generated on the same grid as the GemPy model (120 × 50 × 100 cells), ensuring consistency between the structural model and the property model.
5. Back-transformation: The resulting Gaussian random field values were back-transformed to the original TOC scale using the inverse of the normal-score transformation established during data preparation. The previously identified depth trend was then added back to obtain the final TOC distribution.

Multiple realizations of the random field were generated to assess the uncertainty in the TOC distribution, following the uncertainty quantification approaches outlined by Caers (2011) and Scheidt et al. (2018). This stochastic simulation approach provides a probabilistic assessment of TOC distribution, capturing the inherent uncertainty in areas between conditioning points while maintaining geological plausibility, as demonstrated in similar source rock characterization studies by Kuchinskiy et al. (2013) and Esmaeilzadeh et al. (2020).

3.2.3.3.5 XG-Boost interpolation

In addition to traditional geostatistical methods, we implemented a machine learning approach using the XGBoost algorithm (Chen and Guestrin, 2016) to predict TOC distribution. XGBoost is an efficient implementation of gradient boosted decision trees that has demonstrated superior performance in various regression tasks, including spatial prediction problems.

The XGBoost model was configured and trained as follows:

1. Feature engineering: The input features included the spatial coordinates (X, Y, Z) and additional contextual variables derived from the GemPy model, such as distance to formation boundaries and relative stratigraphic position within the Itapema Formation. This approach builds upon work by Karimpouli et al. (2020) and Bestagini et al. (2017), who demonstrated improved prediction performance when incorporating geological context into machine learning models for reservoir property prediction.
2. Hyperparameter optimization: Rather than using fixed parameters, we employed RandomizedSearchCV to systematically search for optimal hyperparameter combinations (Table 4). The optimization process evaluated 20 different parameter combinations using 5-fold cross-validation with root mean squared error (RMSE) as

the evaluation metric. This approach enables more robust model fitting by exploring a broader range of parameter combinations than would be feasible with manual tuning (Bergstra and Bengio, 2012).

3. Training: The model was trained on the detrended TOC values from the synthetic wells, with the spatial coordinates and contextual variables as predictors. All input features were standardized to ensure optimal convergence during training.
4. Uncertainty estimation: To quantify prediction uncertainty, we implemented quantile regression by training additional XGBoost models with quantile loss functions (α = 0.16 and α = 0.84), following methodologies proposed by Meinshausen (2006) and Zhang et al. (2019) for uncertainty estimation in machine learning predictions. The difference between the upper and lower quantile predictions provides an estimation of the prediction uncertainty, approximately corresponding to one standard deviation in a normal distribution, a technique validated for geospatial prediction by Hengl et al. (2018) and Kirkwood et al. (2022).
5. Prediction: The trained models were applied to predict TOC values at all grid cells within the Itapema Formation. The predictions were then back-transformed and the depth trend was added back to obtain the final TOC distribution.

Table 4 - Hyperparameters of the XGBoost Regressor Algorithm with Corresponding Search Ranges Used for RandomizedSearchCV Optimization.

Hyperparameter	Search Range	Description
n_estimators	50-300	Number of gradient boosted trees in the ensemble
max_depth	3-7	Maximum depth of each tree, controlling model complexity
learning_rate	0.01-0.31	Step size shrinkage used to prevent overfitting
min_child_weight	1-6	Minimum sum of instance weight needed in a child node
subsample	0.6-1.0	Fraction of samples used for fitting individual trees
colsample_bytree	0.6-1.0	Fraction of features used when constructing each tree

The XGBoost approach offers several advantages over traditional geostatistical methods: (1) it efficiently captures non-linear relationships between spatial position and TOC values; (2) it integrates multiple geological constraints simultaneously without requiring explicit variogram modeling; (3) it handles multivariate dependencies that might not be fully captured by two-point statistics in variogram-based approaches; and (4) it provides flexible uncertainty quantification through quantile regression. These capabilities are particularly valuable for modeling complex source rock properties in structurally heterogeneous settings like the pre-salt section, as demonstrated by recent studies applying machine learning to reservoir characterization (Karpatne et al., 2017; Bergen et al., 2019).

3.2.3.3.6 Kriging

As a comparative baseline, we implemented ordinary kriging to predict TOC distribution based solely on the well data, following the geostatistical workflows established by Goovaerts (1997) and Armstrong (1998). This kriging was selected ordinary kriging because our trend analysis provided a reliable global mean for the detrended residuals, making the constant mean assumption appropriate for this application (Chiles and Delfiner, 2012; Deutsch, 2002). Furthermore, this approach provides smoother interpolation results with minimized estimation variance, making it suitable as a conservative baseline for comparison with more complex methods, and has been successfully applied to source rock characterization by Bohling and Dubois (2003) and more recently by Caetano et al. (2017) in carbonate reservoir settings.

The kriging implementation followed these steps:

1. Variogram model: The same variogram model selected in the vertical variogram analysis was used, extended to 3D with the anisotropy ratios derived from the lateral correlation length estimation.
2. Detrended data: Kriging was performed on the detrended and normal-score transformed TOC residuals to ensure stationarity assumptions were satisfied.
3. Neighborhood definition: A neighborhood of 24 closest points was used for local kriging, with a maximum search radius of 5000 meters to ensure computational efficiency while maintaining prediction accuracy.
4. Kriging implementation: The GSTools implementation of simple kriging was used, which efficiently handles the sparse conditioning data and anisotropic variogram model, with options for pseudo-inverse calculation to enhance numerical stability.
5. Back-transformation: The kriged estimates were back-transformed to the original TOC scale, and the depth trend was added back to obtain the final TOC distribution.
6. Kriging variance: The kriging variance was calculated at each prediction location, providing a measure of prediction uncertainty based on the spatial configuration of the conditioning data and the variogram model. This uncertainty quantification approach follows methodologies described by Chilès and Delfiner (2012) and has been applied to source rock property modeling by Liu et al. (2017) and Dubrule (2003), enabling probabilistic assessment of TOC distribution in areas with limited data constraints.

The simple kriging results serve as a reference point for evaluating the performance of the more complex methods (random field simulation and XGBoost regression), providing a baseline for both prediction accuracy and uncertainty quantification.

3.2.4 Results and discussion

In this study, we present a conceptually integrated modeling workflow for the parameterization of Total Organic Carbon (TOC) distribution in the Santos basin pre-salt section. The workflow incorporates 2D depth-migrated seismic data, seismic attributes, and geostatistical workflows to arrive at a representative stochastic model that accounts for geological heterogeneity in the subsurface. The model was designed for numerical investigation of the TOC distribution in the Santos basin pre-salt section, which has been recognized as one of the most prolific petroleum systems in the South Atlantic margin

(Moreira et al., 2007; Gomes et al., 2009). Geologic heterogeneity presents significant challenges for modeling TOC distribution in lacustrine source rocks of the pre-salt sequence due to the complex depositional environment characterized by isolated lacustrine systems with variable organic matter input and preservation conditions (Antonello et al., 2021; Mello et al., 2022). Recent studies have highlighted the importance of integrating heterogeneity derived from seismic data as a key approach to improve 3D characterization for numerical modeling purposes (Thomas et al., 2019; Micallef et al., 2020). Our model achieves this by integrating the geo-structural framework, synthetic TOC data, and amplitude pixel attributes as constraints for geostatistical modeling of TOC distribution.

The use of amplitude pixel attributes as secondary data for TOC distribution modeling was designed to compensate for the limited well data available in the Santos basin pre-salt section. While current exploration efforts have yielded significant discoveries, deep well penetrations with comprehensive TOC measurements remain sparse across the basin (Berton et al., 2020; Rodriguez et al., 2018). The seismic data provided data points spread over 54 km² in the study area, allowing for spatial characterization beyond the limited well control. Amplitude attributes enable the definition of variogram properties across multiple orientations, which would otherwise not be possible given the limited number and distribution of available wells. This approach is supported by previous studies that have demonstrated relationships between seismic amplitude attributes and organic-rich intervals in source rocks (Løseth et al., 2011; Carcione et al., 2015). While our approach uses synthetic TOC data rather than measured values, this methodology provides a framework that can be refined as additional well data becomes available. The moderate correlation observed between the synthetic TOC values and amplitude pixel attributes is considered acceptable for deriving variogram properties, though not sufficient for direct property prediction, consistent with approaches used in similar geostatistical studies with limited conditioning data (Deutsch and Pyrcz, 2014; Caers, 2011).

3.2.4.1 Geo-structural Model

The geo-structural model is constrained by four major formation boundaries (Ariri, Barra Velha, Itapema, and Piçarras), forming a pseudo-3D representation of the Santos Basin pre-salt section. Two criteria were considered in selecting the boundaries: (1) the formations yield coherent seismic reflections that could be easily identified on the seismic section, (2) they subdivide the model domain into the key stratigraphic units of the pre-salt petroleum system, with Itapema Formation representing the primary source rock and Barra Velha Formation containing the main carbonate reservoir facies (Mancini et al., 2008; Gomes et al., 2020). The 3D stratigraphic framework shown in Fig. 15 was generated using a single interpreted seismic line with crossline extrapolation to create a pseudo-3D volume encompassing approximately 54 km². The model domain was discretized into 120 × 50 × 100 cells, resulting in 600,000 total active cells for computation (Liu et al., 2021).

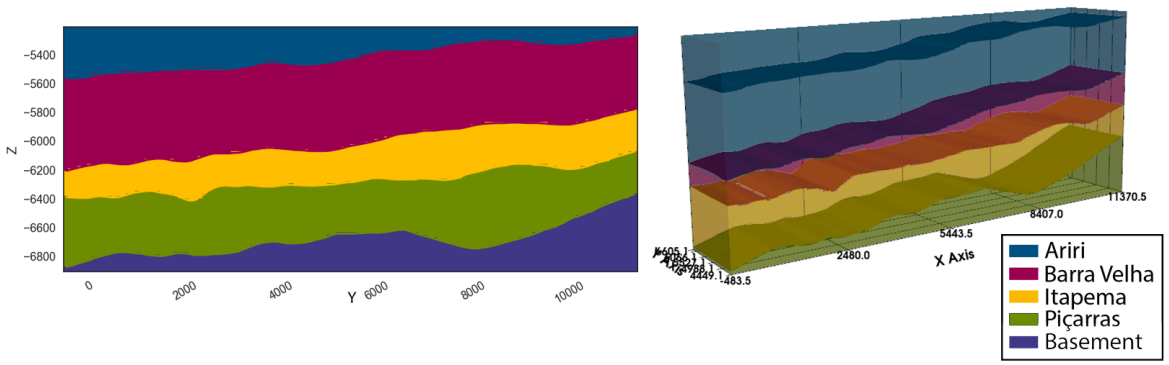


Fig. 15 - Geo-structural Model. 2D cross-section of the interpolated model showing formation boundaries generated from the surface and orientation points, and pseudo-3D visualization of the final geological model showing the spatial relationships between the pre-salt formations and basement.

The bounding surfaces are continuous across the model domain, with gentle dips characteristic of the Santos Basin's passive margin setting. The formations display a seaward-thickening wedge geometry, with maximum dip orientation toward the southeast. The cross-sectional view (Fig. 15) reveals that the formation boundaries maintain relatively consistent thicknesses across the model domain, with some localized variations in the Barra Velha Formation. This structural configuration aligns with previous studies describing the Santos Basin's pre-salt section as a series of relatively tabular carbonate and siliciclastic units deposited in a lacustrine environment with limited tectonic deformation post-deposition (Buckley et al., 2015; Wright and Barnett, 2015).

The seismic characteristics of these formations provide insights into their depositional environments, as seen in Fig. 16. The Ariri Formation (purple) displays high-amplitude, continuous reflectors representing the evaporitic seal deposited during marine incursion. The Barra Velha Formation (blue) shows more heterogeneous reflection patterns, consistent with its complex carbonate facies architecture including microbialites, shrubs, and spherulites (Arienti et al., 2018). The Itapema Formation (green) exhibits relatively continuous, moderate-amplitude reflections that correspond to organic-rich calcilutites, marls, and coquinas, making it the primary TOC-bearing interval in the sequence (Gonçalves et al., 2015). The basal Piçarras Formation (orange) displays discontinuous reflections typical of syn-rift siliciclastic deposits.

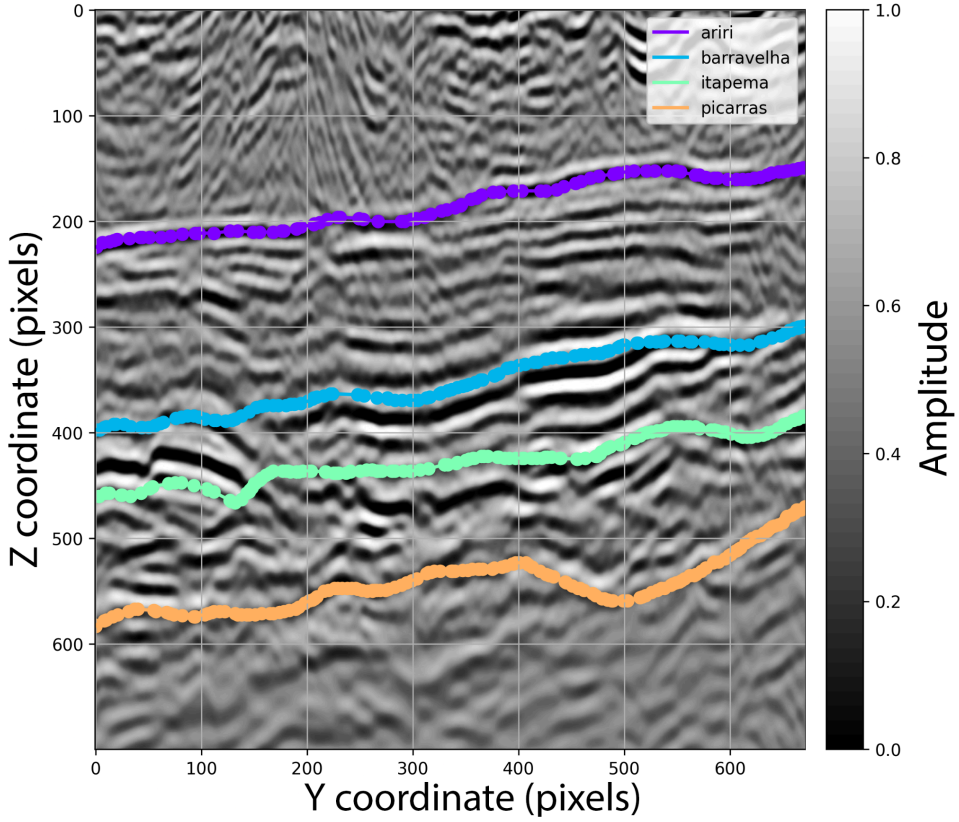


Fig. 16 - Seismic interpretation of key pre-salt formations. Interpreted horizons from the 3D seismic dataset showing the main stratigraphic units of the Santos Basin pre-salt section (Ariri, Barra Velha, Itapema, and Piçarras formations). The interpretation serves as the primary input for the 3D geological modeling workflow.

This geo-structural framework provides the foundation for subsequent property modeling, ensuring that TOC distribution honors the key stratigraphic relationships of the Santos Basin pre-salt petroleum system. While the pseudo-3D approach introduces some uncertainty in the crossline dimension, it represents a pragmatic solution given the limited availability of closely-spaced 2D seismic data in the study area.

3.2.4.2 Geostatistical modeling - Variogram model

Prior to geostatistical analysis, we performed extensive processing of the synthetic TOC data to ensure statistical validity. The TOC data exhibited a depth-related exponential trend (Fig. 17), which was modeled as TOC(z) = 0.0749 x exp(-0.00073 x z), with an R^2 value of 0.0647. While the correlation is relatively weak, this trend was implemented as part of our synthetic data generation approach. It's worth noting that in real geological settings, TOC distributions can be influenced by various factors including depositional conditions, preservation environments, and thermal maturity (Tissot & Welte, 1984; Peters & Cassa, 1994; Bohacs et al., 2000). For our methodology testing purposes, this exponential relationship provides a reasonable baseline to demonstrate the trend removal process required in geostatistical workflows, though it may not reflect the actual TOC-depth relationships in the pre-salt Santos Basin formations.

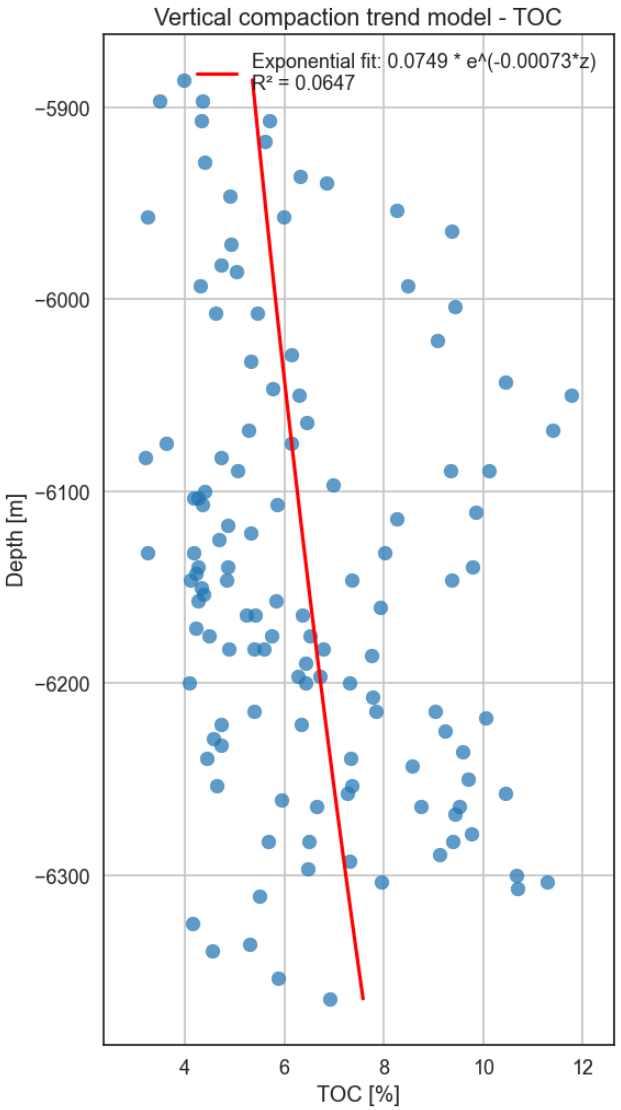


Fig. 17 - Vertical synthetic TOC trend showing the exponential relationship between TOC and depth. Blue points represent synthetic TOC measurements from wells in the Itapema Formation, while the red line shows the exponential fit (TOC(z) = 0.0749 x exp(-0.00073 x z), R² = 0.0647).

After identifying this vertical trend, we performed detrending to isolate the residual TOC variations that are independent of depth (Fig. 18). This step is crucial for variogram analysis, as it ensures that the spatial correlation structure is not dominated by the global depth trend (Viera et al., 2010). The detrended values exhibit a more stationary behavior, which is a necessary condition for reliable geostatistical modeling (Miguel et al., 2012).

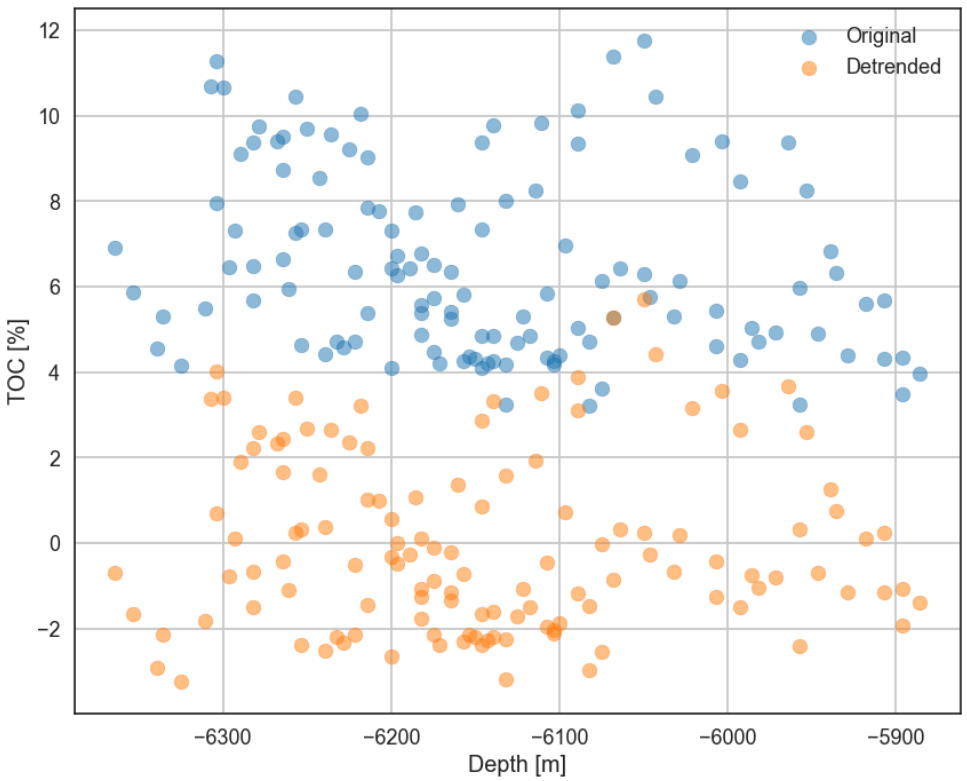


Fig. 18 - Comparison of original and detrended TOC values versus depth. Blue points represent the original TOC measurements, while orange points show the corresponding detrended values after removing the exponential depth trend.

Following detrending, we applied normal-score transformation to convert the detrended TOC residuals to follow a standard normal distribution (Fig. 19). This transformation is essential for variogram modeling and subsequent Sequential Gaussian Simulation, as it satisfies the Gaussian assumption required for these geostatistical methods (Van Den Boogaart et al, 2017). The histograms and cumulative distribution functions (CDFs) before and after transformation demonstrate the successful normalization of the data. (Robertson et al., 2006)

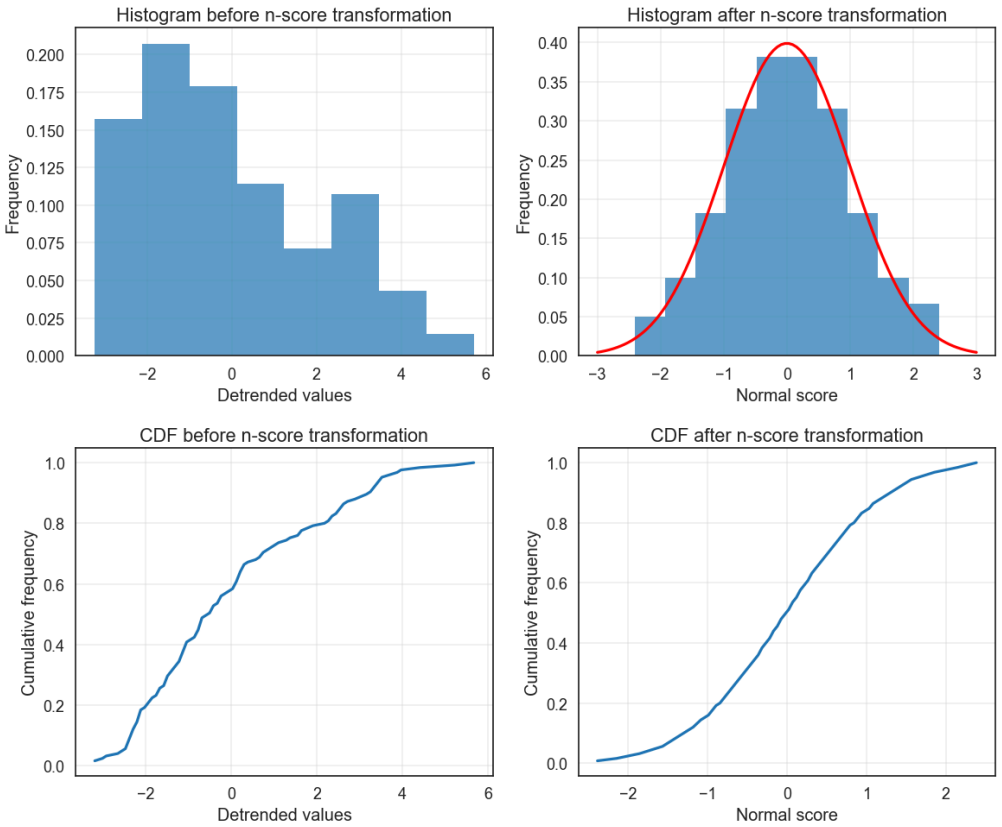


Fig. 19 - Normal-score transformation of detrended TOC values. Top row shows histograms before (left) and after (right) transformation, with the red curve on the right representing a standard normal distribution. Bottom row shows the cumulative distribution functions (CDFs) before and after transformation, illustrating the conversion to a normal distribution.

The same preprocessing workflow was applied to the seismic amplitude data. The pixel amplitude values also exhibited a weak depth-dependent trend (Fig. 20), modeled as Amplitude(z) = 0.0351 x exp(-0.00042 x z), with an R^2 of 0.0068. Despite the weak correlation, removing this trend ensures that subsequent spatial analyses are not influenced by systematic depth variations (Zhang et al., 2002).

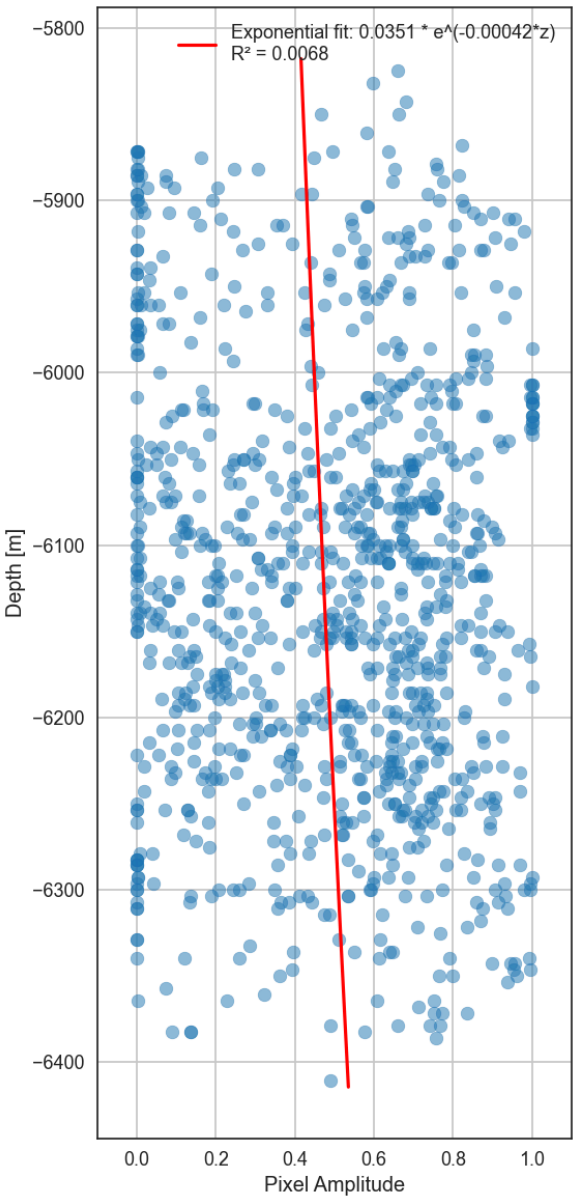


Fig. 20 - Vertical trend of pixel amplitude values with depth. Blue points represent seismic amplitude measurements, while the red line shows the exponential fit (Amplitude(z) = 0.0351 * exp(-0.00042 * z), R^2 = 0.0068).

After detrending the amplitude data (Fig. 21), we performed normal-score transformation to convert the residuals to a standard normal distribution (Fig. 22). The bimodal distribution observed in the original detrended values was successfully transformed into a normal distribution (Cheadle et al., 2003), as evidenced by the histogram and CDF plots.

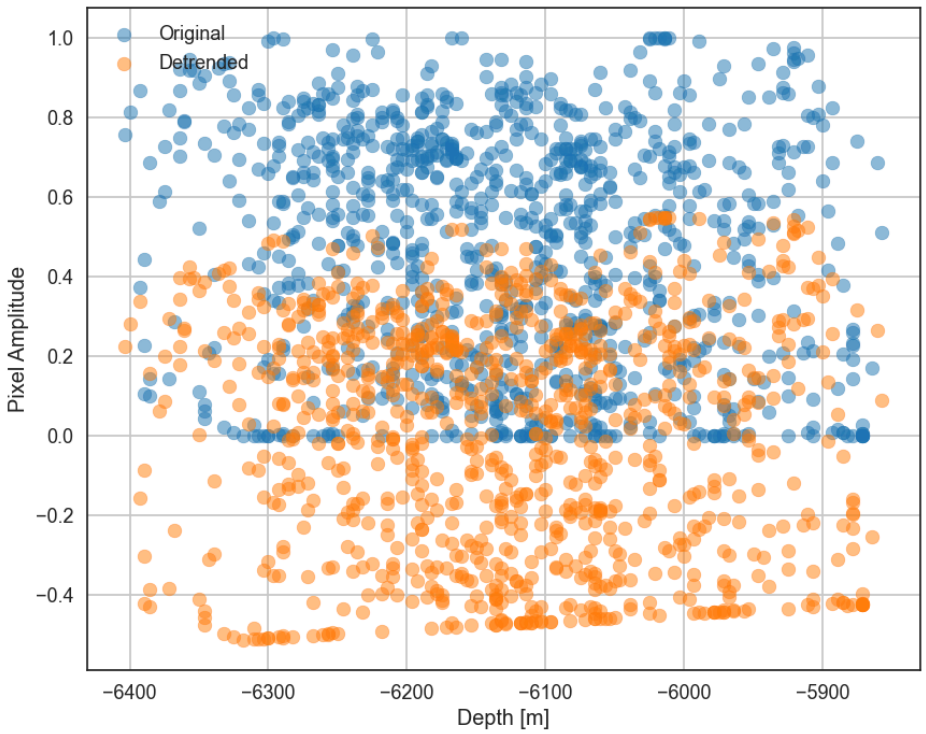


Fig. 21 - Comparison of original and detrended pixel amplitude values versus depth. Blue points represent the original amplitude measurements, while orange points show the corresponding detrended values after removing the exponential depth trend.

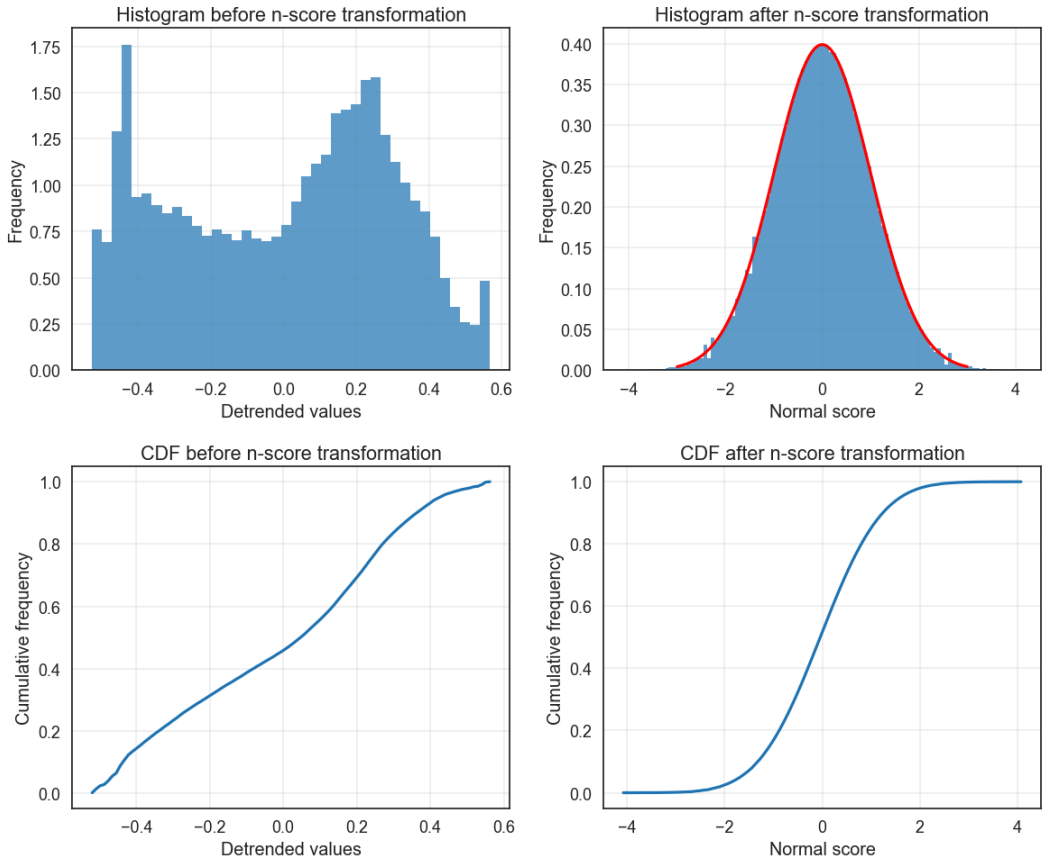


Fig. 22 - Normal-score transformation of detrended pixel amplitude values. Top row shows histograms before (left) and after (right) transformation, with the red curve on the right representing a standard

normal distribution. Bottom row shows the cumulative distribution functions (CDFs) before and after transformation.

Using the normalized data, we performed variogram analysis to characterize the spatial correlation structure of both TOC and seismic amplitude. For the vertical variogram, we determined the optimal number of bins to be 9 through statistical analysis, which balanced resolution with sufficient data pairs per bin (minimum 32 pairs per bin, average 969 pairs per bin). For the directional variogram, we used 20 bins as determined by the optimal bin calculation (average 69,838,072 pairs per bin, minimum 3,305,973 pairs per bin).

The vertical variogram analysis revealed a correlation range of 79.8 meters with a mean semivariogram value of -0.003 (Fig. 23, left panel). For horizontal directions, the directional variogram analysis identified the angle of maximum continuity at 0° with a correlation range of 44.2 meters (Fig. 23, center panel). The anisotropy rose diagram (Fig. 23, right panel) illustrates the spatial correlation structure, with the principal direction of continuity aligned with the 0° axis.

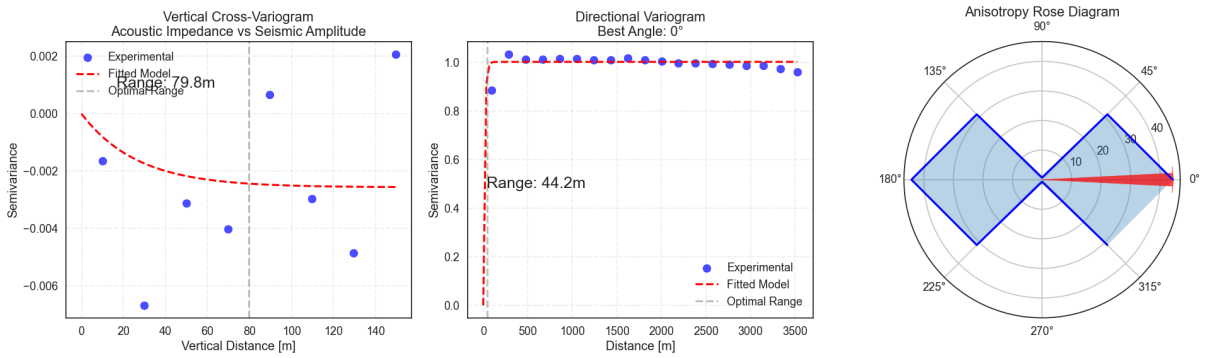


Fig. 23 - Variogram analysis results. Left: Vertical variogram showing a correlation range of 79.8m. Center: Directional variogram at 0° (angle of maximum continuity) showing a correlation range of 44.2m. Right: Anisotropy rose diagram illustrating the directional variability of spatial correlation.

This anisotropy ratio (approximately 1:1.8 for horizontal-to-vertical) is notably lower than initially expected for the Santos Basin pre-salt lacustrine depositional environment. For this synthetic dataset, the calculated ratio was approximately 1:1.8 (horizontal-to-vertical), which differs from what literature suggests for lacustrine carbonate depositional systems. Studies by Di et al. (2016) and Chitale et al. (2015) have documented that in natural lacustrine carbonate settings, lateral continuity often exceeds vertical heterogeneity by one to two orders of magnitude (typically ranging from 20:1 to 100:1) due to episodic depositional processes and subsequent diagenetic modifications. This difference highlights a limitation of our synthetic approach, as the generated data doesn't fully capture the pronounced anisotropy that would be expected in actual pre-salt formations. However, the methodology demonstrated here could be applied to real data in future studies to better characterize the true spatial heterogeneity of the Itapema Formation. The variogram parameters derived from this analysis—including the principal range of 44.2m, secondary range of 0.7m, vertical range of 79.8m, sill of 1.002, and nugget effect of 0—were subsequently used to constrain the TOC distribution modeling via SGS, kriging, and machine learning approaches.

Three property modeling methods were implemented and compared for TOC distribution (Fig. 24): Kriging (SK), Random Field Generation (RFG) using Sequential Gaussian Simulation, and XGBoost machine learning regression. All methods incorporated the vertical trend model, with RFG additionally integrating the seismic-derived spatial correlation structure. The results show distinct statistical characteristics, with kriging producing the smoothest distribution (TOC range: 4.52% to 7.97%) compared to the more heterogeneous distributions from RFG (TOC range: 2.58% to 10.97%) and XGBoost (TOC range: 4.19% to 10.81%). This behavior is expected, as kriging is known to minimize estimation variance (Chilès and Delfiner, 2012; Goovaerts, 1997), while stochastic simulation and machine learning approaches better preserve the inherent variability of the property being modeled (Deutsch and Journel, 1998; Mariethoz and Caers, 2014).

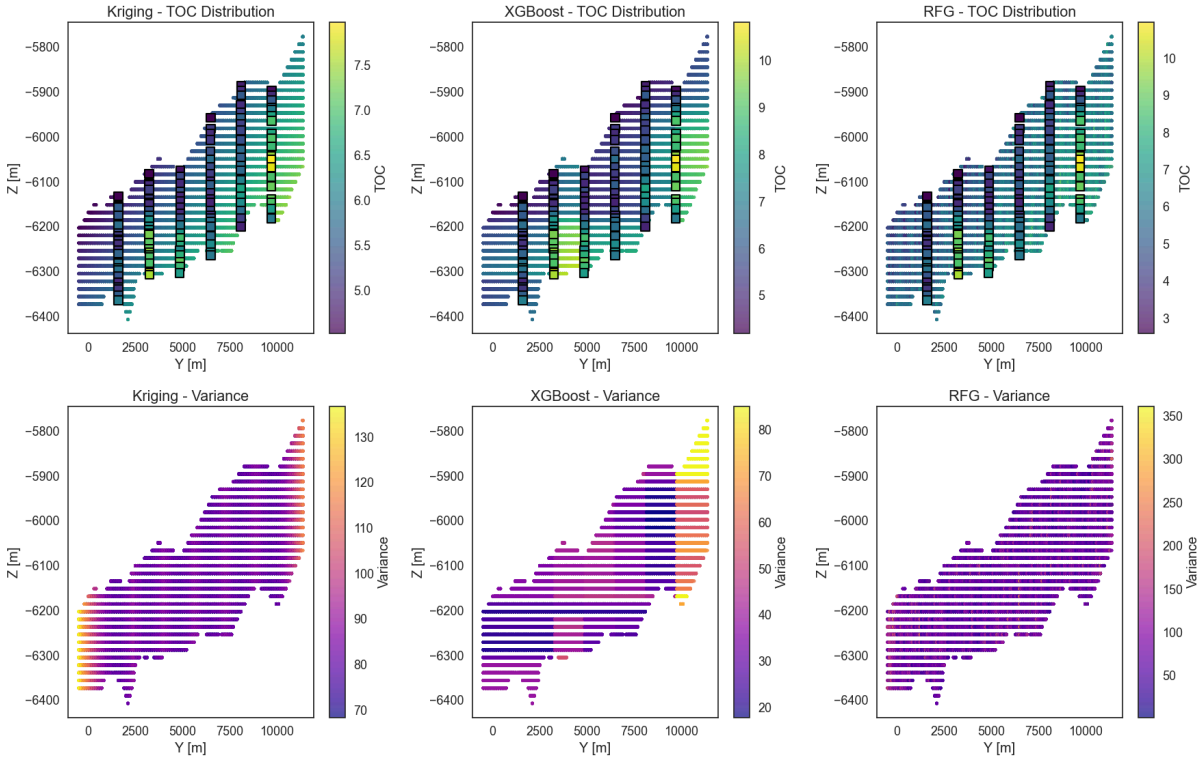


Fig. 24. - Comparison of interpolation methods for TOC distribution in the Itapema Formation. Top row shows TOC distribution for Kriging (left), XGBoost (center), and RFG (right). Bottom row shows the corresponding prediction variance for each method, highlighting different uncertainty characterization approaches.

The variance maps (Fig. 24, bottom row) further illustrate the fundamental differences between these methods. Kriging variance is primarily controlled by data configuration, showing lowest uncertainty near well locations and highest uncertainty in areas distant from conditioning data, consistent with theoretical expectations (Isaaks and Srivastava, 1989; Goovaerts, 1997). The XGBoost variance reflects the algorithm's confidence based on feature similarity, with an overall lower uncertainty range but more complex spatial patterns, characteristic of machine learning approaches that capture non-linear relationships (Chen and Guestrin, 2016; Hengl et al., 2018). The RFG approach produces the highest variance (up to 350 compared to 130 for kriging and 80 for XGBoost), characteristic of stochastic simulation methods that aim to reproduce the full spectrum of spatial variability rather than providing smoothed estimates (Journel, 2002; Lantuéjoul, 2013).

Vertical analysis of TOC predictions against well data (Fig. 25) reveals how each method handles the depth trend and local variations. Kriging produces a smooth, consistent depth trend that closely follows the global exponential model with minimal local deviations, reflecting its property of exact interpolation at data locations (Webster and Oliver, 2007). The narrow range of kriging predictions (4.52% to 7.97%) represents a significant smoothing effect compared to the original data range (3.21% to 11.77%), indicative of the method's tendency to minimize estimation variance at the expense of reproducing extreme values (Pyrcz and Deutsch, 2014). XGBoost shows more discrete clustering of predictions with distinct zones of higher TOC values, reflecting its tree-based architecture that tends to group similar instances (Friedman, 2001; Chen and Guestrin, 2016). Despite hyperparameter optimization (colsample_bytree: 0.926, learning_rate: 0.030, max_depth: 3, min_child_weight: 4, n_estimators: 62, subsample: 0.674), the XGBoost model exhibited poor validation metrics (RMSE: 2.1452, R^2: -0.0182), indicating challenges in establishing meaningful correlations between spatial coordinates and TOC values. The RFG results demonstrate the greatest local variability while still honoring the conditioning data, with prediction ranges (2.58% to 10.97%) most closely matching the original data distribution (3.21% to 11.77%), consistent with the behavior of stochastic simulation methods that aim to reproduce the input data statistics (Goovaerts, 1997; Deutsch and Journel, 1998).

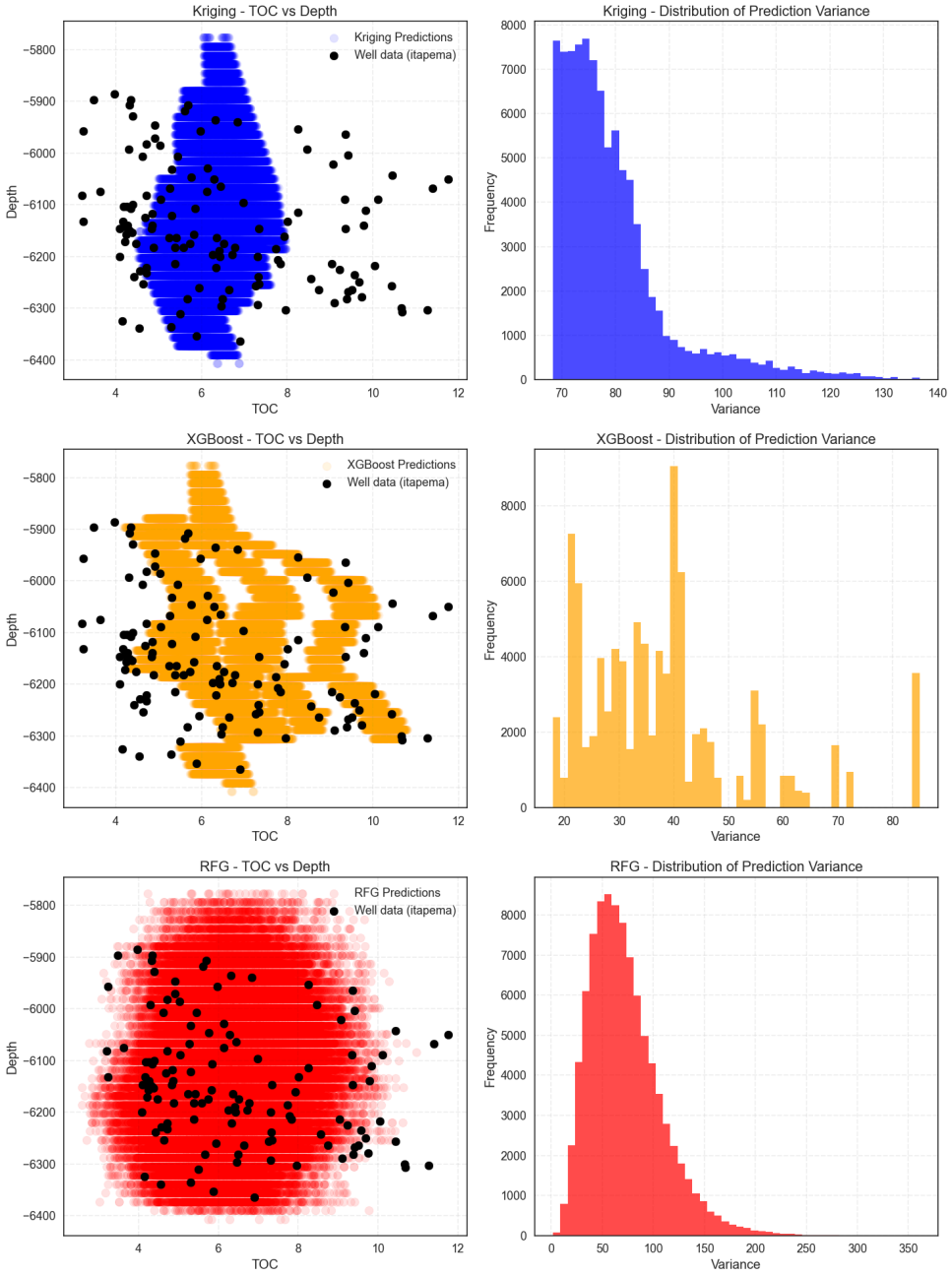


Fig. 25. - TOC versus depth plots comparing predictions with well data for all three methods. Left column shows predicted values against actual well measurements. Right column shows the distribution of prediction variance, highlighting the different uncertainty quantification approaches of each method.

These differences in vertical trends are further illustrated in the method comparison plot (Fig. 26), where Kriging and RFG show remarkably similar mean trends throughout most of the depth interval, while XGBoost demonstrates greater deviation, particularly in the middle sections (-6000m to -6300m) where it predicts higher TOC values. This behavior reflects the fundamentally different mathematical approaches of these methods, with kriging providing minimum-variance estimates based on weighted averages (Armstrong, 1998), XGBoost capturing potentially non-linear relationships through hierarchical decision trees (Chen and Guestrin, 2016), and RFG honoring both the conditioning data and the spatial correlation structure through stochastic simulation (Deutsch and Journel, 1998).

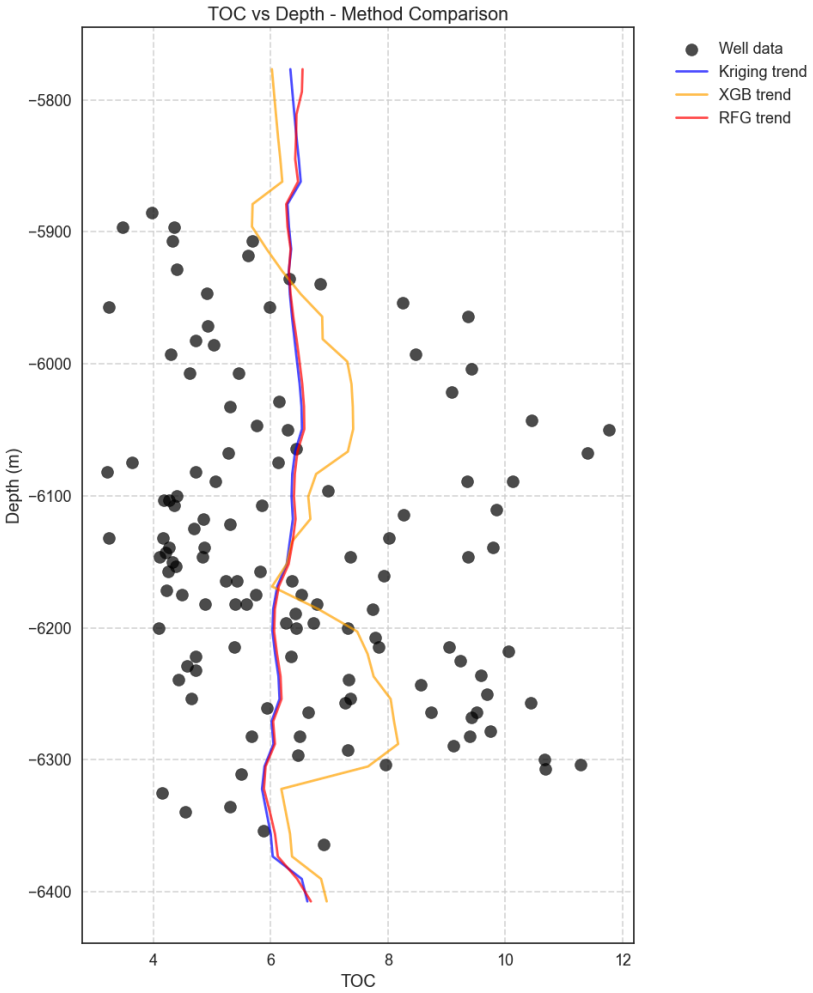


Fig. 26. - Method comparison of vertical TOC trends versus depth. The plot shows well data (black dots) alongside the mean predicted TOC trends for Kriging (blue), XGBoost (orange), and RFG (red), illustrating the different vertical characterization patterns of each approach.

Cross-sectional views of TOC distribution through the six synthetic wells (Fig. 27) highlight the spatial heterogeneity captured by each method. The kriging result (Fig. 27a) shows a very smooth, laterally continuous TOC distribution with gradual transitions, characteristic of deterministic interpolation methods (Isaaks and Srivastava, 1989). The XGBoost result (Fig. 27b) exhibits more abrupt transitions between TOC zones, particularly in the central portion of the Itapema Formation, reflecting the decision tree structure of gradient boosting algorithms (Friedman, 2001). The RFG result (Fig. 27c) displays the most heterogeneous distribution, with discontinuous high-TOC patches that better reflect the expected facies variability in lacustrine source rock deposits of the pre-salt section (Arienti et al., 2020; Wright and Barnett, 2015).

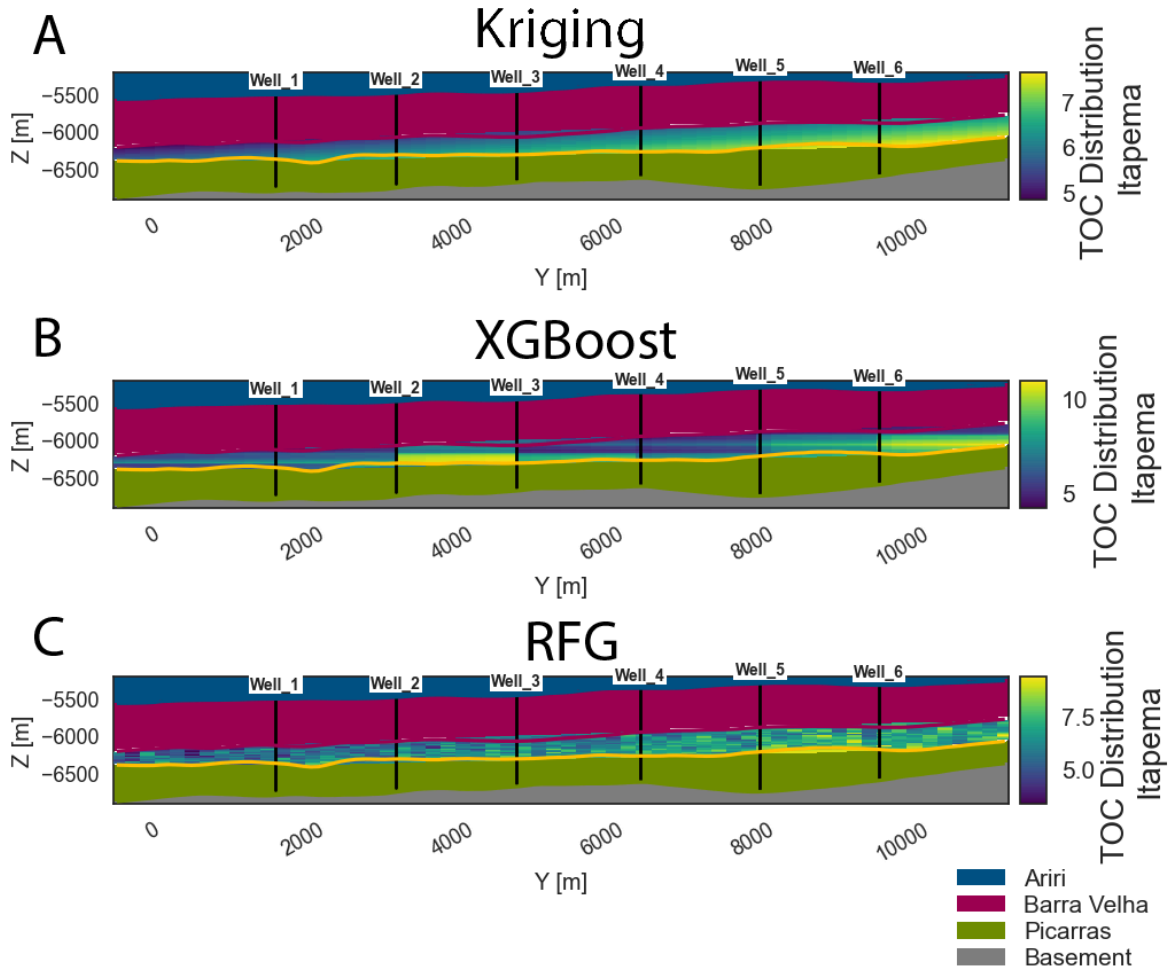


Fig. 27 - Cross-sectional views of TOC distribution through the six synthetic wells. (A) Kriging result showing smooth, laterally continuous TOC distribution. (B) XGBoost result showing more distinct zonation patterns. (C) RFG results displaying the highest heterogeneity and most complex spatial patterns among the three methods.

The integration of seismic-derived spatial correlation information proved particularly valuable in the RFG approach, enabling the model to capture not only the vertical TOC variations observed in wells but also the lateral continuity patterns evident in the seismic data. This multidisciplinary integration approach represents a significant advancement over traditional well-based interpolation methods by incorporating geologically reasonable patterns of spatial heterogeneity across the entire model domain, similar to integrated approaches described by Dubrule (2003), Doyen (2007), and more recently by Al-Mudhafar (2017) and Karimpouli et al. (2020).

3.2.4.3 TOC model

The 3D TOC model for the Itapema Formation (Fig. 28) represents the spatial distribution of organic carbon content within this primary source rock interval of the Santos Basin pre-salt section. The three modeling approaches—Kriging, RFG, and XGBoost—produced broadly similar large-scale patterns but differed significantly in their representation of small-scale heterogeneity and prediction ranges.

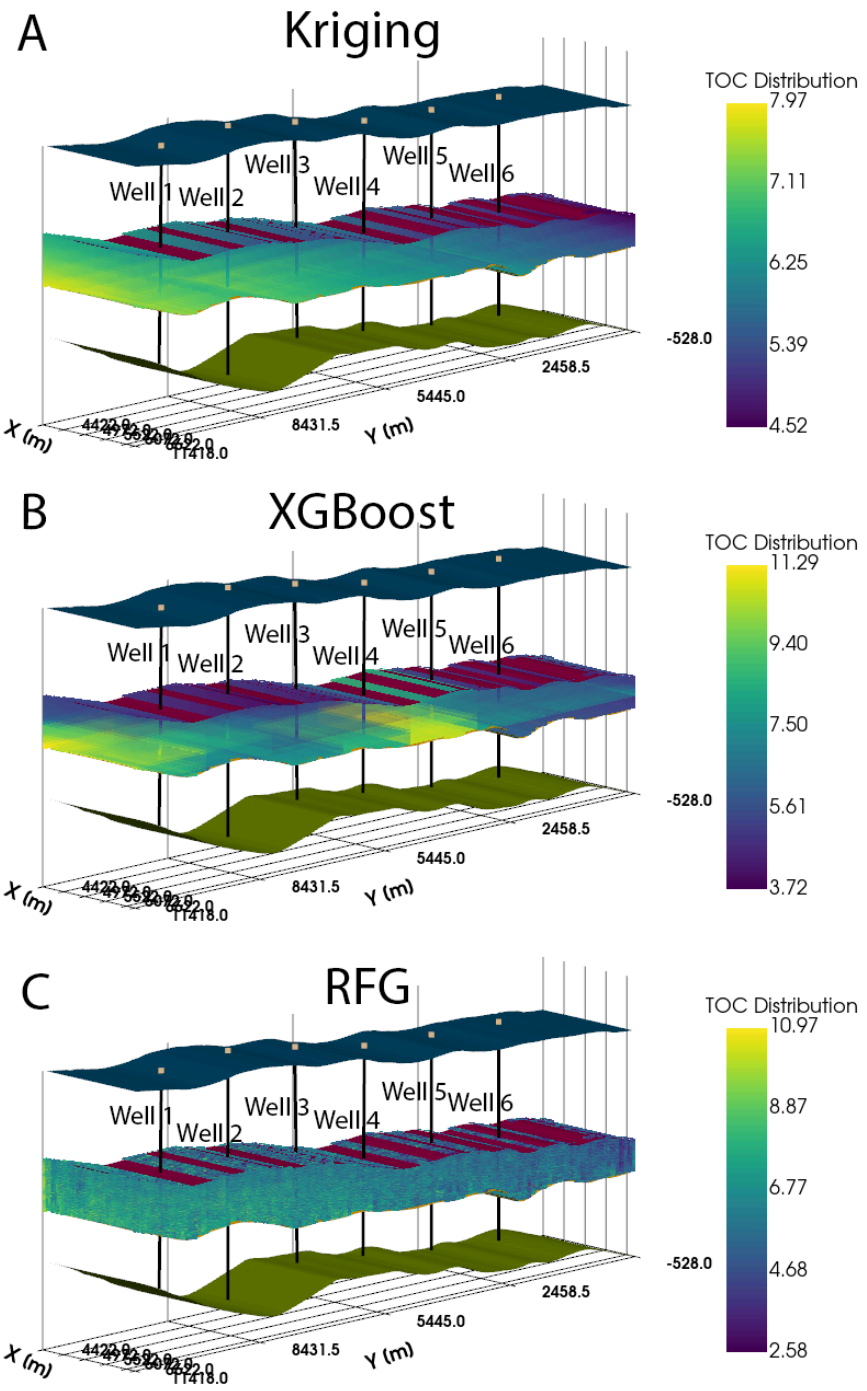


Fig. 28 - 3D visualization of TOC models for the Itapema Formation: (A) Kriging model showing smooth TOC distribution ranging from 4.52% to 7.97%. (B) XGBoost model showing TOC distribution ranging from 4.19% to 10.81% with distinct spatial patterns. (C) RFG model showing heterogeneous TOC distribution ranging from 2.58% to 10.97%.

The Kriging TOC model (Fig. 28a) exhibits a smooth distribution pattern with values ranging from 4.52% to 7.97%, displaying a strong depth-dependent trend and limited lateral variability. This smooth character is consistent with the minimum-variance property of kriging interpolation, which tends to produce conservative estimates by dampening extreme values (Pyrcz and Deutsch, 2014). The TOC distribution shows a general increase toward the top of the Itapema Formation, consistent with the expected thermal maturation pattern in which

deeper sections experience greater thermal alteration and consequent reduction in TOC (Tegelaar and Noble, 1994; Peters et al., 2005). The limited range of kriging predictions represents a significant smoothing effect compared to the original data range (3.21% to 11.77%), emphasizing broad trends at the expense of local heterogeneity.

The XGBoost model (Fig. 28b) produces an intermediate result with TOC values ranging from 4.19% to 10.81%, capturing more heterogeneity than kriging but with more structured patterns than RFG. The machine learning approach creates distinct zones of similar properties rather than the continuously varying fields produced by geostatistical methods, which may better represent compartmentalized depositional environments characteristic of lacustrine settings. This zonation pattern is particularly evident near transitional areas between the Itapema Formation and adjacent units (Karpatne et al., 2017), where the algorithm detects subtle changes in the input features that might correspond to facies boundaries or depositional shifts. While the model struggles with overall prediction accuracy, its ability to identify potential geological compartments offers complementary insights when integrated with the other modeling approaches.

The RFG model (Fig. 28c) displays considerably more heterogeneity, with TOC values ranging from 2.58% to 10.97%. This range most closely approximates the original data distribution (3.21% to 11.77%), consistent with the method's aim to reproduce input data statistics. This approach captures both the overall trends and the local variability expected in organic-rich lacustrine deposits, where TOC distribution is influenced by multiple factors including sediment input variations, water column stratification, and preservation conditions (Bohacs et al., 2000; Katz and Lin, 2014). The RFG implementation incorporates the seismic-derived spatial correlation structure (principal range: 44.2m, secondary range: 0.7m, vertical range: 79.8m, sill: 1.002, nugget effect: 0), resulting in a geologically plausible representation of TOC heterogeneity. This stochastic approach is particularly valuable for uncertainty quantification in subsequent basin modeling applications, as it can generate multiple equally probable realizations (10 realizations were generated in this study) to assess the impact of TOC variability on hydrocarbon generation and migration (Hantschel and Kauerauf, 2009; Baur et al., 2019).

All three modeling approaches show the highest TOC values concentrated in the middle to upper sections of the modeled Itapema Formation. While this pattern appears to align with the expected stratigraphic evolution documented in the Santos Basin (Gonçalves et al., 2020) where optimal conditions for organic matter preservation occurred during specific phases of lacustrine development (Wright and Barnett, 2015; Mello et al., 2022), it's important to note that this consistency is incidental to our synthetic data generation approach rather than confirmatory evidence. Future studies using real well data could investigate whether actual TOC distributions follow similar patterns and potentially validate the understanding of organic-rich intervals being associated with optimal lake development phases when water column stratification promoted organic matter preservation.

The model results provide critical input for petroleum system modeling of the Santos Basin pre-salt play, enabling more accurate prediction of hydrocarbon generation volumes and timing. The spatial variability captured in the models, particularly in the RFG approach, allows for refined assessment of source rock heterogeneity's impact on migration pathways and accumulation patterns (Hantschel and Kauerauf, 2009; Peters et al., 2012).

Furthermore, the integration of seismic attributes with well data demonstrates a workflow that can be readily updated as additional data becomes available, providing a flexible foundation for ongoing exploration and development activities in this prolific petroleum province.

3.2.5 Conclusion

This study presents an integrated workflow for three-dimensional Total Organic Carbon (TOC) modeling in the Santos Basin pre-salt section, demonstrating the successful integration of seismic attributes, synthetic well data, and advanced geostatistical techniques. Three modeling approaches—Kriging, Random Field Generation (RFG) using Sequential Gaussian Simulation, and XGBoost machine learning—were implemented and compared, each offering distinct advantages for characterizing spatial TOC distribution.

The TOC models developed through this workflow reveal distinct patterns of organic richness within the Itapema Formation, with all methods consistently identifying higher TOC values in the middle to upper sections. This pattern aligns with our understanding of the depositional evolution of the Santos Basin, where optimal conditions for organic matter preservation occurred during specific phases of lacustrine development. The RFG approach produced the most heterogeneous TOC distribution (2.58-10.97%), capturing the expected facies variability in lacustrine source rocks, while Kriging produced a smoother distribution (4.52-7.97%) that emphasizes the broad trends. The XGBoost results (4.19-10.81%) offered an intermediate level of heterogeneity with distinctive spatial patterns but suffered from poor validation metrics (R^2: -0.0182).

Each modeling approach demonstrated unique uncertainty characteristics. Kriging variance was primarily controlled by data configuration, showing lowest uncertainty near well locations. XGBoost variance reflected the algorithm's confidence based on feature similarity, with an overall lower variance range but more complex spatial patterns. The RFG approach produced the highest overall variance (up to 350 compared to 130 for kriging and 80 for XGBoost), characteristic of stochastic methods that aim to reproduce the full spectrum of spatial variability.

The integration of seismic-derived spatial correlation information proved particularly valuable in constraining the TOC distribution away from well control. By leveraging amplitude attributes to establish variogram models with a vertical range of 79.8m and a principal horizontal range of 44.2m, we identified an anisotropy ratio of approximately 1:1.8 for horizontal-to-vertical correlation. The resulting anisotropy ratio was derived from these synthetic data. It's important to note that this ratio is used primarily to validate the workflow methodology rather than to draw definitive conclusions about the actual depositional patterns of the Itapema Formation. With real data, such analysis could potentially provide insights into whether the formation exhibits compartmentalized depositional patterns or more laterally continuous facies as commonly described in the literature.

Our findings have important implications for petroleum system modeling in the Santos Basin pre-salt play. The detailed 3D TOC distributions generated through this workflow can

significantly improve hydrocarbon generation volume calculations, thermal maturity modeling, and migration pathway predictions. The heterogeneous TOC distribution captured in the RFG model, in particular, enables more realistic simulation of differential hydrocarbon generation and expulsion rates across the source rock interval. When incorporated into basin modeling software, these detailed TOC models would allow for more accurate timing predictions for hydrocarbon generation, potentially reducing exploration risk through improved charge assessments.

Furthermore, the spatial uncertainty quantification provided by all three methods offers valuable input for probabilistic petroleum system modeling. The variance maps could be used to generate multiple equiprobable TOC realizations for Monte Carlo simulations, providing a range of plausible hydrocarbon generation scenarios that better capture subsurface uncertainties. This probabilistic approach is particularly valuable in frontier areas with limited well control, where characterizing the range of possible outcomes is crucial for risk assessment.

It is important to acknowledge the limitations of this study. The use of synthetic TOC data, while necessary given the data constraints in the Santos Basin pre-salt section, introduces uncertainties that would require careful calibration in real-world applications. In production settings, actual geochemical data including hydrogen index, oxygen index, and kerogen type would provide important additional constraints, potentially modifying both the spatial patterns and absolute values of TOC. The relationship between seismic attributes and TOC distribution would also benefit from calibration with measured TOC data, potentially requiring multivariate analysis to identify the optimal combination of seismic attributes for TOC prediction.

Additionally, our model focused primarily on the spatial distribution of present-day TOC without explicitly modeling the transformation of original organic matter through thermal maturation. In comprehensive petroleum system studies, this TOC model would need to be integrated with thermal history models to reconstruct original TOC distributions and calculate transformation ratios through geological time.

Despite these limitations, the workflow presented here provides a robust foundation for TOC modeling in data-constrained environments. The comparative analysis of multiple modeling approaches offers insights into the benefits and limitations of different techniques, enabling informed selection based on specific project requirements and risk tolerance. Future work could extend this methodology in several key directions: incorporating real well TOC data instead of synthetic measurements would significantly enhance model calibration and reliability; expanding the seismic dataset to include multiple images would provide more comprehensive pixel attributes and spatial constraints across the basin; and integrating additional constraints from core analysis, biomarker studies, and basin-scale depositional models. These enhancements would collectively improve the spatial resolution and geological fidelity of TOC predictions, particularly in areas between well control points, while providing a more statistically robust foundation for characterizing lateral heterogeneity. Such developments would further refine our understanding of source rock distribution in this prolific petroleum province, ultimately contributing to more accurate resource assessments and reduced exploration risk.

3.2.6 References

Al-Mudhafar, W.J., 2017. Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms. Journal of Petroleum Exploration and Production Technology 7:1023-1033.

Antonello, L.L., Carrasquilla, A., Pires, D., Lemos, T., 2021. Statistical analysis of petrophysical properties of the Santos Basin presalt carbonate reservoirs. Journal of Petroleum Science and Engineering 201:108395.

Arienti, L.M., Castro, J.C., Silva, C.M.A., 2020. Facies model for the Itapema Formation carbonates (Santos Basin, Brazil): A contribution to understanding Brazilian pre-salt reservoirs. Journal of South American Earth Sciences 104:102831.

Arienti, L.M., Souza, R.S., Viana, S.M., Falcão, L.C., 2018. Microbialites of the Barra Velha Formation, Santos Basin, Brazil. AAPG Bulletin 102(3):457-477.

Armstrong, M., 1998. Basic linear geostatistics. Springer, Berlin.

Baur, F., Di Primio, R., Lampe, C., Littke, R., 2019. Mass balance calculations for different models of hydrocarbon migration in the Gifhorn Trough, northern Germany. Marine and Petroleum Geology 106:310-326.

Bergen, K.J., Johnson, P.A., de Hoop, M.V., Beroza, G.C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science 363(6433):eaau0323.

Berton, F., Vesely, F.F., Pitthan, E., 2020. Controls on the spatial distribution of microbialite reservoirs in the Santos Basin pre-salt carbonates, offshore Brazil. Marine and Petroleum Geology 112:104107.

Bestagini, P., Lipari, V., Tubaro, S., 2017. A machine learning approach to facies classification using well logs. SEG Technical Program Expanded Abstracts 2017:2137-2142.

Bhandari, A., Flemings, P., Polito, P., Cronin, M., Bryant, S., 2015. Anisotropy and Stress Dependence of Permeability in the Barnett Shale. Transport in Porous Media 108:393-411.

Bialik, O., Wang, X., Zhao, S., Waldmann, N., Frank, R., Li, W., 2018. Mg isotope response to dolomitization in hinterland-attached carbonate platforms: Outlook of δ26Mg as a tracer of basin restriction and seawater Mg/Ca ratio. Geochimica et Cosmochimica Acta.

Bizzi, A.L., Schobbenhaus, C., Vidotti, R.M., Gonçalves, J.H., (n.d.). Geology, Tectonics and Mineral Resources of Brazil: Text, Maps and GIS. CPRM - Geologic Service of Brazil.

Bogner, K., Pappenberger, F., Cloke, H., 2011. Technical Note: The normal quantile transformation and its application in a flood forecasting system. Hydrology and Earth System Sciences 16:1085-1094.

Bohacs, K.M., Carroll, A.R., Neal, J.E., Mankiewicz, P.J., 2000. Lake-basin type, source potential, and hydrocarbon character: an integrated sequence-stratigraphic-geochemical framework. In: Gierlowski-Kordesch, E.H., Kelts, K.R. (Eds.), Lake Basins through Space and Time. AAPG Studies in Geology 46:3-34.

Bohacs, K.M., Grabowski, G., Carroll, A.R., 2003. Lake-basin type, source potential, and hydrocarbon character: an integrated-sequence-stratigraphic-geochemical framework. In: Mertz, K.A. (Ed.), Cenozoic Systems of the Rocky Mountain Region. SEPM, Denver, pp. 328-354.

Bohling, G.C., Dubois, M.K., 2003. An integrated application of neural network and Markov chain techniques to prediction of lithofacies from well logs. Kansas Geological Survey Open-file Report 2003-50.

Boisvert, J.B., Deutsch, C.V., 2011. Programs for kriging and sequential Gaussian simulation with locally varying anisotropy using non-Euclidean distances. Computers & Geosciences 37:495-510.

Boisvert, J.B., Manchuk, J.G., Deutsch, C.V., 2009. Kriging in the presence of locally varying anisotropy using non-Euclidean distances. Mathematical Geosciences 41:585-601.

Bostanabad, R., Kearney, T., Tao, S., Apley, D., Chen, W., 2018. Leveraging the nugget parameter for efficient Gaussian process modeling. International Journal for Numerical Methods in Engineering 114:501-516.

Buckley, J.P., Bosence, D., Elders, C., 2015. Tectonic setting and stratigraphic architecture of an Early Cretaceous lacustrine carbonate platform, Sugar Loaf High, Santos Basin, Brazil. Geological Society, London, Special Publications 418(1):175-191.

Caers, J., 2011. Modeling uncertainty in the earth sciences. Wiley-Blackwell, Chichester.

Caers, J., Zhang, T., 2004. Multiple-point geostatistics: a quantitative vehicle for integrating geologic analogs into multiple reservoir models. AAPG Memoir 80:383-394.

Caetano, H., Marques, F., Costa, A., Pinheiro, L., Soares, A., 2017. Geostatistical seismic inversion for frontier exploration. Interpretation 5:SL43-SL52.

Carcione, J.M., Helle, H.B., Avseth, P., 2015. Source-rock seismic characterization. In: Bjørlykke, K. (Ed.), Petroleum Geoscience: From Sedimentary Environments to Rock Physics, 2nd edn. Springer, Berlin, pp. 423-450.

Carroll, A., Bohacs, K., 2001. Lake-type controls on petroleum source rock potential in nonmarine basins. AAPG Bulletin 85:1033-1053.

Carroll, A.R., Bohacs, K.M., 1999. Stratigraphic classification of ancient lakes: Balancing tectonic and climatic controls. Geology 27(2):99–102.

Chang, H.K., Assine, M.L., Corrêa, F.S., Tinen, J.S., Vidal, A.C., Koike, L., 2008. Sistemas petrolíferos e modelos de acumulação de hidrocarbonetos na bacia de Santos. Brazilian Journal of Geology 38(2):29–46.

Chang, H.K., Kowsmann, R.O., Figueiredo, A.M.F., 1990. Novos conceitos sobre o desenvolvimento das bacias marginais do leste brasileiro. Origem e evolução de bacias sedimentares. PETROBRAS, Rio de Janeiro, 269–289.

Cheadle, C., Vawter, M., Freed, W., & Becker, K. (2003). Analysis of microarray data using Z score transformation.. The Journal of molecular diagnostics : JMD, 5 2, 73-81 .

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, pp. 785–794.

Chilès, J.P., Delfiner, P., 2012. Geostatistics: Modeling Spatial Uncertainty, 2nd edn. Wiley, New York.

Chitale, V., Alabi, G., Kasten, R., Tinnin, B., Prasad, T., Sil, S., 2015. Reservoir characterization challenges due to multiscale spatial heterogeneity in the Presalt carbonate sag formation, North Campos Basin, Brazil. Interpretation 3(1):SV33-SV45.

Chopra, S., Marfurt, K.J., 2007. Seismic attributes for prospect identification and reservoir characterization. Society of Exploration Geophysicists, Tulsa.

Chopra, S., Sharma, R.K., Keay, J., Marfurt, K.J., 2018. Shale gas reservoir characterization workflows. In: Unconventional Resources Technology Conference, Houston, Texas, URTeC-2903038.

Comunian, A., Giudici, M., 2021. Improving the robustness of the comparison model method for the identification of hydraulic transmissivities. Computers & Geosciences 149:104705.

Contreras, J., Zühlke, R., Bowman, S., Bechstädt, T., 2010. Seismic stratigraphy and subsidence analysis of the southern Brazilian margin (Campos, Santos and Pelotas basins). Marine and Petroleum Geology 27:1952-1980.

Corbett, P.W.M., Geiger, S., Borges, L., Garayev, M., Gonzalez Camejo, J.G., Valdez, C., 2012. Limitations in numerical well test modelling of fractured carbonate rocks. In: Spence, G.H., Redfern, J., Aguilera, R., Bevan, T.G., Cosgrove, J.W., Couples, G.D., Daniel, J.M. (eds) Advances in the study of fractured reservoirs. Geological Society London Special Publications 374:147-161.

Creaney, S., Passey, Q., 1993. Recurring Patterns of Total Organic Carbon and Source Rock Quality within a Sequence Stratigraphic Framework. AAPG Bulletin 77:386-401.

De Mahiques, M., Schattner, U., Lazar, M., Sumida, P., Souza, L., 2017. An extensive pockmark field on the upper Atlantic margin of Southeast Brazil: spatial analysis and its relationship with salt diapirism. Heliyon 3.

De Oliveira Nardi Leite, C., De Assis Silva, C., Ros, L., 2020. Depositional and diagenetic processes in the pre-salt rift section of a Santos Basin area, SE Brazil. Journal of Sedimentary Research.

de la Varga, M., Schaaf, A., Wellmann, F., 2019. GemPy 1.0: open-source stochastic geological modeling and inversion. Geoscientific Model Development 12:1–32.

Delbari, M., Afrasiab, P., Loiskandl, W., 2009. Using sequential Gaussian simulation to assess the field-scale spatial uncertainty of soil water content. Catena 79:163-169.

Deutsch, C.V., 2002. Geostatistical reservoir modeling. Oxford University Press, New York.

Deutsch, C.V., Journel, A.G., 1998. GSLIB: Geostatistical Software Library and User's Guide, 2nd edn. Oxford University Press, New York.

Deutsch, C.V., Pyrcz, M.J., 2014. Geostatistical reservoir modeling, 2nd edn. Oxford University Press, Oxford.

Di, H., Alfarraj, M., AlRegib, G., 2016. 3D curvature analysis of seismic waveform and its interpretational implications. In: 2016 SEG International Exposition and Annual Meeting. SEG.

Dorn, G.A., Shimeld, J.W., 2020. Seismic interpretation of crustal structure: A machine learning approach. Interpretation 8(2):SM115-SM126.

Doyen, P.M., 2007. Seismic reservoir characterization: an earth modelling perspective. EAGE Publications, Houten.

Dubrule, O., 2003. Geostatistics for seismic data integration in earth models. Society of Exploration Geophysicists and European Association of Geoscientists and Engineers, Tulsa.

Dunn, P., Smyth, G., 1996. Randomized Quantile Residuals. Journal of Computational and Graphical Statistics 5:236-244.

Dupont, E., Zhang, T., Tilke, P., Liang, L., Bailey, W., 2018. Generating realistic geology conditioned on physical measurements with generative adversarial networks. arXiv preprint arXiv:1802.03065.

Eltom, H., Saraih, N., Esteva, O., Kusuma, L., Ahmed, S., Yassin, M., 2020. Three-Dimensional Modeling and Fluid Flow Simulation for the Quantitative Description of Permeability Anisotropy in Tidal Flat Carbonate. Energies.

Emery, X., 2010. Iterative algorithms for fitting a linear model of coregionalization. Computers & Geosciences 36:1150-1160.

Emmel, B., Baskoro, A., De Jager, G., Grøver, A., Roli, O., 2018. The influence of paleo-bathymetry on total organic carbon distribution tested in the Cretaceous Hammerfest Basin, Barents Sea. Marine and Petroleum Geology.

Esmaeilzadeh, S., Salehi, S., Hetz, G., 2020. Spatiotemporal geostatistical modeling of reservoir data with random effects. Journal of Petroleum Science and Engineering 190:107032.

Farias, F., Szatmari, P., Bahniuk, A., França, A., 2019. Evaporitic carbonates in the pre-salt of Santos Basin – Genesis and tectonic implications. Marine and Petroleum Geology.

Faugères, J., Mézerais, M., Stow, D., 1993. Contourite drift types and their distribution in the North and South Atlantic Ocean basins. Sedimentary Geology 82:189-203.

Ford, M., Vergés, J., 2020. Evolution of a salt-rich transtensional rifted margin, eastern North Pyrenees, France. Journal of the Geological Society 178.

Freitas, J.T.R., 2006. Ciclos deposionais evaporíticos da bacia de Santos: uma análise cicloestratigráfica a partir de dados de 2 poços e de traços de sísmica. Master's thesis, Universidade Federal do Rio Grande do Sul.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 29:1189-1232.

Frykman, P., 2001. Spatial variability in petrophysical properties in Upper Maastrichtian chalk outcrops at Stevns Klint, Denmark. Marine and Petroleum Geology 18:1041-1062.

Gamboa, L.A.P., Machado, M.A.P., Silveira, D.P., Freitas, J.T.R., Silva, S.R.P., Mohriak, W., Szatmari, P., Anjos, S., 2008. Evaporitos estratificados no Atlântico Sul: interpretação sísmica e controle tectono-estratigráfico na bacia de Santos. Sal: Geologia e Tectônica, Exemplos nas Bacias Brasileiras, 340–359.

Garcia, S., Letouzey, J., Rudkiewicz, J., Filho, A., Lamotte, D., 2012. Structural modeling based on sequential restoration of gravitational salt deformation in the Santos Basin (Brazil). Marine and Petroleum Geology 35:337-353.

Gomes, J.P., Bunevich, R.B., Tedeschi, L.R., Tucker, M.E., Whitaker, F.F., 2020. Facies classification and patterns of lacustrine carbonate deposition of the Barra Velha Formation, Santos Basin, Brazilian Pre-salt. Marine and Petroleum Geology 113:104176.

Gomes, P.O., Kilsdonk, B., Minken, J., Grow, T., Barragan, R., 2009. The outer high of the Santos Basin, Southern São Paulo Plateau, Brazil: pre-salt exploration outbreak, paleogeographic setting, and evolution of the syn-rift structures. AAPG International Conference and Exhibition, Cape Town, South Africa.

Gonçalves, F.T.T., Araújo, C.V., Penteado, H.L.B., Gilberto, A., Frota, E.S.T., Soldan, A.L., 2015. Organic geochemistry and paleoenvironmental assessment of the pre-salt oil-source system of Santos Basin, Brazil. AAPG Annual Convention and Exhibition, Denver, Colorado.

Gonçalves, F.T.T., Penteado, H.L.B., Dores, F.B., Caputo, M.V., Lima, F.D., 2020. Source rock potential of the Itapema Formation in the Santos Basin, Brazil: Insights from geochemical analysis and basin modeling. Marine and Petroleum Geology 122:104646.

Gonzalez, R.C., Woods, R.E., 2018. Digital image processing, 4th edn. Pearson Education, London.

Goovaerts, P., 1997. Geostatistics for natural resources evaluation. Oxford University Press, New York.

Gringarten, E., Deutsch, C.V., 2001. Teacher's aide: variogram interpretation and modeling. Mathematical Geology 33:507-534.

Groenigen, J., 2000. The influence of variogram parameters on optimal sampling schemes for mapping by kriging. Geoderma 97:223-236.

Groshong, R.H., 2006. 3-D structural geology: A practical guide to quantitative surface and subsurface map interpretation. Springer, Berlin.

Gust, D., Biddle, K., Phelps, D., Uliana, M., 1985. Associated middle to late Jurassic volcanism and extension in southern South America. Tectonophysics 116:223-253.

Hami-Eddine, K., Klein, P., Richard, L., Furniss, A., Lallier, F., 2015. Using well log-derived properties to build a robust stratigraphic model. The Leading Edge 34(1):20-22.

Hantschel, T., Kauerauf, A.I., 2009. Fundamentals of Basin and Petroleum Systems Modeling. Springer, Berlin.

Hartmann, K., Krois, J., Waske, B., 2018. E-learning project SOGA: statistics and geospatial data analysis. Department of Earth Sciences, Freie Universitaet Berlin.

Heine, C., Zoethout, J., Muller, R., 2013. Kinematics of the South Atlantic rift. Solid Earth 4:215-253.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6:e5518.

Heße, F., Prykhodko, V., Schlüter, S., Attinger, S., 2014. Generating random fields with a truncated power-law variogram: a comparison of several numerical methods. Environmental Modelling & Software 55:32–48.

Hinnov, L.A., 2013. Cyclostratigraphy and its revolutionizing applications in the earth and planetary sciences. Geological Society of America Bulletin 125:1703–1734.

Hood, A., Gutjahr, C.C., Heacock, R.L., 1975. Organic metamorphism and the generation of petroleum. AAPG Bulletin 59(6):986-996.

Hubral, P., Schleicher, J., Tygel, M., 1996. A unified approach to 3-D seismic reflection imaging, Part I: Basic concepts. Geophysics 61:742-758.

Hunt, J.M., 1996. Petroleum Geochemistry and Geology, 2nd edn. W.H. Freeman, New York.

Hutton, A.C., 1987. Petrographic classification of oil shales. International Journal of Coal Geology 8:203-231.

Isaaks, E.H., Srivastava, R.M., 1989. An introduction to applied geostatistics. Oxford University Press, New York.

Jackson, C., Jackson, M., Hudec, M., Rodriguez, C., 2015. Enigmatic structures within salt walls of the Santos Basin—Part 1: Geometry and kinematics from 3D seismic reflection and well data. Journal of Structural Geology 75:135-162.

Jones, D.M., Head, I.M., Gray, N.D., Adams, J.J., Rowan, A.K., Aitken, C.M., Bennett, B., Huang, H., Brown, A., Bowler, B.F.J., Oldenburg, T., Erdmann, M., Larter, S.R., 2015. Crude-oil biodegradation via methanogenesis in subsurface petroleum reservoirs. Nature 451:176-180.

Journel, A.G., 2002. Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses. Mathematical Geology 34:573-596.

Journel, A.G., Alabert, F., 1988. Non-Gaussian data expansion in the earth sciences. Terra Nova 1:123-134.

Journel, A.G., Huijbregts, C.J., 1978. Mining geostatistics. Academic Press, London.

Kaczmarek, S.E., Fullmer, S.M., Hasiuk, F.J., 2017. A universal classification scheme for carbonate porosity. Journal of Sedimentary Research 87:543-565.

Kanan, C., Cottrell, G.W., 2012. Color-to-grayscale: Does the method matter in image recognition? PloS one 7(1):e29740.

Kapageridis, I., 2015. Variable lag variography using k-means clustering. Computers & Geosciences 85:49-63.

Karimpouli, S., Tahmasebi, P., Ramandi, H.L., 2020. A deep learning perspective of 3D rock typing. Computers & Geosciences 137:104406.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., Kumar, V., 2017. Machine learning for the geosciences: Challenges and opportunities. IEEE Transactions on Knowledge and Data Engineering 31(8):1544-1554.

Katz, B.J., 1995. A survey of rift basin source rocks. Geological Society, London, Special Publications 80(1):213–240.

Katz, B.J., Lin, F., 2014. Lacustrine basin unconventional resource plays: Key differences. Marine and Petroleum Geology 56:255-265.

Keighley, D., Flint, S., Howell, J., Moscariello, A., 2003. Sequence stratigraphy in lacustrine basins: a model for part of the Green River Formation (Eocene), southwest Uinta Basin, Utah, USA. Journal of Sedimentary Research 73:987-1006.

Kelts, K., 1988. Environments of deposition of lacustrine petroleum source rocks: an introduction. Geological Society, London, Special Publications 40(1):3–26.

Kelts, K., 1989. Environments of deposition of lacustrine petroleum source rocks: an introduction. Geological Society, London, Special Publications 40(1):3–26.

Kirkwood, C., Economou, T., Odbert, H., Pugeault, N., 2022. Bayesian deep learning for spatial interpolation in the presence of auxiliary information. Mathematical Geosciences 54:507-531.

Kramer, P.R., Kurbanmuradov, O., Sabelfeld, K., 2007. Comparative analysis of multiscale Gaussian random field simulation algorithms. Journal of Computational Physics 226(1):897-924.

Kuchinskiy, V., Gechter, D., Ratcliffe, K., 2013. Integrating high resolution chemostratigraphy and facies analysis to identify systems tracts and sequence boundaries. International Petroleum Technology Conference, Beijing, China, IPTC-16797-MS.

Kupfersberger, H., Deutsch, C.V., 1999. Methodology for integrating analog geologic data in 3-D variogram modeling. AAPG Bulletin 83:1262-1278.

Lantuéjoul, C., 2013. Geostatistical simulation: models and algorithms. Springer Science & Business Media, Berlin.

Leuangthong, O., McLennan, J.A., Deutsch, C.V., 2004. Minimum acceptance criteria for geostatistical realizations. Natural Resources Research 13:131-141.

Li, X., Al-Tous, H., Hajri, S., Tirkkonen, O., 2024. Enhanced Weighted K-Nearest Neighbor Positioning. 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring), 1-6.

Lima, B., Ros, L., 2019. Deposition, diagenetic and hydrothermal processes in the Aptian Pre-Salt lacustrine carbonate reservoirs of the northern Campos Basin, offshore Brazil. Sedimentary Geology.

Liu, F.T., Ting, K.M., & Zhou, Z.H. (2012). Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data, 6(1), 1-39.

Liu, K., Liu, J., Huang, X., 2021. Coupled stratigraphic and petroleum system modeling: Examples from the Ordos Basin, China. AAPG Bulletin 105:1-28.

Liu, Q., Zhu, D., Jin, Z., Liu, C., Zhang, D., He, Z., 2017. Coupled thermal-hydraulic-mechanical modeling of hot dry rock reservoir stimulation and EGS heat extraction. In: Proceedings of the 42nd Workshop on Geothermal Reservoir Engineering, Stanford University, SGP-TR-212.

Løseth, H., Wensaas, L., Gading, M., Duffaut, K., Springer, M., 2011. Can hydrocarbon source rocks be identified on seismic data? Geology 39:1167-1170.

Lourenço, J., Menezes, P.T.L., Barbosa, V.C.F., 2014. Connecting onshore-offshore Campos Basin structures: Interpretation of high-resolution airborne magnetic data. Interpretation 2(4):SJ35–SJ45.

Lovecchio, J., Abdelmalak, M., Planke, S., Silio, O., Rohais, S., Arismendi, S., Vera, E., Kulhanek, D., Bolatti, N., Ramos, V., 2024. Mesozoic Rifting in SW Gondwana and Breakup of the Southern South Atlantic Ocean. Geological Society, London, Special Publications.

Lovecchio, J., Rohais, S., Joseph, P., Bolatti, N., Ramos, V., 2020. Mesozoic rifting evolution of SW Gondwana: A poly-phased, subduction-related, extensional history responsible for basin formation along the Argentinean Atlantic margin. Earth-Science Reviews.

Mälicke, M., 2021. SciKit-GStat 1.0: a SciPy-flavored geostatistical variogram estimation toolbox written in Python. Geoscientific Model Development 15(6):2505–2532.

Manchuk, J.G., Deutsch, C.V., 2012. A flexible sequential Gaussian simulation program: USGSIM. Computers & Geosciences 41:208-216.

Mancini, E., Obid, J., Badalí, M., Liu, K., Parcell, W., 2008. Sequence-stratigraphic analysis of Jurassic and Cretaceous strata and petroleum exploration in the central and eastern Gulf coastal plain, United States. AAPG Bulletin 92:1655-1686.

Marchant, B., Lark, R., 2007. Robust estimation of the variogram by residual maximum likelihood. Geoderma 140:62-72.

Mariethoz, G., Caers, J., 2014. Multiple-point geostatistics: stochastic modeling with training images. John Wiley & Sons, Chichester.

Mariethoz, G., Renard, P., Straubhaar, J., 2010. The Direct Sampling method to perform multiple-point geostatistical simulations. Water Resources Research 46:W11536.

Meinshausen, N., 2006. Quantile regression forests. Journal of Machine Learning Research 7:983-999.

Mello, M.R., Bender, A.A., Vieira, R.A.B., 2022. Santos Basin, Brazil: The New Giant in town - pre-salt discoveries and petroleum systems. AAPG Bulletin 106(2):281-315.

Mello, Maxwell, J., 1990. Organic Geochemical and Biological Marker Characterization of Source Rocks and Oils Derived from Lacustrine Environments in the Brazilian Continental Margin: Chapter 5. 50:77-99.

Micallef, A., Person, M., Haroon, A., Weymer, B.A., Jegen, M., Schwalenberg, K., et al., 2020. 3D characterisation and quantification of an offshore freshened groundwater system in the Canterbury Bight. Nature Communications 11(1):1372.

Miguel A. Cuba, O. Leuangthong and J. Ortiz. "Detecting and quantifying sources of non-stationarity via experimental semivariogram modeling." Stochastic Environmental Research and Risk Assessment, 26 (2012): 247-260. https://doi.org/10.1007/s00477-011-0501-9.

Milani, E.J., Brandão, J.L., Zalán, P.V., Gamboa, L.A.P., 2000. Petróleo na margem continental Brasileira: Geologia, exploração, resultados e perspectivas. Revista Brasileira de Geofisica 18(3):351–396.

Milani, E.J., Rangel, H.D., Bueno, G.V., Stica, J.M., Winter, W.R., Caixeta, J.M., Neto, O.P., 2007. Bacias sedimentares brasileiras: cartas estratigráficas. Boletim de Geociências da PETROBRAS 15(2):183–205.

Mishra, S., Datta-Gupta, A., 2017. Applied Statistical Modeling and Data Analytics: A Practical Guide for the Petroleum Geosciences. Elsevier, Amsterdam.

Mohriak, W.U., Mello, M.R., Dewey, J.F., Maxwell, J.R., 1990. Petroleum geology of the Campos Basin, offshore Brazil. Geological Society, London, Special Publications 50(1):119–141.

Moreira, J.L.P., Madeira, C.V., Gil, J.A., Machado, M.A.P., 2007. Bacia de Santos. Boletim de Geociências da Petrobras 15(2):531–549.

Moulin, M., Aslanian, D., Unternehr, P., 2005. A new starting point for the South and Equatorial Atlantic Ocean. Earth-Science Reviews 98(1-2):1-37.

Müller, S., Schüler, L., 2021. GeoStat-Framework/GSTools: Pastas integration. Zenodo.

Murphy, J., O'Brien, L., 1977. The correlation of peak ground acceleration amplitude with seismic intensity and other physical parameters. Bulletin of the Seismological Society of America.

Neto, A., Mota, B., Belém, A., Albuquerque, A., Capilla, R., 2016. Seismic peak amplitude as a predictor of TOC content in shallow marine sediments. Geo-Marine Letters 36:395-403.

Neves, I., Lupinacci, W., Ferreira, D., Zambrini, J., Oliveira, L., Azul, M., Ferrari, A., Gambôa, L., 2019. Presalt reservoirs of the Santos Basin: Cyclicity, electrofacies, and tectonic-sedimentary evolution. Interpretation.

Nussbaumer, R., Mariethoz, G., Gravey, M., Gloaguen, E., Holliger, K., 2018. Accelerating sequential Gaussian simulation with a constant path. Computers & Geosciences 112:121-132.

Olea, R.A., 2018. A practical primer on geostatistics. U.S. Geological Survey, Open-File Report 2009-1103.

Olsen, P.E., 1990. Tectonic, climatic, and biotic modulation of lacustrine ecosystems–examples from Newark Supergroup of eastern North America. In Lacustrine Basin Exploration case Stud. Mod. Analog., 209–224.

Passey, Q.R., Bohacs, K.M., Esch, W.L., Klimentidis, R., Sinha, S., 2010. From oil-prone source rock to gas-producing shale reservoir—geologic and petrophysical characterization of unconventional shale-gas reservoirs. SPE International Oil and Gas Conference and Exhibition, Beijing, China, SPE-131350-MS.

Pereira, M.J., Feijó, F.J., 1994. Bacia de Santos. Boletim de Geociências da PETROBRAS 8(1):219–234.

Peters, K.E., Hosford Scheirer, A., Magoon, L.B., 2012. Basin and petroleum system modeling. AAPG Bulletin 96:2147-2148.

Peters, K.E., Walters, C.C., Moldowan, J.M., 2005. The Biomarker Guide: Volume 2, Biomarkers and Isotopes in Petroleum Exploration and Earth History. Cambridge University Press, Cambridge.

Piper, D.Z., Calvert, S.E., 2009. A marine biogeochemical perspective on black shale deposition. Earth-Science Reviews 95:63-96.

Pyrcz, M.J., Deutsch, C.V., 2014. Geostatistical Reservoir Modeling, 2nd edn. Oxford University Press, Oxford.

Riccomini, C., Sant'anna, L.G., Tassinari, C.C.G., 2012. Pré-sal: geologia e exploração. Revista USP (95):33–42.

Ringrose, P., Bentley, M., 2015. Reservoir model design: a practitioner's guide. Springer, Dordrecht.

Robertson, R., Mueller, U., & Bloom, L. (2006). Direct sequential simulation with histogram reproduction: A comparison of algorithms. Comput. Geosci., 32, 382-395. https://doi.org/10.1016/j.cageo.2005.07.002.

Rodriguez, K., Hodgson, N., Intawong, A., 2018. The complex puzzle of the pre-salt Santos Basin. GEO ExPro 15(2):46-50.

Rodrigues, S., Hernández-Molina, F., Kirby, A., 2021. A Late Cretaceous mixed (turbidite-contourite) system along the Argentine Margin: Paleoceanographic and conceptual implications. Marine and Petroleum Geology 123:104768.

Romero-Sarmiento, M.F., Ducros, M., Carpentier, B., Lorant, F., Cacas, M.C., Pegaz-Fiornet, S., Wolf, S., Rohais, S., Moretti, I., 2013. Quantitative evaluation of TOC, organic porosity and gas retention distribution in a gas shale play using petroleum system modeling: application to the Mississippian Barnett Shale. Marine and Petroleum Geology 45:315-330.

Santos, T.P., Lisboa, L.P., Carreira, V., Venancio, I.M., Bernardes, M.C., Belem, A.L., Díaz, R.A., Moreira, M., Lopes, A.A.O., Santos, T.L., Souza, I.V., Spigolon, A.L.D., Albuquerque, A.L.S., 2023. Orbitally-driven palaeogene to neogene deposition in the western south Atlantic (Espírito Santo basin) and its correlation with global sea level. Sedimentology 70:1–27.

Scheidt, C., Li, L., Caers, J., 2018. Quantifying uncertainty in subsurface systems. Wiley, Hoboken.

Schlumberger, 2010. Petrel simulation software manuals. Schlumberger, Houston.

Schubert, E., Sander, J., Ester, M., Kriegel, H.P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems, 42(3), 1-21.

Scotese, C.R., 2013. Paleomap Project: Earth History, Paleogeography, and Paleoclimate. Evanston, IL. Retrieved from http://www.scotese.com

Sinan, S., Glover, P., Lorinczi, P., 2020. Modelling the Impact of Anisotropy on Hydrocarbon Production in Heterogeneous Reservoirs. Transport in Porous Media 133:413-436.

Slatt, R.M., Rodriguez, N.D., 2012. Comparative sequence stratigraphy and organic geochemistry of gas shales: Commonality or coincidence? Journal of Natural Gas Science and Engineering 8:68-84.

Sun, S., Mao, W., Ouyang, W., Du, M., Yang, M., 2022. Amplitude-Preserving Imaging Condition for Scattering-Based RTM in Acoustic VTI Media. IEEE Geoscience and Remote Sensing Letters 19:1-5.

Talbot, M.R., 1988. The origins of lacustrine oil source rocks: evidence from the lakes of tropical Africa. Geological Society, London, Special Publications 40(1):29–43.

Tänavsuu-Milkeviciene, K., Sarg, J.F., 2012. Evolution of an organic-rich lake basin – stratigraphy, climate and tectonics: Piceance Creek basin, Eocene Green River Formation. Sedimentology 59:1735-1768.

Tearpock, D.J., Bischke, R.E., 2003. Applied subsurface geological mapping with structural methods, 2nd edn. Prentice Hall, Upper Saddle River.

Tegelaar, E.W., Noble, R.A., 1994. Kinetics of hydrocarbon generation as a function of the molecular structure of kerogen as revealed by pyrolysis-gas chromatography. Organic Geochemistry 22:543-574.

Theano Development Team, 2016. Theano: A Python framework for fast computation of mathematical expressions. arXiv preprint arXiv:1605.02688.

Thomas, A.T., Reiche, S., Riedel, M., Clauser, C., 2019. The fate of submarine fresh groundwater reservoirs at the New Jersey shelf, USA. Hydrogeology Journal 27:2673-2685.

Thompson, D., Stilwell, J., Hall, M., 2015. Lacustrine carbonate reservoirs from Early Cretaceous rift lakes of Western Gondwana: Pre-Salt coquinas of Brazil and West Africa. Gondwana Research 28:26-51.

Trindade, L.A.F., Dias, J.L., Mello, M.R., 1995. Sedimentological and geochemical characterization of the Lagoa Feia Formation, rift phase of the Campos Basin, Brazil. In Petroleum source rocks, 149–165. Springer.

Tucker, W., 1990a. Carbonate Depositional Systems I: Marine Shallow-Water and Lacustrine Carbonates. In Sedimentary Petrology: An Introduction to the Origin of Sedimentary Rocks (Chapter 4, pp. 101–227). John Wiley & Sons, Ltd.

Tucker, W., 1990b. Carbonate Depositional Systems II: Deeper-Water Facies of Pelagic and Resedimented Limestones. In Sedimentary Petrology: An Introduction to the Origin of Sedimentary Rocks (Chapter 5, pp. 228–283). John Wiley & Sons, Ltd.

Vail, P.R., Mitchum, R.M., Jr., Thompson, S., 1977a. Seismic Stratigraphy and Global Changes of Sea Level, Part 3: Relative Changes of Sea Level from Coastal Onlap. In Seismic Stratigraphy — Applications to Hydrocarbon Exploration. American Association of Petroleum Geologists.

Vail, P.R., Mitchum, R.M., Jr., Thompson, S., III, 1977b. Seismic stratigraphy and global changes of sea level: Part 3. Relative changes of sea level from coastal onlap: section 2. Application of seismic reflection configuration to stratigraphic interpretation.

Valença, L., Neumann, V., Mabesoone, J., 2003. An overview on Callovian-Cenomanian intracratonic basins of Northeast Brazil : onshore stratrigraphic record of the opening of the southern Atlantic. Geologica Acta 1:261-275.

Van Den Boogaart, K., Mueller, U., & Tolosana-Delgado, R. (2017). An Affine Equivariant Multivariate Normal Score Transform for Compositional Data. Mathematical Geosciences, 49, 231-251. https://doi.org/10.1007/s11004-016-9645-y.

Verma, S., Zhao, T., Marfurt, K.J., Devegowda, D., 2016. Estimation of total organic carbon and brittleness volume. Interpretation 4:T373-T385.

Vieira S., J. R. P. D. Carvalho, M. Ceddia and A. Gonzalez. "Detrending non stationary data for geostatistical applications." Bragantia, 69 (2010): 1-8. https://doi.org/10.1590/S0006-87052010000500002.

Walker, J.D., Geissman, J.W., Bowring, S.A., Babcock, L.E., 2013. The Geological Society of America Geologic Time Scale. GSA Bulletin 125(3-4):259-272.

Wang, G., Li, P., Wang, J., Hao, F., Wang, H., Zou, H., 2015. Source rock characteristics and resource potential of the Lower Cambrian black shales in the southeast Yangtze Block, China. Marine and Petroleum Geology 70:30-44.

Wang, K., Ye, S., Gao, P., Yao, X., Zhao, Z., 2022. Optimization of Numerical Methods for Transforming UTM Plane Coordinates to Lambert Plane Coordinates. Remote Sensing 14:2056.

Wang, Y., Li, M., 2016. Reservoir characterization using multiscale information from seismic inversion, well logging, and core analysis: a case study. Open Geosciences 8:21-32.

Wang, Z., Di, H., Shafiq, M.A., Alregib, G., Deriche, M., 2021. Seismic attribute selection for unsupervised machine learning. Geophysics 86(2):1-62.

Waples, D.W., 2000. The kinetics of in-reservoir oil destruction and gas formation: constraints from experimental and empirical data, and from thermodynamics. Organic Geochemistry 31:553-575.

Watson, M., 1986. Univariate detrending methods with stochastic trends. Journal of Monetary Economics 18:49-75.

Webster, R., McBratney, A., 1989. On the Akaike Information Criterion for choosing models for variograms of soil properties. European Journal of Soil Science 40:493-496.

Webster, R., Oliver, M.A., 2007. Geostatistics for environmental scientists, 2nd edn. John Wiley & Sons, Chichester.

Wen, Z., Jiang, S., Song, C., Wang, Z., He, Z., 2019. Basin evolution, configuration styles, and hydrocarbon accumulation of the South Atlantic conjugate margins. Energy Exploration & Exploitation 37:1008-992.

Williams, G.E., 1993. History of the Earth's obliquity. Earth-Science Reviews 34(1):1–45.

Wright, V.P., Barnett, A.J., 2015. An abiotic model for the development of textures in some South Atlantic early Cretaceous lacustrine carbonates. Geological Society, London, Special Publications 418:209-219.

Wu, Z., Huang, N., Long, S., Peng, C., 2007. On the trend, detrending, and variability of nonlinear and nonstationary time series. Proceedings of the National Academy of Sciences 104:14889-14894.

Zalán, P.V., 2007. Evolução Fanerozóica das Bacias Sedimentares Brasileiras. In Geologia da Plataforma Sul-Americana (Chapter 23, pp. 595–613). Petrobras, Rio de Janeiro.

Zech, A., Dietrich, P., Attinger, S., Teutsch, G., 2021. A field evidence model: how to predict transport in heterogeneous aquifers at low investigation level. Hydrology and Earth System Sciences 25(1):1-15.

Zhang, H., Chen, J., Zeng, H., 2019. Applying machine learning for reservoir quality prediction: a case study of Lower Cretaceous sandstones in Linxing area, eastern Ordos Basin, China. Journal of Petroleum Science and Engineering 176:148-157.

Zhang, J., Lay, T., Zaslow, J., & Walter, W. (2002). Source Effects on Regional Seismic Discriminant Measurements. Bulletin of the Seismological Society of America, 92, 2926-2945. https://doi.org/10.1785/0120010242.

Zhao, P., H., Rasouli, V., Liu, W., Cai, J., Huang, Z., 2017. An improved model for estimating the TOC in shale formations. Marine and Petroleum Geology 83:174-183.

Zhao, T., Ramachandran, K., Marfurt, K.J., Nissen, S., 2015. Interval prediction of TOC and brittleness logs in shale plays. In: SEG Technical Program Expanded Abstracts, pp. 2884-2888.

Zuo, G., Wang, H., Fan, G., Zhang, J., Zhang, Y., Wang, C., Yang, L., Ding, L., Pang, X., Zuo, Y., 2022. Geochemical Characteristics and Distribution of the Subsalt Source Rocks in the Santos Basin, Brazil. ACS Omega 7:25715-25725.

4. General Conclusions

This research developed innovative approaches for TOC prediction and spatial modeling in the Santos Basin pre-salt section, yielding three key contributions:

First, machine learning algorithms significantly outperformed traditional TOC estimation methods, with GBDT showing superior performance by effectively capturing complex relationships between well-logs and TOC content. Model performance improved substantially with more homogeneous datasets, with a reduction from five to three wells resulting in 59.39% lower RMSE and 53.73% more data within the acceptable error margin.

Second, our integrated geostatistical workflow successfully created 3D TOC models that honor both synthetic well data and seismic-derived spatial patterns. Comparing modeling approaches revealed distinct advantages: Kriging produced smooth distributions highlighting broad trends, RFG captured heterogeneity representative of lacustrine source rocks, and XGBoost provided intermediate results with distinctive spatial patterns.

Third, the integration of seismic attributes demonstrated how lateral continuity patterns can be constrained in data-limited scenarios, though our synthetic dataset's anisotropy ratio (1:1.8 horizontal-to-vertical) served primarily to validate the methodology rather than characterize actual depositional patterns.

Several limitations were identified throughout this research. Data imbalance in the machine learning component, with 98.77% of samples representing TOC values below 3%, reduced model sensitivity to higher TOC values. The use of synthetic TOC data for geostatistical modeling, while necessary given data constraints, introduces uncertainties that would require calibration in real-world applications. Additionally, the pseudo-3D approach for structural modeling introduces some uncertainty in the crossline dimension, representing a pragmatic but imperfect solution given the limited availability of closely-spaced 2D seismic data.

These findings enhance petroleum system modeling capabilities by improving TOC prediction accuracy from well logs and providing more realistic 3D TOC distributions for basin modeling applications. Future work should address limitations by: (1) expanding well datasets, particularly for high-TOC ranges; (2) incorporating real TOC data; (3) utilizing more comprehensive seismic surveys; (4) implementing advanced techniques to address class imbalance; and (5) integrating additional geophysical parameters beyond conventional logs.

The methodologies developed here provide templates applicable to other basins and geological settings, contributing to the broader field of quantitative petroleum geoscience.

5. References

Bialik, O., Wang, X., Zhao, S., Waldmann, N., Frank, R., & Li, W. (2018). Mg isotope response to dolomitization in hinterland-attached carbonate platforms: Outlook of δ26Mg as a tracer of basin restriction and seawater Mg/Ca ratio. Geochimica et Cosmochimica Acta, 235, 189-207.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. Nature Methods, 15(4), 233–234.

Caers, J. (2011). Modeling uncertainty in the earth sciences. Wiley-Blackwell, Chichester.

Carroll, A.R., & Bohacs, K.M. (1999). Stratigraphic classification of ancient lakes: Balancing tectonic and climatic controls. Geology, 27(2), 99–102.

Carroll, A., & Bohacs, K. (2001). Lake-type controls on petroleum source rock potential in nonmarine basins. AAPG Bulletin, 85, 1033-1053.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.

de la Varga, M., Schaaf, A., & Wellmann, F. (2019). GemPy 1.0: open-source stochastic geological modeling and inversion. Geoscientific Model Development, 12, 1–32.

Deutsch, C.V., & Pyrcz, M.J. (2014). Geostatistical reservoir modeling (2nd ed.). Oxford University Press.

Ester, M., Kriegel, H.P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp. 226-231). AAAI Press.

Fernandes, E. (2017). Bacia de Santos: Sumário Geológico e área em oferta. ANP – Agência Nacional de Petróleo, Gás Natural e Biocombustíveis – Seminário Técnico.

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Friedman, J.H. (2002). Stochastic gradient boosting. Computational Statistics and Data Analysis, 38(4), 367–378.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Hood, A., Gutjahr, C.C.M., & Heacock, R.L. (1975). Organic metamorphism and the generation of petroleum. AAPG Bulletin, 59, 989–996.

Mälicke, M. (2021). SciKit-GStat 1.0: a SciPy-flavored geostatistical variogram estimation toolbox written in Python. Geoscientific Model Development, 15(6), 2505–2532.

Moreira, J.L.P., Madeira, C.V., Gil, J.A., & Machado, M.A.P. (2007). Bacia de Santos. Boletim de Geociências da Petrobras, 15(2), 531–549.

Müller, S., & Schüler, L. (2021). GeoStat-Framework/GSTools: Pastas integration. Zenodo.

Passey, Q.R., Bohacs, K.M., Esch, W.L., Klimentidis, R., & Sinha, S. (2010). From oil-prone source rock to gas-producing shale reservoir—geologic and petrophysical characterization of unconventional shale-gas reservoirs. Society of Petroleum Engineers, SPE-131350.

Passey, Q.R., Creaney, S., Kulla, J.B., Morretti, F.J., & Stroud, J.D. (1990). A Practical Model for Organic Richness from Porosity and Resistivity Logs. AAPG Bulletin, 74(12), 1777-1794.

Peters, K.E., & Cassa, M.R. (1994). Applied source rock geochemistry. In L.B. Magoon & D.G. Dow (Eds.), The Petroleum System—From Source to Trap (pp. 93–120). AAPG Memoir.

Rodrigues, S., Hernández-Molina, F., & Kirby, A. (2021). A Late Cretaceous mixed (turbidite-contourite) system along the Argentine Margin: Paleoceanographic and conceptual implications. Marine and Petroleum Geology, 123, 104768.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536.

Tissot, B.P., & Welte, D.H. (1984). Petroleum Formation and Occurrence (2nd ed.). Springer-Verlag, Berlin.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520-525.

Wright, V.P., & Barnett, A.J. (2015). An abiotic model for the development of textures in some South Atlantic early Cretaceous lacustrine carbonates. Geological Society, London, Special Publications, 418, 209-219.

Zhang, S. (2012). Nearest neighbor selection for iteratively kNN imputation. Journal of Systems and Software, 85(11), 2541-2552.

Zhu, L., Zhang, C., Zhang, C., Zhang, Z., Nie, X., Zhou, X., Liu, W., & Wang, X. (2019). Forming a new small sample deep learning model to predict total organic carbon content by combining unsupervised learning with semi-supervised learning. Applied Soft Computing, 83, 105596.