Applied Computing and Geosciences 22 (2024) 100161

Contents lists available at ScienceDirect

Applied Computing and Geosciences

journal homepage: www.sciencedirect.com/journal/applied-computing-and-geosciences

Fracture density reconstruction using direct sampling multiple-point statistics and extreme value theory

Ana Paula Burgoa Tanaka [a,b,*], Philippe Renard [a], Julien Straubhaar [a]

[a] Centre for Hydrogeology and Geothermics, University of Neuchâtel, Rue Emile-Argand 11, 2000, Neuchâtel, Switzerland
[b] Petrobras, Rua Marquês de Herval 90, Santos, 11010-310, São Paulo, Brazil

ARTICLE INFO

Dataset link: https://github.com/randlab/xgapfilling

Keywords:
Fracture
Reservoir
Geostatistics
Modeling
Well log

ABSTRACT

The aim of this work is to present a methodology for the reconstruction of missing fracture density within highly fractured intervals, which can represent preferential fluid flow pathways. The lack of record can be very common due to the intense presence of fractures, dissolution processes, or data acquisition issues. The superposition of numerous fractures makes the definition of fracture surfaces impossible, as a consequence, modeling such zones is challenging. In order to address this issue, the usage of direct sampling multiple-point statistics to perform gap filling in well logs is demonstrated as an alternative to other techniques. It reproduces data patterns and provides several models representing uncertainty. The method was tested in intervals from a highly fractured well, by removing previously known fracture density data, and simulating different scenarios with direct sampling. Simulation results are compared to the observed data using cross-validation and continuous rank probability score. The reference scenario training data set consists in one well and two variables: fracture density and fracture occurrence. A sensitivity analysis is carried out considering additional variables, additional wells, different intervals, resampling with extremes, and other gap filling techniques. The auxiliary variable plays an important role in pattern matching, but adding wells and logs increases the complexity of the method without improving pattern retrieval. Best results are obtained applying extreme values theory for stochastic process with the enrichment of the fracture density data at the tail region, followed by resampling of the new values. The enriched data is used for the gap filling resulting in lower continuous rank probability score, and the achievement of extreme fracture density values.

1. Introduction

1.1. The importance of missing record in highly fractured zones

The challenge and importance of this work is related to the fact that missing records often happen in nature, not only because the tools do not have resolution to detect some events but also because the products of some processes are not materialized as rock, but in general as voids spaces, such as unconformities, bedding, karst, unfilled faults and fractures.

A highly fractured zone is described by the superposition of numerous fracture surfaces, where sometimes it is impossible to identify each fracture, measure orientation, dip, aperture, roughness, or even count how many fractures cross a line to calculate fracture density. One can infer highly fracture intervals in subsurface from low amplitudes signals in borehole acoustic image logs, where the sinusoidal forms with higher dips cannot be precisely individualized (Fig. 1A), or by the fragmented rock core samples from wells (Fig. 1B).

In Fig. 1C, a single cored fragment of the reservoir reinforces the indication of a highly fractured zone with the sampling of a fault plane with slickensides, indicating movement but uncertain about the attitude, as the core and consequently the fault cannot be oriented due to high fragmentation.

The lack of well log data can also be due to cost cutting, operational drilling problems, and image acquisition issues, either in logging while drilling or in wireline logging. Core samples are not often planned at highly fracture zones because of the low recovery. The consequence of the lack of data in making models is that most likely there will be no value of fracture intensity in the most fractured regions. A common practice from industry is to attribute arbitrary values for the maximum fracture density, or interpolate values from two consecutive points to avoid misinterpretation.

In spite of this, the missing record is meaningful to understand phenomena in nature, and in this case fracture density reflect patterns that are important to understand structure development and controls

* Corresponding author at: Centre for Hydrogeology and Geothermics, University of Neuchâtel, Rue Emile-Argand 11, 2000, Neuchâtel, Switzerland.
E-mail address: ana.burgoa@unine.ch (A.P.B. Tanaka).

https://doi.org/10.1016/j.acags.2024.100161
Received 22 September 2023; Received in revised form 18 March 2024; Accepted 22 March 2024
Available online 23 March 2024

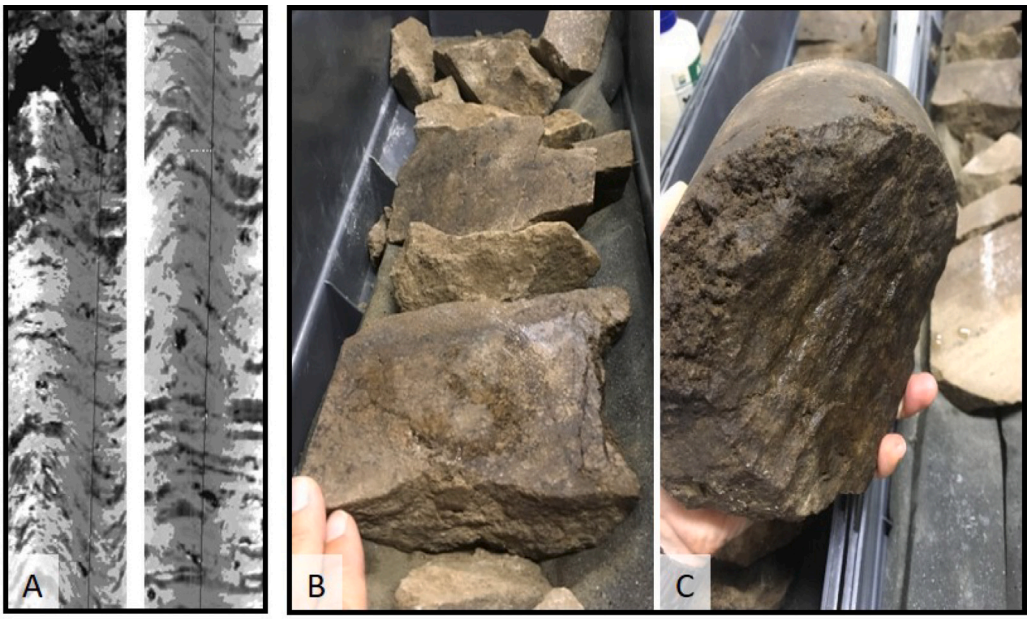


Fig. 1. Fractured intervals in a well: (A) Borehole acoustic image logs, the low amplitudes are showed in dark representing open fractures, highlighting a fracture cluster at the top left. (B) Core sample fragments from a fractured interval with partial recovery. (C) Fault plane with slickensides from a core fragment of the same well, indicating that the low recovery is due to high fracture density.

in diagenetic processes for karstic development, the percolation of hydrothermal fluids, and to characterize uncertainty for reservoir modeling. Fracture distribution, the multi-scale representation of structures, and flow characterization require fracture density measurement or estimation. When fractures occur in sufficient spacing or length their effect on fluid flow becomes crucial (Nelson, 2001; Philip et al., 2005; Sanderson and Nixon, 2018). Whereas connectivity depend mainly on fracture density, length, and angular scatter (Ozkaya and Al-Fahmi, 2022). To asses this effect, it is important to know how many of these fractures exist in a given reservoir volume, and understand how it evolves with depth.

Further examples that highlight the importance of fracture characterization in well logs for fracture modeling and fluid flow of naturally fractured carbonates, from the Santos Basin, can be found in Fernandez-Ibanez et al. (2022), de Jesus et al. (2016), Mimoun and Fernández-Ibáñez (2023), Tanaka et al. (2022, 2018) and Wennberg et al. (2023). In a broad sense fracture modeling is important for hydrocarbon recovery, mining and nuclear waste disposal, water resources assessment and for the development of geothermal systems.

1.2. Data reconstruction with direct sampling multiple-point statistics

The application of direct sampling multiple-point statistics for missing data reconstruction is exemplified in previous works from Mariethoz and Renard (2010), Mariethoz et al. (2012) and Oriani et al. (2016). Mariethoz and Renard (2010) demonstrate that the application of direct sampling recovers the statistical content of the missing data, and generates complex signal structures, with the possible use of auxiliary information.

The concept of multiple-point statistics was developed in the early 1990s, as a non-parametric statistical framework developed to represent heterogeneity. At the beginning it could be thought as a way to work in under-informed situations (Mariethoz, 2018). The first approaches (Guardiano and Srivastava, 1993; Journel, 1993) brought tools that made the inclusion of interpretative knowledge in spatial models practical, as the multiple-point statistics framework had the novelty of carrying the concept and this knowledge within the training image of what was to be modeled. The first successful applications took place in fields where data are typically few, uncertain and expensive, such as reservoir modeling, soil science or mining. Multiple-point statistics is widely applied because it reproduces the structures and patterns present in a training data set, allowing its composition with few hard data. In such cases, the design of the training data set becomes a very important part of the modeling work.

Mariethoz and Renard (2010) propose to use direct sampling for the reconstruction of partially informed images in a wide range of applications. The idea is to extract the statistics directly from the data set, that becomes a training data set. The inference of consistent multiple-point statistics is possible even when the training data set accounts to less than 1% of a 3D volume to reconstruct. Five reconstruction examples are presented: (1) spatial repartition of the training data set using a categorical image of sand channels in a clay matrix, (2) continuous variable example using synthetic transmissivity field with reference images built from aerial photographs of braided channels, (3) 3D synthetic example with the reference image of an object-based simulation of turbidites, (4) 3D case of a quarry in the sediments from a point-bar deposit formed by a meandering river, and (5) borehole image case, where the goal was to fulfill gaps between the pads of a borehole microresistivity image. In the last example, the authors highlight that borehole images are a powerful way to obtain information about bedding and fractures.

Mariethoz et al. (2012) explore an approach to produce spatially continuous fields from discontinuous data focusing on the reconstruction of gaps of satellite-based Earth observations. Using synthetic imagery derived from a regional climate model, they demonstrate that the method is straightforward, as it requires minimum parameter adjustment and user intervention. The approach could represent complex spatial patterns and the fine-scale data structure. Direct sampling provided more realistic spatially continuous fields than other kriging methods. As kriging might result in interpolated areas that are clearly distinct from the rest of the image, presenting unrealistic continuous textures and, possibly artifacts.

Oriani et al. (2016) propose the reconstruction of a flow rate time-series, considering several missing data scenarios with auxiliary data. A multivariate direct sampling setup was proposed including two out-of-phase periodic triangular functions, an indicator variable, and the flow rate time-series as the main variable. It generates more realistic simulations in comparison with an autoregressive moving average with

exogenous variable. The prediction becomes better when a correlated flow rate time-series is used. Direct sampling allows the simulation of complex natural process by sampling the available data set where sufficiently similar patterns can be found, without the requirement of a high amount of data. The efficiency depends on finding the good ensemble of variables, best suitable for the application.

1.3. Extreme value theory applied for stochastic processes

The lack of extreme values is an intrinsic characteristic of the missing record, as the highest values might not be available in the considered data set. The importance of extreme events in environmental hazards and in assessing risk, is described by Haan and Ferreira (2006) for the Netherlands, which territory is approximately 40% below sea level. In this case, the government required that dikes would be as high, so that the probability of the seawater level exceeding the top of the dike in a given year would be very low, 10^{-4}. The question was how to estimate the probability of a once in thousand years event when the observations of high-tide water levels during storms were limited to 100 years? Going beyond that range seemed to be almost impossible. However, the solution arises with the use of extreme value theory, since it provides a theoretical basis and framework for extrapolation. It restricts the behavior of the distribution function in the tail to resemble a limited class of functions that can be fitted to the tail of the distribution function (Haan and Ferreira, 2006).

Different applications of statistical analysis of extreme values to hydrology and environmental sciences, finance and insurance, material sciences, and human sciences can be found in Reiss et al. (1997). The authors describe the importance of the extreme part of the samples as it may exhibit a larger risk potential of random events such as floods, hurricanes, high concentration of air pollutants, price shocks, and others. Reactions to the likelihood of a future catastrophe may help to prevent a greater disaster to happen. Therefore, the statistical insight gained from extremes can be considered decisive in daily business, for the solution to ecological or technical issues.

Fracture density data generally shows heavy tail distribution, the sampled data close to the extreme values present higher spacing among the points. A proper reconstruction of the missing data requires to account for the lack of the extreme values in the data set. The standard direct sampling algorithm does not generate new values beyond the observed range. To overcome that issue, the proposition is to enrich the original data set in the tail region and resample with lifting based on functional extreme value theory, allowing the extrapolation of the observed data towards yet unobserved high quantiles with direct sampling as proposed by Opitz et al. (2021).

The semi-parametric resampling with extremes (Opitz et al., 2021) aims to enrich the original data with new values at the tails of the distribution by generating an independent sample of the marginal distribution while keeping the rank order of the observed data, and apply classical resampling algorithms fixing a target range of return levels of a magnitude variable and resampling the magnitudes constrained to that range. The authors use the method to generate heatwave scenarios over France, based on daily temperature reanalysis. The generation of extreme values is showed to be crucial for climate-related applications to understand its impact on the historical record and in future events (Opitz et al., 2021).

1.4. Objective

The objective of the article is to propose and test a methodology to infer missing fracture information along borehole well logs. The methodology considers the available information from well logs to fulfill fracture density from highly fractured areas using direct sampling multiple-points statistics and tools from extreme value theory for stochastic processes.

Direct sampling was chosen because it is able to simulate realistic natural patterns based on the data set, and it requires few parameters. It provides several models that can represent the intrinsic uncertainty of this type of data. Resampling with extremes is applied to reach extreme values in the reconstruction of missing data.

The methodology was tested in a real data set from a naturally fractured reservoir. But the aim of the work is to propose a methodology that could be broadly applied in various different settings. For petrophysical and structural geology studies, to facilitate the routine work of log processing and fracture modeling from different sites, by improving the quality of well log gap filling and making models more predictive.

2. Methodology

2.1. Data set and study site

The data set used is composed of 24 wells with directional surveys, conventional well logs (gamma ray, resistivity, neutron porosity, density, and sonic logs), top reservoir stratigraphic markers, and fracture interpretation from borehole acoustic and resistivity image logs.

The wells were drilled in an isolated naturally fractured lacustrine carbonate Aptian reservoir located on the continental margin of SE Brazil, in the South Atlantic Ocean (Tanaka et al., 2022). Specifically in the central portion of the Santos Basin, on the top of a long-lived basement paleo structural high known as the Santos External High (Carminatti et al., 2009) (Fig. 2). The area shows dominantly NNE-striking normal faults and the north domain of the reservoir is highly fractured (Tanaka et al., 2018, 2022).

2.2. Selection of variables for the training dataset

The fracture density P10 is the main variable of interest. It was calculated with a sampling window of 1 m and corrected using the borehole deviation (Terzaghi, 1965). As convention P10 [m^{-1}] is the 1-d (linear) density defined as the number of fractures per unit length of the sampling line, according to the system of measures that describes fracture abundance (Dershowitz and Herda, 1992), where P stands for persistence, followed by subscripts designating the dimension of the measurement region, or sampling domain (1: borehole) and the dimension of the measure or fracture attribute (0: count of a number of fractures).

For the reconstruction of fracture density, the available P10 was used both as conditioning data and training data set (Fig. 3). For the first approach only P10 was used as a training data set, other auxiliary variables were gradually added to compare results.

Several auxiliary variables were tested to deal with the non-stationarity of the training data set. The use of auxiliary variables in multiple-point statistics simulations was first suggested by Chugunova and Hu (2008). For this work the selection of auxiliary variables was inspired by Oriani et al. (2014) that applied direct sampling to simulate daily rainfall time series in order to reproduce the complexity of the rainfall signal up to the decennial scale, with the aid of a multivariate data set.

The idea is to select variables that describe the low-frequency fracture density trend. Several variables were tested: the P10 densities calculated with different sampling windows (A1 = 10 m), classes of fracture occurrence (OC = 0,1,2,3) calculated from the P10, and additional well logs such as the compression sonic log (DTCO). An example of the training data set selection for well-11 is showed in Fig. 3. The highlighted gaps were used for cross-validation, as intervals of less (LFG), intermediate (IFG), and higher (HFG) mean fracture density P10.

The low frequency trend A1 was built from a sample window of 10 m, as derived from the main variable P10. Another auxiliary variable was built as classes of fracture occurrence defined in terms of fracture intensity with class 0: P10=0, class 1: $0 < P10 < 5$, class 2: $5 \leq P10$

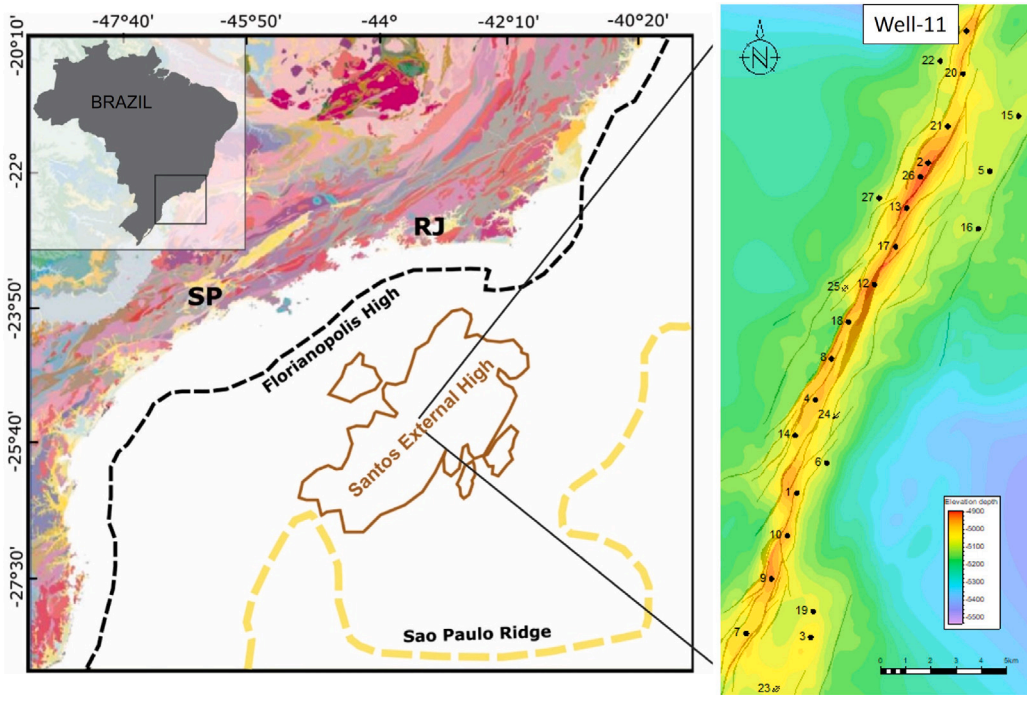


Fig. 2. Location of the study area in the Santos Basin External High in SE Brazil. The right figure shows a zoom on the structural elevation map of the top reservoir, with the black traces representing interpreted faults. The numbered black dots show the location of the drilled wells.
Source: Modified from Tanaka et al. (2022).

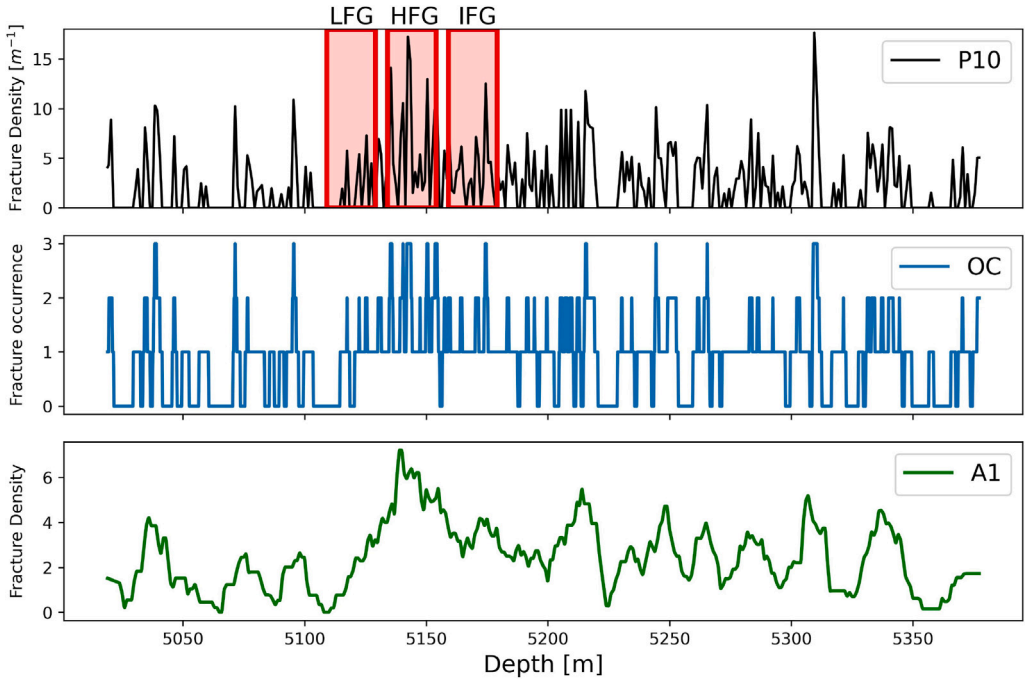


Fig. 3. Training data set from well-11 for gap filling: fracture density (P10), fracture occurrence (OC) and low frequency trend (A1). Highlighted gaps in red at the P10 track, where data set was excluded for cross-validation at the less fractured interval (LFG), higher fractured interval (HFG), and intermediate fracture interval (IFG).

< 10, and class 3: P10 ≥ 10. The occurrence is a relative measure that can be obtained from the wells and when the fracture intensity is not available it can be interpreted from other well logs, production data, and conceptual knowledge from the area. The identification of lower amplitudes signals might be possible at the acoustic image log, where fracture occurrence might be estimated by relative comparing the amplitudes. To fulfill gaps with adequate values and patterns, the previous knowledge if either the interval has a higher fracture intensity is ideal, but not required, as it is possible to work with auxiliary variables with different resolutions.

Table 1
Summary of the parameters used in the direct sampling setup.

Variable	Distance threshold (t)	Number of neighbors (n)	Maximal scan fraction (f)
Main	0.05	30	0.5
Auxiliary	0.08	1	0.5

The fracture occurrence generation was inspired by the dry/wet sequence presented in Oriani et al. (2014), a variable indicating the position of a day inside a rainfall pattern of wet, dry, solitary wet, and wet at the beginning or end of a wet spell, that allowed the simulation to reproduce realistic patterns and reach higher values.

2.3. Direct sampling setup

Among the multiple-point statistics techniques, the direct sampling operates in a multivariate framework and constitutes a very flexible tool for reproducing complex features within and between variables. It can handle both categorical and continuous data (Mariethoz et al., 2010; Straubhaar et al., 2020). The simulation process consists in sequentially populating the missing cells in the simulation grid with values sampled from the training data set. The algorithm relies on the comparison of patterns using a distance function defined accordingly to the type of variable (categorical or continuous) and taking values between 0 and 1, interpreted as a mismatch rate (Mariethoz et al., 2010; Straubhaar et al., 2020). The main parameters of the algorithm are the maximal number of neighbors in the pattern (n), the maximal distance from compatible patterns or acceptation threshold (t), and the maximal fraction (f) of the training data set that can be scanned for the simulation of each node. This last parameter allows to control the simulation time, if the acceptation threshold is not reached after having scanned a fraction f, the search is stopped and the best candidate is retained (Mariethoz et al., 2010; Straubhaar et al., 2020).

The gap filling simulations were conducted using the direct sampling multiple-point statistics DeeSee algorithm from the Python Package Geone.[1] The DeeSse algorithm can handle multiple variables simultaneously, auxiliary categorical and continuous variables that can guide the simulation of the main variable of interest, which improves the quality of the simulation results.

Different direct sampling parameters were tested, but in the following we present the results obtained with $n = 30$, $t = 0.05$ for the primary variable, $n = 1$, $t = 0.08$ for the auxiliary variables, and $f = 0.5$ for all the scenarios (Table 1). For each scenario, 20 stochastic simulations were generated.

The parameters were selected based on tests and what has been used in previous works (Oriani et al., 2014, 2016). Oriani et al. (2014) use $t = 0.05$ for all variables from the setup, and a low maximal number of neighbors $n = 1$, and $n = 10$ for auxiliary variables to condition dry/wet patterns on the small scale and to reproduce realistic patterns and reach higher values. For the reconstruction of flow rates Oriani et al. (2016) use for two auxiliary variables a threshold $t = 1$, and a low maximal number of neighbors $n = 1$. The maximal fraction $f = 0.5$ was selected in order to scan 50% of the training data set. Ranges were tested but the best continuous rank probability score (CRPS) came from the presented setup.

2.4. Extreme values resampling

To enhance the capture of extreme values, semi-parametric resampling with extremes is used alongside the gap filling method. While initially tested for reconstructing fracture density data, this approach holds broader applicability in settings dealing with extreme values. Resampling with extremes is implemented using algorithms proposed by Opitz et al. (2021).

[1] https://github.com/randlab/geone.

The key steps involve: (1) estimating the univariate probability density function to model and estimate the marginal distribution, and (2) choosing between enriching the tail using a technique called naïve resampling or enriching with a lifting mechanism.

For additional details, the original names of the algorithms developed by Opitz et al. (2021) are cited for each step, summarized as follows:

(1) To estimate the marginal distribution, first the data is modeled with a kernel density estimation (KDE). Subsequently, the marginal right tail of the distribution is modeled with a generalized Pareto distribution requiring a probability (p_u) of exceedance over the threshold (u). The extreme value theory restricts the behavior of the distribution function in the tail to resemble a limited class of functions.

Fig. 4 illustrates the distribution of P10 fracture density data, from well-11, modeled with a kernel density function. The chosen probability of exceedance $p_u = 0.3$ is emphasized in the cumulative distribution function (CDF) derived from the KDE (Fig. 4).

The univariate density model is then expressed as:

$$\hat{f}(x) = \begin{cases} \bar{f}(x), & x \le u, \\ p_u f_{GP}(x - u \mid \hat{\sigma}_u, \hat{\xi}_u), & x > u, \end{cases} \tag{1}$$

where $\bar{f}$ is provided by the KDE, and f_{GP} is the density of a generalized Pareto distribution with scale (σ) and shape (ξ) parameters, whose the survival function is defined for $y > 0$ as

$$P(Y > y \mid \sigma, \xi) = \int_y^{+\infty} f_{GP}(z \mid \sigma, \xi) dz = \begin{cases} (1 + \xi y/\sigma)^{-1/\xi}, & \xi \neq 0, \\ exp(-y/\sigma), & \xi = 0 \end{cases} \tag{2}$$

The scale parameter is first set to $\hat{\sigma}_u = p_u / \bar{f}(u)$ to ensure the continuity of the modeled density, and then the estimated shape ($\hat{\xi}_u$) is obtained using a tail index estimator such as maximum likelihood estimator computed from the data above u. Hence, the univariate density model is defined from the data, a KDE and the given probability of exceedance p_u (Algorithm 1, Estimation of the univariate probability density function, Opitz et al., 2021).

(2) Enriching data can be achieved by naïve resampling as follows. A new set of data values, of same size as the training data set, is generated by sampling from the modeled density $\hat{f}$. The vector of ranks in the sorted list of the training data values is used to reorder the new data values, to reproduce the patterns and keep spatial (or temporal) consistency. (Algorithm 2, Naïve resampling, Opitz et al., 2021).

A lifting mechanism can be applied as an alternative to naïve resampling. First, the distribution X of data values, of CDF $\widehat{F}$ modeled in step 1, is transformed to $X^{(P)} = 1/(1 - \widehat{F}(X))$, which is a standard Pareto distribution of shape parameter $\alpha = 1$, whose cumulative density is $F(z) = 1 - 1/z^{\alpha}$. Then, a uniform marginal scale is used by considering the variable $U = -1/X^{(P)} = \widehat{F}(X) - 1$ which follows a uniform distribution on the interval $[-1, 0)$. The threshold value for the new variable U is then $v = \widehat{F}(u) - 1 < 0$. The lifting mechanism requires two additional parameters $v_1 \le v_2 \le 0$ (and $v_1 < 0$ if $v_2 = 0$) and consists in drawing a new value $\tilde{v}$ uniformly between v_1 and v_2, and considering the lifted variable $\widetilde{U} = s \cdot U$, with the scaling factor $s = \tilde{v}/v$. If $\tilde{v} > v$, then $0 < s < 1$ (since both $\tilde{v}$ and v are negative) and the variable U is "uplifted" (Algorithm 3, Lifting with uniform data margins, Opitz et al., 2021). The lifting operation is only applied to (uniform scale) data values exceeding the threshold v, followed by a post-processing to avoid discontinuities around the threshold value (and values less than -1 in case of "downlifting") (Algorithm 4, Lifting with uniform data margins and postprocessing, Opitz et al., 2021). Finally, the result is back-transformed in the original scale via $\widetilde{X} = \widehat{F}^{-1}(1 + \widetilde{U})$. To

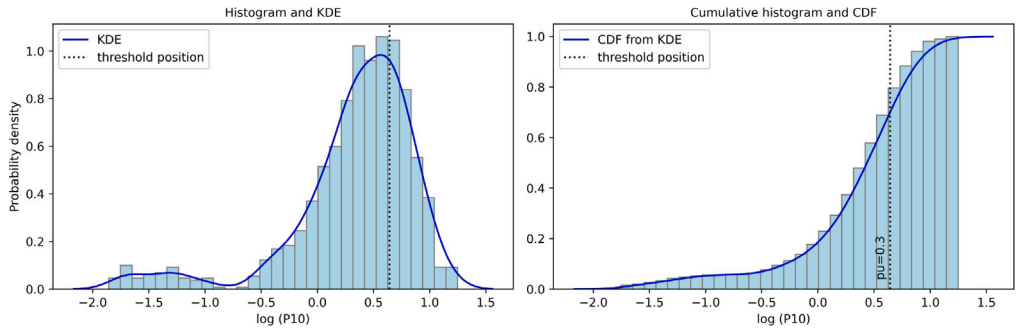


Fig. 4. Histogram of log fracture density (P10) from well-11, fitted kernel density estimate (KDE), and cumulative density (CDF) from kernel density estimate.

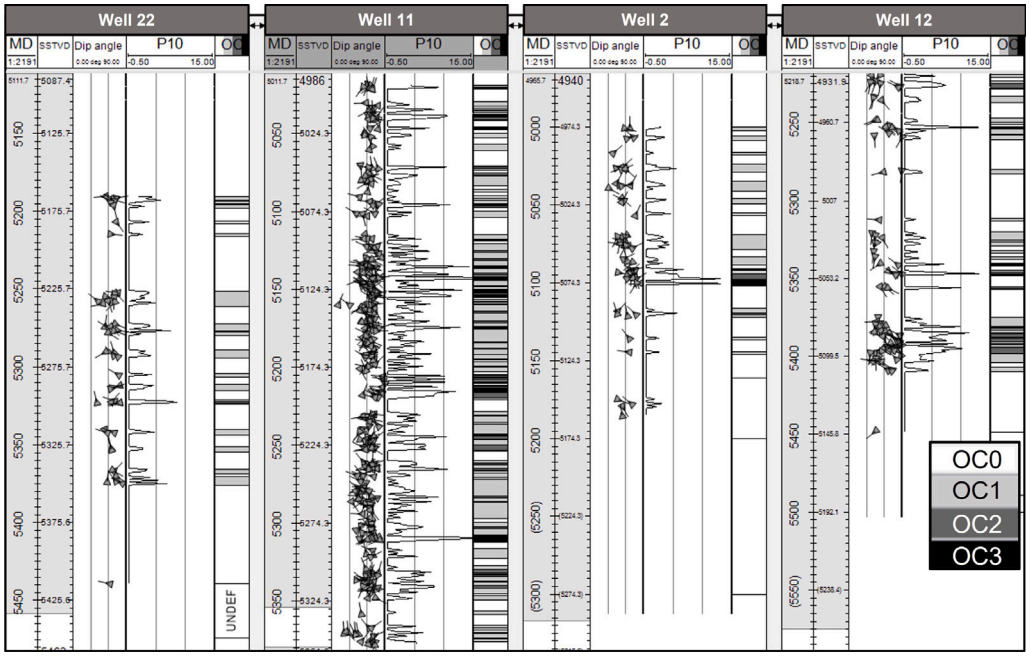


Fig. 5. Well correlation with measured depth (MD), vertical depth (SSTVD), tadpoles as fracture interpretation (Dip Angle), fracture density (P10), and fracture occurrence (OC). Wells locations in Fig. 2.

ensure better spatial consistency, this method can be applied to a set of realizations, by first applying lifting to get an enriched data set and then generating new realizations by resampling based on the enriched data (Algorithm 5, Resampling dependent extremes, Opitz et al., 2021). In addition, following this procedure allows to easily account for local constraints (conditioning data points) in the realizations.

2.5. Testing procedure and quality criteria

Well-11 was chosen for cross-validation as it characterizes longer intervals of highly fractured reservoir. A well section illustrates different fracture densities from correlated wells of the same structural domain (Figs. 2, 5).

Three intervals of 20 m each were selected for testing and divided according to the mean fracture density, described from the base to the top as: intermediate fractured gap (IFG) = 3.50 m^{-1}, higher fractured gap (HFG) = 5.92 m^{-1}, and less fractured gap (LFG) = 1.55 m^{-1} (Fig. 3). The observed P10 values in these intervals are known, but removed from the data set for cross-validation. Gap filling algorithm is run and the results are compared to the observed data.

Among all the tests a *reference scenario* is defined. It corresponds to the test of a training data set (TDS) that includes the main variable P10 and the auxiliary variable fracture occurrence (OC), for the filling of higher fractured interval, using only the data from well-11. Five other scenarios, with variants, are presented for the sensitivity analysis and compared to the reference (Table 2).

(1) *Selecting variables*: no auxiliary variable for comparison, one auxiliary variable (A1 or DTCO) is tested instead of the fracture occurrence variable, and additional auxiliary variables are included to the reference scenario training data set. The training data corresponds to well-11 only.
(2) *Adding wells*: training data from 5 and 24 wells are included, considering only the variables P10 and fracture occurrence.
(3) *Different intervals*: the same training data set is used from the reference scenario but the filled gaps are located in less fractured intervals.
(4) *Resampling with extremes*: gap filling considers P10 values enriched with naïve resampling or lifting with extremes.
(5) *Other techniques*: different types of interpolation (linear, piecewise cubic hermite, and cubic spline), two-point statistics (ordinary kriging, cokriging and sequential gaussian simulation), and machine learning (neural network and random forest) are applied.

The quality of the tests was evaluated with the continuous rank probability score (CRPS) as proposed by Juda et al. (2020) for multiple point statistics simulations. It is a scoring rule originally described

Table 2
Summary of the scenarios, sensitivity analysis and comparison with other techniques.

Scenarios		Wells	Variables	Interval	Method	Mean CRPS
Reference	a	1	(2) P10, OC	HFG	Direct sampling	0.79
(1) Selecting variables	b	1	(1) P10	HFG	Direct sampling	5.06
	c	1	(2) P10, A1	HFG	Direct sampling	3.15
	d	1	(2) P10, DTCO	HFG	Direct sampling	3.95
	e	1	(3) P10, OC, A1	HFG	Direct sampling	0.77
	f	1	(3) P10, OC, DTCO	HFG	Direct sampling	0.98
	g	1	(4) P10, OC, A1, DTCO	HFG	Direct sampling	0.99
(2) Adding wells	h	5	(2) P10, OC	HFG	Direct sampling	0.95
	i	24	(2) P10, OC	HFG	Direct sampling	0.97
(3) Different intervals	j	1	(2) P10, OC	LFG	Direct sampling	0.39
	k	1	(2) P10, OC	IFG	Direct sampling	0.70
(4) Resampling with extremes	l	1	(2) P10, OC	HFG	Direct sampling, naïve	0.73
	m	1	(2) P10, OC	HFG	Direct sampling, lifting	0.77
	n	1	(2) P10, A1	HFG	Direct sampling, naïve	2.97
	o	1	(2) P10, A1	HFG	Direct sampling, lifting	3.07
	p	1	(3) P10, OC, A1	HFG	Direct sampling, naïve	0.76
	q	1	(3) P10, OC, A1	HFG	Direct sampling, lifting	0.78
(5) Other techniques	r	1	(1) P10	HFG	Linear interpolation	4.21
	s	1	(1) P10	HFG	Piecewise cubic hermite	4.52
	t	1	(1) P10	HFG	Cubic spline	16.94
	u	1	(1) P10	HFG	Ordinary kriging	3.55
	v	1	(1) P10, OC	HFG	Cokriging	2.35
	w	1	(1) P10	HFG	Sequential guassian simulation	3.94
	x	1	(1) P10, OC	HFG	Sequential guassian simulation	2.69
	y	1	(1) P10	HFG	Neural network	4.15
	z	1	(1) P10	HFG	Random forest	5.27

by Gneiting and Raftery (2007) and Gneiting et al. (2007), that quantifies the quality of a probabilistic forecast by comparing it with a single true value. If x represents an observation, the cumulative distribution function (CDF) F associated with the empirical probabilistic forecast (such that $F(y) = P[X \leq y]$) is considered to compute the CRPS between x and F, defined as:

$$CRPS(F, x) = \int_{-\infty}^{\infty} (F(y) - \mathcal{H}(y - x))^2 dy \tag{3}$$

where $\mathcal{H}$ is the Heaviside step function taking the value of 1 if the real argument is positive or zero, or the value of 0 otherwise. The CRPS is expressed in the same unit as the observed variable and generalizes the mean absolute error (Gneiting and Raftery, 2007). In this sense, lower scores indicate a better match between ensemble simulations and the hard data at every observed data point. The comparison has been conducted on a large ensemble of scenarios.

3. Results and discussion

3.1. Fracture density reconstruction with direct sampling

Table 2 summarizes the scenarios and sensitivity analysis done to test the methodology, and compare results from each different combination of training data sets, without and with extreme resampling, and with other techniques to fulfill gaps in fracture density.

Fig. 6 illustrates results from one single simulation of each scenario of well-11 gap filling without extremes, at the higher fractured interval. From the reference scenario (a) to scenarios 1 and 2 (b to h) where the training data set is composed of different variables and additional wells.

For the selection of other auxiliary variables, an example of simulation without an auxiliary variable (Scenario 1b) is presented with high mean CRPS of 5.06 and pattern mismatch (Fig. 6b), highlighting the importance of the auxiliary variable.

When auxiliary variable A1 or the sonic log DTCO are selected alone as auxiliary variables (Scenario 1c and 1d), the mean CRPS are high 3.15 and 3.95, respectively, but patterns do not match (Fig. 6 c, d).

By adding the low frequency trend as auxiliary variable to the reference training data set (Scenario 1e), there is a slight improvement in comparison to the reference scenario (Fig. 6e). With pattern matching and lower CRPS, as it changes from 0.79 from the reference scenario to 0.77.

When well logs are added, e.g. the sonic log DTCO, to the reference scenario, and the low frequency trend as auxiliary variable, (Scenario 1f and 1g), the results do not improve, as CRPS turns to 0.98 and 0.99. In both cases the pattern still matches, but higher values are not reached (Fig. 6f, g). The curve becomes slightly more erratic, that could be due to the different resolution. Another aspect is that well logs are the result of various natural events imprinted on the rock and that they can also reflect the characteristics from the fluid. Therefore, the well log records are not directly related to fracturing. On the opposite, the fracture occurrence variable relates fracture density with mechanical stratigraphy in a sense that is not captured by the other logs, other than a borehole image log interpretation. This is why working with the auxiliary variable of fracture occurrence improves pattern predictability but it could be a limitation if there is no previous knowledge about the most fractured zones.

By considering more wells in the training data set (Scenario 2), there is not an improvement as CRPS turns to 0.95, and 0.97 by considering 5 (Fig. 6h) or 24 wells, respectively. Although the results are still good. As well-11 is the highest fracture well in the area, the additional data does not improve the matching of the highest values. Figs. 2, 5 illustrate the position of the wells and the difference of fracture intensity between wells.

When intervals with different fracture density are submitted to gap filling (Scenario 3), as expected, the best result are obtained for the less fractured interval (Fig. 7j) with mean fracture density of 1.55 and CRPS of 0.39, in comparison to the reference scenario with mean fracture density of 5.92 and CRPS of 0.79. The intermediate fracture interval is in between both cases, with mean fracture density of 3.50 and CRPS of 0.70 (Fig. 7k). In general, power law distribution, log-normal, and exponential laws describe rather well the frequency distribution of fracture density (Bonnet et al., 2001; Davy et al., 2013). As lower fracture densities are more frequent than higher density, lower P10 values are easier to sample for the direct sampling algorithm. Besides that, the zero values can be precisely matched in the simulations with the aid of the auxiliary variable (Fig. 7).

The fracture occurrence showed to be an important guide to retrieve accurate patterns of peaks and valleys (Figs. 6, 7).

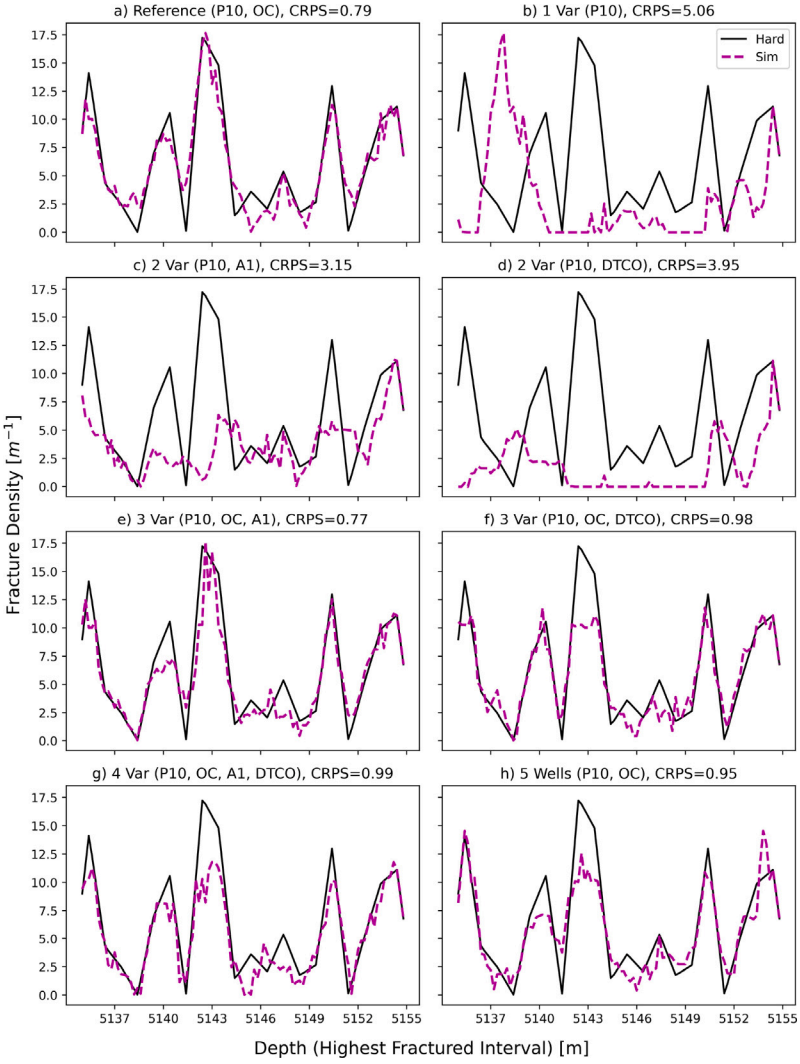


Fig. 6. Results from different scenarios of fracture density P10 reconstruction with direct sampling from well-11 at a zoom at the higher fractured interval, with the correspondent scores. The observed data is shown in black solid line and the simulations in magenta dashed line. (a) Reference scenario with one well and two variables, (b) Scenario with one well and one main variable, (c) Scenario with two variables, with a low frequency trend as auxiliary variable, (d) Scenario with two variables, with the sonic log as auxiliary variable, (e) Scenario with three variables, adding the low frequency trend to the reference scenario (f) Scenario with three variables, adding the sonic log to the reference scenario, (g) Scenario with four variables, (h) Scenario with five wells and two variables.

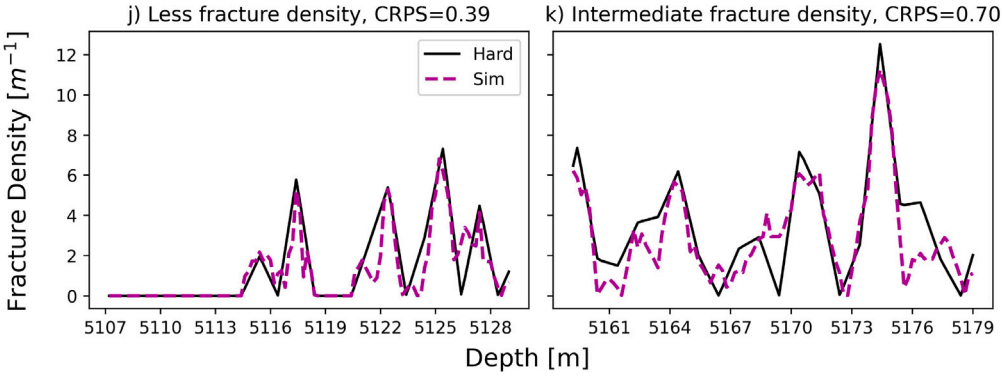


Fig. 7. Results from fracture density P10 reconstruction from well-11 with direct sampling, using one well and two variables, at different intervals of lower fracture densities, with the correspondent scores. The observed data is shown in black solid line and the simulations in magenta dashed line. (j) Scenario of gap filling of the less fractured interval selected for cross-validation, (k) Scenario of gap filling of the intermediate fractured interval.

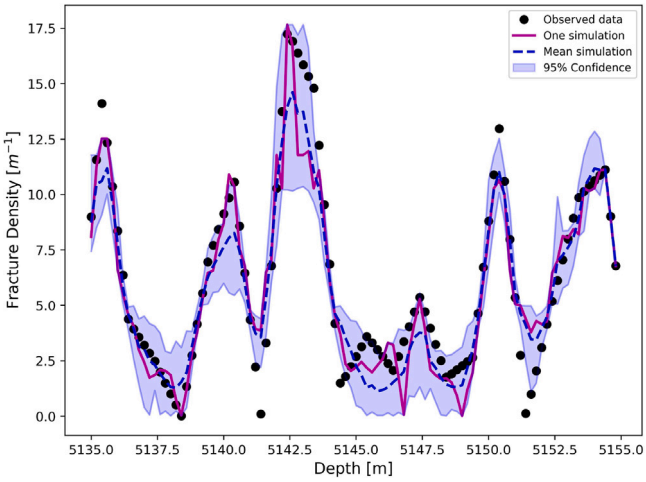


Fig. 8. Results from the gap filling of well-11 at the higher fractured interval, reference scenario. The points are the observed data, the magenta line is one simulation, the blue traced line is the mean from the ensemble simulations and the shaded area is the confidence intervals.

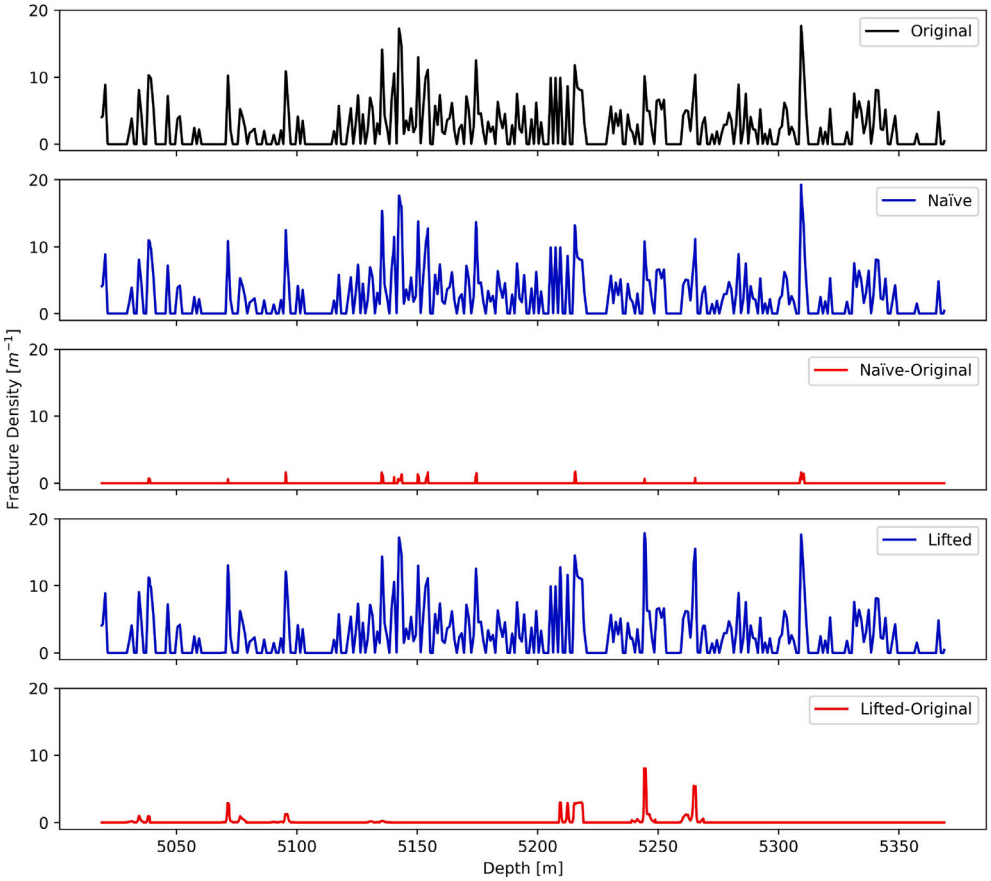


Fig. 9. Fracture density curves (P10), from top to bottom: original observed data before resampling, enriched data after naïve resampling, difference between the enriched curve with naïve resampling and the original, enriched data after lifting, and difference between the enriched curve with lifting and the original data.

Fig. 8 shows results from 20 simulations, with confidence interval, for the reconstruction of P10 at the higher fractured interval from well-11 (Reference scenario 1a). Only a few points of extreme fractured densities are outside the range obtained by the simulations.

Most of the higher and lower P10 values that characterize the described patterns could be reached with well-11 data only (Fig. 8), although it still misses to match all the highest values, for example at 5135.6 m and 5150.6 m. Simulating the highest values of the distribution are a challenge for the gap filling with direct sampling, such values can be achieved by resampling with extremes.

3.2. Resampling fracture density with extremes

Resampling with extremes was performed using data from well-11.

After the original data was enriched either by naive resampling or lifting with extremes, gap filling was applied for the reference scenario by replacing the original P10 with the new enriched P10 values (Scenario 4) (Table 2). The modeling considered only the enrichment of the values from the right tail, the higher values.

Fig. 9 show the original P10 curve for well-11, the enriched P10 curve after naive resampling, the difference between them, the enriched

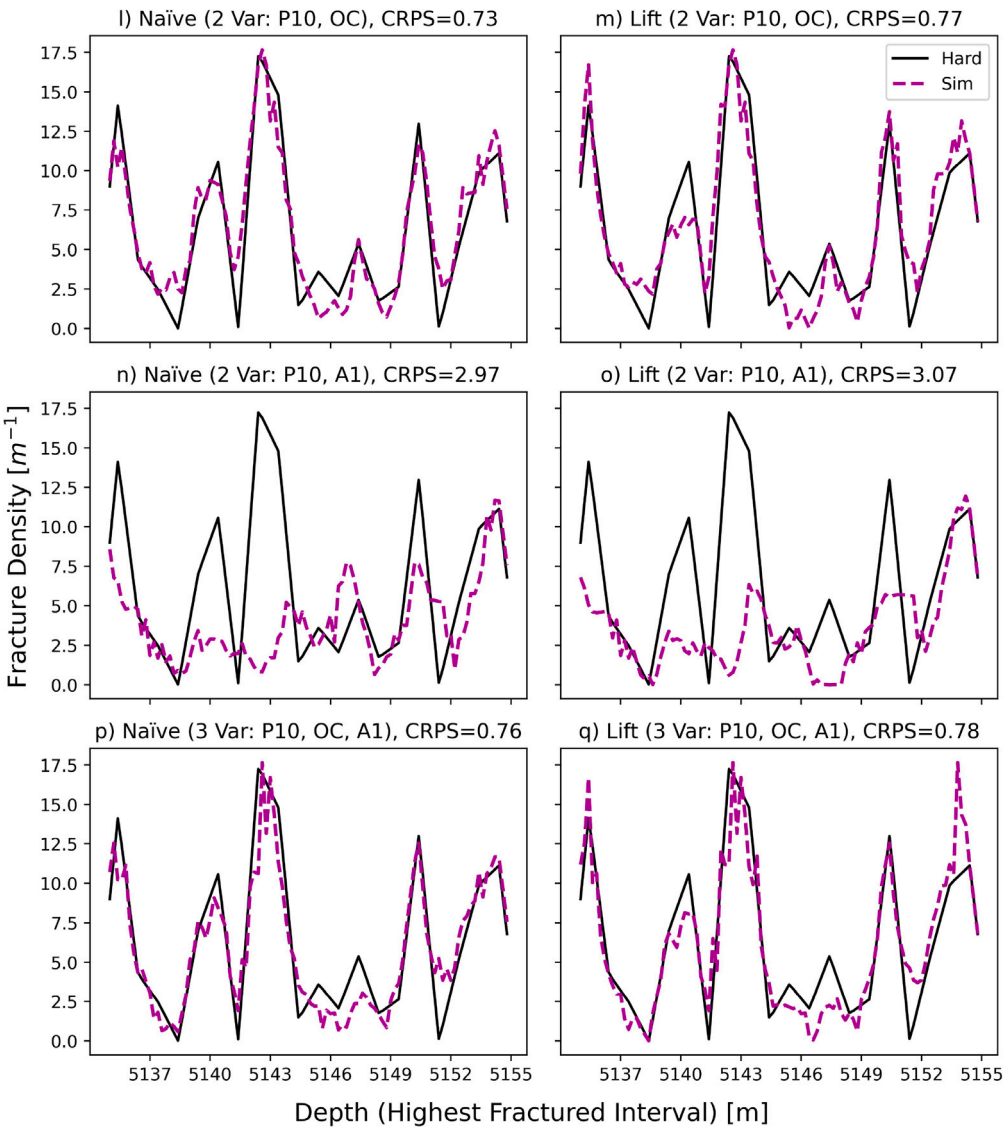


Fig. 10. Results from different scenarios of fracture density P10 reconstruction with direct sampling and resampling with extremes from well-11 at a zoom at the higher fractured interval, with the correspondent scores. (l) Scenario with two variables like the reference scenario, where P10 is enriched by naïve resampling, (m) Scenario with two variables like the reference scenario, where P10 is enriched by lifting, (n) Scenario with two variables, with a low frequency trend as auxiliary variable, and enriched P10 by naïve resampling, (o) Scenario with two variables, with a low frequency trend as auxiliary variable, and enriched P10 by lifting, (p) Scenario with three variables, and enriched P10 by naïve resampling, (q) Scenario with three variables, and enriched P10 by lifting.

P10 curve after lifting, and the difference between the lifted and the original curve.

In both cases the substitution of the original P10 curve for the new resampled values took place where the new values were higher than the original ones.

Fig. 10 illustrates the gap filling with the enriched P10 curves with different auxiliary variables, the best result is from naïve resampling (Scenario 4l), with CRPS 0.73 (Fig. 10l). The training data set consists in 1 well, and 2 variables: enriched P10 and fracture occurrence. Lifting was also applied to the same setting (Scenario 4m) obtaining CRPS of 0.77 (Fig. 10m).

When the fracture occurrence is substitute by the low frequency trend A1, the CRPS increase to 2.97 with naïve resampling (Scenario 4n) and 3.07 with lifting (Scenario 4o) (Fig. 10n, o).

With the addition of a third auxiliary variable (Fig. 10p, q) results are still good with CRPS of 0.76 and 0.78, respectively. Showing a small gain from the reference scenario.

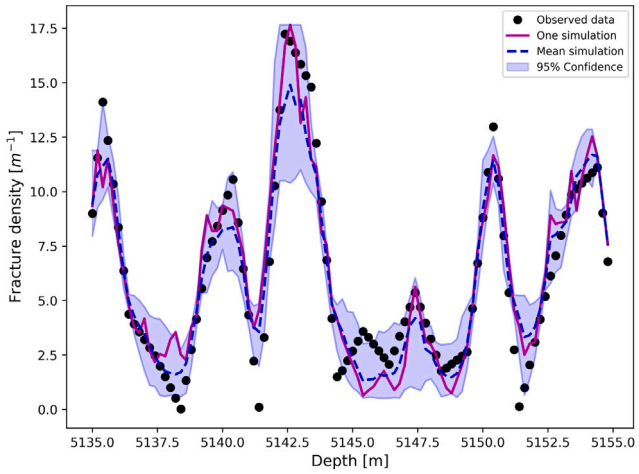
The ensemble simulation for Scenarios 4l and 4m, built with naïve resampling and lifting match most of the highest values and reach extreme high values in both cases (Fig. 11).

The improvement in CRPS of 0.73 and 0.77 respectively for the naïve resampling and lifting, in comparison to the reference scenario without extremes of 0.79, might seem to be minor but the major gain of the application of both approaches is reaching the high P10 value at 5135.6 m. This high value was not achieved previously without the extremes, being now reached due to the extremes (Figs. 8 and 11). For the two cases it is possible to reach values unobserved by the original data. Although with lifting it was also possible to match high P10 value at 5150.6 m (Fig. 11).

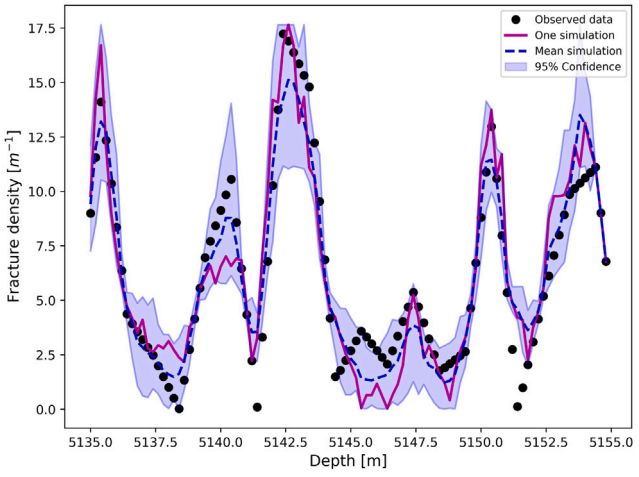
3.3. Comparison with other methods of gap filling

The interpolation of well logs is a routine task for the processing of petrophysical data. The most common interpolation technique used is linear interpolation which is most suitable for small gaps (maximum of 5 values) and not for a large number of missing values (Churikov and Grafeeva, 2018).

To compare various reconstruction gap filling we use different interpolation techniques such as linear interpolation, piecewise cubic



(a) Naïve Resampling



(b) Lifting mechanism

Fig. 11. Results from the gap filling of well-11 at the higher fractured interval with direct sampling and resampling with extremes ($p_u = 0.3$). The points are the observed hard data, the magenta line is the result from one simulation, the blue dashed line is the mean from the ensemble simulations and the shaded area is the confidence intervals from the simulations.

hermite interpolating polynomial (PCHIP), and cubic spline (Scenario 5r to 5t). Additionally, for two-point statistics, ordinary kriging, cokriging and sequential gaussian simulation (SGS) were used (scenarios 5u to 5x), and for machine learning, neural network and random forest were explored (Scenario 5y to 5z). The continuous rank probability score was calculated for each case (Table 2). While the techniques serve for gap filling by generating values to fill the gaps, they may not succeed in retrieving patterns within the gap.

Linear interpolation and piecewise cubic hermite are quick methods for obtaining extrapolated values to fill gaps, but they fall short in capturing the patterns of fracture density and its uncertainties. Kriging and SGS perform better than interpolations and allow the use of auxiliary variables, although even with the auxiliary variable OC pattern retrieval is not better then the reference scenario. Machine learning methods, such as gradient boosting and random forest, demand the configuration of a larger number of parameters to split data and set up tree learning.

Fig. 12 illustrates the comparison between methods. Direct sampling approach for the reference scenario shows better results in terms of pattern match and with lower CRPS.

In terms of time spent in the algorithm setup, the proposed approach of gap filling with direct sampling is in between the standard interpolation techniques and the machine learning methods. It does not require a large amount of data set or data augmentation and it outputs several simulation that can be used to characterize uncertainties.

Lopes and Jorge (2018) present well log gap prediction using machine learning methods such as artificial neural networks, gradient tree boosting, and random forests, in comparison to algorithms of linear regression. They gather a basic suite of logs from thousand wells for the prediction of neutron porosity missing gaps of a single well. Concluding that the random forests and gradient boosting performed better than the linear approaches. They show that the greater the gap, the higher the average error, as it could be expected due to less amount of data remaining for the training, typically showing more variance in the target values.

More recently Shakiba et al. (2022) present a study of one-dimensional fracture arrangements with multi-scale data analytics using Ripley's K-function, as a measure of spatial interaction. They use it to fill spatial gaps where data is absent by taking into account the spacing between fractures and position along a scanline. The idea is to randomly choose and relocate a fracture along the interval by comparing the original k-function to the one from the new fracture arrangement, if the mismatch is reduced the new arrangement is accepted. For restoring fracture gaps it needs reference values of the k-function, for their case study the authors assume that fracture arrangement in the gap at the middle of the scanline will be similar to the rest of the area. The arrangements were initially limited to 10 m, but after data imputation they could extend to 30 m.

In the context of fracture density reconstruction, the target variable is derived from borehole image log interpretation. Gaps can be in the scale of tens of meters, so greater gaps are more difficult to match the observed data. Importantly, the objective of gap filling extends beyond merely generating values for intervals without data. It encompasses the additional goal of recovering fracture density patterns capable of characterizing zones indicative of preferential fluid flow paths within the gaps.

4. Conclusions

The results show that direct sampling is suitable for fracture density gap filling, as an alternative for common interpolation, with the parametrization of few settings such as the acceptation threshold (t), maximal number of neighbors (n) and maximal fraction (f).

The summarized steps for the reconstruction of missing fracture density in well logs with direct sampling are: (i) selection of the training data set, (ii) enriching data using extreme value theory, (iii) running the simulation with enriched main variable for gap filling, and (iv) validating results with cross-validation and evaluating suitable scenarios with CRPS.

The training data set selection of two variables, fracture density P10 as the main variable and fracture occurrence being the auxiliary variable, is simple and showed good results for gap filling of a highly fractured interval of well-11. The auxiliary variable plays an important role as it guides pattern match. When more variables are added, the complexity in the training data set increases, although the results do not necessarily improve.

Resampling with extremes improve the reconstruction of missing data and could be further applied for various cases where gaps are due to extreme events. Values beyond the observed data range could be achieved. For the resampling with extremes a probability of exceedance (p_u) is also required.

Other than reservoir modeling, the proposed methodology could be extended for scanlines in structural geology and geothermal studies, and for the post processing and gap filling of various types of well logs and log-derived data of different settings.

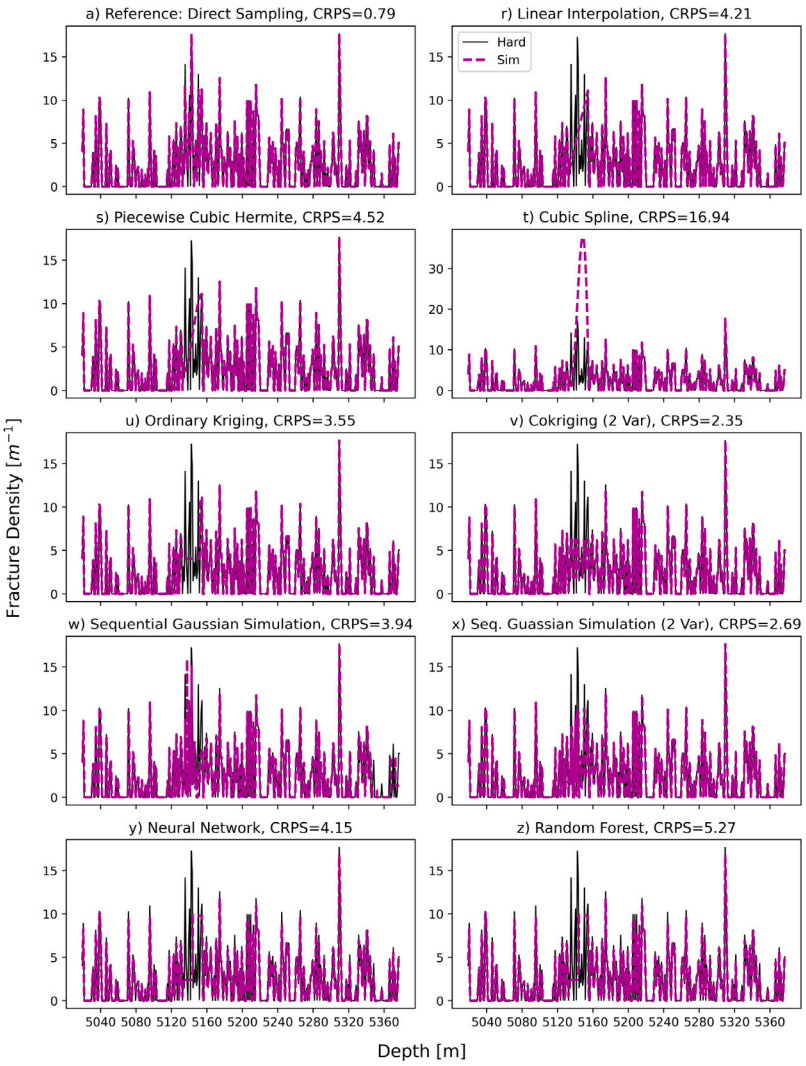


Fig. 12. Results from scenarios of fracture density P10 reconstruction with different techniques from well-11 at a zoom at the higher fractured interval, with the correspondent scores. The observed data is shown in black solid line and the simulations in magenta dashed line. Interpolation scenarios: (a) Reference scenario with direct sampling, (r) Linear interpolation, (s) Piecewise cubic hermite interpolation, (t) Cubic spline interpolation. Two point statistics scenario: (u) Ordinary kriging, (v) Cokriging, (w) Sequential guassian simulation, (x) Sequential guassian simulation with auxiliary variable. Machine learning scenarios: (y) Neural network, (z) Random forest.

For further application, the code and examples of the reconstruction of fracture density data with extremes are available in Python. [2]

CRediT authorship contribution statement

Ana Paula Burgoa Tanaka: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Philippe Renard:** Writing – review & editing, Validation, Supervision, Software, Methodology, Formal analysis, Conceptualization. **Julien Straubhaar:** Writing – review & editing, Validation, Supervision, Software, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ana Paula Burgoa Tanaka reports a relationship with Petrobras that includes: employment.

[2] https://github.com/randlab/xgapfilling.

Data availability

The original data used in this study are not available due to restrictions on sharing. However, a repository with examples using synthetic data is available in Python https://github.com/randlab/xgapfilling.

Acknowledgments

The authors acknowledge Petrobras for the permission to publish this study. Special thanks to Robin Voland for advice and review, Valentin Dall'Alba-Arnau and Ludovic Schorpp for the thorough review. Thanks to Alex Kobayashi, Alexis Neven, Celia Trunz, Jefter Caldeira, Olinto de Souza, Rita Pozzi, Rogério Brandi, Renato Kramberger, Saulo Pedrinha, and Carlos Eduardo Pereira. Thanks to the editors of the journal, Rodrigo Corrêa and anonymous reviewer for their comments and suggestions to improve the manuscript.

References

Bonnet, E., Bour, O., Odling, N.E., Davy, P., Main, I., Cowie, P., Berkowitz, B., 2001. Scaling of fracture systems in geological media. Rev. Geophys. 39, 347–383.
Carminatti, M., Dias, J., Wolff, B., 2009. From turbidites to carbonates: Breaking paradigms in deep waters. OTC Offshore Technol. Conf. All Days, 4–7.

Chugunova, T.L., Hu, L.Y., 2008. Multiple-point simulations constrained by continuous auxiliary data. Math. Geosci. 40, 133–146.
Churikov, N., Grafeeva, N., 2018. Recovering gaps in the gamma-ray logging method. arXiv preprint.
Davy, P., Le Goc, R., Darcel, C., 2013. A model of fracture nucleation, growth and arrest, and consequences for fracture density and scaling. J. Geophys. Res.: Solid Earth 118 (4), 1393–1407.
de Jesus, C.M., Martins Compan, A.L., Surmas, R., 2016. Permeability estimation using ultrasonic borehole image logs in dual-porosity carbonate reservoirs. Petrophys. - SPWLA J. Form. Eval. Reserv. Descr. 57 (06), 620–637.
Dershowitz, W., Herda, H.H., 1992. Interpretation of fracture spacing and intensity. In: 33rd U.S. Symposium on Rock Mechanics. USRMS 1992, Vol. All Days, pp. ARMA–92–0757.
Fernandez-Ibanez, F., Nolting, A., Breithaupt, C.I., Darby, B., Mimoun, J., Henares, S., 2022. The properties of faults in the Brazil pre-salt: A reservoir characterization perspective. Mar. Pet. Geol. 146, 105955.
Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc. Ser. B Stat. Methodol. 69, 243–268.
Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102, 359–378.
Guardiano, F.B., Srivastava, R.M., 1993. Multivariate geostatistics: Beyond bivariate moments. In: Soares, A. (Ed.), Geostatistics Tróia'92: Volume 1. Kluwer Academic Publications, Dordrecht, pp. 133–144.
Haan, L., Ferreira, A., 2006. Extreme Value Theory: An Introduction, vol. 3, Springer.
Journel, A.G., 1993. Geostatistics: roadblocks and challenges. In: Soares, A. (Ed.), Geostatistics Tróia'92: Volume 1. Kluwer Academic Publications, Dordrecht, pp. 213–224.
Juda, P., Renard, P., Straubhaar, J., 2020. A framework for the cross-validation of categorical geostatistical simulations. Earth Space Sci. 7, 17.
Lopes, R.L., Jorge, A.M., 2018. Assessment of predictive learning methods for the completion of gaps in well log data. J. Pet. Sci. Eng. 162, 873–886.
Mariethoz, G., 2018. When should we use multiple-point geostatistics? In: Handbook of Mathematical Geosciences: Fifty Years of IAMG. Springer International Publishing, pp. 645–653.
Mariethoz, G., McCabe, M.F., Renard, P., 2012. Spatiotemporal reconstruction of gaps in multivariate fields using the direct sampling approach. Water Resour. Res. 48, 13.
Mariethoz, G., Renard, P., 2010. Reconstruction of incomplete data sets or images using direct sampling. Math. Geosci. 42, 245–268.
Mariethoz, G., Renard, P., Straubhaar, J., 2010. The direct sampling method to perform multiple-point geostatistical simulations. Water Resour. Res. 46, 14.
Mimoun, J., Fernández-Ibáñez, F., 2023. Carbonate excess permeability in pressure transient analysis: A catalog of diagnostic signatures from the Brazil pre-salt. J. Pet. Sci. Eng. 220, 111173.
Nelson, R.A., 2001. Geologic Analysis of Naturally Fractured Reservoirs. Gulf Professional Publishing.
Opitz, T., Allard, D., Mariethoz, G., 2021. Semi-parametric resampling with extremes. Spatial Stat. 42, 100445.
Oriani, F., Borghi, A., Straubhaar, J., Mariethoz, G., Renard, P., 2016. Missing data simulation inside flow rate time-series using multiple-point statistics. Environ. Model. Softw. 86, 264–276.
Oriani, F., Straubhaar, J., Renard, P., Mariethoz, G., 2014. Simulation of rainfall time series from different climatic regions using the direct sampling technique. Hydrol. Earth Syst. Sci. 18, 3015–3031.
Ozkaya, S.I., Al-Fahmi, M., 2022. Estimating size of finite fracture networks in layered reservoirs. Appl. Comput. Geosci. 15, 100089.
Philip, Z.G., Jennings, J.W., Olson, J.E., Laubach, S.E., Holder, J., 2005. Modeling coupled fracture-matrix fluid flow in geomechanically simulated fracture networks. SPE Reserv. Eval. Eng. 8 (04), 300–309.
Reiss, R.-D., Thomas, M., Reiss, R., 1997. Statistical Analysis of Extreme Values, vol. 2, Springer.
Sanderson, D.J., Nixon, C.W., 2018. Topology, connectivity and percolation in fracture networks. J. Struct. Geol. 115, 167–177.
Shakiba, M., Lake, L.W., Gale, J.F., Pyrcz, M.J., 2022. Multiscale spatial analysis of fracture arrangement and pattern reconstruction using Ripley's K-function. J. Struct. Geol. 155, 104531.
Straubhaar, J., Renard, P., Chugunova, T., 2020. Multiple-point statistics using multi-resolution images. Stoch. Environ. Res. Risk Assess. 34, 251–273.
Tanaka, A.P.B., Borges, J.P.G., de Matos, G.C., Campos, M.T.R., Cunha, B.M., de Souza, R.B., Caldeira, J.N.d.M., de Oliveira, T.A.S., Marçon, D.R., Lima, A.P.M., 2022. Fault-related fracture modeling in a pre-salt lacustrine carbonate reservoir from Santos basin, offshore Brazil: Predicting preferential fluid flow paths using 3D geological and flow simulation models. Mar. Pet. Geol. 135, 7.
Tanaka, A.P.B., Faria, D.L.d.P., Gomes, J.P.B., de Souza, Jr., O.G., 2018. Geological characterization and modeling of an aptian carbonate reservoir in the Santos basin, Brazil. AAPG Search Discov. 11128, 11128.
Terzaghi, R.D., 1965. Sources of error in joint surveys. Geotechnique 15, 287–304.
Wennberg, O., Ramalho, F.D.O., Mafia, M.V., Lapponi, F., Chandler, A., Cartesio, L.G., Hunt, D.W., 2023. The characteristics of natural open fractures in acoustic borehole image logs from the pre-salt Barra Velha formation, Santos basin, Brazil. J. Struct. Geol. 167, 104794.