Report on implementation and validation protocol for EGS and SHGS

D 5.12

Report on implementation and validation protocol for EGS and SHGS

D 5.12

Responsible author: Gianluca Gola (CNR)

Responsible SP leader: Gylfi Páll Hersir (ÍSOR)

Responsible WP leader: Gylfi Páll Hersir (ÍSOR)

Contributions by: Gianluca Gola (CNR), Adele Manzella (CNR), Eugenio Trumpy (CNR), Alessandro Santilano (CNR), Gylfi Páll Hersir (ÍSOR), Ásdís Benediktsdóttir (ÍSOR)

Work package 5.4

Website: http://www.gemex-h2020.eu

The GEMex project is supported by the European Union's Horizon 2020 programme for Research and Innovation under grant agreement No 727550 and the Mexican Energy Sustainability Fund CONACYT-SENER, project 2015-04-268074

Table of Contents[1]

[1] The content of this report reflects only the authors' view. The Innovation and Networks Executive Agency (INEA) is not responsible for any use that may be made of the information it contains.

List of figures

List of tables

Executive summary

The objective of Deliverable D5.12 within the GEMex Project is: *Report on implementation and validation protocol for EGS and SHGS. This report concludes WP5 activities, retrieving all the results achieved in Task 5.1, 5.2, 5.3, 5.4 and defining a protocol, i.e. a set of procedures to be followed for data integration in geothermal exploration of EGS and SHSG.*

This report describes a procedure aimed at performing an effective integration from different geophysical methods providing an unambiguous, self-constrained model of a geothermal system. In this framework, cross-plotting and clustering procedures are considered as a promising auxiliary tool towards the integration of distinct geophysical datasets. Cluster analyses are applied to the Los Humeros and Acoculco test sites in Mexico and in two geothermal fields in Europe.

1 Introduction

The search for geothermal and tools of search requires innovative concepts for attaining near-term and long-term EGS and SHGS goals and objectives. Resource characterization regards research in geothermal gradients and heat flow; geological structure, including lithology and hydrogeology; tectonics; induced seismicity potentials. Reservoir design and development includes research in fracture mapping and in-situ stress determination; prediction of optimal stimulation zones. Reservoir operation and maintenance includes research in reservoir performance monitoring through the analysis of temporal variation of reservoir properties.

Two main goals are addressed by exploration and investigation: to reduce the mining risk by cutting the exploration cost and increasing the probability of success in identification of EGS and SHGS in prospective areas, and to provide all necessary subsurface information to guarantee the best exploitation efficiency, the sustainability of the resource and the lowest possible environmental impact. Technological challenges targeted to these goals are mainly aimed to: 1) find improved and newly developed methodologies able to map reservoir condition suitable for exploitation, in particular at local scale; 2) provide data integration (static and dynamic) and uncertainty analysis. Integration of technology and multidisciplinary evaluation of data are, therefore, required as a core competency in the geothermal world. Each single petrophysical parameter retrieved by geophysical surveys provides important relations with lithology, fluid saturation and phase and underground physical conditions (Table 1).

Table 1: Degree of relationship between petrophysical parameters (horizontal) and some main exploration targets (vertical).

		Density	Magnetic susceptibility	Electrical resistivity	Seismic velocity (Vp, Vs)	Temperature
	Porosity					
	Water content					
	Fluid phase					
Alteration effect	*Clay content*					
	Magnetic mineral content					
	Metallic mineral content					
	Mechanical properties					
	Subsurface structure					
	Underground temperature distribution					
	Strong	**Moderate**	**Weak**	**None**		

A multivariate approach, i.e. integrating multiple datasets, reinforce the interpretation and provide additional, unequivocal information. For example, since water content is imaged well by both

electrical resistivity and seismic velocity, the combination of magnetotelluric and seismic data provide independent constraints for imaging volumes rich of geothermal fluids.

There are different way to integrate geophysical dataset. In the frame of GEMex D5.10 the integral visualization of the various dataset acquired within the Project have been discussed. Here we describe another way to analyse and query the data, exploring an approach that allows a semi-automatic integration: the cluster procedures of high-dimensional data into homogeneous subgroups. Clustering techniques proved useful in the data analysis in Earth Sciences, e.g. in oil & gas, mining and geothermal exploration. The techniques of clustering have been recognized at least since the 1960s; however, since the 1990s the method has found widespread applications in the field of molecular biology as a way to recognize patterns of gene expression in DNA data (Eisen, et al., 1998). In the field of exploratory data mining few examples exist in literature, e.g. (Lindsey, et al., 2018) (Di Giuseppe, et al., 2018). We critically explored the pre-processing actions and the applicability of clustering techniques tacking advantage of the availability of multiple geophysical dataset resulting from the WP5 activities. The analysis is performed using two separate survey scales enabling the extraction of regional and local geophysical features.

2 Cross-plot and cluster analysis methods

2.1 Definitions

Cross-plot is synonym for scatter plot used primarily in the Earth Sciences to describe a specialized chart that compares multiple measurements made at a single location along two or more axes. The axes of the plot are commonly linear, but may also be logarithmic. Cross-plots are used to interpret geophysical data; they can suggest various kinds of correlations between variables with a certain confidence interval. Correlations may be positive (rising), negative (falling), or null (uncorrelated). In the case of Bi-Variate data, a line of best fit can be drawn in order to study the relationship between the two variables.

Beside the scattered visualization of the data, a similar approach consists in the Density plot. It can be viewed as a generalization of the histogram and for a Bi-Variate dataset, it becomes a 3D histogram. A Density plot requires the choice of an anchor point and bin widths for the variables in order to define a binning grid. In case of a Bi-Variate dataset, the third axis on the 3D histogram reports the number of occurrences in each cell grid. The Density plot investigates the properties of a given dataset and can give valuable indications of such features such as multimodality or clustering in the data.

Cluster analysis or clustering is an ensemble of techniques aimed at classify a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense) to each other than to those in other groups (clusters). The basic properties of these clusters are:

1. All the data points in a cluster should be similar to each other.
2. The data points from different clusters should be as different as possible.

Many different approaches to the clustering problem have been developed and we explored few clustering procedures. To give an overview, inevitably limited, of the different methods, we categorize them in two principal groups: i) unsupervised methods and ii) supervised methods.

Unsupervised clustering methods can be divided into two additional subgroups: 1a) hard clustering, in which each data point either belongs to a cluster completely or not (e.g. the *K-mean* algorithm); and 1b) soft clustering, in which instead of putting each data point into a separate cluster a probability or likelihood of that data point to be in those clusters is assigned (e.g. the *Gaussian Mixture Model*). In clustering, we do not have a target to predict. We look at the data and then try to group similar observations and form different groups. Hence it is an unsupervised problem. Unsupervised methods are useful when we don't know the right answer ahead of time. Then, how to choose the number of components *k*? If we choose *K* to small, we under fit the data, whereas if we choose it too large, we can over fit.

In the supervised methods, predefined classes are assigned by properties. In machine learning and statistics, classification is a supervised learning approach in which the computer program learns

from the data input given to it and then uses this learning to classify new observations. It is a two-step process, comprised of a learning step and a classification step. In the learning step, a classification model is constructed (e.g. *Decision Tree learning method*) and, subsequently, the latter is used to predict the classes for a given multivariate dataset.

2.1.1 K-mean Model

One of the first algorithms aimed at finding clusters in a set of data points is the non-probabilistic technique called the *K-means* algorithm (Lloyd, 1982). The main goal is to partition the dataset into some number *k* of clusters, where the value of *k* is given. Intuitively, a cluster comprises a group of data points whose inter-point distances are small compared with the distances to points outside of the cluster.

The *K-mean* algorithm assumes each data as a location point in the Euclidean space. The dimension of the Euclidean space corresponds to the number of geophysical parameters, $n = \{1,2,...,D\}$. Given a point **p** of coordinates $\{x_1, x_2, ..., x_n\}$ and a cluster **c** with centroid $\{\mu_1, \mu_2, ..., \mu_n\}$, the Euclidean distance is:

(1) $$d(\mathbf{p},\mathbf{c}) = \sqrt{\sum_{i=1}^{n}(c_i - p_i)^2}$$

The *K-mean* algorithm works in the following way (see also Figure 1 for a visual illustration):

1. K centroids $\boldsymbol{C_i}$ are created randomly (based on the predefined value of *k*)
2. *K-means* allocates every data point in the dataset to the nearest centroid (minimizing the Euclidean distances between them), meaning that a data point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other centroid
3. Then *K-means* recalculates the centroids by taking the average of all the observations assigned to that centroid's cluster to obtain K new centroid locations
4. The algorithm iterates between steps 2 and 3 until some criteria is met, e.g. the sum of distances between the data points *x* and their corresponding centroid μ is minimized:

(2) $$\arg\min_{C} \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

Nevertheless, *K-means* presents some disadvantages. One is the *a priori* definition of the number of clusters. Moreover, the results depend on the initial random conditions and they may not be comparable. Furthermore, the boundaries between the *K-means* clusters are linear, which means that this method fails for more complicated boundaries. One notable feature of the *K-means* algorithm is that at each iteration, every data point is assigned uniquely to one, and only one, of the clusters (hard clustering). Whereas some data points will be much closer to a particular center μ_k than to any other center, there may be other data points that lie roughly midway between cluster

centers. In the latter case, it is not clear that the hard assignment to the nearest cluster is the most appropriate.

2.1.2 Gaussian Mixture Model

The *Gaussian Mixture Model* (GMM) is a probabilistic, unsupervised clustering method. In general, for a set of observations **x** in the *n*-dimensional sample space, the Multi-Variate Gaussian probability density function is defined by the mean μ and the covariance matrix Σ:

(3) $$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\Sigma^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}.$$

Formally, a GMM is defined by the sum of several gaussians, each identified by the index $k \in \{1,\ldots,K\}$ where K is the predefined number of clusters, the mean μ that defines its center, the covariance matrix Σ that defines its width and the mixture proportion α (or mixing coefficient) that represent the probability that an observation point belongs to the k^{th} mixture component. Figure 2 illustrates the example of a mixture model for a Uni-Variate (n = 1) Gaussian distributions with k = 3. For the Uni-Variate case, the covariance matrix Σ simplifies to the variance σ^2 (or standard deviation σ).

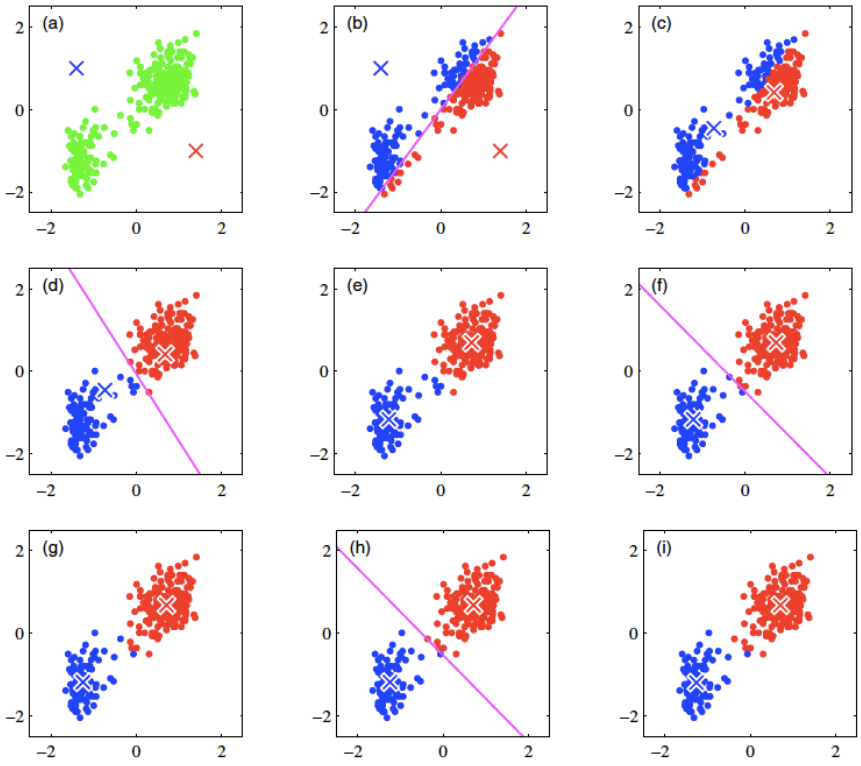

Figure 1: Illustration of the K-means algorithm, modified from (Bishop, 2006). (a) Green points denote the data set in a two-dimensional Euclidean space. The initial choices for centres μ1 and μ2 are shown by the red and blue crosses, respectively. (b) In the initial step, each data point is assigned either to the red cluster or to the blue cluster, according to which cluster centre is nearer. This is equivalent to classifying the points according to which side of the perpendicular bisector of the two cluster centres, shown by the magenta line, they lie on. (c) In the subsequent step, each cluster centre is re-computed to be the mean of the points assigned to the corresponding cluster. (d)–(i) show successive steps through to final convergence of the algorithm.

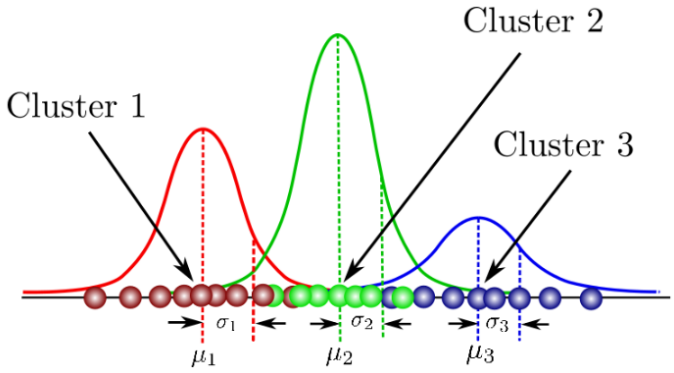

Figure 2: Graphical representation of the Uni-Variate GMM with *k* = 3 together with the parameters *μ* (mean) and *σ* (standard deviation) of each Gaussian function, modified from (Bishop, 2006).

The GMM consist of *k* mixing components each of which corresponds to a Gaussian distribution and is defined as follow:

(4) $$p_{GMM}(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k \; p(x \mid \mu_k, \Sigma_k)$$

The mixture proportions α are probabilities and they must meet the condition:

(5) $$\sum_{k=1}^{K} \alpha_k = 1$$

In the cluster analysis, given a GMM, the goal is to find for each Gaussian function the parameters $\theta_k = \{\mu_k, \Sigma_k, \alpha_k\}$ which maximize the likelihood function *L*. The likelihood function measures the support provided by the data for each possible value of the parameters ϑ_k (μ_k, Σ_k and α_k). If we compare the likelihood functions at N points and find that $L(\vartheta_1|\mathrm{x}) > L(\vartheta_2|\mathrm{x})$ then ϑ_1 is a more plausible value for ϑ than ϑ_2. The maximum likelihood is found using the expectation-maximization algorithm or EM algorithm.

The GMM EM algorithm works as follow (see also Figure 3 for a visual illustration):

1. First it chooses some initial values for the means, covariances, and mixing coefficients.
2. In the expectation step, or E step, it uses the current values for the parameters to evaluate the *a posteriori* probabilities given by:

(6) $$\gamma(z_{nk}) = \frac{\alpha_k \, p(x_n \mid \mu_k, \Sigma_k)}{\sum_{i=1}^{K} \alpha_i \, p(x_n \mid \mu_i, \Sigma_i)}$$

3. In the maximization step, or M step, it uses the current probabilities to re-estimate the means, covariance matrices, and mixing coefficients:

(7) $$\mu_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n$$

(8) $$\Sigma_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(x_n - \mu_k^{\text{new}}\right)\left(x_n - \mu_k^{\text{new}}\right)^T$$

(9) $$\alpha_k^{\text{new}} = \frac{N_k}{N} \text{ where } N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

4. Then, it evaluates the log likelihood:

(10) $$\ln p(\mathbf{x}\,|\,\mu,\Sigma,\alpha) = \sum_{n=1}^{N}\ln\left[\sum_{k=1}^{K}\alpha_k p(x_n\,|\,\mu_k,\Sigma_k)\right]$$

and check for convergence. The algorithm iterates between steps 2 - 4 until the log of the likelihood function falls below some threshold.

One notable feature of the GMM is that at each iteration of the EM algorithm, every data point is not assigned uniquely to one of the clusters. Instead, the probabilities that a data point falls in each cluster are evaluated (soft clustering). Few problems arise in the application of unsupervised mixture models. How to initialize the clusters? This is a tricky point. There's no strategy that is guaranteed to work, but one good option is to initialize the different clusters to have random means and very broad standard deviation. Another option is to initialize the cluster assignments using the *K-mean* algorithm for the centroids, to assume a shared and symmetric covariance matrix and to assign mixture coefficients equally distributed.

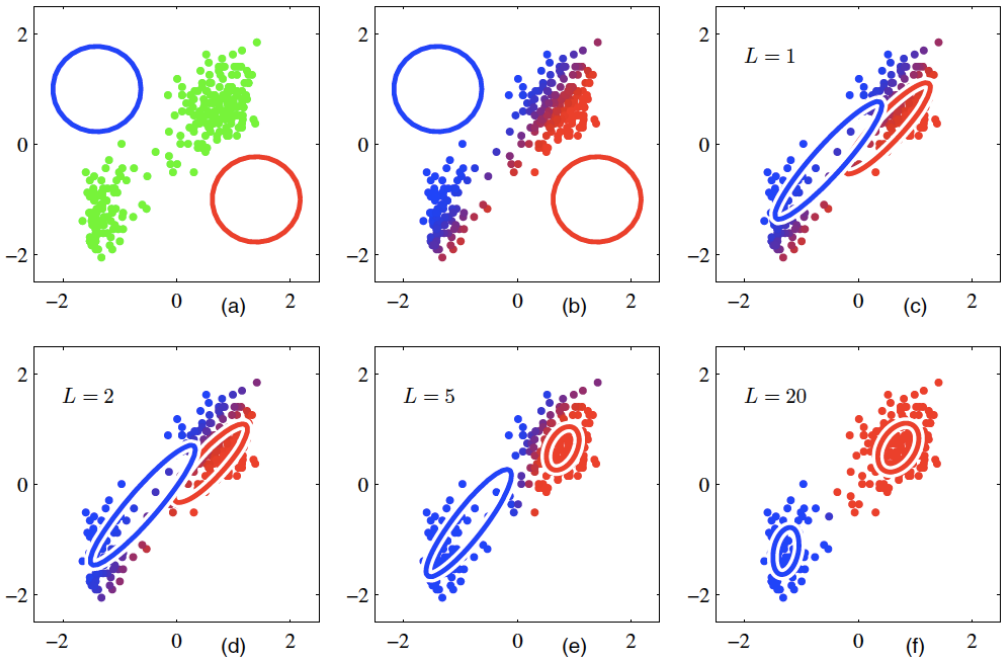

Figure 3: Illustration of the EM algorithm applied to the Bi-Variate GMM, modified from (Bishop, 2006). (a) Green points denote the data set in a two-dimensional Euclidean space. The initial choices for centroids μ_1 and μ_2 are shown by the red and blue circles, respectively. As initial conditions the covariance matrices Σ_1 and Σ_2 are shared and symmetric, the mixing coefficients α_1 and α_2 are evenly allocated. (b) In the E step, the posterior probabilities are evaluated at the observation points for the initial values. The color's tone represents the value of the responsibilities $\gamma(z_{nk})$ associated with data point x_n, obtained by using proportions of red and blue given by $\gamma(z_{nk})$ for k = 1, 2, respectively. (c) In the M step, the means, the covariances and the mixing coefficients are re-computed using the current responsibilities and the likelihood (L) is evaluated. (d)–(f) show successive steps through to final convergence of the algorithm.

2.1.3 Decision Tree Model

The Decision tree is one of the predictive modelling approaches used in statistics, data mining and machine learning. It is a type of supervised learning algorithm (having a predefined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, the population or sample is split into smaller and smaller subsets (or sub-populations) based on most significant decisions which are incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches and a leaf node represents a classification or decision.

The Decision tree algorithm works as follow (see also Figure 4 for a visual illustration):

1. A Training Dataset is given with some feature variables and classification output.
2. Determine the "best feature" in the dataset to split the data on. In classification settings, the split point is defined so that the population in subpartitions are pure as much as possible. In this step the Gini Index is used as the cost function when the rules are defined:

(11) $$Gini(t) = 1 - \sum_{i=1}^{c} \left[p(i \mid t) \right]^2$$

The Gini Index is a score that gives an idea of how good a split is by how mixed the classes are in the two labelled groups created by the split at the node *t*. Here *p* is the relative frequency of class *i* at node *t*. A perfect separation results in a Gini score of 0, when all records belong to one class implying most interesting information. The worst case split results when records are equally distributed among all classes.

3. Recursively generate new tree nodes by using the subset of data created from step 2. The splitting is repeated until the points are classified with the maximum accuracy while minimising the number of nodes.

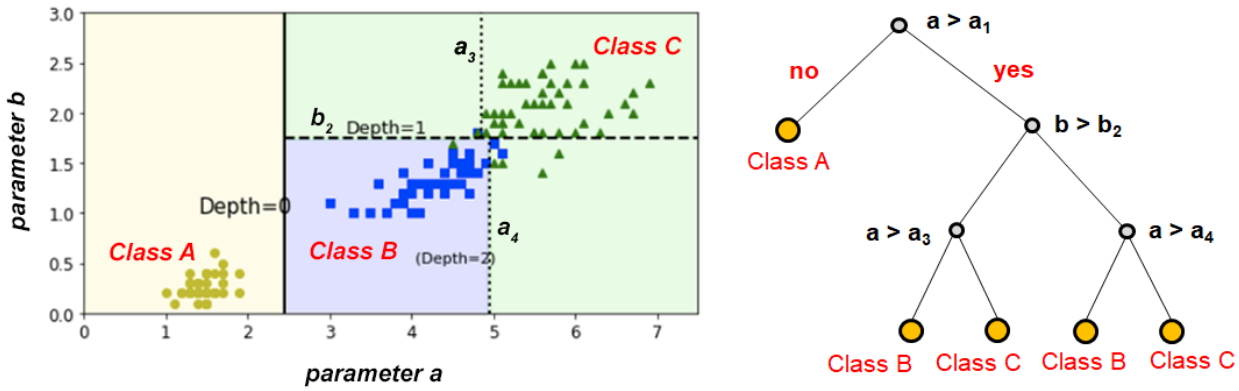

Figure 4: Illustration of the Decision Tree algorithm. Starting from an initial dataset (left) with 2 observed parameters for each data point, a tree structure is built as a sequence of questions at the splitting points or nodes. The answers (branches) determine what the next question is if any data point exists and so on until the decision tree reaches a suitable structure to classify the dataset. The resulting tree plot (right) shows the possible splitting rules that can be used to effectively predict the classes and then assign a correct classification label to the points.

2.2 Workflow

2.2.1 Data interpolation and Q-Q plots

There are two fundamental requirements to perform the cluster analysis. The first is the availability of datasets imaging the spatial distribution of at least two different petrophysical properties, e.g. resistivity, density, Vp, Vs, magnetization. The second is the definition of such models on the same grid. When both the requirements are fulfilled, each node of the grid has two or more observations which can be jointly compared and analysed.

Geophysical methods play a crucial role because they are able to image the continuous distribution of specific physical properties in the underground. Nevertheless, the resolution and extension of such models strictly depend on the location of the acquisition stations as well as on the adopted inversion technique. Very often, the solutions are computed on irregularly spaced nodes defining a grid which becomes coarser toward the lateral and bottom boundaries. As each geophysical model has its own grid (with the only exception for the joint inversions), an appropriate pre-processing

aimed at sampling the original datasets on a shared grid is a crucial aspect. In order to obtain a regularly spaced distribution of values, a triangulation-based linear interpolation is applied.

Usually, the original datasets are oversampled, i.e. the shared grid has a greater resolution. Oversampling (or the opposite and roughly equivalent under sampling) introduces a bias due to the generation of new values. With the aim to evaluate the suitability of the new interpolated datasets a graphical method for comparing the probability distributions of the original and the interpolated values is applied. For this purpose, we applied quantile-quantile (Q-Q) plots for similarity check between original and interpolated distributions. The Q-Q plot compares a dataset to a theoretical model (normal distribution) by plotting their quantiles against each other. In Figure 5 the comparison between the statistical distributions of the original and interpolated data of the regional density and magnetization models in Los Humeros area displayed. We can observe a marked similarity of the two distributions in each Q-Q plot which ensures that the interpolation procedure has retained the original statistical distribution, without introducing artefacts.

2.2.2 Cluster analysis

Once two or more interpolated datasets are available, we are ready to proceed with the cluster analysis. In most applications, the number of clusters (or components) *k* is unknown. The joint visualization of the data by cross plotting does not ensure the detection of clusters and how many components occur. First, the application of an unsupervised method is recommended. Here, the GMM is chosen and the algorithm is running with an increasing number *k*. Given some models, one way to choose the best one is by comparing information criteria, i.e. a measure of the quality of a statistical model which considers how well the model fits the data and the complexity of the model. The most popular information criteria are the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC). AIC and BIC are a likelihood-based measures of the model fit that include the negative loglikelihood (NlogL) and a penalty for the number of estimated parameters (*p*) and observations (n):

(12) $$AIC = 2\,\mathbf{NlogL} + \mathbf{2}\,p$$

(13) $$BIC = 2\,\mathbf{NlogL} + p\log(n)$$

When comparing multiple models, the one with a smaller value of AIC or BIC is better. Nevertheless, the challenge is to find the minimum number of components that will capture the essential patterns in the data. In Figure 6 the trends of the AIC and BIC values as function of the increasing number of components are displayed for the GMM clustering applied to the Regional Density and Magnetization interpolated dataset. AIC and BIC decrease fast when *k* increases from 1 to 4, next a clear change in the slopes occurs. The model with 4 components is the one that better explain the data with the lowest number of clusters.

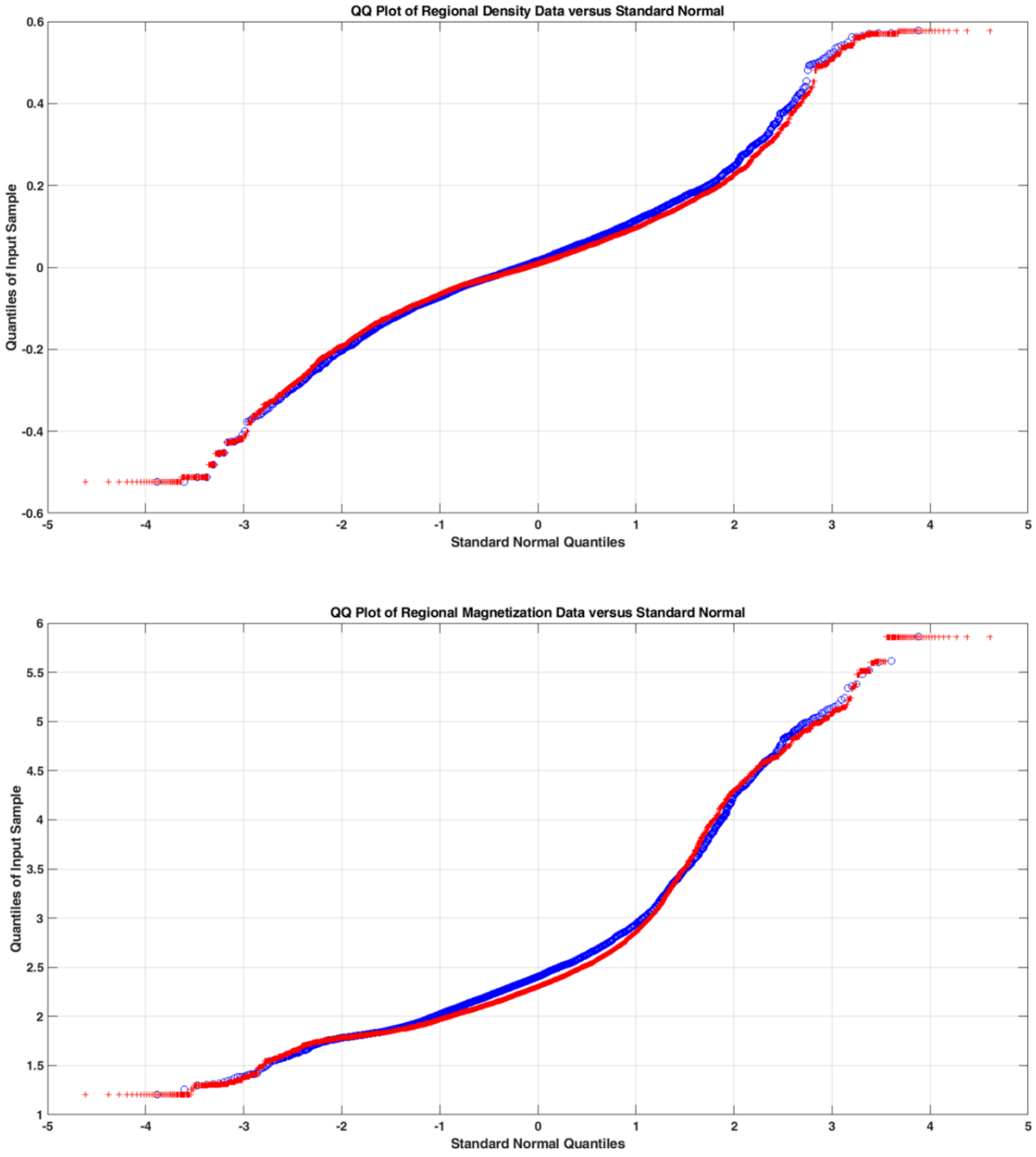

Figure 5: Q-Q Plots of Regional Density (upper) and Regional Magnetization (lower) dataset. The statistical distribution of the original data (blue circles) is compared with the statistical distribution of the new interpolated data (red crosses).

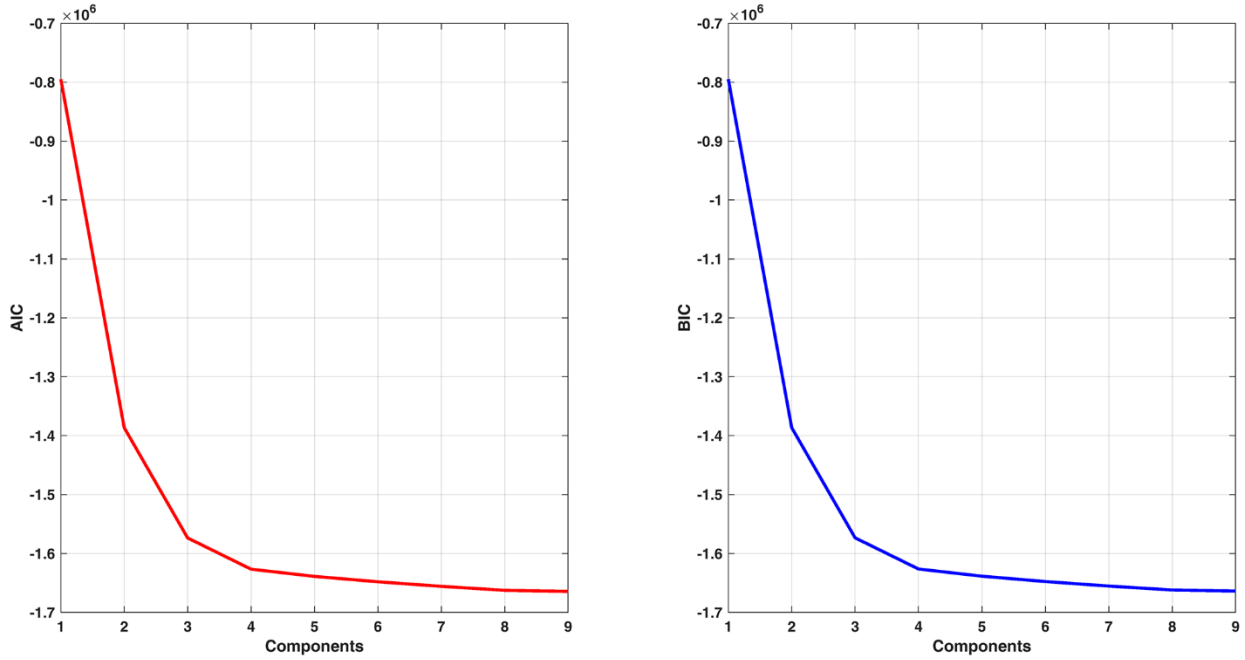

Figure 6: Akaike Information Criterion (AIC, left) and Bayes Information Criterion (BIC, right) evaluated for the GMM with k variable from 1 to 9 applied to the Regional Density and Magnetization interpolated dataset.

2.2.3 Visualization

In the framework of *Task 5.4.7: Advanced 3D model integration of geoscientific data into a conceptual model for Los Humeros and Acoculco* (GEMex, 2020), it was decided to use Paraview as visualization tool of the 3D geoscientific data. In the same way, the results of clustering are imaged in Paraview enabling a direct comparison between the 3D clusters, the geophysical anomalies and the geological/structural models.

3 Cluster analysis of Los Humeros geothermal field

3.1 Geophysical datasets

For the detection of the main geological and geophysical features of Los Humeros geothermal field we take advantage of the availability of different 3D geophysical models computed in the framework of the WP5 (Task 5.1, Task 5.2 and Task 5.3) and of the 3D geological model figured in the framework of WP3 (Task 3.1). The datasets used for the cluster analysis are summarised in Table 2. The areal extension of the different geophysical models together with the modelled faults are shown in Figure 7.

Table 2: List of the geological and geophysical data used in the cluster analysis of Los Humeros area.

Data type	Short description	Partner	Ref.
Local scale			
Geological model	Geological faults and units	BRGM	(Calcagno, et al., 2018)
Resistivity	Resistivity from 3D MT inversion	ISOR	(GEMex, 2019a)
Density	Density contrast from 3D grav inversion	KIT (INE)	(GEMex, 2019b)
Velocity model	Vp and Vp/Vs from seismic tomography	GFZ	(GEMex, 2019c)
Regional scale			
Density	Density contrast from joint 3D grav-mag inversion	CICESE	(Carrillo, et al., 2020)
Magnetization	Magnetization from joint 3D grav-mag inversion	CICESE	(Carrillo, et al., 2020)

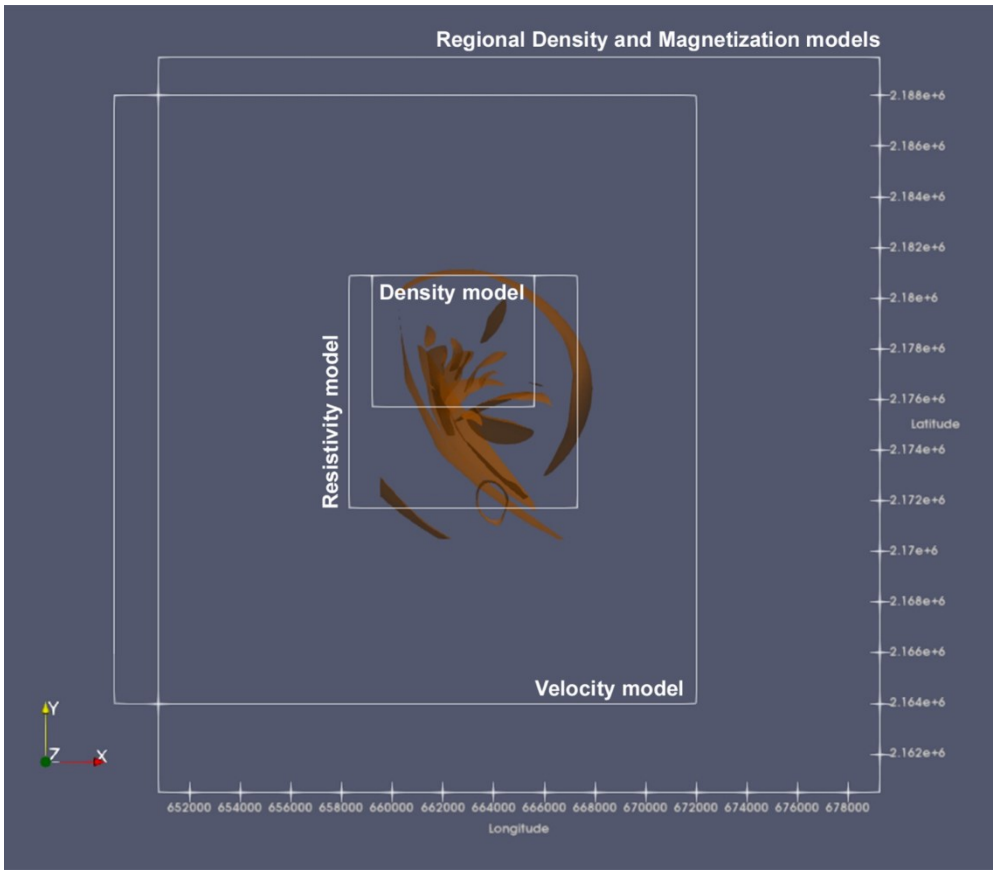

Figure 7: Map view (XY plane) showing the external boundaries of the regional density and magnetization models, local velocity, resistivity and density models together with the structural features of the geological model in Los Humeros area.

3.2 Cross-plots and Density plots

In the following sections we present the results obtained from the application of the GMM to the data coming from i) the regional joint inversion of gravity and magnetic data and ii) the local resistivity and Vp/Vs distributions resulting from MT and earthquake tomography inversions, respectively. In Figure 8 the cross-plot and density plot of the regional density vs magnetization data are displayed. Density is expressed in terms of density contrast against a reference value of 2.67 g/cm^3. The density contrast values range between – 0.55 and + 0.55 g/cm^3. The rock magnetization values fall in the interval 1.25 – 6 A/m. The density plot highlights a principal cluster centred around 0 g/cm^3 and 2 A/m. A subordinate cluster locates in the upper part of the plot and appears characterized by a negative correlation between density and magnetization with high magnetization and negative density contrast.

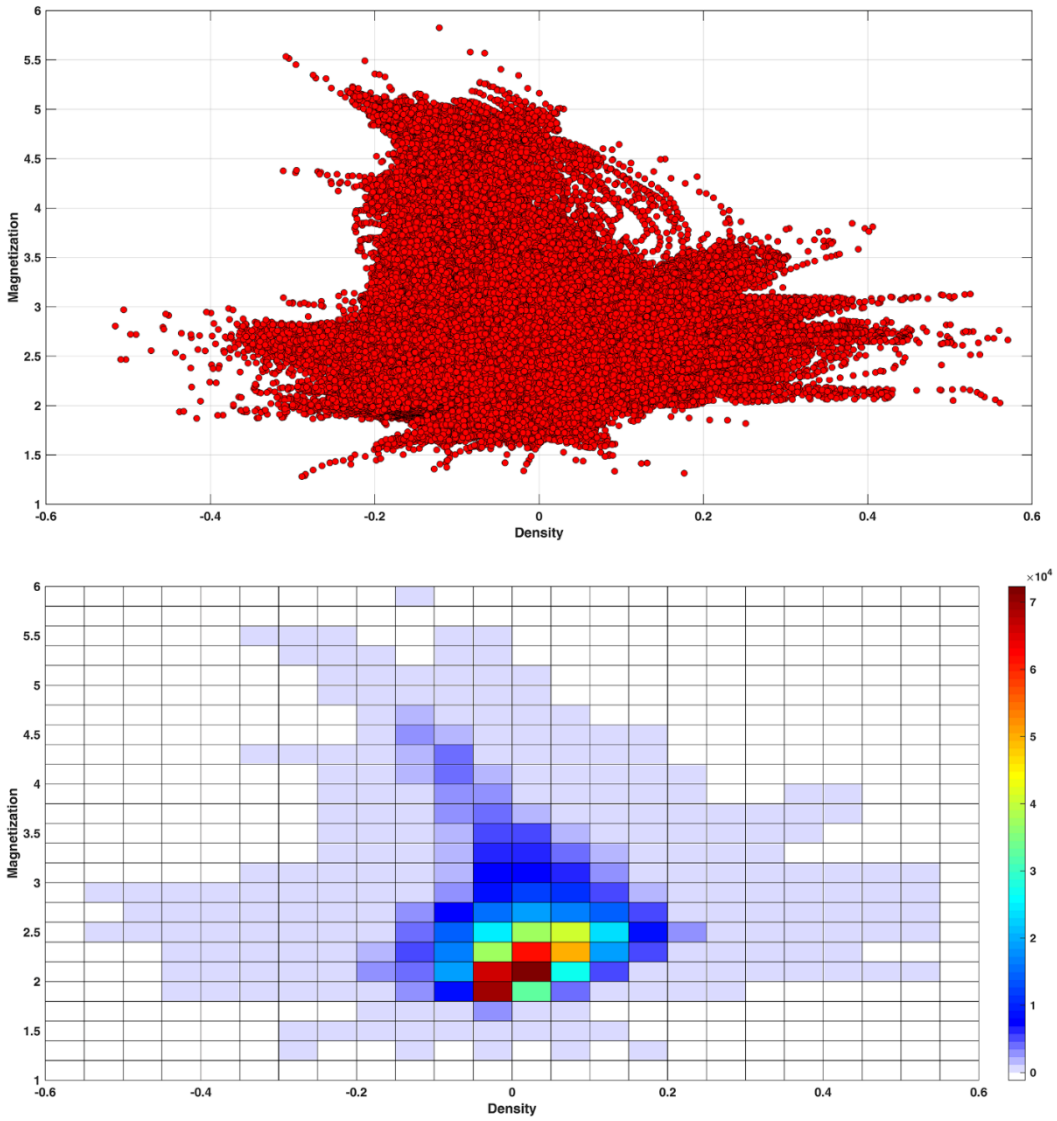

Figure 8: Cross-Plot (upper) and Density-Plot (lower) of the interpolated Regional Density-Magnetization dataset.

In Figure 9 the cross-plot and density plot of the local resistivity vs Vp/Vs data are displayed. Resistivity is expressed in logarithmic form and the values range between – 0.9 and + 4.2 $\log_{10}(\Omega$ m). The Vp/Vs values fall in the interval 1.42 – 1.85. The density plot highlights a principal cluster centred at about (100 Ω m and 1.52). For this dataset, a roughly positive correlation between resistivity and Vp/Vs values exists. The values of Vp/Vs above 1.75 are no longer continuously distributed and take discrete values of 1.775, 1.8, 1.825 and 1.85. This distribution depends on the maximum depth resolved by the seismic tomography which, in turn, is controlled by the wave paths from the wave sources (earthquakes) to the seismometers at surface. Locally, the initial velocity model (a horizontally layered velocity model) doesn't change due to the lack of seismic rays.

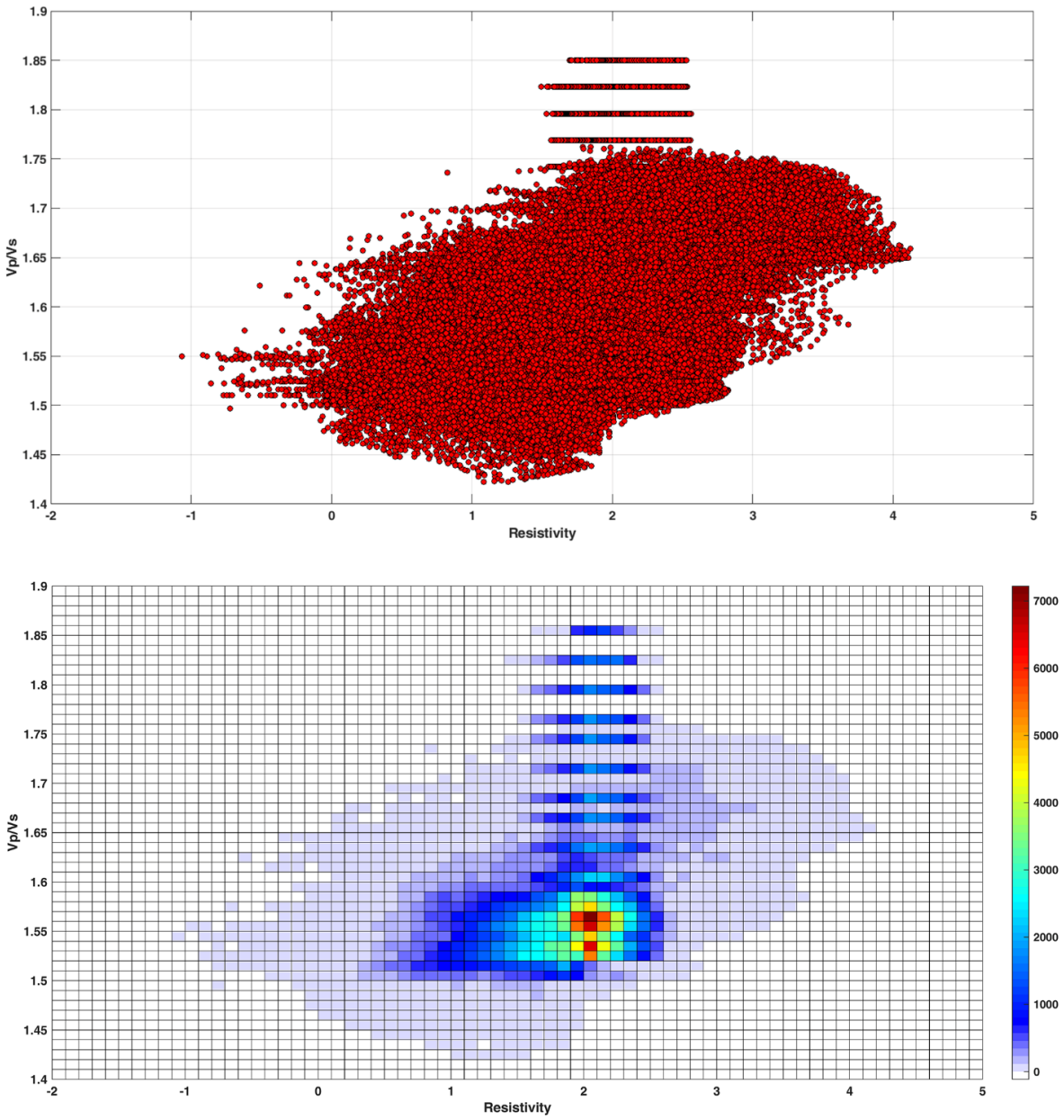

Figure 9: Cross-Plot (upper) and Density-Plot (lower) of the interpolated Local Resistivity-Vp/Vs dataset.

3.3 Unsupervised clustering

3.3.1 Regional model

We took advantage of the availability of the 3D Joint Inversion of Gravity and Magnetic Data in Los Humeros to test the GMM at regional scale. In Figure 10 the cluster distributions evaluated with the GMM for a number of components variable between 1 and 9 are displayed. In order to choose the best fitting model having the minimum number of components we used the information criteria reported in Figure 6. The model with four components is that one capable to capture the main features of the data using the minimum number of clusters. In Figure 11 the statistical distribution of the values of each cluster is reported for the above-mentioned solution.

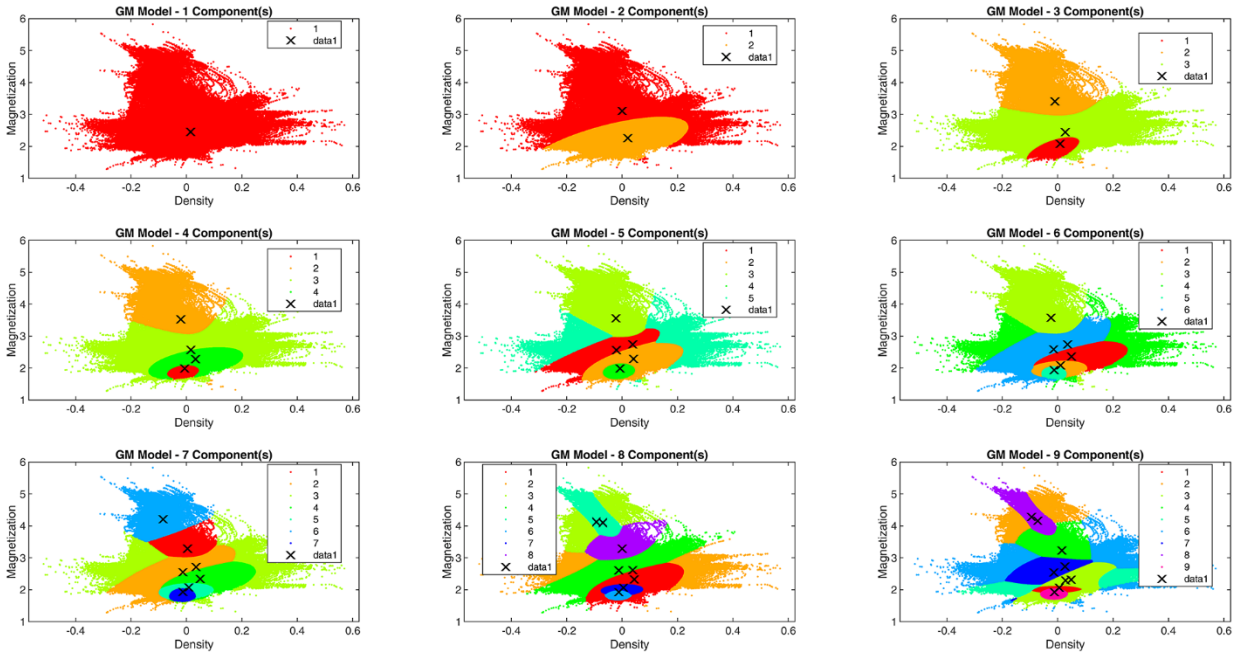

Figure 10: Unsupervised clustering using the GMM with k variable from 1 to 9 applied to the Regional Density and Magnetization interpolated dataset.

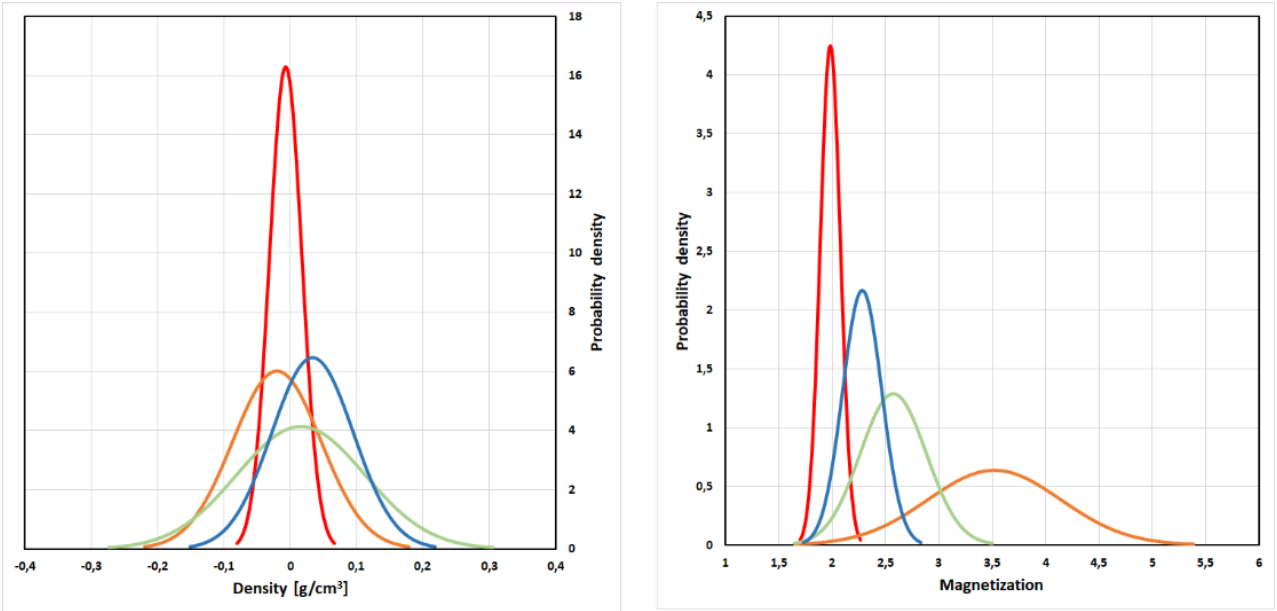

Figure 11: Statistics of clusters (k = 4) of the Regional Density (left) and Magnetization (right) interpolated dataset.

In Figure 12 the 3D spatial distribution of the clusters is displayed. The comparison between the main geological structures coming from Task 3.1 and the spatial location of the clusters along two cross-sections is presented in Figure 13.

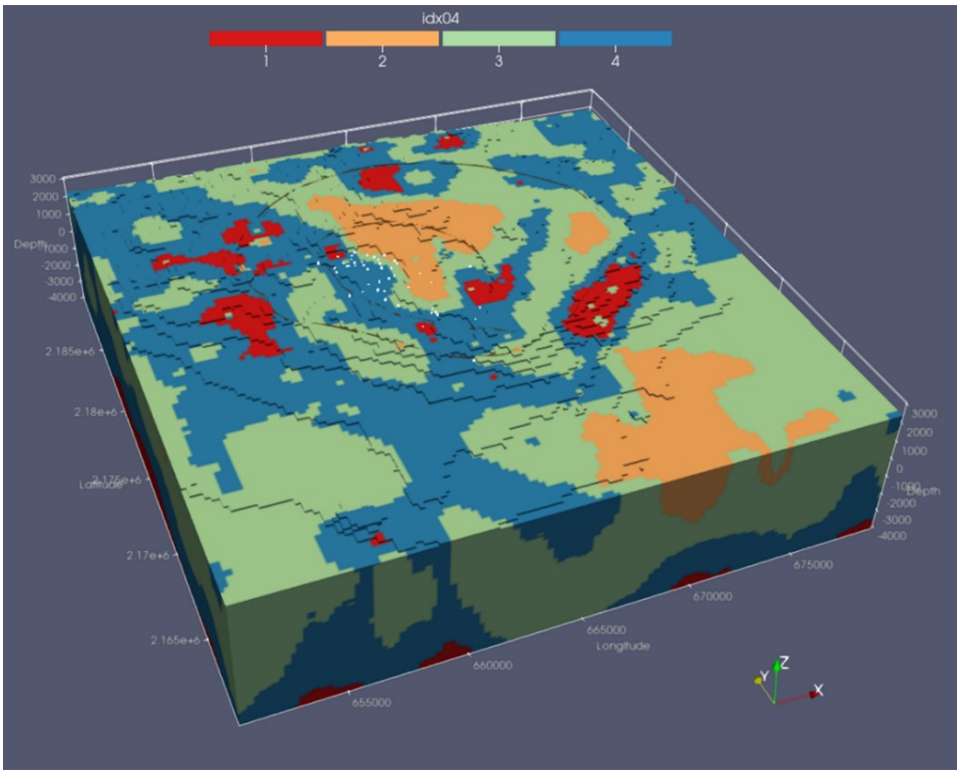

Figure 12: 3D visualization of clusters (k = 4) computed for the Regional Density and Magnetization interpolated dataset.

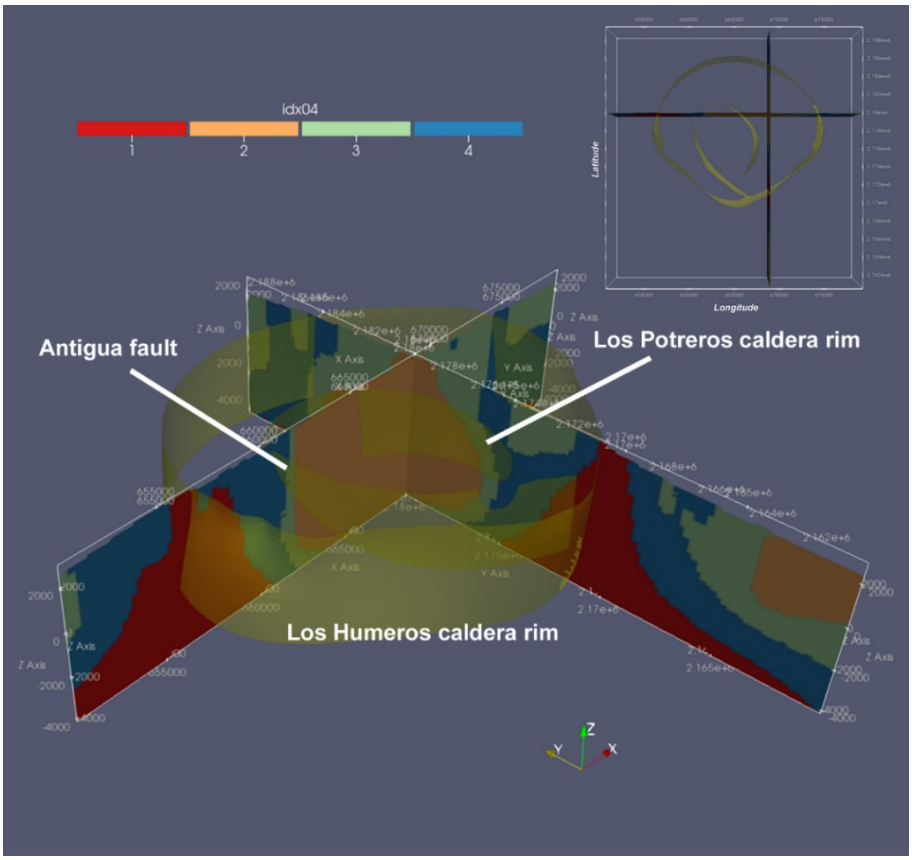

Figure 13: Visualization of clusters (k = 4) along two selected E-W and N-S cross sections for the Regional Density and Magnetization interpolated dataset together with the main structural features.

Two clusters present an interesting pattern and characteristic values of density and magnetization. We refer to the cluster 1 and cluster 2 with density and magnetization mean values of 2.66 g/cm^3, 2.65 g/cm^3 and 1.98 A/m, 3.52 A/m, respectively.

According to (Carrillo, et al., 2020) the gravity anomaly appears more diffuse, instead the magnetic anomaly is clearly delimited, in particular at depth. Along the fault zones, comparably low magnetizations between 1-2 A/m are dominant. The cluster 1 (Figure 14) mimics the Los Humeros caldera rim with characteristic two structural highs approaching the surface in the western and south-eastern portions of the caldera. The cluster 1 results bounded (Figure 13) by the cluster 4 having the highest density (mean value of 2.70 g/cm^3) and medium-low magnetization (mean value of 2.28 A/m). A further interesting structure is that one imaged by the cluster 2 (Figure 15) with mean medium-low density (2.65 g/cm^3) and the highest magnetization (mean 3.52 A/m and maxima values > 5 A/m). The cluster 2 highlights two magnetized bodies, one inside the Los Humeros caldera rim and intersecting the Los Potreros caldera rim, the other one outside the caldera area to the south.

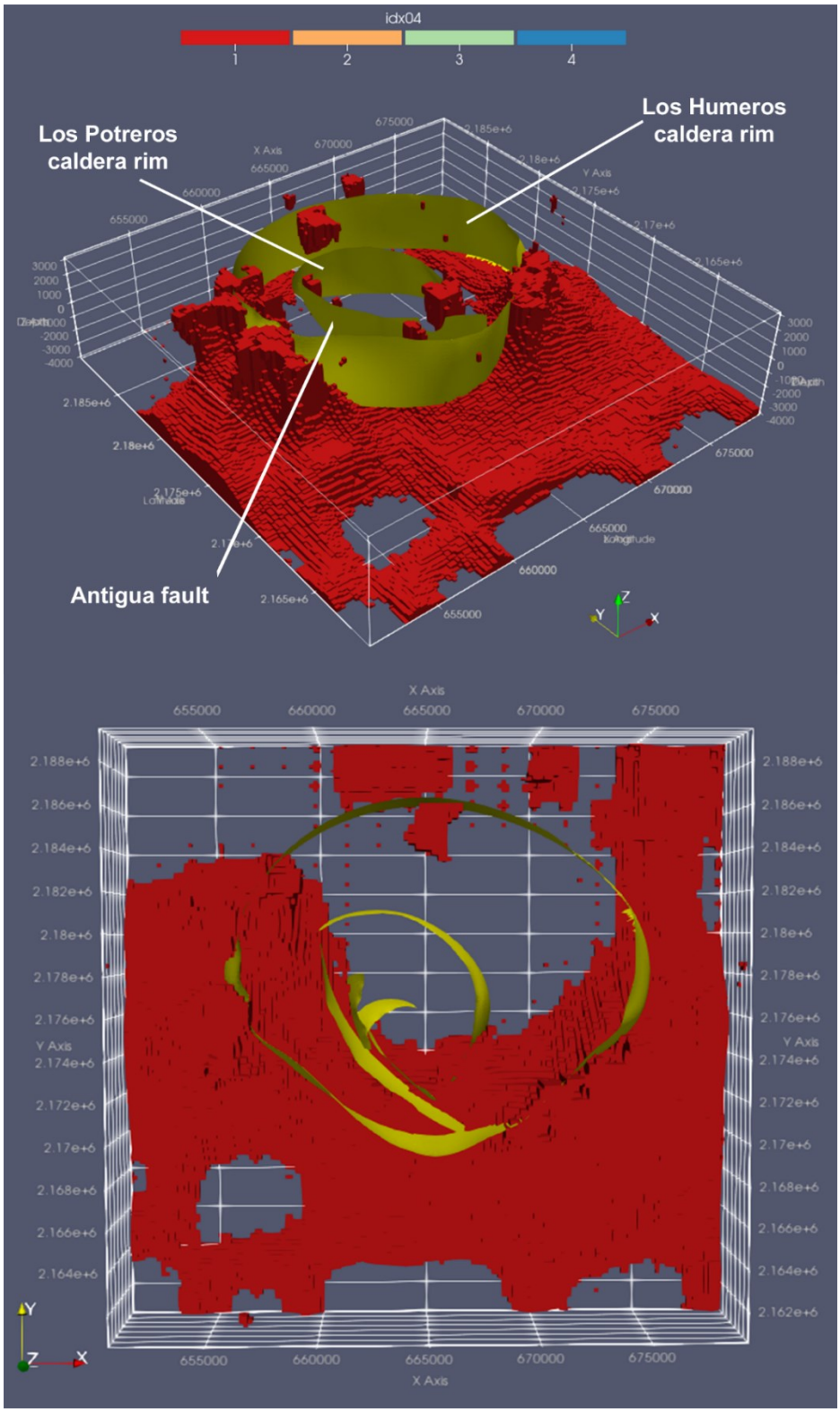

Figure 14: 3D visualization (upper) and top view (lower) of the cluster k = 1 together with the main structural features.

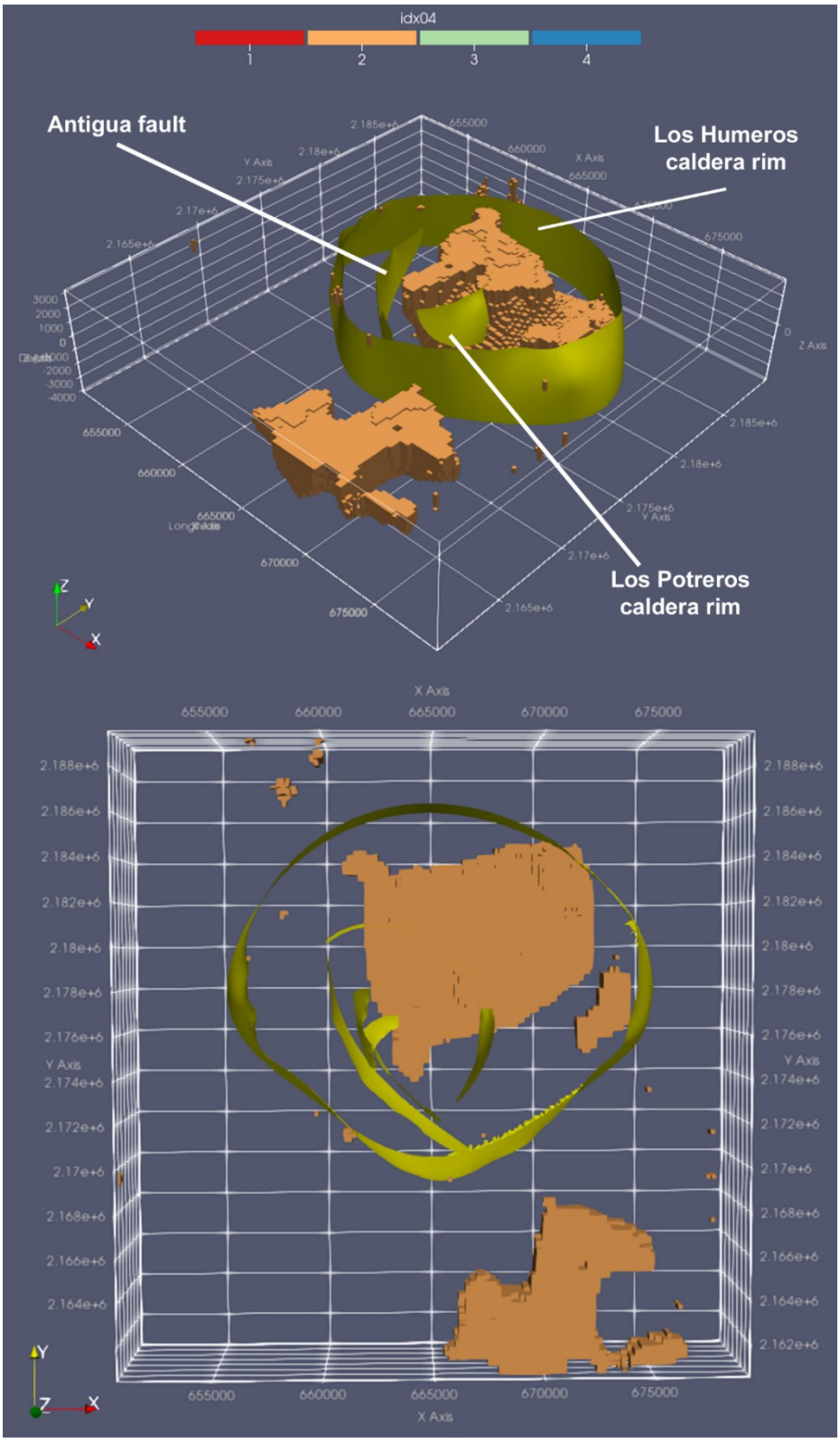

Figure 15: 3D visualization (upper) and top view (lower) of the cluster k = 2 together with the main structural features.

3.3.2 Local model

The local scale cluster analysis in Los Humeros has been performed by means of the resistivity structure obtained from the MT data processing (GEMex, 2019a) and the Vp/Vs structure computed by travel-time tomography of the recorded local seismicity (GEMex, 2019c). In Figure 16 the cluster distributions evaluated with the GMM for a number of components variable between 1 and 9 are displayed. In order to choose the best fitting model having the minimum number of components we used the information criteria reported in Figure 17Figure 6. The model with four components is that one capable to capture the main features of the data using the minimum number of clusters. In Figure 18 the statistical distribution of the values of each cluster is reported for the above-mentioned solution. The cluster 1 has medium-low resistivity (mean value of 66 Ω m) and medium-high Vp/Vs (mean value of 1.61). The cluster 2 has the lowest resistivity (mean value of 38 Ω m) and the lowest Vp/Vs (mean value of 1.54). The cluster 3 has the highest resistivity (mean value of 124 Ω m) and the highest Vp/Vs (mean value of 1.74). The cluster 4 has medium-high resistivity (mean value of 118 Ω m) and medium-high Vp/Vs (mean value of 1.56).

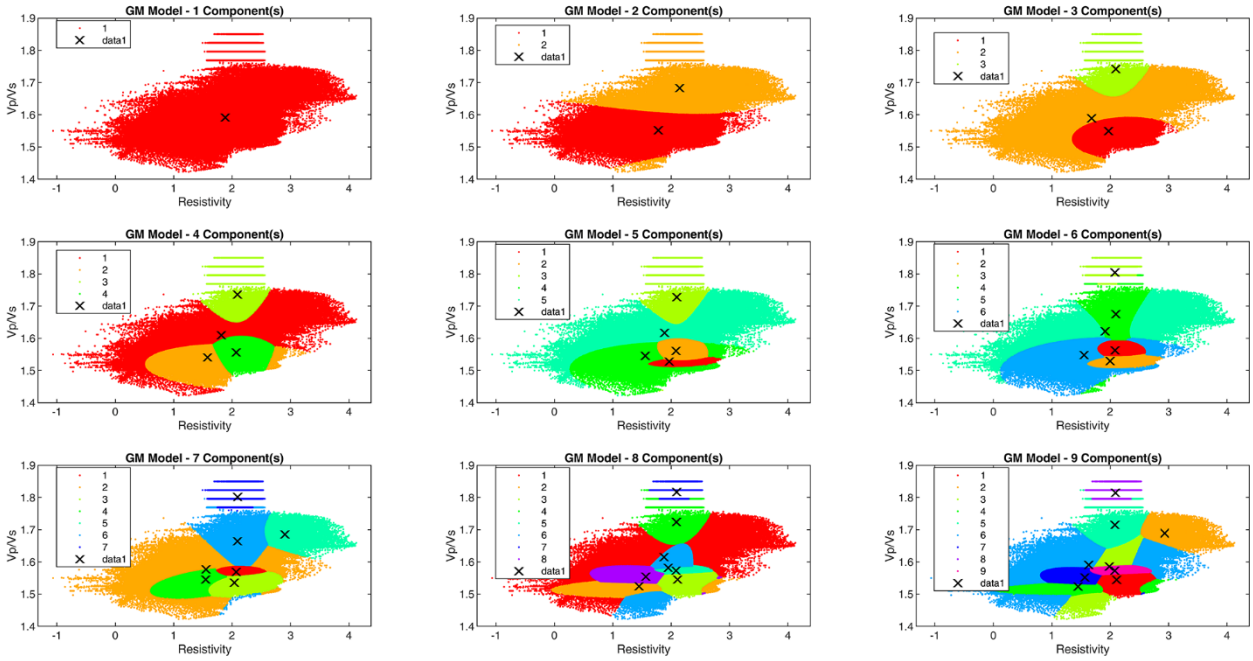

Figure 16: Unsupervised clustering using the GMM with k variable from 1 to 9 applied to the Local Resistivity-Vp/Vs interpolated dataset.

The main results are summarized in Figure 19. Los Homeros exhibits the typical structure of high temperature geothermal systems characterized by a thick low-resistivity cap dome underlain by a resistive core (GEMex, 2019d). The low-resistivity regions have two distinct Vp/Vs characters: one has medium-high Vp and Vs ratio, the second has low Vp/Vs ratio. Essentially, the clusters 1 and 2 highlight the above-mentioned characteristics, respectively. As pointed out in (GEMex, 2020), low resistivity is seen close to the surface in the main production area. Here, the cluster 2 appears very shallow, almost coincident with the topographic surface. Outside this area, the cluster 1 occupies

the shallower levels and mainly set above the cluster 2. The latter deepens in the eastern sector and mimics the modulation in depth of the 50 Ω m resistivity iso-surface. The cluster 1 appears again at depths between 0 and +1000 m a.s.l. in the NW and SE sectors of the Potreros caldera. In the NW sector, this cluster mainly locates spatially within the productive layers. Here, the indication of the possible occurrence of fluids highlighted by medium-high Vp/Vs ratio is proven by the presence of several production wells.

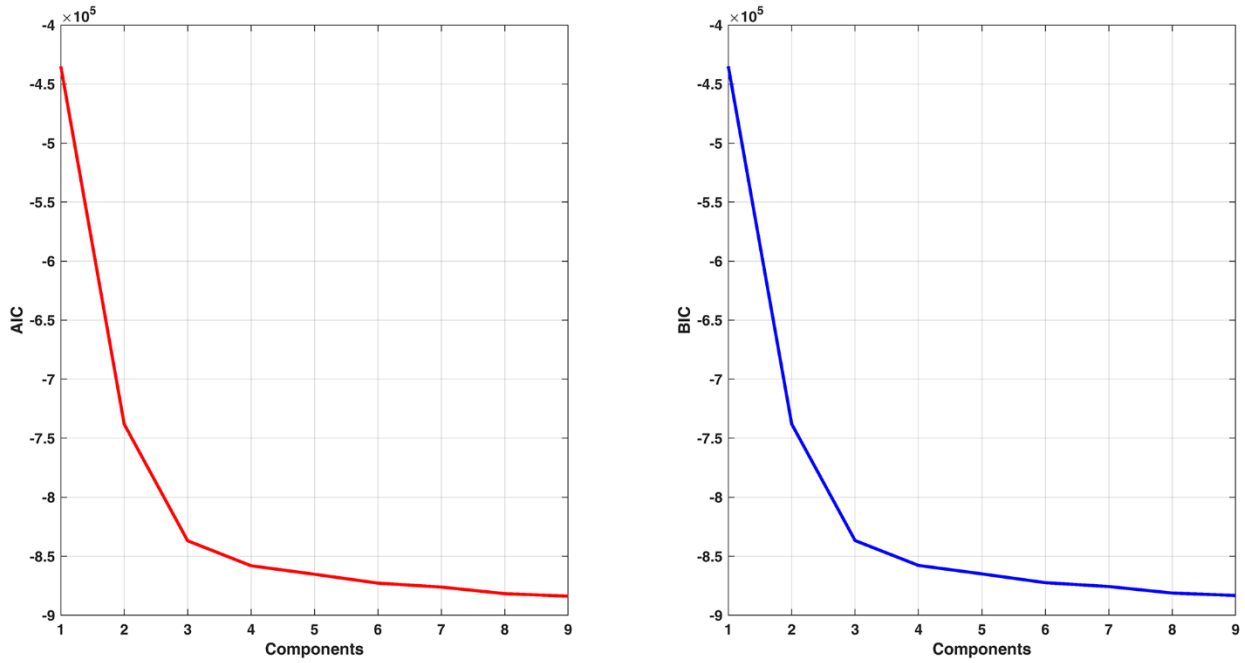

Figure 17: Akaike Information Criterion (AIC, left) and Bayes Information Criterion (BIC, right) evaluated for the GMM with k variable from 1 to 9 applied to the Local Resistivity-Vp/Vs interpolated dataset.

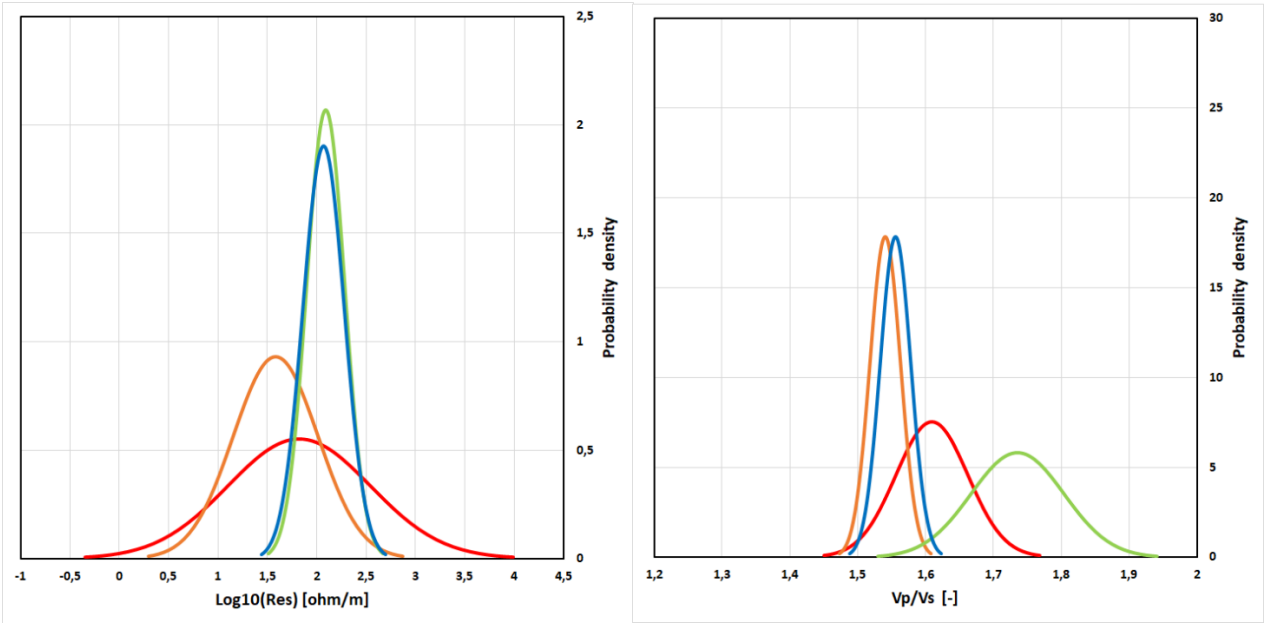

Figure 18: Statistics of clusters (k = 4) of the Local Resistivity (Right) and Vp/Vs (Left) interpolated dataset.

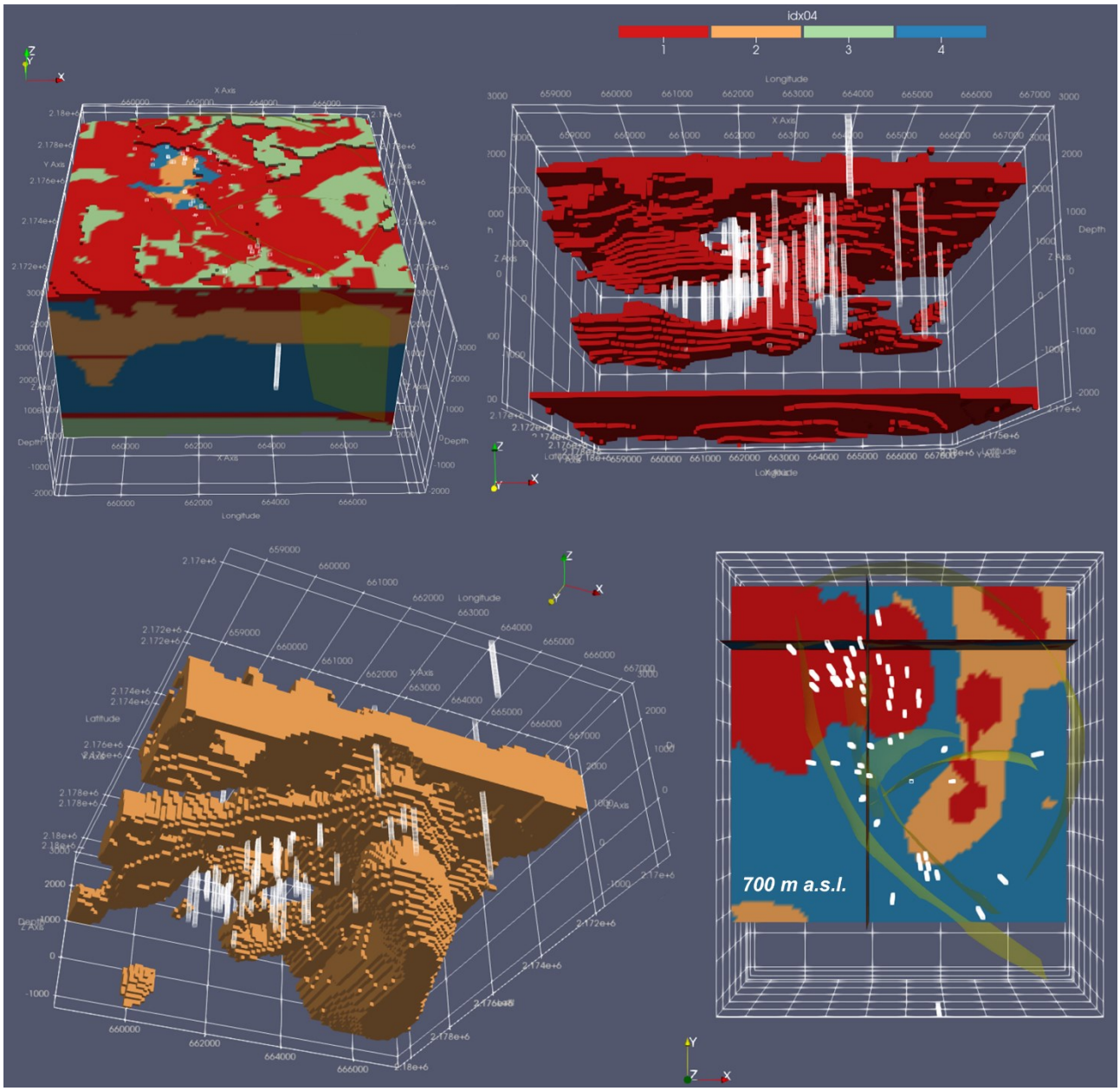

Figure 19: 3D visualizations of clusters (4 components) computed using the Local Resistivity and Vp/Vs interpolated dataset (upper left), spatial distribution of the cluster number 1 (upper right) and 2 (lower left), horizontal section (lower right) at 700 m a.s.l. together with the main structural features. The wells (white lines) are also reported.

4 Cluster analysis of Acoculco geothermal field

4.1 Geophysical datasets

For the detection of the main geological and geophysical features of Acoculco geothermal field we take advantage of the availability of different 3D geophysical models developed in the framework of the WP5 (Task 5.1 and Task 5.3) and of the 3D geological model developed in the framework of WP3 (Task 3.1). The datasets used for the cluster analysis are summarised in Table 3.

Table 3: List of the geological and geophysical data used in the cluster analysis of Acoculco area

Data type	Short description	Partner	Ref.
Local scale			
Geological model	Geological faults and units	BRGM	(Calcagno, et al., 2018)
Resistivity	Resistivity from 3D MT inversion	ISOR	(GEMex, 2019a)
Density	Density contrast from 3D grav inversion	KIT (INE)	(GEMex, 2019b)
Regional scale			
Density	Density contrast from joint 3D grav-mag inversion	CICESE	(Carrillo, et al., 2020)
Magnetization	Magnetization from joint 3D grav-mag inversion	CICESE	(Carrillo, et al., 2020)

4.2 Cross-plots and Density plots

In the following sections we present the results obtained from the application of the supervised clustering to the data coming from the local resistivity and density distributions resulting from MT and gravity inversions, respectively. In Figure 20 the cross-plot and density plot of the local density vs resistivity data are displayed. Density is expressed in terms of density contrast against a reference value of 2.67 g/cm^3. The density contrast values range between –0.75 and +1.25 g/cm^3. The resistivity values fall in the interval –0.4 – 3.7 $\log_{10}$(Ω m). The density plot highlights a principal cluster centred around 0 g/cm^3 and 100 Ω m. It is difficult to recognize a clear correlation between density and resistivity as the data points scatter in a cross-like pattern.

4.3 Supervised clustering

4.3.1 Local model

As described in the section *Decision Tree Model*, an *a priori* classification is needed as input to instruct the algorithm. For this purpose, a Training dataset resulting from a random extraction of data from the whole dataset is used. The chosen classification consists of 9 groups defined by the threshold values –0.05, +0.05 g/cm^3 and 60, 150 Ω m for the density contrast and resistivity, respectively (Table 4). Low-resistivity together with low-density coupled evidences should be indicative of occurrence of geothermal fluids (liquid phase) hosted in high porosity rocks.

Table 4: Supervised classification of the resistivity and density dataset in Acoculco area.

	D < –0.05 g/cm³	–0.05 <= D < +0.05 g/cm³	D >= +0.05 g/cm³
R < 60 Ω m	11	21	31
60 <= R < 150 Ω m	12	22	32
R > 150 Ω m	13	23	33

In Figure 21 the results of the supervised clustering are reported. Cluster 11, corresponding to the low density-low resistivity class, images isolated bodies underlying the shallower cluster 21, corresponding to the medium density-low resistivity class. The latter has a roughly dome-shaped character with minimum thickness in proximity of the drilled wells (EAC-1 and EAC-2) and thickening toward the outer borders. In vicinity of the wells the cluster 33, corresponding to high density-high resistivity class, occurs. The top of this anomalous body, referred as the anomaly (A1) in the (GEMex, 2020) corresponds to bottom of volcanites so that the anomaly falls into the metamorphosed rocks (skarn and hornfels) following the stratigraphy by (Pulido, et al., 2010). Another interesting body is that one depicted by the cluster 23, corresponding to medium density-high resistivity, located at the centre of the area of study and underlying the cluster 33. This cluster should mimic the shape of the granitic intrusion drilled in both the wells.

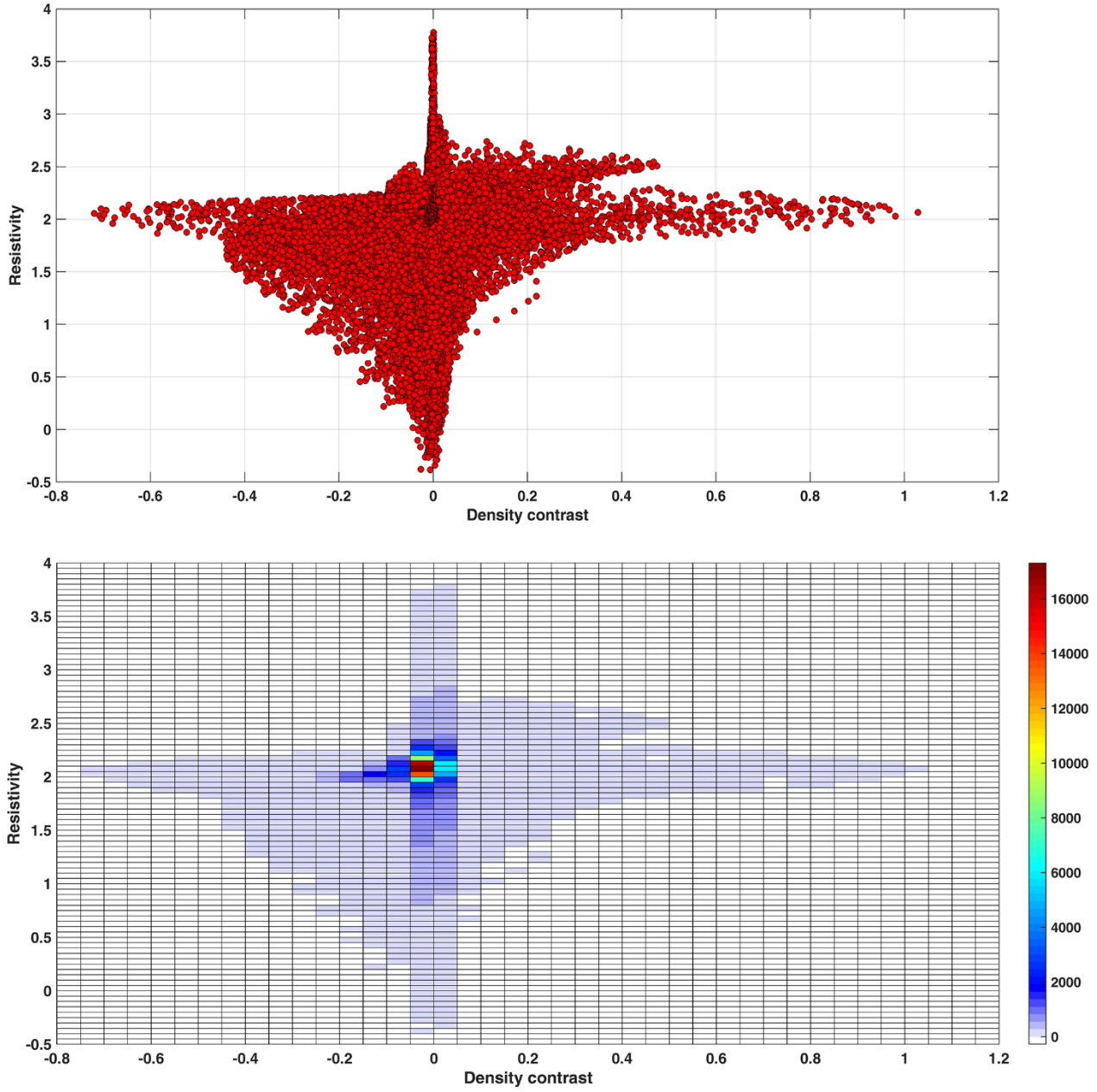

Figure 20: Cross-Plot (upper) and Density-Plot (lower) of the Acoculco Local Resistivity and Density interpolated dataset.

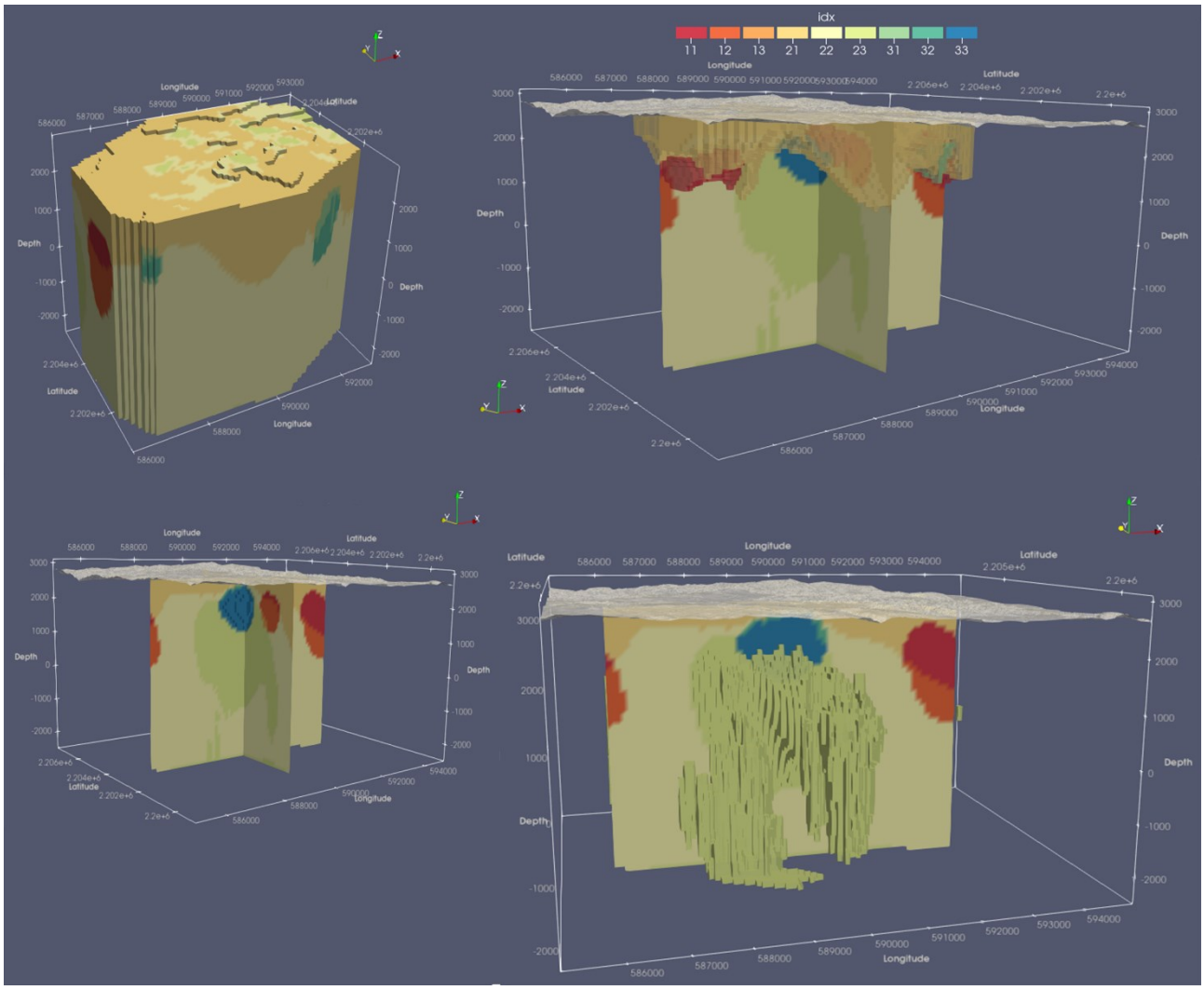

Figure 21: 3D visualizations of supervised clusters (k = 9) computed for the Local Resistivity and Density interpolated dataset (upper left), clusters k = 11, 21 and 31 (upper right), k = 33 (lower left) and k = 23 (lower right).

5 Validation of the protocol

5.1 Case study 1: Krafla (Iceland)

The Krafla region set within the Northern Volcanic Zone of Iceland. The Krafla volcanic system consists of a central volcano bisected by an NNE-SSW trending fissure swarm which accommodates most of the crustal spreading. A caldera 110 ka old occurs in the middle of the area. A comprehensive overview of the magma-hydrothermal-tectonic system of Krafla can be found in (Arnason, 2020). We applied the unsupervised GMM to the available resistivity (ISOR internal report) and velocity (Schuler, et al., 2015) 3D models (Table 5). In Figure 22 the cross-plot and density-plot of resistivity and Vp/Vs data are reported. The resistivity spans between 1 to 10000 Ω m, the Vp/Vs ranges from 1.65 to 1.85. The density-plot highlights a principal cluster around 100 Ω m and a Vp and Vs ratio of 1.75. In Figure 23 the cluster distributions evaluated with the GMM for a number of components variable between 1 and 9 are displayed. In order to choose the best fitting model having the minimum number of components we used the information criteria reported in Figure 24Figure 6. A net change in the slope of the information criteria as function of the number of the components is not clearly recognizable.

Table 5: List of the geophysical data used for validating the protocol in the Islandic site

Data type	Short description	Delivered by
Resistivity	Resistivity from 3D MT inversion	ISOR
Velocity model	Vp and Vp/Vs from seismic tomography	ISOR

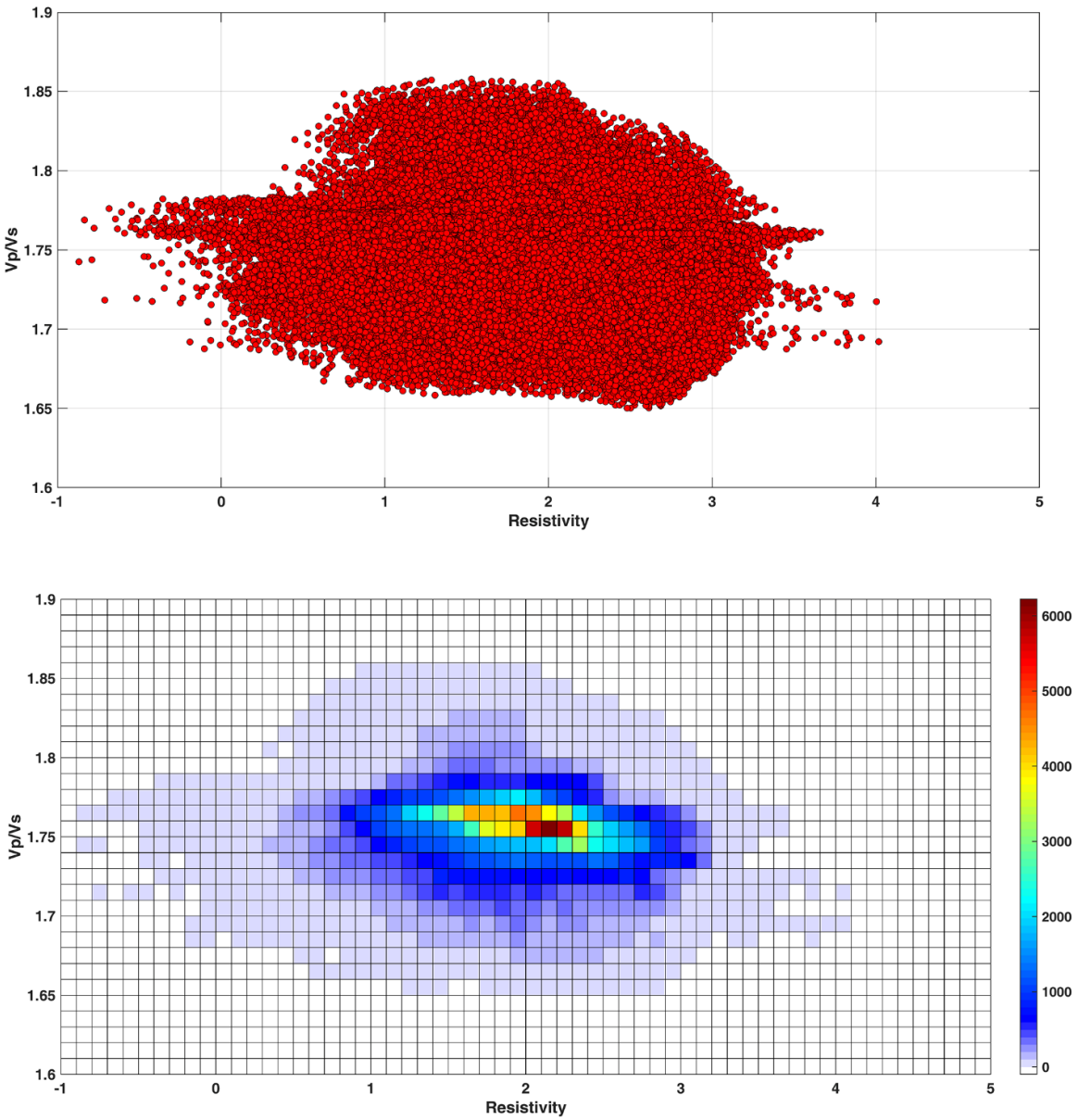

Figure 22: Cross-Plot (upper) and Density-Plot (lower) of the Krafla Resistivity and Vp/Vs interpolated dataset.

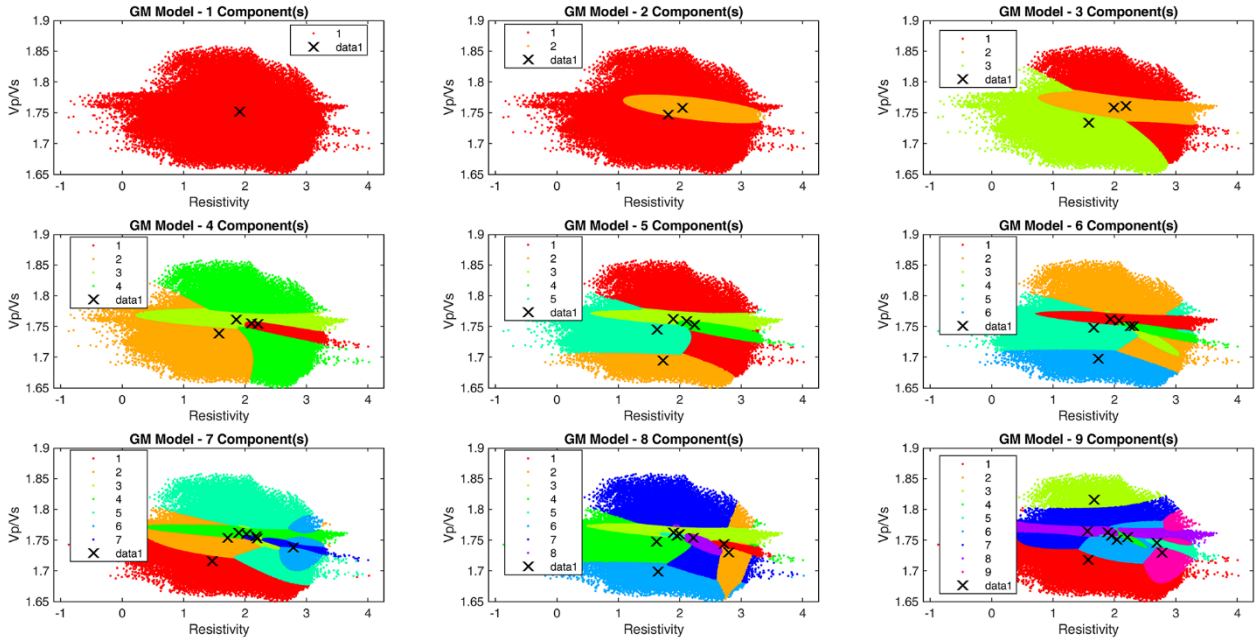

Figure 23: Unsupervised clustering using the GMM with k variable from 1 to 9 applied to the Krafla Resistivity and Vp/Vs interpolated dataset.

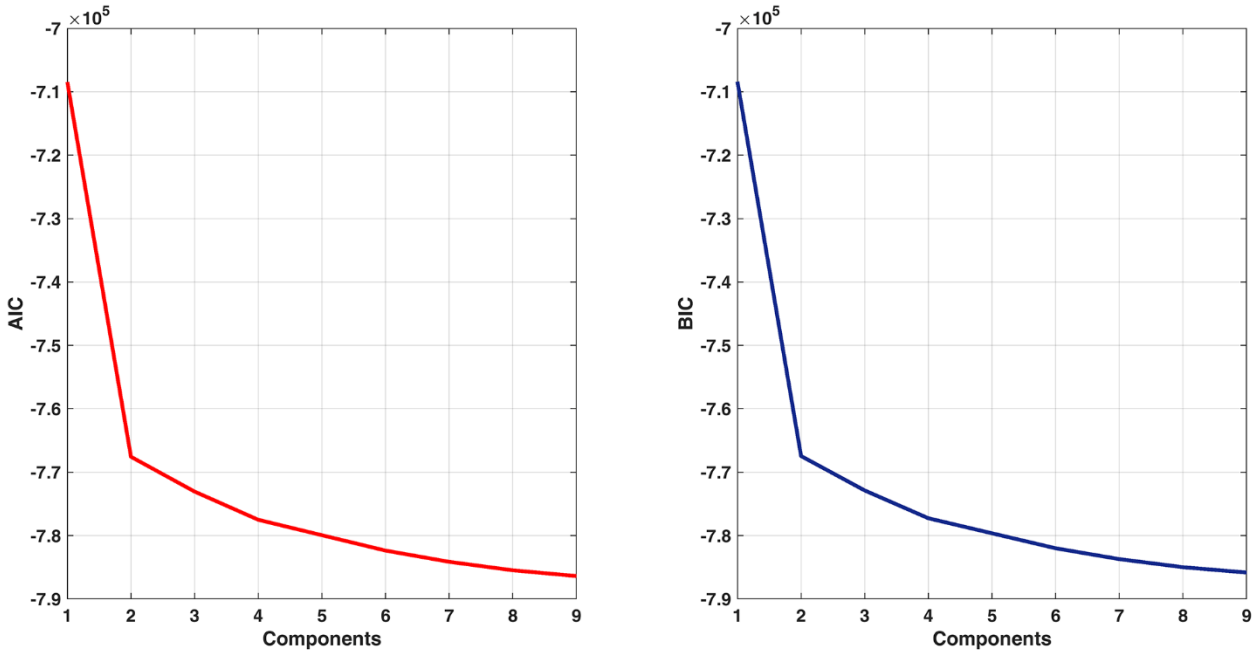

Figure 24: Akaike Information Criterion (AIC, left) and Bayes Information Criterion (BIC, right) evaluated for the GMM with k variable from 1 to 9 applied to the Krafla Resistivity and Vp/Vs interpolated dataset.

The goodness of the model fit continuously increase with the growing number of clusters. For this reason, we investigated the results obtained by the 9 components analysis. In Table 6 the statistics of each cluster are reported for the above-mentioned solution.

Table 6: Statistics of the clusters (k = 9) of the Krafla Resistivity and Vp/Vs interpolated dataset .

Cluster	Resistivity [Log_{10}(Ω m)]		Vp/Vs [-]	
	Mean	St.dev.	Mean	St.dev.
1	1.5671	0.3780	1.7177	0.0006
2	1.9485	0.0667	1.7571	0.0000
3	1.6695	0.1511	1.8157	0.0003
4	2.2073	0.0365	1.7545	0.0001
5	2.6866	0.0604	1.7457	0.0001
6	2.0362	0.1547	1.7507	0.0003
7	1.5579	0.2305	1.7646	0.0003
8	1.8872	0.3055	1.7626	0.0000
9	2.7734	0.0341	1.7295	0.0006

Clusters number 1, 3 and 9 are visualized in Figure 25. Cluster 1 has low resistivity (mean value of 37 Ω m) and low Vp/Vs (mean value of 1.72). Cluster 3 has low resistivity (mean value of 47 Ω m) and high Vp/Vs (mean value of 1.82). Cluster 9 has high resistivity (mean value of 593 Ω m) and low Vp/Vs (mean value of 1.73). The cluster 1 is characterized by two distinct bodies which set at the top and the bottom of the investigated volume connected by a narrow vertical structure. At the top, the mean values of resistivity and Vp/Vs suggest that it should correspond to the alteration cap-rock.

With the aim to highlight possible volumes hosting geothermal fluids we focused on the coupled parameters low resistivity and high Vp/Vs ratio. The cluster 3 fulfils the above mentioned requirements. Within this volume, the resistivity and Vp/Vs values should indicate the possible occurrence of fluid at liquid phase. The cluster 3 relies above the deep body characterized by a bi-modal distribution of the physical parameters. At depths between 2 and 3 km b.s.l. cluster 1 and cluster 9 appear connected. The resistive and almost dry cluster 9 should correspond to a crystallized intrusion. The deeper cluster 1 should correspond to the same intrusive body hosting a minor amount of residual, hypersaline magmatic fluids which drop down the resistivity but do not change the overall velocity distribution.

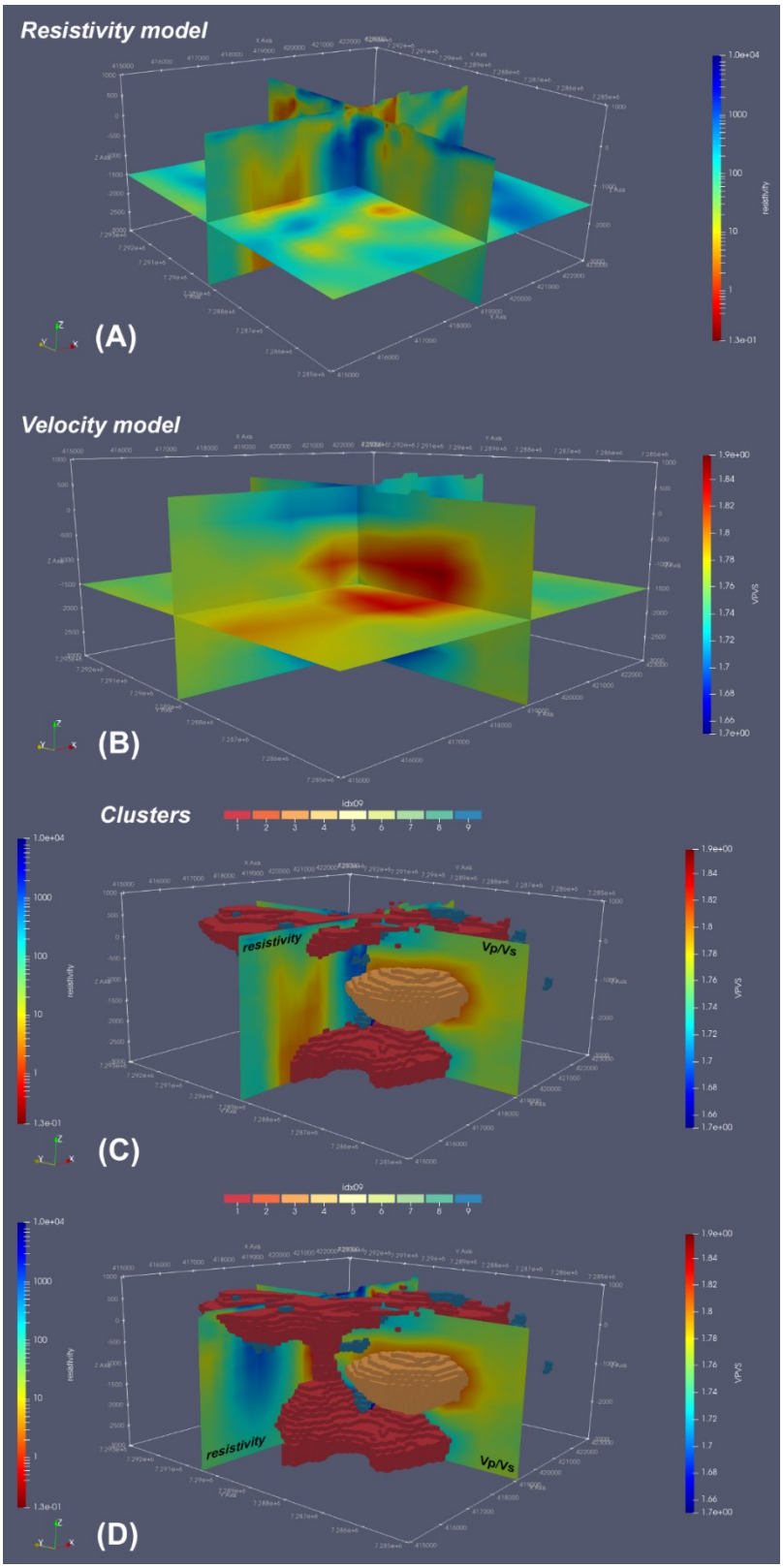

Figure 25: 3D visualization of the resistivity (A) and Vp/Vs (B) distributions along two N-S and E-W vertical sections intersecting themselves in the vicinity of the ICDP-1 well and on the horizontal slice set approximatively at the bottom of the well (-1500 m b.s.l.). The 3D spatial distributions of the clusters 1, 3 and 9 (C and D) are reported together with the resistivity and Vp/Vs structures on the vertical sections. In the figure C the sections are located as in A and B, in the figure D the E-W resistivity section is moved northward by few kilometres intersecting the vertical conductive anomaly highlighted by the cluster 1.

5.2 Case study 2: Mensano (Italy)

In the framework of the EU H2020 GECO Project, the Mensano geothermal permit has been selected as the Italian case study to set-up and test technologies to lower the emissions from geothermal power generation by capturing them for either reuse or storage (Trumpy, et al., 2020). The Italian GECO demonstration site locates in the northeaster side of the Larderello geothermal area. The geothermal anomaly characterizing the Larderello geothermal area s.l. should be therefore framed in the magmatic and tectonic evolution of the inner Northern Apennines, also characterised by the so-called Tuscan magmatic province. A recent overview of the geothermal system in Larderello and surrounding areas can be found in (Gola, et al., 2017).

We applied the unsupervised cluster analysis and supervised classification to the available resistivity, density and magnetization (Magma Energy internal report) 3D models (Table 7). In Figure 26 the datasets are illustrated along a SW-NE section. In Figure 27 the cross-plot and density-plot of resistivity and density data are reported. The resistivity spans between 1 to 10000 Ω m, the density ranges from 2.2 to 2.9 g/cm^3. The density-plot highlights a principal asymmetric cluster with centroid around 100 Ω m and 2.7 g/cm^3 elongated toward more resistive values (> 10^3 Ω m) and slightly greater density (about 2.8 g/cm^3). Two minor distinct clusters are present and centered at about 20 Ω m – 2.6 g/cm^3 and 7 Ω m – 2.4 g/cm^3. In Figure 28 the density-plot of density vs magnetization and resistivity vs magnetiation data are reported. Magnetization is very low within the investigated volume with some anomalies highlighted by values higher than 0.008 SI. Few points assume magnetizations as high as 0.035 SI.

Table 7: List of the geophysical data used for validating the protocol in the Italian site

Data type	Short description	Delivered by
Resistivity	Resistivity from 3D MT inversion	CNR
Density	Density contrast from 3D grav inversion	CNR
Magnetization	Magnetization from 3D mag inversion	CNR

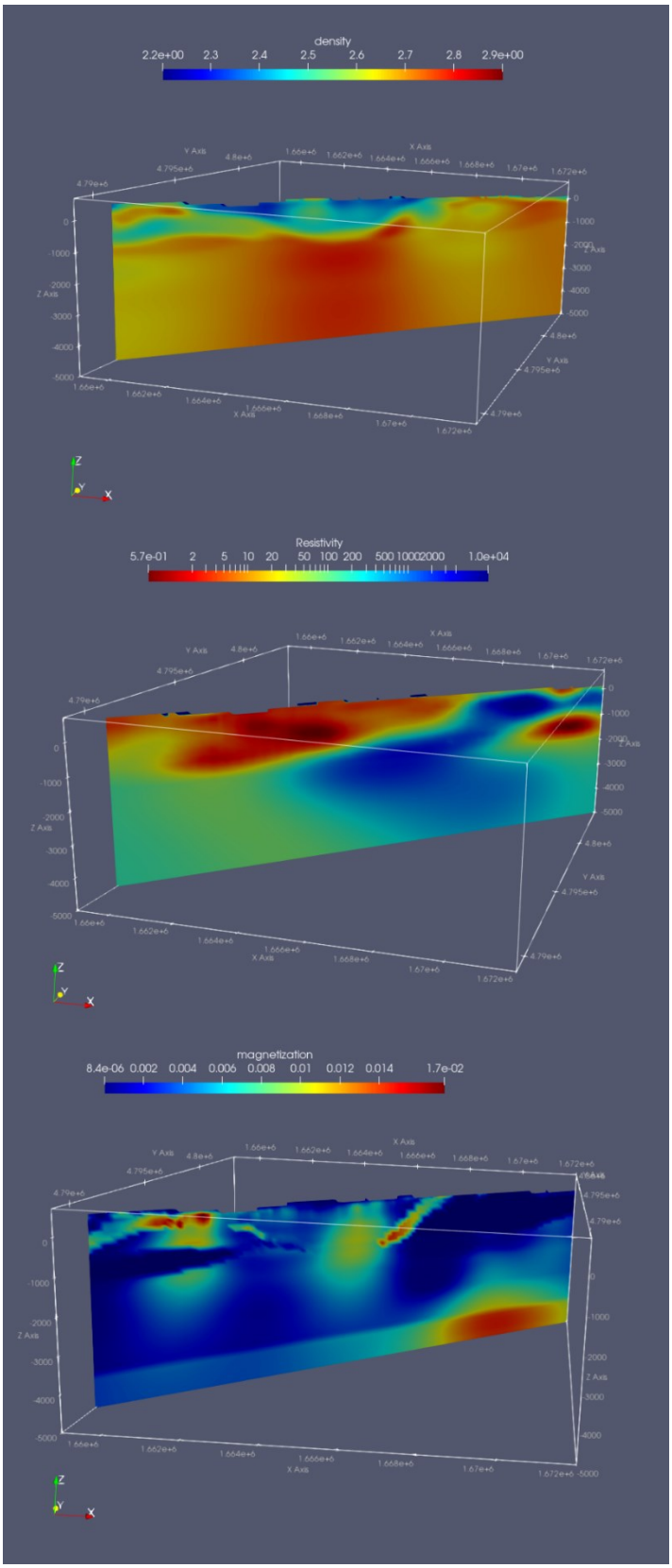

Figure 26: Distributions of density (upper), resistivity (middle) and magnetization (lower) along a SW-NE section throughout the 3D models.

In Figure 29 the cluster distributions evaluated with the GMM for a number of components variable between 1 and 9 are displayed. In order to choose the best fitting model having the minimum number of components we used the information criteria reported in Figure 30. A net change in the slope of the information criteria as function of the number of the components is not clearly recognizable. The goodness of the model fit continuously increase with the growing number of clusters. For this reason, we investigated the results obtained by the 9 components analysis. In Table 8 the statistics of each cluster are reported for the above-mentioned solution.

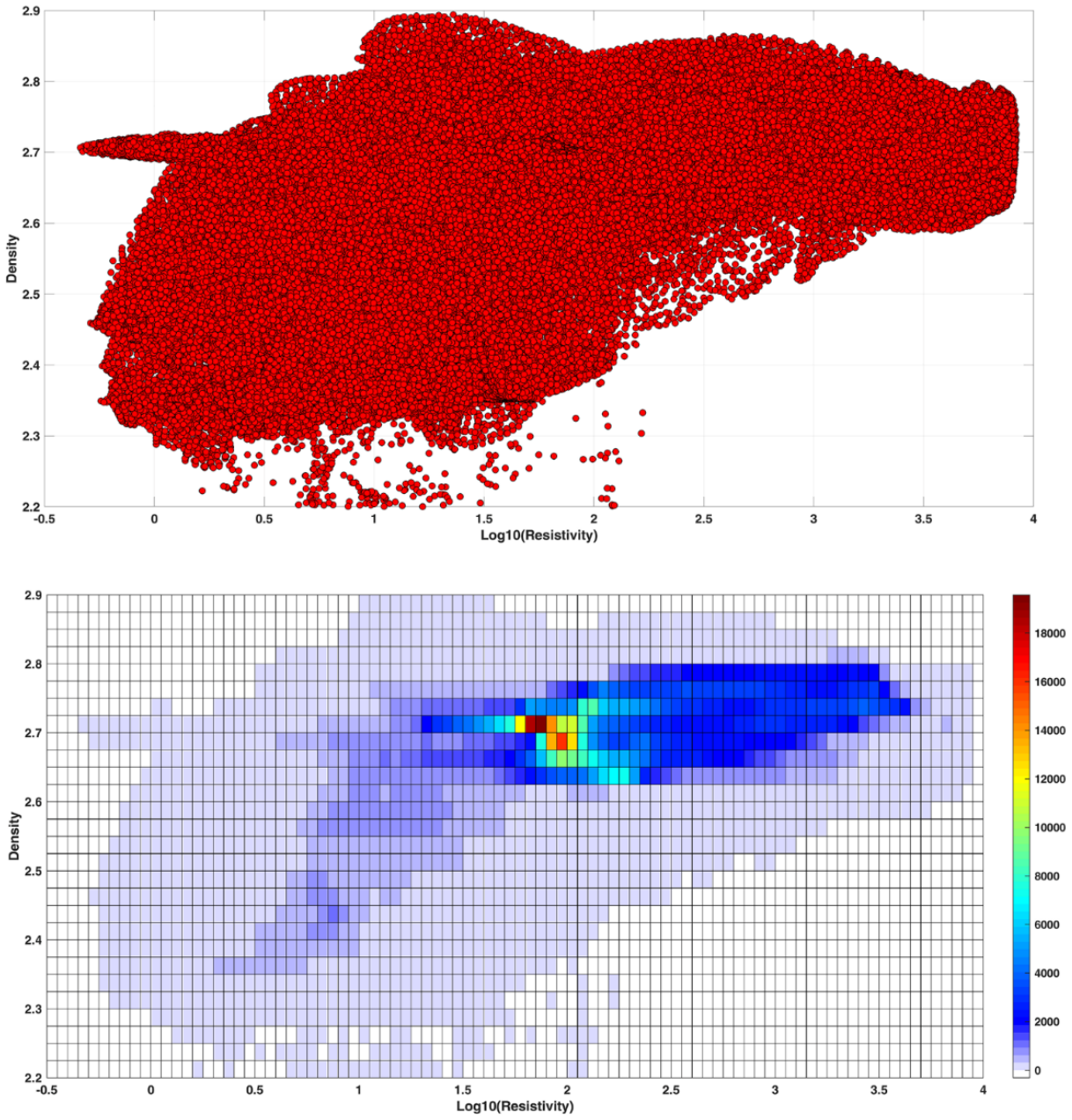

Figure 27: Cross-Plot (upper) and Density-Plot (lower) of the Mensano Resistivity and Density interpolated dataset.

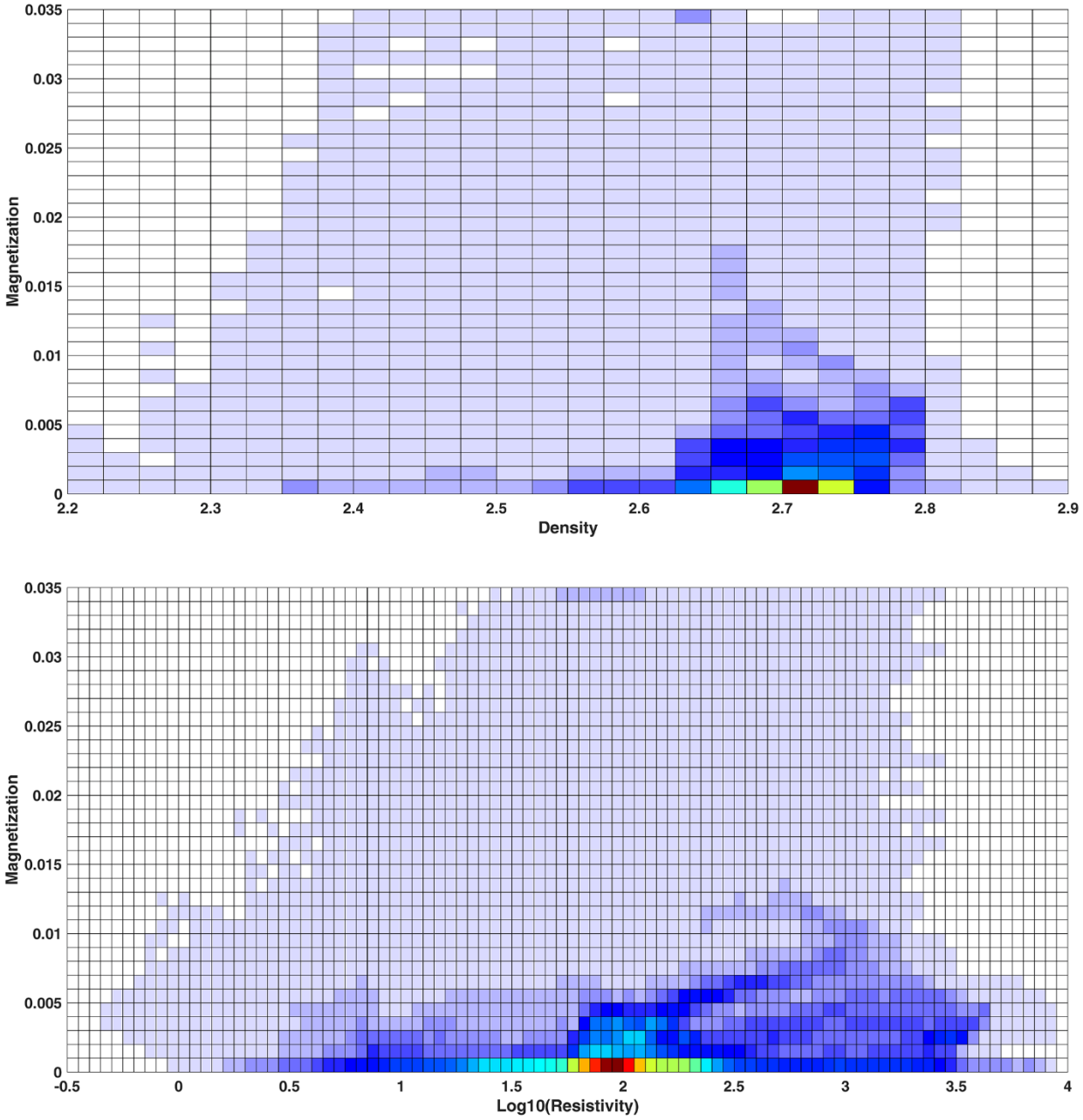

Figure 28: Density-Plot of the Mensano Density-Magnetization (upper) and Resistivity-Magnetization (lower) interpolated dataset.

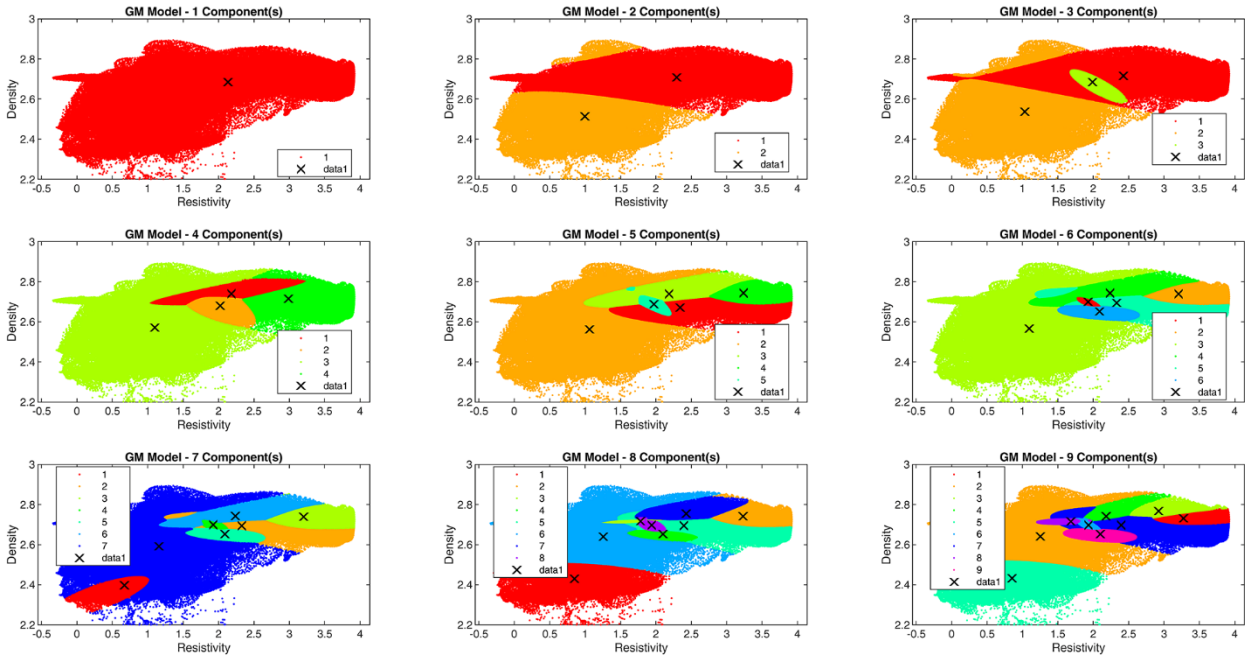

Figure 29: Unsupervised clustering using the GMM with k variable from 1 to 9 applied to the Mensano Resistivity and Density interpolated dataset.

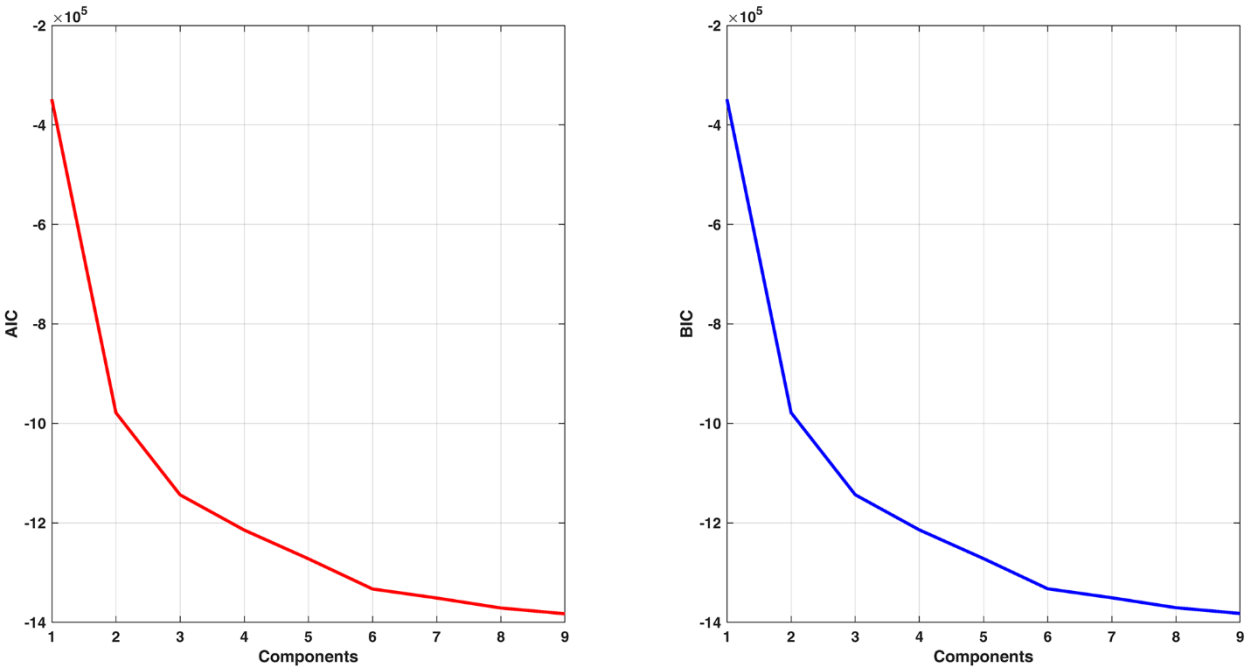

Figure 30: Akaike Information Criterion (AIC, left) and Bayes Information Criterion (BIC, right) evaluated for the GMM with k variable from 1 to 9 applied to the Mensano Resistivity and Density interpolated dataset.

Table 8: Statistics of the clusters (k = 9) of the Mensano Resistivity and Density interpolated dataset.

Cluster	Resistivity [$Log_{10}(\Omega$ m)]		Density	
	Mean	St.dev.	Mean	St.dev.
1	3.27445	0.0610	2.7315	0.0004
2	1.2524	0.2162	2.6406	0.0061
3	2.9223	0.1020	2.7678	0.0003
4	2.1835	0.0556	2.7422	0.0006
5	0.8536	0.1686	2.4326	0.0041
6	1.9333	0.0081	2.6984	0.0003
7	2.3948	0.2362	2.6975	0.0010
8	1.6828	0.0424	2.7159	0.0000
9	2.0998	0.0496	2.6520	0.0002

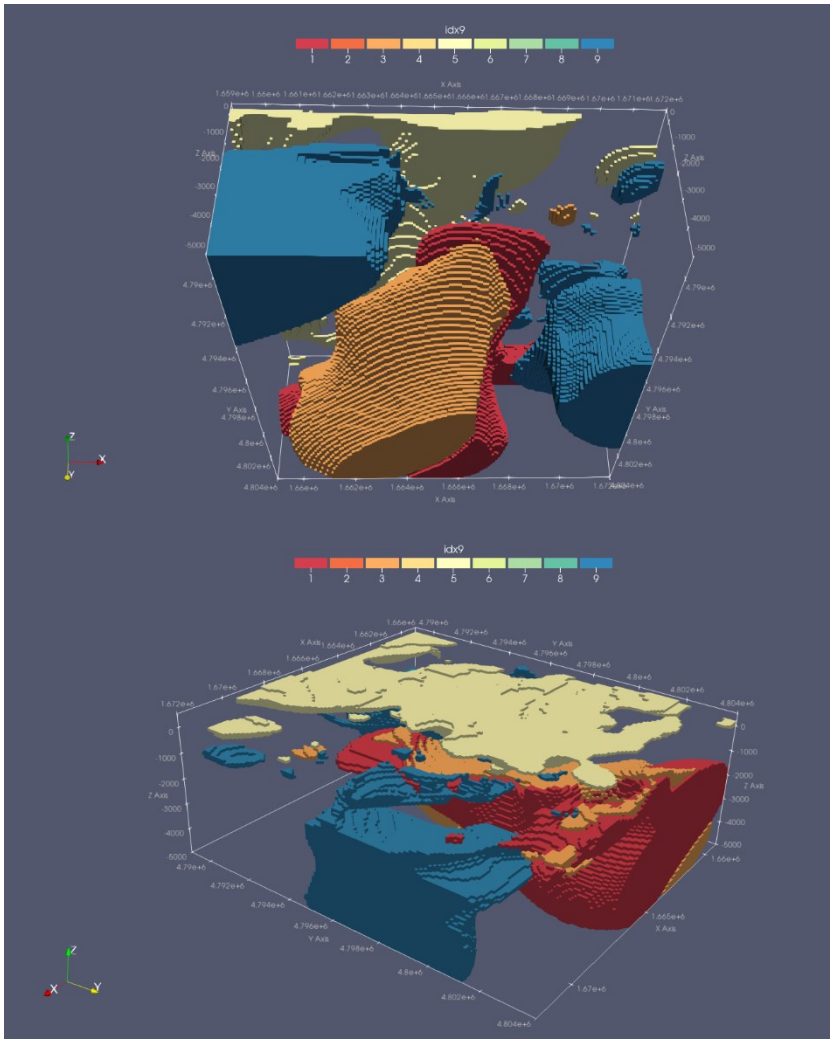

Figure 31: Akaike Information Criterion (AIC, left) and Bayes Information Criterion (BIC, right) evaluated for the GMM with k variable from 1 to 9 applied to the Mensano Resistivity and Density interpolated dataset.

The unsupervised clustering highlights predominantly NW-SE structures which mimic the NW-SE oriented principal alignments of the Apennine thrust belt (Figure 31). The uppermost cluster is characterized by low density and low resistivity (cluster 5) that fit well the sedimentary basin developed during the Miocene extensional tectonic phase. The underlying, mainly metamorphic, basement has high resistivity and high density with the exception of the SW and NE domains within the investigated volume and characterized by slightly lower values. This anomaly should be related to brittle deformation that increased porosity. Moreover, in the NE sector, the upper portion of the metamorphic basement outcrop. The clusters 1 and 3 mimic this geological feature.

Beside the GMM, we applied the supervised classification method in order to jointly integrate all the available data. We categorized in two main groups the magnetization data by a threshold value of 0.008 SI. The density is classified in three classes by the threshold values of 2.5 and 2.65 g/cm^3. Finally, the resistivity values are classified by the threshold values of 30 Ω m and 200 Ω m. This scheme results into 18 classes.

Table 9: Supervised classification of the magnetization, resistivity and density dataset in Mensano area.

M < 0.008 SI			
	D < 2.5 g/cm^3	**2.5 ≤ D < 2.65 g/cm^3**	**D ≥ 2.65 g/cm^3**
R < 30 Ω m	1	4	7
30 ≤ R < 200 Ω m	2	5	8
R ≥ 200 Ω m	3	6	9
M ≥ 0.008 SI			
R < 30 Ω m	10	13	16
30 ≤ R < 200 Ω m	11	14	17
R ≥ 200 Ω m	12	15	18

In Figure 32 we highlighted the magnetized bodies filtered by the low density and low resistivity ones.

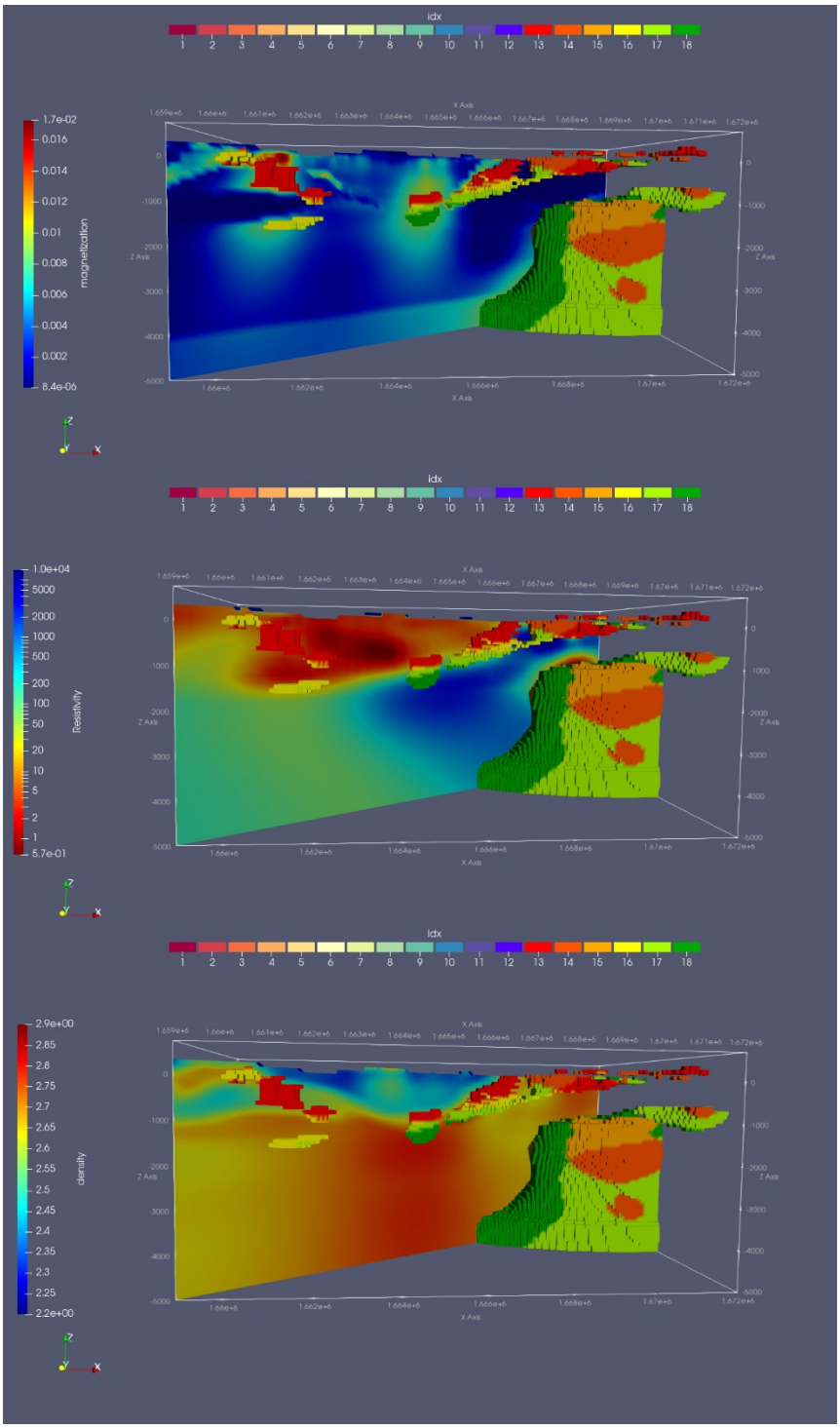

Figure 32: 3D visualization of the high magnetization, medium to high resistivity and medium to high density bodies together with the magnetization (upper), resistivity (middle) and density (lower) distribution along the SW-NE section.

6 Conclusions

The application of the clustering method to datasets in various geothermal area has shown that this approach is an effective method to quickly retrieve local relationships between distinct physical parameters. With respect to the visual integration described in D5.10, this approach is much faster: in a few hours it is possible to recognize various branches of clusters that, once mapped back to the space domain, provide easily recognizable volumes of joint parameters.

Velocity is not the only advantage. This approach is analytical, and reduce the bias due to subjectivity. Beside uncertainty, which refers to data quality - how uncertain are the data based on the type of technique employed, including inversion techniques, the joint interpretation of data is often subjective, i.e. shaped by the personal opinions and feelings of operators. With this approach the operator is left with only the decision on the number of clusters in the unsupervised approach, or the definition of classes in the supervised classification approach.

7 Acknowledgments

We acknowledge the Comisión Federal de Electricidad (CFE) for kindly providing support and advice and for granting access to their geothermal fields. Information on well locations has been kindly provided by CFE.

We also acknowledge our Mexican colleagues for their help and collaboration. Special thanks to J. Carrillo, M. A. Pérez-Flores, and L.A. Gallardo for the provision of the result of the joint inversion for regional density and magnetisation.

The authors thank the colleagues working in WP5 activities, who provided important hints in developing the method and the Los Humeros and Acoculco dataset on which this analysis is based. Landsvirkjun and Magma Energy courteously provided the dataset used on the Krafla and Mensano areas.

8 Bibliography

Arnason, K., 2020. New conceptual model for the magma-hydrothermal-tectonic system of Krafla, NE Iceland.. *Geosciences,* 10(34), p. 27.

Bishop, C. M., 2006. *Pattern Recognition and Machine Learning.* s.l.:Springer.

Calcagno, P. et al., 2018. Preliminary 3-D geological models of Los Humeros and Acoculco geothermal fields (Mexico) – H2020 GEMex Project. *Adv. Geosci.,* Volume 45, pp. 321-333.

Carrillo, J. et al., 2020. *3D Joint Inversion of Gravity and Magnetic Data in Los Humeros and Acoculco Unconventional Geothermal Systems.* Reykjavik, Iceland, April 26 – May 2, 2020, Proceedings World Geothermal Congress 2020.

Di Giuseppe, M. G. et al., 2018. A geophysical k-means cluster analysis of the Solfatara-Pisciarelli volcano-geothermal system, Campi Flegrei (Naples, Italy). *Journal of Applied Geophysics,* Volume 156, pp. 44-54.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D., 1998. Cluster analysis and display of genome-wide expressions patterns. *Proc. Natl. Acad. Sci. U. S. A,* Issue 95, pp. 14863-14868.

GEMex, 2019a. *GEMEX-Deliverable D5.2: Report on resistivity modelling and comparison with other SHGS.,* GEMex Project (Horizon 2020, grant agreement No 727550): http://www.gemex-h2020.eu.

GEMex, 2019b. *GEMEX-Deliverable D5.6: Report on gravity modelling,* GEMex Project (Horizon 2020, grant agreement No 727550): http://www.gemex-h2020.eu.

GEMex, 2019c. *GEMEX-Deliverable D5.3: Seismic structures of the Acoculco and Los Humeros geothermal fields.,* GEMex Project (Horizon 2020, grant agreement No 727550): http://www.gemex-h2020.eu.

GEMex, 2019d. *GEMEX-Deliverable D5.8: Report on 3D resistivity modelling with external constraints.* http://www.gemex-h2020.eu, GEMex Project (Horizon 2020, grant agreement No 727550).

GEMex, 2020. *GEMEX-Deliverable D5.10: Report on integrated geophysical model of Los Humeros and Acoculco.,* GEMex Project (Horizon 2020, grant agreement No 727550): http://www.gemex-h2020.eu.

Gola, G. et al., 2017. Data integration and conceptual modelling of the Larderello geothermal area, Italy. *Energy Procedia,* Issue 125, pp. 300-309.

Lindsey, C. R. et al., 2018. Cluster analysis as a tool for evaluating the exploration potential of Known Geothermal Resource Areas. *Geothermics,* Volume 72, pp. 358-370.

Lloyd, S. P., 1982. Leat squares quantization in PCM. *IEEE Transactions on Information Theory,* 28(2), pp. 129-137.

Pulido, C., Armenta, M. & Silva, G., 2010. *Characterization of the Acoculco Geothermal Zone as a HDR System.* s.l., GRC Transaction, Vol. 34.

Schuler, J. et al., 2015. Seismic imaging of the shallow crust beneath the Krafla central volcano, NE Iceland. *Journal of Geophysical Research: Solid Earth,* Issue 120, p. 7156–7173.

Trumpy, E. et al., 2020. *Geological Assessment of Castelnuovo (Italy) Demonstration Site for CO2 Reinjection in Deep Geothermal Reservoir. H2020 GECO Project.* Reykjavik, Iceland, April 26 – May 2, 2020, Proceedings World Geothermal Congress 2020.

Coordination Office. GEMex project

Helmholtz-Zentrum Potsdam
Deutsches GeoForschungsZentrum

Telegrafenberg. 14473 Potsdam

Germany