



# The Hydrothermal System of Solfatara Crater (Campi Flegrei, Italy) Inferred From Machine Learning Algorithms

Stefano Bernardinetti<sup>1</sup> and Pier Paolo G. Bruno<sup>2\*</sup>

<sup>1</sup> Laboratorio di Geofisica Applicata, Centro di GeoTecnologie, Università degli Studi di Siena, San Giovanni Valdarno, Italy;

<sup>2</sup> Dipartimento di Scienze della Terra, dell'Ambiente e delle Risorse, Università di Naples Federico II, Naples, Italy

## OPEN ACCESS

### Edited by:

Susanne Buiter,  
Geological Survey of Norway, Norway

### Reviewed by:

Luca De Siena,  
Johannes Gutenberg University  
Mainz, Germany  
Cleyton Carneiro,  
University of São Paulo, Brazil

### \*Correspondence:

Pier Paolo G. Bruno  
pierpaolo.bruno@unina.it

### Specialty section:

This article was submitted to  
Solid Earth Geophysics,  
a section of the journal  
Frontiers in Earth Science

**Received:** 29 July 2019

**Accepted:** 18 October 2019

**Published:** 06 November 2019

### Citation:

Bernardinetti S and Bruno PPG (2019)  
The Hydrothermal System of Solfatara  
Crater (Campi Flegrei, Italy) Inferred  
From Machine Learning Algorithms.  
*Front. Earth Sci.* 7:286.  
doi: 10.3389/feart.2019.00286

Two machine learning algorithms were applied to three multivariate datasets acquired at Solfatara volcano. Our aim was to find an unbiased and coherent synthesis among the large amount of data acquired within the crater and along two orthogonal vertical NNE- and WNW-trending cross-sections. The first algorithm includes a new approach for a soft K-means clustering based on the use of the silhouette index to control the color palette of the clusters. The second algorithm which uses the self-organizing maps incorporates an alternative method for choosing the number of nodes of the neural network which aims to avoid the need for downstream clustering of the results of the classification. Both methods achieved an objective characterization of the shallow hydrothermal system of the volcano, enhancing and highlighting subtle geophysical anomalies likely correlated to structural pathways of deep magmatic degassing. Comparison between the results of K-means and self-organizing maps on the datasets with the largest number of nodes confirms that, with respect to the K-means, self-organizing maps compress the data in a way that better highlights finer details of the original data. However, the choice of the coloring scheme of the neurons is critical for an effective visualization of the results. Unsupervised integration of the three multivariate datasets allowed us to spatially correlate, with a high-degree of confidence, the geophysical anomalies recorded at the surface of the crater with those recorded at the subsurface along the two cross-sections. It also allowed us to associate those anomalies to different hydrothermal features such as shallow gas-saturated and water-saturated zones and their underlying fractures/faults feeding system. Our results suggest that the main shallow structural patterns, which influence the hydrothermal dynamics at Solfatara volcano, remained substantially unchanged in the last 13 years. Our approach shows that the use of clustering methods to interpret multivariate data reduces interpretation uncertainties and achieves an improved understanding of the complex dynamics occurring in volcanoes.

**Keywords:** Solfatara, machine learning, self-organizing maps, K-means, Campi Flegrei, seismic methods, potential methods, hydrothermal systems

## INTRODUCTION

Solfatara crater is at present the most active area of Campi Flegrei, a large, 13 km wide, nested caldera located within the metropolitan area of Naples, Italy. Underneath this small (i.e.,  $\sim 600 \times 700$  m) but dynamic maar-diatreme structure (Isaia et al., 2015; Bruno et al., 2017), deep magmatic CO<sub>2</sub>-rich fluids mix with meteoric water and form a hydrothermal plume that feeds fumaroles and mud pools (Caliro et al., 2007). Each day  $\sim 3,350$  t of water vapor is discharged and  $\sim 1,500$  t of CO<sub>2</sub> is released through soil diffuse degassing (Chiodini et al., 2001, 2005; Chiodini, 2009). Most of the water vapor condenses near the surface, producing a thermal power flux of  $\sim 100$  MW, and contributing notably to the total water input into the CF hydrothermal system (Chiodini et al., 2005).

Hydrothermal manifestations at Solfatara are connected to the deep structure of the volcano via a complex network of E-W, NE and NW, sub-vertical faults/fractures (see Isaia et al., 2015), which allow fluids to migrate from the deeper hydrothermal reservoir (Bruno et al., 2007, 2017; De Siena et al., 2018). Both ground deformations and seismicity occurring at Solfatara are likely controlled by the pressure and temperature increase of the hydrothermal system, due to repeated, impulsive transfers of high amount of magmatic fluids from depth that exceed the degassing capabilities of the geological medium (Chiodini et al., 2017). It is therefore essential to obtain high resolution images of the faults/fractures and to locate and monitor areas of subsurface fluid accumulation. Active-source, geophysical exploration methods can be profitably used for this task, however, in harsh volcanic environments they usually are not able to achieve an acceptable signal-to-noise (S/N) ratio, which adds up to their intrinsic interpretative non-uniqueness. Improvements in both field data acquisition, processing and interpretation have been tested to overcome these limits. A key role to reduce non-uniqueness and improve the geological interpretation is played by multivariate co-located geophysical data. A first attempt of interpretation of co-located geophysical and geochemical data aimed at imaging the subsurface and elucidate patterns in the shallow subsurface degassing at Solfatara is provided by Bruno et al. (2007). However, when visually comparing low-S/N multivariate datasets bias may be introduced by pre-existing ideas and/or assumptions made by the interpreter.

Unsupervised learning, also known as machine learning, can provide not only useful insights for unbiased geological interpretation but also a feedback to cooperative inversion by finding a statistically robust link between different geophysical parameters. In the last decades, different machine learning algorithms have been used for multivariate geophysical data interpretation (Lary et al., 2016 and references therein, Rodriguez-Galiano et al., 2015). Because of its conceptual simplicity and robustness, the K-means clustering method (Lloyd, 1982; Bock, 2007) is one of the most popular and widely used clustering techniques. Partitioning techniques such as the K-means clustering are known to be less susceptible to outliers and to be computationally more efficient than hierarchical methods (Tronicke et al., 2004). The recent literature provides several examples of effective integration of low-dimensional (i.e., 2D

or 3D) multivariate geophysical datasets, based on K-means algorithms (e.g., Tronicke et al., 2004; Bernardinetti et al., 2017). On the other hands, artificial neural networks are known to be more computationally efficient with higher-dimensional datasets (e.g., Roden et al., 2015). Among these, self-organizing (or Kohonen) maps (SOM: Kohonen, 2013) are clustering methods based on a competitive learning approach. SOM are regarded as one of the most important tools for unsupervised seismic facies analysis (Taner et al., 2001; Coléou et al., 2003).

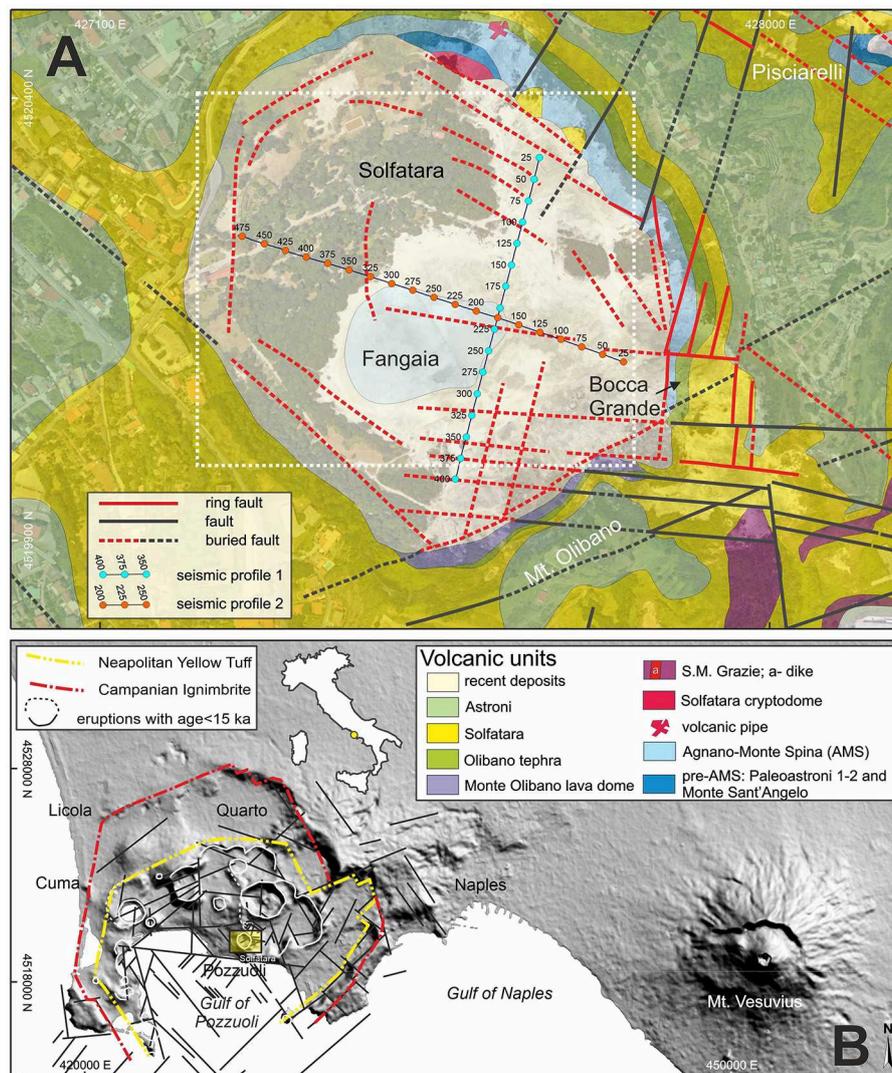
Hereinafter we show and discuss three examples of application of K-means and SOM clustering on multivariate data acquired in the recent past within the Solfatara tuff cone. In the first part of our paper we show and discuss the K-means clustering of a three-dimensional dataset (i.e., spatial variation of seismic noise, Bouguer anomaly, CO<sub>2</sub> flux) recorded at the surface of the crater and discussed in Bruno et al. (2007).

The surface data was also used to plan for the optimal location of the 2D and 3D seismic arrays used in the MED-SUV RICEN project (see Bruno et al., 2017; De Landro et al., 2017; Amoroso et al., 2018 among others). In particular, using the data from the RICEN project Bruno et al. (2017), were able to provide for the first time a high-resolution seismic image of the first 0.8 km of Solfatara crater, while Gresse et al. (2017), estimated a detailed three-dimensional electric resistivity model of Solfatara from the inversion of several electric measurements overlapping to the 2D and 3D seismic arrays. In the second part of our paper, we link, with K-means clustering, the tomographic P-wave velocity profiles from Bruno et al. (2017), with the overlapping electric resistivities estimated by Gresse et al. (2017). This analysis aims at elucidating the near-surface saturation state (i.e., gas vs. fluid) of Solfatara and compare it with the results of the spatial clustering obtained at the surface of the crater.

In the third part of this paper we merge three seismic attributes (i.e., similarity, energy and dip), from the seismic reflection profiles of Bruno et al. (2017), with an additional attribute (GLCM texture: see Haralick et al., 1973) for a two-fold purpose: (1) compare SOM vs. K-means clustering algorithms on this four-dimensional dataset and (2) provide a detailed information about the shallower hydrothermal features (i.e., 0–500 m) beneath the Solfatara.

## GEOLOGICAL SETTING

Solfatara is one of the many monogenic volcanoes of Campi Flegrei (**Figure 1**), an active resurgent caldera (Vitale and Isaia, 2014 and references therein) carved by two massive eruptions: the Campanian Ignimbrite (40 ka, De Vivo et al., 2001) and the Neapolitan Yellow Tuff (15 ka, Deino et al., 2004). CF caldera has undergone recurring inflation/deflation episodes (e.g., De Natale et al., 1991). Sea level measurements, made on ancient Roman artifacts in Pozzuoli suggest a slow deflation occurring in the area. In 1530 AD a 7-m inflation culminated with the Monte Nuovo eruption in 1538 (Di Vito et al., 1987). After 1538, a new period of deflation lasted until 1968, interrupted by two rapid inflations in 1969–1972 (+170 cm) and in 1982–1984 (+182 cm: Berrino et al., 1984). Inflation is still ongoing there. A marked seismic activity



**FIGURE 1 | (A)** Geological and structural map of Solfatara area (Isaia et al., 2015) overlain to a 2D image of the area (© 2019 Google); in red are the ring faults while in green are the tectonic faults; buried faults are represented by dashed lines. The white dashed rectangle shows the locations of the surface geophysical and geochemical measurements (see **Figure 2**). **(B)** DEM image showing the relationships of Solfatara (yellow-filled rectangle) with CF caldera and Somma-Vesuvius. The structural framework of Campi Flegrei Caldera is from Orsi et al. (1996), Vitale and Isaia (2014). Regional faults are inferred by geophysical studies and morphological structures (Florio et al., 1999; Milla et al., 2000; Bruno et al., 2003; Sacchi et al., 2009).

characterizes inflation periods: earthquakes mostly cluster at Solfatara and within the bay of Pozzuoli; they remain contained within the caldera margins and abruptly stop at  $\sim 4$  km deep, suggesting a sharp brittle-to-ductile rock transition. Seismicity is likely to be triggered by an upward migration of an excess of fluid pressure front from magmatic intrusion, and by the brittle readjustment of the inflated system occurring along some lubricated structures (e.g., Gaeta et al., 1998; Bianco et al., 2004; Saccorotti et al., 2007; Cusano et al., 2008).

Solfatara is located  $\sim 1.2$  km to the east of Pozzuoli town (**Figure 1**). The volcano formed during the most recent epoch of volcanic activity at Campi Flegrei at ca. 4200 yr B.P. (Isaia et al., 2009; Smith et al., 2011), which was dominated by explosive

activity, mainly in the central-eastern caldera sector, climaxing with the Plinian eruption of Agnano-Monte Spina (De Vita et al., 1999) at ca. 4.5 ka. The Solfatara eruption was preceded and followed by explosive eruptions from nearby centers (Astroni, Mt. Olibano, S-M- Grazie, Agnano-Monte Spina, and pre-Astroni) whose products are found in the area around the present crater rims (**Figure 1**).

Geophysical, geochemical and geological data allow to infer, with different resolution and degree of uncertainty, the structure and stratigraphy of the upper part of the crater, together with the main features of the upper part of its complex hydrothermal system, made up of a mix of uprising magmatic fluids and meteoric water. High-resolution seismic reflection

profiles (Bruno et al., 2017), show a  $\sim 400$  m deep asymmetrical crater filled by volcanoclastic sediments and rocks and carved within an overall non-reflective pre-eruptive basement pierced by intrusive bodies: their seismic data clearly show several steep and segmented collapse faults affecting the crater filling. Faults generally have normal kinematics and dip toward the crater center. Resistivity tomography surveys, combined with mappings of diffuse  $\text{CO}_2$  flux, ground temperature and self-potential (Byrdina et al., 2014; Gresse et al., 2017) delineate three shallow plume structures: two liquid-dominated conductive plumes below the Fangaia mud-pool and the Pisciarelli fumarole and a gas-dominated plume below Bocca Grande fumarole. In the shallow steam-heated part of Solfatara crater, a predominant argillitic alteration occurs, where sulfuric acid is created at, or above, the water table by the oxidation of  $\text{H}_2\text{S}$  (Rye, 2005). The hydrothermal activity at the surface is characterized by an intense soil diffuse degassing, both inside and outside of the crater (Cardellini et al., 2017 and reference therein). Acidic (pH  $\sim 1.7$ ) high-temperature ( $\geq 160^\circ\text{C}$ ) fumaroles are mainly located in the eastern part of the crater, whereas hot springs, steam-heated pools ( $45\text{--}95^\circ\text{C}$ ), and fumarolic vents are concentrated in its center (Glamoclija et al., 2004; Valentino and Stanzone, 2004; Chiodini et al., 2011). The absence of vegetation (**Figure 1A**) best depicts the area of diffuse soil degassing.

## MATERIALS AND METHODS

### Datasets and Integration

As briefly outlined before, we use three different multivariate datasets in this paper. The first is made of three co-located spatial geophysical and geochemical data shown in **Supplementary Figure 1** and whose acquisition and processing is discussed in detail in Bruno et al. (2007). Specifically, we had four data available: (a) the high-resolution Bouguer anomaly; (b) the  $\log_{10}$  of the  $\text{CO}_2$  flux; (c): the soil temperature; and (d): the mean environmental seismic noise level measured within the 10–15 Hz frequency window on the amplitude spectrum. However, **Supplementary Figure 1** shows that both the  $\text{CO}_2$  flux and the soil temperature bear the same type of information: i.e., temperature anomalies at Solfatara are generated by diffuse degassing of hot  $\text{CO}_2$  gas through soil. Therefore, we decided to remove one of the two datasets (i.e., soil temperature) before clustering to avoid assigning a higher “a priori” weight to the high-temperature degassing phenomenon with respect to the other parameters.

The second, bi-variate, dataset is made by merging the shallow P-wave velocities estimated by Bruno et al. (2017) from the tomographic inversion of the first-arrival traveltimes (SRT) with the electric resistivity (ERT) estimated by Gresse et al. (2017) from the inversion of Wenner/Schlumberger profiles that are in part overlapping with the two seismic arrays. For both ERT and SRT images the co-located grid has the same progressive of the seismic dataset (**Figure 1**). The tomographic and electric profiles are shown in **Supplementary Figures 2, 3**. Electric resistivity and p-wave velocity do not show evident correlations.

The third dataset is made of four seismic attributes (**Supplementary Figures 4, 5**). Similarity, energy and dip

attributes are computed and described by Bruno et al. (2017), on the depth-converted CRS stacks along the two profiles of **Figure 1**. We merged them with a new attribute (GLCM texture: see Haralick et al., 1973), computed with a step out  $\text{nr} = 3$  and using a GLCM size of  $32 \times 32$ . All four chosen attributes were computed using a depth gate of 10 m. The energy attribute is a measure of reflectivity strength within the chosen depth-gate, while similarity is a multi-trace attribute that returns trace-to-trace similarity properties. It ranges between 0 and 1: a similarity of 1 means that the trace segments are identical in waveform and amplitude. Similarity is the best indicator of structural discontinuity: faults and fractures are generally visible as narrow low-similarity areas. The dip-angle attribute provides the apparent angle of dip (in degree) of seismic features within the profiles. The dip angle is computed from the dip-steering process that produces a steering cube, i.e., a volume that stores information about the seismic dip of coherent events at every sample position. Finally, the entropy attribute is part of the Texture-Directional attributes. This class of attributes uses the gray level co-occurrence matrix (GLCM) and its derived attributes are tools for image classification that were initially described by Haralick et al. (1973). The GLCM entropy attribute measures the disorderliness (or roughness) of the patch of seismic amplitude values; maximum entropy occurs when all probabilities of values are equal and therefore result in a random distribution of values (Malleswar et al., 2010).

The four seismic attributes described above were chosen among many possible others with the aid of Principal Component Analysis (Abdi and Williams, 2010; Kassambara, 2017), a preliminary analysis needed, for assessing the most representative attributes to use for the subsequent processing. Principal component analysis represents a rotation of the multi-dimensional point cloud so that the maximum variability is projected onto the pair-wise combination of axes (Prasad et al., 2005). The contribution of attributes and their quality were evaluated in the correlation circle (Abdi and Williams, 2010), shown in **Supplementary Figure 6**. Energy and similarity showed a better representation on the compared principal component, while dip angle showed the lowest values. However, the calculated values do not differ so much to justify the exclusion of dip angle from the analysis. The variances for the first two dimensions are respectively 34.7 and 29.1% (**Supplementary Figure 6**), their ratio is equal to 1.19, and following Kohonen (2014) these values were used to create the neural networks.

All processing steps were performed using algorithms written in MATLAB<sup>®</sup> and R<sup>®</sup>. As first preprocessing step, we removed from the three dataset those measurements that were outside the overlapping areas; then overlapping datasets were interpolated and resampled over a common grid. Many different approaches can be chosen for this step. For example, to avoid a loss of data to preserve the information content of both models Bedrosian et al. (2007), interpolated a 2D magnetotelluric and seismic dataset using the grid dimension of the higher-resolution data. For the tri-variate spatial dataset we followed a similar approach: we obtained a  $50 \times 50$  square matrix (see **Figure 1**, dashed box) with an individual cell-size of  $10 \times 8.6$  m. For the bi-variate ( $V_p$ ,  $\rho$ )

dataset, ERT and SRT were sampled with a uniform grid spacing of  $3 \times 3$  m, a compromise between the more resolved SRT and the less resolved ERT. The 4D seismic attributes were instead exported from OpenDtect with a common depth sampling of 1 m and spatial sampling of 1 CDP (i.e., 1 m). Seismic attributes were limited to a depth of 600 m below the datum (i.e., 97 m a.s.l.) instead of the original 800 m because very few seismic features of interest are within the 600–800 m depth range. We obtained therefore two very dense four-dimensional matrices sampled along grids of  $372 \times 600$  nodes for Profile 1 and  $451 \times 600$  nodes for profile 2.

Since the geophysical and geochemical measurements have different ranges and standard deviations, then one dataset might dominate the distance used in clustering. Therefore, some preliminary data standardization is necessary. We used a technique known as “data sphering” (Koivunen and Kostinski, 1999), a linear transformation that converts a vector with known covariance matrix into a set of new variables whose covariance is the identity matrix, meaning that they are uncorrelated and each one has a variance of 1. For each multidimensional matrix discussed above, the p-dimensional sample mean  $\bar{x}$  and the covariance matrix  $S$  were computed, and finally, the data were sphered using the following transformation (Martinez and Martinez, 2005):

$$Z_i = \Delta^{\frac{1}{2}} Q^T (x_i - \bar{x}); \quad (1)$$

where  $Q$  are the eigenvectors obtained from the covariance matrix  $S$ ,  $\Delta$  is a diagonal matrix of corresponding eigenvalues,  $x_i$  is the  $i$ -th sample for any geophysical variable and  $\bar{x}_m$  is its average value.  $Z_i$  is the new scaled multivariate dataset suitable as input for data mining processing.

## K-Means Clustering

The K-means (Lloyd, 1982), is an unsupervised method used in data exploratory analysis to find similar observations. K-means is a partitional, non-hierarchical and unsupervised clustering algorithm, which allows separating a dataset in “ $k$ ” clusters based on distances among points. The objective function of the algorithm is to minimize the within-cluster variance and to maximize it among different clusters. The within-class scatter matrix,  $S_w$ , is defined as Martinez and Martinez (2005):

$$S_w = \frac{1}{n} \sum_{j=1}^m \sum_{i=1}^n I_{ij} (x_i - \bar{x}) (x_i - \bar{x})^T; \quad (2)$$

where  $I_{ij}$  is one if  $x_i$  belongs to group  $j$  and zero otherwise, and  $m$  is the number of groups. The criterion that is minimized in the K-means is the sum of the diagonal elements of  $S_w$ . Everitt and Dunn (2001) show that minimizing the sum of the diagonal elements of  $S_w$  is equivalent to minimizing the sum of the squared Euclidean distances between the individual elements  $x_i$  and their group mean.

In general, the algorithm is initialized by randomly defining an initial number “ $K$ ” of centroids and assigning each observation to its closest centroid using the Euclidean distance between the observation and the cluster centroid. The second step of the procedure is to calculate the new centroids (i.e., the new mean

values) using the assigned observations. These steps are repeated until there are no changes in cluster membership or until the centroids do not change (Späth, 1980).

Like many other types of numerical minimizations, the algorithm may converge into a local minimum. This often depends on the initial choice of the centroids. To choose the initial centroids based on the data, we initialized the algorithm performing a preliminary clustering on a random 10% subsample of the entire dataset, with the option “start” and “cluster”. Moreover, MATLAB’s “kmeans” function allows the use of the parameter “Replicates” to overcome the above-mentioned problem of falling into a local minima. Setting a parameter higher than one in “Replicates” instructs the algorithm to begin from a different set of initial centroids, therefore, even if sometimes “kmeans” finds more than one local minimum, the final solution that the function returns is the one with the lowest total sum of distances, over all replicates. We used 20 replicates for our analysis as it was the minimum number that returned the same final results.

Another issue with the K-means algorithm is that the choice of the number of clusters,  $K$ , is arbitrary, meaning that the optimal number  $K$  has to be found using some statistical criterion. Many tools have been used in assessing the quality and optimal number of clustering as well as the degree with which a clustering scheme fits a specific data set (Halkidi et al., 2001); each tool has its own advantages and disadvantages. To assess both the optimal number  $K$  and the degree with which our clustering scheme fits our specific data sets, we used the Silhouette Index (SI: Kaufman and Rousseeuw, 1990; Brock et al., 2008). The silhouette index compares the distances of every  $i$ -th observation within a cluster with the average extra-cluster distances. In other words, the silhouette index measures the degree of confidence in the clustering assignment of a given observation. For an observation  $i$ , the silhouette index is defined as Rawashdeh and Ralescu (2012):

$$SI_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}; \quad (3)$$

where:  $a_i$  is the average dissimilarity of  $i$  to all other observations in its own cluster and  $b_i$  is the minimum value of all average dissimilarity of  $i$  to all observations in any other cluster  $c$ . The average silhouette index is found by averaging  $SI_i$  over all observations:

$$\bar{SI} = \frac{1}{n} \sum_{i=1}^n SI_i; \quad (4)$$

The silhouette index in Equation 3 ranges from  $-1$  to  $1$ . Values  $<0$  and close to  $1$  mean that the observation is well-matched to its own cluster, values near to  $0$  highlight observations with unclear assignment (either to the current cluster or to the nearest one), while values  $<0$  are typical of misclustered observations. Compared with the R-Squared index (Sharma, 1996), which reveals the optimal clustering at the “knee,” the silhouette index provides less chances to choose the wrong number of clusters: this happens because finding the “knee” in noisy data is not as obvious as it could appear. While using the silhouette index, the

optimal number of clusters  $K$  is where the average silhouette index (Equation 4) reaches its maximum value.

Therefore, not only the silhouette index allows evaluating the optimal number of cluster but also it provides an assessment of cluster membership. We used a hard-clustering technique for our K-means analysis, meaning that each measurement can belong only to one cluster. A hard-clustering representation in the original space shows all observations belonging to the same cluster with the same color. Therefore, transactions among clusters are graphically represented by color changes. However, to account for uncertainty in cluster membership we merged the results of the silhouette index with the K-means by saturating the color of each single measurement according to its SI. This technique is similar in principle to the technique applied by Paasche et al. (2010), where the color saturation of their images is also based on a membership function.

## Self-Organizing Maps

Self-Organizing Maps (SOM) are a type of neural network suitable for unsupervised learning (Kohonen, 1997), that uses a competitive learning strategy. The SOM transforms a feature vector of arbitrary dimension drawn from the given feature space into a simplified, generally two-dimensional, discrete map (Klose, 2006). This is achieved in a manner that neurons physically located close to each other in the output layer of the SOM have similar input patterns (Kalteh et al., 2008). The SOM network preserves the original topology and delivers an intuitive visual representation of the clustering; the mapping produced by SOM is usually of the type many-to-one, i.e., the projection images on the SOM are local averages of the input data, which is comparable to the K-means averages (Gersho, 1982; Gray, 1984).

A SOM neural network is structured in two layers: an input layer and a Kohonen layer. In most applications the Kohonen layer represents a structure with a single two-dimensional map consisting of neurons arranged in rows and columns. Each neuron of the Kohonen layer is fixed and is fully connected with all source neurons in the input layer. The input variables, can be represented as vectors of the type  $\vec{x} = \{x_1, x_2, x_3, \dots, x_n\}^T$  in the space  $R^n$  where  $n$  is the dimension of the input space (Wehrens and Buydens, 2007). In our case, the vector space  $n$  has four-dimensions (i.e., the four seismic attributes used for our analysis). The objective of the algorithm is to organize the input seismic attributes into a geometric structure. If the map has  $q$  neurons there are  $q$  prototype vectors, or weights, defined as:

$$\vec{\omega}_k = \left\{ \omega_{(k,S)}, \omega_{(k,E)}, \omega_{(k,D)}, \omega_{(k,H)} \right\}^T \quad (5)$$

Where:  $S$  is the Similarity attribute,  $E$  is the Energy,  $D$  is the Dip,  $H$  is the Entropy and  $K=1, 2, \dots, q$ . The codebook vectors are initialized as random values.  $\vec{\omega}_k$  connects the  $n$  input layer neurons to the total number of neurons  $q$  in the Kohonen layer. Learning occurs during the self-organizing procedure as the input vectors  $\vec{x} = \{S, E, D, H\}^T$  are presented to the input layer of the network. The weights  $\vec{\omega}_k$  are used to determine only

one stimulated neuron in the Kohonen layer after the “winner-takes-all” principle that can be summarized as follows: for each  $\vec{x}$ , the Kohonen neurons compute their respective values of a discriminant function (i.e., Euclidean distance ( $\|\vec{x}_i - \vec{\omega}_k\|$ )). These values are used to define the winner neuron. That means the network determines the index  $j$  of that neuron, whose weight  $\vec{\omega}_k$  is the closest to vector  $\vec{x}_i$  by:

$$j(\vec{x}_i) = \arg \min \|\vec{x}_i - \vec{\omega}_k\| \quad |k=1, 2, \dots, m, \dots, q \quad (6)$$

Afterwards, the learning procedure modifies the weights  $\vec{\omega}_j$  of the winner neuron and the winner neighborhood.

$$\vec{\omega}_k(t+1) = \vec{\omega}_k(t) + \eta(t) \cdot h_{k,j(\vec{x}_i)}(t) \cdot [\vec{x}_i(t) - \vec{\omega}_k(t)]; \quad (7)$$

where  $t$  denotes the iteration number,  $\eta(t)$  is the learning-rate parameter during the calculation step  $t$ , and  $h_{k,j(\vec{x}_i)}$  is the neighborhood function centered around the winning neuron  $j(\vec{x}_i)$ . The learning rate  $\eta(t)$  is usually a small value in the order of 0.05, which decreases during the training so that the map converges. The size of the neighborhood function also decreases during training and eventually only the winning units are modified (Wehrens and Buydens, 2007). Thus, the codebook vectors are updated at each iteration and the algorithm terminates after a predefined number of iterations.

The initialization parameters have their importance, because also the SOM, as for the K-means, can get trapped in a local minimum solution. As for the K-means, repeated training of a SOM will lead sometimes to a rather different mapping, because of the random initialization. However, the conclusions drawn from the map remain remarkably consistent among different initialization of the same data, which makes the SOM a very useful tool in many circumstances (Wehrens and Buydens, 2007).

The initialization parameters (such as the grid size and the number of iterations) are important to a successful SOM analysis. As recommended by Vesanto and Alhoniemi (2000), the grid should have a number of nodes well above the number of real clusters in the dataset. One may have to test several grid sizes to check if the cluster structures are shown with a sufficient resolution and statistical accuracy (Kohonen, 2014). To define the shape of the map we need to compute the first two principal dimensions in which the variances of the input dataset are. The ratio between the number of neurons in the two directions of the grid is proportional to the ratio between the two principal components (Kohonen, 2014). Here, following Abdi and Williams (2010), we assessed the quality of the representation of the variables on the factor map by using the “square cosine, squared coordinates” (i.e.,  $\cos^2$ ) method. A high  $\cos^2$  highlights a good representation of the variable on the principal component (i.e., the variable is close to the circumference of the correlation circle). A low  $\cos^2$  suggests that the variable is not perfectly represented by the principal components (i.e., it is close to the center of the circle).

The nets sizes used by us were computed taking into account the ratio of the number of neurons according to the following

formula (Kohonen, 2014):

$$R = \frac{\text{var1}}{\text{var2}} \cong \frac{n1}{n2} \quad (8)$$

where *var1* and *var2* are the variance for the first and second principal component respectively, while *n1* and *n2* are the number of neurons along the x and y directions of the neural network. The relationship between the neurons constituting the two dimensions of the network is derived from the ratio of variance between the first two principal components. Through Equation 8 therefore, we determine the value of *n1* by imposing *n2*. Differently from the empiric approach proposed by several authors (e.g., C er ghino and Park, 2009), where sizing of a neural network is based on the number of observations constituting the multivariate dataset, we prefer to use a smaller number of neurons but significantly greater than the number of expected seismic facies. The optimal size of the network is identified by successive attempts, using Equation 8, and increasing the number of neurons until we obtain a SOM image that best represents the expected geophysical/geological features. Our approach allows to represent all neurons in the network with a different color, through a color map based on the position of the neurons as discussed below. Differently from C er ghino and Park (2009), our approach allows shorter computation times and preserve all the details, avoiding a possible loss of information due to the need (in order to have a concise representation of the starting space) of grouping together the high number of neurons, often resulting from the empirical approach of C er ghino and Park (2009). In fact, Unglert et al. (2016), show that the clustering process applied to the SOM results may fail to regroup the neurons in a manner that is consistent with the input dataset presented to the SOM.

With the aim to allow the neural network to replicate the topological distribution of the dataset, we presented it to the SOM 9,000 times. We inspected the training progress by checking the average distance of an object with the closest codebook vector (Wehrens and Buydens, 2007). Moreover, for the neurons a hexagonal connectivity was adopted resulting with a number of six neighbors for the inner neurons, with a planar topology for the lattice.

## Displaying SOM Maps

Once the SOM algorithm has converged, the two-dimensional feature maps of Kohonen neurons display the following important statistical characteristics of the represented feature space (Klose, 2006):

*Approximation*: A feature map represented by a set of weights in the Kohonen layer provides a good approximation to the input space. *Topological ordering*: The two-dimensional feature map is topologically ordered in the sense that similar Kohonen layer neurons correspond to similar feature vectors of the higher dimensional input space. *Density matching*: The feature map reflects variations in the statistics of the distribution of the original feature space: regions in the input space from which sample vectors are drawn with a high probability of occurrence are mapped onto larger domains in the Kohonen layer, and therefore with better resolution than regions in the input space

where sample vectors are drawn with a low probability of occurrence.

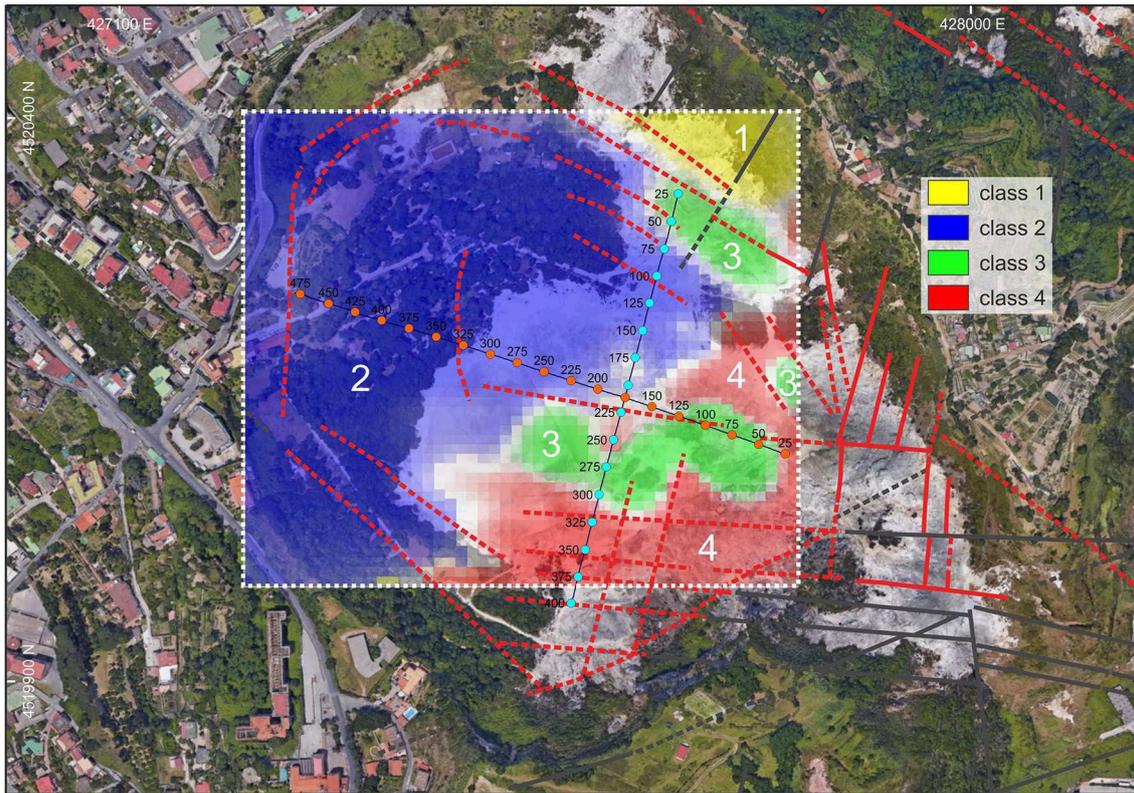
Since the SOM preserve the topological ordering of the input space (i.e., similar input features are classified within the same neuron or within adjacent neurons) it makes sense to display the results using a color map based on the relative position of neurons. The most used method to visualize the cluster structure of a SOM is based on the unified distances matrix (U-matrix; Ultsch, 1993). This is a visualization technique that allows to represent the structural properties of the high dimensional input space with a color scale based on the relative distance between the prototype vectors and the winner unit and it is conceptually similar to the gradational color scales used to represent data from digital elevation models. A synoptic view of the input multivariate dataset is also needed for result interpretation, to identify features across the neurons and highlight correlations that in our case could be interpreted as signatures of hydrothermal features. Here we adopted the "heatmaps" (see Pryke et al., 2007) as a way to represent the average values of input observations classified into each neuron. The heatmaps, computed by averaging the variables classified into each neuron, are powerful ways to show the distribution of the different variables (i.e., seismic attributes) across the neural network, thus facilitating a visual interpretation of the classification provided by the SOM. The variable range is represented by a gradational color scale, in which we assign cold colors to the lowest values and hot colors for the highest values (**Supplementary Figure 7**).

The method we propose to graphically represent distances between neurons uses a RGB color gradation based on both the position of neurons and on the values of U-matrix, to consistently image input seismic facies in the original space in the output 2D lattice (**Supplementary Figure 7A**). The *x,y* neuron coordinates in the 2D lattice provide the R and G values while the U-matrix value provides the B value. All RGB values are scaled to fit the range from 0–255. This RGB color map will show similar colors for adjacent neurons, while the B U-matrix value is useful to highlight distances of neurons, showing how far are the observations. This method is analog to the coloring scheme used by Langer et al. (2009) with the difference that Langer et al. (2009) did not use the U-matrix value in coding the RGB colors.

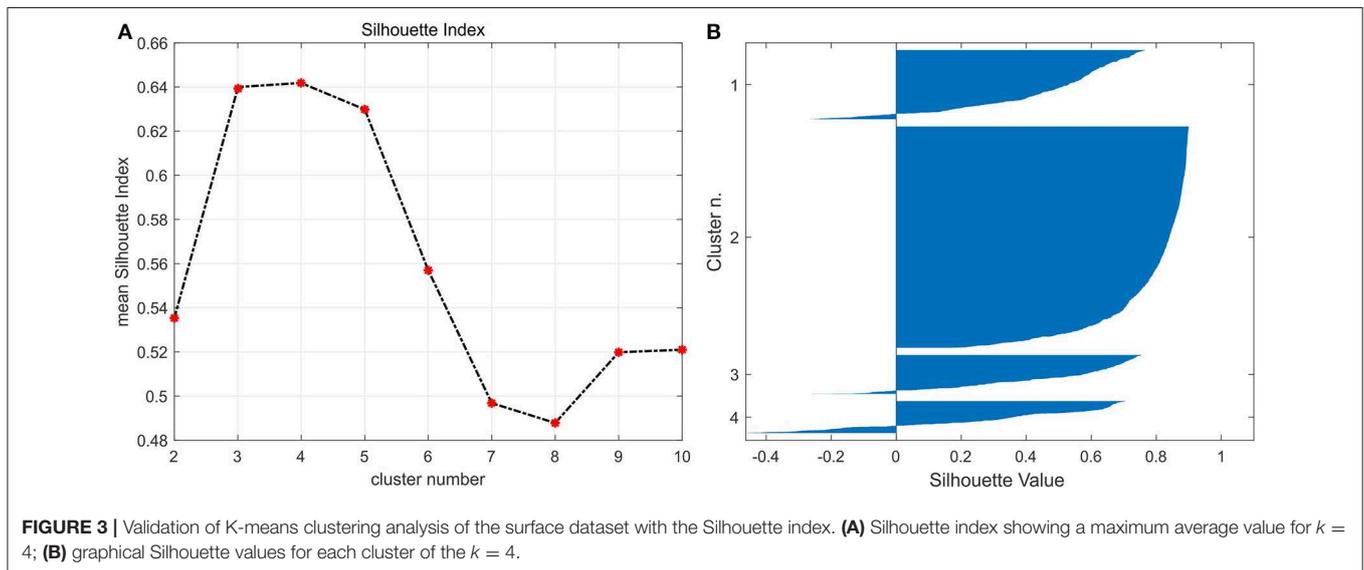
## RESULTS

### Surface Data Integration

The results obtained by K-means on the tri-variate dataset composed by the Bouguer anomaly, CO<sub>2</sub> flux and seismic noise are in **Figure 2**. The optimal cluster number, assessed by the silhouette index is *K* = 4 (**Figure 3**). Light colored pixels are those characterized by a low silhouette index and therefore of uncertain collocation between adjacent clusters. Misclustered pixels (i.e., SI < 0) are colored in white. The clusters boundaries show a prevailing NE and WNW trend and an excellent agreement with the patterns of the intracratonic tectonic structures mapped by Isaia et al. (2015). This agreement clearly demonstrates that the surface distribution of the analyzed geophysical and geochemical data, which reflects the undergoing



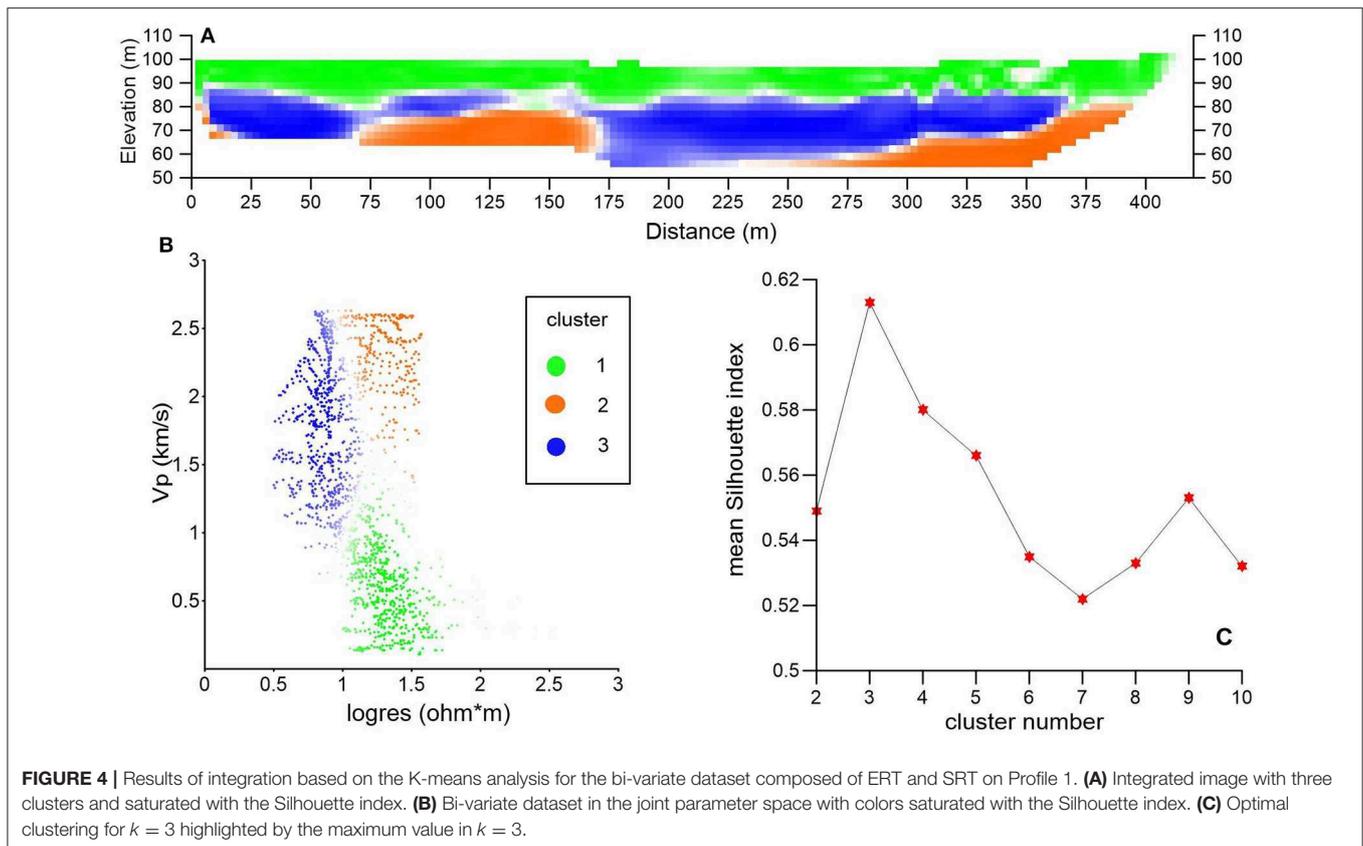
**FIGURE 2** | 2D image of the Solfatara area (© 2019 Google) as in **Figure 1** with results of surface dataset integration by means of K-means clustering overlaid with the main faults from Isaia et al. (2015). Dots in light blue are the CDP metrics of Profile 1 while dots in orange are the CDP metrics of Profile 2.



**FIGURE 3** | Validation of K-means clustering analysis of the surface dataset with the Silhouette index. **(A)** Silhouette index showing a maximum average value for  $k = 4$ ; **(B)** graphical Silhouette values for each cluster of the  $k = 4$ .

hydrothermal activity at Solfatara, is mainly influenced by the tectonic framework. This evidence is clearer on the outcomes of the cluster analysis (**Figure 2**) rather than on the individual measurements of **Supplementary Figure 1**.

**Supplementary Figure 1** overlays the hard cluster partitioning obtained by the K-means analysis to the spatial distribution of the three variables used for the clustering (plus the soil temperature, which was not included in the analysis). A



qualitative assessment of the average value of each variable in each of the four clusters is reported in the **Table 1**.

Cluster 2 (in blue in **Figure 2**) has the largest surface extension and it is overall characterized by low values of the three analyzed variables. It highlights the most stable sector of Solfatara not much affected by the hydrothermal dynamics. Vice-versa the most active areas (such as the “Fangaia” and the “Stufe di Nerone” fumaroles) are highlighted by cluster 3 (green), which combines high values of  $\text{CO}_2$  flow and seismic noise with low values of the Bouguer anomaly. Cluster 3 most likely emphasizes the surface location of the main degassing pathways.

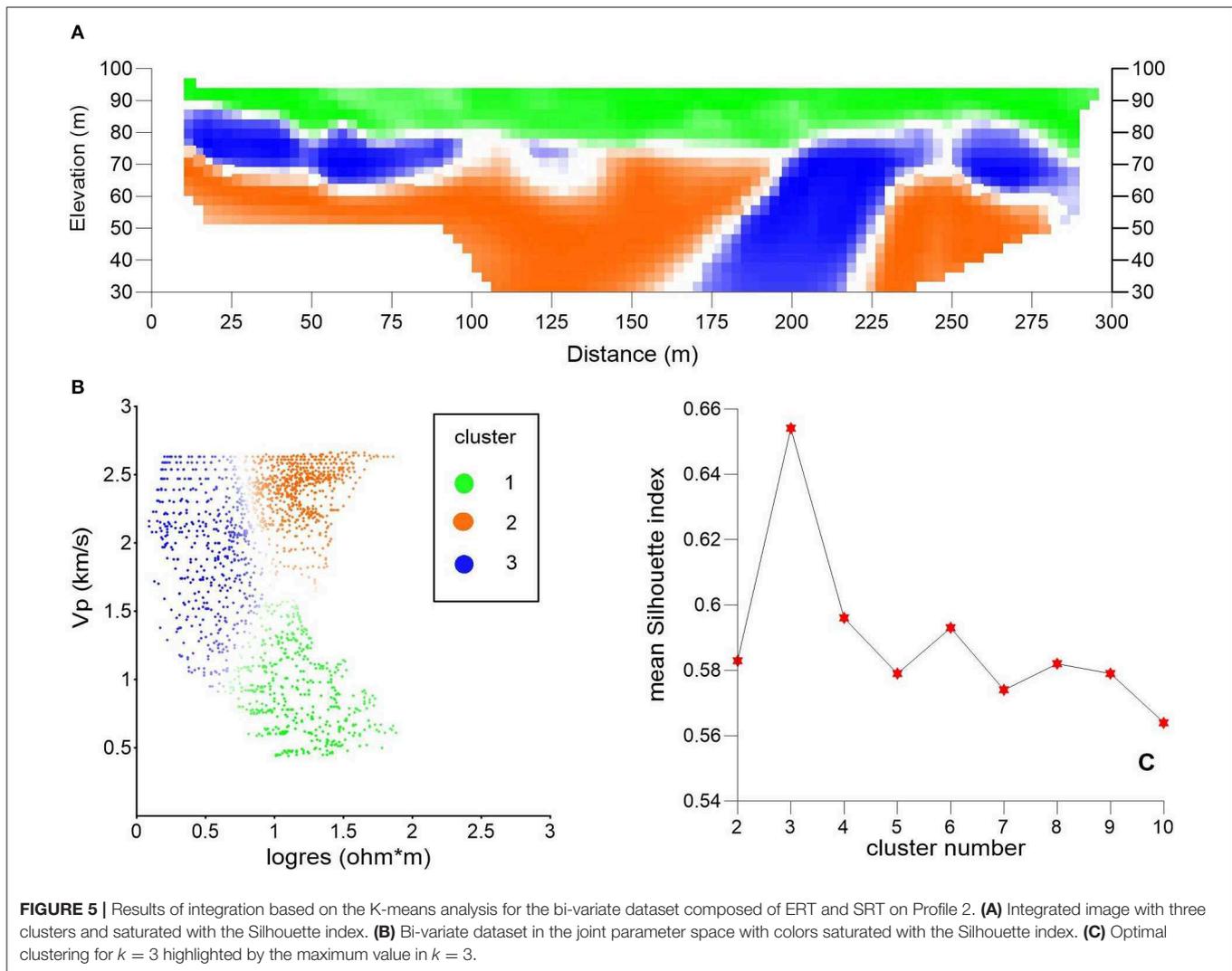
Cluster 4 (red) alternates spatially with cluster 3 and is characterized by low- to very-low values of Bouguer anomaly and of  $\text{CO}_2$ , which it is considered to be, together with the surface temperature, the main indicator of magmatic degassing, and by the highest values of seismic noise. In general, the high levels of seismic noise characterizing both clusters 3 and 4 can be associated with the intense vibrational activity connected with the uprising fluxes.

Finally, cluster 1, characterized by very high values of Bouguer anomaly and average values of the other two variables, has the smallest extension and it is only found in the NE corner of the investigated area, close to the Solfatara cryptodome and to a volcanic pipe (**Figure 1**). The high bouguer anomaly within cluster 4 can be therefore explained by a higher rock density of the crater rims with respect to the volcanoclastic material filling the crater.

## Near-Surface Resistivity and P-Wave Velocity Integration

The clustering of the electrical resistivity and the P-wave seismic tomography along the two profiles shown in **Figure 1** was done with the purpose of assessing the hydrological properties (i.e., gas-saturated vs. water saturated porous media and degree of saturation) of the near surface, since both the electrical resistivity and the P-wave velocity are sensitive to changes of water/gas saturation. The analysis of the silhouette index revealed that the optimal number clusters is consistently 3 for both Profile 1 (**Figure 4**) and Profile 2 (**Figure 5**). Moreover, both the cross-plots in **Figures 4B**, **5B** show that three clusters are well-separated. As for the previous analysis we report in the **Table 2** a qualitative assessment of the average value of each variable in each of the three clusters.

In both cross-plots, cluster 1 (in a green color) combines low values of P-wave velocity with fair to very high values of electric resistivity. These values are typical of loose, unsaturated to partly saturated tephra affected by diffuse  $\text{CO}_2$  degassing therefore confirming the findings of Chiodini et al. (2001), Bruno et al. (2007, 2017), and Gresse et al. (2017). As it can be seen in both **Figures 4A**, **5A**, cluster 1 occupies the topmost part of both profiles for an average thickness of 10–30 m. Clusters 2 (orange) and cluster 3 (blue) are both characterized by a fair-to-very high P-wave velocity (higher for cluster 2) but while cluster 3 has very low values of resistivity, cluster 2 is instead characterized by a higher resistivity. We tentatively associate cluster 2 with



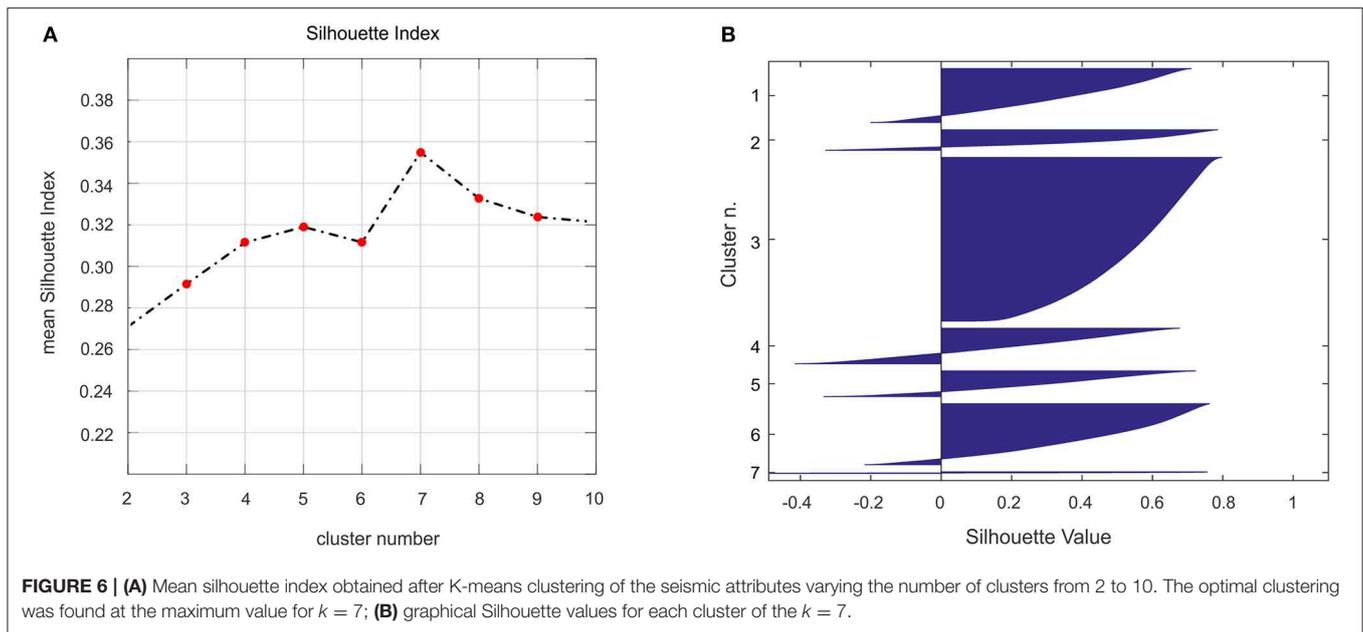
**FIGURE 5** | Results of integration based on the K-means analysis for the bi-variate dataset composed of ERT and SRT on Profile 2. **(A)** Integrated image with three clusters and saturated with the Silhouette index. **(B)** Bi-variate dataset in the joint parameter space with colors saturated with the Silhouette index. **(C)** Optimal clustering for  $k = 3$  highlighted by the maximum value in  $k = 3$ .

predominantly gas-saturated porous media and cluster 3 with predominantly water-saturated porous media. P-wave velocity should decrease in the presence of gas: therefore, high-values of P-wave velocity in cluster 2 are a bit odd and can be explained hypothesizing an overall high-water saturation (i.e., presence of a two-phase fluid) even in the predominantly gas-saturated area. Both clusters are found at depths  $<10\text{--}30\text{ m}$  in the integrated images in **Figures 4A, 5A**.

### Seismic Attribute Integration

Seismic attributes were integrated using both the K-means and SOM techniques. For the K-means clustering, the mean Silhouette Index indicates an optimal value of 7 clusters (**Figure 6**), for which we qualitatively assess the average ranges of each of the four seismic attributes in **Table 3**. The subsurface 7-cluster distribution can be observed in **Figures 7, 8**, where we can also compare the results between SOM and K-means. For example, on both profiles (**Figures 7, 8**) and on the depth range  $250\text{--}450\text{ m}$  cluster 1 (red) highlights elongated features

characterized by high to very high entropy, dip and similarity that can be interpreted as intrusive bodies (dikes). Instead, only on profile 1 and in a shallower ( $100\text{--}300\text{ m}$ ) depth range, cluster 5 (beige) and to a lesser extent cluster 4 (yellow) and 1, show some peculiar sub-vertical thin patterns that cross-cut or are found below areas belonging to cluster 2. These features are not evident with the same clearness on Profile 2 (**Figure 8**) because, as discussed in Bruno et al. (2017), this profile is characterized by a lower S/N ratio possibly correlated to a higher structural complexity of the Solfatara along the WNW-ESE direction (Isaia et al., 2015). In both profiles, clusters 2 (dk green) and 7 (purple) highlight sub-horizontal structures with high energy and similarity. Cluster 3 (green) features areas overall characterized by high-reflectivity in profile 1 while it is not discriminant of any facies in profile 2, due to lower S/N ratio. Cluster 3 and 6 (lt. green) have similar average values, however in profile 1 cluster 3 has a spatial distribution that highlights the upper filling of the crater down to a depth of  $\sim 350\text{ m}$ .



**TABLE 1 |** Qualitative average values of Bouguer anomaly, CO<sub>2</sub> flow and Seismic noise amplitude in each of the four clusters.

Cluster	1 (Yellow)	2 (Blue)	3 (Green)	4 (Red)
Bouguer anomaly	Very high	Fair/low	Low/very low	Very low
CO <sub>2</sub> flow (log <sub>10</sub> )	Fair to low	Fair	High/very high	Low/very low
Seismic noise amp.	Fair/high Uncertain	Low/very low Stable	High Unstable (high)	Very high Unstable

**TABLE 2 |** Qualitative average values of electric resistivity and P-wave velocity in each of the three clusters.

Cluster	1 (Green)	2 (Orange)	3 (Blue)
Electric resistivity	Fair/very high	Fair/high	Low/very low
P-wave velocity	Very low/low Unsaturated	High/very high Water-saturated	Fair/very high Gas-saturated

As discussed above, we also performed the SOM analysis on the seismic attribute datasets using different neural networks with increasing size ( $6 \times 5$ ,  $8 \times 7$  and  $12 \times 10$  nodes). The  $12 \times 10$ -node network provided the best details in the integrated images for both profiles 1 and 2. Those node maps are also shown in **Figures 7, 8** and overall display more consistent results and better details than the corresponding K-means images. As explained before, in order to reflect the distribution of attributes across the lattice, the SOM results are displayed by using an RGB colormap based on the geometry of the neural network and on the U-matrix values. By matching the SOM and K-means images and by a synoptic comparison of RGB colormap (**Supplementary Figure 7A**), heatmaps (**Supplementary Figures 7C–F**) and K-means clusters

of the two integrated profiles (**Figures 7, 8**) five main volcanic/hydrothermal facies could be inferred. These facies include all the 120 SOM neurons of **Supplementary Figure 7A**.

Facies 1 groups neurons 1–4, 13–16, 25–27, 37–39, 49–50 in **Supplementary Figure 7A**, and roughly corresponds to cluster 2 (i.e., dark green) and 7 (i.e., purple) of the K-means analysis. Such facies is characterized by fair to very high energy, very high similarity, very low values of entropy and by sub-horizontal dips (see **Supplementary Figures 4, 5**) and is visible on Profiles 1–2 as sub-horizontal to slightly dipping, dark-colored zones in the depth range 50–300 m on the two profiles (**Figures 7, 8**). We interpret them as sub-horizontal areas of fluid (both gas and water) accumulation within the crater.

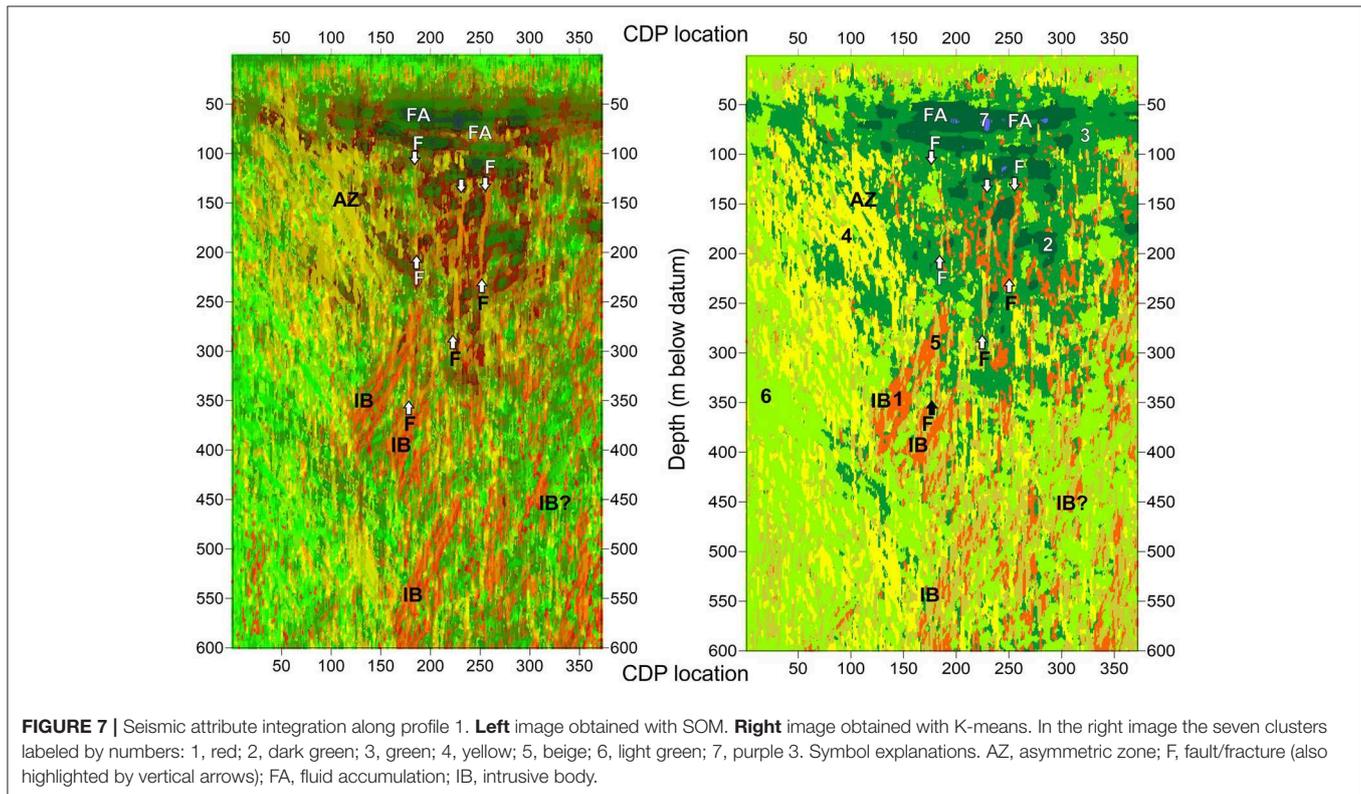
Neurons n. 57–59, 69–72, 82–84, 96, 106–108 in **Supplementary Figure 7A**, with high dip, very low energy, high entropy and low similarity, are instead grouped into Facies 2. Such facies includes cluster 5 and in part 1 and 4 of the K-means analysis. Neurons here show characteristic sub-vertical thin patterns and are interpreted as narrow, faults/fractures filled with fluids of hydrothermal origin. They are extremely well evident at CDP 100–300 and at depths of 100–500 on the image of profile 1 (**Figure 7A**).

Facies 3 groups neurons 81, 94, and 95 that have fair to low energy, high right-dip and entropy and fair similarity and are mainly present on the left side of the section from the depth of about 50 m up to 300 m, creating an asymmetric zone labeled as AZ (i.e., asymmetric zone) on the section of Profile 1 (**Figure 7A**). These features, which are comparable in part to cluster 4 (i.e., yellow) in **Table 3** may represent old, low permeability faults/fractures or other dipping bodies (intrusions).

Neurons 22, 23, 33–35, 46, 47 are grouped in Facies 4, overall characterized by high dip, mainly left-dipping, fair to low similarity, low energy and fair to high entropy. This facies roughly corresponds to cluster 1 of the K-means analysis. We

**TABLE 3** | Qualitative average values of similarity, abs(dip), energy, and entropy in each of the seven clusters.

Cluster	1 (Red)	2 (dk green)	3 (Green)	4 (Yellow)	5 (Beige)	6 (lt. green)	7 (Purple)
Similarity	High	Very high	High	Fair/low	Low/very low	High	Very high
abs(dip)	High/very high	Very low	Fair	High/very high	High/very high	Low	Very low
Energy	Low	Fair/high	Low/fair	Very low	Very low	Very low	Very high
Entropy	Very high	Low/fair	Very high	High	High/very high	High	Low/fair
	Fault or intrusion	Fluid accumulation	High reflectivity zone	Low reflectivity zone	Fault or fracture	Non-reflective zone	Bright spot



associate it to intrusions with high-dip angle (dykes) visible mainly in the 300–600 m depth range.

Finally, facies 5 groups instead all remaining neurons in **Supplementary Figure 7A**, which, in lack of well logs and other geological constraint we were unable to associate to any peculiar seismic feature. It correspond roughly to clusters 4 and 6 of the K-means analysis. In general observations colored with a light green-yellow in **Figures 7, 8** have very low energy, variable dip, fair similarity and entropy and are distributed at the borders of the images.

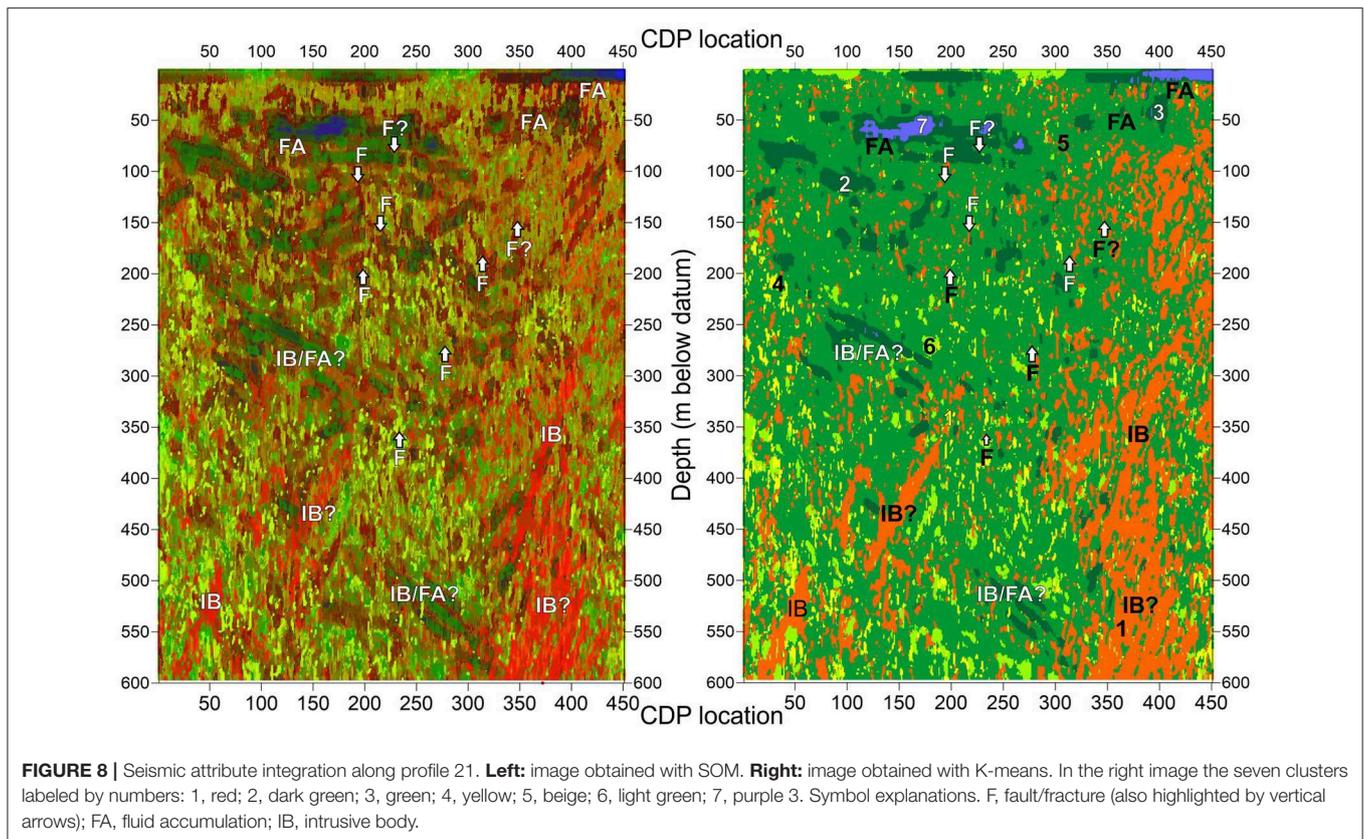
## DISCUSSION

### Results Comparison

**Figures 9, 10** show two three-dimensional cross-sections of Solfatara crater along the NNE and WNW direction (i.e., profiles 1–2). With these figures we compare: (1) the K-means

clustering of the surface data with (2) the K-means clustering of resistivity and P-wave velocity and (3) with the SOM analysis on seismic attributes. For the last dataset we choose to display the SOM results and not the K-means clustering because, as better explained in the following sections, the SOM results are far more informative than the corresponding K-means images, especially along profile 2.

The SOM images show, at 50–350 m deep, well-definite areas of fluid accumulation (Facies 1, labeled as FA in **Figures 9, 10**). Bruno et al. (2017) interpret these zones as stratigraphic traps in which uprising hydrothermal fluids (both gas and water) accumulate, fed by an underlying sub-vertical network of faults and fractures (i.e., Facies 2). These faults are evident in our SOM results as narrow sub-vertical features (i.e., Facies 2) starting at depths of ~100 m (i.e., 0 m a.s.l.) and intersecting the deeper areas of fluid accumulation (e.g., FA in **Figure 9**). The overlain K-means analysis of resistivity/P-wave velocity allows us to discern,



within the upper areas of fluid accumulation, predominantly gas-saturated media (i.e., cluster 2) and predominantly water-saturated media (i.e., cluster 3). We note that in both profiles, gas-saturated (i.e., orange) areas are found below the water-saturated (i.e., blue) zones. Overall, three main lateral heterogeneities are visible at a distance of  $\sim 60$  m and  $\sim 170$  m on NNE trending profile 1 of **Figure 9** and at a distance of  $\sim 175$ – $200$  m on the WNW trending profile 2 of **Figure 10**. We argue that these lateral variations in the distribution of clusters 2–3, are controlled by fault/fractures affecting the very shallow hydrothermal circulation, Vice-versa cluster 1 (in green) which marks the unsaturated medium, does not show any significant lateral variation along the two cross-sections.

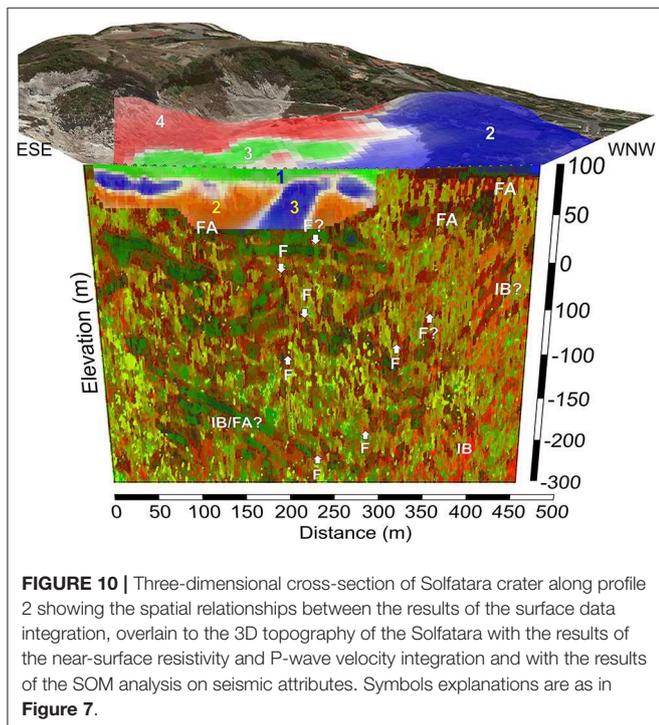
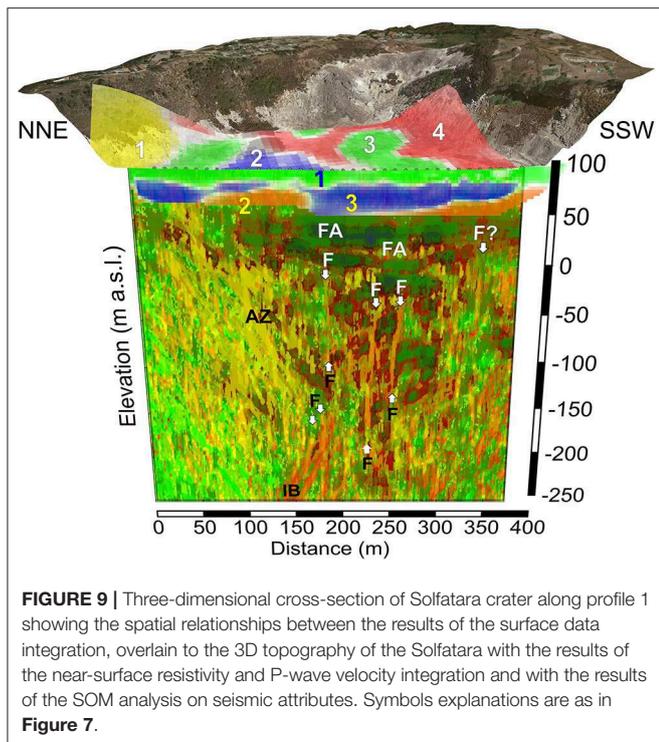
Next, we correlate the variations in the K-means clustering along the two profiles with the pattern of the surface K-means clusters within the crater (**Figure 2**), which we recall show an excellent agreement with the distribution of the intracratere tectonic structures mapped by Isaia et al. (2015). In doing this comparison we need to keep in mind that the two datasets were acquired in different periods and are made of entirely different geophysical parameters (i.e., Bouguer anomaly,  $\text{CO}_2$  flux and seismic noise amplitude vs. resistivity and P-wave velocity). Despite these facts, on profile 1 (**Figure 9**) we observe an overall good agreement between the patterns of subsurface and surface cluster boundaries. This agreement worsens on profile 2, probably because of its lower-S/N ratio. However, we can still notice on this latter profile that the boundary between

surface clusters 3–4 (which highlight the most unstable areas of the crater) and cluster 2 (i.e., the most stable) occurs at a distance of  $\sim 200$  m, above the aforementioned lateral transaction between cross-sectional cluster [2 (i.e., gas saturated) and 3 (i.e., water-saturated)].

It is worth noting that the surface datasets were acquired between in 2000 and 2004 and the subsurface geophysical (i.e., seismic and geoelectrical) data were acquired on in 2014–2015. A minor lack of correlation, noticeable among the different datasets, can be therefore associated to the high dynamism of the Solfatara volcano. Anyway, the overall good agreement between the three sets of geophysical data examined with the two machine learning techniques whitens the fact that the main shallow structural patterns, which influence the hydrothermal dynamics at Solfatara volcano, remained mainly unchanged in the last 15 years. This is not a minor outcome of this work.

### K-Means vs. Self-Organizing Maps

One of the purposes of this paper was compare the performance of two machine-learning algorithms in integrating the four seismic attributes computed from the orthogonal seismic reflection profiles of Bruno et al. (2017) which are macroscopically characterized by a different S/N ratio. The two datasets have a high number of nodes ( $372 \times 600$  for profile 1 and  $451 \times 600$  for profile 2). The process of reducing the dimensionality of the vectors, which is essentially a data compression technique, was done through SOM and K-means



clustering. The idea behind both algorithms is to map high-dimensional vectors onto a smaller dimensional (typically a 2D) space (Klose, 2006). In addition to data compression, a SOM, however, creates a network that stores information in such a way that any topological relationships within the training datasets is

maintained (Kalteh et al., 2008). At a more intuitive level, both K-means and SOM are moving nodes toward denser areas of the high-dimensional space. With K-means, the nodes move freely, with no direct relationship to each other. With SOM, when a node moves toward the data, it pulls neighboring nodes in the 2D lattice along with it. This naturally maintains a topology embedded in the data space. Therefore, unlike the K-means, with the SOM, vectors that are close in the high-dimensional space also end up being mapped to nodes that are close in 2D space. SOM, therefore, preserve the topology of the original data because the distances among neurons in 2D space reflect those in the high-dimensional space. However, in the case of a low number of nodes (such as our surface and near-surface data) K-means provides results very similar to SOM because it forces every vector to match an existing node, acting as a prototype/centroid, without any room for divergence. In the case of a high number of nodes, with SOM there is instead a margin for showing transitioning zones, which mimic the space among prototypes/centroids, thus modeling the transformed topological space among samples. In this way, the relative distances are “preserved” at expenses of a larger iteration time. This is a desirable feature, especially when dealing with data characterized by low S/N ratios.

Preservation of the original topological space translates into preservation of subtle features; this is evident when we compare the results of K-means and SOM with the seismic attribute datasets in Figures 7, 8 or Figures 9, 10. On profile 1, which holds a better S/N ratio, the comparison between SOM and K-means (Figure 7) shows an overall good agreement and conveys the same type of information. Therefore, even if the SOM image is more complex and detailed, when geological constraints are not available, such as in our case, the colormap simplification provided by K-means might actually help in synthesizing a simple geological interpretation. However, on profile 2 (Figure 8), which is characterized instead by a much lower S/N ratio, the preservation of the original topological space provided by the SOM results in an image that is far more informative than the K-means image. For instance, hints of possible sub-vertical faults (labeled as “F” in Figure 8), highlighted by subtle color changes on the SOM image, are totally missing on the K-means image (see the corresponding labels pointing at the exact locations in Figure 8). Therefore, in this latter profile the color map simplification provided by K-means is a flaw and SOM results hold more potential for geological interpretation.

As stated above, the coloring scheme of the SOM neural network must reflect the distribution of variables (i.e., seismic attributes) across the lattice. With our coloring scheme, based on the position of neurons and the neighborhood distances of neurons (i.e., U-matrix), nearby features have similar color gradations. One of the limits of this approach is the absence of a quantitative criteria to assess the validity of the results (de Matos et al., 2006; Roden et al., 2015); moreover, when dealing with a high number of neurons only subtle color differences characterize neighboring neurons (i.e., similar classified observation). The lack of a sharp color change in some cases makes it hard to reveal differences among observations (see again Figure 8).

## Soft K-Means Clustering by the Silhouette Index

Both the tri-variate geophysical dataset measured at the surface of the crater and the bivariate tomographic dataset estimated along the two profiles are characterized by a low number of nodes. As discussed above, SOM maps with a small number of nodes are expected to behave in a way that is similar to K-means, therefore we chose the latter and faster method to analyze the above datasets. For these analyses we employed a hard-clustering algorithm but we modified the colormap of each observation by linking the Silhouette Index to the pixel color saturation as shown in **Figures 2, 4A, 5A**. As discussed above, the Silhouette Index ranges from  $-1$  to  $1$ . In our color maps light color saturations reveal clusters with values of the Silhouette Index near to  $0$ , which highlights observations with unclear assignment. Values  $<0$ , which are typical of misclustered observations are colored in white. Our approach is therefore similar to a soft clustering (Paasche et al., 2010) and helps in reducing well-known limits of the K-means with clustering data at the edges of the clusters. Moreover, a Silhouette Index-based color map can be helpful in highlighting intra-cluster zones where geophysical parameters are changing. This improvement can provide more complete and realistic results with respect to the typical hard clustering algorithms since it allows to locate areas where observations are well/poorly clustered, spot local dissimilarities and transition zones.

## CONCLUSIONS

In this work we tested the K-means and the SOM methods, widely used in multivariate clustering analysis, for the integration of different geophysical surveys acquired within Solfatara, on different scales and with different sensitivities, with the aim of providing robust geophysical models of the shallow hydrothermal system and processes. In applying these techniques, we explored two innovative approaches: for the K-means we propose to modify the color intensity of the clusters based on the Silhouette Index in order to graphically locate the membership uncertainty of the individual observations. This is useful to account for differences between observations who might be proximal or distant from their relative centroids and consequently allows us to go beyond a rigid classification which is typical of K-means clustering. This approach, detailed in the previous section, allows to visualize graphically the uncertainty of cluster assignment, which is not possible for K-means, this being a hard-clustering method. It also allows to highlight intra-cluster dissimilarities, which might indicate variability of geophysical characteristics within the same cluster number.

For the Self-Organizing Maps, we propose an alternative method for choosing the number of nodes of the neural network which aims to avoid the need for downstream clustering of the results of the classification. Our approach is based on the dimensioning of a neural network with a limited number of neurons but much greater than the dimensions of the expected seismic facies. In this way a possible loss of information is minimized with respect to larger number of nodes who have

necessarily to be grouped after SOM analysis, through classic clustering techniques. Moreover, our approach makes it possible to represent all the neurons on the original space of the dataset (in our case CDPs vs. depth), by means of a color map based on: (1) the position of the neurons and (2) on the U-matrix values, maintaining all the details of the individual neurons.

Overall, our results show that both methods hold a great potential to aid an unbiased interpretation of geophysical data in complex geological settings. Similarly to other active volcanic areas, Solfatara hydrothermal dynamics leaves subtle but detectable footprints in each investigated geophysical dataset. While these hints are not easy to spot by classical interpretation schemes based on a single-method approach, their detection is highlighted by merging and compressing the vast amount of geophysical information with both the unsupervised learning methods analyzed in this paper. This is evident in **Supplementary Figures 1–3** where we show the hard-cluster classification from the K-means overlain on each of the original data used for the analysis. This type of plot is very useful for post-analysis qualitative assessment of results. For instance, **Supplementary Figure 1** allows us to understand that each of the three-surface datasets used provided an important contribution to clusters classification, even if the main contribution was delivered by the  $\text{CO}_2$  flux. Similarly, **Supplementary Figures 2, 3** show that the ERT data were dominant in the subsurface K-means clustering, while P-wave tomography provided a smaller contribution. Nevertheless, from a thorough analysis of those figures it is evident that unsupervised learning techniques were highly successful in synthesizing the complex geophysical information provided by each single dataset in a simpler, more meaningful model of the surface and of the shallow subsurface of the volcano. A proof of this can be found in **Figure 2** that show us how well the patterns of the surface clusters correlate with the strike of intracrateric tectonic structures mapped by Isaia et al. (2015).

Another important point worth highlighting is that generally color maps used to represent the individual datasets, such as the ERT or the P-wave tomography, are subjective; i.e., their choice does influence our interpretation. In other words, by changing the colormap the resulting image changes slightly and the interpretation is affected by this change. For instance, the colormap used to represent our P-wave velocities in **Supplementary Figures 2, 3** does not enhance subtle lateral variations while the colormap used to represent ERT data well-enhances them: however, by changing those color maps we could influence the interpretation of lateral and or vertical heterogeneities. This is clear by comparing the ERT and P-wave images with the overlain K-means clusters: it is impossible to determine the exact position of the cluster boundaries in an objective manner only from the ERT and P-wave images.

To further highlight the improvement in achieving objective characterization of the shallow hydrothermal system of the volcano, we use **Supplementary Figures 8, 9** to show a comparison performed, across the two profiles, between the results obtained from K-means and SOM algorithms and the original seismic reflection and ERT data: it is evident the great aid provided by the unsupervised learning methods for

detection of the subsurface areas of fluid accumulation and their structural pathways underneath Solfatara. For instance, while the classical interpretation schemes, based on the analysis of reflection offset and reflection strength can profitably be used on the seismic reflection data of **Supplementary Figure 8** to pinpoint the location of sub-vertical faults and bright spots associated to fluids, the details provided by SOM and K-means images allows to increase our confidence in the interpretation. This support is particularly valuable when the interpreter has to deal with data at low S/N ratio, such as those shown in **Supplementary Figure 9**, where the results provided by machine learning algorithms on multivariate dataset can provide important constraints in minimizing the interpretation risk.

Considering that Solfatara and the neighboring areas are experiencing an new period of volcanic unrest and considering the vast amount of diverse geophysical, geochemical and geological data acquired in these area in the recent years, we hope that our case study will further promote the use of unsupervised learning techniques with the purpose of minimizing the interpretation risk and of achieving an improved understanding of the complex dynamics occurring in volcanoes.

## DATA AVAILABILITY STATEMENT

The datasets analyzed during the current study are freely available from the MEDSUV-RICEN website ([www.med-suv.eu](http://www.med-suv.eu)) and from the corresponding author on reasonable request.

## REFERENCES

- Abdi, H., and Williams, L. J. (2010). Principal component analysis. *WIREs Comp. Stat.* 2, 433–459. doi: 10.1002/wics.101
- Amoroso, O., Festa, G., Bruno, P. P., D'Auria, L., De Landro, G., Di Fiore, V., et al. (2018). Integrated tomographic methods for seismic imaging and monitoring of volcanic caldera structures and geothermal areas. *J. Appl. Geophys.* 156, 16–30. doi: 10.1016/j.jappgeo.2017.11.012
- Bedrosian, P. A., Maercklin, N., Weckmann, U., Bartov, Y., Ryberg, T., and Ritter, O. (2007). Lithology-derived structure classification from the joint interpretation of magnetotelluric and seismic models. *Geophys. J. Int.* 170, 737–748. doi: 10.1111/j.1365-246X.2007.03440.x
- Bernardinetti, S., Maraio, S., Bruno, P. P. G., Cicala, V., Minucci, S., Giannuzzi, M., et al. (2017). Potential shallow aquifers characterization through an integrated geophysical method: multivariate approach by means of K-means algorithms. *Ital. J. Groundwater* 6, AS21–A278. doi: 10.7343/as-2017-278
- Berrino, G., Corrado, G., Luongo, G., and Toro, B. (1984). Ground deformation and gravity changes accompanying the 1982 Pozzuoli uplift. *Bull. Volcanol.* 47, 187–200. doi: 10.1007/BF01961548
- Bianco, F., Del Pezzo, E., Saccorotti, G., and Ventura, G. (2004). The role of hydrothermal fluids in triggering the July–August 2000 seismic swarm at Campi Flegrei, Italy: evidence from seismological and mesostructural data. *J. Volcanol. Geother. Res.* 133, 229–246. doi: 10.1016/S0377-0273(03)00400-1
- Bock, H. H. (2007). “Clustering methods: a history of k-means algorithms,” in *Selected Contributions in Data Analysis and Classification. Studies in Classification, Data Analysis, and Knowledge Organization*, eds P. Brito, G. Cucumel, P. Bertrand, and F. de Carvalho (Berlin; Heidelberg: Springer).
- Brock, G., Pihur, V., Datta, S., and Datta, S. (2008). clValid: an R package for cluster validation. *J. Stat. Softw.* 25, 1–32. doi: 10.18637/jss.v025.i04

## AUTHOR CONTRIBUTIONS

SB wrote MATLAB and R routines for SOM and K-means processing of the seismic attributes and of the P-wave velocity and resistivity data. PB interpreted the results, wrote the paper, and coded MATLAB software for the K-means clustering of the intra-crater spatial data and for data homogenization and interpolation. Both authors were equally involved in figure preparation and manuscript editing and reviewing.

## FUNDING

This work used part of the datasets acquired under the Med-Suv project. MED-SUV has received funding from the European Union's Seventh Program for research, technological development and demonstration under the grant agreement No. 308665.

## ACKNOWLEDGMENTS

The authors thank Vincenzo di Fiore, Marceau Gresse and Jean Vandemeulebrouck for making it available the electric measurements data along the two profiles.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/feart.2019.00286/full#supplementary-material>

- Bruno, P. P., Ricciardi, G. P., Petrillo, Z., Di Fiore, V., Troiano, A., and Chiodini, G. (2007). Geophysical and hydrogeological experiments from a shallow hydrothermal system at Solfatara Volcano, Campi Flegrei, Italy: response to caldera unrest. *J. Geophys. Res. Solid Earth* 112, 1–17. doi: 10.1029/2006JB004383
- Bruno, P. P. G., Maraio, S., and Festa, G. (2017). The shallow structure of Solfatara Volcano, Italy, revealed by dense, wide-aperture seismic profiling. *Sci. Rep.* 7:17386. doi: 10.1038/s41598-017-17589-3
- Bruno, P. P. G., Rapolla, A., and Di Fiore, V. (2003). Structural setting of the Bay of Naples (Italy) seismic reflection data: implications for Campanian volcanism. *Tectonophysics*, 372, 193–213. doi: 10.1016/j.tecto.2003.09.002
- Byrdina, S., Vandemeulebrouck, J., Cardellini, C., Legaz, A., Camerlynck, C., Chiodini, G., et al. (2014). Relations between electrical resistivity, carbon dioxide flux, and self-potential in the shallow hydrothermal system of Solfatara (Phlegrean Fields, Italy). *J. Volcanol. Geother. Res.* 283, 172–182. doi: 10.1016/j.jvolgeores.2014.07.010
- Caliro, S., Chiodini, G., Moretti, R., Avino, R., Granieri, D., Russo, M., et al. (2007). The origin of the fumaroles of La Solfatara (Campi Flegrei, South Italy). *Geochim. Cosmochim. Acta* 71, 3040–3055. doi: 10.1016/j.gca.2007.04.007
- Cardellini, C., Chio, G., Frondini, F., Avino, R., Bagnato, E., Caliro, S., et al. (2017). Monitoring diffuse volcanic degassing during volcanic unrests: the case of Campi Flegrei (Italy). *Sci. Rep.* 7, 1–15. doi: 10.1038/s41598-017-06941-2
- Céréghino, R., and Park, Y. S. (2009). Review of the self-organizing map (SOM) approach in water resources: commentary. *Environ. Modell. Softw.* 24, 945–947. doi: 10.1016/j.envsoft.2009.01.008
- Chiodini, G. (2009). CO<sub>2</sub>/CH<sub>4</sub> ratio in fumaroles a powerful tool to detect magma degassing episodes at quiescent volcanoes. *Geophys. Res. Lett.* 36:L02302. doi: 10.1029/2008GL036347

- Chiodini, G., Avino, R., Caliro, S., and Minopoli, C. (2011). Temperature and pressure gas geoindicators at the Solfatara fumaroles (Campi Flegrei). *Anna. Geophys.* 54, 151–160. doi: 10.4401/ag-5002
- Chiodini, G., Frondini, F., Cardellini, C., Granieri, D., Marini, L., and Ventura, G. (2001). CO<sub>2</sub> degassing and energy release at Solfatara volcano, Phlegraean Fields, Italy. *J. Geophys. Res. Solid Earth* 106, 16213–16221. doi: 10.1029/2001JB002046
- Chiodini, G., Granieri, D., Avino, R., Caliro, S., Costa, A., and Werner, C. (2005). Carbon dioxide diffuse degassing and estimation of heat release from volcanic and hydrothermal systems. *J. Geophys. Res. Solid Earth* 110:B08204. doi: 10.1029/2004JB003542
- Chiodini, G., Selva, J., Del Pezzo, E., Marsan, D., De Siena, L., D'Auria, L., et al. (2017). Clues on the origin of post-2000 earthquakes at Campi Flegrei Caldera (Italy). *Sci. Rep.* 7:4472. doi: 10.1038/s41598-017-04845-9
- Coléou, T., Poupon, M., and Azbel, K. (2003). Unsupervised seismic facies classification: a review and comparison of techniques and implementation. *Lead. Edge* 22, 942–953. doi: 10.1190/1.1623635
- Cusano, P., Petrosino, P., and Saccorotti, G. (2008). Hydrothermal origin for sustained Long-Period (LP) activity at Phlegraean Fields volcanic complex, Italy. *J. Volcanol. Geother. Res.* 177, 1035–1044. doi: 10.1016/j.jvolgeores.2008.07.019
- De Landro, G., Serlenga, V., Russo, G., Amoroso, O., Festa, G., Bruno, P. P., et al. (2017). 3D ultra-high resolution seismic imaging of shallow Solfatara crater in Campi Flegrei (Italy): new insights on deep hydrothermal fluid circulation processes. *Sci. Rep.* 7, 1–10. doi: 10.1038/s41598-017-03604-0
- de Matos, M. C., Osorio, P. L., and Johann, P. R. (2006). Unsupervised seismic facies analysis using wavelet transform and self-organizing maps. *Geophysics* 72, P9–P21. doi: 10.1190/1.2392789
- De Natale, G., Pingue, F., Allard, P., and Zollo, A. (1991). Geophysical and geochemical modelling of the 1982–1984 unrest phenomena at Phlegraean Fields caldera (southern Italy). *J. Volcanol. Geother. Res.* 48, 199–222. doi: 10.1016/0377-0273(91)90043-Y
- De Siena, L., Sammarco, C., Cornwell, D. G., La Rocca, M., Bianco, F., Zaccarelli, L., et al. (2018). Ambient seismic noise image of the structurally controlled heat and fluid feeder pathway at Campi Flegrei caldera. *Geophys. Res. Lett.* 45, 6428–6436. doi: 10.1029/2018GL078817
- De Vita, S., Orsi, G., Civetta, L., Carandente, A., D'Antonio, M., Deino, A., et al. (1999). The Agnano-Monte Spina eruption (4100 years BP) in the restless Campi Flegrei caldera (Italy). *J. Volcanol. Geother. Res.* 91, 269–301. doi: 10.1016/S0377-0273(99)00039-6
- De Vivo, B., Rolandi, G., Gans, P. B., Calvert, A., Bohrsen, W. A., Spera, F. J., et al. (2001). New constraints on the pyroclastic eruptive history of the Campanian volcanic Plain (Italy). *Mineral. Petrol.* 73, 47–65. doi: 10.1007/s007100170010
- Deino, A. L., Orsi, G., de Vita, S., and Piochi, M. (2004). The age of the Neapolitan Yellow Tuff caldera-forming eruption (Campi Flegrei caldera-Italy) assessed by <sup>40</sup>Ar/<sup>39</sup>Ar dating method. *J. Volcanol. Geother. Res.* 133, 157–170. doi: 10.1016/S0377-0273(03)00396-2
- Di Vito, M., Lirer, L., Mastrolorenzo, G., and Rolandi, V. (1987). The 1538 Monte Nuovo eruption (Campi Flegrei, Italy). *Bull. Volcanol.* 49, 608–615. doi: 10.1007/BF01079966
- Everitt, B. S., and Dunn, G. (2001). *Applied Multivariate Data Analysis*, Vol. 2. London: Arnold.
- Florio, G., Fedi, M., Cella, F., and Rapolla, A. (1999). The Campanian Plain and Phlegraean Fields: structural setting from potential field data. *J. Volcanol. Geother. Res.* 91, 361–379. doi: 10.1016/S0377-0273(99)00044-X
- Gaeta, F. S., De Natale, G., Peluso, F., Mastrolorenzo, G., Castagnolo, D., Troise, C., et al. (1998). Genesis and evolution of unrest episodes at Campi Flegrei caldera: the role of thermal fluid-dynamical processes in the geothermal system. *J. Geophys. Res.* 103, 20921–20933. doi: 10.1029/97JB03294
- Gersho, A. (1982). On the structure of vector quantizers. *IEEE Trans. Inform. Theory* 28, 157–166. doi: 10.1109/TIT.1982.1056457
- Glamoclija, M., Garrel, L., and Berthon, J. (2004). Biosignatures and bacterial diversity in hydrothermal deposits of Solfatara Crater, Italy. *Geomicrobiol. J.* 21, 529–541. doi: 10.1080/01490450490888235
- Gray, R. (1984). Vector quantization. *IEEE Assp Magazine* 1, 4–29. doi: 10.1109/MASSP.1984.1162229
- Gresse, M., Vandemeulebrouck, J., Byrdina, S., Chiodini, G., Revil, A., Johnson, T. C., et al. (2017). Three-dimensional electrical resistivity tomography of the Solfatara Crater (Italy): implication for the multiphase flow structure of the shallow hydrothermal system. *J. Geophys. Res.* 122, 8749–8768. doi: 10.1002/2017JB014389
- Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *J. Intell. Inform. Syst.* 17, 107–145. doi: 10.1023/A:1012801612483
- Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* 3, 610–621. doi: 10.1109/TSMC.1973.4309314
- Isaia, R., Marianelli, P., and Sbrana, A. (2009). Caldera unrest prior to intense volcanism in Campi Flegrei (Italy) at 4.0 ka B.P.: implications for caldera dynamics and future eruptive scenarios. *Geophys. Res. Lett.* 36, 1–6. doi: 10.1029/2009GL040513
- Isaia, R., Vitale, S., Di Giuseppe, M. G., Iannuzzi, E., D'Assisi Tramparulo, F., and Troiano, A. (2015). Stratigraphy, structure, and volcano-tectonic evolution of Solfatara maar-diatreme (Campi Flegrei, Italy). *Geol. Soc. Am. Bull.* 127, 1–20. doi: 10.1130/B31183.1
- Kalteh, A. M., Hjorth, P., and Berndtsson, R. (2008). Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application. *Environ. Modell. Softw.* 23, 835–845. doi: 10.1016/j.envsoft.2007.10.001
- Kassambara, A. (2017). *Practical Guide To Principal Component Methods in R*. Vol. 2. STHDA.
- Kaufman, L., and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: John Wiley & Sons, Inc.
- Klose, C. D. (2006). Self-organizing maps for geoscientific data analysis: geological interpretation of multidimensional geophysical data. *Comput. Geosci.* 10, 265–277. doi: 10.1007/s10596-006-9022-x
- Kohonen, T. (1997). “Exploration of very large databases by self-organizing maps,” in *Proceedings of International Conference on Neural Networks (ICNN'97)*, Vol. 1 (Houston, TX: IEEE), PL1–PL6.
- Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Netw.* 37, 52–65. doi: 10.1016/j.neunet.2012.09.018
- Kohonen, T. (2014). *MATLAB Implementations and Applications of the Self-Organizing Map*. Helsinki: Unigrafia Oy.
- Koivunen, A. C., and Kostinski, A. B. (1999). The feasibility of data whitening to improve performance of weather radar. *J. Appl. Meteorol.* 38, 741–749.
- Langer, H., Falsaperla, S., Masotti, M., Campanini, R., Spampinato, S., and Messina, A. (2009). Synopsis of supervised and unsupervised pattern classification techniques applied to volcanic tremor data at Mt Etna, Italy. *Geophys. J. Int.* 178, 1132–1144. doi: 10.1111/j.1365-246X.2009.04179.x
- Lary, D. J., Alavi, A. H., Gandomi, A. H., and Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geosci. Front.* 7, 3–10. doi: 10.1016/j.gsf.2015.07.003
- Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theory* 28, 129–137. doi: 10.1109/TIT.1982.1056489
- Malleswar, Y., Marfurt, K. J., and Matson, S. (2010). Seismic texture analysis for reservoir prediction and characterization. *Lead. Edge* 29, 1116–1121. doi: 10.1190/1.3485772
- Martinez, W. L., and Martinez, A. R. (2005). *Exploratory Data Analysis with MATLAB*. New York, NY: CHAPMAN & HALL/CRC. doi: 10.1201/9780203483374
- Milia, A., Torrente, M. M., and Giordano, F. (2000). Active deformation and volcanism offshore Campi Flegrei, Italy: new data from high-resolution seismic reflection profiles. *Marine Geol.* 171, 61–73. doi: 10.1016/S0025-3227(00)00111-0
- Orsi, G., De Vita, S., and Di Vito, M. (1996). The restless, resurgent Campi Flegrei nested caldera (Italy): constraints on its evolution and configuration. *J. Volcanol. Geother. Res.* 74, 179–214. doi: 10.1016/S0377-0273(96)00063-7
- Paasche, H., Tronicke, J., and Dietrich, P. (2010). Automated integration of partially collocated models: Subsurface zonation using a modified fuzzy c-means cluster analysis algorithm. *Geophysics* 75, P11–P22. doi: 10.1190/1.3374411
- Prasad, M., Fabricius, I. L., and Olsen, C. (2005). Rock physics and statistical well log analyses in marly chalk. *Lead. Edge* 24, 491–495. doi: 10.1190/1.1926806
- Pryke, A., Mostaghim, S., and Nazemi, A. (2007). “Heatmap visualization of population based multi objective algorithms,” in *Evolutionary Multi-Criterion Optimization. EMO 2007. Lecture Notes in Computer Science*, Vol 4403, eds S. Obayashi, K. Deb, C. Poloni, T. Hiroyasu, and T. Murata (Berlin; Heidelberg: Springer), 361–375.

- Rawashdeh, M., and Ralescu, A. (2012). “Crisp and fuzzy cluster validity: generalized intra-inter silhouette index,” in *Fuzzy Information Processing Society (NAFIPS), 2012 Annual Meeting of the North American* (Berkeley, CA: IEEE), 1–6. doi: 10.1109/NAFIPS.2012.6290969
- Roden, R., Smith, T., and Sacrey, D. (2015). Geologic pattern recognition from seismic attributes: principal component analysis and self-organizing maps. *Interpretation* 3, SAE59–SAE83. doi: 10.1190/INT-2015-0037.1
- Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., and Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol. Rev.* 71, 804–818. doi: 10.1016/j.oregeorev.2015.01.001
- Rye, R. O. (2005). A review of the stable-isotope geochemistry of sulfate minerals in selected igneous environments and related hydrothermal systems selected igneous environments and related hydrothermal systems. *Chem. Geol.* 215, 5–36. doi: 10.1016/j.chemgeo.2004.06.034
- Sacchi, M., Alessio, G., Aquino, I., Esposito, E., Molisso, F., Nappi, R., et al. (2009). Risultati preliminari della campagna oceanografica CAFE\_07-Leg 3 nei Golfi di Napoli e Pozzuoli, Mar Tirreno orientale. *Quaderni di geofisica.* 64:26.
- Saccorrotti, G., Petrosino, S., Bianco, F., Castellano, M., Galluzzo, D., La Rocca, M., et al. (2007). Seismicity associated with the 2004–2006 renewed ground uplift at Phlegraean Fields Caldera, Italy. *Phys. Earth Planet. Inter.* 165, 14–24. doi: 10.1016/j.pepi.2007.07.006
- Sharma, S. C. (1996). *Applied Multivariate Techniques*. New York, NY: John Wiley and Sons.
- Smith, V. C., Isaia, R., and Pearce, N. J. G. (2011). Tephrostratigraphy and glass compositions of post-15 kyr Campi Flegrei eruptions: implications for eruption history and chronostratigraphic markers. *Quat. Sci. Rev.* 30, 3638–3660. doi: 10.1016/j.quascirev.2011.07.012
- Späth, H. (1980). *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*. Chichester: Horwood.
- Taner, M. T., Walls, J. D., Smith, M., Taylor, G., Carr, M. B., and Dumas, D. (2001). “Reservoir characterization by calibration of selforganized map clusters,” in *2001 SEG Annual Meeting. Society of Exploration Geophysicists* (San Antonio, TX). doi: 10.1190/1.1816406
- Tronick, J., Holliger, K., Barrash, W., and Knoll, M. D. (2004). Multivariate analysis of cross-hole georadar velocity and attenuation tomograms for aquifer zonation. *Water Resour. Res.* 40:W01519. doi: 10.1029/2003WR002031
- Ultsch, A. (1993). “Self-organizing neural networks for visualisation and classification,” in *Information and Classification*, eds O. Opitz, B. Lausen, and R. Klar (Berlin; Heidelberg: Springer), 307–313. doi: 10.1007/978-3-642-50974-2\_31
- Unglert, K., Radić, V., and Jelinek, A. M. (2016). Principal component analysis vs. self-organizing maps combined with hierarchical clustering for pattern recognition in volcano seismic spectra. *J. Volcanol. Geother. Res.* 320, 58–74. doi: 10.1016/j.jvolgeores.2016.04.014
- Valentino, G. M., and Stanzione, D. (2004). Geochemical monitoring of the thermal waters of the Phlegraean Fields. *J. Volcanol. Geother. Res.* 133, 261–289. doi: 10.1016/S0377-0273(03)00402-5
- Vesanto, J., and Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Trans. Neural Netw.* 11, 586–600. doi: 10.1109/72.846731
- Vitale, S., and Isaia, R. (2014). Fractures and faults in volcanic rocks (Campi Flegrei, southern Italy): insight into volcano-tectonic processes. *Int. J. Earth Sci.* 103, 801–819. doi: 10.1007/s00531-013-0979-0
- Wehrens, R., and Buydens, L. M. (2007). Self-and super-organizing maps in R: the Kohonen package. *J. Stat. Softw.* 21, 1–19. doi: 10.18637/jss.v021.i05

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Bernardinetti and Bruno. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.