

Research Article

Search Space Reduction for Determination of Earthquake Source Parameters Using PCA and k -Means Clustering

Seongjae Lee and Taehyoun Kim 

Department of Mechanical and Information Engineering and Department of Smart Cities, University of Seoul, Seoul 02504, Republic of Korea

Correspondence should be addressed to Taehyoun Kim; thkim@uos.ac.kr

Received 21 July 2020; Revised 14 August 2020; Accepted 23 August 2020; Published 7 September 2020

Academic Editor: Sang-Hoon Hong

Copyright © 2020 Seongjae Lee and Taehyoun Kim. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The characteristics of an earthquake can be derived by estimating the source geometries of the earthquake using parameter inversion that minimizes the L2 norm of residuals between the measured and the synthetic displacement calculated from a dislocation model. Estimating source geometries in a dislocation model has been regarded as solving a nonlinear inverse problem. To avoid local minima and describe uncertainties, the Monte-Carlo restarts are often used to solve the problem, assuming the initial parameter search space provided by seismological studies. Since search space size significantly affects the accuracy and execution time of this procedure, faulty initial search space from seismological studies may adversely affect the accuracy of the results and the computation time. Besides, many source parameters describing physical faults lead to bad data visualization. In this paper, we propose a new machine learning-based search space reduction algorithm to overcome these challenges. This paper assumes a rectangular dislocation model, i.e., the Okada model, to calculate the surface deformation mathematically. As for the geodetic measurement of three-dimensional (3D) surface deformation, we used the stacking interferometric synthetic aperture radar (InSAR) and the multiple-aperture SAR interferometry (MAI). We define a wide initial search space and perform the Monte-Carlo restarts to collect the data points with root-mean-square error (RMSE) between measured and modeled displacement. Then, the principal component analysis (PCA) and the k -means clustering are used to project data points with low RMSE in the 2D latent space preserving the variance of original data as much as possible and extract k clusters of data with similar locations and RMSE to each other. Finally, we reduce the parameter search space using the cluster with the lowest mean RMSE. The evaluation results illustrate that our approach achieves 55.1~98.1% reductions in search space size and 60~80.5% reductions in 95% confidence interval size for all source parameters compared with the conventional method. It was also observed that the reduced search space significantly saves the computational burden of solving the nonlinear least square problem.

1. Introduction

In the past decades, interferometric synthetic aperture radar (InSAR) has been a powerful method to acquire geophysical features such as surface deformation or topography by comparing the phases of at least two complex-valued SAR images obtained from different location or time. Since SAR provides high-resolution images, InSAR can measure the surface deformation with centimetric or even millimetric accuracy. This accurate deformation map enables the observation of ocean and ground surface changes, the measurement of ice

drift and glacier elevations, and the analysis of seismic deformation or volcanic activities [1].

Surface deformation acquired from InSAR measurements provides essential information for studying earthquakes and volcanic activities. To derive the characteristics of an earthquake, we can estimate the source geometries of an earthquake by performing parameter inversion that minimizes the L2 norm of residuals between measured and modeled displacement calculated from a dislocation model [2]. In many cases, the root-mean-square error (RMSE) values of residuals are used for the evaluation of the inversion result.

A dislocation model provides the mathematically calculated coseismic displacement of the earth's surface. The most popular dislocation model is a rectangular dislocation model (Okada model) [3] that assumes isotropic homogeneous half-space and finite rectangular source. Other dislocation models that provide closed analytical expression include the prolate spheroid model (Yang model) [4] and the spherical source model (Mogi model) [5].

In general, since the dislocation model cannot express the surface displacement linearly, estimating source geometries is a nonlinear inverse problem that can be solved by a nonlinear least square method. The nonlinear least square method iteratively searches local minima based on the partial derivatives of the objective function that indicates the misfit between observed and synthetic data. Hence, the solution may vary according to the starting point of parameters, and thus, we cannot assure that the solution is a global minimum [6]. To avoid local minima and describe uncertainties, the Monte-Carlo restarts are often used to solve the problem [7].

In solving a nonlinear least square problem with the Monte-Carlo restarts, we should designate a search space by setting boundary conditions of each parameter to pick random starting points. Since a nonlinear least square method iteratively explores the parameter search space, the search space size significantly affects the accuracy and execution time of the algorithm. Previous studies [8, 9] used the initial range of search space relying on the focal mechanism from seismologists. However, the accuracy of seismological studies can influence the result of an inverse problem. Besides, a dislocation model with many source parameters describing fault leads to bad data visualization and thus analyzing the distribution of the results becomes quite tricky.

This paper presents a new search space reduction algorithm based on machine learning techniques to come up with the challenges mentioned above. Our proposed method proceeds as follows:

- (1) Initially, assume a wide range of search space and perform the Monte-Carlo restarts of five thousand iterations with randomly initialized starting points
- (2) Then, choose the samples that preserve maximum variance and have a low RMSE and project them to the two-dimensional (2D) space using PCA
- (3) Decide hyperparameter k based on the distribution of data on the 2D space and cluster samples using k -means clustering
- (4) Choose the cluster with the lowest mean RMSE and reduce the search space using the cluster
- (5) Perform additional Monte-Carlo restarts of five thousand iterations in the reduced search space and determine the final source parameters

Our search space reduction algorithm was compared with the conventional approach that uses geological studies to define the search space in the same study area, the 2017 Pohang earthquake. The evaluation was performed in terms of the size of search space, RMSE calculated from determined

source parameters, a 95% confidence interval of each source parameter, and the average iteration number of a nonlinear least square.

This paper is organized as follows. Section 2 introduces the related work and background of our approach. In Section 3, our machine learning-based search space reduction algorithm is described in detail. Section 4 presents the results and discussion of our approach in the study area, Pohang, Korea. Finally, Section 5 concludes the paper.

2. Related Work

2.1. Interferometric SAR and Earthquake Source Parameter Determination. InSAR is the method that extracts geophysical features such as surface topography and deformation by comparing the phase offsets of at least two complex-valued SAR images [1]. Comparing SAR images measured at different time instants, InSAR can retrieve the surface displacement at specific time instants, e.g., the surface deformation induced from earthquakes or volcanic activities, landslides [10], thaw-derived slope failure [11], and glacier elevations and changes [12]. However, conventional InSAR has a limitation that it can measure only the line-of-sight (LOS) component of displacement [13]. To overcome the limitation and measure 3D surface displacement accurately, the offset tracking method [14] and stacking InSAR and multiple aperture interferometry (MAI) [15] were proposed.

The source geometries of an earthquake can also be estimated by performing parameter inversion that finds the best approximate source parameters having the minimum misfit between the measured displacement and the synthetic 3D surface displacement of a theoretical dislocation model. Typical dislocation models are as follows. Mogi's model [5] creates a 2D displacement map in radial and tangential direction with four input parameters assuming ideal semi-infinite elastic crust and spherical source. The Yang model [4] generates a 3D displacement map in east, north, and vertical directions with eight source parameters assuming finite prolate spheroid source in an elastic half-space. Okada's model [3] builds a 3D displacement map similar to the Yang model, but it assumes a finite rectangular source and uses ten source parameters. All these dislocation models are nonlinear because the relation between the source parameters and surface displacement cannot be expressed as linear equations.

To estimate the source parameters using a nonlinear dislocation model, we can use the nonlinear least square method that minimizes the Euclidean norm of residuals between the measured and the modeled displacement. Let us denote the dislocation model as G , source parameters as \mathbf{m} , and observed displacement as \mathbf{d} , respectively. Then, the objective function of the nonlinear least square method is formulated as Equation (1) [6].

$$f(\mathbf{m}) = \|G(\mathbf{m}) - \mathbf{d}\|_2. \quad (1)$$

Source parameters \mathbf{m} can have boundary constraints. As for the parameter indicating an angle, the range of parameters is physically limited. For example, the parameter Strike that means the horizontal angle of the fault should be on

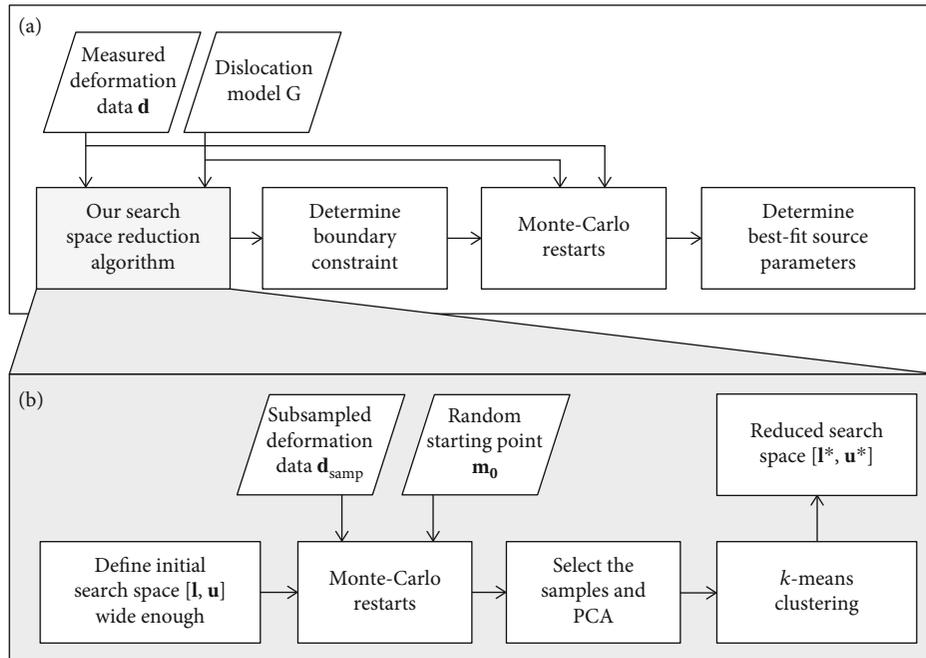


FIGURE 1: Schematic workflow of our approach. (a) is a big workflow that estimates earthquake source parameters. (b) shows our machine learning approach, included in (a), to reduce search space.

TABLE 1: Source parameters of the Okada model.

Parameter	Unit	Description
E	km	Distance from the reference point to the east
N	km	Distance from the reference point to the north
Depth	km	Depth of source
Strike	Degree	Angle of fault relative north
Dip	Degree	Angle between the fault and a horizontal plane
Length	km	Length of fault
Width	km	Width of fault
Rake	Degree	Angle of slip relative to the Width direction
Slip	mm	Dislocation in rake direction
Open	km	Dislocation in tensile component

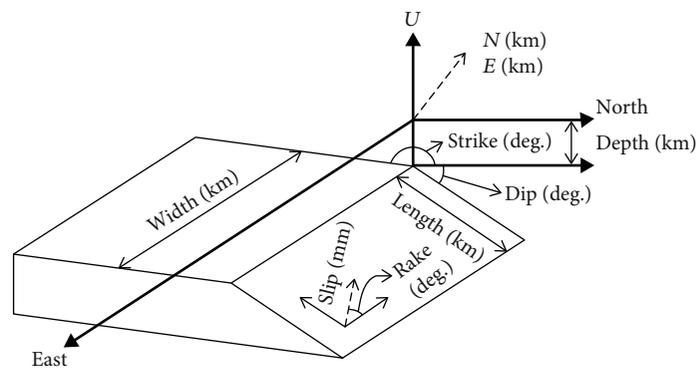


FIGURE 2: Fault geometry of the Okada model.

TABLE 2: Initial search space of source parameters.

	E (km)	N (km)	Depth (km)	Strike (deg.)	Dip (deg.)	Length (km)	Width (km)	Rake (deg.)	Slip (mm)
LB	-40	-40	1	0	0	0.01	0.01	0	-10000
UB	40	40	60	360	90	120	120	180	10000

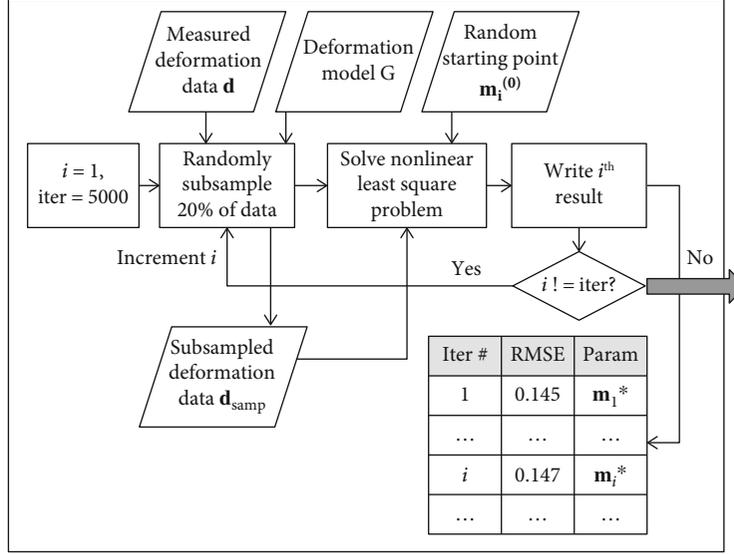


FIGURE 3: Data acquisition in the initial search space using the Monte-Carlo restarts.

TABLE 3: Options of MATLAB lsqnonlin() function.

Option	Value
Algorithm	Trust-region-reflective
MaxIter	500
TolX	0.01
TolFun	0.01

the range $[0^\circ, +360^\circ]$. Also, seismological studies such as the focal mechanism can be used to establish additional boundary conditions. Equation (2) shows the objective function of a nonlinear least square with lower bound \mathbf{l} and upper bound \mathbf{u} .

$$f(\mathbf{m}) = \|\mathbf{G}(\mathbf{m}) - \mathbf{d}\|_2, \quad \mathbf{l} \leq \mathbf{m} \leq \mathbf{u}. \quad (2)$$

The nonlinear least square method iteratively explores the search space from a specific starting point to local minima, so the method cannot derive global minima. To avoid local minima and describe uncertainties of parameters, we can use the Monte-Carlo restarts that generate many initial starting points and then solve the nonlinear least square problem for each starting point. Obtaining results from multiple starting points helps to find the global minimum and show the probabilistic descriptions such as standard deviation and histogram. Many studies employed the Monte-Carlo restarts to obtain the best-fit source parameters of earthquakes or volcanic activities [8, 9, 16–18]. Besides,

Bayesian estimation of source parameters can be used to determine the source parameters [19, 20].

Our approach used stacking InSAR and multiaperture InSAR (MAI) [15] and assumed a wide initial search space. Then, we performed Monte-Carlo restarts for randomly initialized starting points to get data that includes the evidence of the initial search space.

2.2. Machine Learning Applications for Remote Sensing.

Machine learning has been used widely in remote sensing [21]. It can be classified into supervised learning and unsupervised learning by whether a given set of training data includes the desired solution called *labels*, or not. In supervised learning, the label y corresponding to the data x exists, and the problem that the data x is provided without the label is called unsupervised learning. Typical supervised learning algorithms include the support vector machine (SVM), random forest (RF), artificial neural network (ANN), and k -nearest neighbor (KNN). In remote sensing, these algorithms have been used for estimating PM2.5 concentrations [22], modeling aboveground biomass of maize [23], and land cover classification [24].

Unsupervised learning techniques have also been used broadly in remote sensing. Principal component analysis (PCA), one of the representatives among them, reduces the dimension of the given dataset. PCA finds a lower dimensional hyperplane that preserves the maximum variance to maintain the maximum amount of original information [25]. In remote sensing, PCA has been used to process hyperspectral images acquired from airborne or spaceborne

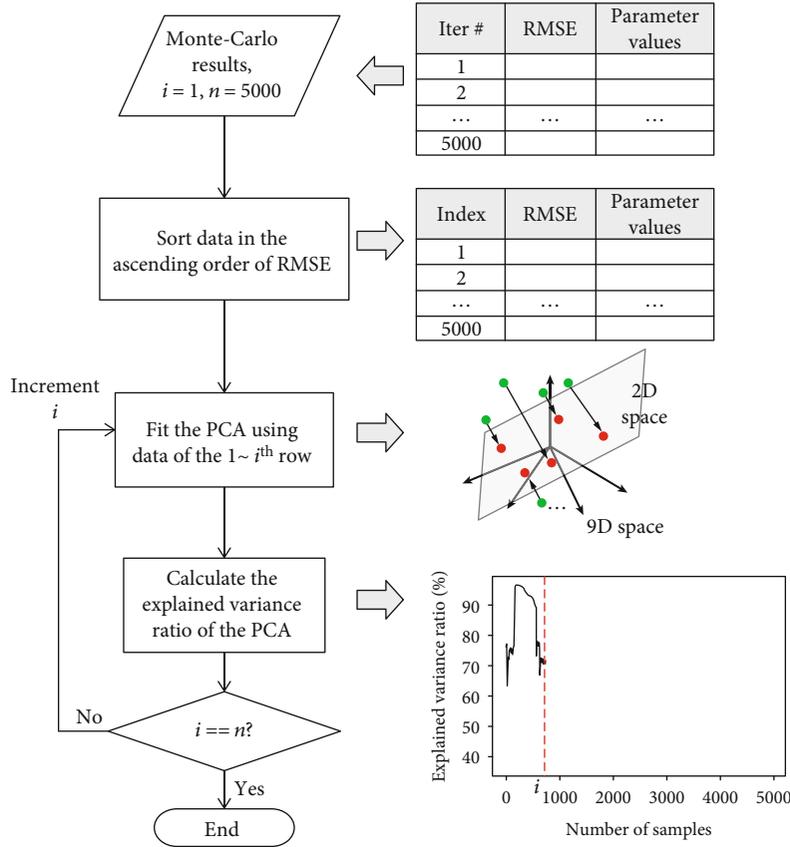


FIGURE 4: PCA procedure for dimensionality reduction of Monte-Carlo results.

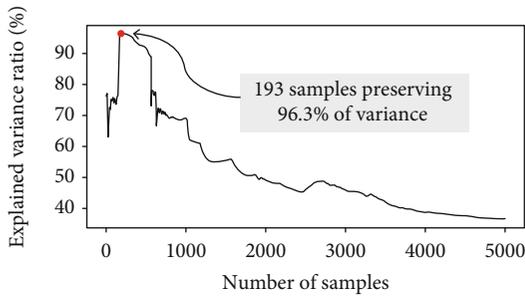


FIGURE 5: Explained variance ratio according to the number of samples.

TABLE 4: Principal components of each parameter.

Parameters	PC 1	PC 2
E	-0.16955	-0.94018
N	-0.34902	0.14051
Depth	0.35540	0.01463
Strike	0.35263	0.05568
Dip	0.34854	-0.01957
Length	-0.35430	0.03954
Width	-0.33501	0.27958
Rake	0.35549	0.04171
Slip	0.33640	-0.10564

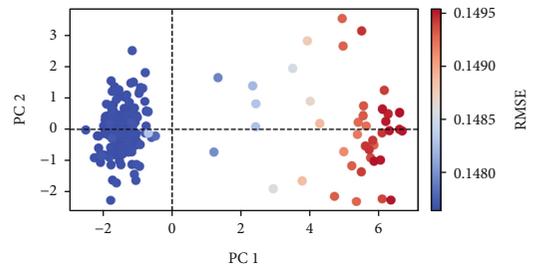


FIGURE 6: Samples projected into 2D space.

vehicles [26–28] and applied in various fields dealing with parameter inversion [29, 30].

k -means clustering is another popular unsupervised learning algorithm that groups data into the k clusters. The parameter k is a hyperparameter that should be defined manually, and data are grouped by minimizing the sum of the squared error of distances between the centroid of the cluster and the data points in the cluster [31]. In this paper, we used PCA to project the data in the initial search space to 2D latent space and performed k -means clustering to cluster the data.

3. Proposed Approach

This paper is aimed at reducing the parameter search space for the determination of earthquake source parameters with

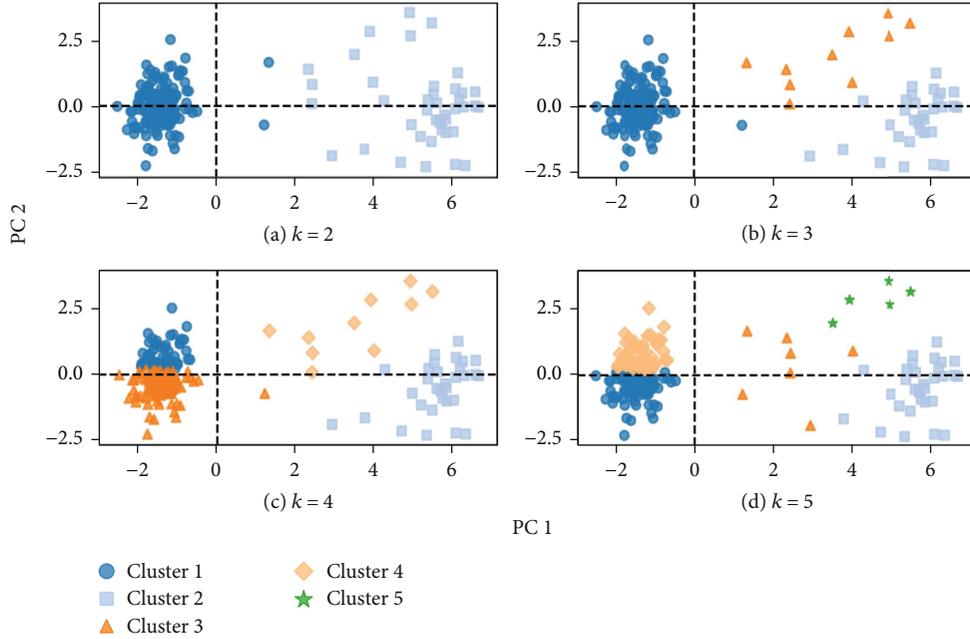


FIGURE 7: Results of k -means clustering for varying k .

TABLE 5: Statistics of each cluster.

Cluster no.	Mean of RMSE (cm)	Standard deviation of RMSE (cm)	Number of samples
1	0.14767	6.08458×10^{-5}	151
2	0.14934	2.13527×10^{-4}	32
3	0.14867	4.96763×10^{-4}	10

better computation time efficiency and data visualization. Figure 1 outlines our proposed search space reduction algorithm. Figure 1(a) illustrates the workflow that determines the source parameters using the Monte-Carlo restarts, and Figure 1(b) presents the proposed method that reduces the parameter search space included in Figure 1(a).

As the first step of our approach, we set a wide initial search space. Second, we perform the Monte-Carlo restarts of five thousand iterations to associate the RMSE and source parameters in the initial search space. In performing Monte-Carlo restarts, we randomly subsample measured deformation to reduce computation time. Third, we reduce the parameter search space using the PCA and the k -means clustering. In doing so, we first apply PCA to choose the samples that preserve the maximum variance of the original information with a low RMSE and perform the k -means clustering with varying k ranging from 2 to 5. Finally, we reduce the search space based on the data in the cluster having the lowest mean RMSE.

3.1. Datasets and Dislocation Model. Our approach is applied to the 2017 Pohang earthquake [32]. To acquire an accurate 3D surface deformation, we applied the stacking InSAR and

the MAI methods [15] to four SAR images obtained from the CSK and ALOS2. The size of the measured displacement is 18.84×15.54 km, and the pixel size is 30 m^2 .

Our study used the Okada model [3] as the dislocation model G. The Okada model assumes a finite rectangular source and isotropic homogeneous half-space. Table 1 describes the physical source parameters defining a rectangular source, and Figure 2 shows the fault geometry of the Okada model, respectively. The parameter E indicates the horizontal, i.e., E-W direction, distance of the upper-left point of the satellite image and fault, and N indicates the vertical one. The upper-left point of the measured deformation map is located at $36^{\circ}10'57''\text{N}$, $129^{\circ}17'03''\text{E}$. We fix the value of parameter Open, i.e., tensile component, to 0 and optimize the other nine parameters to determine the source parameter values.

3.2. Monte-Carlo Restarts in the Initial Search Space. As the first step, our algorithm conducts the Monte-Carlo restarts for a wide initial search space described in Table 2. Since the parameters Strike, Dip, and Rake represent angles, the lower and upper bounds for them are defined as the whole range physically available. Other parameters illustrating the length unit of the fault geometry are defined as large enough to cover the maximum movement of the fault possible in the study area.

Then, the Monte-Carlo restarts of five thousand iterations are performed to extract the [data point, RMSE] pairs that will be used in the next machine learning step. Figure 3 shows the procedure of the Monte-Carlo step. We randomly subsample 20% of measured deformation data \mathbf{d}_{samp} and define loose termination tolerance of the nonlinear least method to save the computational burden. In solving the nonlinear least square problem, we aim to minimize the objective function $\|G(\mathbf{m})_{\text{samp}} - \mathbf{d}_{\text{samp}}\|_2$ to retrieve the best-

TABLE 6: Search space determined from seismological approaches.

Parameters	E (km)	N (km)	Depth (km)	Strike (deg.)	Dip (deg.)	Length (km)	Width (km)	Rake (deg.)	Slip (mm)
LB	4.8	-12.0	4.1	220.0	33.0	4.0	4.0	120.0	10
UB	10.0	-6.0	5.7	235.0	55.0	6.5	6.7	150.0	300

TABLE 7: Reduced search space of three clusters.

Parameters	Cluster 1		Cluster 2		Cluster 3	
	LB	UB	LB	UB	LB	UB
E (km)	6.26911	6.41034	6.27024	6.37718	6.22817	6.3297
N (km)	-7.9826	-7.87204	-8.2007	-8.06074	-8.10795	-7.98743
Depth (km)	4.83734	5.19459	5.28377	5.54763	5.18088	5.49678
Strike (deg.)	224.9835	231.7157	232.4768	236.8167	231.5858	236.5544
Dip (deg.)	42.61413	44.68897	44.99918	46.47305	44.35434	46.34576
Length (km)	4.19294	4.92414	1.13009	3.04056	1.7581	4.19284
Width (km)	5.7689	6.45088	4.67679	5.51209	5.34835	6.08954
Rake (deg.)	127.4311	131.1753	132.5249	135.0445	131.1863	135.0024
Slip (mm)	170	230	330	950	220	520

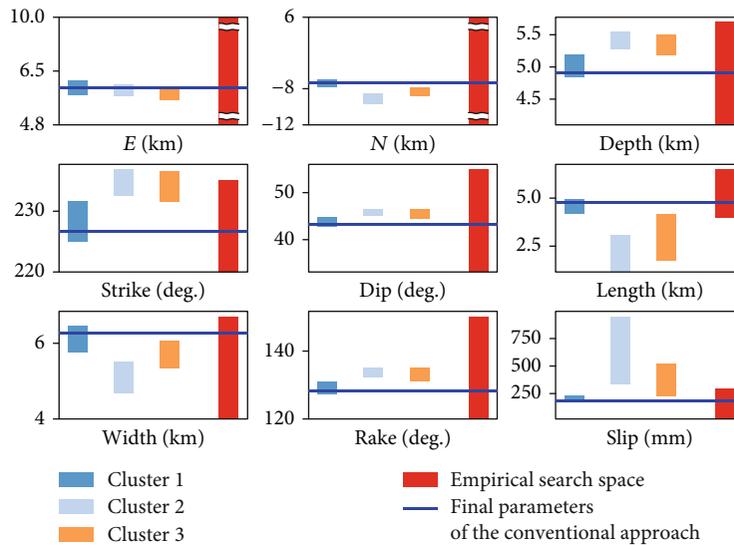


FIGURE 8: Comparison of search spaces between conventional approach and our approach.

TABLE 8: Search space size derived from conventional seismological studies and our proposed algorithm.

Parameters	Conventional approach	Our approach	Reduction ratio (%)
E (km)	5.2	0.1412	97.3
N (km)	6.0	0.1106	98.2
Depth (km)	1.6	0.3573	77.7
Strike (deg.)	15.0	6.7322	55.1
Dip (deg.)	22.0	2.0748	90.6
Length (km)	2.5	0.7312	70.8
Width (km)	2.7	0.6820	74.7
Rake (deg.)	30.0	3.7442	87.5
Slip (mm)	290	60	79.3

fit parameters with minimum residuals. As a result, optimization result \mathbf{m}_i^* and corresponding RMSE are calculated in each iteration. The nonlinear least square problem is solved by using the `lsqnonlin()` function of the MATLAB R2019a with the trust-region-reflective algorithm [33]. Table 3 specifies the option parameters used in the function. Option parameters not specified are set to the default values.

3.3. Search Space Reduction Using Machine Learning Approach. This study uses the PCA and the k -means clustering to reduce the search space. At first, we perform the PCA to reduce the dimensionality of the Monte-Carlo results, as described in Figure 4. Before the PCA fitting, the results obtained from the Monte-Carlo step are sorted in the ascending order of the RMSE value to select well-optimized

TABLE 9: Determined parameters and RMSE of conventional approach and our approach with a 95% confidence interval.

Parameters	Conventional approach	Our ML approach
E (km)	6.346 ± 0.098	6.341 ± 0.034
N (km)	-7.904 ± 0.107	-7.922 ± 0.029
Depth (km)	4.921 ± 0.172	4.901 ± 0.052
Strike (deg.)	226.563 ± 3.276	226.364 ± 1.059
Dip (deg.)	43.162 ± 0.985	43.04 ± 0.394
Length (km)	4.825 ± 0.465	4.786 ± 0.123
Width (km)	6.288 ± 0.724	6.21 ± 0.141
Rake (deg.)	128.424 ± 2.35	128.373 ± 0.65
Slip (mm)	175.243 ± 36.221	177.618 ± 7.417
RMSE (cm)	0.147637	0.147634

samples. Then, we perform the PCA fitting iteration for the sorted data in the index range of $[1, i]$ where $1 \leq i \leq 5000$ and calculate the explained variance ratio for the PCA results. The explained variance ratio is the ratio of the variance of projected data on the reduced dimensions to the total variance of the original information. Also, the explained variance ratio is regarded as a metric to evaluate the usefulness of the principal components extracted by the PCA, and thus a higher explained variance ratio means the information loss is kept smaller. For this reason, our algorithm chooses the samples showing the highest explained variance ratio for further processing.

PCA enables visualization of high-dimensional data in the low-dimensional space while preserving information of the original data as much as possible. In general, a 2D scatter matrix is used to visualize high-dimensional data, but each element of the scatter matrix represents only partial information. Furthermore, since the number of two-dimensional

scatter plots of p -dimensional parameters is $\binom{p}{2} = p(p-1)$,

visualizing high-dimensional data using a scatter matrix is cumbersome [34]. For example, visualizing the results of Monte-Carlo restarts by a scatter matrix requires 36 plots for the Okada model with nine parameters. To overcome this problem, this paper uses the PCA visualizing the nine-dimensional data on the 2D hyperplane preserving the original information so that we can analyze the [data, RMSE] pairs with just one 2D scatter plot.

For the data projected to the 2D space, we cluster the data using the k -means clustering. k -means clustering identifies k number of centroids and allocates data points to the nearest cluster, where k is a hyperparameter that a user should define manually. Our algorithm applies k -means clustering to the data on the 2D plane for varying k to find the most appropriate k that well divides projected samples with similar RMSE and position to each other. Once deciding the parameter k , we reduce the search space for each source parameter using the data cluster showing the lowest mean RMSE for the given k .

If the PCA preserves most of the variance of the Monte-Carlo results and all source parameters are on the same scale,

the distances between original data points in the latent 2D space can be directly assessed [35]. It means that the k -means clustering using the projected data can group the data in the adjacent location of the original nine-dimensional space because this algorithm minimizes the squared error between the centroid of the cluster and data points in the cluster. In summary, by using the PCA and the k -means clustering, we can extract the similarity of the nine-dimensional data on the 2D space with better data visualization.

Before the PCA fitting, each parameter is scaled to have a standard normal distribution $N(0, 1)$. PCA finds the principal components maximizing the variance. If parameters are unscaled, the principal component loading vector will be biased to a parameter showing a large variance. For example, the PCA fitting for the unscaled Monte-Carlo results may lead the principal component to have a significant loading for the Strike parameter that has the largest parameter search space among the source parameters of the Okada model. Furthermore, the distances between data points in the 2D space cannot be assessed. In this paper, we implemented our machine learning-based algorithm using the Scikit-learn package [36] in Python.

4. Experimental Results and Discussion

To evaluate the effectiveness of our machine learning-based search space reduction algorithm, this paper applied the proposed machine learning-based algorithm to the 2017 Pohang earthquake case [32]. The evaluation was performed in terms of the size of search space, RMSE values derived from determined source parameters, a 95% confidence interval of each source parameter, and the number of iterations for each non-linear least square of the Monte-Carlo restarts. The results were compared with those of the conventional approach, which refers to the seismological studies.

4.1. Principal Component Analysis and k -Means Clustering.

As mentioned in Section 3, we first conducted dimensionality reduction using PCA from nine-dimensional source parameters of the Okada model to the 2D space. Figure 5 illustrates the explained variance ratio for the number of samples used for PCA fitting. As described in Section 3, the data samples were presorted in the ascending order of the RMSE values. The result presents that the samples with high RMSE values interfere with the dimensionality reduction. The result also shows that the number of samples equals 193 produces the highest explained variance ratio, i.e., 96.3%. It means that the projection results with only 193 samples maintain most of the variance of the original data while keeping a low RMSE value, so the relation between those samples in the 2D space can be directly evaluated.

Table 4 shows the principal components (PCs) of each parameter, and Figure 6 presents the data projected into 2D space, respectively. With PCA, we can easily visualize the distribution of samples with only one 2D scatter plot. As shown in Figure 6, the samples with low RMSE values are clustered to the left side, while samples with high RMSE values are distributed from the center of the plot to the right side.

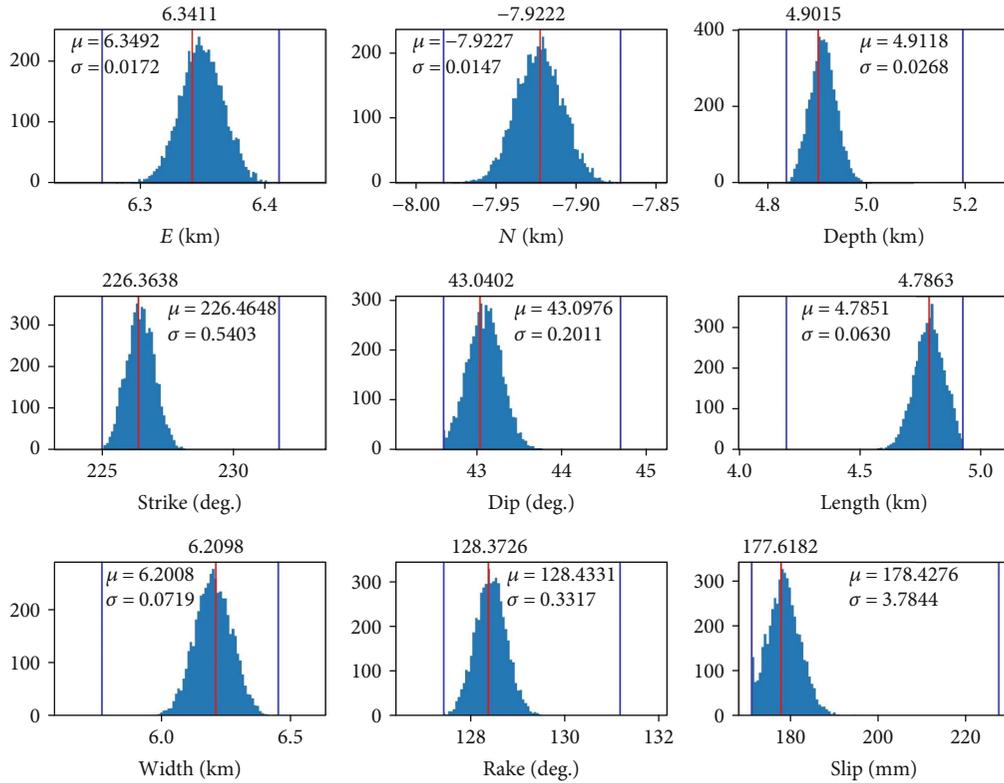


FIGURE 9: Histogram of determined parameters.

Then, we performed k -means clustering for $k = 2 \sim 5$ on the 2D hyperplane. Figures 7(a)–7(d) illustrate the k -means clustering results for each k , respectively. It was observed that the result for $k = 2$ could not distinguish the samples in the upper center with medium RMSE from those in the bottom right side with high RMSE. The results for $k = 4, 5$ also divide low-RMSE samples in the left corner. As a result, in our test case, data samples are best partitioned when $k = 3$. Table 5 also shows the mean RMSE and standard deviation of each cluster obtained for $k = 3$.

4.2. Search Space Reduction. We then reduced the search space using the minimum and maximum values of data in each cluster. Tables 6 and 7 show the search space derived from the seismological studies [37] and three reduced search spaces determined from our proposed algorithm for each source parameter. Figure 8 illustrates the results that compare the reduced search spaces derived from our approach with the empirically defined search space referring to seismological approaches. In Figure 8, the red bar shows empirically defined search space, and the others illustrate the search spaces determined by the three clusters. The blue horizontal line denotes the determined best-fit source parameters acquired by Monte-Carlo restarts of ten thousand iterations in the empirically defined search space. Figure 8 shows that, for all the parameters, the reduced search spaces for Cluster 1 include the source parameters determined from the Monte-Carlo restarts on the empirical parameter search. In other words, the reduced search space having the lowest mean RMSE is in good agreement with the conventional

research results based on seismological studies while effectively reducing the search space. For this reason, the rest of the experiments were conducted using the search space from Cluster 1.

Table 8 summarizes the reduction of search space for each source parameter after applying the proposed approach. From the results, the search space size for each parameter decreased by about 55.1~98.1%, compared with the search space from the conventional approach.

4.3. Determination of Source Parameters. As the final step, we conducted additional five thousand iterations of Monte-Carlo restarts in the reduced search space based on Cluster 1 and determined the source parameters as the best-fit parameter of the results of the Monte-Carlo restarts with a 95% confidence interval. Table 9 summarizes the source parameters determined and RMSE of our approach and the conventional approach. The results from two approaches seem similar to each other. However, it can be remarked that our machine learning approach produces more reliable results than the conventional one because a 95% confidence interval of our approach is much smaller than the conventional approach. The interval size reduction was about 60~80.5% for each parameter.

Figure 9 illustrates the distribution of the Monte-Carlo results derived from reduced search space using our machine learning-based approach. Two blue vertical lines of each histogram represent the lower and upper bounds of each parameter, and the red vertical line indicates the best-fit source parameter. In the reduced search space, all parameters

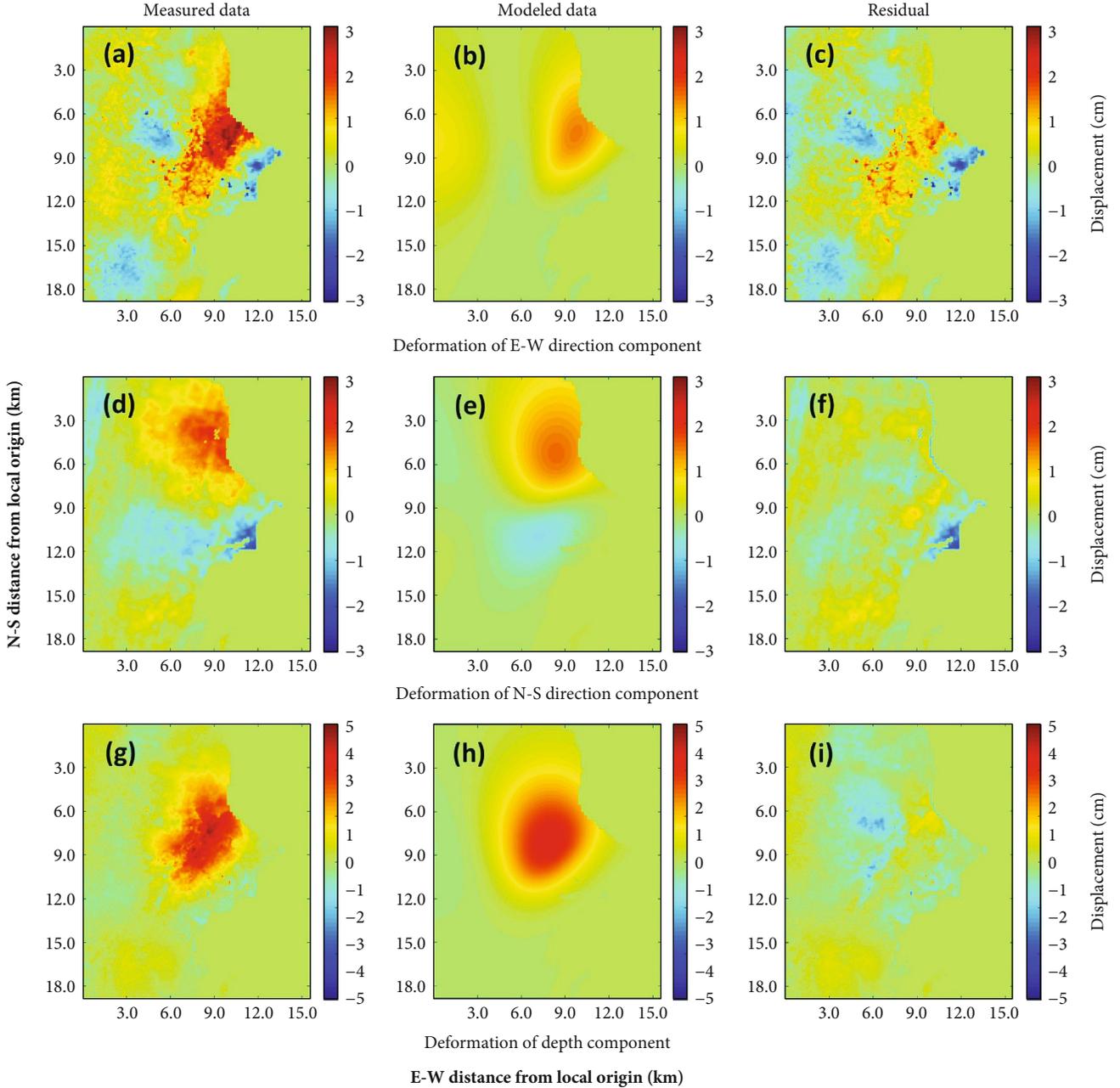


FIGURE 10: Final results of source modeling.

TABLE 10: Average number of iterations for nonlinear least square from the Monte-Carlo restarts.

Conventional approach	Our approach
4.93	2.25

feature a unimodal histogram fit in the search space appropriately with small standard deviations.

Figure 10 compares the synthetic model's deformation maps using the parameters acquired from our approach with those obtained from InSAR measurements. Each row of Figure 10 describes the deformation of E-W, N-S, and depth

direction component. The first column and the second column in Figure 10 show the measured surface deformation map for each direction component using InSAR and modeled deformation map from the Okada model. The third column shows a residual between the measured displacements and the modeled displacements. It can be concluded that the determination of source parameters using the Okada model and our reduced search space fit the measured displacement correctly because RMSE values for E-W, N-S, and depth components were 0.1863, 0.1544, and 0.0829 cm.

4.4. Computational Efficiency. For the computational efficiency validation, the average iterations of a nonlinear least

square from the Monte-Carlo restarts were also compared. As shown in Table 10, the average iterations of a nonlinear square for the reduced search space produced by our approach are much smaller than that of the conventional approach.

5. Conclusions and Future Works

This paper presented a new search space reduction algorithm using machine learning techniques for the earthquake source parameter determination. Our algorithm proceeded in the order of the Monte-Carlo restarts, PCA, k -means clustering, and search space reduction. PCA leads to better visualization of the Monte-Carlo results using only one 2D scatter plot, so we can find the hyperparameter k of the k -means clustering that groups data for them to have similar RMSE and close location to each other.

We compared the proposed approach with the conventional approach that defines the parameter search space by referring to seismological studies in terms of size of the search space, RMSE, a 95% confidence interval of Monte-Carlo results, and the average iteration number of nonlinear least square from the Monte-Carlo restarts. The experimental results for the 2017 Pohang earthquake test case showed that our approach achieves significant reductions in search space size and a 95% confidence interval size for all source parameters, i.e., about 55.1~98.1% and 60~80.5%, respectively. Also, the average iteration number of a nonlinear least square from our reduced search space was much smaller than that of a conventionally defined search space. Consequently, the results demonstrated that the proposed approach effectively reduces the parameter search space size and the computational burden.

Our future challenge is to develop the automatic search space reduction algorithm that uses the hierarchical clustering algorithm for determining the number of clusters.

Data Availability

The surface deformation maps were provided by Prof. Jung and Mr. Baek. They have generated the maps from the ALOS PALSAR-2 interferometric data that were provided through the JAXA's ALOS-2 research program (RA4, PI No. 1412).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the 2019 sabbatical year research grant of the University of Seoul.

References

- [1] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papathan, "A tutorial on synthetic aperture radar," *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 1, pp. 6–43, 2013.
- [2] K. L. Feigl, "Estimating earthquake source parameters from geodetic measurements," in *International Handbook of Earthquake and Engineering Seismology*, vol. 81A, pp. 1–14, Elsevier, 2002.
- [3] Y. Okada, "Surface deformation due to shear and tensile faults in a half-space," *Bulletin of the Seismological Society of America*, vol. 75, no. 4, pp. 1135–1154, 1985.
- [4] X.-M. Yang, P. M. Davis, and J. H. Dieterich, "Deformation from inflation of a dipping finite prolate spheroid in an elastic half-space as a model for volcanic stressing," *Journal of Geophysical Research: Solid Earth*, vol. 93, no. B5, pp. 4249–4257, 1988.
- [5] K. Mogi, *Relations between the eruptions of various volcanoes and the deformations of the ground surfaces around them*, Earthquake Research Institute, 1958.
- [6] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems*, Elsevier, 2nd edition, 2018.
- [7] M. Sambridge and K. Mosegaard, "Monte Carlo methods in geophysical inverse problems," *Reviews of Geophysics*, vol. 40, no. 3, 2002.
- [8] W. Qu, B. Zhang, Z. Lu et al., "Source parameter estimation of the 2009 Ms6.0 Yao'an Earthquake, Southern China, using InSAR observations," *Remote Sensing*, vol. 11, no. 4, p. 462, 2019.
- [9] M.-J. Jo, H.-S. Jung, and J.-S. Won, "Detecting the source location of recent summit inflation via three-dimensional InSAR observation of Kilauea volcano," *Remote Sensing*, vol. 7, no. 11, pp. 14386–14402, 2015.
- [10] Y. Kang, C. Zhao, Q. Zhang, Z. Lu, and B. Li, "Application of InSAR techniques to an analysis of the Guanling landslide," *Remote Sensing*, vol. 9, no. 10, p. 1046, 2017.
- [11] B. Hu, Y. Wu, X. Zhang et al., "Monitoring the thaw slump-derived thermokarst in the Qinghai-Tibet Plateau using satellite SAR interferometry," *Journal of Sensors*, vol. 2019, Article ID 1698432, 8 pages, 2019.
- [12] Y. Sun, L. Jiang, L. Liu, Q. Sun, H. Wang, and H. Hsu, "Mapping glacier elevations and their changes in the western Qilian Mountains, northern Tibetan Plateau, by bistatic InSAR," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 1, pp. 68–78, 2016.
- [13] J. Hu, Z. W. Li, X. L. Ding, J. J. Zhu, L. Zhang, and Q. Sun, "Resolving three-dimensional surface displacements from InSAR measurements: a review," *Earth-Science Reviews*, vol. 133, pp. 1–17, 2014.
- [14] M. Furuya and T. Yasuda, "The 2008 Yutian normal faulting earthquake (Mw 7.1), NW Tibet: non-planar fault modeling and implications for the Karakax fault," *Tectonophysics*, vol. 511, no. 3–4, pp. 125–133, 2011.
- [15] H. S. Jung, Z. Lu, J. S. Won, M. P. Poland, and A. Miklius, "Mapping three-dimensional surface deformation by combining multiple-aperture interferometry and conventional interferometry: application to the June 2007 eruption of Kilauea volcano, Hawaii," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 1, pp. 34–38, 2011.
- [16] M.-J. Jo, H.-S. Jung, and S.-H. Yun, "Retrieving precise three-dimensional deformation on the 2014 M6.0 South Napa earthquake by joint inversion of multi-sensor SAR," *Scientific Reports*, vol. 7, no. 1, p. 5485, 2017.
- [17] G. Diczelis, M. Assumpção, J. Kellogg, P. Pedraza, and F. Dias, "Estimating the 2008 Quetame (Colombia) earthquake source parameters from seismic data and InSAR measurements,"

- Journal of South American Earth Sciences*, vol. 72, pp. 250–265, 2016.
- [18] V. De Novellis, R. Castaldo, C. De Luca et al., “Source modeling of the 2015 Wolf volcano (Galápagos) eruption inferred from Sentinel 1-A DInSAR deformation maps and pre-eruptive ENVISAT time series,” *Journal of Volcanology and Geothermal Research*, vol. 344, pp. 246–256, 2017.
- [19] M. Bagnardi and A. Hooper, “Inversion of surface deformation data for rapid estimates of source parameters and uncertainties: a Bayesian approach,” *Geochemistry, Geophysics, Geosystems*, vol. 19, no. 7, pp. 2194–2211, 2018.
- [20] R. Dutta, S. Jónsson, T. Wang, and H. Vasyura-Bathke, “Bayesian estimation of source parameters and associated Coulomb failure stress changes for the 2005 Fukuoka (Japan) earthquake,” *Geophysical Journal International*, vol. 213, no. 1, pp. 261–277, 2017.
- [21] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, “Machine learning in geosciences and remote sensing,” *Geoscience Frontiers*, vol. 7, no. 1, pp. 3–10, 2016.
- [22] G. Chen, S. Li, L. D. Knibbs et al., “A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information,” *Science of the Total Environment*, vol. 636, pp. 52–60, 2018.
- [23] L. Han, G. Yang, H. Dai et al., “Modeling maize above-ground biomass based on machine learning approaches using UAV remote-sensing data,” *Plant Methods*, vol. 15, no. 1, p. 10, 2019.
- [24] P. T. Noi and M. Kappas, “Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery,” *Sensors*, vol. 18, no. 1, p. 18, 2017.
- [25] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O’Reilly Media, 2019.
- [26] A. Agarwal, T. El-Ghazawi, H. El-Askary, and J. Le-Moigne, “Efficient hierarchical-PCA dimension reduction for hyperspectral imagery,” in *2007 IEEE International Symposium on Signal Processing and Information Technology*, pp. 353–356, Giza, Egypt, 2007.
- [27] X. Kang, X. Xiang, S. Li, and J. A. Benediktsson, “PCA-Based edge-preserving features for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7140–7151, 2017.
- [28] Z. Wu, Y. Huang, and K. Zhang, “Remote sensing image fusion method based on PCA and curvelet transform,” *Journal of the Indian Society of Remote Sensing*, vol. 46, no. 5, pp. 687–695, 2018.
- [29] M. Zuo, G. Dai, L. Peng, L. Chen, X. Chen, and Z. Song, “Global optimisation of multiple gravity assist spacecraft trajectories based on search space exploring and PCA,” in *2016 IEEE Congress on Evolutionary Computation (CEC)*, pp. 2655–2660, Vancouver, BC, Canada, 2016.
- [30] S. Karimi, N. Samani, and Z. Mohammadi, “Characterization of Semnan thermal springs using principal component analysis and geochemical inverse modeling,” *Arabian Journal of Geosciences*, vol. 12, no. 24, p. 777, 2019.
- [31] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.
- [32] K.-H. Kim, J.-H. Ree, Y. H. Kim, S. Kim, S. Y. Kang, and W. Seo, “Assessing whether the 2017Mw5.4 Pohang earthquake in South Korea was an induced event,” *Science*, vol. 360, no. 6392, pp. 1007–1009, 2018.
- [33] T. F. Coleman and Y. Li, “An interior trust region approach for nonlinear minimization subject to bounds,” *SIAM Journal on Optimization*, vol. 6, no. 2, pp. 418–445, 1996.
- [34] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013.
- [35] R. Bro and A. K. Smilde, “Principal component analysis,” *Analytical Methods*, vol. 6, no. 9, pp. 2812–2831, 2014.
- [36] F. Pedregosa, G. Varoquaux, A. Gramfort et al., “Scikit-learn: machine learning in Python,” *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [37] K.-K. Lee, *Final report of the Korean Government Commission on relations between the 2017 Pohang earthquake and EGS project*, The Geological Society of Korea, 2019.