scientific reports

OPEN

Rational partitioning of spectral feature space for effective clustering of massive spectral image data

Yusei Ito[1], Yasuo Takeichi[1], Hideitsu Hino[2] & Kanta Ono[1,3]✉

We have successfully proposed and demonstrated a clustering method that overcomes the "needle-in-a-haystack problem" (finding minuscule important regions from massive spectral image data sets). The needle-in-a-haystack problem is of central importance in the characterization of materials since in bulk materials, the properties of a very tiny region often dominate the entire function. To solve this problem, we propose that rational partitioning of the spectral feature space in which spectra are distributed, or defining of the decision boundaries for clustering, can be performed by focusing on the discrimination limit defined by the measurement noise and partitioning the space at intervals of this limit. We verified the proposed method, applied it to actual measurement data, and succeeded in detecting tiny (~ 0.5%) important regions that were difficult for human researchers and other machine learning methods to detect in discovering unknown phases. The ability to detect these crucial regions helps in understanding materials and designing more functional materials.

In bulk materials, the properties of a very tiny region often dominate the entire function. Examples of functionally essential tiny regions include the following:

1. Metal accumulation on the surface of catalytic materials degrades the performance[1].
2. Defects in battery electrodes lead to efficiency loss[2,3].
3. Nanoscale impurities embedded within the structure of solar cell materials lead to reduced solar cell performance[4].

Understanding the role of ultramicro regions is essential to understanding materials and designing more functional materials. However, these regions are only present in extremely small regions compared to the overall size of the material, and where they are located is not known. Therefore, developing nanoscale or submicron-scale measurements over the entire material is essential, i.e., a characterization method that can simultaneously achieve a large field of view and a high resolution, and a data analysis method that can find a small number of phases in the measurement data without overlooking them.

The development of multimodal measurement techniques with a large field of view and a high resolution has been actively studied in various fields. Spectroscopic imaging measurement techniques such as hyperspectral imaging[5], infrared spectroscopic imaging[6,7], Raman imaging[8–10], and X-ray absorption spectroscopic imaging[11–16] enable the acquisition of spectra reflecting chemical and electronic states at each point in space. In particular, X-ray absorption spectroscopic imaging is widely used for analyzing the reaction process of battery materials because the obtained spectra reflect the valence and local structure of the absorbing elements. Our newly developed gigapixel X-ray microscope (G-XRM) can measure spectra with a submicron spatial resolution in a centimeter-order field of view. As a result, approximately one billion spectra can be obtained in one measurement. However, data analysis methods for finding minuscule important regions from massive spectral image data sets have yet to be studied. Despite the development of measurement techniques, there is currently no practical solution to the "needle-in-a-haystack problem" of data analysis (finding tiny regions compared to the entire enormous data set).

There are two main approaches to spectral data analysis by machine learning techniques: supervised learning[17–21] and unsupervised learning[22–29]. In supervised learning, some methods use predefined standard

[1]Department of Applied Physics, Osaka University, 2-1 Yamadaoka, Suita 565-0871, Osaka, Japan. [2]The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190- 8562, Japan. [3]Institute of Materials Structure Science, High Energy Accelerator Research Organization (KEK), Tsukuba, Japan. ✉email: ono@ap.eng.osaka-u.ac.jp

spectra and perform linear regression for each point in space[17,18]. Moreover, methods that use random forest[19] or neural networks[20,21] to infer specific physical properties have been developed. However, linear regression is impossible to use when dealing with phases whose spectra are unknown. In scenarios where the goal is to visualize the spatial distribution of a physical quantity obtained by spectral analysis, such as inferring the magnetic moment from an X-ray magnetic circular dichroism (XMCD) spectrum, inferring a specific physical quantity from the spectrum is suitable. However, this method cannot be applied when the intrinsic physical properties of minority phases governing the bulk properties are unknown. In contrast, unsupervised learning methods use only measured data and classify regions considering the spectral similarity for each measured spatial point. The researcher then analyzes these classified regions to understand the material. In contrast to supervised methods, regions can be identified even when the spectra of phases contained throughout the material are unknown in advance. Unsupervised classification methods can be broadly divided into non-density-based methods and density-based methods. Non-density-based methods can be further divided into centroid-based methods represented by the K-means method[22–25], neural networks[26,27], and hierarchical clustering[28]. The K-means method is an unsupervised classification method that performs with a given number of clusters and is known for its high computational efficiency. In recent years, methods that optimize computational efficiency further have been developed[25], making it easier to apply to large datasets. Additionally, there are some research to develop methods that can accurately determine the number of clusters[29,30], which has been a limitation of the traditional K-means method. However, these methods tend to overlook minority phases because they minimize the sum of intra-cluster variances as a loss function, where the contribution of the minority phase is insignificant. Neural network-based methods perform dimensionality reduction to obtain effective latent representation of input data, then use typical clustering method in the latent space. However, the number of data points remains same even if the dimension is reduced, therefore the same problem of K-means method arises. In contrast, hierarchical clustering calculates the distance between each data point and classifies them based on a certain threshold. This method overcomes the drawback of centroid-based methods in identifying minority phases, as it calculates the distances between all points. However, it requires $O\left(N^2\right)$ computations, which makes it impossible to apply to large datasets (more than 1 M pixels). In density-based clustering methods, DBSCAN[31–35] and MeanShift[36] are considered the most representative approaches. These methods classify data based on density, which eliminates the need to designate the number of clusters and allows for accurate classification even when there is significant bias within the clusters. While both methods typically have a computational complexity of $O\left(N^2\right)$, recent advancements in computational efficiency and the use of appropriate approximations of DBSCAN have reduced this complexity to $O\left(N\log N\right)$or $O\left(N^2\right)$ with small constant[33–35]. However, a common drawback of these methods is the need to specify the non-trivial hyperparameter of density threshold, which greatly affects the results depending on its adjustment. As discussed above, it is difficult to address the "needle-in-a-haystack problem" by using current machine learning methods.

In this study, we focused on the low-dimensional spectral feature space in which all possible X-ray absorption spectroscopy (XAS) spectra are distributed. Previous studies have demonstrated that the valence and coordination state of an absorbing atom can be estimated by carefully selecting hand-crafted features based on the researcher's expertise[37]. Thus, considering a spectrum as a point in high-dimensional data space with dimensions of the measured energy points or as a point in infinite-dimensional functional space, the possible spectrum would be distributed in a thin membrane-like low-dimensional manifold (spectral feature space) rather than densely filling the high-dimensional data space. This intrinsic low dimensionality of high-dimensional natural data, called the manifold hypothesis[38], is believed to hold not only for scientific measurement data such as XAS spectra but also for image data such as handwritten numbers[39,40] and crystal structures[41]. As shown in Fig. 1, this paper proposes an unsupervised classification method by rationally partitioning the spectral feature space in high-dimensional data space. Because of partitioning the spectral feature space in advance, even minority phases can be detected. To realize rational partitioning, we considered the measurement noise. If there is no measurement noise, then even slightly different spectra should be classified into different clusters. However, noise exists in actual measurements, which means that the discrimination limit can be determined by the amount of noise. Therefore, rational partitioning of the spectral feature space is realized by performing partitioning at intervals of approximately the discrimination limit. In addition, to ensure robustness against irregular noise in actual measurements, such as sample thickness variations and sample drift during measurement, information on the spatial correlation, where adjacent points have similar chemical states, is incorporated.

First, we applied the developed method to synthetic data to verify that it solves the "needle-in-a-haystack problem," i.e., analyzes entire materials without missing a tiny minor phase. The study was further validated using materials with heterogeneous chemical state distributions. Actual materials such as iron ore sinter[42,43] were used in the study. Iron ore sinter, a crucial raw material in iron making, is produced by liquid-phase sintering of iron ore with limestone and coke. Due to its natural origin, it has an inherently heterogeneous internal structure from micrometer to millimeter scale and a mixture of different chemical phases. The minority phases within the structure are responsible for the degradation of the reduction performance in a blast furnace[44]. However, the details of this structure still need to be clarified. Therefore, extracting intrinsically important regions from a large spectral data set that measures a large area with a high resolution is essential to understanding the ironmaking process and making it more efficient. In this study, approximately 1 billion spectra measured by G-XRM were analyzed by the proposed method, and phases present in tiny amounts, which have been overlooked in conventional analysis, were successfully found for the first time.

Results

Evaluation of the spectral feature space

In this paper, XAS spectra are considered as vectors in high-dimensional data space, which is Euclidean space with dimensions corresponding to the measured energy points. In XAS spectrum, the constant sum added

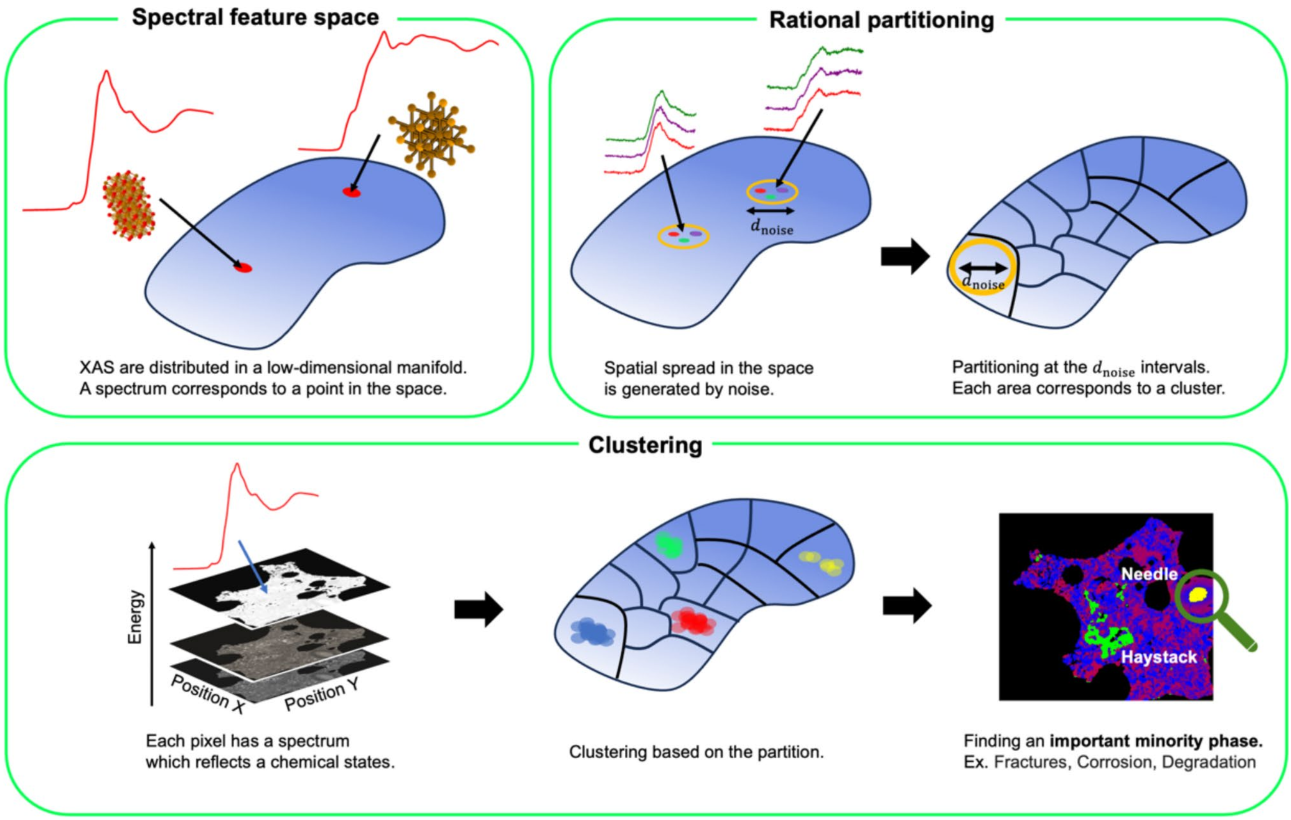
Spectral feature space
XAS are distributed in a low-dimensional manifold.
A spectrum corresponds to a point in the space.
Rational partitioning
d_{noise}
Spatial spread in the space
is generated by noise.
d_{noise}
Partitioning at the d_{noise} intervals.
Each area corresponds to a cluster.
Clustering
Energy
Position X
Position Y
Each pixel has a spectrum
which reflects a chemical states.
Clustering based on the partition.
Needle
Haystack
Finding an **important minority phase.**
Ex. Fractures, Corrosion, Degradation

Fig. 1. Concept of the proposed clustering method. (Upper left) A spectrum corresponds to a point in the low-dimensional space (spectral feature space). (Upper right) For measurement noise, spatial spread in the space is generated and rational partitioning of the space is dividing at interval of the spread. (Lower) Each pixel has a spectrum which corresponds to a point in the spectral feature space. Because of partitioning the spectral feature space in advance, even minority phases can be detected.

to the spectrum represents the background, such as the absorption of nontarget elements, and the constant multiplication is proportional to the density of the absorbing element. Since this paper aims to classify spectra by the chemical state, we standardized the spectra to ensure invariance to the chemical state. In terms of the similarity between two spectra, this standardization is equivalent to calculating the Pearson similarity between spectra before standardization, and the Pearson similarity has already been reported to be a good similarity metric for XAS spectra[45].

Taking the Fe-K edge XAS spectra as an example, all XAS spectra have similar spectral shapes, with a sharp step-like increase at the Fe-K edge energy and a wavy shape afterward. This similar spectral shape demonstrates that the XAS spectra are not densely distributed in high-dimensional data space but on a low-dimensional spectral feature space, as shown in Fig. 1 (upper left). We verify the low dimensionality of the XAS spectral feature space with a standard spectral data set. The intrinsic dimensionality for the Fe-K edge (178 energy points) 61 XAS data set from the NIMS MDR XAFS database[46] is estimated by the Fukunaga-Olsen method[47] to be 3-dimensional which is much smaller than the dimensionality of the data space.

Partitioning of the spectral feature space

Solving a classification problem is equivalent to partitioning the spectral feature space or defining decision boundaries. In physical measurements, noise must be present, and its amount can be estimated based on the measurement conditions. If the distance between two spectra is less than the magnitude of the measurement noise, then they are indistinguishable. Therefore, partitioning the spectral feature space into intervals corresponding to the measurement noise is reasonable. As shown in Fig. 2, this partitioning is approximately realized by arranging the Voronoi diagram site points at intervals of the measurement noise under the constraint of all possible XAS spectra and performing Voronoi tessellation[48]. From an implementation standpoint, clustering based on Voronoi tessellation is realized by performing a nearest neighbor search for each measured spectrum, with the Voronoi diagram site points as the search targets, which means that each pixel can be handled independently, therefore clustering can be performed with an $O(N)$ computational complexity. Note that although the data has dimensionality of more than 150, the 'curse of dimensionality'[49] is not a problem due to the low dimensionality of the spectral feature space (approximately 3-dimensional).

We assumed that a linear combination of standard spectral data sets can represent spectra in the spectral feature space that consists of all possible XAS spectra. Therefore, we used the linear combination of spectra in

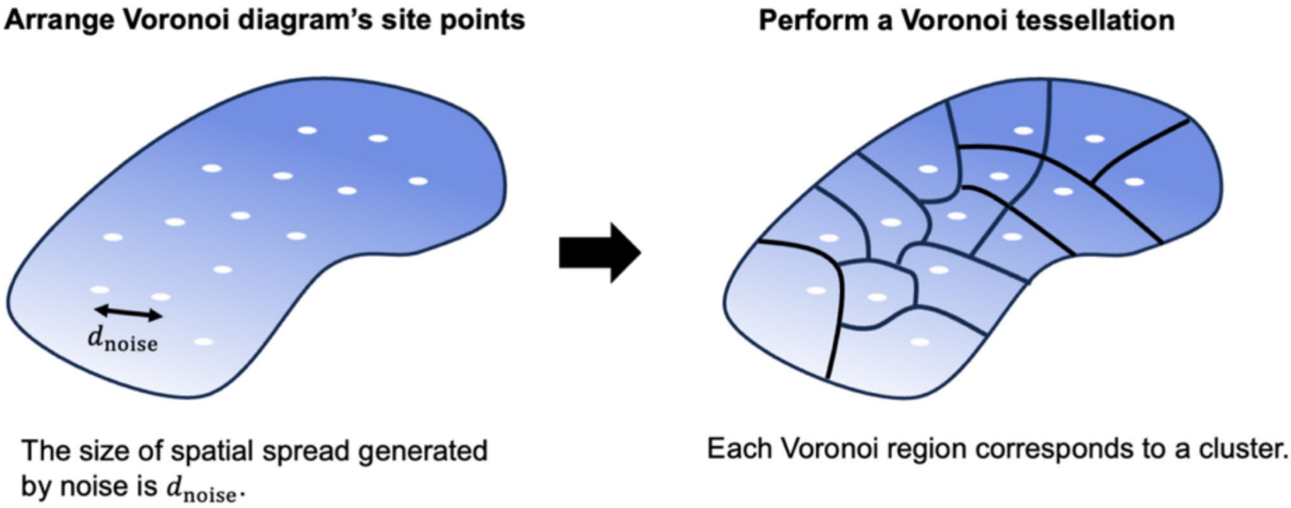
Arrange Voronoi diagram's site points
Perform a Voronoi tessellation
d_{noise}
The size of spatial spread generated by noise is d_{noise}.
Each Voronoi region corresponds to a cluster.

Fig. 2. Realizing rational partitioning. By arranging Voronoi diagram site points at intervals of measurement noise and performing Voronoi tessellation, it can be approximately realized to divide the spectral feature space at the interval.

the data set as candidates for Voronoi site points. We then calculated the spacing of the Voronoi site points by using the metric determined by the measurement noise, which is formally described below. As discussed, the spectra measured for identical chemical state spread out to the distance corresponding to the measurement noise in the spectral feature space. Therefore, we reduced the number of candidates for site points so that the size of the partitioning area was the same as the noise. The following describes the metric of distances between candidate site points for Voronoi tessellation that considers measurement noise. For the candidate site point spectra $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, let us consider the Voronoi tessellation of the spectral feature space based on these two points: the Voronoi region containing $\boldsymbol{S}_1$ is V_1, and the region containing $\boldsymbol{S}_2$ is V_2. If the distance between these two site points is smaller than the spread of the noise in the spectral feature space, then spectrum $\boldsymbol{S}_{1\mathrm{g}}$ generated by adding the measurement noise to spectrum $\boldsymbol{S}_1$ is included in V_2. In situations where the classification is based on Voronoi tessellation, this results in a misclassification, and the probability can be thought of as the similarity between the site points. The similarity of the two site points can be described by Eq. (1), which shows the misclassification rate when spectra are generated by adding noise to spectra $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ and classified by Voronoi tessellation.

$$s_{\mathrm{error1}}\left(\boldsymbol{S}_1,\ \boldsymbol{S}_2\right) = \left(\int_{d\left(\boldsymbol{S}_1,\ \boldsymbol{S}_{\mathrm{g}}\right)>d\left(\boldsymbol{S}_2,\ \boldsymbol{S}_{\mathrm{g}}\right)} p_1\left(\boldsymbol{S}_{\mathrm{g}}\right) d\boldsymbol{S}_{\mathrm{g}} + \int_{d\left(\boldsymbol{S}_1,\ \boldsymbol{S}_{\mathrm{g}}\right)>d\left(\boldsymbol{S}_2,\ \boldsymbol{S}_{\mathrm{g}}\right)} p_2\left(\boldsymbol{S}_{\mathrm{g}}\right) d\boldsymbol{S}_{\mathrm{g}}\right)/2 \tag{1}$$

where $d\left(\boldsymbol{S}_{\mathrm{a}},\ \boldsymbol{S}_{\mathrm{b}}\right)$ is the Euclidean distance after standardization between the two spectra $\boldsymbol{S}_{\mathrm{a}}$ and $\boldsymbol{S}_{\mathrm{b}}$. $p_1\left(\boldsymbol{S}_{\mathrm{g}}\right)$ represents the probability that spectrum $\boldsymbol{S}_1$ plus noise determined from the measurement conditions will produce spectrum $\boldsymbol{S}_{\mathrm{g}}$. Similarly, $p_2\left(\boldsymbol{S}_{\mathrm{g}}\right)$ represents the probability that spectrum $\boldsymbol{S}_{\mathrm{g}}$ is generated from spectrum $\boldsymbol{S}_2$. Hierarchical clustering[28] was performed by quantifying the measurement noise, approximating this equation as described in the Methods section, and calculating the similarity between all candidate site points to reduce the number of site points that could be reasonably divided. Sixty-one standard spectra were obtained from the XAS database, and 5967 candidate site points were generated by linear combination of the database spectra as described in the Methods section. A similarity matrix of candidate points was calculated by approximating Eq. (1), and 1528 sites were selected through reduction based on hierarchical clustering with a threshold set at 0.0001, which means the misclassification probability is approximately 0.01%.

To confirm that the Voronoi site points are densely distributed in the spectral feature space, a synthetic spectrum was created by randomly selecting N spectra from the standard spectral data set and linearly combining them by random weights. The maximum values for the similarity between the synthetic spectra and the Voronoi site points were calculated by using Eq. (1). Since we used linear combination of 3 (= estimated intrinsic dimensionality of the spectral feature space) components for generating the candidate of Voronoi site points, we checked the denseness by spectra generated by more than 3 components: $N = 3$ to 5, 10,000 times in each N, 30,000 times in total. Supplementary Fig. 1 shows the distribution of s_{error} values where blue color range is in the threshold, and red color range is over. The rate of over the threshold is approximately 2%. This result confirms that the site points of Voronoi tessellation are mostly, although not perfectly, realized in a dense distribution, which means the 'curse of dimensionality' does not arise.

Verification with synthetic data and improvement of noise redundancy using spatial correlation

The ability of the proposed method to classify minority spectra (< 1%) without overlooking them was verified using synthetic data. Figure 3(a) shows a XAS data set for validating the classification of a minority spectrum

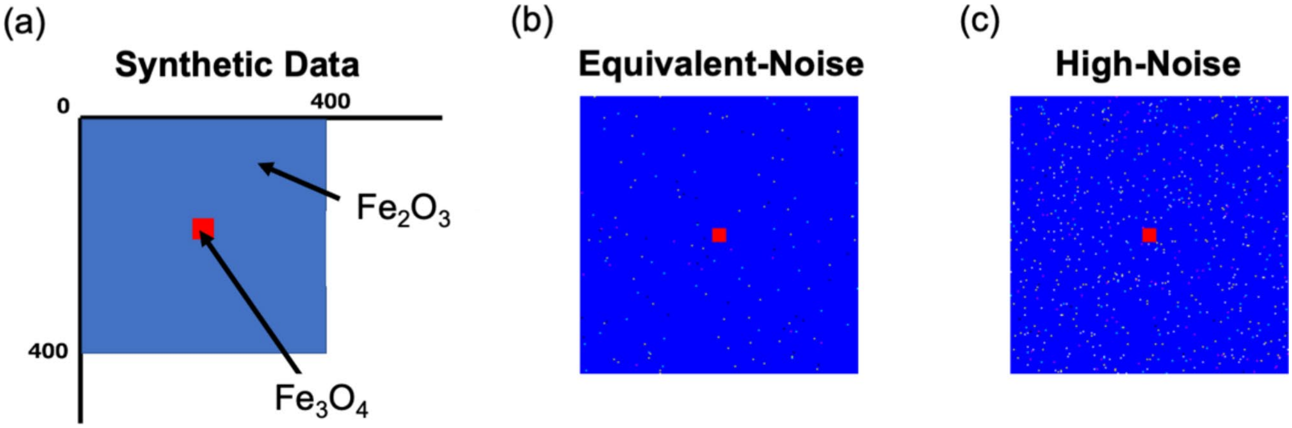

Fig. 3. Result of synthetic data analysis. (**a**) Schematic diagram of the prepared of synthetic data. 400 × 400 pixels are filled with Fe_2O_3 spectrum in all regions except for the central 10 × 10 pixels containing Fe_3O_4 spectrum. (**b**) Clustering result by the proposed method where the amount of added noise is equivalent to that used for site points reduction. (**c**) Clustering result where the amount of the noise is high.

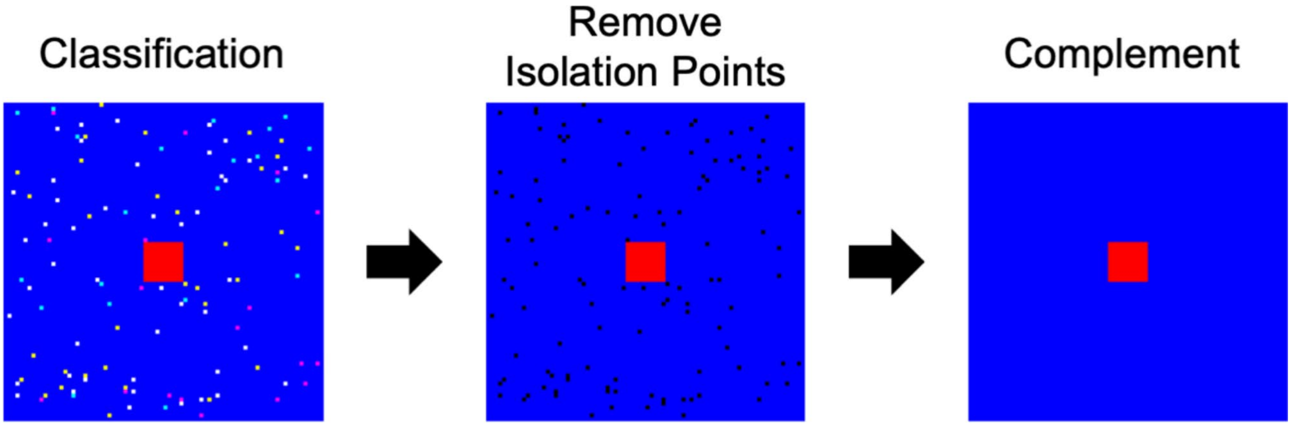

Fig. 4. Result of improvement of noise redundancy using spatial correlation. First, remove the isolation points, then complement the points.

occupying 1/1600 of total; 400 × 400 pixels are filled with Fe_2O_3 spectrum in all regions except for the central 10 × 10 pixels containing Fe_3O_4 spectrum which is obtained by measuring standard samples (very low noise measurement). It is a common situation for minority phases to cohere in one area in real systems, rather than being scattered. Noise was added to all spectra, as described in the Methods section. Classification was performed on these data using the proposed method. Figure 3(b) shows the case where the amount of added noise is equivalent to that used for site point reduction, Fig. 3(c) shows the case where the amount of added noise is high $K = 2$. When more noise is added than that used for site points reduction, the number of classes becomes 11 with an error rate of 1.9%, which is significantly different from the correct number of classes of 2. This occurs because the number of site points is determined by allowing a misclassification rate of 0.01%, which would result in approximately ten misclassifications when applied to around 100,000 spectra. Furthermore, the grand-truth spectra of the clusters only partially match the site points.

To solve this problem, we integrated the spatial information by assuming that adjacent points have similar chemical states. We determined the minimum size for the spatial region of each cluster, and if the size was not met, then the region was considered misclassified, and the points in that region were complemented based on the surrounding information. The details are described in the supplementary materials. We used K_{size} =5 and P_{num}=23 for the parameters of the method in the following analysis. Figure 4 shows the results of using this method in the synthetic data analysis. The data, which were classified into 11 clusters when only the classification was performed, were successfully classified into 2 clusters by integrating spatial information and the data. Supplementary Fig. 3 and Supplementary Table 1 show the image of the results of our proposed method and statistical analysis, along with comparisons to the K-means method as shown before and the DBSCAN method, which is considered the state of the ate method for identifying minority phases due to its density-based approach and relatively low computational complexity ($O\left(N\log N\right)$). In the K-means method, it cannot be said that the method correctly classifies the data since it recognizes the majority phase as two separate phases rather than one due to the small contribution of minority spectra to the loss function, as described above. The DBSCAN method, on the other hand, roughly classifies the data correctly, however, there remains the issue of the minority phases being divided into several phases. The next section demonstrates that when applied to real data, it leads to fatal misclassification. Additionally, this method is required non-trivial hyperparameter tuning. For the synthetic data used in this study, we obtained what we considered to be the optimal parameters after approximately ten trials. However, performing such extensive tuning on large-scale real materials data, where the correct parameters are completely unknown, is not realistic.

Refer to Supplementary Fig. 4 for the number of classes and misclassification rate dependent on the amount of added noise. These results show that the proposed method including the spatial information realizes a high noise tolerance. The minimal size of the detection limit is theoretically determined by K_{size}. The rates of theoretical detection limits reached when varying the amount of noise are shown in Supplementary Fig. 5. The amount of noise was approximately 0.04 in the experimental data discussed later. Therefore, this method is confirmed to be able to detect up to the theoretical limit in practical conditions.

In summary, clustering that overcomes the needle-in-a-haystack problem can be realized by the rational partitioning of low-dimensional spectrum feature space by accounting for measurement noise. Furthermore, the integration of spatial information ensures robustness against the measurement noise. When the proposed method was used for synthetic data, the minority phase could be detected even with more noise than expected.

Application to large-scale measurement data of actual materials

We adapted the proposed method to classify G-XRM measurement data of iron ore sinter to understand the critical microstructure. As mentioned before, extracting small, important regions from massive spectral data sets is essential to understand and improve the ironmaking process.

We performed G-XRM measurements at the Fe K-edge of sintered ore provided by a steel mill. The measurement area was 0.8 cm x 1.2 cm, and each pixel was approximately 500 nm x 500 nm, resulting in 450 million pixels. The analysis with the proposed method could be performed in approximately 20 h by AMD Ryzen Threadripper 3990 × 64-Core Processor. The analysis resulted in a total of 57 clusters. Careful examination of the mean spectra of each cluster by human experts confirmed flattening of the peak areas that occurs when the sample thickness is not appropriate. Furthermore, some of the spectra could be described by a linear combination of two other spectra. Therefore, the spectra with flattened peaks were replaced with the proper cluster, and the spectra that could be considered a linear combination of two other spectra were detected and removed. Then, component ratios of the removed points were calculated by performing a nonnegative linear regression using their basis spectra. Finally, the number of components was determined to be 5. Currently, the measured data classification for actual materials is partially automated since each spectrum is validated based on expert knowledge. However, automatically obtaining 57 different components from 450 M spectra based on the measurement noise and organizing the classification results based on researchers' objectives and domain knowledge is one possible solution to the "needle-in-a-haystack problem". The results are shown in Fig. 5, where the colors correspond to clusters representing different chemical states, and the shading shows the component ratios and iron density. The microregion shown in yellow in the figure, which accounts for only 0.5% of the total, is the most exciting classification result. From careful inspection to the component spectrum, the yellow region was suggested as

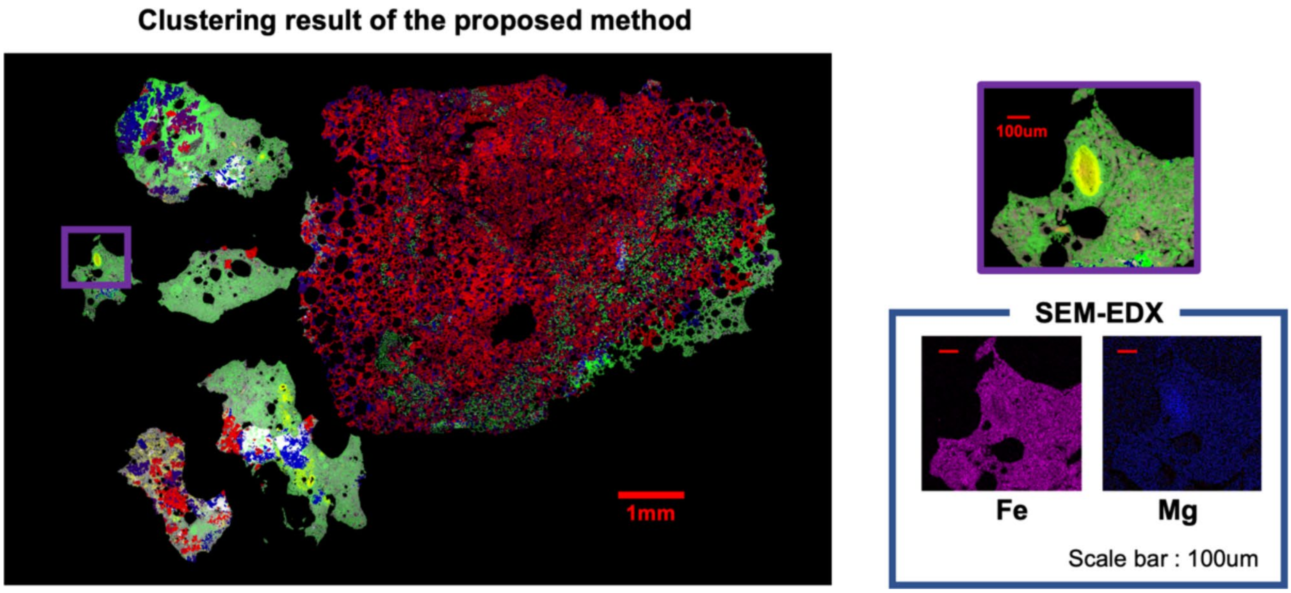

Fig. 5. Clustering result of the large-scale measurement data of an actual material and results of SEM-EDX measurements. The colors correspond to clusters representing different chemical states, and the shading shows the component ratios and iron density in the clustering result image. SEM-EDX measurements show that the yellow phase is Mg-rich phase.

magnesium ferrite phase. Scanning electron microscopy with energy-dispersive X-ray spectroscopy (SEM-EDX) measurement showed a specific concentration of Mg surrounded by a high-Fe layer at the yellow region. This feature corresponds to the magnesioferrite layer around unmolten dolomite or serpentine[50] and the adhesive layer surrounding iron ore relicts[51]. It is worth emphasizing that the newly discovered component was derived only by the spectral analysis. As shown in the application to the practical data, the proposed method based on the division of the spectral feature space has successfully discovered a tiny phase out of large-scale (~ 1G pixel) measurement data, without pre-defining the spectral feature or region of interest. Additionally, Supplemental Fig. 6 shows comparing results with existing methods, K-means and DBSCAN, on the cropped data including this intriguing yellow region.

Discussion

In this study, we developed a clustering method that overcomes the "needle-in-a-haystack problem" by focusing on the spectral feature space of XAS spectra and partitioning the space at intervals of the discrimination limit determined by the measurement noise. We assumed that all the possible spectra can be considered to represent the linear combinations of spectra in the database. Using this assumption and considering low dimensionality of the spectral feature space, we can enumerate all the possible spectra that can be discriminated. As a result, we acquired a set of approximately a thousand spectra and then partitioned the spectral feature space for clustering by Voronoi tessellation by using these spectra as the site points. Here, we discuss the potential for application to other measurement data with the same data structure and measurement optimization considering the discrimination limit.

First, we discuss the potential limitations and possible future works.

1. The proposed method can be used when the spectral feature space is low-dimensional, and its basis spectra are known. Even with other measurement methods, the latent spectral feature space is considered a low-dimensional manifold based on the manifold hypothesis, and methods for automatically determining the manifold from data have been proposed by using neural networks[52,53]. Combining these methods enables classification to be performed using only measurement data and can be applied for measurement methods for which an extensive database is available.
2. In the proposed method, to ensure robustness against noise, clusters that are not spatially localized are considered as misclassifications. As a result, it is not possible to achieve accurate classification when the chemical state changes on a scale comparable to the pixel size. If, instead of using spatial information, we determine whether each cluster's spectrum has a high level of noise, it might be possible to achieve classification at the single-pixel level.

Next, we discuss the design of the optimal measurement experiment. In the present method, the number of candidates for Voronoi site points is reduced by focusing on the discrimination limit and quantifying it by Eq. (1). Conversely, suppose that the pairs of spectra that need to be distinguished are predetermined. In this case, combined with active learning[54,55], this method can be applied to automated stopping of the measurement when the discrimination limit reaches the difference between the spectra of the pairs.

In conclusion, we proposed rational partitioning of spectral feature space by accounting for measurement noise for effective clustering of massive spectral image data. By arranging the Voronoi site points at intervals of approximately the measurement noise, we achieved a classification in which the spectra within the same cluster can be regarded as the same spectrum. Thus, the probability of misclassification due to measurement noise is low. We applied this method to both synthetic and measured data of X-ray absorption spectroscopic imaging and confirmed that this method could solve the needle-in-a-haystack problem. We used the database for understanding the spectral feature space, however combining latent representation methods makes it possible to applied to other measurement methods which have no appropriate database. Additionally, our proposed metric based on the measurement noise is important not only for clustering problem but also measurement optimization.

Methods

Sample preparation

Iron ore sinter used in this study is provided by a steel mill and was studied as 'Sinter5 in Ref[56]. The sample was resin embedded, polished to 30- μ m-thick, and then lifted off from the glass slide to prepare a stand-alone slice for transmission mode X-ray observation.

Experiments

G-XRM measurements of iron ore sinter were performed at beamline NW2A of the Photon Factory Advanced Ring (PF-AR) of the High Energy Accelerator Research Organization (KEK)[57]. Figure 6 shows the measurement setup for G-XRM. X-rays generated by the undulator were monochromatized by a double crystal monochromator and irradiated on the sample. The X-rays transmitted from the sample were converted to visible light by a scintillator, magnified by a 10x lens, and detected by an sCMOS camera (Hamamatsu Photonics ORCA-Quest). The X-ray energy was scanned to measure the two-dimensional XAS data. The sCMOS camera had 2304 × 4096 pixels with an effective pixel size of 0.46 μm. The exposure time at each energy point was set to 1.5 s, and 178 energy points were measured from 7009.2 eV to 7321.2 eV around the Fe-K absorption edge. Including the moving time of the double crystal monochromator, the measurement time for one field of view was approximately 10 min. We also measured the reference images at each energy points, then calculated an absorbance of each pixel by using Eq. (2).

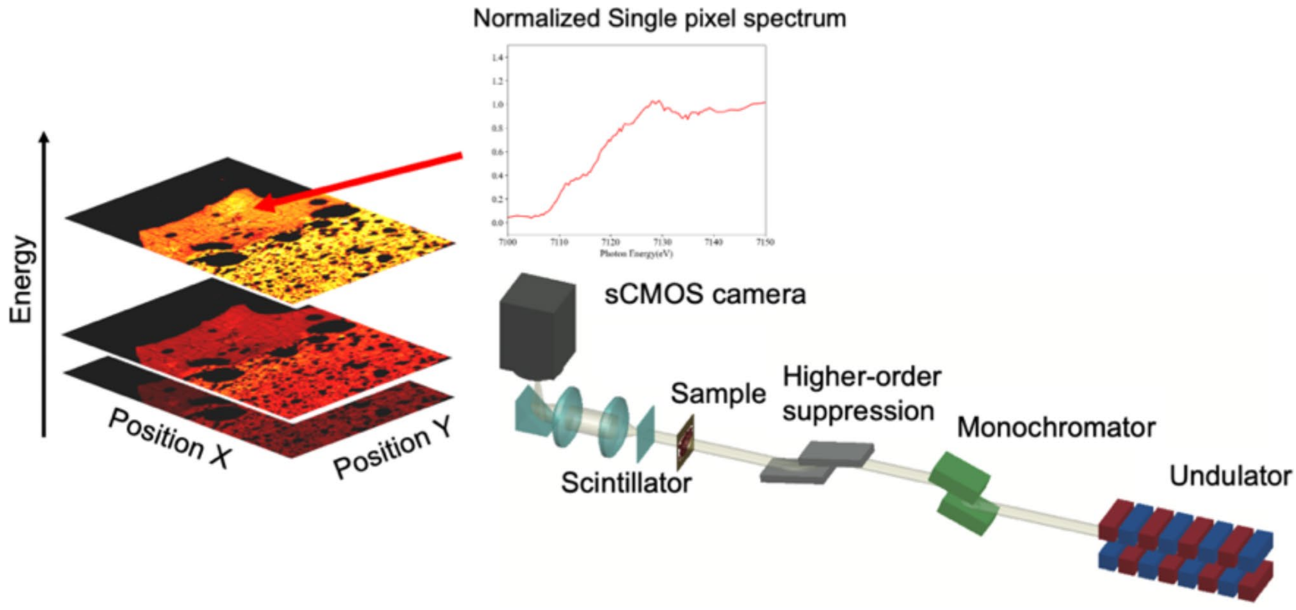

Fig. 6. Experimental setup of G-XRM. X-ray absorption spectrum of each pixel in the space is obtained by taking the reference and transmitted light at each X-ray energy.

$$\mu t = \ln\left(\frac{c_{\mathrm{ref}}}{c_{\mathrm{trans}}}\right), \tag{2}$$

where c_{ref} and c_{trans} represents the counts of reference and transmitted X-rays, respectively.

By scanning the sample position and measuring for a total of 56 fields of view, the XAS spectrum of each point in space in an area of 0.8 cm × 1.2 cm was measured. We used the method described in Ref[58] to correct the energy shift, normalized the spectra by conventional method[59], and stitched the images of each measurement by the ImageJ stitching plugin[60]. The number of pixels after stitching was approximately 450 M pixels.

Quantification of measurement noise

To measure the signal to noise ratio of the sCMOS camera, we repeatedly performed measurements with the same field of view and exposure time and then estimated the standard deviation for the number of counts. Assuming that the noise is proportional to the square root of the number of counts and background noise exists, the relationship between the standard deviation σ and the number of counts N can be represented as $\sigma = \sqrt{A \times N - B}$ with A and B as parameters. We performed least square fitting; the results were $A = 9.3131$ and $B = 63.003$. By considering the propagation of uncertainty[61], the standard deviation of the absorbance calculated by Eq. (2) was estimated as shown in Eq. (3).

$$\sigma_{\mu t} = \sqrt{\frac{A\times c_{\mathrm{ref}} - B}{c_{\mathrm{ref}}^2} + \frac{A\times c_{\mathrm{trans}} - B}{c_{\mathrm{trans}}^2}} \tag{3}$$

Since we perform conventional normalization[59] after calculating an absorbance by Eq. (2), we need to modify this equation for calculating the standard deviation of the normalized absorbance. As shown in Supplemental Fig. 7, normalization processes are two steps: subtract the pre-edge line and divide by the post-edge line. We defined $\mu\, t_{\mathrm{low}}$ as mean absorbance value of the pre-edge region, and $\mu\, t_{\mathrm{high}}$ as mean absorbance value of the post-edge region. For calculating the standard deviation of the normalized absorbance, reference counts value, $\mu\, t_{\mathrm{low}}$, $\mu\, t_{\mathrm{high}}$ and normalized absorbance value are needed. By using these values, the standard deviation of the normalized absorbance can be calculated by the following steps : First, calculating c_{trans} by $c_{\mathrm{ref}} \exp\left\{-\mu\, t_{\mathrm{normalize}} \left(\mu\, t_{\mathrm{high}} - \mu\, t_{\mathrm{low}}\right) - \mu\, t_{\mathrm{low}}\right\}$, where $\mu\, t_{\mathrm{normalize}}$ is the normalized absorbance value. Then, the standard deviation of the normalized absorbance can be obtained by dividing the $\sigma_{\mu\,\mathrm{t}}$ by $\mu\, t_{\mathrm{high}} - \mu\, t_{\mathrm{low}}$ corresponding to the second step of normalization.

In the G-XRM measurement, the measurement noise is different for each pixel because each pixel has a different c_{ref}, $\mu\, t_{\mathrm{low}}$ and $\mu\, t_{\mathrm{high}}$. However, we used the average of these values to reduce the number of site points, so that all pixels were classified using the same Voronoi tessellation. From our G-XRM experimental conditions, $\left(\overline{c_{\mathrm{ref}}}, \overline{\mu t_{\mathrm{low}}}, \overline{\mu t_{\mathrm{high}}}\right) = \left(2.7 \times 10^4,\ 0.36,\ 1.56\right)$.

Approximation of Eq. (1)

This section describes the approximation method for the first term of Eq. (1) as shown in Eq. (4). The second term can be calculated in the same way.

$$s_{\mathrm{error1}}\left(\boldsymbol{S}_1,\ \boldsymbol{S}_2\right) = \int_{d\left(\boldsymbol{S}_1,\, \boldsymbol{S}_{\mathrm{g}}\right)>d\left(\boldsymbol{S}_2,\, \boldsymbol{S}_{\mathrm{g}}\right)} p_1\left(\boldsymbol{S}_{\mathrm{g}}\right) d\boldsymbol{S}_{\mathrm{g}} \tag{4}$$

where $d\left(\boldsymbol{S}_{\mathrm{a}}, \boldsymbol{S}_{\mathrm{b}}\right)$ is the Euclidean distance calculated after standardizing spectra $\boldsymbol{S}_{\mathrm{a}}$ and $\boldsymbol{S}_{\mathrm{b}}$, and $p_1\left(\boldsymbol{S}_{\mathrm{g}}\right)$ is the probability that spectrum $\boldsymbol{S}_{\mathrm{g}}$ is generated by adding measurement noise to spectrum $\boldsymbol{S}_1$. Using the definition of the Pearson similarity, we obtain $d\left(\boldsymbol{S}_1,\ \boldsymbol{S}_{\mathrm{g}}\right) > d\left(\boldsymbol{S}_2,\ \boldsymbol{S}_{\mathrm{g}}\right) \Leftrightarrow s_{\mathrm{pearson}}\left(\boldsymbol{S}_1,\ \boldsymbol{S}_{\mathrm{g}}\right) < s_{\mathrm{pearson}}\left(\boldsymbol{S}_2,\ \boldsymbol{S}_{\mathrm{g}}\right)$, where $s_{\mathrm{pearson}}\left(\boldsymbol{S}_{\mathrm{a}},\ \boldsymbol{S}_{\mathrm{b}}\right)$ is the Pearson similarity between $\boldsymbol{S}_{\mathrm{a}}$ and $\boldsymbol{S}_{\mathrm{b}}$. Since Eq. (4) represents the probability that spectrum $\boldsymbol{S}_{\mathrm{g}}$ generated by adding noise to spectrum $\boldsymbol{S}_1$ satisfies $s_{\mathrm{pearson}}\left(\boldsymbol{S}_1,\ \boldsymbol{S}_{\mathrm{g}}\right) < s_{\mathrm{pearson}}\left(\boldsymbol{S}_2,\ \boldsymbol{S}_{\mathrm{g}}\right)$, Eq. (4) can be calculated based on the probability distribution of $s_{\mathrm{pearson}}\left(\boldsymbol{S}_2,\ \boldsymbol{S}_{\mathrm{g}}\right) - s_{\mathrm{pearson}}\left(\boldsymbol{S}_1,\ \boldsymbol{S}_{\mathrm{g}}\right)$ and the cumulative probability that it is greater than 0. Given that the absorbance at each energy point follows a normal distribution, the probability distribution of each element in $\boldsymbol{S}_{\mathrm{g}}$ can be represented as $S_{\mathrm{g}i} \sim N\left(\mu_{\mathrm{g}i}, \sigma_{\mathrm{g}i}^2\right)$. Assuming that the absorbance at each energy point follows an independent normal distribution, the required mean value and norm for standardizing $\boldsymbol{S}_{\mathrm{g}}$ can be approximated as constant values $\mu_{\ g}$ and N_g. In this case, the probability distribution of $s_{\mathrm{pearson}}\left(\boldsymbol{S}_2,\ \boldsymbol{S}_{\mathrm{g}}\right) - s_{\mathrm{pearson}}\left(\boldsymbol{S}_1,\ \boldsymbol{S}_{\mathrm{g}}\right)$ can be written as in Eq. (5).

$$s_{\mathrm{pearson}}\left(\boldsymbol{S}_2,\ \boldsymbol{S}_{\mathrm{g}}\right) - s_{\mathrm{pearson}}\left(\boldsymbol{S}_1,\ \boldsymbol{S}_{\mathrm{g}}\right) \cong \sum_i \frac{\left(S'_{2i} - S'_{1i}\right)}{N_g} S_{gi} - \sum_i \left(S'_{2i} - S'_{1i}\right)\frac{\mu_g}{N_g} \sim N(\mu, \sigma^2) \tag{5}$$

where

$$\mu \ = \sum\nolimits_i \frac{\left(S'_{2i} - S'_{1i}\right)}{N_{\mathrm{g}}} \mu_{\ \mathrm{g}i} - \sum\nolimits_i \left(S'_{2i} - S'_{1i}\right)\frac{\mu_{\ \mathrm{g}}}{N_{\mathrm{g}}},\ \ \sigma^{\ 2} = \sum\nolimits_i \left(\frac{\left(S'_{2i} - S'_{1i}\right)}{N_{\mathrm{g}}}\right)^2 \sigma_{\ \mathrm{g}i}^{\ 2}.$$

As a result, Eq. (4) can be approximated by the cumulative probability of $s_{\mathrm{pearson}}\left(\boldsymbol{S}_2,\ \boldsymbol{S}_{\mathrm{g}}\right) - s_{\mathrm{pearson}}\left(\boldsymbol{S}_1,\ \boldsymbol{S}_{\mathrm{g}}\right)$ which follows a normal distribution.

Generation of the candidate site points for Voronoi tessellation

The threshold of similarity calculated by Eq. (1) is set to 0.0001 for the reduction of site points; therefore, the candidate points are generated such that the similarity between the points is larger than this threshold over the entire spectral feature space. Selecting two spectra with duplicates from the standardized spectra data set and performing a one-to-one linear combination, we initially chose 1891 spectra as the candidates. As mentioned in the Results section, the intrinsic dimensionality of the spectral feature space is 3, which means that points of the spectral feature space can be represented by a linear combination of approximately three spectra in the data set. Hence, the following operation was repeated 10,000 times to generate the site points: We randomly selected three spectra from the data set, and then, the spectra were generated by linear combination using randomly generated weights. The similarities between the generated spectrum and the candidate site points can be calculated by Eq. (1). If the similarity was less than 0.01 (> 0.0001), the spectrum was added to the candidate site points. As a result, we obtained 5967 candidate points.

Adding noise to the synthetic data

Gaussian noise was added to the reference and transmitted X-ray counts to calculate the simulated absorbance to approximate the noise generated in actual measurements. The conditions used in the analysis when reducing the number of site points were ($c_{\mathrm{ref}},\ \mu\, t_{\mathrm{low}},\ \mu\, t_{\mathrm{high}}$) = ($2.7 \times\ 10^4$, 0.36, 1.56) which is the condition of our G-XRM experiment. Additionally, highly noisy data were generated under the conditions ($c_{\mathrm{ref}},\ \mu\, t_{\mathrm{low}},\ \mu\, t_{\mathrm{high}}$) = ($2.0 \times\ 10^4$, 0.5, 2.5).

Data availability

Minimum datasets generated and/or analyzed during the current study and the implementation of our method are available in the GitHub repository https://github.com/quantumbeam/Clustering-Method-for-Massive-Spectral-Image-Data. All datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Received: 6 May 2024; Accepted: 23 September 2024
Published online: 29 September 2024

References

1. Liu, Y. et al. Relating structure and composition with accessibility of a single catalyst particle using correlative 3-dimensional micro-spectroscopy. *Nat. Commun.*7, 12634 (2016).
2. Yu, Y. et al. Nonequilibrium pathways during Electrochemical phase transformations in single crystals revealed by dynamic Chemical Imaging at Nanoscale Resolution. *Adv. Energy Mater.* **5**, 1402040 (2015).
3. Mao, Y. et al. High-voltage charging-Induced strain, heterogeneity, and Micro-Cracks in secondary particles of a Nickel-Rich Layered Cathode Material. *Adv. Funct. Mater.* **29**, 1900247 (2019).
4. Macpherson, S. et al. Local nanoscale phase impurities are degradation sites in halide perovskites. *Nature.* **607**, 294–300 (2022).
5. Jahr, W. et al. Hyperspectral light sheet microscopy. *Nat. Commun.* **6**, 7990 (2015).
6. Amenabar, I. et al. Hyperspectral infrared nanoimaging of organic samples based on Fourier transform infrared nanospectroscopy. *Nat. Commun.* **8**, 14402 (2017).
7. Knez, D. et al. Spectral imaging at high definition and high speed in the mid-infrared. *Sci. Adv.* **8**, eade4247 (2022).
8. Ideguchi, T. et al. Coherent Raman spectro-imaging with laser frequency combs. *Nature.* **502**, 355–358 (2013).
9. Ozeki, Y. et al. High-speed molecular spectral imaging of tissue with stimulated Raman scattering. *Nat. Photonics.* **6**, 845–851 (2012).

10. Yano, T. et al. Tip-enhanced nano-raman analytical imaging of locally induced strain distribution in carbon nanotubes. *Nat. Commun.* **4**, 2592 (2013).
11. Guttmann, P. et al. Nanoscale spectroscopy with polarized X-rays by NEXAFS-TXM. *Nat. Photonics*. **6**, 25–29 (2012).
12. Yu, Y. S. et al. Three-dimensional localization of nanoscale battery reactions using soft X-ray tomography. *Nat. Commun.* **9**, 921 (2018).
13. Ade, H. & Stoll, H. Near-edge X-ray absorption fine-structure microscopy of organic and magnetic materials. *Nat. Mater.* **8**, 281–290 (2009).
14. Pfeiffer, F. X-ray ptychography. *Nat. Photonics*. **12**, 9–17 (2018).
15. Shapiro, D. A. et al. Chemical composition mapping with nanometre resolution by soft X-ray microscopy. *Nat. Photonics*. **8**, 765–769 (2014).
16. Shapiro, D. A. et al. An ultrahigh-resolution soft x-ray microscope for quantitative analysis of chemically heterogeneous nanomaterials. *Sci. Adv.* **6**, eabc4904 (2020).
17. Liu, Y. et al. Five-dimensional visualization of phase transition in BiNiO3 under high pressure. *Appl. Phys. Lett.* **104**, 043108 (2014).
18. Nelson Weker, J. et al. Operando Spectroscopic Microscopy of LiCoO2 cathodes outside standard operating potentials. *Electrochim. Acta*. **247**, 977–982 (2017).
19. Torrisi, S. B. et al. Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships. *Npj Comput. Mater.* **6**, 109 (2020).
20. Liu, Y. et al. Mapping XANES spectra on structural descriptors of copper oxide clusters using supervised machine learning. *J. Chem. Phys.* **151**, 164201 (2019).
21. Timoshenko, J. et al. Neural Network Approach for characterizing structural transformations by X-Ray absorption fine structure spectroscopy. *Phys. Rev. Lett.* **120**, 225502 (2018).
22. Boesenberg, U. et al. Mesoscale phase distribution in single particles of LiFePO $_4$ following Lithium Deintercalation. *Chem. Mater.* **25**, 1664–1672 (2013).
23. Zhang, K. et al. Finding a needle in the Haystack: identification of functionally important minority phases in an operating battery. *Nano Lett.* **17**, 7782–7788 (2017).
24. Zhang, F., Du, B., Zhang, L. & Zhang, L. Hierarchical feature learning with dropout k-means for hyperspectral image classification. *Neurocomputing*. **187**, 75–82 (2016).
25. Xia, S. et al. A fast adaptive k-means with no bounds. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 87–99 (2020).
26. Nalepa, J. et al. Unsupervised segmentation of hyperspectral images using 3-D Convolutional Autoencoders. *IEEE Geosci. Remote Sens. Lett.* **17**, 1948–1952 (2020).
27. Zhao, W. & Du, S. Spectral–spatial feature extraction for hyperspectral image classification: a dimension reduction and Deep Learning Approach. *IEEE Trans. Geosci. Remote Sens.* **54**, 4544–4554 (2016).
28. Murtagh, F. & Contreras, P. Algorithms for hierarchical clustering: an overview. *Wiley Interdisc Rev. Data Min. Knowl. Discov*. **2**, 86–97 (2012).
29. Pelleg, D. & Moore, A. W. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proc. 17th Int. Conf. on Mach. Learn.* **1**, 727–734 (2000).
30. Andrei, R. Inertia-based indices to determine the number of clusters in K-Means: an experimental evaluation. *IEEE Access*. **12**, 11761–11773 (2024).
31. Duan, X. et al. Unsupervised Data Mining in Nanoscale X-ray Spectro-Microscopic Study of NdFeB Magnet. *Sci. Rep.* **6**, 34406 (2016).
32. Song, S., Zhou, H., Yang, Y. & Song, J. Hyperspectral anomaly detection via Convolutional Neural Network and low Rank with density-based clustering. *IEEE J. Sel. Top. Appl. Earth Obs Remote Sens.* **12**, 3637–3649 (2019).
33. Cheng, D. et al. GB-DBSCAN: a fast granular-ball based DBSCAN clustering algorithm. *Inf. Sci.* **674**, 120731 (2024).
34. Li, C. et al. Fast density peaks clustering algorithm based on improved mutual K-nearest-neighbor and sub-cluster merging. *Inf. Sci.* **647**, 119470 (2023).
35. Chen, Y. et al. BLOCK-DBSCAN: fast clustering for large scale data. *Pattern Recogn.* **109**, 107624 (2021).
36. Comaniciu, D. et al. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 5 (2002).
37. Guda, A. A. et al. Understanding X-ray absorption spectra by means of descriptors and machine learning algorithms. *Npj Comput. Mater.* **7**, 203 (2021).
38. Rifai, S. et al. The Manifold Tangent Classifier. In *Adv. Neural. Inf. Process. Syst.* (2011).
39. Wang, Y., Yao, H. & Zhao, S. Auto-encoder based dimensionality reduction. *Neurocomputing*. **184**, 232–242 (2016).
40. Cohen, U., Chung, S., Lee, D. D. & Sompolinsky, H. Separability and geometry of object manifolds in deep neural networks. *Nat. Commun.* **11**, 746 (2020).
41. Suzuki, Y., Taniai, T., Saito, K., Ushiku, Y. & Ono, K. Self-supervised learning of materials concepts from crystal structures via deep neural networks. *Mach. Learn. : Sci. Technol.* **3**, 045034 (2022).
42. Wang, Y. Z., Zhang, J. L., Liu, Z. J. & Du, C. B. Recent Advances and Research Status in Energy Conservation of Iron Ore Sintering in China. *JOM*. **69**, 2404–2411 (2017).
43. Wang, W. et al. Research progress on multiscale structural characteristics and characterization methods of iron ore sinter. *J. Iron Steel Res. Int.* **27**, 367–379 (2020).
44. Kimura, M. et al. Non-empirical identification of trigger sites in heterogeneous processes using persistent homology. *Sci. Rep.* **8**, 3553 (2018).
45. Suzuki, Y., Hino, H., Kotsugi, M. & Ono, K. Automated estimation of materials parameter from X-ray absorption and electron energy-loss spectra with similarity measures. *Npj Comput. Mater.* **5**, 39 (2019).
46. Ishii, M. et al. MDR XAFS DB. Materials Data Repository, National Institute for Materials Science (2021). https://doi.org/10.48505/nims.1447
47. Fukunaga, K. & Olsen, D. R. An algorithm for finding intrinsic dimensionality of data. *IEEE Trans. Comput.* **C–20**, 176–183 (1971).
48. Aurenhammer, F. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Comput. Surv.* **23**, 345–405 (1991).
49. Bishop, C. M. Pattern recognition and machine learning 33–38Springer New York, (2006).
50. Matsumura, M., Hoshi, M. & Kawaguchi, T. Improvement of Sinter Softening Property and Reducibility by Controlling Chemical compositions. *ISIJ Int.* **45**, 594–602 (2005).
51. Mežibrický, R., Fröhlichová, M., Findorák, R. & Goettgens, V. Ore Assimilation and secondary phases by sintering of Rich and High-Gangue Iron Ores. *Minerals*. **9**, 128 (2019).
52. P Kingma, D. & Welling, M. Auto-Encoding Variational Bayes. *Arxiv*. https://doi.org/10.48550/ARXIV.1312.6114 (2013).
53. P Kingma, D., J Rezende, D., Mohamed, S. & Welling, M. Semi-supervised learning with Deep Generative models. *Arxiv*. https://doi.org/10.48550/ARXIV.1406.5298 (2014).
54. Ueno, T. et al. Adaptive design of an X-ray magnetic circular dichroism spectroscopy experiment with gaussian process modelling. *Npj Comput. Mater.* **4**, 4 (2018).
55. Ueno, T., Ishibashi, H., Hino, H. & Ono, K. Automated stopping criterion for spectral measurements with active learning. *Npj Comput. Mater.* **7**, 139 (2021).

56. Murakami, T., Wakabayashi, H., Maruoka, D. & Kasai, E. Effect of Hydrogen Concentration in reducing gas on the changes in Mineral Phases during reduction of Iron Ore Sinter. *ISIJ Int.* **60**, 2678–2685 (2020).
57. Niwa, Y., Takeichi, Y., Watanabe, T. & Kimura, M. Development of spectromicroscopes for multiscale observation of heterogeneity in materials at photon factory, IMSS, KEK. *AIP Conf. Proc.* **2054**, 050003 (2019).
58. Katayama, M. et al. Development of a two-dimensional imaging system of X-ray absorption fine structure. *J. Synchrotron Rad.* **19**, 717–721 (2012).
59. Liu, Y. et al. TXM-Wizard: a program for advanced data collection and evaluation in full-field transmission X-ray microscopy. *J. Synchrotron Rad.* **19**, 281–287 (2012).
60. Preibisch, S., Saalfeld, S. & Tomancak, P. Globally optimal stitching of tiled 3D microscopic image acquisitions. *Bioinformatics*. **25**, 1463–1465 (2009).
61. Taylor, J. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements 73–79* (University Science Books, 1997).

Acknowledgements

This work is partly supported by the JST-Mirai Program, Grant Number JPMJMI19G1, the MEXT Program: Data Creation and Utilization-Type Material Research and Development Project (Digital Transformation Initiative Center for Magnetic Materials) Grant Number JPMXP1122715503, the MEXT as “Developing a Research Data Ecosystem for the Promotion of Data-Driven Science”, the JSPS Grant-in-Aid for Transformative Research Areas (A) 22H05109, and the JST Moonshot R&D (Grant Number JPMJMS2236). This article is partly based on results obtained from a project, JPNP14004, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). The experiment was performed under the approval of the Photon Factory Program Advisory Committee (Proposal Nos. 2022G059 and 2022S2-001). The authors thank Prof. T. Murakami of Tohoku University for providing iron ore sinter samples.

Author contributions

Y.I. led the algorithm development, conducted the experiment, analysis and prepared the manuscript. Y.T. designed the G-XRM system, contributed to experimental design and interpreted the resulting spectral image data. H.H. led the formulation of the mathematical framework of the algorithm. K.O. conceived and led the direction of the project. All authors contributed to discussion and writing the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-74016-0.

Correspondence and requests for materials should be addressed to K.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.