Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **510,** 334–350 (2022) https://doi.org/10.1093/mnras/stab3478
Advance Access publication 2021 December 3

Evaluation of the classification of pre-solar silicon carbide grains using consensus clustering with resampling methods: An assessment of the confidence of grain assignments

Grethe Hystad [1]★ Asmaa Boujibar,[2] Nan Liu,[3] Larry R. Nittler[4] and Robert M. Hazen[4]

[1]*Department of Mathematics and Statistics, Purdue University Northwest, 2200 169th Street Hammond, IN 46323-2094, USA*
[2]*Department of Geology / Department of Physics and Astronomy, Western Washington University, 516 High Street, MS-9080, Bellingham, WA 98225-9164, USA*
[3]*Department of Physics, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130, USA*
[4]*Earth and Planets Laboratory, Carnegie Institution for Science, 5251 Broad Branch Rd, NW, Washington, DC 20015, USA*

Accepted 2021 November 22. Received 2021 November 21; in original form 2021 June 9

ABSTRACT

We report the use of several cluster analysis techniques to evaluate the classification of pre-solar silicon carbide (SiC) grains. The stability of clusters and the confidence of individual cluster assignments of grains are assessed using consensus clustering with resampling methods. Our analysis shows that pre-solar SiC grains can be divided into seven groups that are found to be highly stable with most of the grains being assigned to the same cluster for at least 90 per cent of the time over multiple aggregated clustering. Among the seven groups, two groups are dominated by AB grains, three groups by MS grains, one group by Z grains, and one group by X grains. The further division of X grains into two groups is highly dependent on the chosen algorithm and is therefore uncertain. Z and Y grains are clustered jointly with MS grains, with one group dominated by Z grains, pointing to their common origins from low-mass asymptotic giant branch stars. The most stable N grain-containing clusters are dominated by ^{15}N-rich AB grains. However, some methods assign N grains with X grains, but in less stable clusters. The suggested genetic relationship among ^{15}N-rich AB, N, and X grains is in line with the recent proposal that all three types of pre-solar SiC grains came from core collapse supernovae. We discuss the results from different clustering techniques based on our assessment of the cluster stabilities and the extent to which the cluster assignments overlap across the different methods.

Key words: nuclear reactions, nucleosynthesis, abundances – methods: statistical.

1 INTRODUCTION

The discovery of microscopic pre-solar grains in meteorites in the mid-1980s has led to a deepened understanding of the physical environment and nucleosynthetic processes occurring in stars. These pre-solar grains are characterized by isotopic compositions that deviate from Solar system compositions by up to several orders of magnitude, thus pointing to their extrasolar stellar origin. Pre-solar grains are divided into different groups based on these anomalous isotopic compositions, and the detailed division scheme has been continuously evolving because of advances in isotope analysis techniques and astrophysical models (for reviews, see Zinner 2014; Nittler & Ciesla 2016). Multi-element isotope data for the best-studied type of pre-solar grains, silicon carbide (SiC), have been used to define numerous sub-groups (e.g. mainstream – MS, AB, X, Y, Z, etc.) which have been proposed to originate from various stellar sources (Zinner 2014). In a recent paper by Boujibar et al. (2021), cluster analysis was employed to evaluate the classification of SiC grains based on an updated pre-solar SiC grain data base (PGD; Stephan et al. 2020) that was initially compiled about 10 yr ago (Hynes & Gyngard 2009). Cluster analysis is the umbrella term for a collection of techniques in statistics and data science used to divide data into groups or clusters, where data in the same cluster possess similarities in their attributes. Using a mixture of normal distributions in concert with the Bayesian Information Criteria (BIC), Boujibar et al. (2021) identified nine groups in their cluster analysis based on the SiC grains' C, N, and Si isotope compositions and inferred initial ^{26}Al/^{27}Al ratios compiled in the PGD. The authors discussed the astrophysical implications of their cluster analysis results in comparison to the original classification scheme for pre-solar SiC. This clustering analysis enabled identification of four clusters of SiC grains likely formed in AGB stars, with different ranges of metallicities, with combinations of MS, Y, and Z grains. These clusters include a compact cluster of MS grains with a narrow range of isotopic compositions. This analysis also pointed to two pairs of clusters of AB and X grains, and another cluster containing putative nova grains (N grains) and AB grains.

The robustness and stability of the nine identified clusters, however, were not further evaluated by Boujibar et al. (2021). A common problem in cluster analysis is that different methods often detect different numbers of optimal clusters and yield different cluster classifications for a data set. Additionally, repeatedly running the same clustering algorithm may lead to a different classification of the data because of the partitioning variability. As a result, two data points that are initially assigned to the same cluster may be assigned

★ E-mail: ghystad@pnw.edu

to two different clusters if the cluster algorithm is rerun, pointing to instabilities in the defined clusters. Variations in the clustering results may arise from both noise in the data and the suitability of the clustering algorithm to a particular data set (Henelius et al. 2016). Perturbations to the data, subtraction or addition of some data points for example, may also lead to different clusters. Ideally, a different random sample taken from the same population or the use of a sub-sample of the original data should provide the same assignment to a cluster. However, this optimal situation is often not the case (James et al. 2013), and the robustness and stability of the cluster results of Boujibar et al. (2021) should thus be evaluated.

In this study, we extend the results of Boujibar et al. (2021) by evaluating the stability of the clusters and assessing the confidence of the classification of the pre-solar SiC grains. Compared to the study of Boujibar et al. (2021), we utilized additional clustering methods. Also, we identified grains that constitute the stable parts of clusters and compared the results across different techniques. The goal of this paper is to provide further insight into the classification of pre-solar SiC grains by employing state-of-the-art clustering methods. The stabilities of the clusters are evaluated using consensus clustering, which is a method that combines the clustering results from multiple clustering algorithms or reruns of the same algorithm into a single consensus clustering. A cluster ensemble represents a collection of multiple clustering of a data set that is used to create a single consensus clustering (Strehl & Ghosh 2002; Monti et al. 2003; Fred & Jain 2005; Ghosh & Acharya 2011; Wang, Shan & Banerjee 2011). Consensus clustering methods arose from the machine-learning literature around two decades ago and have rapidly gained interest with a burgeoning literature in the last 10 yr.

The paper is organized as follows. The clustering methods are described in Section 2. The clustering results using a variety of clustering techniques are presented in Section 3. In Section 4, we discuss the stabilities of the clusters by the different clustering methods and their astrophysical implications. We also compare our partitioning results to the classification given in Boujibar et al. (2021) in this section. The conclusions are summarized in Section 5. The R-code created for the different consensus clustering techniques used in this paper is posted on GitHub.[1]

2 METHODS AND CLUSTERING

2.1 Pre-solar SiC isotope data

In this paper, we used a sample of 1478 observations for four measured isotope ratios, $^{12}C/^{13}C$, $^{14}N/^{15}N$, $\delta(^{29}Si/^{28}Si)$, and $\delta(^{30}Si/^{28}Si)$, from the data base $PGD\_SiC\_2020-01-30$ (Stephan et al. 2020) to further evaluate the classification of pre-solar SiC grains. Each observation also contains information about previously assigned grain subtype (MS, X, AB, N, Z, Y). The rare C and U grains and 25 observations of contaminated grains are excluded from the analysis as in the study of Boujibar et al. (2021). We calculated the $^{i}Si/^{28}Si$ ratios based on the $\delta(^{i}Si/^{28}Si)$ values compiled in the PGD, with $\delta(^{i}Si/^{28}Si)$ defined as

$$\delta\left(^{i}\mathrm{Si}/^{28}\mathrm{Si}\right) = \left(\frac{^{i}\mathrm{Si}/^{28}\mathrm{Si}}{\left(^{i}\mathrm{Si}/^{28}\mathrm{Si}\right)_{\mathrm{solar}}} - 1\right) \times 1000,$$

in which i denotes 29 or 30 and $(^{i}Si/^{28}Si)_{solar}$ the Solar ratio. The isotopic ratios for the 1478 observations of SiC grains were transformed into logarithmic units and then scaled to a mean of zero and unit variance. These isotopic ratios were used in the cluster analysis. The R-library DPLYR (Wickham et al. 2021) was used for data manipulation. Fig. 1 shows a plot of the nitrogen and carbon isotopic compositions of the 1478 pre-solar SiC grains from the updated PGD (Stephan et al. 2020).

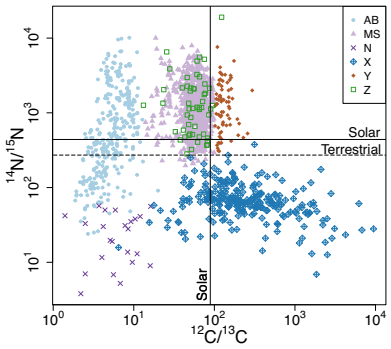

Figure 1. Plot of nitrogen and carbon isotopic compositions of 1478 pre-solar SiC grains from the updated PGD (Stephan et al. 2020).

2.2 Clustering techniques

Several model-based cluster analysis techniques were compared by means of BIC and the Integrated Completed Likelihood (ICL) for model selection and the detection of the optimal number of clusters. Model-based clustering analysis is based on mixture models which are probability distributions that represent hidden sub-populations without observed information for which sub-population each data point belongs to. For finite mixture models, BIC is defined as $BIC = 2\log p(y \mid \hat{\theta}, M) - k\log(n)$, were $p(y \mid \hat{\theta}, M)$ is the maximum of the mixture likelihood for the data y and model M, $\hat{\theta}$ is the maximum-likelihood estimate for the parameter vector θ, k is the number of model parameters to be estimated, and n is the number of observations. BIC prevents overfitting by including a penalty term for the number of parameters in the model. ICL is defined as $ICL = BIC - E(M)$, where $E(M)$ is the expected entropy of the clustering from model M. Entropy measures the uncertainty of the clustering and is high when the uncertainty is large, while zero if there is no uncertainty. For model-based clustering, the model with the largest BIC or ICL is selected. The BIC is devised to select the number of components in the mixture model, which may be different than the number of clusters if the clusters are strongly non-Gaussian. The ICL is often a better measure for the detection of the optimal number of clusters, given that the ICL tends to favour more clearly separated clusters. ICL selects a model with the same number or smaller number of clusters than BIC. We refer the reader to Bouveyron et al. (2019, p. 51–55) for a detailed description of BIC and ICL for model selection. For an example illustrating how a non-Gaussian cluster is represented by a mixture of Gaussian mixture components using BIC while only one Gaussian component, using ICL see Bouveyron et al. (2019, p. 99). We focused on the following model-based clustering methods: mixture of normal distributions, mixture of normal distributions with a uniform component to represent outliers or noise, and mixture of t-distributions. We also modelled a mixture of restricted skew normal distributions and a mixture of generalized hyperbolic distributions. For these latter two models, although the model results both seem to be robust in terms of BIC and ICL, we found that the number of clusters detected based on ICL was lower than the number of grain types in the original classification and is

[1] https://github.com/ghystad/Consensus-clustering.presolar.SiC.grains/tree/master

therefore considered as too low. We therefore did not perform further analysis using these models.

The mixture of normal distributions for modelling noise adds a uniform component to represent the outliers. The mixture of t-distributions models elliptical clusters but with heavier tails than the mixture of normal distributions. We refer the reader to McLachlan & Peel (2000) and Bouveyron et al. (2019) for details regarding these models and methods. The statistical software package R (R Core Team 2016) has libraries for all of these models.

Finally, we used spectral clustering, which takes the form of a graph partitioning problem. The algorithm is based on transforming the observations into eigenvectors computed from the Laplacian obtained by a similarity graph. The matrix of eigenvectors corresponding to the k lowest eigenvalues is then used as an input to the k-means clustering algorithm. The k-means clustering algorithm then assigns the observations to the clusters for which the Euclidean distance from the clusters centroid to the observations is the shortest (James et al. 2013). We used the k-nearest neighbours of the observations to obtain the similarity graph accompanied by the random walk normalized Laplacian (von Luxburg 2007). The R-library KKNN (Schliep & Hechenbichler 2016) was used to perform the spectral clustering. The spectral clustering algorithm assumes no particular form of the clusters and is able to detect non-convex clusters and clusters formed as intertwined spirals (von Luxburg 2007).

We created a k-nearest neighbour graph by finding the k nearest neighbours for each observation based on the Euclidean distances between data points. The resulting adjacency matrix was converted into an undirected graph using the IGRAPH R library (Csárdi & Nepusz 2006). We then computed the number of connected components in the graph as a function of k. We found that the graph is connected for $k > 4$ with a right-skewed degree distribution. We used spectral clustering with a random walk normalized Laplacian for different values of $k > 4$. We found the algorithm to be unstable for small values of k. We decided to use $k = 130$ in the spectral algorithm by noticing that values of $k > 115$ gave similar results. There is no established theory for the relationship between the number of clusters and the number of nearest neighbours used to create the similarity graph (von Luxburg 2007). According to von Luxburg (2007), a normalized Laplacian should be employed if the vertices have different degrees, for which a random walk Laplacian is recommended. The paper also suggested to use a connected similarity graph for the spectral algorithm unless one is certain that the connected components in the graph correspond to the correct clusters (von Luxburg 2007).

2.3 Reference labels

Running the clustering algorithm on resampled data will often produce different labelling of the clusters; a problem referred to as the label correspondence problem. The Hungarian algorithm (Kuhn 1955) used in this paper is a standard way to solve the cluster label correspondence problem by aligning the cluster labels across different runs of the data to a reference labelling (Dudoit & Fridlyand 2003). We applied the Hungarian algorithm to keep a consistent labelling of the clusters across different clustering algorithms, thereby providing results that can be directly compared. Let A and B be among the clustering methods used in this paper. We created an assignment matrix for aligning the clustering labels from clustering method A with the labels from clustering method B. The Hungarian algorithm with the R-function HungarianSolver in the R-library RCPPHUNGARIAN(Silverman 2019) was then used to find an optimal assignment of rows to columns, where the cluster labels from method A were mapped in a one-to-one correspondence to the cluster labels from method B for the same number of clusters. In order to map the cluster labels from seven to nine clusters, the HungarianSolver function was again used. Now the cluster labels from method A were mapped on to the cluster labels from method B, as the latter method yielded a larger number of clusters. These mappings allowed us to have matching colours for the clusters across different techniques and different numbers of clusters. We finally relabelled the clusters for nine clusters such that the cluster labels from seven clusters could be in order from one to seven. The labels obtained from clustering with a mixture of normal distributions with seven clusters were used as the overall reference labels.

2.4 Consensus clustering methods

We use consensus clustering with resampling techniques to identify the stable clusters of pre-solar SiC grains and the stable parts of the clusters. Cluster ensembles that are formed from different resampling methods improve the precision of the clustering method by using averaging to decrease the variance of the clustering results. It also provides a way to assess the stability of the clusters and to evaluate the confidence of the classification of the observations (Dudoit & Fridlyand 2003).

In this paper, we focus on three different consensus clustering methods that are based on resampling. The first two, titled BagClust1 and BagClust2, were introduced by Dudoit & Fridlyand (2003) 'to improve the accuracy of a clustering procedure.' BagClust1 and BagClust2, use a technique called bagging, which aggregates multiple clustering based on bootstrapping. Bootstrapping (Efron & Tibshirani 1993) refers to methods that use samples obtained by taking random samples with replacement from the original sample. These samples have the same sample size as the original sample. However, in this study we deleted duplicated observations such that the number of observations in each sample was on average 63.2 per cent of the original sample size (Efron & Tibshirani 1997). BagClust1 produces a final partitioning of the data based on a majority vote across bootstrap samples. BagClust2 produces a dissimilarity matrix-based on the proportion of co-occurrences in clusters of each pairwise observation across bootstrap samples, which is subsequently used in a new clustering algorithm. By deleting duplicates, we used a modification of these methods in this paper. The third method, which was introduced by Henelius et al. (2016), uses a form of clustering aggregation to identify core clusters. A core cluster is the largest set of observations inside a cluster that co-occur in the same cluster with a probability of at least $1 - \alpha$ for a significance level of $0 < \alpha < 1$ when the clustering algorithm is run on resampled data. Finally, the individual cluster stabilities are also evaluated by computing the distribution of the Jaccard coefficient as a similarity measure for each original cluster compared to the most similar cluster in the resampled data (Hennig 2007).

2.4.1 BagClust1

Bagging uses averaging over aggregated clustering to decrease the variability of the clustering algorithm (Dudoit & Fridlyand 2003). In the BagClust1 procedure (illustrated in an example in Fig. 2), the original cluster analysis algorithm was applied to the data in order to obtain cluster labels for each observation using a fixed number of clusters. We refer to this partitioning as the reference clustering. From the original data, $B = 500$ bootstrap samples were selected, but with the duplicate observations deleted. For each sub-sample, the

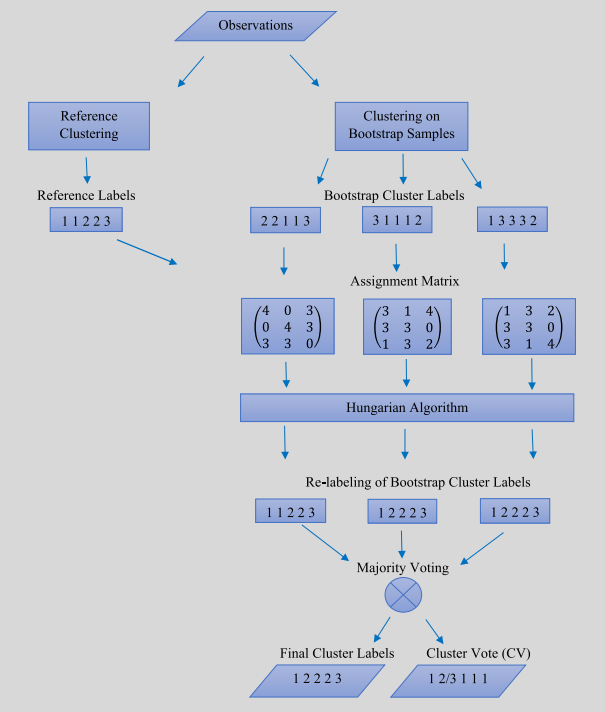

Figure 2. Illustration of the BagClust1 procedure by Dudoit & Fridlyand (2003) and the alignment of bootstrap cluster labels to the reference labels for a sample of five observations and three clusters. For example, in the first assignment matrix, the $(1, 2)^{th}$ entry is 0 because two observations are occurring in bootstrap cluster 1, two observations are occurring in reference cluster 2, and these clusters also have two observations in common (i.e. $(2-2) + (2-2) = 0$). The end product is the final cluster partition found by majority votes as well as the cluster votes for the individual observations. For example, observation number 2 is assigned by majority vote to cluster 2 with a cluster vote of 2/3 since it occurs in cluster 2 for two of the three bootstrap samples.

cluster algorithm was run, and the clustering labels were obtained for each observation in the sub-sample. In order to obtain similar cluster labels for each run, an assignment matrix was calculated. The $(i, j)^{th}$ entry of this matrix is the number of elements that are in cluster i formed from the sub-sample and not in reference cluster j plus the number of elements that are in reference cluster j and not in cluster i, i.e. the symmetric difference of cluster i and cluster j. The Hungarian algorithm with the R-function HungarianSolver in the R-library RCPPHUNGARIAN (Silverman 2019) was then used to maximize the overlap between the reference clusters and the clusters formed from the sub-sample. The labels for each cluster calculated from the sub-sample were subsequently mapped in a one-to-one correspondence to the labels of the reference clusters. Each observation was finally assigned to the cluster for which it occurred most frequently over the B subsamples (majority vote). If there was a tie in the frequency, we randomly assigned the observation to the most frequent clusters. In addition, a cluster vote (CV) was calculated for each observation, which is the proportion of times the observation was assigned to its winning cluster, accounting for the number of times the observation occurred in the B sub-samples. The R-code for the BagClust1 procedure is posted on GitHub.[2] The code for the cluster label correspondence problem included in the BagClust1 procedure was based on the code of Rösler (2012) with some slight modifications.

2.4.2 *BagClust2*

The BagClust2 procedure (Dudoit & Fridlyand 2003) avoids the cluster label correspondence problem by constructing a connectivity matrix, M, from B bootstrap samples, where $B = 500$ was used in this paper, but with duplicates deleted. If n is the number of observations, the connectivity matrix is an n-by-n matrix that is formed by recording for each pair of observations the number of times they were found together in the clusters calculated from the B sub-samples. In addition, an indicator matrix IN was calculated to keep track of the number of times each pair of observations co-occurred in the B sub-samples. A dissimilarity matrix was then formed from $D = I - M/IN$, where I is the identity matrix. The dissimilarity matrix D was subsequently used as an input to the Partitioning Around Medoids (PAM) clustering algorithm. The PAM algorithm is based on minimizing the pairwise distance between the observations in the cluster and their medoid (center) of the cluster. We used the R-library CLUSTER (Maechler et al. 2021) for the PAM cluster algorithm. The internal stability of the clusters can then be evaluated by computing the silhouette width of each cluster as well as the overall average silhouette width of all the clusters. The silhouette width is a number between -1 and 1, for which a number closer to 1 indicates good classification of the data. Kaufman & Rousseeuw (1990) specified that a silhouette width above 0.5 is a reasonable classification of the data, while a number below 0.2 points to a lack of cluster structure (Everitt et al. 2011). A negative silhouette width signifies that the cluster assignment is incorrect, while a value of zero indicates that the observation is located in between two clusters, pointing to an uncertain assignment. The average silhouette width can also be used to detect the number of clusters (Everitt et al. 2011). The method for creating the consensus matrix M/IN is illustrated in an example in Fig. 3. The R-code for the BagClust2 procedure is posted on GitHub.[3]

2.4.3 *Core cluster identification*

The consensus matrix M/IN, described in Section 2.4.2, is also used to identify the core clusters by creating an undirected graph with the observations as the nodes and an edge between two nodes if the co-occurrence probability is at least $1 - \alpha$ for a significance level $0 < \alpha < 1$. We used the R-library IGRAPH (Csárdi & Nepusz 2006) to convert the consensus matrix to an adjacency matrix, where the $(i, j)^{th}$ entry is one if the co-occurrence probability is at least $1 - \alpha$ and zero otherwise. This matrix was then subsequently converted to an undirected graph, where two observations are connected by an edge if their $(i, j)^{th}$ (and $(j, i)^{th}$ by symmetry) entry is one. The core clusters are found as the largest maximal clique for each cluster in the graph (Henelius et al. 2016). A clique is a sub-graph that consists of vertices that are all connected to each other via edges, while a maximal clique is 'not a sub-set of a larger clique' (Kolaczyk & Csárdi 2014). Thus, the observations in a core cluster co-occur in the same cluster with a probability of at least $1 - \alpha$. Some clusters had a few largest maximal clique and we then selected one of them as they only differed in one or a few observations (with the exception of cluster 7 for the mixture of normal distributions of nine clusters

[2]https://github.com/ghystad/Consensus-clustering.presolar.SiC.grains/blob/master/BagClust1_with_spectral_algorithm.R

[3]https://github.com/ghystad/Consensus-clustering.presolar.SiC.grains/blob/master/BagClust2_with_spectral_algorithm.R

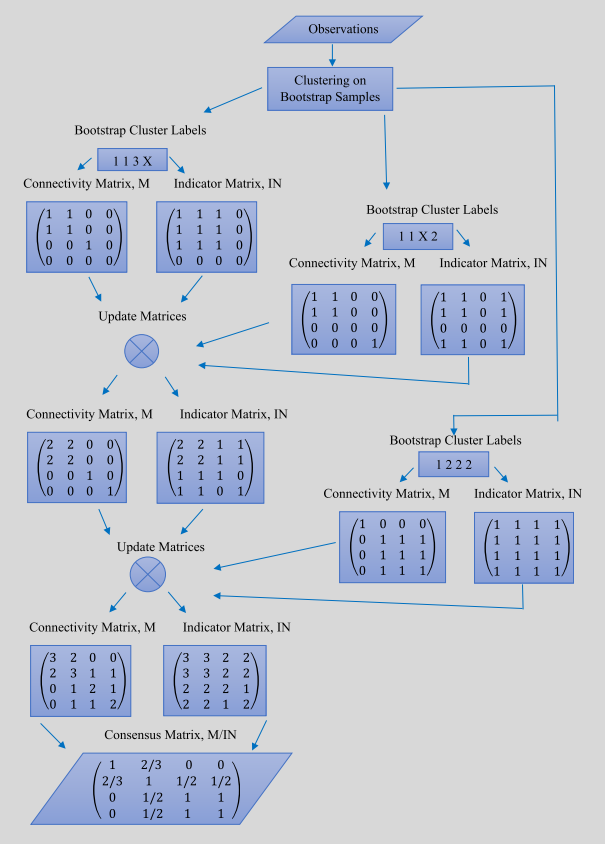

Figure 3. Illustration of the creation of a consensus matrix *M*/*IN* for four observations and three bootstrap samples. The *X* in the bootstrap sample indicates that the observation was not included in the bootstrap sample. For example the $(1, 2)^{th}$ and $(2, 1)^{th}$ entries of the consensus matrix *M*/*IN* is 2/3 since observations number 1 and 2 occur in the same cluster for two of the three bootstrap samples. The consensus matrix is used in both the BagClust2 procedure by Dudoit & Fridlyand (2003) and the core cluster identification method by Henelius et al. (2016).

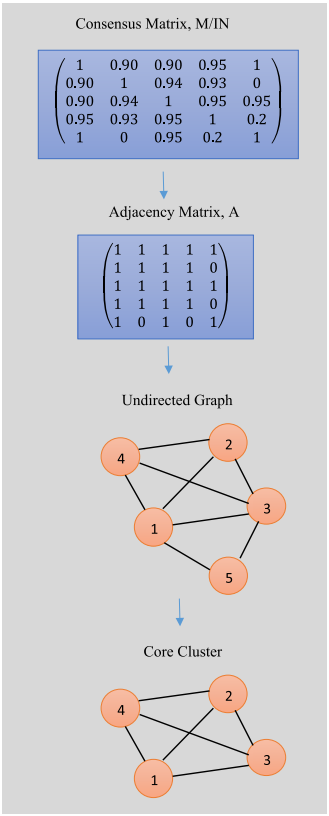

Figure 4. Illustration of the creation of a core cluster (Henelius et al. 2016) from a consensus matrix *M*/*IN*. The consensus matrix is here assumed created from one cluster where the $(i, j)^{th}$ entry (and $(j, i)^{th}$ entry by symmetry) is the proportion of times the i^{th} and j^{th} observations occur in the same cluster. The $(i, j)^{th}$ entry of the adjacency matrix, *A*, is 1 if the $(i, j)^{th}$ entry of the consensus matrix *M*/*IN* is at least 0.90, and 0 otherwise. There is a link between observations *i* and *j* in the undirected graph if the $(i, j)^{th}$ entry of the matrix *A* is 1. We omitted all the loops in the graph. In this figure, there is no link between observation 5 and observations 2 and 4, and hence, observation 5 is not part of the core cluster. In the core cluster, all observations are linked with each other. This means that these observations occur in the same cluster for at least 90 per cent of the times.

for which two maximal cliques differed by 10 and 11 observations.) Fig. 4 shows an example for the creation of a core cluster from a consensus matrix using $1 - \alpha = 0.90$. There is also an R-library titled CORECLUSTER for identifying the core clusters. However, here we created our own code for this, which is posted on GitHub.[4]

2.4.4 *Jaccard similarity coefficient*

Finally, we evaluated the stability of the individual clusters by computing the average Jaccard (similarity) coefficient for each cluster obtained from the BagClust1 procedure and the most similar cluster in the sub-sample calculated over $B = 100$ bootstrap samples with duplicates deleted (Hennig 2007). The Jaccard coefficient for two clusters *i* and *j* is defined as

$$J = \frac{c}{a + b + c},$$

where *a* is the number of elements that are occurring in cluster *i* and not cluster *j*, *b* is the number of elements that are occurring in cluster *j* and not cluster *i*, and *c* is the number of elements that the two clusters have in common. In other words, the denominator in the Jaccard coefficient is the union of the elements in cluster *i* and cluster *j*. The Jaccard coefficient is a number between 0 and 1, for which a number closer to one indicates that the two clusters are similar. There is also an R-library titled FPC (Hennig 2020) for evaluating the clusterwise stabilities using the Jaccard coefficient as a similarity measure. However, here we created our own code for this, which is posted on GitHub.[5]

2.5 Stability of clusters

For a fixed number of clusters, the stability of the clusters were evaluated and compared across different clustering algorithms. In addition, we were able to identify the observations that tend to reoccur in the same cluster when we ran the clustering algorithm repeatedly, thereby providing a confidence of the cluster assignment. We identified the stable part of a cluster to consist of the observations

[4]https://github.com/ghystad/Consensus-clustering.presolar.SiC.grains/blob/master/Corecluster_identification_with_spectral_clustering.R

[5]https://github.com/ghystad/Consensus-clustering.presolar.SiC.grains/blob/master/Jaccard_similarity_coefficient_with_spectral_algorithm.R

of pre-solar SiC grains that have a CV of at least 0.90 computed using the BagClust1 procedure as described in Section 2.4.1. These observations occurred in its winning cluster assignment more than 90 per cent of the times when running the clustering algorithm over the 500 sub-samples obtained in the bootstrap scheme. We defined a cluster as stable if the average CV of its observations was at least 0.90. Subsequently, we computed the Jaccard coefficient described in Section 2.4.4 for the clustering results obtained from the BagClust1 procedure as another measure for the stability of a cluster. An average Jaccard similarity coefficient of at least 0.85 characterizes a highly stable cluster (Hennig 2020), whereas a value below 0.6 indicates an unstable cluster. We also identified the pre-solar SiC grains that occur in core clusters as the observations that co-occur in the same cluster with a probability of at least 0.90, i.e. for at least 90 per cent of the 500 sub-samples using the methods described in Section 2.4.3. Observations outside core clusters were labelled weak points (Henelius et al. 2016). The average silhouette width computed from the BagClust2 procedure as described in Section 2.4.2 was used to compare the internal stability of the clusters.

3 RESULTS

3.1 Clustering results

Employing the R-library MCLUST (Scrucca et al. 2016), the best model selected for a mixture of normal distributions was a variable volume, shape, and orientation (VVV) model with nine optimal clusters detected by BIC (−5152.67) and seven optimal clusters detected by ILC (−5575.51). Using the R-library TEIGEN (Andrews & McNicholas 2012; Andrews et al. 2018), the best model selected for a mixture of t-distributions was unconstrained volume, orientation, and shape for the scale matrix, and constrained degrees of freedom (UUUC) with nine clusters selected by BIC (−5161.97) and seven clusters selected by ICL (−5530.08) with all model parameters unconstrained (UUUU). The mixture of t-distributions with seven clusters provided the better model if ICL was used as the selection criteria. On the other hand, if the BIC was used for model selection, the mixture of normal distributions with nine clusters should be selected. Seven clusters plus a noise component were detected when we fitted a mixture of normal distributions with a uniform component for outliers using the nearest neighbour cleaning method in the PRABCLUS R library (Hennig & Hausdorf 2020). Applying a modification of the nearest neighbour cleaning method in the COVROBUST R library (Wang & Raftery 2017), six clusters plus a noise component were detected. Both the BIC and ICL were here lower than for the other models, which led us to instead focus on the mixture of normal distributions with no noise and the mixture of t-distributions. One interesting observation found by modelling the data with a mixture of normal distributions and a noise component is that about half of N grains and some of the X grains were identified as noise; the rest of the X grains were mostly found in one cluster.

For the spectral algorithm, it is not completely clear what the optimal number of clusters is when using the eigengap heuristic as described in von Luxburg (2007) to compute the number of clusters. The number of clusters seems to be either three or four, but the result is ambiguous. As for most methods employed to detect the optimal number of clusters, the eigengap heuristic gives ambiguous results if the clusters are not well separated. We also compared the average silhouette width for different number of clusters computed using BagClust2, which again indicates that the optimal number of clusters is three. However, by comparing to the six groups that were originally proposed for classifying SiC grains, three or four clusters are too low. If we ignore such a small number of clusters, six and seven clusters gave the highest average silhouette widths, where only a slightly higher value was observed for six clusters as compared to seven clusters. Since the model-based methods indicated that seven is the optimal cluster number, we chose seven clusters for the spectral algorithm.

3.2 Illustration of the clustering results

The clustering results for the 1478 pre-solar SiC grains are illustrated in Fig. 5 for the spectral clustering using seven clusters, Fig. 6 for a mixture of *t*-distributions using seven clusters, Fig. 7 for a mixture of normal-distributions using seven clusters, and Fig. 8 for a mixture of normal-distributions using nine clusters. For all figures, the results from the BagClust1 procedure are shown in panels a–c, where panel a shows the clusters in the plot of $^{14}N/^{15}N$ versus $^{12}C/^{13}C$, panel b shows the clusters in the plot of $\delta(^{29}Si/^{28}Si)$ versus $\delta(^{30}Si/^{28}Si)$, and panel c shows a barplot of grain types versus clusters. Panels d–f, and panels g–i show the same set of plots, but for the stable clusters and the core clusters, respectively. The results from the BagClust2 procedure are shown in the panels j–l of Figs 6–8. The BagClust2 clustering results are not shown for the spectral clustering algorithm since they were almost identical to the BagClust1 clustering results. Fig. 9 is a zoom-in of panels a–b of Figs 5–8. The R-library GGPLOT (Wickham 2016) was used for creating the barplots.

Table 1 summarizes for each considered clustering technique, the proportion of pre-solar SiC grains with a CV of at least 0.90, the average Jaccard similarity coefficient, the average silhouette width, the proportion of observations with silhouette width greater than 0.5, and the proportion of observations occurring in core clusters. According to Table 1, spectral clustering yields clusters with the largest proportion of grains with CV values greater than or equal to 0.90, the largest average Jaccard coefficient, the largest average silhouette width, the largest proportion of grains with silhouette width greater than 0.50, and the largest proportion of grains in core clusters. The mixture of t-distributions yields larger values of the above quantities compared to clustering with a mixture of normal distributions, but lower compared to spectral clustering. Table 2 provides for each cluster and for each of the considered clustering technique, the proportion of pre-solar SiC grains with CV values greater than or equal to 0.9, the average Jaccard coefficient, the average silhouette width, and the proportion of grains in core clusters. Table 3 provides for each of the considered clustering technique the proportion of pre-solar SiC grain types with CV values greater than or equal to 0.90 and the proportion of grain types in core clusters.

4 DISCUSSION

4.1 Spectral clustering with seven clusters

The BagClust1 and BagClust2 clustering procedures produce similar results when using spectral clustering with seven clusters. The pre-solar SiC grains are partitioned into two main groups of AB grains (clusters 5 and 6), three main groups of MS grains (clusters 3, 4, and 7), one main group of Z grains (cluster 2), and one main group of X grains (cluster 1; Fig. 5, panels a–c). All N grains, except for one, are included with AB grains in cluster 5. Y and Z grains are spread over several clusters, for which all of the Y grains and about 60 per cent of the Z grains occur jointly with MS grains. In addition, cluster 2

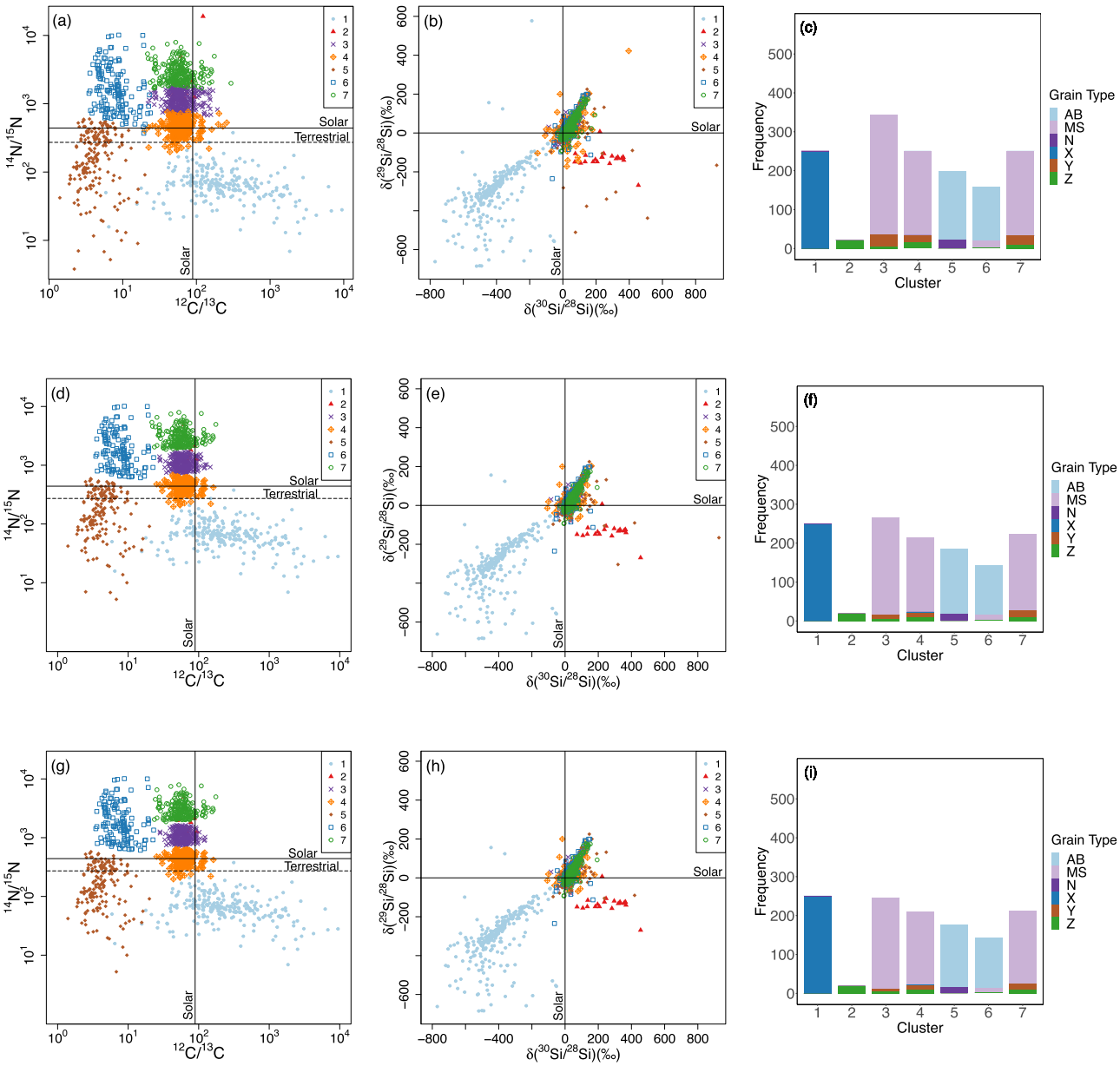

Figure 5. Spectral clustering of pre-solar SiC grains using seven clusters. The results from the BagClust1 procedure are shown in panels (a–c), where the clusters are shown in panels a–b and a barplot of grain types versus clusters is shown in panel c. The stable clusters are depicted in panels d–e, with a barplot of grain types versus the stable clusters shown in panel f. The core clusters are shown in panels g–h, with a barplot of grain types versus the core clusters shown in panel i.

is dominated by Z grains. The clusters are highly stable with large average Jaccard coefficients and with the majority of the observations having *CV* values of at least 0.90. The stable clusters (Fig. 5, panels d–f) and the core clusters (Fig. 5, panels g–i) are very similar, but there are a few observations that are not in the intersection of the core clusters and the stable parts of the clusters. Each of the clusters also has large average silhouette width. All X grains and the majority of AB, MS, N, and Z grains have CV values of at least 0.9 and occur in core clusters. Y grains are the least stable grain types for which just above half have CV of at least 0.9 and more than half of them are classified as weak points (Table 3).

Consistent with the results of Boujibar et al. (2021), our spectral clustering results point out that the original division of pre-solar SiC grains into three major groups, MS, AB and X, is quite robust, but the original classification of the additional minor groups, N, Y, Z, is questionable. Our results here, however, differ from Boujibar et al. (2021) in detail as summarized below. (1) MS grains are divided into three clusters (3, 4, 7) mainly based on the $^{14}N/^{15}N$ ratio here, in contrast to the three clusters of MS grains identified by Boujibar et al. (2021) that had different ranges of Si and C isotope ratios. (2) X grains identified here (cluster 1) are consistent with their original definition in the literature (Fig. 1), in contrast to the two clusters of X grains reported by Boujibar et al. (2021). (3) AB grains are divided into two groups with the divider lying around the solar $^{14}N/^{15}N$ value, which is slightly different from the detailed division given by Boujibar et al. (2021) but agrees with the classification scheme recommended by Liu et al. (2016, 2017a,b, 2018) based on the isotope data of a larger number of elements (C, N, Si, Ti,

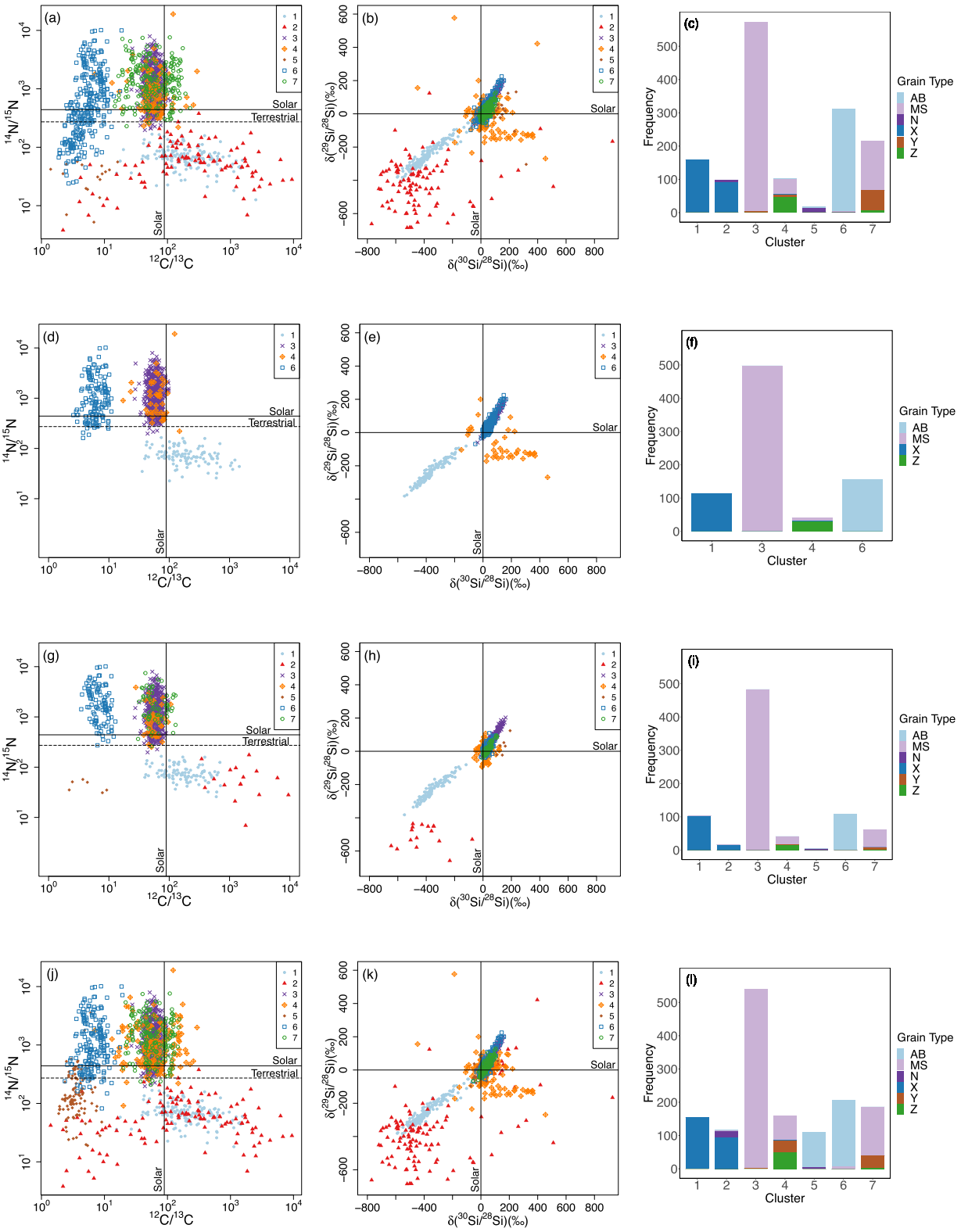

Figure 6. Clustering of pre-solar SiC grains with a mixture of t-distributions using seven clusters. The results from the BagClust1 procedure are shown in panels a–c, where the clusters are shown in panels a–b and a barplot of grain types versus clusters is shown in panel c. The stable clusters are depicted in panels d–e, with a barplot of grain types versus the stable clusters shown in panel f. The core clusters are shown in panels g–h, with a barplot of grain types versus the core clusters shown in panel i. The results from the BagClust2 procedure are shown in panels j–l, where the clusters are shown in panels j–k and a barplot of grain types versus clusters is shown in panel l.

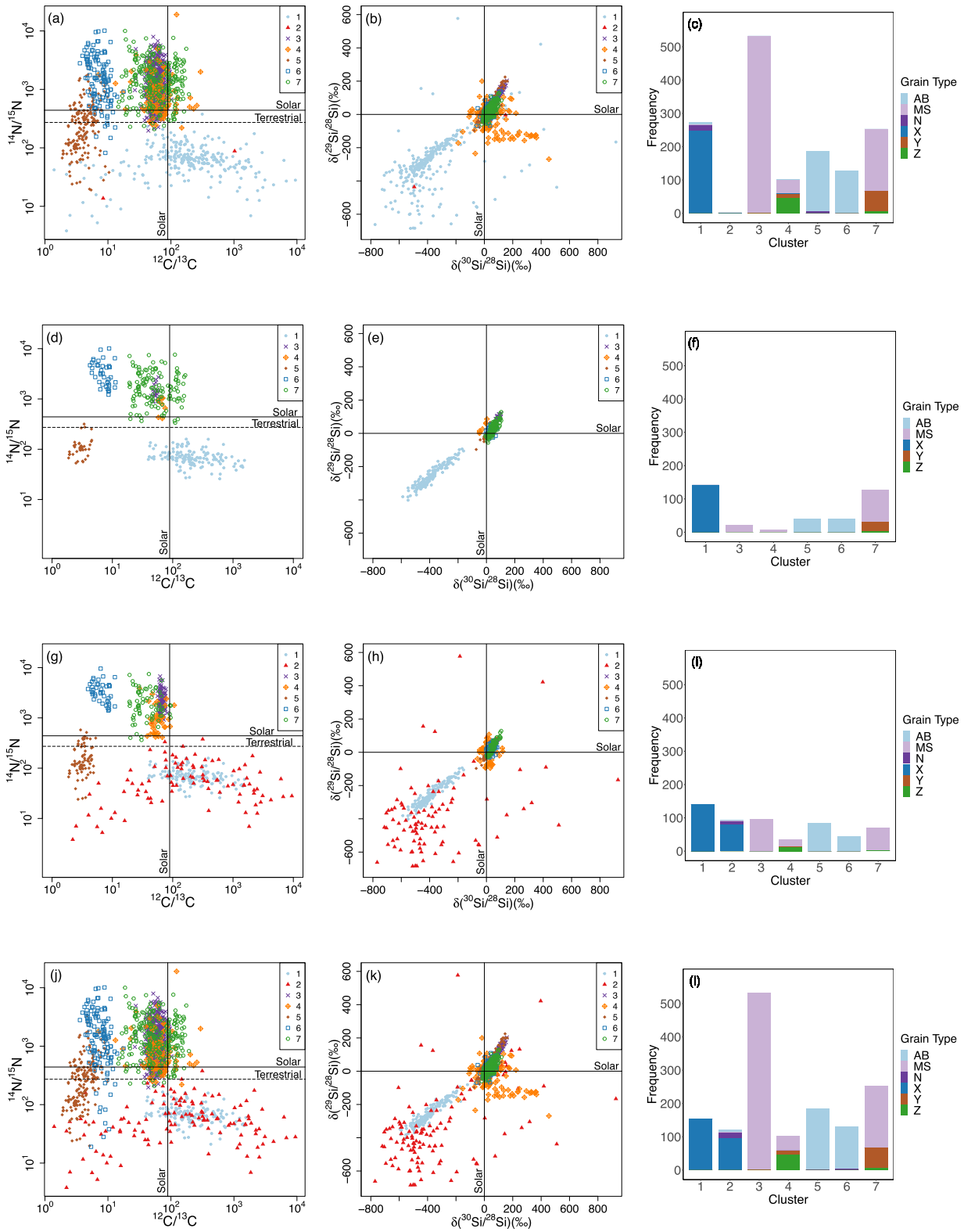

Figure 7. Clustering of pre-solar SiC grains with a mixture of normal distributions using seven clusters. The results from the BagClust1 procedure are shown in panels a–c, where the clusters are shown in panels a–b and a barplot of grain types versus clusters is shown in panel c. The stable clusters are depicted in panel d–e, with a barplot of grain types versus the stable clusters shown in panel f. The core clusters are shown in panels g–h, with a barplot of grain types versus the core clusters shown in panel i. The results from the BagClust2 procedure are shown in panels j–l, where the clusters are shown in panels j–k and a barplot of grain types versus clusters is shown in panel l.

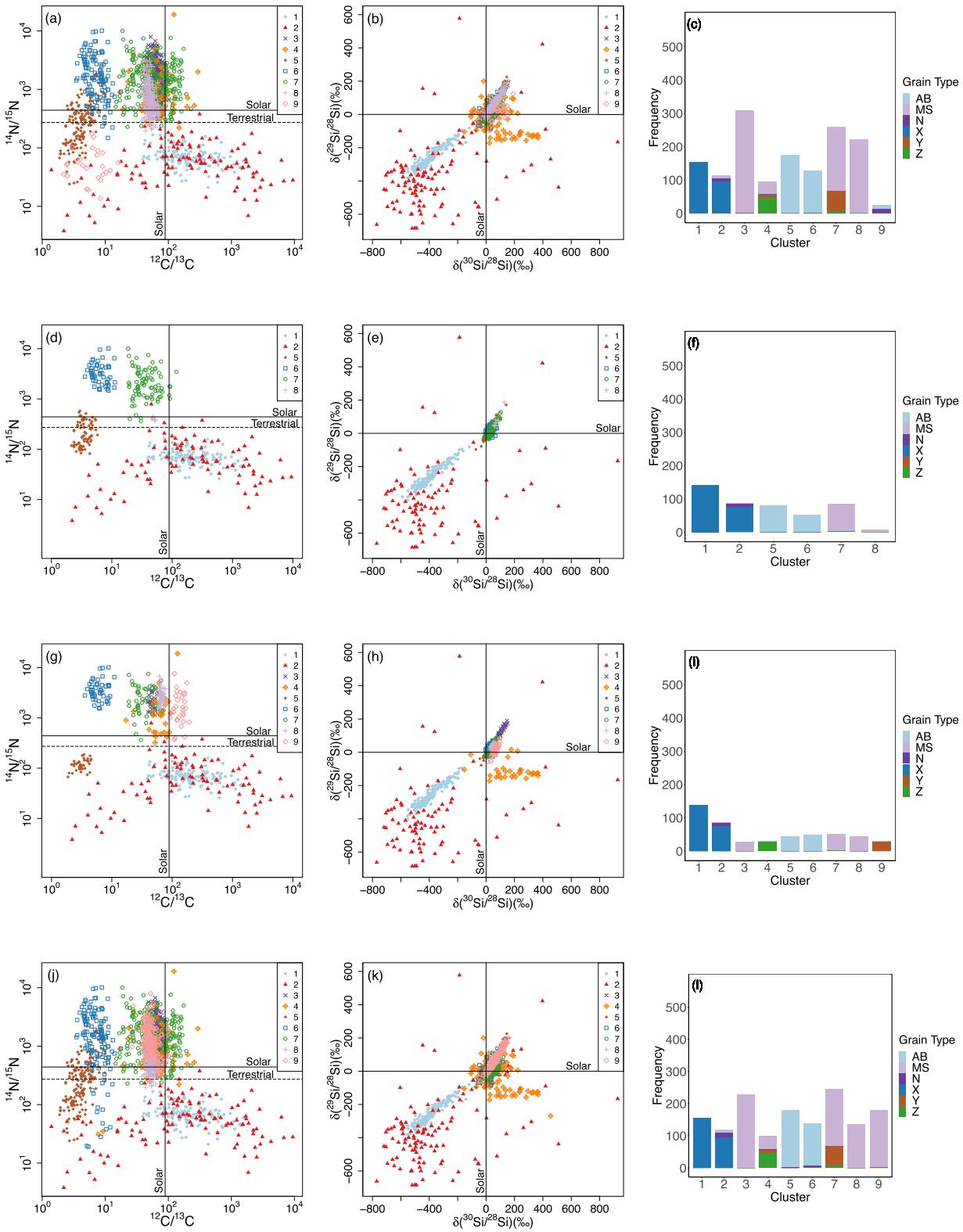

Figure 8. Clustering of pre-solar SiC grains with a mixture of normal distributions using nine clusters. The results from the BagClust1 procedure are shown in panels a–c, where the clusters are shown in panels a–b and a barplot of grain types versus clusters is shown in panel c. The stable clusters are depicted in panels d–e, with a barplot of grain types versus the stable clusters shown in panel f. The core clusters are shown in panels g–h, with a barplot of grain types versus the core clusters shown in panel i. The results from the BagClust2 procedure are shown in panels j–l, where the clusters are shown in panels j–k and a barplot of grain types versus clusters is shown in panel l.

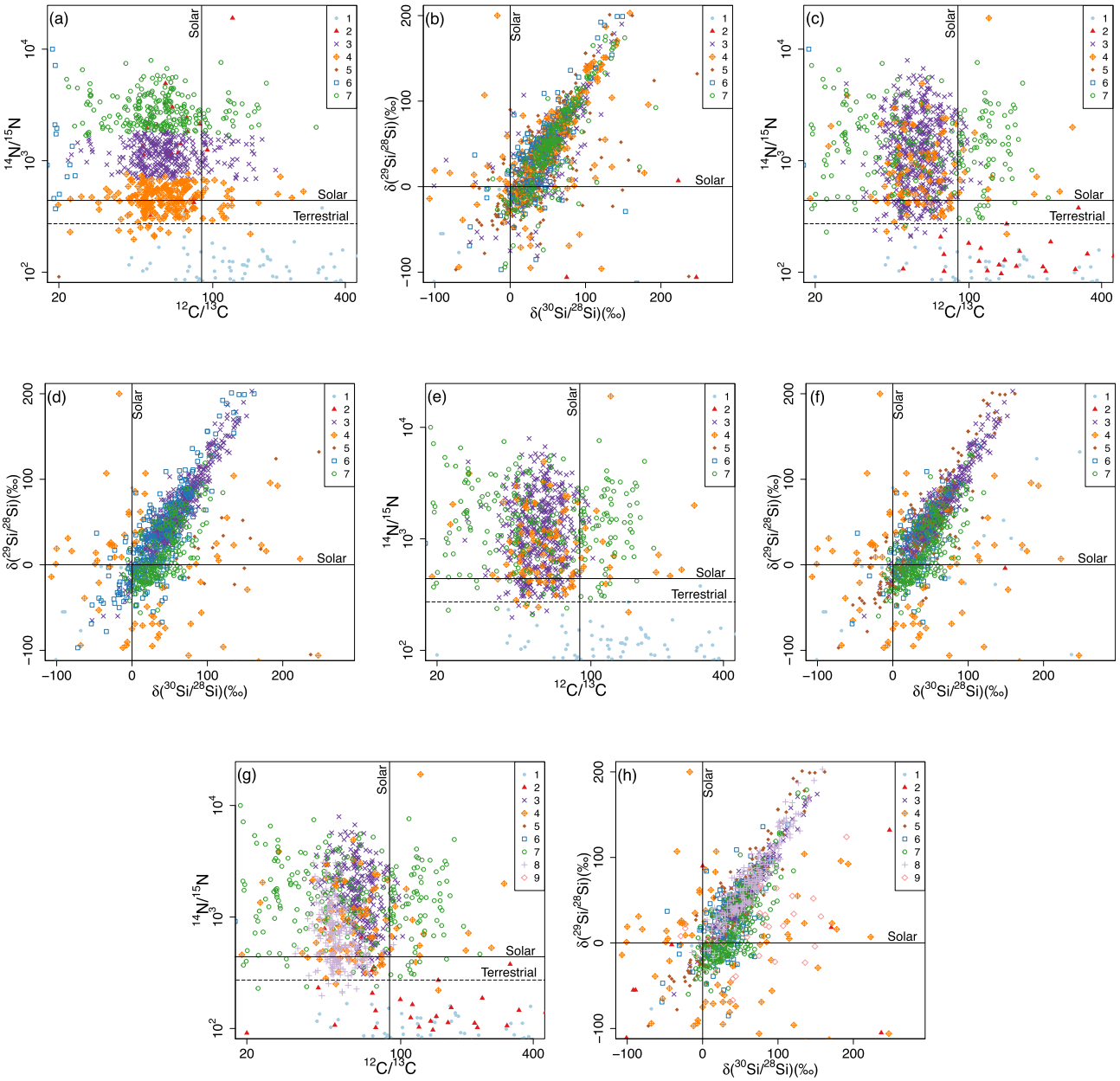

Figure 9. Zoom-in of panels a–b of Figs 5–8 for spectral clustering (panels a–b), a mixture of t-distributions (panels c–d), a mixture of normal distributions (panels e–f), all with seven clusters, and a mixture of normal distributions with nine clusters (panels g–h).

Table 1. The proportion of pre-solar SiC grains with cluster vote $CV \geq 0.90$, the average Jaccard coefficient, the overall average silhouette width, the proportion of grains with silhouette width, $SW > 0.50$, and the proportion of grains that occur in core clusters obtained for the spectral algorithm (Spectral), the mixture of t-distributions (t(UUUU)), and the mixture of normal distributions (Normal (VVV),7), all with seven clusters, and the mixture of normal distributions with nine clusters (Normal (VVV),9).

Methods/Model	$CV \geq 0.90$	Average Jaccard coefficient	Average silhouette width	$SW > 0.50$	Core clusters
Spectral	0.882	0.936	0.895	0.963	0.852
t (UUUU)	0.548	0.647	0.740	0.851	0.554
Normal (VVV),7	0.256	0.628	0.728	0.909	0.380
Normal (VVV),9	0.307	0.699	0.613	0.708	0.336

and Mo isotope ratios and inferred initial $^{26}Al/^{27}Al$ ratios) for AB grains. It is noteworthy that the spectral clustering results for AB and MS grains depend strongly on the $^{14}N/^{15}N$ ratio and could reflect some artefact in the grain data. This is because it was shown that the literature $^{14}N/^{15}N$ isotope data of pre-solar SiC grains sampled different proportions of solar or terrestrial nitrogen contamination (Liu et al. 2021). However, it is interesting that the identified cluster 4 exhibits both the lowest range of $^{14}N/^{15}N$ ratios among the three

Table 2. The proportion of pre-solar SiC grains in clusters with $CV \geq 0.90$, the average Jaccard similarity coefficient, average silhouette width, and the proportion of grains in core clusters for each cluster obtained from spectral clustering (Spectral), a mixture of t-distributions (t (UUUU)), a mixture of normal distributions (Normal (VVV),7) all with seven clusters, and a mixture of normal distributions (Normal (VVV),9) with nine clusters.

Cluster	Methods/Model	$CV \geq 0.90$	Jaccard coefficient	Silhouette width	Proportion in core clusters
1	Spectral	0.996	0.994	0.989	0.996
	t (UUUU)	0.713	0.858	0.839	0.636
	Normal (VVV),7	0.522	0.764	0.870	0.946
	Normal (VVV),9	0.916	0.924	0.906	0.932
2	Spectral	0.909	0.927	0.913	0.909
	t (UUUU)	0.000	0.579	0.573	0.165
	Normal (VVV),7	0.000	0.016	0.751	0.750
	Normal (VVV),9	0.772	0.844	0.758	0.696
3	Spectral	0.773	0.879	0.814	0.706
	t (UUUU)	0.869	0.862	0.913	0.885
	Normal (VVV),7	0.039	0.745	0.727	0.174
	Normal (VVV),9	0.000	0.625	0.493	0.095
4	Spectral	0.849	0.925	0.866	0.847
	t (UUUU)	0.402	0.653	0.463	0.390
	Normal (VVV),7	0.069	0.753	0.709	0.347
	Normal (VVV),9	0.000	0.780	0.757	0.437
5	Spectral	0.930	0.966	0.936	0.894
	t (UUUU)	0.000	0.284	0.589	0.250
	Normal (VVV),7	0.218	0.742	0.694	0.500
	Normal (VVV),9	0.466	0.757	0.690	0.232
6	Spectral	0.911	0.952	0.934	0.889
	t (UUUU)	0.503	0.789	0.729	0.354
	Normal (VVV),7	0.313	0.646	0.621	0.283
	Normal (VVV),9	0.417	0.697	0.612	0.340
7	Spectral	0.888	0.909	0.881	0.851
	t (UUUU)	0.000	0.507	0.605	0.259
	Normal (VVV),7	0.502	0.727	0.719	0.294
	Normal (VVV),9	0.324	0.705	0.713	0.280
8	Normal (VVV),9	0.0317	0.504	0.544	0.203
9	Normal (VVV),9	0.000	0.457	0.176	0.232

MS clusters and the large variations from the MS line defined by the other two MS clusters in the Si 3-isotope plots. This observation suggests that in addition to contamination, the close-to-solar and-terrestrial $^{14}N/^{15}N$ ratios observed in this cluster of grains also reflect some real nucleosynthetic effect. MS grains are commonly argued to have come from low-mass asymptotic giant branch (AGB) stars with close-to-solar and/or higher than solar metallicities (Cristallo et al. 2020; Lugaro et al. 2020). The MS grains' $^{14}N/^{15}N$ ratios, however, cannot be explained by stellar nucleosynthesis models for such low-mass AGB stars, because larger ^{14}N enrichments are expected for their stellar envelope composition (Palmerini et al. 2011). The inconsistency therefore suggests that this cluster of grains did not originate from carbon-rich low-mass AGB stars, or that their parent AGB stars were born from materials that experienced heterogeneous Galactic chemical evolution (GCE). It is interesting to note that when clustering with C, N, Si, and inferred Al isotopic ratios, Boujibar et al. (2021) found two groups of MS grains with distinct ranges of $^{14}N/^{15}N$ ratio. One of these two clusters has a composition similar to cluster 4: $^{14}N/^{15}N$ lower than 1000 and shows significant deviations from the MS line in the Si 3-isotope plot. In addition, this previously identified cluster has relatively high $^{26}Al/^{27}Al$ ratios. The coupled low $^{14}N/^{15}N$ and high $^{26}Al/^{27}Al$ ratios of grains in this cluster is in contrast to the coupled high $^{14}N/^{15}N$ and high $^{26}Al/^{27}Al$ ratios generally predicted by state-of-the-art AGB model calculations (Cristallo et al. 2009; Karakas & Lugaro 2016), which further corroborates the data–model discrepancy in the $^{14}N/^{15}N$ ratio.

4.2 Clustering with a mixture of t-distributions using seven clusters

Both the BagClust1 and BagClust2 clustering procedures cluster the pre-solar SiC grains into two main groups of X grains (clusters 1 and 2), two main groups of MS grains (clusters 3 and 7), and one group dominated by both MS and Z grains (cluster 4) using a mixture of *t*-distributions with seven clusters (Fig. 6). The BagClust1 procedure (Fig. 6, panels a–c) clusters the grains into one main group of AB grains (cluster 6), whereas the BagClust2 procedure (Fig. 6, panels j–l) yields two main groups of AB grains (clusters 5 and 6). More N grains are clustered with the X grains using BagClust2 compared to BagClust1. Y and Z grains are again clustered jointly with the MS grains in two clusters. However, BagClust2 resulted in more of the Y grains to be included in cluster 4 compared to BagClust1. Cluster 1 (X grains) and cluster 3 (almost all MS grains), are highly stable in terms of having a high percentage of grains with CV values greater than or equal to 0.90 and a high average Jaccard coefficient. Cluster 6, which contains AB grains, has some observations with CV of at least 0.90. Cluster 2 (N and X grains), cluster 5 (N and a few AB grains), and cluster 7 (MS, Y, and a few Z grains) are highly unstable. All types of grains, except for MS and Z grains, have less than 50 per cent of the grains in stable clusters. All Y and N grains have $CV < 0.9$ (Table 3). The core clustering procedure (Fig. 6, panels g–i) produces results similar to the BagClust1 procedure, but has slightly different observations in core clusters as in stable clusters (Fig. 6, panels d–f). From the BagClust2 clustering procedure, clusters 1, 3, and 6 have

Table 3. Proportion of pre-solar SiC grain types with $CV \geq 0.90$ and in core clusters obtained from spectral clustering (Spectral), a mixture of t-distributions (t (UUUU)), a mixture of normal distributions (Normal (VVV),7) all with seven clusters, and a mixture of normal distributions (Normal (VVV),9) with nine clusters.

Methods/Model	Grain type	Proportion of types with $CV \geq 0.90$	Proportion of types in core clusters
Spectral	Z	0.830	0.830
	Y	0.549	0.465
	X	1.000	1.000
	N	0.783	0.696
	MS	0.854	0.816
	AB	0.946	0.924
t (UUUU)	Z	0.566	0.321
	Y	0.000	0.113
	X	0.460	0.480
	N	0.000	0.174
	MS	0.663	0.731
	AB	0.498	0.349
Normal (VVV), 7	Z	0.075	0.264
	Y	0.366	0.014
	X	0.568	0.880
	N	0.000	0.435
	MS	0.163	0.245
	AB	0.257	0.406
Normal (VVV), 9	Z	0.038	0.528
	Y	0.014	0.394
	X	0.864	0.844
	N	0.435	0.435
	MS	0.119	0.168
	AB	0.425	0.289

high average silhouette width values, and 12 observations have a negative silhouette width.

Our t-distribution clustering results are generally similar to those of Boujibar et al. (2021), as (1) MS grains are divided into three clusters (clusters 3, 4, and 7) with different ranges of Si isotope ratios, representing their parent stars' different ranges of initial stellar metallicities, (2) X grains are divided into two clusters with one of the clusters exhibiting less correlated Si isotope ratios, and (3) two groups of AB grains with some overlap in their $^{14}N/^{15}N$ isotope ratios for the BagClust2 method. Our stability tests further illustrate that the three clusters, including MS grains with the lower range of Si isotope ratios (cluster 7), ^{13}C, ^{15}N-enriched AB and N grains (cluster 5), and X grains with less correlated Si isotope ratios (cluster 2), are not that statistically robust. In agreement with the Boujibar et al. (2021) conclusion, our results point out that there is no distinct separation of MS, Y, and Z grains. Interestingly, the stable core of ^{14}N-rich AB grains (cluster 6) in panels g–i of Fig. 6 seems to exhibit a weak positive correlation between their ^{13}C and ^{14}N enrichments, which is consistent with H burning signatures at low-stellar temperatures ($<\sim 1 \times 10^8$ K; Palmerini et al. 2011). It is noteworthy that (1) this weakly correlated ^{13}C and ^{14}N enrichments in ^{14}N-rich AB grains is generally supported by all the model-based clustering and the associated stability assessments (Figs 7 and 8) and (2) this correlation of ^{14}N-rich AB grains is in contrast to the weakly correlated ^{13}C and ^{15}N enrichments in ^{15}N-rich AB grains observed in cluster 5 (panels g, j of Fig. 6; panels a, d, g, j of Figs 7 and 8), which points to H burning signatures in explosive environments (e.g, novae, core-collapse supernovae; Liu et al. 2016). The opposite trends observed between ^{14}N-rich and ^{15}N-rich AB grains therefore favour two different stellar formation environments and are more consistent with ^{14}N-rich AB grains dominantly originating from J-type carbon stars or born-again AGB stars (Liu et al. 2017b) and ^{15}N-rich AB grains dominantly from core-collapse supernovae (Liu et al. 2017a, 2018). Note that it is still possible that some of the ^{14}N-rich AB grains that are not included in the stable part of cluster 6 originated from core-collapse supernovae, as suggested by Hoppe et al. (2019).

4.3 Clustering with a mixture of normal distributions using seven clusters

Using a mixture of normal distributions with seven clusters for the BagClust1 clustering procedure, produces one main group of X grains (cluster 1), two main groups of AB grains (clusters 5 and 6), and three main groups of mixed combinations of MS, Y, Z grains (clusters 3, 4, and 7). Cluster 3 is dominated by MS grains, while clusters 4 and 7 have Z and Y grains clustered jointly with MS grains (Fig. 7, panels a–c). Cluster 2 only contains two observations (N and X grains) and is highly unstable. Some runs of the algorithm produce two main groups of X grains (in clusters 1 and 2), but cluster 2 does not reach majority vote for any of the observations, with two exceptions. N grains are mostly clustered jointly with X grains, but some of the N grains are clustered with AB grains. The main difference between the partitioning obtained from the BagClust1 and BagClust2 procedures (Fig. 7, panels j–l) are found in clusters 1 and 2, where the BagClust2 procedure split X-grains into two groups.

Clusters 1 and 7 are the most stable clusters in terms of having the largest proportion of grains with $CV \geq 0.9$. These clusters have moderate Jaccard coefficient values. The stable parts of cluster 1 consist of X grains, while in cluster 7 the stable parts consist largely of MS grains with some Y grains and a few Z grains. Clusters 3, 4, 5, and 6 show some patterns by yielding an average Jaccard coefficient above 0.6, but with a lower proportion of grains with $CV \geq 0.9$. All the grain types have a majority of their observations with $CV < 0.90$ except for X grains. The core clustering procedure (Fig. 7, panels g–i) yielded a somewhat different partitioning than the ones

obtained from the BagClust1 method, but similar to the BagClust2 procedure. Here X grains are again split into two clusters (clusters 1 and 2) in panels g–i, in contrast to zero grains in cluster 2 in the stable cluster results shown in panels d–f. The majority of the grains are characterized as weak points, except for X grains that are mostly occurring in core clusters (panels g–i and Table 3).

The BagClust2 clustering procedure yielded clusters 1, 2, 3, 4, and 7 with moderate to high average silhouette widths. There are only three observations that have negative silhouette width values. Compared to the t-distribution results shown in the previous section, the normal distribution results here (1) point out that the division of X grains into two clusters is not robust and (2) shows the division of AB grains into two clusters enriched or depleted in ^{15}N and ^{13}C, with a strong overlap at ^{14}N/^{15}N $\approx$ 500–2000. Note that AB grains were initially divided into A and B grains by adopting a divider of ^{12}C/^{13}C $= 3.5$ (Hoppe et al. 1994). The stability test further shows that the overlapping parts of the AB grains are not stable and are ambiguous regarding their cluster assignment. Thus, the spectral clustering, clustering with a mixture of t-distributions, and a mixture of normal distributions all support the division of AB grains into two clusters by adopting a divider of around the solar ^{14}N/^{15}N ratio, in line with the proposal of Liu et al. (2017a).

4.4 Clustering with a mixture of normal distributions using nine clusters

Both of the BagClust1 (Fig. 8, panels a–c) and BagClust2 (Fig. 8, panels j–l) procedures cluster the grains into two main groups of X grains (clusters 1 and 2) and two main groups of AB grains (clusters 5 and 6) when clustering with a mixture of normal distributions using nine clusters. BagClust1 yields four main groups of MS–Y–Z grains (clusters 3, 4, 7, and 8), whereas BagClust2 yields five main groups of MS–Y–Z grains (clusters 3, 4, 7, 8, and 9). N grains are now included in clusters with both X and some ^{15}N- rich AB grains. Y and Z grains are clustered jointly with MS grains in two clusters (clusters 4 and 7). Clusters 3 and 8 produced by the two procedures are different in their ranges of C and N isotopic ratios, respectively. Also, cluster 9 in BagClust2 is dominated by MS grains, while in BagClust1, it comprises AB and N grains. The most stable clusters are cluster 1, which contains X grains, and cluster 2, which includes mostly X grains and half of the N grains. Cluster 3 (MS grains), cluster 4 (MS, Y, and Z grains), and cluster 8 (MS grains) are highly unstable because these clusters have no grains (or only a few grains for cluster 8) with CV of at least 0.9. Cluster 9, which contains a few AB grains and half of the N grains from the BagClust1 clustering procedure, is also unstable. Clusters 8 and 9 also yield low average Jaccard coefficients.

There are fewer observations in stable clusters (Fig. 8, panels d–f) than in core clusters (Fig. 8, panels g–i). The main difference between the partitions obtained from the BagClust1 method and the core cluster identification method is in how cluster nine is defined, where core cluster 9 contains Y grains. X grains are the only grain type that has a majority of observations with a CV of at least 0.9 as well as the majority of the observations occurring in core clusters. Slightly more than half of the Z grains occur in core clusters (Table 3).

From the BagClust2 clustering procedure, clusters 1, 2, 4, and 7 have moderate or high average silhouette widths. There are 40 observations with negative silhouette width values, for which almost all are MS grains.

In summary, compared to the normal distribution clustering using seven clusters, the two extra clusters identified in this section mainly include additional X and MS grain clusters. While the addition of X grain cluster (cluster 2) is statistically robust, the additional MS cluster (cluster 8) is highly unstable.

4.5 Comparison of the clusters given by different clustering methods

In previous sections, we have evaluated the classification of pre-solar SiC grains by using different cluster analysis ensemble methods for grain partitioning. Here, we further compare the first seven clusters obtained from spectral clustering, a mixture of t-distributions, and a mixture of normal distributions. Using the Hungarian algorithm as described in Section 2.3, we kept a consistent labelling of the clusters across the different clustering methods. Recall that the cluster labels adopt the labels from one run of a mixture of normal distributions with seven clusters as the reference labels. As a result, a few clusters have no direct identification across the clustering algorithm and methods. Clusters 2, 3, 4, and 7 from the spectral algorithm best matches clusters 9, 7, 8, and 3, respectively, from clustering with a mixture of normal distributions using nine clusters (BagClust1). Cluster 6 from the mixture of t-distribution using BagClust1 is better identified with cluster 5 in some of the other clustering algorithms. It also varies how cluster 5 from the mixture of t-distributions using BagClust1 is identified. The difficulty here arises because the BagClust1 procedure with the mixture of t-distributions resulted in only one AB group, while all the other algorithms and methods produced two main groups of AB grains. A similar issue is seen with cluster 2 (Z-grains) from the spectral algorithm.

(1) Cluster 1, which is dominated by X grains, is the most stable cluster across all the algorithms (with the exception of cluster 3 for the mixture of t-distributions). In particular, it shows a high stability for both the spectral clustering technique with seven groups and the clustering with a mixture of normal distributions using nine groups. Clustering with a mixture of normal distributions using seven clusters assigned both X and N grains to cluster 1, but only X grains are in the stable parts of the cluster. All clustering methods yielded large overlapping stable parts of X grains in cluster 1.

(2) Cluster 2 has a high stability for both the spectral clustering technique and for clustering with a mixture of normal distributions using nine clusters. Cluster 2 contains both X and N grains when using a mixture of normal distributions with nine clusters and t-distributions with seven clusters. In contrast, cluster 2 consists of only Z grains (except for one MS grain) when using the spectral algorithm with seven clusters. Clustering with a mixture of normal distributions using both seven and nine groups also yielded a large number of both X and N grains in core cluster 2. Thus, except for the spectral clustering analysis results, all the other clustering analysis methods point to the clustering of N grains with a portion of X grains from the aspect of statistics. This inferred genetic relationship between N and X grains is strongly supported by new supernova models and detailed isotopic investigation of N grains (Nittler & Hoppe 2005; Pignatari et al. 2015; Liu et al. 2016).

(3) Cluster 3 is most stable for the mixture of t-distributions using seven groups but also has a high stability for spectral clustering with seven groups. Cluster 3 is always dominated by MS grains and has some overlaps across the different clustering techniques. With model-based clustering, cluster 3 is characterized by a relatively narrow range of ^{12}C/^{13}C, and highly correlated Si isotopic ratios compared to other MS-grain-containing clusters. Spectral clustering assigned several Y and a few Z grains to cluster 3, highlighting a common problem in separating MS, Y, and Z grains regarding the original classification scheme and AGB nucleosynthesis models.

(4) Cluster 4 and cluster 7, which consist of mixtures of MS, Y, and Z grains for all the methods, have only high stabilities for the spectral algorithm. The MS, Y, and Z grains in clusters 4 and 7 are split differently for the spectral algorithm compared to the other model-based methods, with a division controlled by the $^{14}N/^{15}N$ ratio. However, as noted earlier, the adoption of the $^{14}N/^{15}N$ ratio alone for dividing MS grains by the spectral algorithm is not reliable given the known problem of nitrogen contamination. Thus, although the spectral analysis results have higher stabilities than the other methods for these clusters, the results may mainly reflect different degrees of nitrogen contamination instead of having significant astrophysical meanings.

(5) Cluster 5 is dominated by AB grains that have a low range of $^{14}N/^{15}N$ for the mixture of normal distributions with both seven and nine clusters (BagClust1 and BagClust2) and the mixture of t-distributions (BagClust2), whereas for the spectral algorithm it consists of both AB and N grains. Cluster 5 only has high stability for the spectral algorithm, but AB grains with CV values $\geq$0.9 show some overlap among the spectral clustering, and clustering with a mixture of normal distributions with both seven and nine clusters.

(6) Cluster 6, which is also dominated by AB grains, has a high stability for the spectral algorithm and has a high Jaccard coefficient for the mixture of t-distributions using seven groups. The stable parts of cluster 6 in different methods have some overlap of AB grains for high values of $^{14}N/^{15}N$.

4.6 Astrophysical implications

Across the different models and clustering techniques used in this paper, we propose that the SiC grains are partitioned into two main groups of AB grains, three main groups of MS–Y–Z grains, and one main group of Z grains. It is uncertain whether X grains should be split into two groups.

4.6.1 MS, Y, and Z grains

The main difference between the spectral clustering method and the model-based methods was found in the splitting of the MS–Y–Z grains; the use of $^{14}N/^{15}N$ ratio as the divider by spectral clustering versus the use of Si and C isotope ratios by the mixture models. The three MS–Y–Z grain groups that are classified by the spectral clustering algorithm, have different ranges of $^{14}N/^{15}N$ ratios (Clusters 3, 4, 7). In contrast, the model-based methods yielded MS groups with different C and Si isotopic signatures. While cluster 3 has a narrow range of $^{12}C/^{13}C$, clusters 4 and 7 show a spread in the $^{12}C/^{13}C$ ratio for the model-based methods. In addition, spectral clustering splits the Z and Y grains over several groups, while the model-based clustering methods include the Z and Y grains mostly in two clusters (clusters 4 and 7, respectively). Z and Y grains are both clustered jointly with MS in all methods, while the spectral algorithm also produces one cluster (cluster 2) that consists mainly of Z grains. These clustering results, overall, highlight the genetic relationship among MS, Y, and Z grains, as they all carry *s*-process isotopic signatures and should have originated from carbon-rich AGB stars (Liu et al. 2019). The inconsistent clustering schemes yielded by these different methods, on the other hand, highlight the difficulties in dividing these AGB dust grains into sub-groups. For instance, although MS, Y, and Z grains show different degrees of ^{30}Si excesses, they exhibit indistinguishable Mo isotopic signatures (Liu et al. 2019). In Stephan et al. (2021) it was also stated that the division between MS, Y, and Z grains is still preliminary and some of the existing criteria seem arbitrary. Isotopic data from more elements are needed in the future to investigate whether there are distinct AGB dust subpopulations.

4.6.2 AB, N, and X grains

N grains occur mostly with ^{15}N-rich AB grains in cluster 5 for the spectral clustering method, while they occur mostly with X grains in cluster 1 for clustering with a mixture of normal distributions using seven clusters with BagClust1 and cluster 2 with BagClust2. Clustering with a mixture of t-distributions and a mixture of normal distributions using nine clusters resulted in N grains to be included with both the X grains in cluster 2 and the AB grains in cluster 5 (for the mixture of t-distributions) and cluster 9 (for the mixture of normal distributions) using BagClust1. These classification schemes are generally consistent with the fact that N grains, ^{15}N-rich AB grains and X grains have all been proposed to originate from core-collapse supernovae (Nittler et al. 1996; Liu et al. 2016, 2017a, 2018). The genetic relationship between N and ^{15}N-rich AB grains, however, is expected to be stronger, because they both recorded strong explosive H-burning isotopic signatures at high temperatures, i.e. large ^{13}C, ^{15}N, ^{26}Al enrichments. This hypothesis is consistent with the fact that cluster 5 identified by the spectral clustering method, consisting of mostly N grains with ^{15}N-rich AB grains, is highly stable, given its large CV (0.930), Jaccard Coefficient (0.966), and silhouette width (0.936) values (Table 2). Besides, the division of X grains into two groups is uncertain as it is only supported by the stability of the second cluster when using a mixture of normal distributions with nine groups (panels d–f in Fig. 8). However, the two groups of X grains seem to be supported by the observation of Stephan et al. (2018) that their two X2 (belonging to cluster 2 in (Normal(VVV),9)) grains and one X1 (cluster 1 in (Normal (VVV),9)) grain had different Sr and Ba isotopic signatures, thus suggesting different formation conditions in Type II supernovae for the two types of X grains. Lin, Gyngard & Zinner (2010) proposed the division of X grains into three sub-groups, X0, X1, and X2, based on their different Si isotopic signatures.

5 CONCLUSION

In this study, we assessed the stability of the clusters and the confidence of the grain assignment to the clusters over several different models and clustering methods with the goal of identifying pre-solar SiC grains that occur in stable clusters.

We demonstrated the use of cluster analysis and cluster ensemble techniques to evaluate the confidence in classifying pre-solar SiC grains. Our spectral clustering method yielded seven clusters with the highest stabilities and reproducibilities, including two main groups of AB grains, three main groups of MS–Y–Z grains, one main group of Z grains, and one main group of X grains. Based on our stability assessment of the clusters, we come to the following conclusions.

(1) The inconsistent clustering of MS grains by different clustering methods and the poor stabilities of the identified MS grain clusters, show the difficulties in dividing the MS group. The three MS grain clusters yielded by the spectral clustering (clusters 3, 4, 7), divided by their different ranges of $^{14}N/^{15}N$ ratios, are most stable, in contrast to the MS grain clusters identified by the model-based clustering methods, which partitioned the grains by their different ranges of Si isotope ratios. While the different $^{14}N/^{15}N$ ratios of the three clusters likely reflect the varying amounts of nitrogen contamination sampled from the grains, cluster 4 exhibits both the lowest $^{14}N/^{15}N$

ratios and least correlated Si isotope ratios in the Si 3-isotope plot, pointing to the effect of true parent stellar nucleosynthesis signatures. Cluster 4 shows some similarity to the DB5 MS cluster 8 with low $^{14}N/^{15}N$ and high $^{26}Al/^{27}Al$ ratios identified by Boujibar et al. (2021). Both of the clusters cannot be explained by state-of-the-art low-mass AGB nucleosynthesis model calculations and together point to either problems in current AGB nucleosynthesis models or the fact that they did not originate from low-mass AGB stars. Given the inconsistency in partitioning MS grains into sub-populations, it is important to obtain higher quality isotope data by suppressing potential contamination and include more attributes like isotopic ratios of other elements, interstellar ages, and grain sizes, which will allow clustering analysis to provide a clearer picture whether distinct groups of MS grains were incorporated into the Solar system.

(2) All the clustering methods partitioned MS, Y, and Z grains into the same clusters, pointing to their genetic relationships, as supported by their common origin in low-mass AGB stars. While Z grains are mostly separated from MS and Y grains and form a single cluster by the different clustering methods, no clear separation was found between MS and Y grains, consistent with the conclusion of Boujibar et al. (2021).

(3) AB grains are generally divided into two groups mostly based on their nitrogen isotope ratios by all the clustering methods. The two AB clusters with the highest stabilities given by the spectral clustering method, adopted a divider around the solar $^{14}N/^{15}N$ ratio, consistent with the previous proposal of dividing AB grains into AB1 and AB2 grains using the solar $^{14}N/^{15}N$ ratio based on a larger set of multi-element isotope systematics. With model-based clustering methods, the two AB clusters are defined by both their N and C isotope ratios, similarly to finding of Boujibar et al. (2021). However, for the model-based methods, the stable and core clusters showed a clear separation of the two groups of AB grains as the AB grains in the unstable parts of the clusters or core clusters could belong to either of the two groups. The ^{14}N-rich AB clusters and/or their cores yielded by the model-based clustering methods, often show weakly correlated ^{14}N and ^{13}C enrichments, while the ^{15}N-rich AB clusters sometimes show weakly correlated ^{15}N and ^{13}C enrichments. The opposite trends observed between these two groups of AB grains point to two distinct stellar formation environments and provide an additional clue to constrain their still ambiguous stellar origins.

(4) X grains are either all clustered in one group with spectral clustering or divided into two clusters with model-based clustering similar to results of Boujibar et al. (2021). One of the two clusters is very stable and has X grains tightly correlated in the 3-Si isotopes, which suggests mixing between material from supernovae outer He/C zone with inner Si/S or Si/C zones (Nittler et al. 1996; Rauscher et al. 2002; Pignatari et al. 2013). The other cluster comprises X grains deviating from this line, indicating contribution of material from other shells. However, the second cluster is not supported by all the clustering methods and is highly unstable for the model-based methods with seven clusters.

(5) N grains are often clustered with X or ^{15}N-rich AB grains, in line with their proposed core-collapse supernova origin. The highly stable cluster 5 given by the spectral clustering method consists mostly of ^{15}N-rich AB and N grains, in agreement with the fact that both types of grains recorded strong H-burning signatures (e.g. ^{13}C, ^{15}N and ^{26}Al enrichments) at high stellar temperatures. This suggested genetic relationship between N and ^{15}N-rich AB grains should be considered in future investigation of their stellar origins.

Cluster analysis is an unsupervised machine-learning technique, where the clustering does not take into account the original classification of grains. In conclusion, the broad division of X/AB/M–Y–Z grains in this paper agree with the broad lines of the existing classification. However, using cluster analysis, we have a quantitative way to classify and to assess the confidence of the classification of the pre-solar SiC grains. As the results can be confirmed with statistical analysis, it will provide further insight to the classification scheme.

ACKNOWLEDGEMENTS

Data-driven studies of mineral evolution and mineral ecology have been supported by the Alfred P. Sloan Foundation, the W. M. Keck Foundation, the John Templeton Foundation (grant #60645), the NASA Astrobiology Institute ENIGMA team (80NSSC18M0093), a private foundation, and the Carnegie Institution for Science. NL acknowledges financial support from NASA (80NSSC20K0387 to NL). LRN acknowledges NASA grant #80NSSC20K0340. Any opinions, findings, or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the National Aeronautics and Space Administration. We also thank the reviewers for helpful comments and suggestions in improving the paper.

DATA AVAILABILITY

The data underlying this article are available in GitHub.[6] The data are from the Pre-solar Grain Data base at https://pre-solar.physics.wustl.edu/pre-solar-grain-data base/

REFERENCES

Andrews J. L., McNicholas P. D., 2012, Stat. Comput., 22, 1021
Andrews J. L., Wickins J. R., Boers N. M., McNicholas P. D., 2018, J. Stat. Softw., 83, 1
Boujibar A. et al., 2021, ApJ, 907, L39
Bouveyron C., Celeux G., Murphy T. B., Raftery A. E., 2019, Model-Based Clustering and Classification for Data Science: With Applications in R, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge Univ. Press, Cambridge
Cristallo S., Straniero O., Gallino R., Piersanti L., Domínguez I., Lederer M. T., 2009, APJ, 696, 797
Cristallo S., Nanni A., Cescutti G., Minchev I., Liu N., Vescovi D., Gobrecht D., Piersanti L., 2020, A&A, 644, A8
Csárdi G., Nepusz T., 2006, The Igraph Software Package for Complex Network Research
Dudoit S., Fridlyand J., 2003, Bioinformatics, 19, 1090
Efron B., Tibshirani R. J., 1993, Monographs on Statistics and Applied Probability, Vol. 57, An Introduction to the Bootstrap. Chapman & Hall/CRC, London
Efron B., Tibshirani R. J., 1997, J. Am. Stat. Assoc., 92, 548
Everitt B. S., Landau S., Leese M., Stahl D., 2011, Cluster Analysis, 5th edition, Wiley Series in Probability and Statistics. John Wiley & Sons, West Sussex
Fred A. L. N., Jain A. K., 2005, Proc. IEEE., 27, 835
Ghosh J., Acharya A., 2011, WIREs Data Mining and Knowledge Discovery. Wiley, New York, p. 305
Henelius A., Puolamäki K., Bostr'om H., Papapetrou P., 2016, Clustering with Confidence: Finding Clusters with Statistical Guarantees. Available at:https://arxiv.org/abs/1612.08714
Hennig C., 2007, Comput. Stat. Data Anal., 52, 258
Hennig C., 2020, fpc: Flexible Procedures for Clustering. Available at: https://CRAN.R-project.org/package = fpc

[6]https://github.com/ghystad/Consensus-clustering.presolar.SiC.grains/blob/master/pre-solar_SiC_data.xlsx

Hennig C., Hausdorf B., 2020, prabclus: Functions for Clustering and Testing of Presence-Absence, Abundance and Multilocus Genetic Data. Available at: https://CRAN.R-project.org/package = prabclus
Hoppe P., Amari S., Zinner E., Ireland T., Lewis R. S., 1994, ApJ, 430, 870
Hoppe P., Stancliffe R. J., Pignatari M., Amari S., 2019, ApJ, 887, 8
Hynes K. M., Gyngard F., 2009, Lunar and Planetary Science Conference. Woodlands, TX, p. 1198
James G., Witten D., Hastie T., Tibshirani R., 2013, An Introduction to Statistical Learning with Application in R, Springer Texts in Statistics. Springer, New York
Karakas A. I., Lugaro M., 2016, ApJ, 825, 26
Kaufman L., Rousseeuw P. J., 1990, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ
Kolaczyk E. D., Csárdi G., 2014, Statistical Analysis of Network Data with R, Use R!. Springer, New York
Kuhn H. W., 1955, Nav. Res. Logist. Q., 2, 83
Lin Y., Gyngard F., Zinner E., 2010, ApJ, 709, 1157
Liu N., Nittler L. R., Alexander C. M. O'D., Wang J., Pignatari M., José J., Nguyen A., 2016, ApJ, 820, 140
Liu N., Nittler L. R., Pignatari M., Alexander C. M. O'D., Wang J., 2017a, ApJ, 842, L1
Liu N. et al., 2017b, ApJ, 844, L12
Liu N. et al., 2018, ApJ, 855, 144
Liu N. et al., 2019, ApJ, 881, 28
Liu N., Barosh J., Nittler L. R., Alexander C. M. O'D., Wang J., Cristallo S., Busso M., Palmerini S., 2021, ApJ, 920, L26
Lugaro M. et al., 2020, ApJ, 898, 96
Maechler M., Rousseeuw P., Struyf A., Hubert M., Hornik K., 2021, cluster: 'finding Groups in Data': Cluster Analysis Extended Rousseeuw et al.. Available at: https://CRAN.R-project.org/package = cluster
McLachlan G., Peel D., 2000, Finite Mixture Models, Wiley Series in Probability and Statistics. John Wiley & Sons, New York
Monti S., Tamayo P., Mesirov J., Golub T., 2003, Mach. Learn., 52, 91
Nittler L. R., Ciesla F., 2016, ARA&A, 54, 53
Nittler L. R., Hoppe P., 2005, ApJ, 631, L89
Nittler L. R., Amari S., Zinner E., Woosley S. E., Lewis R. S., 1996, ApJ, 462, L31
Palmerini S., La Cognata M., Cristallo S., Busso M., 2011, ApJ, 729, 3
Pignatari M. et al., 2013, ApJ, 771, L7
Pignatari M. et al., 2015, ApJ, 808, L43
R Core Team, 2016, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria
Rauscher T., Heger A., Hoffman R. D., Woosley S. E., 2002, ApJ, 576, 323
Rösler R., 2012, Matching clustering solutions using the 'Hungarian method'
Schliep K., Hechenbichler K., 2016, kknn: Weighted K-Nearest Neighbours. Available at: https://CRAN.R-project.org/package = kknn
Scrucca L., Fop M., Murphy T. B., Raftery A. E., 2016, R J., 8, 289
Silverman J., 2019, Rcpphungarian: Solves Minimum Cost Bipartite Matching Problems. Available at: https://CRAN.R-project.org/package = RcppHungarianS
Stephan T. et al., 2018, Geochim. Cosmochim. Acta, 221, 109
Stephan T. et al., 2020, in Lunar and Planetary Science Conference. Available at: https://www.hou.usra.edu/meetings/lpsc2021/pdf/2358.pdf
Stephan T. et al., 2021, in Lunar and Planetary Science Conference, Lunar and Planetary Science Conference. Available at: https://pre-solar.physics.wustl.edu/pre-solar-grain-data base/
Strehl A., Ghosh J., 2002, J. Mach. Learn. Res., 3, 583
von Luxburg U., 2007, Stat. Comput., 17, 395
Wang N., Raftery A., 2017, covrobust: Robust Covariance Estimation via Nearest Neighbour Cleaning. Available at: https://cran.r-project.org/web/packages/covRobust/covRobust.pdf
Wang H., Shan H., Banerjee A., 2011, Stat. Anal. Data Min.: ASA Data Sci. J., 4, 54
Wickham H., 2016, ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York
Wickham H., François R., Henry L., Müller K., 2021, dplyr: A Grammar of Data Manipulation. Available at: https://CRAN.R-project.org/package = dplyr
Zinner E., 2014, Treatise on Geochemistry, 2nd edn. Elsevier, Oxford, p. 181

This paper has been typeset from a TEX/LATEX file prepared by the author.