PosgenPy: An Automated and Reproducible Approach to Assessing the Validity of Cluster Search Parameters in Atom Probe Tomography Datasets

Przemysław Klupś[1*], Daniel Haley[1], Andrew J. London[2], Hazel Gardner[1], James Famelton[1], Benjamin M. Jenkins[1], Jonathan M. Hyde[3], Paul A.J. Bagot[1], Michael P. Moody[1].

[1]Department of Materials, University of Oxford, Oxford OX1 3PH, UK

[2]UK Atomic Energy Authority, Culham Science Centre, Abingdon OX14 3DB, UK

[3]National Nuclear Laboratory, Culham Science Centre, Abingdon OX14 3DB, UK

*corresponding author; e-mail address: przemyslaw.klups@materials.ox.ac.uk

1 Abstract

One of the main capabilities of Atom Probe Tomography (APT) is the ability to not only identify, but also characterise early stages of precipitation at length-scales that are not achievable by other techniques. One of the most popular methods to identify nanoscale clustering in APT data, based on the Density-Based Spatial Clustering of Applications with Noise (DBSCAN), is used extensively in many branches of research. However, it is common that not all of the steps leading to selection of certain parameters used in the analysis are reported. Without knowing the rationale behind parameter selection, it may be difficult to compare cluster parameters obtained by different researchers. In this work, a simple open-source tool, PosgenPy, is used to justify cluster search parameter selection via providing a systematic sweep through parameter values with multiple randomisations to minimise false-positive cluster ratio. The tool is applied to several different microstructures: a simulated material system; two experimental datasets from a low-alloy steel and one from an Fe-9Cr alloy. The analyses show how values for the various parameters can be selected to ensure the calculated cluster number density and cluster composition are accurate.

Key words: atom probe tomography, cluster analysis, parameter selection, DBSCAN, maximum separation method, core-linkage, solute precipitation, MNSP, CRP,

2 Introduction

Atom Probe Tomography (APT) has proven to be useful in identifying solute atom clusters and precipitates at length scales that are not readily available via other characterisation techniques. The effect of solute atom clustering is a subject of interest in many materials, including metals, ceramics, and geological materials (Marquis & Hyde, 2010; Miller & Burke, 1992; Pareige, Auger, Bas, & Blavette, 1995; Reddy et al., 2020; Stiller, Thuvander, Povstugar, Choi, & Andrén, 2016). Methods available to identify solute atom clustering include amongst others; frequency distribution analyses (Eric Camus & Christian Abromeit, 1994; E. Camus & C. Abromeit, 1994; Langer, Bar-on, & Miller, 1975; Moody, Stephenson, Liddicoat, & Ringer, 2007), iso-concentration surfaces (Hellman, Vandenbroucke, Rüsing, Isheim, & Seidman, 2000), and cluster search algorithms (Jonathan M. Hyde & English, 2000; Stephenson, Moody, Liddicoat, & Ringer, 2007; Vaumousse, Cerezo, & Warren, 2003; Wang, Schreiber, Bailey, Hosemann, & Toloczko, 2019). Iso-concentration surfaces are more readily used on larger clusters (with diameter greater than 5 nm), where visual inspection is able to distinguish the locally-enriched regions from the matrix, than in the systems with much smaller clusters. This study focuses on the algorithms based on Density-Based Spatial Clustering of Applications with Noise (DBSCAN) first developed by Ester et al. (Ester, Kriegel, Sander, & Xu), which despite their relative simplicity are still considered to be widely used, effective, and performant in processing complex data (Schubert, Sander, Ester, Kriegel, & Xu, 2017).

Each cluster identification method has intrinsic parameters that define the cluster search, and the user must currently select an appropriate value for each of these parameters, although there is ongoing interest in the development of machine learning approaches to minimise interaction with the operator (Zelenty, Dahl, Hyde, Smith, & Moody, 2017) or the use of other, more refined, cluster search algorithms such as OPTICS (Wang et al., 2019) or

HDBSCAN (Ghamarian & Marquis, 2019) that minimise the number of required input parameters, alongside other advantages. Nonetheless, selection of parameter values is often subjective and user-dependent as well as hypothesis-dependent. If not reported, it can adversely affect the reproducibility of reported cluster data from individual researchers. Ultimately, this means that comparisons between studies can be associated with significant uncertainty or are simply impossible.

For nearest neighbour type algorithms, the main cluster identification parameters can be divided into three groups. The first group includes the parameters that identify which ions can form the skeleton (or core) of each cluster. Parameters in this category include definition of the core ions, which are the solute elements considered to form these clusters; d_{max}, which is the maximum physical separation of core ions considered to be clustered (Hyde & English, 2000); and Order (or KNN) which is the minimum number of the core cluster ions required within d_{max} (Stephenson et al., 2007). Solute ions that themselves do not satisfy the Order criteria but are within d_{max} of any that do, are so-called border points, and are also considered to be a part of the cluster.

The second group of parameters are responsible for adding non-core ions to clusters so that elements from the matrix can be included in the cluster composition. This is also called the *enveloping* procedure. All ions within a user-defined radius, L, of each core ion are included in the cluster (Vaumousse et al., 2003). This method, however, may lead to the erroneous result whereby the defined clusters incorporate a layer of the matrix ions around the surface. To counteract any over-inclusion, an *erosion* step is subsequently applied such that every non-core ion within a user-defined radius, $d_{erosion}$, of each matrix ion may subsequently be removed from the cluster (Vaumousse et al., 2003).

The last parameter can be applied post-processing and provide a cluster size threshold, N_{min}, which is the smallest possible number of solute ions in a cluster.

Minor differences in parameter choices can lead to drastic changes in the count and nature of identified clusters. This parameter sensitivity makes it difficult to directly compare clusters populations analysed by different users, as demonstrated by Marquis et al. (2019). This prior study emphasises the need to carefully document parameterisation of the methods, and how they are selected, in addition to providing comparisons with complementary randomised datasets to ensure reported results are both accurate and meaningful.

Additionally, it is challenging even for the same user to compare clustering results across different datasets. Parameters chosen to optimise the analysis of one dataset cannot necessarily be appropriately translated across to other datasets. Experimental conditions used for APT can vary significantly, even for the same material, and as such affect the results (De Geuser & Gault, 2020). One example of this is how varying emission field correlates with the trajectory aberrations of solute elements and consequently increases or decreases apparent density in the reconstruction (Miller & Hetherington, 1991). Additionally, lateral scatter in the reconstructed ion positions introduces artefacts into the APT data which, together with statistical fluctuations of the positions of the solute ions, can lead to small groups of solute atoms being incorrectly detected as “clusters”. We will refer to these clusters as “false clusters”. In contrast, actual clusters of atoms arising from the original sample from some real feature such as a particle or solute enrichment will be referred to as “true clusters”. The exact number of false clusters cannot be known in the experimental dataset. However, the relative proportions of true and false clusters for a particular set of cluster search parameters can be estimated by an extra step of swapping the mass-to-charge state values of every ion in the reconstruction while maintaining their original spatial coordinates (Styman, Hyde, Wilford, & Smith, 2013). The effect of this is to create a complementary randomised dataset that can be directly compared with the original reconstruction. The cluster analysis is then re-applied to this random comparator. Using this comparison, the cluster search parameters can be varied, in a justified way, e.g., to maximise the ratio of true to false clusters. This assumes that having high certainty that the detected clusters are real clusters is preferred over including all real clusters in the analysis. It is important to understand that other type of false negative results can arise, in which the algorithm fails to detect true clusters in the dataset due to poor parameter selection.

Mapping the parameter space has the potential to reduce user-dependency by highlighting a range of justifiable parameters, or even to show whether appropriate search parameters that can neatly delineate the clusters from the matrix actually exist. Often it is possible to find regions of parameter space where the calculations are relatively insensitive to small changes in the parameter values giving confidence that appropriate values have been identified. However, with current methods, it can be time consuming to sweep the cluster search parameter space. This makes it difficult to systematically explore the parameter space and justify the parameter choice. This barrier-to-entry is often enough to dissuade users from performing the parameter sweep necessary to justify their cluster selection parameters or determine the uncertainty in their calculated results. Therefore, there is a need for a tool which facilitates parameter selection and enables reproducible results to be obtained.

In this paper, a suite of Python scripts for mapping and systematic selection of the cluster search parameters is presented. The effect each parameter has on the cluster distribution detected in a dataset is determined and randomisations of the dataset are used to exclude the contribution from false clusters. This allows a user to make a more informed decision on the cluster search parameters to be implemented. This approach has clear and important applications for atomic-scale characterisation across a wide variety of materials and research fields. By making the justification and reporting of the cluster search process easier, it also facilitates the ability of user to apply the FAIR (findability, accessibility, interoperability, and reusability) principles (Wilkinson et al., 2016).

3 Materials and Methods

3.1 PosgenPy Method

PosgenPy uses the “Posgen” program (D. Haley, 2012) to implement a neighbour-based clustering algorithm (Stephenson et al., 2007). PosgenPy then implements high-level decision-making for cluster data interpretation. In order to compare each of the cluster searches to the relabelled dataset, PosgenPy uses a multiple random mass relabelling strategy, where multiple randomised datasets are prepared, and cluster analysis is applied. This enables estimation of the effects of inhomogeneities in the material due to fluctuations in density caused by imperfect spatial resolution, as well as the effect of statistical fluctuations in the distribution of atoms. The mean number of the clusters in the relabelled data (false positives) is used to find the ratio of real clusters to false ones (Williams, Haley, Marquis, Smith, & Moody, 2013), called the true positive rate (TPR) which is given by:

$$TPR = \frac{N_{detected} - \overline{N_{random}}}{N_{detected}} \qquad \text{Eq. 1}$$

where $N_{detected}$ is the number of identified clusters in the dataset, and $\overline{N_{random}}$ is the average number of false positive clusters detected in the randomised datasets. By using a specified value for TPR as a threshold, N_{min} can be parameterised such that the threshold is achieved by excluding smaller clusters which usually form a greater fraction of clusters detected in the randomised datasets than in the real datasets.

The next stage is to systematically repeat these steps through varying values of a given parameter, combine all results from the iterated cluster searches and summarise the information collectively. Similar approaches were investigated by various studies (Hyde et al., 2017; Hyde, Marquis, Wilford, & Williams, 2011; Jägle, Choi, & Raabe, 2014; Styman et al., 2013; Williams et al., 2013). At this point, the user can see how cluster number density (N_d), TPR, and average composition of the solute in the clusters are affected by the set of values for one parameter; and optimise this parameter by finding justifiable values. Such values should lie in so-called "low-gradient regions", meaning small variations in the parameter value do not result in a significant change in the results. If, after initial iteration through parameter values, the results do not change significantly as a function of small changes in the input parameter, this indicates local stability, and the parameter values can be considered as locally justified. The workflow of PosgenPy is also presented in Fig 1. In the case of more than one local stability due to varying density or different types of clusters, it is recommended to either undertake separate cluster searches for different types of clusters, or consider implementing other potentially more effective cluster search algorithms for the current circumstances, such as OPTICS or HDBSCAN (Ghamarian & Marquis, 2019; Wang et al., 2019). Both of these methods allow for detection of clusters with different densities as well as consider the non-uniformities such as crystallographic poles in the APT reconstructions.

3.2 Datasets

In this paper, three case studies are used to demonstrate the capabilities of PosgenPy. The first one consists of a simulated dataset, which can provide a comparison to the correct values that are known prior to any analysis. The other two analyses involve a low-alloy steel after two different thermal ageing treatments.

The purpose of the simulated dataset is to construct a "ground truth" of known cluster composition, size, and count, which can be used to later inform the analysis of real datasets. In this work, the simulated dataset, shown in Fig 2, was created using ClusterAlphaEdge repository (London, 2019) and contains 100 non-overlapping clusters with 50:50 ratio of Ni and Cu in the Fe matrix resulting in an average compositional uncertainty in each cluster of 1.5 at. %. This dataset also contains 17 face and 3 edge clusters that are smaller and harder to identify than the other 80 in the volume as some parts of them do not lie within the boundary box of the dataset. The true average cluster size is centred around 90 atoms excluding the face and edge clusters. The boundaries of this dataset form a cube with 40 nm edges and contain 1.9 million ions which equates to the density of 30 ions per nm^3. More information about the methods used for creating this simulated this dataset can be found in the study by Jenkins et al. (Jenkins et al., 2020).

Two experimental datasets of low-alloy steel will be used to show how PosgenPy is applicable to real data relevant to the nuclear industry. As this alloy has a high content of copper, as well as other solutes, and was aged for 10,000 hours and 30,000 hours, significant clustering at various stages of nucleation and growth is expected (Odette, Yamamoto, Williams, Nanstad, & English, 2019). Both datasets were reconstructed using IVAS 3.8.6. Crystallographic poles were used for the justifiable selection of the initial tip radius and k-factor assuming ICF = 1.65. The APT sample from the alloy aged for 10,000 hours is shown in Fig 2 with different elements crucial in nucleation and formation of Cu-rich Precipitates (CRPs) and Mn-Ni-Si-rich Precipitates (MNSPs) presented in separate images. In the Cu atom map on Fig 2.d regions locally enriched in Cu can be observed. It is much more difficult to identify discrete clusters visually when looking at the atom maps for other elements in Fig 2.a-c. The second APT sample from the alloy aged for 30,000 hours is presented in Fig 3. In contrast to Fig 2, numerous clusters can be observed for all elements in the atom maps of Fig 3. Therefore, the 10,000-hour-aged sample is a more subtle, challenging case for selecting appropriate cluster search parameters and will be discussed subsequently after the 30,000-hour-aged steel.

4 Results

In all of the analyses presented below, a number of initial assumptions were made. The enveloping (L) and erosion ($d_{erosion}$) steps were used and their values were set at half of the d_{max} used in the given analysis as it was shown to result in the compositions close to the simulated ones in previous studies (Hyde et al., 2011; Jägle et al., 2014). Hence, L is the smallest value that allows for including all the bulk ions that lie between the core ions and therefore count most of the bulk ions inside of the clusters. An equal value for $d_{erosion}$ is required to remove the over-inclusion of the bulk ions around each cluster. Bulk ions were defined to be all of the elements in the nominal composition of the materials; Ga was excluded as it is a residue from the sample preparation via focused ion beam (FIB). 20 randomisations were found to be sufficient to identify accurately false clusters. For each order and d_{max} sweep presented in this study, a value for order and d_{max} had to be estimated prior to the analysis. These values were based on the preliminary PosgenPy analyses of d_{max} on order = 1 as well as often reiterated order sweeps with d_{max} values taken from the former. After finding the justifiable parameter range for d_{max} at order = 1, this value was used for further order and d_{max} sweeps presented below.

4.1 Simulated Dataset

For the simulated dataset, two separate analyses of the PosgenPy analyses were conducted. At first, it iterated through different values for order (2-10) at a constant d_{max} value of 0.5 nm (Fig 5. a-c). In Fig 5.a, the number density for each Order is shown. As expected, the N_d count decreases with increasing order as well as with increasing N_{min}. It is also important to notice that N_{min} affects N_d count if N_{min} > order. There is a small exception in the case when N_{min} < order and some of the clusters share the same border ions that will have to be grouped to only one of them; both clusters are then identified, even though in one of them there are fewer solute ions than specified in the order parameter. That is why results for different N_{min} values are the same as N_{min} = 2 for order = N_{min}.

For order smaller than 5, N_d count overestimates the number of clusters due to many false positives, which can be seen from Fig 5.b. As order values exceed 5 or N_{min} is increased to 10, N_d count is underestimated but the false positive ratio is low. This can be explained by the fact that some of the clusters in the simulated data are placed on the edge or border of the dataset, and therefore are more difficult to distinguish from the false positive clusters due to their smaller size.

In Fig 5.b, the TPR for each value is presented. For the lower order values, it is shown that an increased N_{min} is required for adjusting the cluster selection. Otherwise, the number of false positive clusters exceeds the assumed threshold. Justified values of N_{min} for each Order were selected so that TPR for each swept parameter exceeds 95%. The concentration of core-ions in the clusters is plotted in Fig 5.c. The composition for both Ni and Cu is below the true value of 50 at. %. In contrast to the N_d count, with increasing order, the concentration approaches the true value. This, therefore, represents the trade-off between accurate N_d count and accurate cluster concentration. It is also worth pointing out that the cluster composition is close to the true value whereas N_d count can vary by orders of magnitude if parameters are chosen unwisely. Using Fig 5.b, one can assume that for the analysis of the dataset in order to minimise use of other parameters such as N_{min}, in this instance the order selected should be higher than 4. As none of the other values drastically changes the results in these three graphs, the range of 5-10 should be a reasonable choice for the further analysis of this dataset.

To further understand the effect of the cluster search parameters on this dataset, d_{max} values were incremented from 0.20 nm to 0.70 nm in 0.02 nm steps (Fig 5.d-f). Using the previous results, the order for this iteration was fixed to be 6 core-ions. As in Fig 5.a, Fig 5.d shows rapidly changing N_d count for small values of d_{max} as small clusters are identified at $d_{max} < 0.25$ nm which are removed by increased N_{min}. Another interesting feature is the peak at 0.32 nm which drops for all N_{min} values at 0.44 nm. This was a result of larger clusters being erroneously identified as several separate smaller clusters, joining at $d_{max} > 0.44$ nm. For 0.44 nm < d_{max} < 0.60 nm, the N_d count is slightly underestimated when compared to the true Nd of the simulated dataset, due to the small number of edge and border clusters as in Fig 5.a. For larger d_{max} values and $N_{min} = 2$, the N_d count starts increasing again suggesting that small clusters are incorrectly included in the analysis. This effect can be explained by an increased chance of finding false positive using these distances. It can be also confirmed by looking at the drop in the TPR for these d_{max} values in Fig 5.e.

However, Fig 5.f shows that for d_{max} values greater than 0.4 nm, both Cu and Ni solute compositions are underestimated relative to the true value. Contrary to the order, higher d_{max} correlates with lower, less-accurate concentrations of the core ions. This can be explained by more bulk ions being included in the clusters due to both slightly higher number of core ions in the clusters as well as proportional enveloping parameter L which equals to $\frac{d_{max}}{2}$. In contrast to the order effect, there is no trade-off between the N_d count and accurate concentration of the core ions.

The trend of the concentration being closer to the true value for higher order is, however, caused by higher density of the solute required for the same d_{max}. For the same total volume of clustered ions (d_{max} is constant), a higher number of solute ions is required for them to be included. Many of the smaller clusters are rejected due to the order = N_{min} property. Out of all the other clusters that are present, only their denser cores that also have higher concentrations of solute ions are identified by the algorithm whilst concentration of bulk ions is lowered.

This leads to decreasing underestimation of the overall cluster concentration with increasing order values.

Considering these results, the d_{max} values in the range of 0.44 nm and 0.6 nm can be considered as justifiable selections. Even though the concentrations plateau for higher d_{max} this should not be taken as evidence of these parameters as better choices; in order to keep the concentration as close to the true value as possible, one might want to minimise this impact by choosing the smallest reasonable d_{max} value, which in this case is equal to 0.44 nm.

4.2 Low-alloy Steel Aged for 30,000h

The first material used in the case study of PosgenPy is a low-alloy steel with high content of copper, thermally aged for 30,000 hours at 330°C. As shown in Fig 4, Cu-rich clusters are visible in the APT atom maps.

Similarly to the simulated dataset, there were two PosgenPy analyses made on the steel aged for 30,000 hours to investigate the effect of order and d_{max}, respectively, on the N_d count, TPR, and solute composition. For the former, d_{max} of 0.50 nm was chosen whereas for the latter, the Order of 7. In both cases Mn, Ni, Si, Cu, and P were considered to be solute elements; the rest of the elements were considered as bulk. The values for the order were in the range of 2 and 10 and d_{max} was varied between 0.20 and 0.70 nm with increments of 0.02 nm.

As in the previous case study, Fig 6.a shows there is a negative correlation between the order and N_d count, which together with its high sensitivity to N_{min} and the TPR graph in Fig 6.b implies a large number of small false positive clusters are detected. The ratio does not reach the 95% threshold for $2 \leq \text{order} \leq 5$ values, even at $N_{min} = 50$. High N_{min} is also needed for the rest of the order values in order to increase TPR.

The steep decrease in the Ni composition in Fig 6.c is due to the program not being able to find a suitable N_{min} that would satisfy the 95% threshold at order < 5. This means that only the order values from 5 onward, should be considered. As in Fig 5.c, Fig 6.c also shows a small positive correlation between the order value and composition of solutes.

The graph in Fig 6.d shows the effect of d_{max} on measured N_d count. For smaller values between 0.20 and 0.40 nm N_d count significantly increases from which point it plateaus before increasing again for d_{max} values above 0.6 nm. There is also a significant sensitivity to the N_{min} parameter, which is less prominent in the plateau region, implying this range of values should be considered for the further inspection. Fig 6.e shows the TPR for these values. The threshold is satisfied for the first range of values up to d_{max} = 0.40 nm. This is due to the smaller number of detected false-positive clusters in this region, compared to the higher number of detected false-positive clusters detected at high values of d_{max}. For $0.40 \leq d_{max} \leq 0.60$ the TPR starts do decrease for smaller N_{min} and from d_{max} = 0.60 nm, a justified N_{min} value cannot be found as the ratio of false positive clusters is too high.

The composition in Fig 6.f shows decreasing concentration of all the solutes with increasing d_{max} as, similarly to the simulated dataset, more bulk atoms are included in the clusters due to $L = d_{erosion} = \frac{d_{max}}{2}$ relationship. As justified N_{min} values could not be found for $d_{max} > 0.60$ nm, the composition of solutes is higher due to higher number of smaller false-positive clusters included in the composition calculations. It is, therefore, suitable to select d_{max} values for further analysis in the range of 0.40 nm to 0.60 nm.

4.3 Low-alloy Steel Aged for 10,000h

The next case study materials for PosgenPy is the same low-alloy steel as in the previous example, but at the earlier stages of thermal ageing. This sample was subjected to 330°C for 10,000 hours and shows much smaller and less clearly defined clusters than its 30,000 hours counterpart. Two PosgenPy analyses were run to see the effect of Order (2-10) and d_{max} (0.20 to 0.70 nm at 0.02 nm increments). The d_{max} of 0.50 nm and order of 7 were used respectively. The same solutes were chosen. To further emphasise the main advantage of the PosgenPy, which is a systematic iteration through any DBSCAN cluster search parameter, another sweep through $d_{erosion}$ = L (0.0 to 0.50 nm with 0.05 increments) was included.

Whilst Fig 7 looks initially similar to Fig 6, there are key differences apparent on closer inspection. The first difference is a lack of plateau regions for the Nd count in both Fig 7.a and Fig 7.d. There is a slight change in the gradient of the slopes between 0.46 nm and 0.54 nm and similar sensitivity to the Nmin parameter. The TPR in Fig 7.b shows an earlier increase at lower order values (8 compared to 10), and similarly in Fig 7.e the TPR drops below the threshold for lower dmax values at 0.30 nm. These results imply that the clusters found in this dataset are much harder to distinguish from the APT artefacts and stochastic nature of the solute atoms due to their smaller size and number density. Even at more conservative ranges of the parameters, the ratio of the false positive clusters is high. The graphs investigating the effect of these parameters on the composition of the solutes show similar trends. The clear differences between Fig 7.c and Fig 6.c, and Fig 7.f and Fig 6.f are the rates of change of the composition in regard to Order and dmax respectively. Whereas the influence of order on solute concentration is more prominent in 10,000-hours-aged sample, the dmax affects it to a lesser extent. This might be caused by a combination of different Nmin values selected for each data point, smaller solute density of the clusters in the early stages of ageing, and smaller clusters sizes, which also correlate with fewer bulk atoms. Considering all the graphs in the Fig 7, it is not a straightforward task to justify the selection of a region of reasonable values for further analysis. However, it can be determined that values for an order lower than 5, and values of dmax lower than 0.40 and higher than 0.55 nm would not be recommended. To investigate how stable the cluster analyses are with these parameter values, another sweep through derosion = L was undertaken. As expected, this did not result in any changes in the number density. There were small insignificant fluctuations in the TPR. The cluster composition remained stable for values equal or greater to half of dmax used. The latter being stable mainly due to high order value as it has been already noted in other studies that the concentration is affected by enveloping and erosion parameters for lower order values (Jägle et al., 2014; Marquis et al., 2019).

5 Discussion

5.1 Number Density

When calculating N_d, there are two aspects to consider. Firstly, have all the real clusters been correctly identified (i.e., individually) and secondly have the false clusters been successfully eliminated.

In the first part, d_{max} and the order are the key parameters and here the trade-off between the values of the two parameters must be considered. Increasing the order will also require increasing d_{max} (and vice-versa). In practice, what is required is the lowest d_{max} for a given order, or the highest order for a given d_{max}, that accurately characterises the microstructure. Either of these combinations will enable the cluster edges to be most clearly defined. Increasing d_{max} unnecessarily will result in solute atoms in the matrix being identified in the clusters. The same will occur when decreasing the order at a given d_{max}. The challenge is to identify the lowest appropriate dmax for a given order, or the highest appropriate order for a given d_{max}. This is exactly the type of analysis that is possible from the data presented in Fig 5 to Fig 7. For instance, on the simulated data, with a dmax of 0.5 nm and order 10 the correct N_d is identified. With an order of 6, and a d_{max} of 0.4 nm also yields the correct answer. However, reducing dmax further either results in a higher N_d as single clusters are characterised as several clusters (using a small Nmin) or lower N_d (using a larger Nmin which excludes many of the clusters solute atoms). This latter point also illustrates the importance of selecting N_{min} appropriately as discussed below.

5.2 True Positive Rate

The TPR plots are extremely useful in identifying the smallest N_{min} that still yields statistically significant N_d results. It was also used in previous studies by Williams et al. (Williams et al., 2013). In this work a TPR of 95% was selected, but the precise choice will depend on the metallurgical system under investigation and the perceived importance of characterising the smallest clusters (in some systems the material properties may be dominated by the larger clusters – and so it may be better to choose a higher threshold, in others the very small clusters may be more influential).

The TRP is determined by randomising the dataset multiple times and, after each randomisation, identifying any clusters using the selected parameters. Small clusters will exist in a random solid solution, but the precise number identified will depend on the exact distribution of atoms and without repeating the randomisation it is not possible to determine the mean accurately. The advantages of multiple random relabellings are detailed in work by Styman et al. (Styman et al., 2013).

Increasing the N_{min}, however, cannot always provide a high TPR – in some cases the N_d count of randomised dataset is equal or higher than N_d count in the real dataset, and there are no ranges of parameters available for the TPR to satisfy the threshold. This could also potentially imply that such a dataset does not contain statistically significant clustering. If the true solute distribution in the dataset is random to begin with, then ($N_{detected}$ – N_{random}) will be zero, as randomisation will have no effect on data entropy. It can occur for both a fraction and full ranges of the parameters.

5.3 Concentration of Solute Ions

Once the TPR is used to find the justified N_{min} for each swept parameter, the solute concentrations for each parameter can be shown. PosgenPy enables fast, systematic analysis that is necessary to highlight the trade-off between accurate cluster composition and accurate number density. Even though the higher order values provide concentrations closer to the true value, the underestimation of N_d count increases as the order increases. The user can then use this information to select the parameters most suitable for their analysis.

Similarly to the N_d graphs, the usefulness of the composition graphs comes from identifying any anomalies which do not fit our expectations such as ones described above. An example of that occurs in extreme values for the lowest d_{max} 0.2-0.3 nm in Fig 5.f. The composition difference between Ni and Cu reaches 4 percentage points: 48% to 52% whereas the true value is 50%:50%. By comparing it with the same parameters in other graphs (Fig 5.d-e), one can assume that the size of the clusters which are identified for 0.2 nm d_{max} are limited, as the N_d count is much lower, and they do not include any bulk ions as the sum of both compositions equals 100%. Another key issue is the systematic difference between the composition levels of Ni and Cu. Further analysis of this phenomenon showed that the systematic difference of about 2% will occur in any individual simulated dataset using this method, due to random fluctuations. It is, therefore, an artefact of the simulation and the true value for the concentration is 49% Cu to 51% Ni.

The steel case studies in Fig 6.f and Fig 7.f highlight the benefits of the systematic analysis enabled by PosgenPy. At lower d_{max} values, the average Cu concentration in the clusters is higher as only the high-density core regions are selected. It is worth noting that due to trajectory aberrations, ions that lie in the centre of these clusters seem denser in the APT reconstructions. This effect is due to the preferential evaporation of the clustering solute ions at the surface leading to local magnification trajectory aberrations and this effect is well documented in these systems (De Geuser & Gault, 2020). This would imply that by systematic changes in d_{max}, the change in composition of solute elements with size of the clusters can be identified. In this case, more Cu atoms are present in the core, which is in an agreement with the general findings of the RPV (reactor pressure vessel) community on the Cu-rich precipitates (CRP) (Odette et al., 2019). This is consistent with previously identified mechanisms that describe Cu taking a critical role in precipitating and nucleating first, ahead of other elements in the steel, due to its lower solubility in Fe. Cu clusters then create heterogeneous nucleation sites for other solute elements, therefore hastening the MNSP nucleation. This only emphasises the importance of lowering the Cu content in these steels as both CRPs and MNSPs are the crucial factor in hardening embrittlement of these steels.

The PosgenPy method, however, cannot be used as an alternative to proxigrams, which are commonly used along with the isoconcentration surface method for cluster analysis. As one can expect such results while knowing the structure of the cluster in advance, there are possible scenarios where other separate cluster with different compositions could affect this metric in the same way and the user would not be able to distinguish between them by looking at this data. Similar conclusions of the difficulty in finding reproducible compositional values in cluster search were previously reported (Dong et al., 2019; Marquis et al., 2019).

5.4 General recommendations

PosgenPy systematically evaluates the effect of cluster search parameters on the N_d, TPR, and solute composition, enabling more informed parameter selection. The fully automated approach saves time (e.g., an average sweep used in this study of 7 Order values and 20 randomisations took approximately 15 minutes during which no user interaction was required) and enables analyses that would not be possible with current commercially available software due to prohibitive run times. The user is advised to select a range of values to be swept that encompass the expected range of justifiable values in order to make sure that the full picture is seen in the dataset.

As one of the parameters, N_{min}, can be adjusted after the analysis, it is recommended to experiment and use much higher number of N_{min} values than in this study (they were omitted for the sake of clarity of the figures). There is also a trade-off between the binning, or number of cluster searches, and time saving aspect of PosgenPy, and it is up to the user to choose the optimal strategy for each dataset; either high binning from start, or two-step verification with smaller increments and range in the second stage.

Although this methodology helps to justify and to some extent optimise the choice of parameters, it cannot guarantee that only the result ranges of values are correct; there will be greater uncertainty, however, on the significance of the cluster analysis results. Analogously, the justified choice of parameters can only provide the most accurate answer given the methodology, but it still might be different from the ground truth.

6 Conclusions

In this study, a method for systematic assessment of the parameters used for the density-based cluster search, PosgenPy, was presented. The results were shown for a simulated dataset with known N_d count and composition of clusters in order to compare and verify how PosgenPy can be used to find suitable ranges of parameters. To prove its applicability in experimental data, two datasets from low-alloy steels subjected to 10,000 and 30,000 hours of thermal ageing at 330°C were studied and the results discussed. The thermally-aged materials are of particular interest to the nuclear industry as steels with similar compositions are known to undergo hardening embrittlement (Michael K. Miller & Russell, 2007). Accurate characterisation and quantification of clusters in these alloys allows for more accurate models and predictions of the lifetime of structural nuclear components to be made; this is crucial from both an economical and safety standpoint.

To conclude, PosgenPy:

- automatically iterates through multiple values of a given cluster analysis parameter and systematically saves the results for the later comparison while emphasising a strong mass relabelling randomisation for accurate estimation of potentially false clusters arising from the fluctuations and artefacts of the dataset
- demonstrates how significantly N_d, TPR, and solute composition all change with respect to changes in the value of a given parameter, while using TPR to optimise N_{min}
- through the case studies, it displays the opposite effects of increasing d_{max} or decreasing Order on increasing N_d via including more clusters; decreasing TPR through including more clusters in the randomised datasets; and increasing underestimation of the solute composition in clusters via including more bulk ions
- facilitates finding a so-called “low-gradient region” for the cluster analysis by either pointing out a range of parameter values in the plateau region or showing that such an optimal range cannot be found and there is no optimal value
- highlights trade-offs and/or potential conflicts in cluster analysis parameter selection when selecting parameters to optimise, for example, N_d and TPR, and focuses on providing an overview of how significantly the parameter selection is influencing final results
- is also an open and extensible solution that allows for improved reproducibility and consistency within the solute clustering research in APT which also facilitates application of FAIR principles (Wilkinson et al., 2016) in the APT community.

The code can be found in the GitHub repository:

https://github.com/PrzemyslawKlups/posgenpy

7 Acknowledgements

We are thankful to Rolls-Royce Plc. for providing the necessary material for the case study. The atom probe facilities at the University of Oxford are funded by the EPSRC (EP/M022803/1). The authors acknowledge use of characterisation facilities within the David Cockayne Centre for Electron Microscopy, Department of Materials, University of Oxford, alongside financial support provided by the Henry Royce Institute (Grant ref EP/R010145/1). The research used UKAEA’s Materials Research Facility, which has been funded by and is part of the UK’s National Nuclear User Facility and Henry Royce Institute for Advanced Materials. PK would like to acknowledge the financial support of EPSRC and Rolls-Royce Plc. BMJ would like to acknowledge funding from EPSRC program grant MIDAS (EP/S01702X/1). AJL acknowledges the EPSRC Grant (EP/T012250/1).

8 References

Camus, E., & Abromeit, C. (1994). Analysis of conventional and three-dimensional atom probe data for multiphase materials. *Journal of Applied Physics, 75*(5), 2373-2382. doi:10.1063/1.356258

Camus, E., & Abromeit, C. (1994). Correlation and contingency analysis of atom probe data: Diffusion-controlled dissolution of precipitates.

De Geuser, F., & Gault, B. (2020). Metrology of small particles and solute clusters by atom probe tomography. *Acta Materialia, 188*, 406-415. doi:https://doi.org/10.1016/j.actamat.2020.02.023

Dong, Y., Etienne, A., Frolov, A., Fedotova, S., Fujii, K., Fukuya, K., . . . Marquis, E. A. (2019). Atom Probe Tomography Interlaboratory Study on Clustering Analysis in Experimental Data Using the Maximum Separation Distance Approach. *Microscopy and Microanalysis, 25*(2), 356-366. doi:10.1017/S1431927618015581

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise*.

Ghamarian, I., & Marquis, E. A. (2019). Hierarchical density-based cluster analysis framework for atom probe tomography data. *Ultramicroscopy, 200*, 28-38. doi:10.1016/j.ultramic.2019.01.011

Haley, D. (2012). Posgen. Retrieved from https://sourceforge.net/projects/apttools/files/

Hellman, O. C., Vandenbroucke, J. A., Rüsing, J., Isheim, D., & Seidman, D. N. (2000). Analysis of Three-dimensional Atom-probe Data by the Proximity Histogram. *Microscopy and Microanalysis, 6*(5), 437-444. doi:10.1007/S100050010051

Hyde, J. M., DaCosta, G., Hatzoglou, C., Weekes, H., Radiguet, B., Styman, P. D., . . . Pareige, P. (2017). Analysis of Radiation Damage in Light Water Reactors:

Comparison of Cluster Analysis Methods for the Analysis of Atom Probe Data. *Microsc Microanal, 23*(2), 366-375. doi:10.1017/S1431927616012678

Hyde, J. M., & English, C. A. (2000). An Analysis of the Structure of Irradiation induced Cu-enriched Clusters in Low and High Nickel Welds. *MRS Proceedings, 650*, R6.6. doi:10.1557/PROC-650-R6.6

Hyde, J. M., Marquis, E. A., Wilford, K. B., & Williams, T. J. (2011). A sensitivity analysis of the maximum separation method for the characterisation of solute clusters. *Ultramicroscopy, 111*(6), 440-447. doi:https://doi.org/10.1016/j.ultramic.2010.12.015

Jägle, E. A., Choi, P.-P., & Raabe, D. (2014). The Maximum Separation Cluster Analysis Algorithm for Atom-Probe Tomography: Parameter Determination and Accuracy. *Microscopy and Microanalysis, 20*(6), 1662-1671. doi:10.1017/S1431927614013294

Jenkins, B. M., London, A. J., Riddle, N., Hyde, J. M., Bagot, P. A. J., & Moody, M. P. (2020). Using alpha hulls to automatically and reproducibly detect edge clusters in atom probe tomography datasets. *Materials Characterization, 160*, 110078.

Langer, J. S., Bar-on, M., & Miller, H. D. (1975). New computational method in the theory of spinodal decomposition. *Physical Review A, 11*(4), 1417-1429. doi:10.1103/PhysRevA.11.1417

London, A. (2019). Cluster Alpha Edge Retrieved from https://github.com/andyroo101/Cluster-Alpha-Edge

Marquis, E. A., Araullo-Peters, V., Dong, Y., Etienne, A., Fedotova, S., Fujii, K., . . . London, A. (2019). *On the use of density-based algorithms for the analysis of solute clustering in atom probe tomography data.* Paper presented at the Proceedings of the 18th International Conference on Environmental Degradation of Materials in Nuclear Power Systems–Water Reactors.

Marquis, E. A., & Hyde, J. M. (2010). Applications of atom-probe tomography to the characterisation of solute behaviours. *Materials Science and Engineering: R: Reports, 69*(4), 37-62. doi:https://doi.org/10.1016/j.mser.2010.05.001

Miller, M. K., & Burke, M. G. (1992). An atom probe field ion microscopy study of neutron-irradiated pressure vessel steels. *Journal of nuclear Materials, 195*(1), 68-82. doi:https://doi.org/10.1016/0022-3115(92)90364-Q

Miller, M. K., & Hetherington, M. G. (1991). Local magnification effects in the atom probe. *Surface Science, 246*(1), 442-449. doi:https://doi.org/10.1016/0039-6028(91)90449-3

Miller, M. K., & Russell, K. F. (2007). Embrittlement of RPV steels: An atom probe tomography perspective. *Journal of nuclear Materials, 371*(1-3), 145-160.

Moody, M. P., Stephenson, L. T., Liddicoat, P. V., & Ringer, S. P. (2007). Contingency table techniques for three dimensional atom probe tomography. *Microscopy research and technique, 70*(3), 258-268. doi:https://doi.org/10.1002/jemt.20412

Odette, G. R., Yamamoto, T., Williams, T. J., Nanstad, R. K., & English, C. A. (2019). On the history and status of reactor pressure vessel steel ductile to brittle transition temperature shift prediction models. *Journal of nuclear Materials, 526*, 151863. doi:https://doi.org/10.1016/j.jnucmat.2019.151863

Pareige, P., Auger, P., Bas, P., & Blavette, D. (1995). Direct observation of copper precipitation in a neutron irradiated FeCu alloy by 3D atomic tomography. *Scripta Metallurgica et Materialia, 33*(7), 1033-1036. doi:https://doi.org/10.1016/0956-716X(95)00329-T

Reddy, S. M., Saxey, D. W., Rickard, W. D. A., Fougerouse, D., Montalvo, S. D., Verberne, R., & Van Riessen, A. (2020). Atom probe tomography: development and application to the geosciences. *Geostandards and Geoanalytical Research, 44*(1), 5-50.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS), 42*(3), 1-21.

Stephenson, L. T., Moody, M. P., Liddicoat, P. V., & Ringer, S. P. (2007). New techniques for the analysis of fine-scaled clustering phenomena within atom probe tomography (APT) data. *Microscopy and Microanalysis, 13*(6), 448.

Stiller, K., Thuvander, M., Povstugar, I., Choi, P.-P., & Andrén, H. O. (2016). Atom probe tomography of interfaces in ceramic films and oxide scales. *MRS Bulletin, 41*(1), 35.

Styman, P. D., Hyde, J. M., Wilford, K., & Smith, G. D. W. (2013). Quantitative methods for the APT analysis of thermally aged RPV steels. *Ultramicroscopy, 132,* 258-264. doi:https://doi.org/10.1016/j.ultramic.2012.12.003

Vaumousse, D., Cerezo, A., & Warren, P. J. (2003). A procedure for quantification of precipitate microstructures from three-dimensional atom probe data. *Ultramicroscopy, 95*, 215-221. doi:https://doi.org/10.1016/S0304-3991(02)00319-4

Wang, J., Schreiber, D. K., Bailey, N., Hosemann, P., & Toloczko, M. B. (2019). The Application of the OPTICS Algorithm to Cluster Analysis in Atom Probe Tomography Data. *Microsc Microanal, 25*(2), 338-348. doi:10.1017/S1431927618015386

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*(1), 1-9.

Williams, C. A., Haley, D., Marquis, E. A., Smith, G. D., & Moody, M. P. (2013). Defining clusters in APT reconstructions of ODS steels. *Ultramicroscopy, 132*, 271-278. doi:10.1016/j.ultramic.2012.12.011

Zelenty, J., Dahl, A., Hyde, J., Smith, G. D. W., & Moody, M. P. (2017). Detecting Clusters in Atom Probe Data with Gaussian Mixture Models. *Microsc Microanal, 23*(2), 269-278. doi:10.1017/S1431927617000320