https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

Technical Note: Cluster Analysis of Inverse Thermochronology Models

Tobias Stephan[1], Taís F. Pinto[2], and Eva Enkelmann[2]

[1]Lakehead University, Department of Geology, Thunder Bay, ON, Canada P7B 5E1
[2]University of Calgary, Department of Earth, Energy and Environment, Calgary, AB, Canada T2N 1N4

Correspondence: Tobias Stephan (tstephan@lakeheadu.ca)

Abstract. Thermochronological inverse modeling may produce none-unique solution, that can group different thermal histories. Objective identification of such groups, also referred as "path families", is challenging and greatly benefits from dimension-reducing exploratory data analysis tools. This article proposes a statistical algorithm to overcome these challenges. We show that Hausdorff and Fréchet distances are viable dissimilarity measures for ordered point sets, such as time-temperature paths. Clustering the pairwise dissimilarities between modeled thermal histories reveals distinct groups of thermal histories for a given sample or set of samples. As demonstrated by clustering a natural example, automated path-clustering allows for an objective and reproducible interpretation and maybe particularly useful for samples with poor prior knowledge of the time-temperature history. To allow adoption of the method by the thermochronology community, the methods introduced in this article are freely available through the package software `thermoclustr`, written in the programming language R.

1 Introduction

Thermochronological inverse thermal history modeling (using HeFTy or QTQt) produces time-temperature (t-T) paths used to infer the thermal history of a region. For example, HeFTy uses a Monte Carlo method to generate independent t-T paths which, depending on predefined boundary conditions (t-T constraints), cover a wide range of plausible cooling histories (Ketcham, 2005). Inverse modeling can produce results with significant diversity that are difficult to interpret. This becomes obvious, when all acceptable and good paths are plotted together and the large variability and overplotting of paths makes it difficult to extract meaningful information. There are two ways to overcome this issue (Fig. 1), either by visualizing the density of paths in t-T space, or by identifying groups of similar paths, so-called "path families" (Murray et al., 2022; Stevens Goddard et al., 2023). However, neither of these approaches have been formalized in a quantitative and statistical framework.

Thermal histories can be grouped into "path families" based on user-defined characteristics such as amount of peak heating in a certain time window, timing of cooling through a certain time window, etc. (Murray et al., 2022; Stevens Goddard et al., 2023). Although this analysis approach was proven to aid geologic interpretation of complex datasets (Stevens Goddard et al., 2023; Pinto et al., 2025), the manual definition of path families can be subjective and time-consuming when dealing with large number of samples.

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

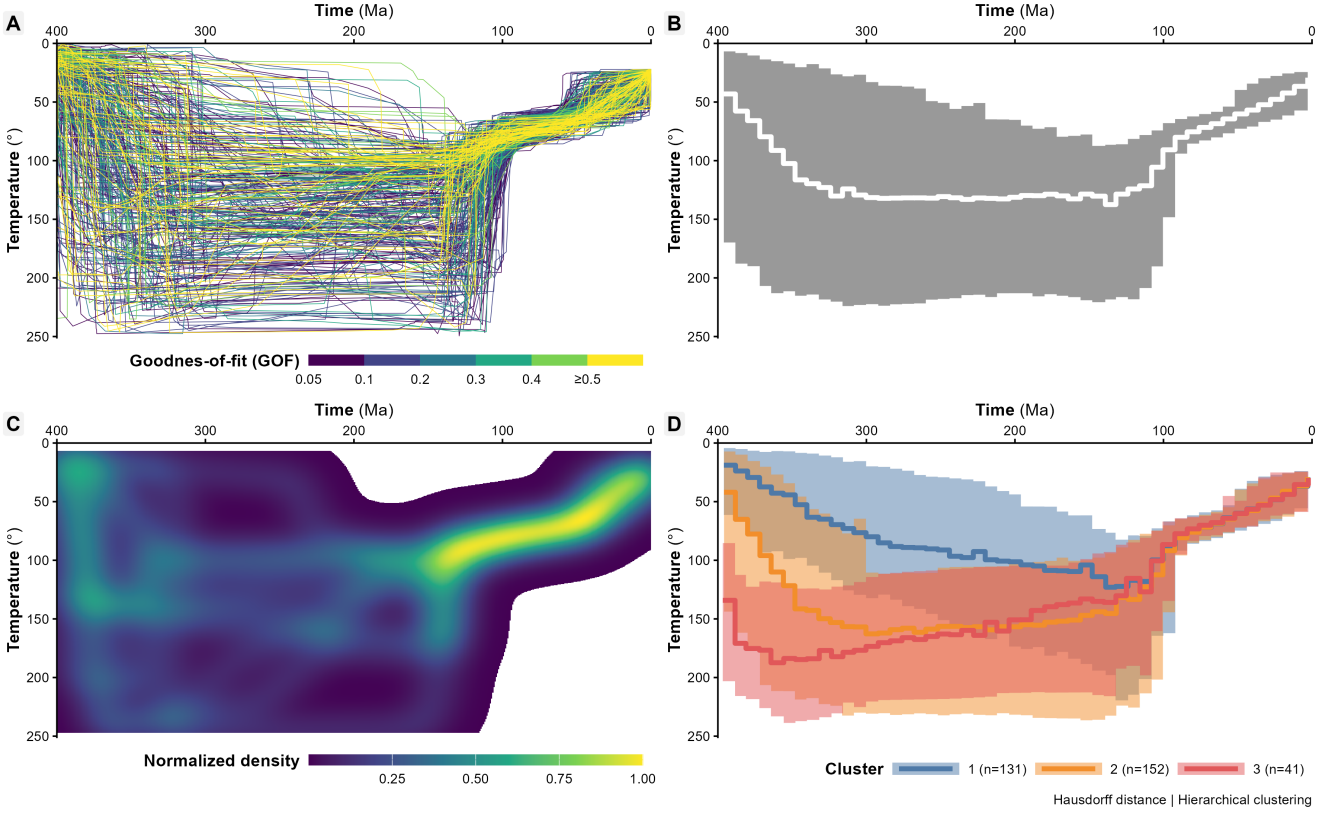

Figure 1. Approaches to visualize t-T path distribution. (A) All paths shown as individual line colored by their goodness-of-fit (GOF) value. (B) 90 % interpercentile range (gray area) and median path (white thick line) calculated from 50 equal-sized time bins. (C) Path density, densities are weighted by the GOF values of the paths. (D) Path clusters. Shaded areas show the clusters' 90 % interpercentile range and colored thick line shows the median paths. Statistics were calculated for 50 equally sized time bins. See details on example time-temperature data in text and Figure A1.

In this study we introduce an algorithm for path densities and automated and objective clustering of thermochronological t-T paths derived from HeFTy inverse thermal history modeling. Clustering is based on measuring the dissimilarity between paths using the Hausdorff and Fréchet distance metrics, which are commonly used for pattern recognitions, such as face matching (Takács, 1998), network analysis (Aksoy et al., 2019), as well as in spatial (-temporal) sciences for clustering satellite trajectories (Chen et al., 2011), road networks (Taha and Hanbury, 2015), or comparing geometries to identify irregularly shaped potatoes (Yu et al., 2022).

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

2 Method

2.1 Path density

Time-temperature paths are discrete paths defined as a two-dimensional set of ordered points. For thermal histories, the paths are in time (t) and temperature (T) space. Hefty reports the path coordinates by the path vertex points only, which can lead to an uneven distribution of points along the path.

Our algorithm for visualizing path densities is based on the following steps: (1) To avoid overrepresentation of paths with a high point density, we first interpolate points along the paths to create paths with identical point densities. The path density is defined by the user as the distance between two consecutive points along the path. The interpolation is done using linear interpolation between two vertex points. This ensures that each path represents the same probability of occurrence. (2) Next, the algorithm performs two-dimensional kernel density estimation. This is done by defining a grid in t-T space and calculating the density of paths that pass through each grid cell. A Gaussian kernel is used to smooth the density estimates by default. (3) Optionally, each path can be weighted by additional constraints, such as HeFTy's goodness-of-fit (GOF) value for each path. This allows to give more importance to paths that better fit the data when calculating the density. (4) Finally, the calculated densities can be visualized as contour plots in t-T space (Fig. 1C).

2.2 Path clustering

Clustering t-T paths require a metric that quantifies how similar the path geometries are. Generally, similar paths should be parallel, have similar time and temperatures, and this similarity should be measured using distance metrics.

2.2.1 Measuring the dissimilarity between two time-temperature paths

The dissimilarity between points in a two-dimensional Cartesian system is commonly measured using the Euclidean distance, which is defined as the square root of the sum of the squared differences between the coordinates of the two points:

$$d(p_a, p_b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2} \tag{1}$$

where (x_a, y_a) and (x_b, y_b) are the Cartesian coordinates of points p_a and p_b, respectively. It is a metric parameter as it satisfies the following properties: (i) non-negativity ($d(p_a, p_b) \geq 0$), (ii) identity ($d(p_a, p_b) = 0$ if $p_a = p_b$), (iii) symmetry ($d(p_a, p_b) = d(p_b, p_a)$), and (iv) triangle inequality ($d(p_a, p_c) \leq d(p_a, p_b) + d(p_b, p_c)$). However, the Euclidean distance is not suitable for quantifying the dissimilarity when the points in a set are ordered, as in paths. Here, a discrete path is defined as a two-dimensional set of ordered points. For cooling histories, the paths are in time (t) and temperature (T) space. The dissimilarity between discretized paths can be measured using the Hausdorff distance or the Fréchet distance.

The Hausdorff distance (HD) is the greatest of all the distances from a point in one path to the closest point in the other path (Fig. 2). Given two paths A and B in $t-T$ space, the Hausdorff distance H between A and B is defined as (Rucklidge, 1996):

$$\mathrm{HD}(A, B) = \max\big[h(A, B),\ h(B, A)\big] \tag{2}$$

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

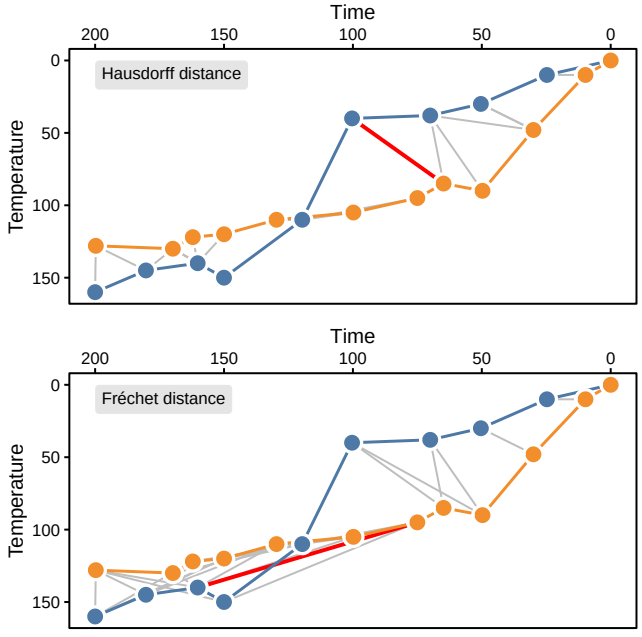

Figure 2. Visual representation of the Hausdorff (top) and Fréchet (bottom) distances between two time-temperature paths (orange and blue lines). The gray lines show the closest neighbors of each point in one path to the points in the other one. The red lines highlight Hausdorff and Fréchet distance, respectively.

which is the larger distance between $h(A,B)$ and $h(B,A)$. $h(A,B)$ expressed that the shortest distance is first found from each point in A to B, and then the largest among the set of shortest distances is selected as the value of $h(A,B)$:

$$h(A,B) = \max_{p_a \in A} \left[\min_{p_b \in B} \left(d(p_a, p_b) \right) \right] \tag{3}$$

$$h(B,A) = \max_{p_b \in B} \left[\min_{p_a \in A} \left(d(p_b, p_a) \right) \right] \tag{4}$$

The Hausdorff distance is (i) non-negative ($\mathrm{HD}(A,B) \geq 0$), (ii) identical ($\mathrm{HD}(A,B) = 0$ if and only if $A = B$), (iii) symmetric ($\mathrm{HD}(A,B) = \mathrm{HD}(B,A)$), and has (iv) triangle inequality ($\mathrm{HD}(A,B) \leq \mathrm{HD}(A,C) + \mathrm{HD}(B,C)$). Therefore, the Hausdorff distance is a metric parameter similar to the Euclidean distance.

Another measure is the Fréchet distance FD, which is a more general measure of similarity between paths (Alt and Godau, 1995). In addition to the ordering of the points along paths, the Fréchet distance takes into account the location and is defined as:

$$\mathrm{FD}(A,B) = \inf_{\alpha,\beta} \max_{t \in [0,1]} \left[d\Big(A\big(\alpha(t)\big),\ B\big(\beta(t)\big) \Big) \right] \tag{5}$$

where $A\big(\alpha(t)\big)$ and $B\big(\beta(t)\big)$ are the temperatures on the paths A and B, respectively, at time t. α and β are continuous non-decreasing functions that map the interval $[0,1]$ to the intervals of the paths. The length of the paths between them at time t is

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

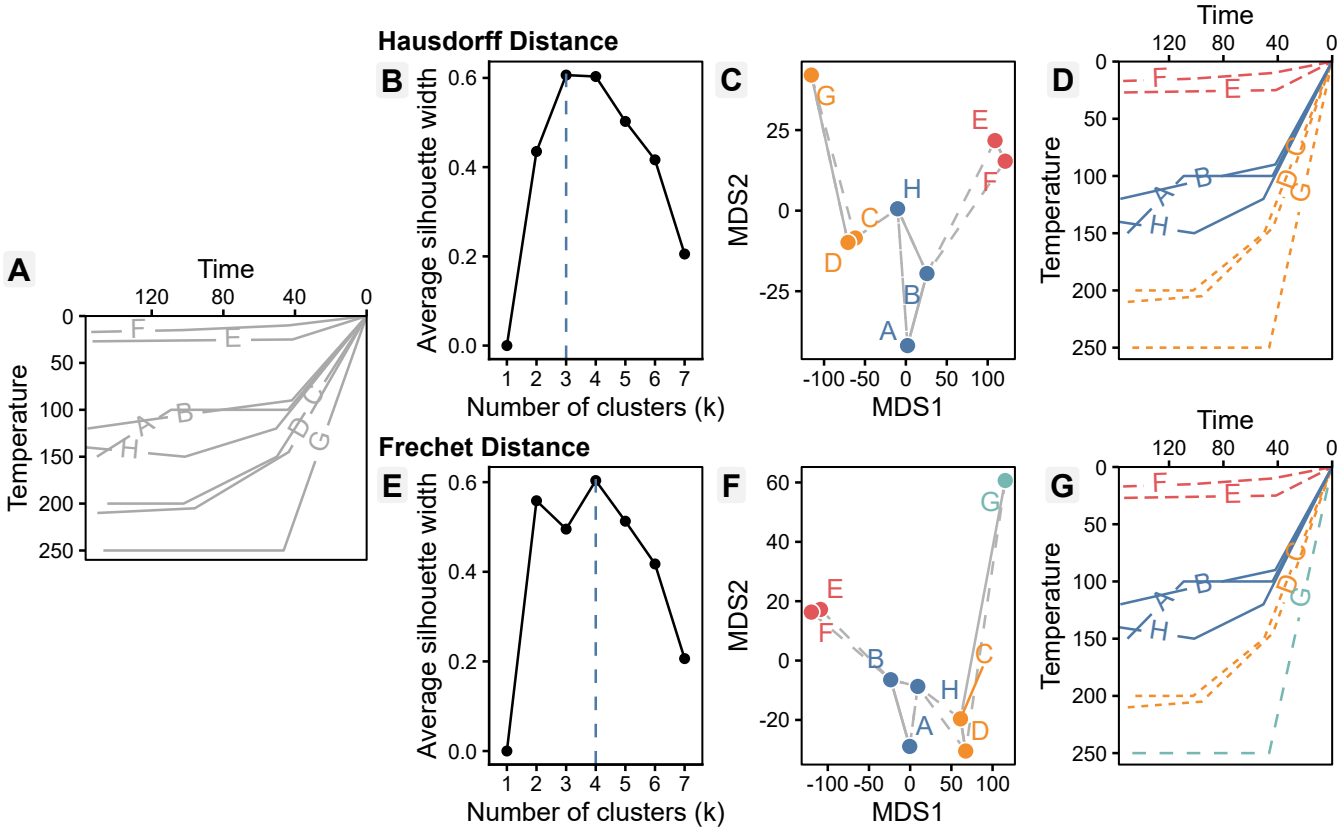

Figure 3. Cluster results for Hausdorff distances (top panel) and Fréchet distance (bottom) of 8 artificial time-temperature paths. Left diagram shows the multidimensional scaling of the dissimilarity matrix. Every letter refers to one path. Solid lines mark the closest neighbors and dashed lines the second-closest neighbors. Center diagram shows the Average Silhouette Width and the optimal number of clusters (vertical dashed line). Right diagram shows the paths clustered by hierarchical clustering using the optimal number of clusters as well as the Hopkins statistic and its p-value.

the distance between $A(\alpha(t))$ and $B(\beta(t))$. The Fréchet distance is also a metric, but is not symmetric in general. However, it can be made symmetric by taking the maximum of the two distances.

The measure is often visualized as the minimum "leash length" required for a person and their dog to walk along two separate paths from start to finish without moving backward (Fig. 2). The Fréchet distance is always greater than or equal to the Hausdorff distance. It is also less sensitive to noise, outliers, and variations in path shape (Fig. 3). However, because t-T paths are inherently ordered by time (i.e., they cannot go backward) and the Fréchet algorithm is computationally more demanding, the Hausdorff distance is sufficient and more suitable for applications in thermochronology.

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

2.2.2 Cluster tendency

The path distances now allow testing whether the n t-T paths are randomly distributed or if groups among the data exist. Such a cluster tendency of t-T paths is assessed by the (generalized) Hopkins statistic that measures the probability that a given t-T data set is generated by a uniform data distribution (Hopkins and Skellam, 1954; Lawson and Jurs, 1990; Banerjee and Dave, 2004; Adolfsson et al., 2019; Cross and Jain, 1982; Coblentz et al., 2024; Wright, 2022a).

The Hopkins statistic (H) compares two sets of data: a set of observations X and a set of uniformly randomly distributed data points Y, which lie in the same domain as X. The Hopkins statistic is computed from the points as:

$$H = \frac{\sum_{i=1}^{n} u_i^2}{\sum_{i=1}^{n} (u_i^2 + w_i^2)} \tag{6}$$

where u_i is the Euclidean distance of an observed data point i to its nearest neighbor, and w_i is the distance from a random point to its nearest observed data point. If the data are uniformly distributed (i.e., no meaningful clusters), $\sum_{i=1}^{m} u_i^2 \sim \sum_{i=1}^{m} w_i^2$ and thus H would be 0.5. If the data are strongly clustered, the distances for the random points would be notably larger than those for the observed data, and hence, H will be > 0.5. Since values for H are samples from the Beta-distribution Beta(m, m) (Cross and Jain, 1982), a value for H greater than 0.75 indicates a clustering tendency at the 95 % confidence level.

2.2.3 Dimensionality reduction

Exploring the dissimilarities among data graphically requires dimension-reducing techniques to project the high-dimensional distance matrix containing all pairwise comparisons among n observations into a lower-dimensional Cartesian space. For example, applying dimension-reducing algorithms to $n(n-1)/2 = 4950$ pairwise comparisons of $n = 100$ t-T paths translate the $n \times n$ Hausdorff (or Fréchet) matrix into k dimensions (usually $k = 2$). The coordinates of the resulting output matrix (also called configuration) are arranged in a way to preserve the level of the original similarities in the reduced space which can be visualized as a two-dimensional scatter plot when $k = 2$ (see Fig. 3C,F and 4B).

There are several algorithms for reducing the dimensionality of distance matrices, including Multidimensional scaling, MDS (Kruskal, 1964; Kruskal and Wish, 1978), Uniform Manifold Approximation and Projection, UMAP (McInnes et al., 2018), and t-Distributed Stochastic Neighbor Embedding, t-SNE (Hinton and Roweis, 2002; van der Maaten and Hinton, 2008). Visualizing similarity relationships through the MDS, UMAP, or t-SNE configuration allows evaluating the structure of the data, including, nearest neighbor relationships, gradual transitions, and well-separated clusters (Fig. 4B). A comparison of some dimensionality-reducing projections is shown in Figure A2.

2.2.4 Finding the optimal number of clusters

Except for density-based cluster approaches, most clustering methods require specification of the number of clusters (k). Finding an appropriate value for k can be achieved either through visual inspection (Fig. 5) or statistical evaluation (Fig. 4A). Visual inspection is useful when clusters are obvious and clearly distinguishable in the t-T diagram or in the MDS configuration derived from the dissimilarity matrix.

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

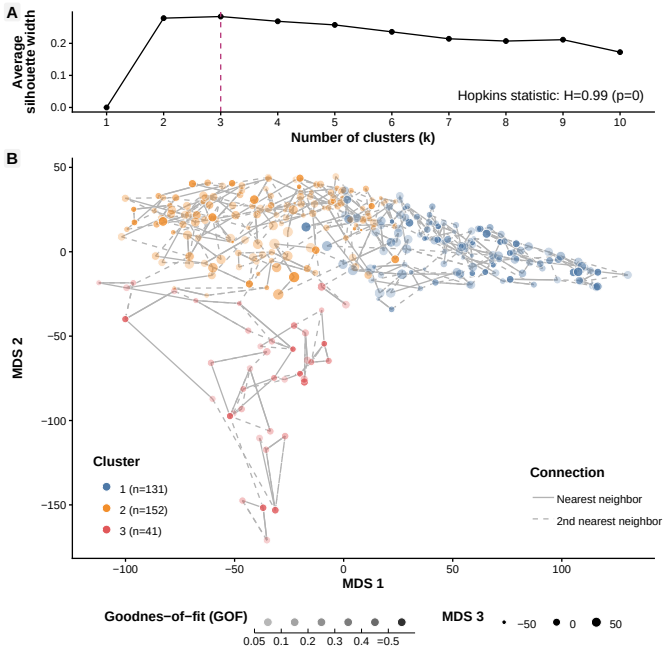

Figure 4. Cluster evaluation for natural t-T paths (data shown in Fig. A1B). (A) Estimation of the optimal number of clusters using the Average Silhouette Width. (B) Path similarities shown by multidimensional scaling (MDS) of the Hausdorff distance matrix. Colors refer to the cluster result using hierarchical clustering of 325 t-T paths. The size of the data dots is proportional to value of the 3rd dimension in the metric MDS configuration. Transparency of the dots is proportional to the goodness-of-fit value (GOF) of the HeFTy modelling.

For more complex data structures (Fig. 5), the optimal number of clusters can be determined statistically using the "Average Silhouette Width" (ASW) (Rousseeuw, 1987; Batool and Hennig, 2021; Kaufman and Rousseeuw, 1990). This metric evaluates how similar each observation is to its assigned cluster compared to other clusters, thereby providing a quantitative measure of clustering quality.

The similarity of an observation to its own cluster is measured by the average distance of a data point i to all other data points in the cluster C_i and is generally defined as:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j) \tag{7}$$

The similarity to other clusters is measured by the distance of i to the closest cluster that i is not assigned to i:

$$b(i) = \min_{j \neq i} \frac{1}{|C_j|} \sum_{l \in C_j} d(i,l) \tag{8}$$

Because of the ordered nature of the path data, the Euclidean distance between two data points d in Eqs. 7 and 8 is replaced by the Hausdorff (Eq. 3) or Fréchet distance (Eq. 5).

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

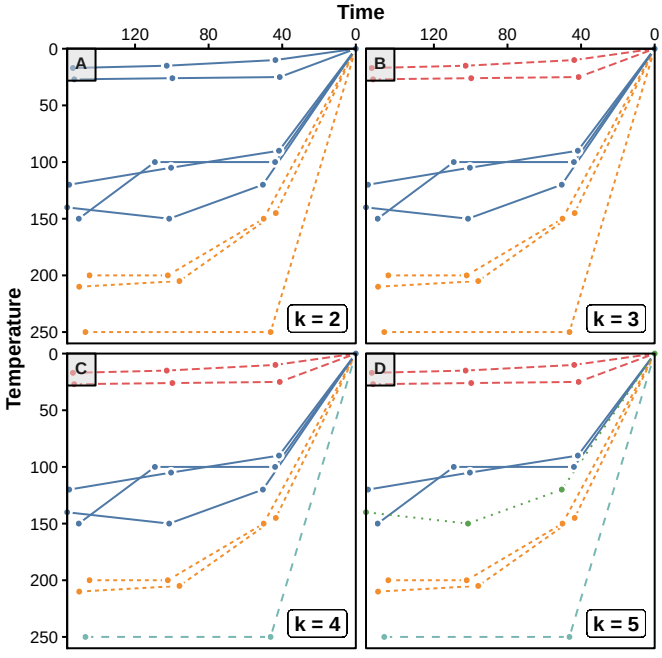

Figure 5. Clustering 8 artificial time-temperature paths (Fig. 3) using hierarchical clustering of Hausdorff distance. Every diagram shows the result for different numbers of clusters (k).

Using both measures, the silhouette width of a data point i is

$$s_i(C,d) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{9}$$

where $s(i) = 0$ if $|C_i| = 1$. The silhouette width ranges from -1 to $+1$, where a high value indicates that the data point is well-matched to its own cluster and poorly matched to neighboring clusters. If most points have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

The Average Silhouette Width (ASW) of a clustering C is

$$\bar{S}(C,d) = \frac{1}{n}\sum_{i=1}^{n} s_i(C,d) \tag{10}$$

A large value of $s_i(C,d)$ means that $b(i)$ is much larger than $a(i)$, and that consequently i is much closer to the observations in its own cluster than to the neighboring one.

Because clusters are supposed to be internally homogeneous and well separated from one another, larger values of s_i and $\bar{S}$ indicate higher clustering quality and a more optimal data partitioning. To find an optimal number of clusters k, the ASW is calculated iteratively for a sequence of candidate values of k. The optimal solution corresponds to the largest value of ASW,

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

i.e., the one that that maximizes $\bar{S}$ (Kaufman and Rousseeuw, 1990, p. 87):

$$SC = \max_k \tilde{s}(k) \tag{11}$$

Figure 5 illustrates the iterative calculation of the ASW for different numbers of clusters, where the maximum ASW represents the optimal number of clusters.

2.2.5 Clustering methods

The Hausdorff/Fréchet dissimilarity matrix can be clustered using any type of cluster algorithms, including partitioning methods (e.g., k-means and k-medioids—also called "PAM"), hierarchical clustering (linkages in dendrogram), spectral clustering (using decomposition of data into eigenvectors and eigenvalues, Ng et al. (2001)), fuzzy clustering (Kaufman and Rousseeuw, 1990), and density methods (based on maximally connected components of the set of points that lie within some defined distance from some core object e.g., DBSCAN, Ester et al. (1996)).

Figure 6 shows the cluster results for the different cluster algorithms implemented in the package. Metric multidimensional scaling of all cluster results (Fig. 7) shows that most of the cluster algorithms produce similar clusters, while the density-based algorithms (dbscan and hdbscan) may produce different clusters.

3 Software implementation

The presented method and the example dataset are implemented in the free and open-source software package `thermoclustr` that is written in the computer language R. As an R-based cross-platform program, `thermoclustr` works on all operating systems, including Windows©, MacOS X© and GNU/Linux. Implementation within the R universe provides users automated repetitive workflows, flexible and advanced data visualization, and access to R's vast statistical capabilities.

The source code for the program is made available under the GNU GPL license v3.0, which permits re-use and modification provided that any derived code is released under the same conditions (Free Software Foundation, 2007). The entire source code of `thermoclustr` is accessible through the Comprehensive R Archive Network (CRAN) and can be accessed through R's command line:

```
1: install.packages("thermoclustr")
2: library("thermoclustr")
```

Time-temperature data of the paths can be imported using any `read*` functions available and suitable. For convenience, the function `read_hefty()` imports the modeled thermal history paths (including time, temperature and GOF data) from a HeFTy export (`.txt`) file. This function creates an `"HeFTy''` object to ease subsequent steps.

To calculate the pairwise dissimilarities between cooling paths, the paths a first converted into 2D objects, where the x-axis is time and the y-axis is temperature. Explicitly, `thermoclustr`'s subroutine `path_diss()` converts t-T paths into spatial objects using the functions from the `sf` package (Pebesma, 2018; Pebesma and Bivand, 2023). The Hausdorff or Fréchet

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

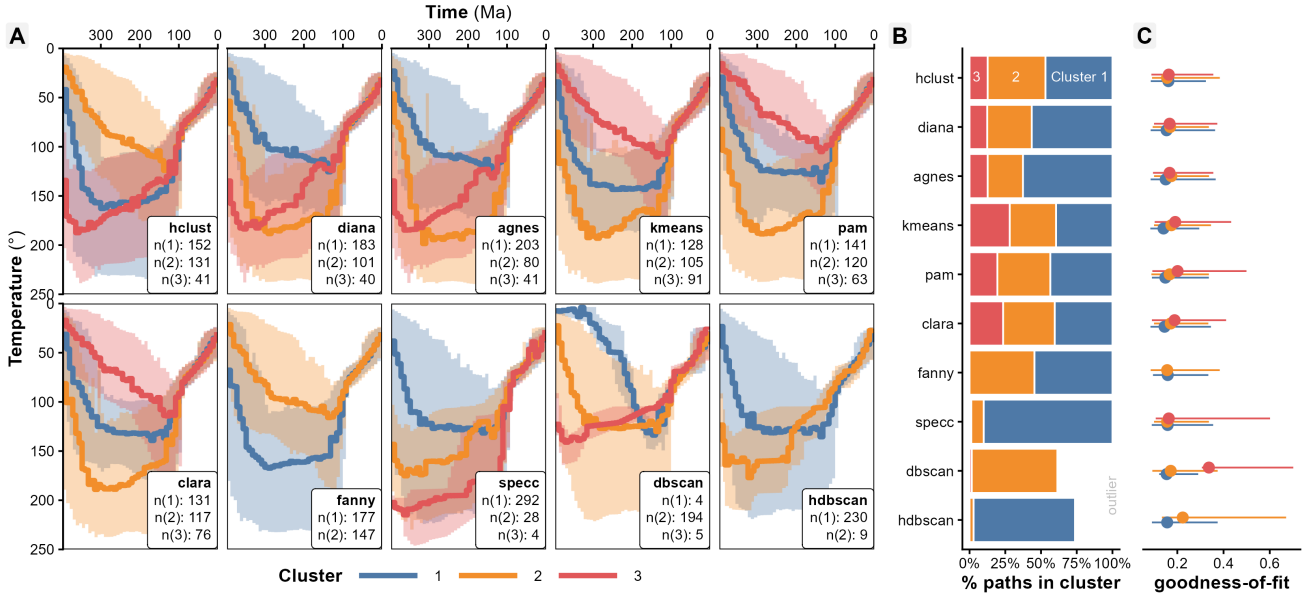

Figure 6. Comparison of cluster results using different algorithms for natural t-T paths (data shown in Fig. A1B). (A) Time-temperature plots showing the clustered path ranges identified by each cluster algorithm. (B) Fraction of paths of total paths within each cluster. Note that the number of paths in density-based clusters do not added up to 100 % because the detected outliers are not counted. (C) Distribution of the goodnes-of-fit (GOF) values in each cluster, where the line and the point indicate the 90 % interpercentile range and the median, respectively. Acronyms of cluster algorithms: hclust – Hierarchical clustering (McQuitty, 1966), diana – DIvisive ANAlysis (hierarchical) Clustering (Kaufman and Rousseeuw, 1990), agnes – Agglomerative Nesting hierarchical clustering (Kaufman and Rousseeuw, 1990), kmeans – K-Means Clustering (Hartigan and Wong, 1979), pam – Partitioning Around Medoids (Kaufman and Rousseeuw, 1990), clara – Clustering Large Applications (Kaufman and Rousseeuw, 1990), fanny – Fuzzy Analysis Clustering (Kaufman and Rousseeuw, 1990), specc – Spectral Clustering (Ng et al., 2001), dbscan – Density-based Spatial Clustering of Applications with Noise (Ester et al., 1996), hdbscan – Hierarchical Density-based Spatial Clustering of Applications with Noise (Campello et al., 2015).

distance between all paths are then calculated using the function `st_distance()` from the package `sf`. This results in a dissimilarity matrix of the pairwise Hausdorff or Fréchet distances between all paths.

The dissimilarity matrix is clustered using conventional cluster algorithm, such as hierarchical clustering or partitioning (e.g., "k-means" and "k-medioids"; implemented in R through `pam()` from the R package "cluster"; Maechler et al. (2025, 1999)), fuzzy clustering (using `fanny()` from the "cluster" package), spectral clustering (through `spec()` from "kernlab" package; Karatzoglou et al. (2004b, a)), or density-based spatial clustering of applications with noise (Hahsler et al., 2019) through `dbscan()` from the "dbscan" package (Hahsler and Piekenbrock, 2015). For convenience, these algorithms are incorporated into the function `path_cluster()` within the `thermoclustr` R package.

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

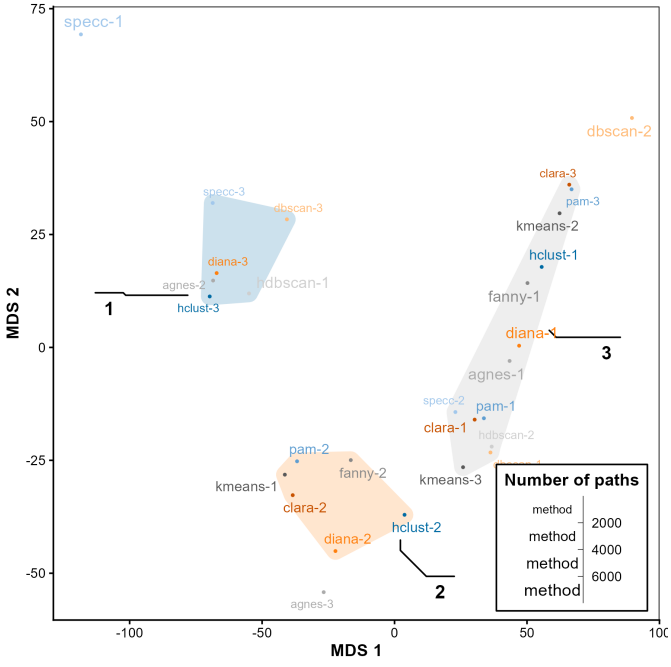

Figure 7. Metric multidimensional scaling plot of the cluster results from the different clustering algorithms applied on natural t-T paths (data shown in Fig. A1B). See Figure 6 for explanation of acronyms for cluster algorithms.

An estimate of the optimal number of clusters can be retrieved using `path_nbclust()` function which implements the `silhouette()` function from the "cluster" package in R. A visualization of the Average Silhouette Width for several candidate numbers of clusters is included in the same function.

The "Hopkins statistic" for evaluating the cluster tendency is computed by the function `path_diss()` which implements functions from the "hopkins" package (Wright, 2022a, b) in R. Because the statistic cannot be applied on a $n \times n$ distance matrix, we first reduce the data to two dimensions using metric multidimensional scaling.

The package `thermoclustr` additionally provides simple plotting tools to visualize the paths, densities, and cluster results. Visualizing the t-T paths can be achieved through `plot_paths()`. The function additionally offers to cluster the paths, and colorize them accordingly. The path densities can be visualized through the functions `plot_path_density()`, `plot_path_density_filled()` and `plot_path_density_filled_weighted()` for a contour plot, rasterized densities and a weighted version of the path density map, respectively. The MDS plot can be created using the configuration coordinates provided in the output of `path_diss()`. This function implements "Classical (metric) MDS" using `cmdscale()` function from R's "stats" package by default. Alternatively, the user is free to chose other dimension-reducing algorithms available in R (Fig. A2), such as Kruskal's non-metric MDS using `isoMDS()` from the "MASS" package (Venables and Ripley, 2002), Sammon's Non-Linear Mapping (Sammon, 1969) using `sammon()` also from the "MASS" package, UMAP using

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

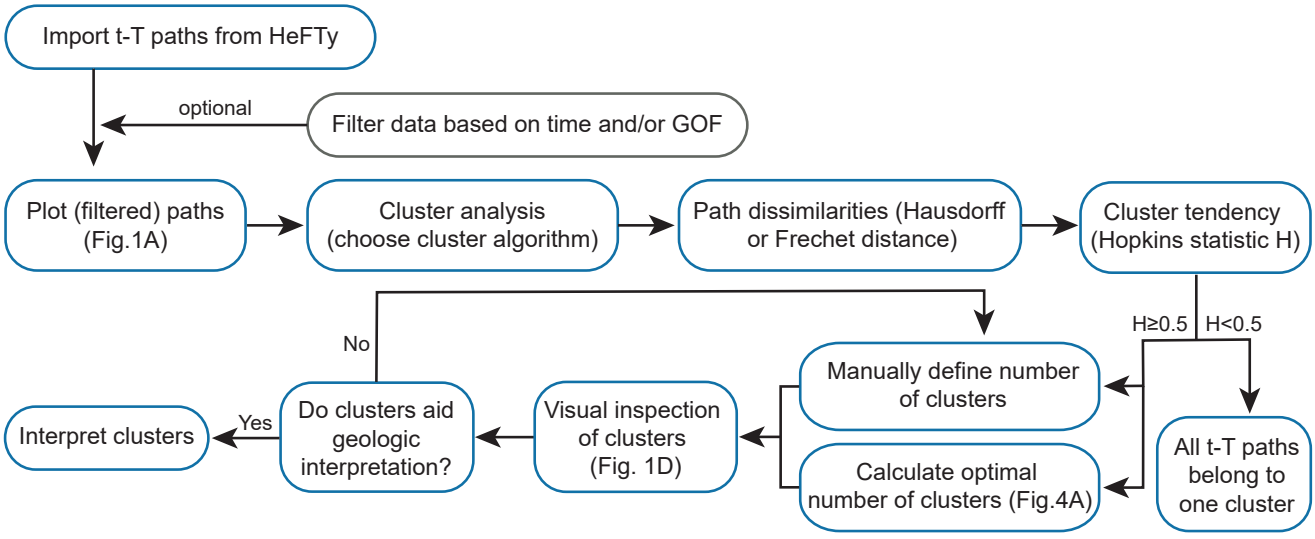

Figure 8. Workflow and decision tree for clustering time-temperature paths.

`umap()` the "uwot" package (Melville, 2019), and t-SNE using the function `Rtsne()` from the package "Rtsne" (Krijthe, 2015).

3.1 Proposed workflow

We suggest the following workflow to provide a structured and reproducible strategy for clustering t-T model data (Fig. 8): As an initial step, datasets may optionally be filtered according to user-defined thresholds for time, temperature and GOF values, allowing the analysis to focus on geologically meaningful solutions. The filtered data are then transformed into a dissimilarity matrix using either the Hausdorff or Fréchet distance, which quantifies similarity among thermal histories. Prior to clustering, the cluster tendency of the dataset is evaluated using the Hopkins statistic to assess whether meaningful grouping is statistically supported. The optimal number of clusters can subsequently be determined through visual inspection or statistical estimation, after which clustering is performed and the results are examined visually within both the t-T diagram or the MDS biplot.

Interpretation remains an iterative process. If the clustering result is unsatisfactory, individual steps of the workflow can be repeated and adjusted, such as modifying filtering thresholds, distance measures, cluster algorithm, or number of clusters (Fig. 8). It should be noted that the number of clusters must be larger than one and should be as small as possible to avoid overfitting and artificial grouping of the data. If repeated optimization fails to produce meaningful results, the user must accept that the data are not significantly clustered. A minimal working example for the recommended workflow using the abovementioned functions is provided in the appendix.

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

4 Application on a natural sample

To show an example of the cluster analysis, we apply the proposed workflow here to a natural sample from a granodiorite of the Canadian Shield in the southwestern Northwest Territories in Canada (sample "112-73" from Pinto et al. (2025); Arne (1991)). The Paleoproterozoic granodioritic basement is overlain by Devonian passive-margin sequences of the Western Canada Sedimentary Basin, that were buried to $\geq 1\,\mathrm{km}$ prior to the Early Cretaceous (Pinto et al., 2025). Early Mesozoic terrane accretion formed the Cordilleran orogenic system and transformed the basin into a foreland basin, followed by erosion and rock cooling.

The sample was analyzed using apatite (U–Th)/He thermochronology (Pinto et al., 2025) and apatite fission-track dating (Arne, 1991). (U–Th)/He dates range from 63 and 36 Ma, whereas fission-track dating yields a mean age of 98 Ma. Thermal histories were generated by inverse modelling using HeFTy v2.1.7 (Ketcham, 2005, 2024). A total of 50,000 paths were set as model end conditions (Fig. A1A), of which 325 paths achieved a GOF greater than 0.05.

To encompass the full range of possible post-crystallization thermal histories, constraints for initial temperatures and timing were set to 300–600 °C and 3400–2500 Ma (Fig. A1A). Because this study focuses on post-depositional thermal histories of the overlying basin, only the time interval constrained by geological information was used for further cluster analysis. Therefore, the model output was truncated to the 0–400 Ma and 0–250 °C time-temperature range (Fig. 1 and A1B). This filtering of t-T paths reduces the weighting effect of poorly constrained and highly variable pre-depositional histories. A detailed discussion of weighting effect and cluster behavior for datasets of varying complexity will be in a future accompanying study.

The filtered t-T paths and their path density distribution (Fig. 1A–C) show substantial variability during the early basin history, especially between 400 and 100 Ma. The median path suggest that maximum temperatures were reached at 350 Ma, followed by prolonged residence at these elevated temperatures until ∼100 Ma and subsequent cooling.

Using the Hausdorff distance metric to measure the path dissimilarities, the Hopkins statistic indicates strong cluster tendency ($H = 0.99$, $p < 0.05$) of the t-T paths. The Average Silhouette Width identifies three clusters (Fig. 4A). Clustering separates the thermal histories into three groups of t-T paths distinguished by heating rates, peak-temperatures, and cooling onset (Fig. 1D and 4B):

- **Cluster 1**: slow heating reaching maximum temperatures (∼200 °C) at ca. 110 Ma by slow heating, followed by cooling,
- **Cluster 2**: moderate heating rates up to ∼230 °C by ca. 300 Ma, remaining at these temperatures until 110 Ma, before cooling,
- **Cluster 3**: relatively rapid heating to ∼230 °C at 350 Ma followed by slow cooling.

All three groups share a similar cooling history after 110 Ma. Cluster 1 is considered unlikely because the stratigraphic record indicates discontinuous sedimentation between the Devonian and the Cretaceous, with multiple erosional unconformities (Morrow et al., 1993). Likewise, Cluster 2 is not supported by documented Carboniferous regression and basin unloading recorded in the regional stratigraphy of the basin (Morrow et al., 1993). Only Cluster 3 agrees with independent constraints from the

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

basin stratigraphy as well and temperature estimates and timing from the nearby Mississippi-Valley–type mineralization at Pine Point (Szmihelsky et al., 2020; Bourdet et al., 2008; Nakai et al., 1993). This example demonstrates that clustering can isolate geologically plausible burial-exhumation scenarios that may be difficult to recognize through visual inspection of HeFTy model outputs alone.

5 Discussion

5.1 Limitations and outlook

Limitations of the automated path-clustering approach primarily arise when datasets become either extremely large or intrinsically unsuitable for clustering. In both cases, results may fail, be biased or become computationally prohibitive. 'Thermal histories with very low dispersion, for example, commonly result from highly precise thermochronology data or strongly constrained, or even overfitted, HeFTy modelling. In such cases, clustering provides little additional insight because the modeled paths already converged toward a narrow thermal history solution.

A practical limitation is computational cost. Despite optimization efforts in `thermoclustr`, including vectorized operations and parallel computation processes, the calculation of pairwise dissimilarities remains the primary bottleneck. For a dataset containing n paths, the computation produces an n^2 data matrix, which is manageable for typical datasets ($n \leq 10{,}000$ paths), but increasingly memory-intensive for very large data ($n > 100{,}000$). Computational demands can be reduced through filtering paths by time, temperature, or GOF thresholds. Furthermore, high-resolution HeFTy models may produce paths with many vertex points, substantially increasing the data size without necessarily improving individual path geometries. Path simplification through down-sampling or smoothing algorithms can therefore be applied prior to reduce complexity without affecting the overall path shape and, hence, cluster result.

Future developments will focus on improving performance for intensive computations and expanding accessibility. Planned improvements include more parallelization using the "future" package (Bengtsson, 2021), implementation of C++ via "Rcpp" (Eddelbuettel and François, 2011; Eddelbuettel et al., 2008), and the potential development of a web-based interface to broaden usability. The package is intended to be an actively evolving, community-driven project designed so that these and other limitations of the existing software can be addressed in the future. Our GitHub repository contains detailed instructions for how to use the code as well as how to make contributions, and we welcome new collaborations for future versions.

5.2 Benefits and potential applications

The advantages of the path-family approach have been discussed in detail by Murray et al. (2022). An automated approach to this workflow expands the applicability by increasing computational speed, reproducibility, and objectivity, while allowing for additional statistical analyzes. It is particularly beneficial for datasets characterized by large analytical or statistical dispersion, where thermal histories are difficult to interpret using traditional visual inspection alone. For example, the approach may be useful when the apatite or zircon thermochronometer were fully or partially reset. The method is well suited for

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

poorly constrained inverse models with broad constraint boxes, as well as detrital samples that inherently represent mixtures of multiple thermal histories and therefore require unmixing approaches. Finally, the method may be also advantageous for deep-time datasets, where increasing uncertainty through geological time leads to more dispersed thermal history solutions and complicated conventional interpretations.

Further geological scenarios and potential applications, together with their implications for the interpretation of complex thermal histories, will be discussed in a future accompanying paper.

6 Conclusions

The use of cluster analysis to aid geologic interpretation of complex thermal histories allows for a more in-depth exploration of modeling results which may elucidate the more variable parts of the thermal history. The automated character of this analysis also prevents biasing and enables large datasets to be more easily approached. Our new method efficiently processes large datasets through a repeatable workflow, that minimizes user bias and ensures reproducible results. Importantly, it allows a quantitative assessment of the statistical significance of identified path families and enables direct comparison of families from different samples of clustering approaches using consistent dissimilarity measures.

All algorithms are implemented in the R package `thermoclustr`, which will be available on CRAN (https://cran.r-project.org/web/packages/thermoclustr) upon publication. The package is intended to be an actively evolving, community-driven project designed to expand access to quantitative thermochronology tools, while promoting reproducible and transparent research.

Code and data availability. The codes and example data used in this study set will be publicly available on CRAN (https://cran.r-project.org/web/packages/thermoclustr) upon publication. A beta version is freely available at https://github.com/tobiste/thermoclustr.

Appendix A: Minimal working example

The following code reproduces the examples used in the study.

```
# Install package
install.packages("thermoclustr")

# Load package
library("thermoclustr")

# Import example data
path2file <- system.file("112-73_30_H1_50-inv.txt", package = "thermoclustr")
tT_paths <- read_hefty(path2file) # replace `path2file' with your own HeFTy file
```

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026


```
10:
11: # Plot the paths
12: plot_paths(tT_paths)
13:
14: # Crop the paths
15: tT_paths_filt <- crop_paths(tT_paths, time = c(0, 400), temperature = c(0, 251))
16:
17: # Densify paths by interpolating equally spaced points
18: tT_paths_filt_dens <- densify_paths(tT_paths_filt, max_distance = 1)
19:
20: # Path statistics
21: # Here, summary statistics are calculated for 50 equally spaced bins
22: path_statistics(tT_paths_filt_dens, breaks = 50)
23:
24: # Plot the path densities
25: plot_path_density(tT_paths_filt) # contour plot
26: plot_path_density_filled(tT_paths_filt) # filled contour plot
27: s
28: # Hausdorff dissimilarity matrix
29: tT_paths_diss <- path_diss(tT_paths_filt, dist = "Hausdorff")
30:
31: # Metric MDS plot
32: plot(tT_paths_diss$mds)
33:
34: # Show Hopkin's statistic
35: print(tT_paths_diss$hopkin)
36:
37: # Optimal number of cluster
38: tT_paths_nb <- path_nbclust(tT_paths_diss)
39: print(tT_paths_nb$optimal)
40:
41: # Average Silhouette Width plot
42: tT_paths_nb$plot
43:
44: # Cluster the paths using hierarchical clustering and
45: # the optimal number for clusters
46: tT_paths_cluster <- cluster_paths(
47:   tT_paths_diss,
```

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026


```
k = tT_paths_nb$optimal,
method = "hclust"
)

# Plot the clustered paths
plot_paths(tT_paths_filt, cluster = tT_paths_cluster)
```

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

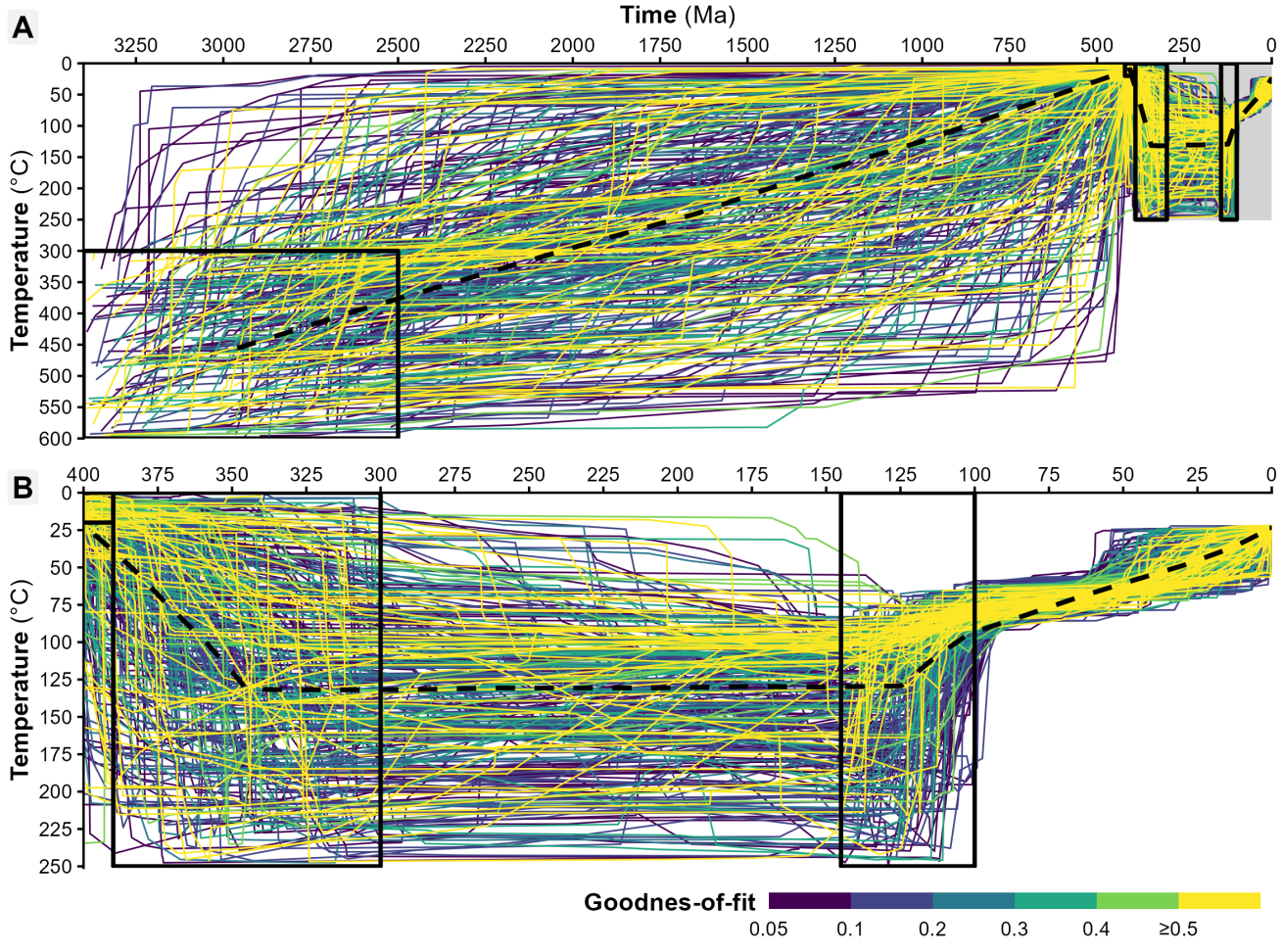

Figure A1. Inverse thermal modeling result of a granodiorite from the Canadian Shield that is overlain by Devonian strata (sample "112-73" from Pinto et al. (2025)) using HeFTy v2.1.7 (Ketcham, 2005, 2024). (A) Full extent of modelled thermal history data. Grey polygon indicates history since Paleozoic deposition. (B) Close-up of the post-depositional thermal history used for cluster analysis. The time-temperature paths are color coded according to the goodness-of-fit value. The dashed line indicates the mean path calculated by HeFTy. The black frames indicate the constraint boxes used for the modelling.

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

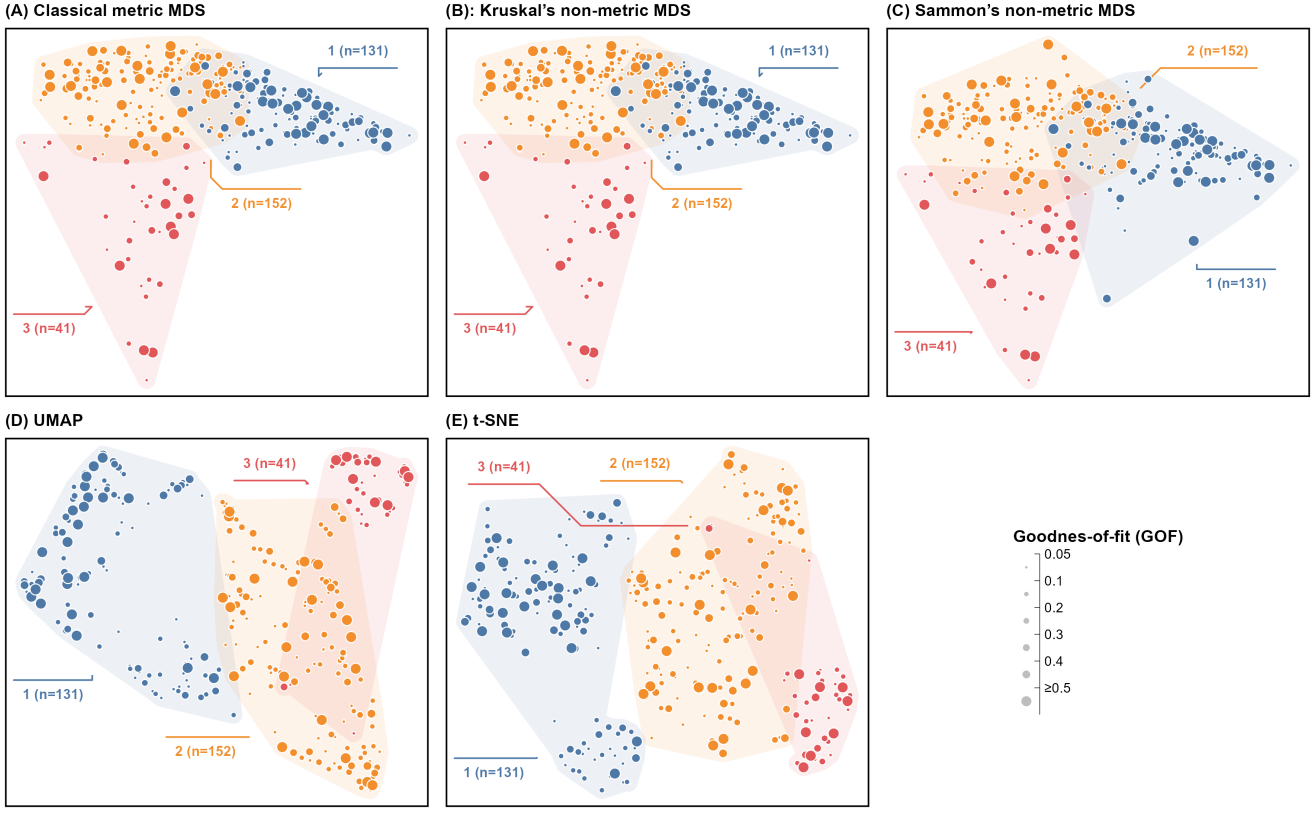

Figure A2. Comparison of dimensional-reducing algorithms applied on the same natural t-T paths as shown in Figure A1B.

Author contributions. All authors contributed to the conceptualization and methodological design of the study. Software development and implementation were undertaken by TS, with validation contributions from TFP and EE. The manuscript was drafted by TS and all authors contributed to the review, editing, or validation of the final version.

Competing interests. The authors have declared that there are no competing interests.

Acknowledgements. This research was funded by the Lakehead University start-up grant to TS and the Natural Sciences and Engineering Research council (NSERC) Discovery Grant #RGPIN-2024-03863 to EE.

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

References

Adolfsson, A., Ackerman, M., and Brownstein, N. C.: To cluster, or not to cluster: An analysis of clusterability methods, Pattern Recognition, 88, 13–26, https://doi.org/10.1016/j.patcog.2018.10.026, 2019.

Aksoy, S. G., Nowak, K. E., Purvine, E., and Young, S. J.: Relative Hausdorff distance for network analysis, Applied Network Science, 4, https://doi.org/10.1007/s41109-019-0198-0, 2019.

Alt, H. and Godau, M.: Computing the Fréchet Distance between Two Polygonal Curves, International Journal of Computational Geometry & Applications, 5, 75–91, https://doi.org/10.1142/s0218195995000064, 1995.

Arne, D. C.: Regional thermal history of the Pine Point area, Northwest Territories, Canada, from apatite fission-track analysis , Economic Geology, 86, 428–435, https://doi.org/10.2113/gsecongeo.86.2.428, 1991.

Banerjee, A. and Dave, R. N.: Validating clusters using the Hopkins statistic, in: IEEE International Conference on Fuzzy Systems (IEEE Cat. No.04CH37542), vol. 1, pp. 149–153, IEEE, https://doi.org/10.1109/fuzzy.2004.1375706, 2004.

Batool, F. and Hennig, C.: Clustering with the Average Silhouette Width, Computational Statistics & Data Analysis, 158, 107 190, https://doi.org/10.1016/j.csda.2021.107190, 2021.

Bengtsson, H.: A Unifying Framework for Parallel and Distributed Processing in R using Futures, The R Journal, 13, 208, https://doi.org/10.32614/rj-2021-048, 2021.

Bourdet, J., Pironon, J., Levresse, G., and Tritilla, J.: Petroleum type determination through homogenization temperature and vapour volume fraction measurements in fluid inclusions, Geofluids, 8, 46–59, https://doi.org/10.1111/j.1468-8123.2007.00204.x, 2008.

Campello, R. J. G. B., Moulavi, D., Zimek, A., and Sander, J.: Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection, ACM Transactions on Knowledge Discovery from Data, 10, 1–51, https://doi.org/10.1145/2733381, 2015.

Chen, J., Wang, R., Liu, L., and Song, J.: Clustering of trajectories based on Hausdorff distance, in: 2011 International Conference on Electronics, Communications and Control (ICECC), pp. 1940–1944, https://doi.org/10.1109/ICECC.2011.6066483, 2011.

Coblentz, D., van Wijk, J., Carmichael, J., Johnson, C., Delorey, A., Chai, C., Maceira, M., and Richardson, R. M.: New approaches to an old problem: addressing spatial gaps in the World Stress Map, Geological Society, London, Special Publications, 546, 47–68, https://doi.org/10.1144/sp546-2023-27, 2024.

Cross, G. R. and Jain, A. K.: Measurement of clustering tendency, pp. 315–320, Elsevier, ISBN 9780080276182, https://doi.org/10.1016/b978-0-08-027618-2.50054-1, 1982.

Eddelbuettel, D. and François, R.: Rcpp: Seamless R and C++ Integration, Journal of Statistical Software, 40, https://doi.org/10.18637/jss.v040.i08, 2011.

Eddelbuettel, D., François, R., Allaire, J. J., Ushey, K., Kou, Q., Russell, N., Ucar, I. n., Bates, D., and Chambers, J.: Rcpp: Seamless R and C++ Integration, https://doi.org/10.32614/cran.package.rcpp, 2008.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 96, 226–231, http://www.dbs.informatik.uni-muenchen.de/dbs/project/publikationen/veroeffentlichun-gen.html, 1996.

Hahsler, M. and Piekenbrock, M.: dbscan: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Related Algorithms, https://doi.org/10.32614/cran.package.dbscan, 2015.

Hahsler, M., Piekenbrock, M., and Doran, D.: dbscan: Fast Density-Based Clustering with R, Journal of Statistical Software, 91, 1–30, https://doi.org/10.18637/jss.v091.i01, 2019.

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

Hartigan, J. A. and Wong, M. A.: Algorithm AS 136: A K-Means Clustering Algorithm, Applied Statistics, 28, 100, https://doi.org/10.2307/2346830, 1979.

Hinton, G. E. and Roweis, S.: Stochastic Neighbor Embedding, in: Advances in Neural Information Processing Systems, edited by Becker, S., Thrun, S., and Obermayer, K., vol. 15, MIT Press, https://proceedings.neurips.cc/paper_files/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf, 2002.

Hopkins, B. and Skellam, J. G.: A New Method for determining the Type of Distribution of Plant Individuals, Annals of Botany, 18, 213–227, https://doi.org/10.1093/oxfordjournals.aob.a083391, 1954.

Karatzoglou, A., Smola, A., and Hornik, K.: kernlab: Kernel-Based Machine Learning Lab, https://doi.org/10.32614/cran.package.kernlab, 2004a.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A.: kernlab – An S4 Package for Kernel Methods in R, Journal of Statistical Software, 11, 1–20, https://doi.org/10.18637/jss.v011.i09, 2004b.

Kaufman, L. and Rousseeuw, P. J.: Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, ISBN 9780470316801, https://doi.org/10.1002/9780470316801, 1990.

Ketcham, R. A.: Forward and Inverse Modeling of Low-Temperature Thermochronometry Data, Reviews in Mineralogy and Geochemistry, 58, 275–314, https://doi.org/10.2138/rmg.2005.58.11, 2005.

Ketcham, R. A.: Thermal history inversion from thermochronometric data and complementary information: New methods and recommended practices, Chemical Geology, 653, 122 042, https://doi.org/10.1016/j.chemgeo.2024.122042, 2024.

Krijthe, J. H.: Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation, https://github.com/jkrijthe/Rtsne, r package version 0.17, 2015.

Kruskal, J. B.: Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis, Psychometrika, 29, 1–27, https://doi.org/10.1007/bf02289565, 1964.

Kruskal, J. B. and Wish, M.: Multidimensional Scaling, SAGE Publications, Inc., ISBN 9781412985130, https://doi.org/10.4135/9781412985130, 1978.

Lawson, R. G. and Jurs, P. C.: New index for clustering tendency and its application to chemical problems, Journal of Chemical Information and Computer Sciences, 30, 36–41, https://doi.org/10.1021/ci00065a010, 1990.

Maechler, M., Rousseeuw, P., Struyf, A., and Hubert, M.: cluster:]]Finding Groups in Data]]: Cluster Analysis Extended Rousseeuw et al., https://doi.org/10.32614/cran.package.cluster, 1999.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K.: cluster: Cluster Analysis Basics and Extensions, https://CRAN.R-project.org/package=cluster, r package version 2.1.8.1 — For new features, see the 'NEWS' and the 'Changelog' file in the package source), 2025.

McInnes, L., Healy, J., and Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, https://doi.org/10.48550/ARXIV.1802.03426, 2018.

McQuitty, L. L.: Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data, Educational and Psychological Measurement, 26, 825–831, https://doi.org/10.1177/001316446602600402, 1966.

Melville, J.: uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction, https://doi.org/10.32614/cran.package.uwot, 2019.

Morrow, D. W., Potter, J., Richards, B., and Goodarzi, F.: Paleozoic burial and organic maturation in the Liard Basin Region, northern Canada, Bulletin of Canadian Petroleum Geology, 41, 17–31, https://doi.org/10.35767/gscpgbull.41.1.017, 1993.

https://doi.org/10.5194/egusphere-2026-1279
Preprint. Discussion started: 26 March 2026

Murray, K. E., Goddard, A. L. S., Abbey, A. L., and Wildman, M.: Thermal history modeling techniques and interpretation strategies: Applications using HeFTy, Geosphere, 18, 1622–1642, https://doi.org/10.1130/ges02500.1, 2022.

Nakai, S., Halliday, A. N., Kesler, S. E., Jones, H. D., Kyle, J. R., and Lane, T. E.: Rb-Sr dating of sphalerites from Mississippi Valley-type (MVT) ore deposits, Geochimica et Cosmochimica Acta, 57, 417–427, https://doi.org/10.1016/0016-7037(93)90440-8, 1993.

Ng, A., Jordan, M., and Weiss, Y.: On Spectral Clustering: Analysis and an algorithm, in: Advances in Neural Information Processing Systems, edited by Dietterich, T., Becker, S., and Ghahramani, Z., vol. 14, MIT Press, https://proceedings.neurips.cc/paper_files/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf, 2001.

Pebesma, E.: Simple Features for R: Standardized Support for Spatial Vector Data, The R Journal, 10, 439–446, https://doi.org/10.32614/RJ-2018-009, 2018.

Pebesma, E. and Bivand, R.: Spatial Data Science: With applications in R, Chapman and Hall/CRC, https://doi.org/10.1201/9780429459016, 2023.

Pinto, T. F., Enkelmann, E., Quadri, S. M. T., and Terlaky, V.: Project Summary — Year 5: Thermal evolution of Phanerozoic sediments of the southwestern Northwest Territories, resreport, Northwest Territories Geological Survey, https://doi.org/10.46887/2025-008, 2025.

Rousseeuw, P. J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, 20, 53–65, https://doi.org/10.1016/0377-0427(87)90125-7, 1987.

Rucklidge, W. J.: Lower bounds for the complexity of the graph of the Hausdorff distance as a function of transformation, Discrete & Computational Geometry, 16, 135–153, https://doi.org/10.1007/bf02716804, 1996.

Sammon, J. W.: A Nonlinear Mapping for Data Structure Analysis, IEEE Transactions on Computers, C-18, 401–409, https://doi.org/10.1109/t-c.1969.222678, 1969.

Stevens Goddard, A. L., Fosdick, J. C., Calderón, M., Ghiglione, M. C., VanderLeest, R. A., and Romans, B. W.: Thermochronological Evidence for Eocene Deformation in the Southern Patagonian Andes: Linking Orogenesis Along the Patagonian Orocline, Tectonics, 42, https://doi.org/10.1029/2022tc007677, 2023.

Szmihelsky, M., Steele-MacInnis, M., Bain, W. M., Falck, H., Adair, R., Campbell, B., Dufrane, S. A., Went, A., and Corlett, H. J.: Mixing of brine with oil triggered sphalerite deposition at Pine Point, Northwest Territories, Canada, Geology, 49, 488–492, https://doi.org/10.1130/g48259.1, 2020.

Taha, A. A. and Hanbury, A.: An Efficient Algorithm for Calculating the Exact Hausdorff Distance, IEEE Transactions on Pattern Analysis and Machine Intelligence, 37, 2153–2163, https://doi.org/10.1109/TPAMI.2015.2408351, 2015.

Takács, B.: Comparing face images using the modified Hausdorff distance, Pattern Recognition, 31, 1873–1881, https://doi.org/10.1016/s0031-3203(98)00076-4, 1998.

van der Maaten, L. and Hinton, G.: Visualizing Data using t-SNE, Journal of Machine Learning Research, 9, 2579–2605, http://jmlr.org/papers/v9/vandermaaten08a.html, 2008.

Venables, W. N. and Ripley, B. D.: Modern Applied Statistics with S, Springer, New York, fourth edn., https://www.stats.ox.ac.uk/pub/MASS4/, iSBN 0-387-95457-0, 2002.

Wright, K.: Will the Real Hopkins Statistic Please Stand Up?, The R Journal, 14, 282–292, https://doi.org/10.32614/rj-2022-055, 2022a.

Wright, K.: hopkins: Calculate Hopkins Statistic for Clustering, https://doi.org/10.32614/cran.package.hopkins, 2022b.

Yu, Y., Jiang, H., Zhang, X., and Chen, Y.: Identifying Irregular Potatoes Using Hausdorff Distance and Intersection over Union, Sensors, 22, 5740, https://doi.org/10.3390/s22155740, 2022.