Reducing Ambiguities in Line-based Density Plots by Image-space Colorization — Supplemental Material

Yumeng Xue, Patrick Paetzold, Rebecca Kehlbeck, Bin Chen, Kin Chung Kwan, Yunhai Wang, and Oliver Deussen

1 Partial derivative

The partial derivative of the stress S with respect to the angle θ_k of a point k is:

$$\frac{\partial S}{\partial \theta_k} = S\sum_{t=1}^{n}\left(\frac{\hat{A}_t}{\sum_{i<j}\left(\delta\left(x_{i,j}\right)-d_{i,j}\right)^2} - \frac{\hat{B}_t}{\sum_{i<j} d_{i,j}^2}\right) \tag{1}$$

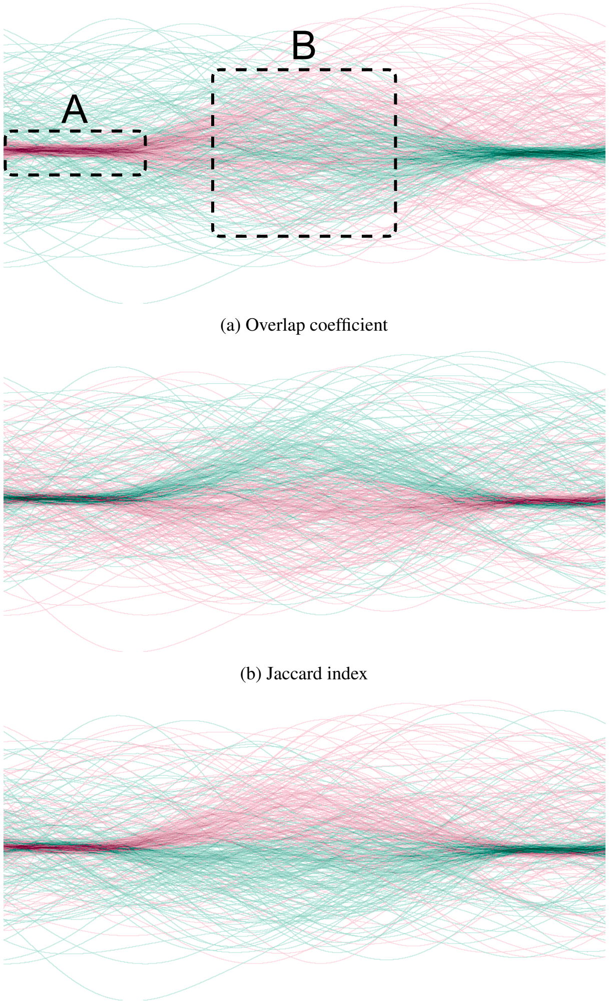
Here, $\hat{A}_t$ and $\hat{B}_t$ are defined as follows:

$$\begin{aligned}\hat{A}_t &= \begin{cases} -d_{t,k}+\delta\left(x_{t,k}\right), & \left|\theta_t-\theta_k\right| \leq \pi \text{ and } \theta_t \geq \theta_k \\ d_{t,k}-\delta\left(x_{t,k}\right), & \left|\theta_t-\theta_k\right| \leq \pi \text{ and } \theta_t < \theta_k \\ d_{t,k}-\delta\left(x_{t,k}\right), & \left|\theta_t-\theta_k\right| > \pi \text{ and } \theta_t \geq \theta_k \\ -d_{t,k}+\delta\left(x_{t,k}\right), & \left|\theta_t-\theta_k\right| > \pi \text{ and } \theta_t < \theta_k \end{cases} \\ \hat{B}_t &= \begin{cases} -d_{t,k}, & \left|\theta_t-\theta_k\right| \leq \pi \text{ and } \theta_t \geq \theta_k \\ d_{t,k}, & \left|\theta_t-\theta_k\right| \leq \pi \text{ and } \theta_t < \theta_k \\ d_{t,k}, & \left|\theta_t-\theta_k\right| > \pi \text{ and } \theta_t \geq \theta_k \\ -d_{t,k}, & \left|\theta_t-\theta_k\right| > \pi \text{ and } \theta_t < \theta_k \end{cases}\end{aligned} \tag{2}$$

2 Illustration of Different Similarity Metrics

Here we compare the effect of different metrics for measuring similarity between sets on our method. As mentioned in our paper, in addition to the overlap coefficient we finally used, we also tried the commonly used Jaccard index $\mathrm{J}(A,B) = \frac{|A \cap B|}{|A \cup B|}$ and Sørensen-Dice coefficient $\mathrm{DSC}(A,B) = \frac{2|A \cap B|}{|A|+|B|}$ as similarity metrics. A ratio close to 1 indicates a high degree of similarity. We use a simple example to illustrate the difference between these metrics and explain why the overlap factor is more applicable to our problem.

We use each of the three similarity metrics to cluster the pixels using the same parameters. As shown in Fig. 1, they exhibit different clustering qualities. This is due to the fact that in a density plot, the patterns that attract attention are generally denser than the surroundings, so the similarity between bins should be measured using a metric that is insensitive to the size of the feature set. Taking this data as an example, there are two sets of line bundles. The lines in Fig. 1 are clustered in area “A”, while in area “B” the lines are sparse. These two line bundles each consist of 200 lines. We use 0 to 199 for the line number of the first line bundle and 200 to 399 for the line number of the other line bundle. As an example, suppose we take a bin from the dense part of “A” whose feature set S_1 is $\{1,4,5,6,8,10,15,23,30\}$. It’s obviously a bin belong to line bundle 1. Then, suppose we take two bins from the sparse part “B”, and they are $\mathrm{S}_2 = \{8,23,289\}$ and $\mathrm{S}_3 = \{289,356\}$. Since the area “B” is the area where the two line bundles cross, it is possible for one bin to include lines from both line bundles at the same time. Thus, we have the following results: $\mathrm{J}(\mathrm{S}_1,\mathrm{S}_2) = 0.2$, $\mathrm{J}(\mathrm{S}_2,\mathrm{S}_3) = 0.25$, $\mathrm{DSC}(\mathrm{S}_1,\mathrm{S}_2) = 0.3333$, and $\mathrm{DSC}(\mathrm{S}_2,\mathrm{S}_3) = 0.4$. We can see that since both coefficients are influenced by the set size, their results both show that S_2 and S_3 are more similar to each other than S_1 and S_2. This leads to an undesirable clustering of pixels for our use case because we expect S_1 to be clustered with S_2 since $2/3$ of the line IDs in S_2 are also included in S_1. This will cause the bin in area “B” to be grouped into the same cluster as its neighboring bins, which leads to the undesirable clustering results in Fig. 1b and 1c. The overlap coefficient can solve this issue very well. As $\mathrm{overlap}(\mathrm{S}_1,\mathrm{S}_2) = 0.6667$ and $\mathrm{overlap}(\mathrm{S}_2,\mathrm{S}_3) = 0.5$, the similarity between S_1 and S_2 is higher than S_2 and S_3. Since the overlap coefficient is not affected by the



(a) Overlap coefficient

(b) Jaccard index

(c) Sørensen-Dice coefficient

Fig. 1: Clustering results of different metrics

- *Yumeng Xue is with University of Konstanz and Shandong University. E-mail: yumeng.xue@uni-konstanz.de.*
- *Patrick Paetzold, Rebecca Kehlbeck, Bin Chen, and Oliver Deussen are with University of Konstanz. E-mail: firstname.lastname@uni-konstanz.de.*
- *Kin Chung Kwan is with California State University Sacramento. E-mail: kwan@csus.edu*
- *Yunhai Wang is with Shandong University. E-mail: cloudseawang@gmail.com.*
- *Oliver Deussen and Yunhai Wang are joint corresponding authors.*

difference in set sizes, it is well-suited to calculate the similarity of bins with different densities. It generates the more desirable clustering results shown in Fig. 1a.

3 User Study Details

Figs. 3 to 12 show, on the one hand, the line-based density plots we used in our motivational user study and, on the other hand, the individual trends the line-based density is composed of. The detailed figures of the 10 datasets of the user study are given in Figs. 3-5 (illusionary patterns), Figs. 6-8 (ambiguous continuation), and Figs. 9-12 (disconnected clusters).

4 Case Study Details

The line-based plot of Hellenic Trench AIS dataset [3] is shown in Fig. 2. Figs. 13a and 13b show the two additional by our method separated clusters of temperature datasets which we could not include in the paper due to space limitations.

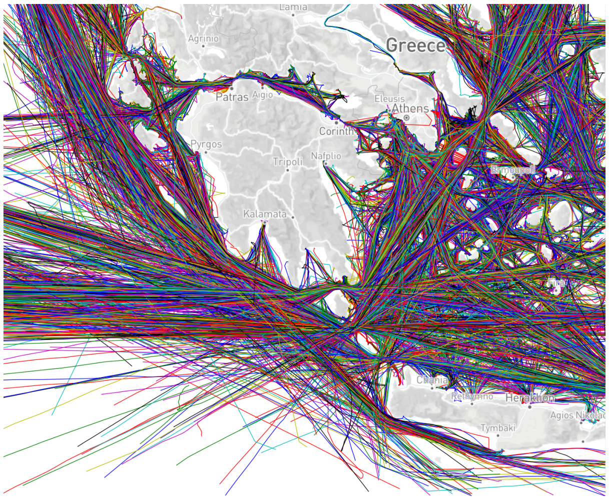


Fig. 2: Line-based plot of the Hellenic Trench AIS data.

Similarly, Figs. 14 and 15 show the remaining separated clusters for the Hellenic Trench AIS dataset, and the Beijing Taxi Trajectory dataset.

5 Comparison to Line Clustering Approaches on Real-world Datasets

Due to space constraints, we could not include the Figs. 16 - 19 comparing the results of the line clustering method to our results in our paper. We utilize the line clustering method [2], which was previously employed in Section 5.1 of our paper, to process the four real-world datasets mentioned in our paper. We use the same number of clusters for the line clustering method as in the case study in Section 5 of our paper. These results in Figs. 16 - 19 demonstrate the main difference between our method and the line clustering method, i.e., our method is more concerned with patterns in the density plot and is a visual clustering that finds "visible" clusters, while the line clustering method is more concerned with the data itself and tends to find clusters that are "invisible".

Temperature data. Fig. 16 shows the five individual clusters identified by the line-based clustering method. It divides the majority of lines into the three different clusters shown in Figs. 16c - 16e. These clustering results do not clearly show that the line bundles in the separate Figures cross. However, our method is able to show the crossing marked by "A" in Fig. 4c in the paper. The clusters in Fig. 16d and 16e look very similar and thus provide only limited additional insight. Besides, the clusters shown in Fig. 16a and 16b contain only a small number of shorter lines, which is because the line clustering method considers the overall characteristics of the lines, so the length of the lines can have a significant impact on the clustering results.

Stock market data. For the stock market dataset, the line clustering method generates results similar to our method. However, there are still differences between the results of the line clustering method shown in Fig. 17 and the results generated by our method in Fig. 5c of the paper. Unlike our method, which focuses on patterns, the line clustering method focuses on the global characteristics of the line, Fig. 17a, only contains the stocks listed from 2012, and there are many lines that are not part of the pattern. Our method, on the other hand, identifies the lines that form this pattern as much as possible, as evidenced in Fig. 17b, where the pattern to the right (highlighted as "A") is still partially present, and in Fig. 5e of our paper this pattern does not exist.

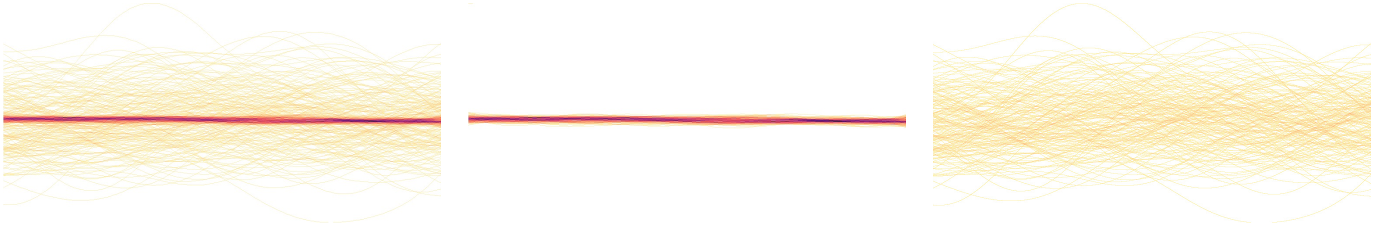
Ship trajectories. The line clustering method does not separate the underlying patterns present in the ship trajectory dataset well, as multiple subfigures of Fig. 18 contain dense, superimposed patterns. Figs. 18e and 18g are visually very similar and do not successfully decompose the "U"-shaped pattern. Whereas our method is able to show that the "U"-shape is actually not a continuous pattern but instead consists of multiple combining and diverging bundles of trajectories. In addition, we can see that the line clustering method generates mainly two groups of visually similar density plots, which also indicates that it is not effective in discovering additional meaningful clusters.

Taxi trajectories. Applied to the taxi trajectory dataset, the line clustering method splits the central ring-shaped pattern present in Fig. 1b in our paper. However, the line-based clustering shown in Fig. 19 only extracted the shorter trajectories along the airport highway (Fig. 19b). This further corroborates the line clustering method's focus on the overall characteristics of the lines rather than the high-density patterns. Also, in Fig. 19d, the high-density pattern along the airport highway does not disappear. In our approach, we successfully extracted trajectories along the airport highway and towards various parts of the city (Fig. 7a of our paper), while in other clusters (Fig. 7b of our paper and Fig. 15), the high-density patterns along the airport highway disappeared.

6 Further Examples of Our Method

We present three additional synthetic datasets (Fig. 20, 21, and 22) containing a total of 400, 900, and 2500 lines distributed in 2, 3, and 5 bundles, respectively. The first two datasets demonstrate the characteristics of illusionary patterns and ambiguous continuation and are disambiguated effectively by our method. The third synthetic dataset (Fig. 22) is more complex, consisting of five line bundles with 500 lines each, combined in a visually ambiguous way. Applying our method separates the three line bundles (Fig. 23c, 23d, and 23e) with 485, 492, and 505 lines, which closely approximate their actual size. However, for the filtered line clusters in Fig. 23a and 23b, it can be seen that their number of lines is more different from 500, which is because the two bundles have more overlap. Some lines were therefore assigned to the wrong cluster. Nonetheless, the trends can still be observed. Overall, these synthetic datasets demonstrate the efficacy of our method.

In addition to the already introduced real-world dataset mentioned in the paper, Fig. 24c shows the results of our method applied to an additional time series dataset [1]. It contains temperature records (SMART 194) for millions of hard drives. Of the 10,000 drives we sampled, 9982 drives remained for analysis after data cleaning. Fig. 24 shows the plot of the raw lines, a line-based density plot, and the results generated by our method. The four filtered clusters are given in Fig. 25. As we can see, the lines filtered by the magenta and brown cluster (Fig. 25b and 25d) are not present within the whole time interval. This cannot be seen from the traditional line-based plot (Fig. 24a) and the line density plot in Fig. 24b. We also applied our method to the Mediterranean Sea Trajectory dataset [4]. It simulates the horizontal movement of larvae of different fish species on the water surface. We used 10021 trajectories of this dataset. The results obtained by plotting the raw lines, the line-based density plot, and our method result are shown in Fig. 26. Several filtered clusters are given in Fig. 27. It can be seen that compared to the traditional density map, our method is able to show cluster boundaries more clearly, and lines corresponding to the clusters are filtered out well.

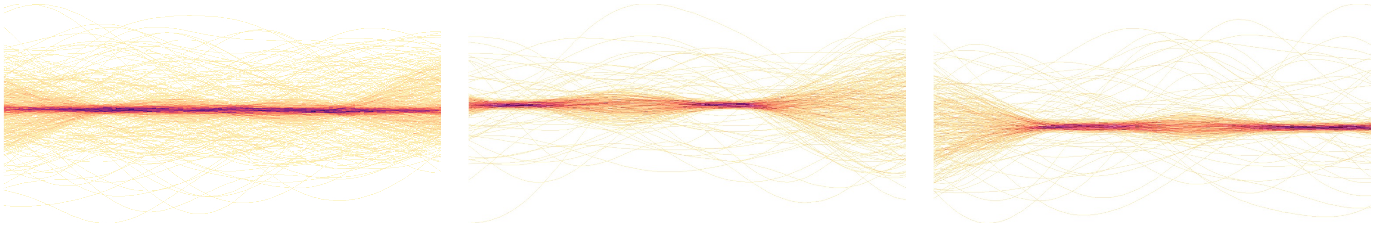


(a) Image for user study

(b) Main trend

(c) Noise

Fig. 3: User study: Dataset 1 (illusionary patterns).

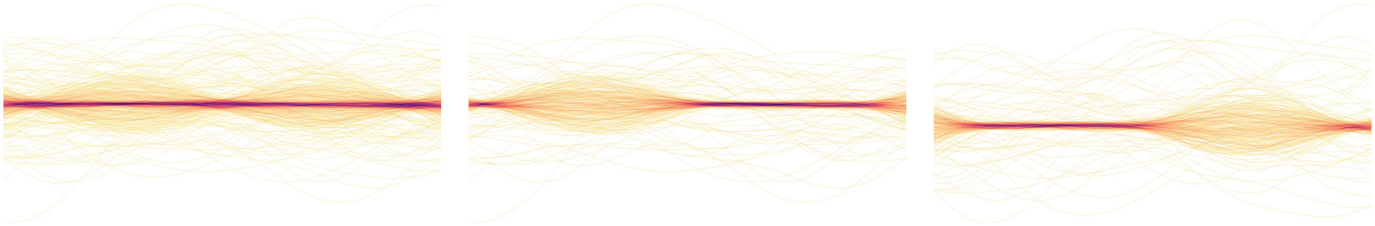


(a) Image for user study

(b) Trend 1 with noise

(c) Trend 2 with noise

Fig. 4: User study: Dataset 2 (illusionary patterns).

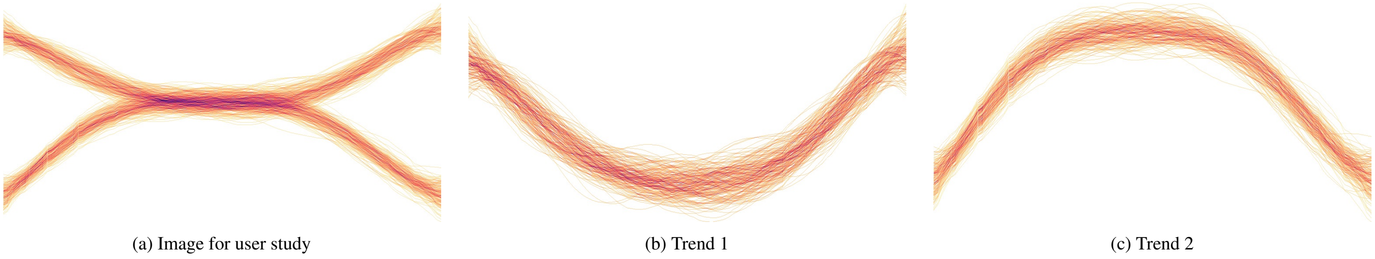


(a) Image for user study

(b) Trend 1 with noise

(c) Trend 2 with noise

Fig. 5: User study: Dataset 3 (illusionary patterns).

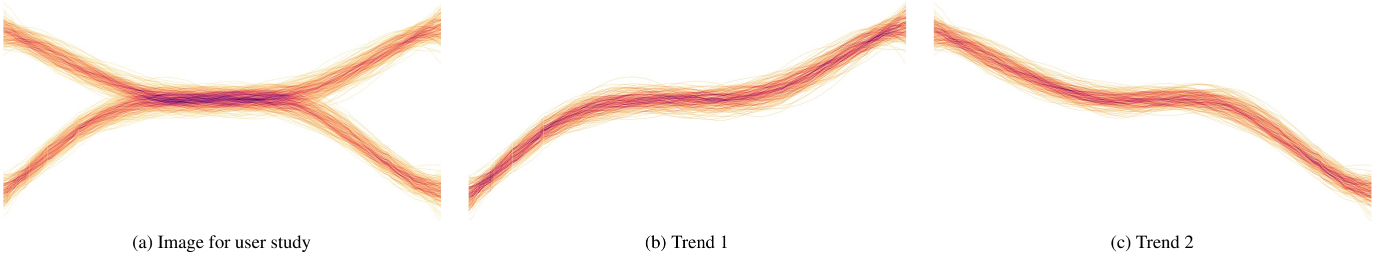


(a) Image for user study

(b) Trend 1

(c) Trend 2

Fig. 6: User study: Dataset 4 (ambiguous continuation).

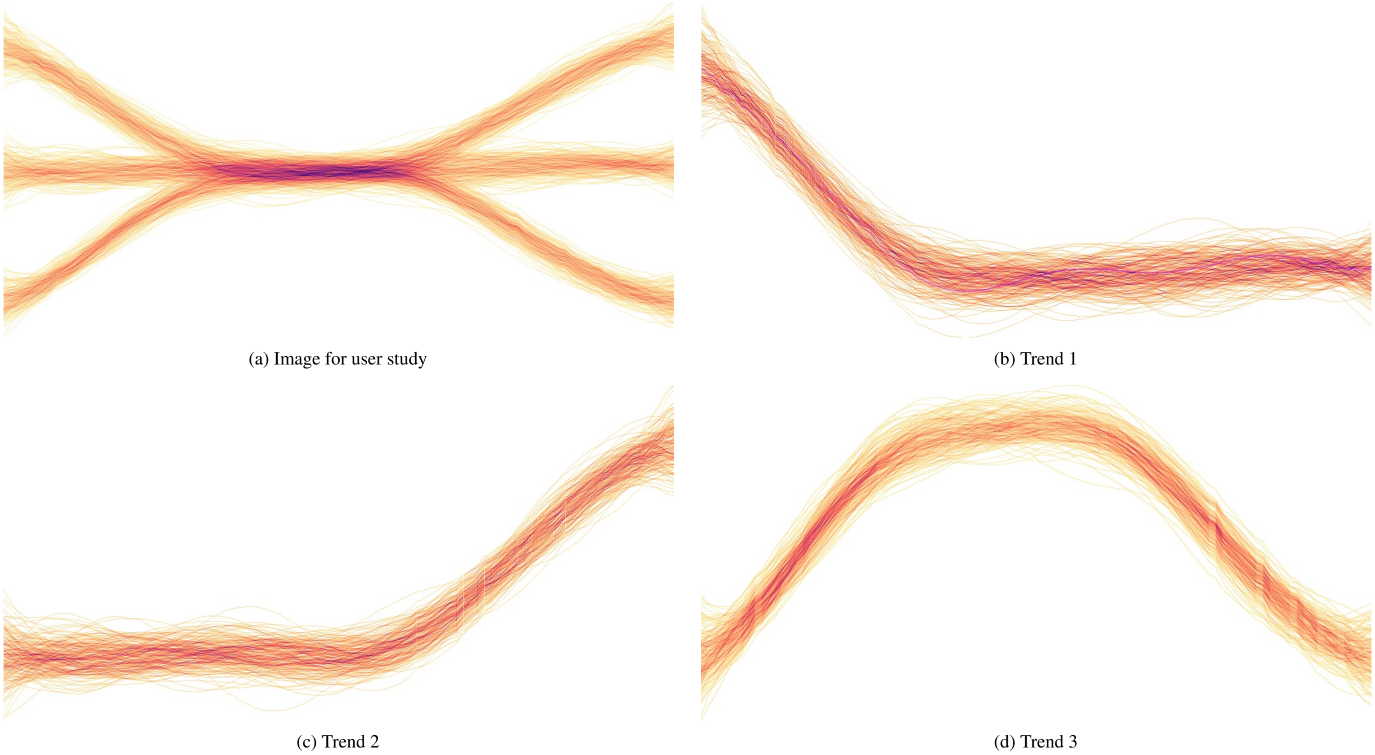


(a) Image for user study

(b) Trend 1

(c) Trend 2

Fig. 7: User study: Dataset 5 (ambiguous continuation).



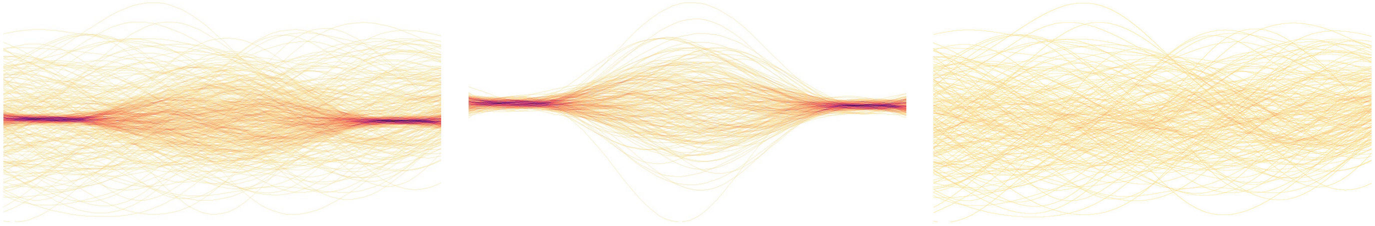
(a) Image for user study

(b) Trend 1

(c) Trend 2

(d) Trend 3

Fig. 8: User study: Dataset 6 (ambiguous continuation).

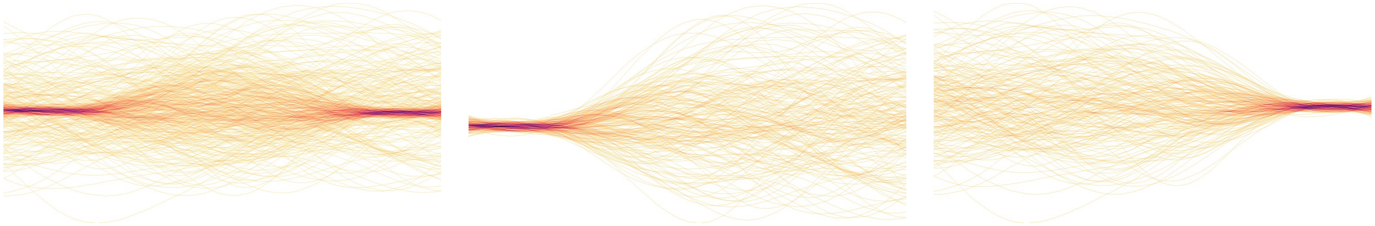


(a) Image for user study

(b) Main trend

(c) Noise

Fig. 9: User study: Dataset 7 (disconnected clusters).

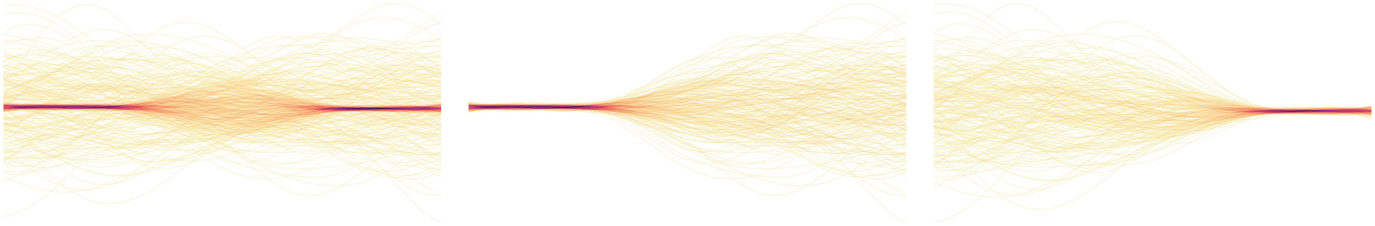


(a) Image for user study

(b) Trend 1

(c) Trend 2

Fig. 10: User study: Dataset 8 (disconnected clusters).

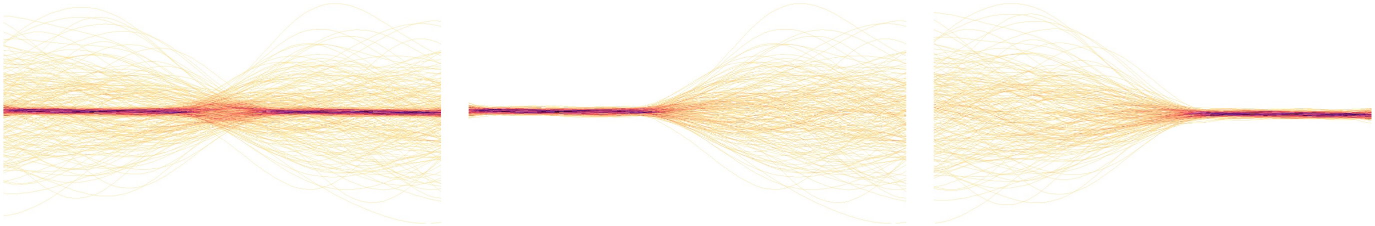


(a) Image for user study

(b) Trend 1

(c) Trend 2

Fig. 11: User study: Dataset 9 (disconnected clusters).

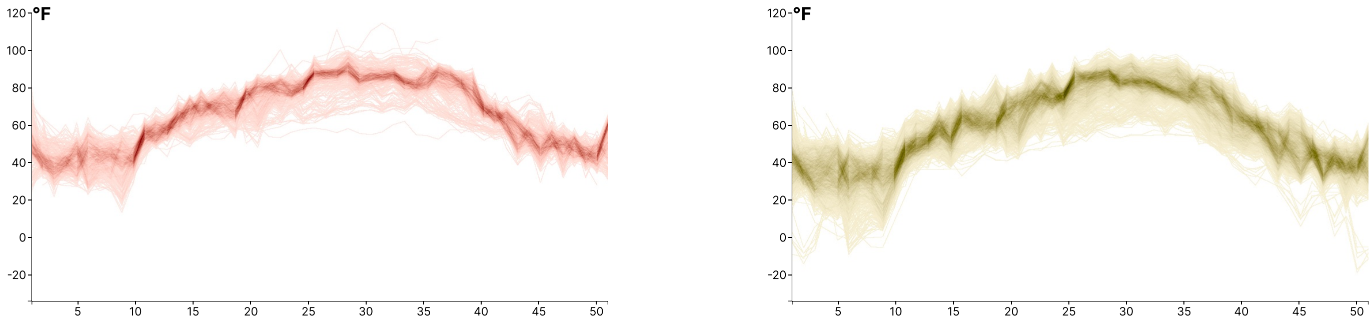


(a) Image for user study

(b) Trend 1

(c) Trend 2

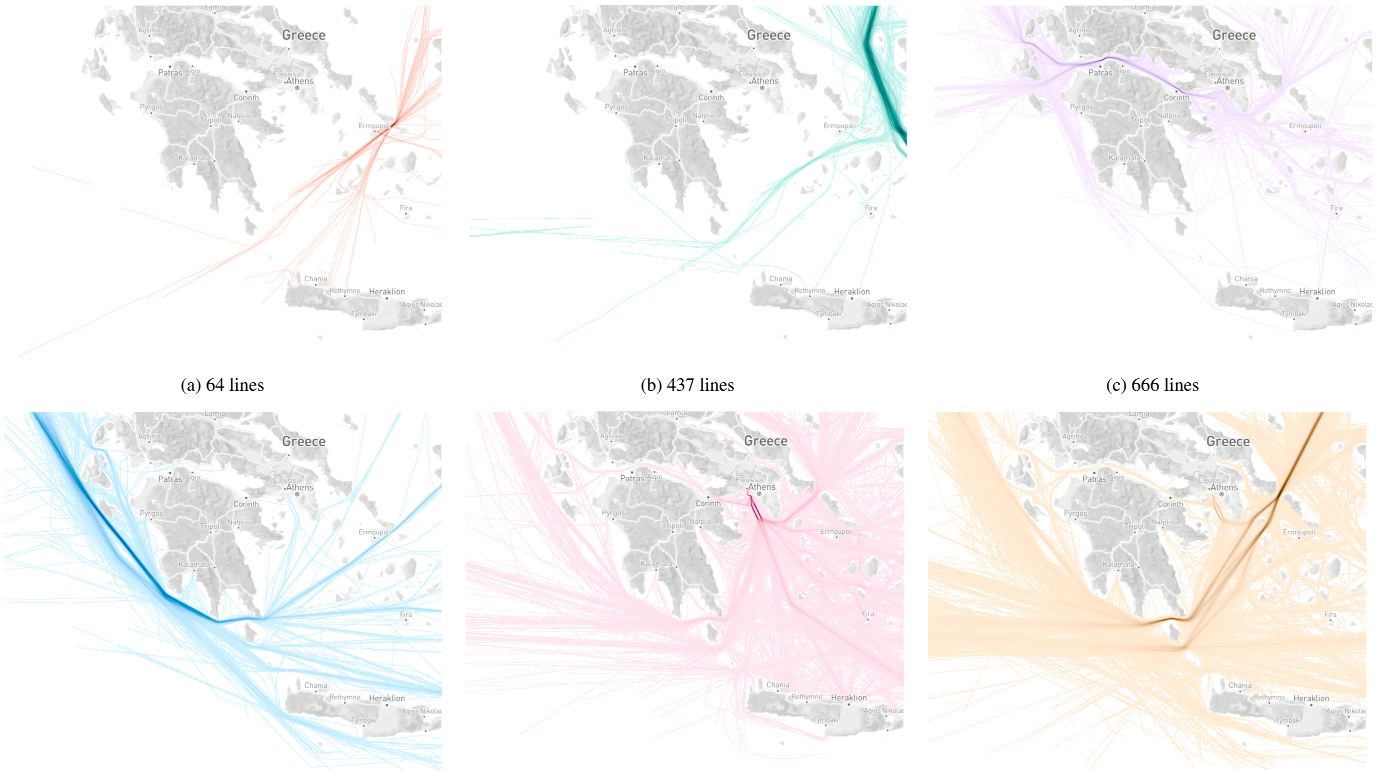
Fig. 12: User study: Dataset 10 (disconnected clusters).



(a) 642 lines

(b) 1830 lines

Fig. 13: Filtered lines from two clusters of the temperature data shown in use cases of our paper.



(a) 64 lines

(b) 437 lines

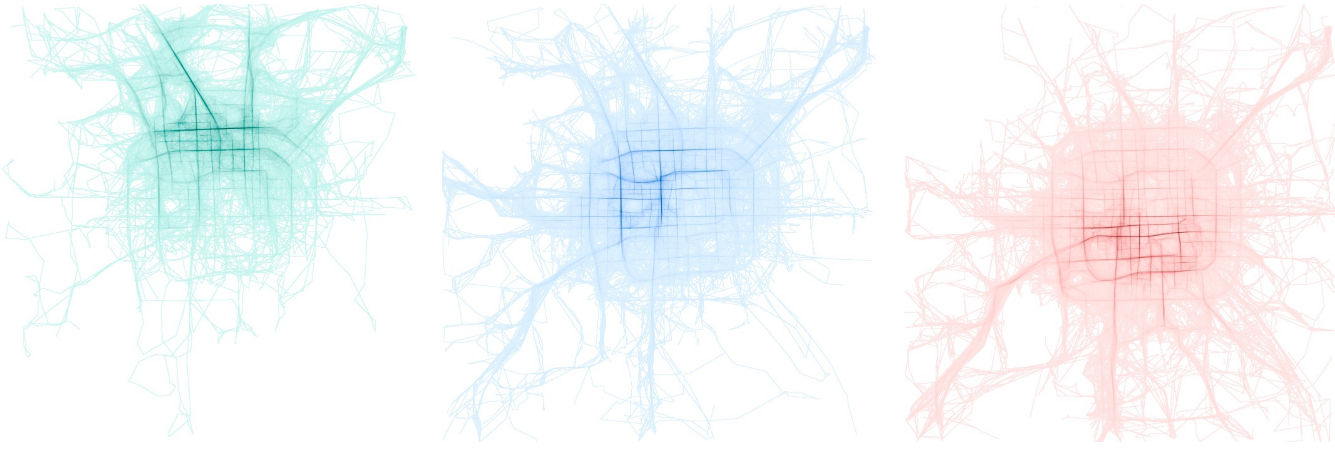
(c) 666 lines

(d) 868 lines

(e) 1804 lines

(f) 3557 lines

Fig. 14: Filtered lines from six clusters of the Hellenic Trench AIS data shown in the use cases of our paper.

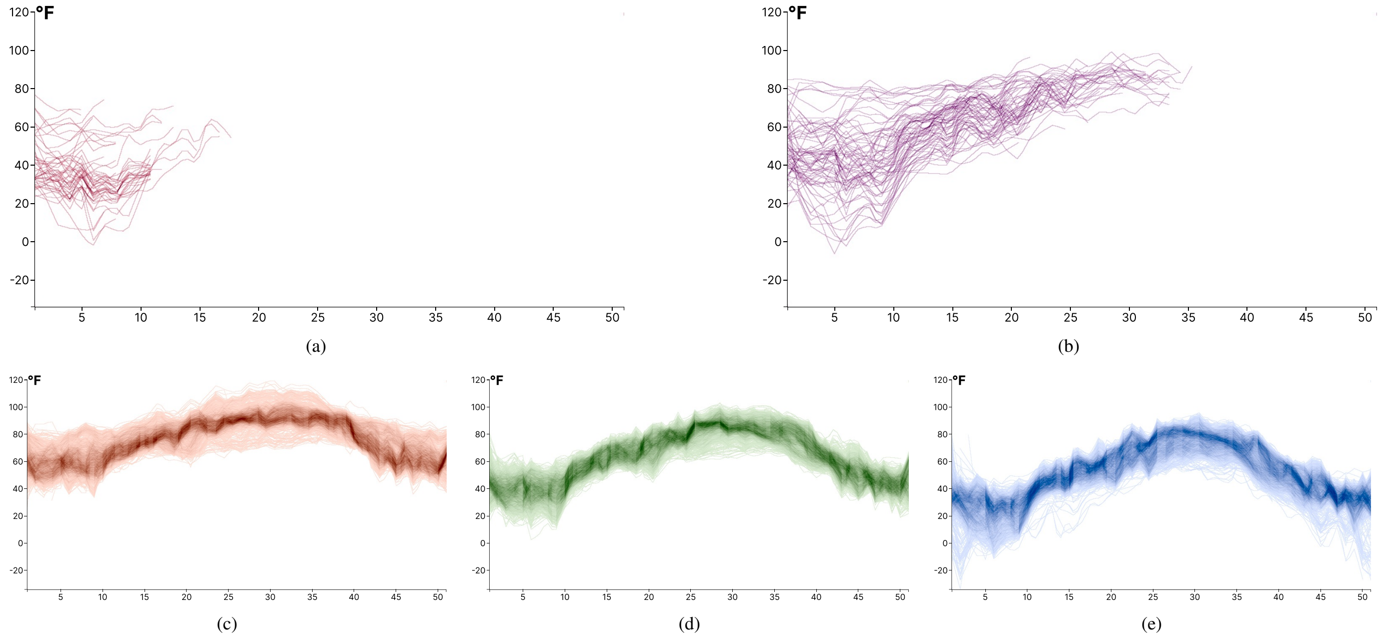


(a) 886 lines

(b) 1128 lines

(c) 1483 lines

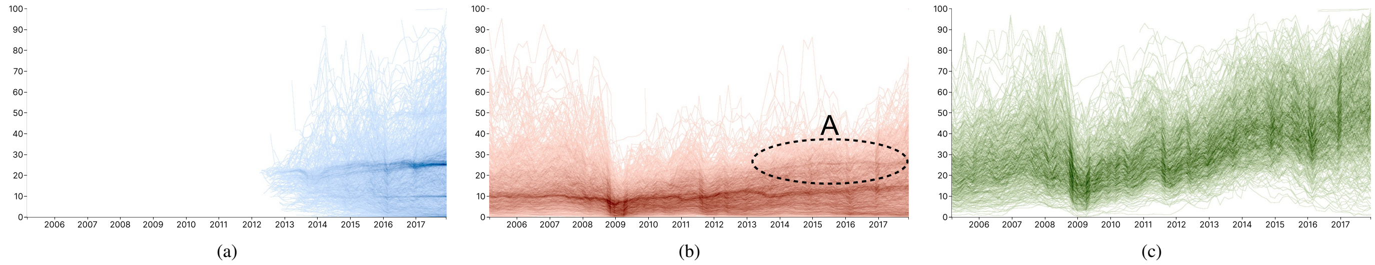
Fig. 15: Filtered lines from three clusters of the Beijing Taxi Trajectory data shown in the teaser of our paper.



(a) (b)

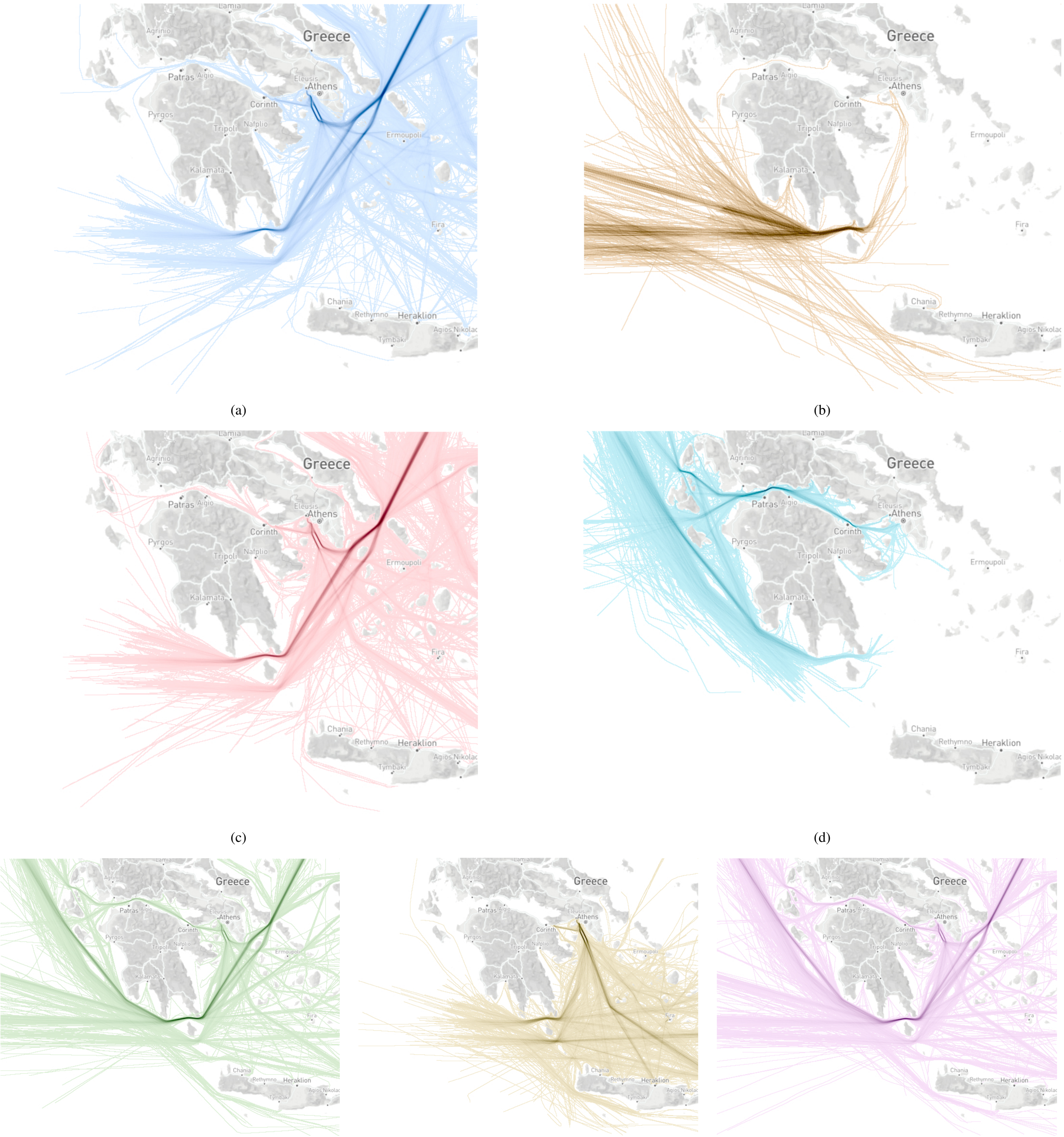
(c) (d) (e)

Fig. 16: Clustering results of the line clustering method on the temperature data.



(a) (b) (c)

Fig. 17: Clustering results of the line clustering method on the New York Stock Exchange data.



(a)

(b)

(c)

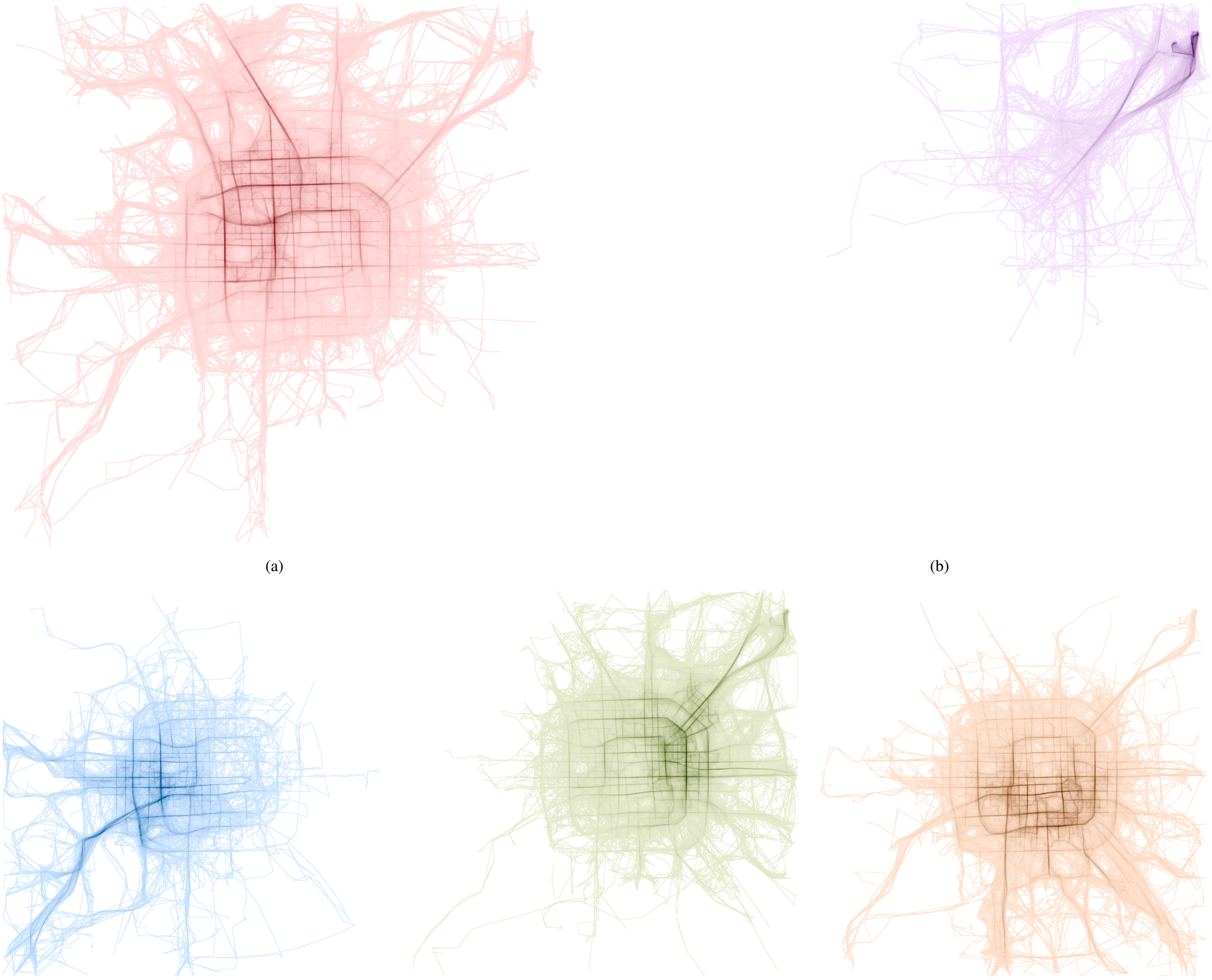
(d)

(e)

(f)

(g)

Fig. 18: Clustering results of the line clustering method on the Hellenic Trench AIS data.



(a)

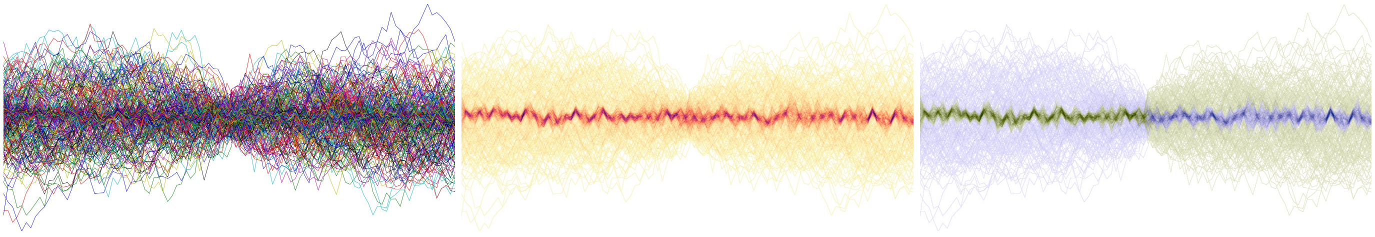
(b)

(c)

(d)

(e)

Fig. 19: Clustering results of the line clustering method on the Beijing Taxi Trajectory data.

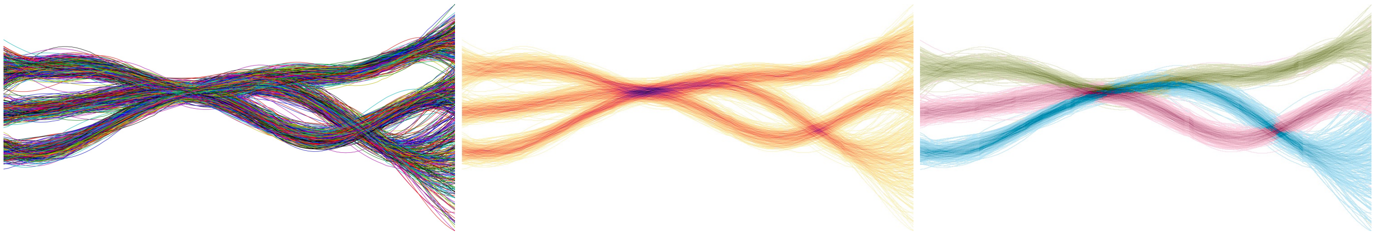


(a) Line-based plot

(b) Line-based density plot

(c) Ours: Colored line-based density plot

Fig. 20: Synthetic data consisting of two line bundles with a total of 400 lines.

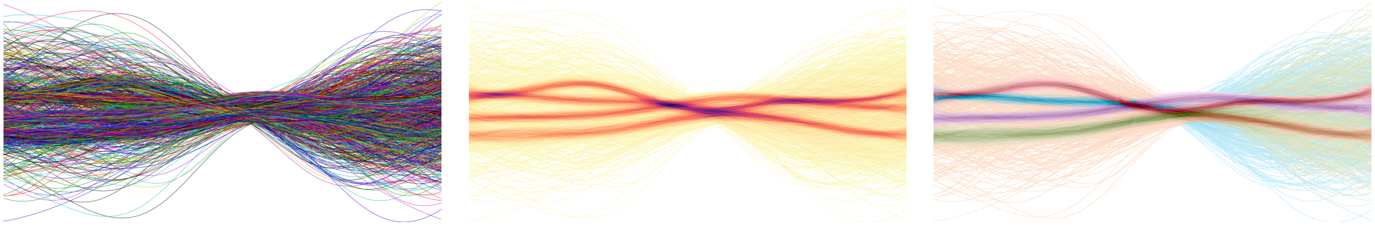


(a) Line-based plot

(b) Line-based density plot

(c) Ours: Colored line-based density plot

Fig. 21: Synthetic data consisting of three line bundles with a total of 900 lines.

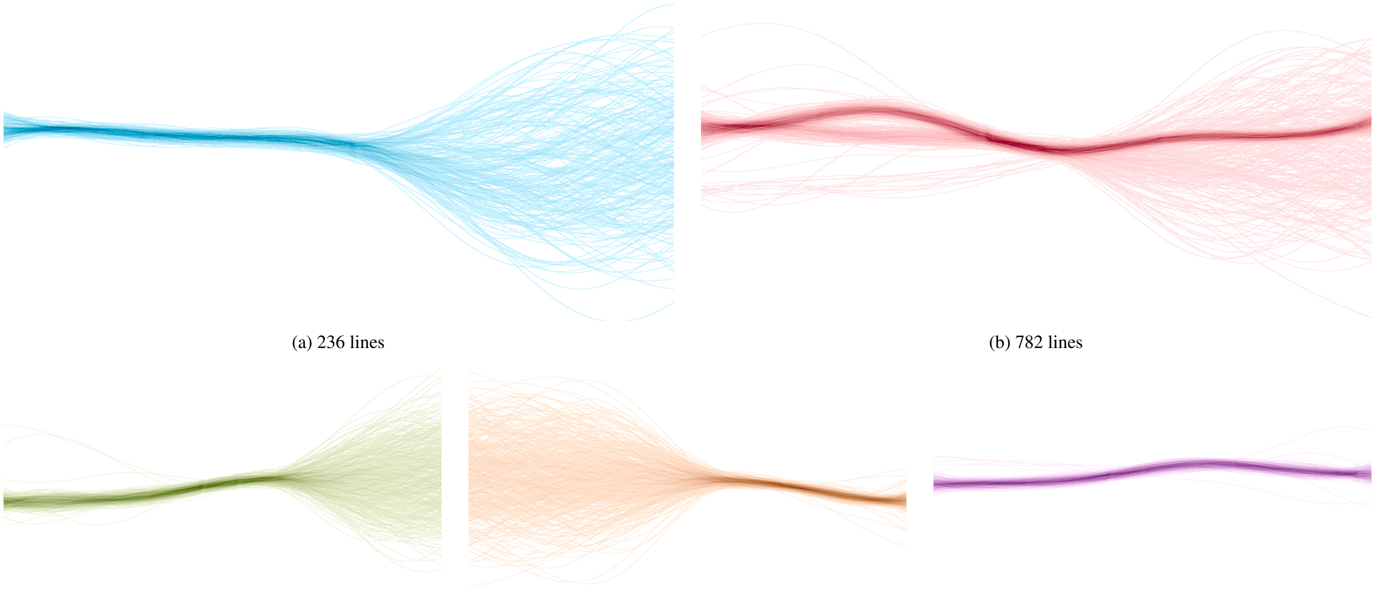


(a) Line-based plot

(b) Line-based density plot

(c) Ours: Colored line-based density plot

Fig. 22: Synthetic data consisting of five line bundles with a total of 2500 lines.



(a) 236 lines

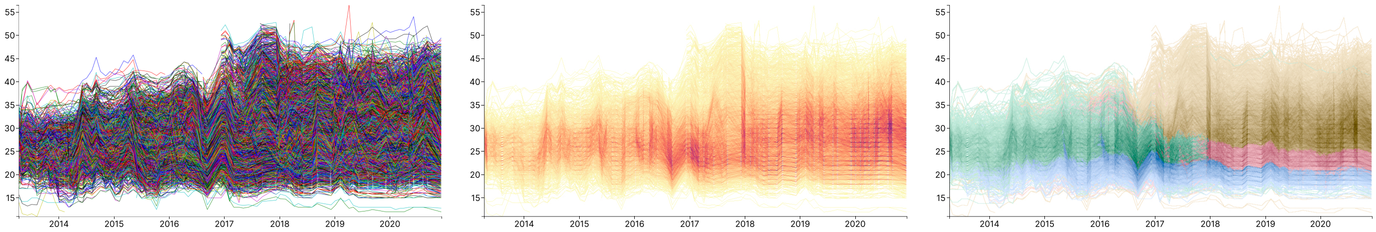
(b) 782 lines

(c) 485 lines

(d) 492 lines

(e) 505 lines

Fig. 23: Filtered lines from five clusters of the synthetic data in Fig. 22.

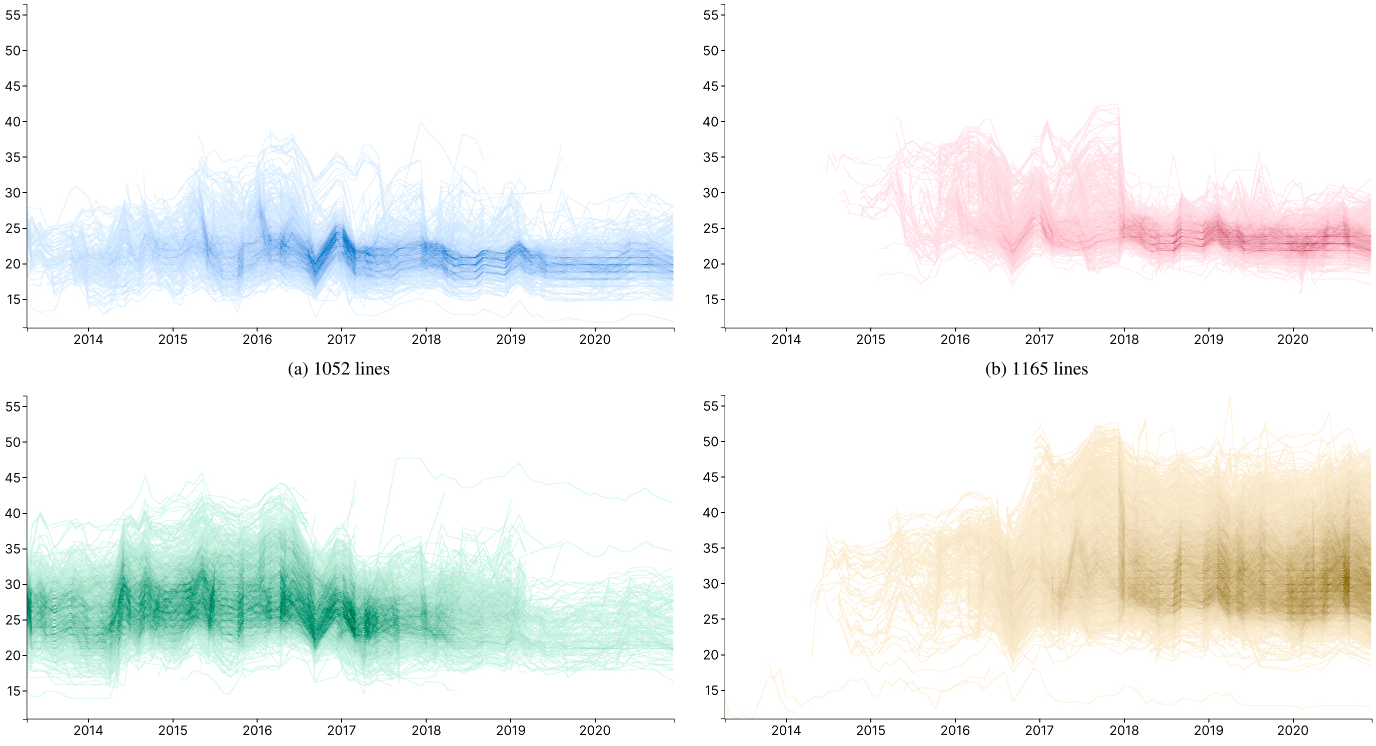


(a) Line-based plot

(b) Line-based density plot

(c) Ours: Colored line-based density plot

Fig. 24: Hard drive temperature data (9982 lines).



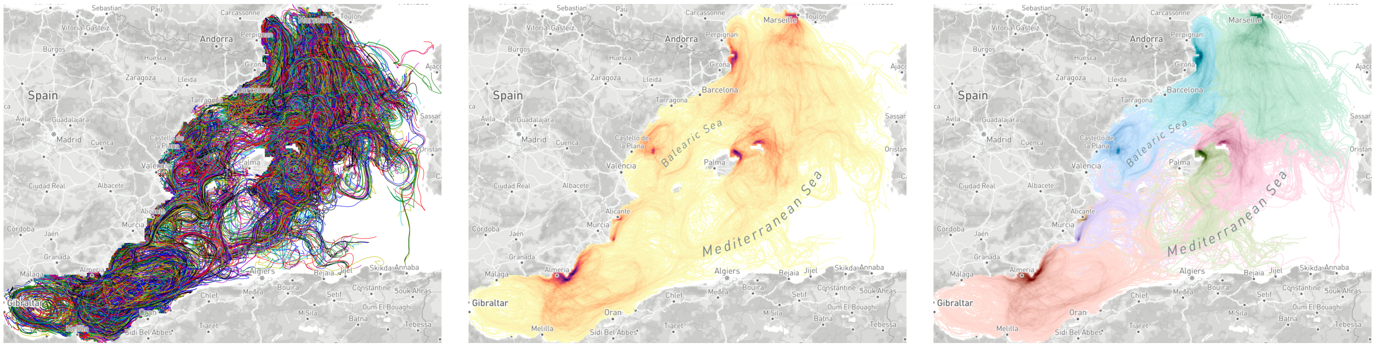
(a) 1052 lines

(b) 1165 lines

(c) 2091 lines

(d) 5674 lines

Fig. 25: Filtered lines from four clusters of the hard drive temperature data

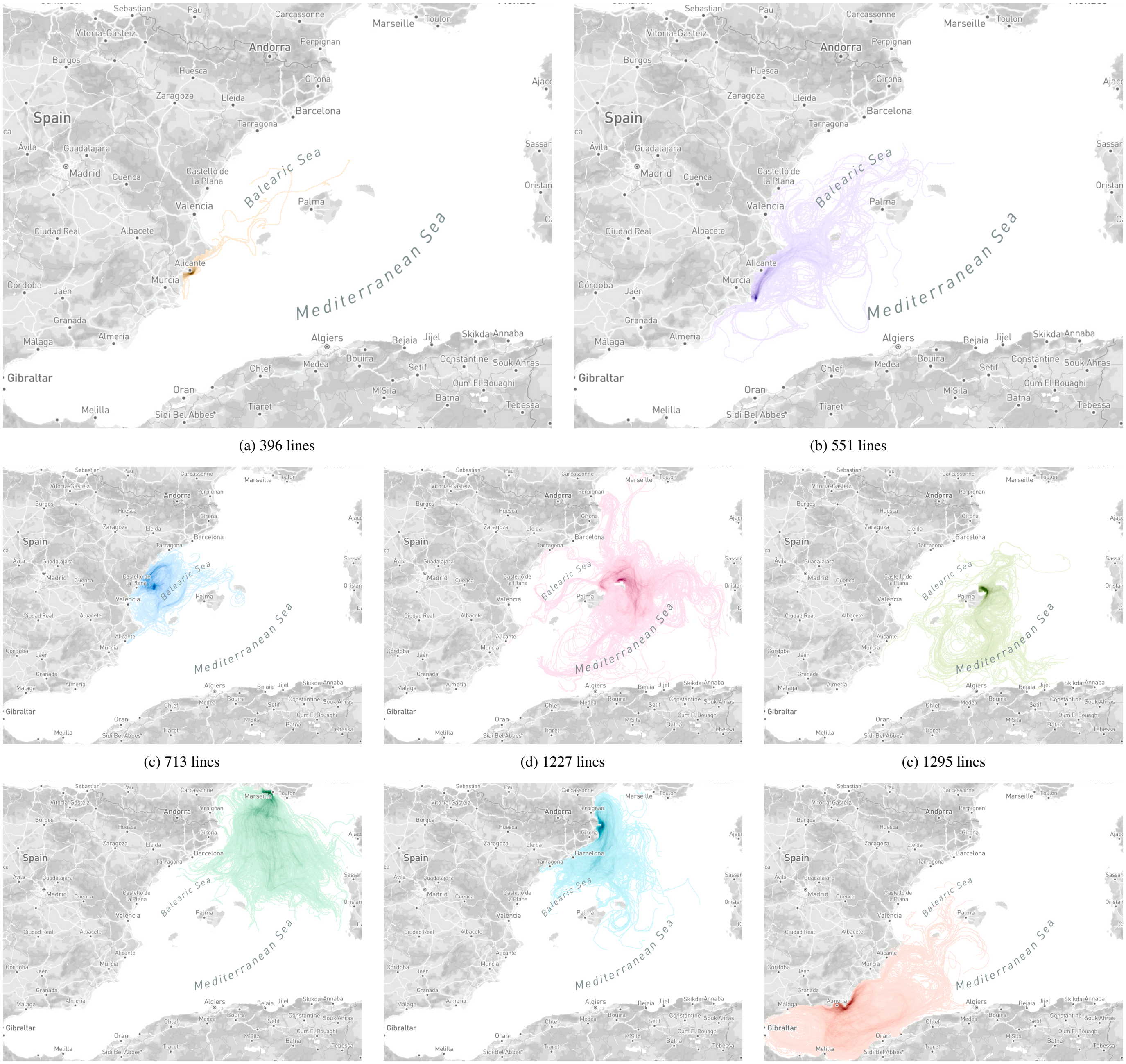


(a) Line-based plot

(b) Line-based density plot

(c) Ours: Colored line-based density plot

Fig. 26: Mediterranean Sea Trajectory data (10021 lines).



(a) 396 lines

(b) 551 lines

(c) 713 lines

(d) 1227 lines

(e) 1295 lines

(f) 1365 lines

(g) 1655 lines

(h) 2819 lines

Fig. 27: Filtered lines from eight clusters of the Mediterranean Sea Trajectory data

References

[1] Hard drive data and stats. (2013). https://www.backblaze.com/b2/hard-drive-test-data.html, 2013. 2

[2] F. Ferstl, K. Bürger, and R. Westermann. Streamline variability plots for characterizing the uncertainty in vector field ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):767–776, 2015. 2

[3] A. Frantzis, R. Leaper, P. Alexiadou, A. Prospathopoulos, and D. Lekkas. Hellenic trench ais data. 2018. doi: 10.17882/57040 2

[4] W. Rath, C. Schmidt, and S. Rühs. Mediterranean sea trajectory data examples. http://dx.doi.org/10.5281/zenodo.4650317, 2021. doi: 10.5281/zenodo.4650317 2