sciendo

Forestry Studies | Metsanduslikud Uurimused, Vol. 82, Pages 20–27

Research paper

Assessment of an optimal parameter space for spatial cluster detection of SMEAR Estonia flux footprint data using unsupervised learning algorithms

Steffen M. Noe[1]*, Anuj Thapa Magar[1] and Emílio Graciliano Ferreira Mercuri[1,2]

Noe, S.M., Thapa Magar, A., Mercuri, E.G.F. 2025. Assessment of an optimal parameter space for spatial cluster detection of SMEAR Estonia flux footprint data using unsupervised learning algorithms. – Forestry Studies | Metsanduslikud Uurimused 82, 20–27, ISSN 1736-8723. Journal homepage: http://mi.emu.ee/forestry.studies

Abstract. Understanding the spatial variability of ecosystem-atmosphere fluxes is essential for accurate carbon and water cycle assessments in forested landscapes. This study investigates the optimal parameter space for spatial cluster detection of flux footprint data from the SMEAR Estonia station using unsupervised learning algorithms. We applied DBSCAN and HDBSCAN clustering methods to half-hourly x-y coordinates of maximal flux contributions, derived from Kljun's footprint model, over a six-year period. The data were scaled using both standard and robust scalers to mitigate the effects of large coordinate values and outliers. We systematically evaluated clustering performance across a range of hyper-parameters, using silhouette and Davies-Bouldin scores to assess cluster quality. Our results indicate that HDBSCAN, particularly with robust scaling, provides more consistent and interpretable clusters, with lower sensitivity to noise and computational demands compared to DBSCAN. The findings highlight the importance of hyper-parameter selection and scaling in cluster analysis of flux footprint data and demonstrate the utility of density-based clustering for identifying spatial patterns in ecosystem flux measurements. These insights can inform future studies on carbon and water dynamics in heterogeneous forest environments and support the development of climate-smart forestry strategies.

Key words: DBSCAN, HDBSCAN, hyperparameters, unsupervised learning, flux footprint, forest ecosystem, atmosphere.

Authors' addresses: [1]Institute of Forestry and Engineering, Estonian University of Life Sciences, Kreutzwaldi 5, Tartu 51006, Estonia; [2]Department of Environmental Engineering, Federal University of Parana (UFPR), Centro Politécnico, Curitiba 81530-000, Brazil; *e-mail: steffen.noe@emu.ee

Introduction

Determining the exchange of matter and energy between terrestrial ecosystems and the atmosphere is a pivotal task in understanding the processes of carbon exchange under the changing climate. Large-scale measurement stations like SMEAR (Station for Measuring Ecosystem-Atmosphere Relations) in Estonia and Finland (Hari & Kulmala, 2005; Kulmala, 2018; Noe *et al.*, 2015), European research infrastructures like ICOS (Franz *et al.*, 2018) and globally operating networks like Fluxnet (Keenan *et al.*, 2021; Keenan *et al.*, 2019) provide vast numbers of eddy flux data that allow to verify carbon uptake and release from terrestrial ecosystems. Addressing the carbon sink capacity of forests is important to develop strategies for carbon sequestration, to mitigate climate change and develop climate-smart forestry solutions.

The concept of a flux footprint (Kljun *et al.*, 2015), the area that is relevant for assessing the ecosystems' exchange processes, is used in determining the source or sink capacity of a terrestrial ecosystem when applying the eddy covariance technique (Baldocchi, 2014; Rannik *et al.*, 2004). The footprint is subject to constant change by natural processes like forest growth, disturbances like windbreak and by human activities like land use change by, e.g. forest management (Kollo *et al.*, 2023). A footprint is not a static area but a dynamic region that is influenced by the wind direction and speed,

DOI: 10.2478/fsmu-2025-0002

atmospheric stability, and the surface roughness of the vegetation (Yu *et al.*, 2018). The footprint area can be estimated by using different models like the Lagrangian particle dispersion model (Hsieh *et al.*, 2000), the backward Lagrangian stochastic model (Kormann & Meixner 2001) or the parameterized footprint model by Kljun *et al.* (2015). Because it is applicable to a wider range of atmospheric stability and the relative low computational cost the latter is used in this work to determine the flux footprint characteristics of SMEAR Estonia.

The footprint's shape and the location of the strongest signal in the x-y plane that are detecable by a flux tower depend strongly on the wind patterns, weather phenomena, turbulent transport, tower properties, geography and land cover of the site (Rey-Sanchez *et al.*, 2022). Therefore, not every potential point within the flux footprint contributes at the same scale to a timely integrated carbon flux signal (Rey-Sanchez *et al.*, 2022). The source regions tend to form clusters and the contribution to the ecosystem-atmosphere exchange of the trees located within these cluster regions will be higher than that of the areas of less coverage by wind patterns.

Density-based spatial clustering of applications with noise (DBSCAN) (Ester *et al.*, 1996) and the hierarchical variation (HDBSCAN) (Campello *et al.*, 2013) are cluster recognition algorithms optimised for large spatial data with noise. They also allow the recognition of clusters via unsupervised machine learning and are applicable to the data without presetting, e.g. the number of clusters necessary for the k-means algorithm (Wu, 2012). Density-based cluster detection also allows outliers, i.e., points in the given spatial data that do not fulfil the connectivity criteria. Eddy covariance relies on the turbulent mixing of the atmosphere (Kowalski & Serrano-Ortiz, 2007) and we will therefore have stable regions and prevailing wind directions as well as rapidly changing wind speeds and directions, very likely causing spatially distributed outliers. We note here that in the context of our data an outlier is just a data point that does not belong to a specific cluster and it does not imply that this data point should be removed from the dataset.

Both clustering algorithms use two hyper-parameters: the maximum distance between two samples for one to be considered as in the neighborhood of the other ϵ and the minimum number of points in a region minPts to determine cluster points. In this study, we aimed a) to assess the best hyper-parameter pairs according to the given flux data, and b) to assess the feasibility of the method to find spatially distributed clusters of the maximal flux signal in the tower's footprint area.

Materials and Methods

Study area

The study area is located in Järvselja in the southeastern part of Estonia and data was obtained from SMEAR Estonia (Noe *et al.*, 2015). The terrain in the study area is flat and the flux tower is at a height of 36 m asl. We used eddy covariance data measured at 70 m above ground, covering an area of up to 4 km distance from the tower reaching up to 3,332 ha whereof about 88% or 2,900 ha are forest land. The other major land cover types are agricultural land (3.6%), clear cuts (2.5%), roads and access ways (2%), water bodies (1.3%), raised bog (1.1%), power lines (0.7%) and the rest are buildings and yards (Noe *et al.*, 2021; Kollo *et al.*, 2023). Flux footprint data represents the exchange of gases from the Järvselja experimental and training forest, which is a drained peatland forest and contains the dominant tree species found in the hemiboreal zone: Scots pine (*Pinus sylvestris* L.), Norway spruce (*Picea abies* (L.) H. Karst.), silver birch (*Betula pendula* Roth), downy birch (*Betula pubescens* Ehrh.), European aspen (*Populus tremula* L.), common alder (*Alnus glutinosa* (L.) Gaertn.) and grey alder (*Alnus incana* (L.) Moench) (Mercuri *et al.*, 2023).

Locating the point of maximal flux in x-y coordinates

The footprint function defines a 3D space where the location of the flux tower is at the origin of the projected 2D footprint area in the x-y plane. The flux footprint is defined as the mathematical transfer function between the sources and sinks, characterized as passive scalars in a surface area, and the turbulent flux that is measured at the tower (Schmid, 2002; Kljun *et al.*, 2015). Applying Kljun's footprint model (Kljun *et al.*, 2015) allows to either obtain an area depending on the calculated footprint function over a time interval or for each footprint function calculated at a discrete time of the vector of the footprint function's maximal point in x-y coordinates. To assess the clustering, we chose the latter option and created a dataset of x-y coordinates for each half hourly flux in our input data. Because the algorithm gives

the distance with respect to the wind direction, we construct the cartesian coordinates by:

$$(x, y) = (\sin(v_H \frac{\pi}{180}), \cos(v_H \frac{\pi}{180}))f_{\max}, \quad (1)$$

where f_{max} is the distance of the footprint's maximum in the horizontal wind direction v_H as given in the output of the footprint model.

Cluster algorithms

The choice of cluster algorithms to assess groups of data that follow "human intention" has several difficulties. The intuitive concept of a cluster is poorly defined and depends on the application (McInnes & Healy, 2017). Our choice for DBSCAN and HDBSCAN follows several implications our flux data entail. Knowing that the horizontal wind direction is the major factor to determine the location of the flux footprint's maximum and the previous research on the SMEAR Estonia flux footprint and horizontal wind direction densities (Figures 4 and 6 in Kollo *et al.* (2023)) we can assume that our data will have:

- clusters of arbitrary shape
- clusters with varying densities
- clusters of different sizes
- noise and outliers

All these features in our data will rule out algorithms like k-Means where we need to fix the number of clusters beforehand and the assumtion that clusters are spherical and equal in size and density.

The chosen algorithms, DBSCAN and HDBSCAN, are thoroughly described (McInnes & Healy, 2017; Campello *et al.*, 2013; Ester *et al.*, 1996) and have proven their feasibility in the detection of clusters, especially for spatial data. To this point, we would like to note that an intuitive criterion for a cluster is the density of data points. In dense regions, the probability of a cluster is high and in sparce regions the probability for a cluster is low. This can be used to motivate a statistical argument where the number of points within a certain range ϵ is linked to the cluster probability (McInnes & Healy, 2017). We may note further that the core distance is smaller or equal to ϵ and there is a probability density function PDF(x) ≥ λ and we have $\lambda = 1/\epsilon$. The λ criteron can be used to construct the hierarchy of the HDBSCAN algorithm.

Another metric that impacts on the quality of estimation of the PDF from data are the minimal number of points that determine what is graded a valid cluster. Values that are too low yield to a very "bumpy" PDF where the noise can be graded as mini-clusters. To avoid that, a threshold is set that is called minPts which determines the maximal size of a bump to be graded a peak in the PDF and by that a contributing cluster.

ϵ and minPts are the two hyper-parameters for the cluster determination we need to optimize for our data.

Data processing and cluster hyper-parameter assessment

To assess the best hyper-parameter set for our measurement station we used data from April 2015 until December 2020. Half-hourly x-y coordinates of the 70 m high eddy flux system's location of maximal contribution, accounting for a total of 68,266 coordinate points, were used to determine ϵ and minPts for monthly clusters. The x-y coordinates are given in metres from the measurement tower in cartesian coordinates using the Estonian plane coordinate system L-EST97 (Maa- ja Ruumiamet, 2019).

We created for each algorithm a Jupyter notebook (Kluyver *et al.*, 2016) to read the monthly data, scale the x-y coordinates according to Scikit-learn's (Pedregosa *et al.*, 2011) RobustScaler or StandardScaler function and apply either the DBSCAN or HDBSCAN functions with ϵ = [0, 1] in steps of 0.05 and minPts = [1, 50]. The cutoff at 50 minPts was chosen because the number of clusters was either one large cluster or no cluster was detected.

The scaling functions have the purpose of avoiding large numbers as given by the L-EST97 grid that may lead to numerical difficulties in the cluster estimation procedure. The scalers are defined by:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}, \quad (2)$$

where μ and σ denote the mean and the standard deviation of the coordinate ranges in the x-y plane for the standard scaler and by:

$$X_{\text{scaled}} = \frac{X - Q_1}{Q_3 - Q_1}, \quad (3)$$

with the first quartile Q_1 (25th percentile) and third quartile Q_3 (75th percentile) for the robust scaler. The robust scaler reduces the impact of possible outliers in the data.

We calculated the silhouette (Rousseeuw, 1987) and Davies-Bouldin (Davies & Bouldin, 1979) scores, both possible on unlabeled data.

The silhouette score is defined as:

$$S = \frac{b - a}{\max(a, b)}, \quad (4)$$

where a is the intra-cluster distance and b the nearest cluster distance and S is a metric on how different the data point is from points in other clusters. The Davies-Bouldin score is defined as:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max\left(\frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)}\right), \quad (5)$$

with a dataset $X = X_1, X_2, X_3,...$ and where $\Delta(X_k)$ is the intra-cluster distance and $\delta(X_i, X_j)$ the inter-cluster distance calculating the average cluster similarity of each cluster to the cluster most similar with it. The score shows the ratio between intra-cluster and inter-cluster distances.

We used these scores in selection of clustering parameters where lower values denote better estimates of the clusters. Further, we determined the number of clusters and outlier points that do not belong to clusters. The number of core points is also used as a selection criterion while the number of outliers may be used in quality controlling and assessing the cluster connectivity.

The software used for data processing and cluster determination was Python 3.10.9 (Van Rossum & Drake 2009), Pandas 1.5.3 (McKinney, 2010), Numpy 1.24.2 (Harris *et al.*, 2020), Matplotlib 3.6.3 (Hunter, 2007), and Scikit-learn 1.4.1.post1. (Pedregosa *et al.*, 2011).

Results and Discussion

Determining clusters per month on all data by using the two algorithms gave us an estimation for the possible ranges of the hyper-parameters related to the SMEAR Estonia stations footprint data. These findings are summarized in Table 1.

Table 1. Summary of the parameters obtained from applying clustering algorithms to the strongest signal location within the SMEAR Estonia footprint. The averages and standard deviations of discrete variables have been rounded to the nearest integer. The execution time is measured in seconds per iteration.

	DBSCAN		HDBSCAN	
Scaler	Standard	Robust	Standard	Robust
ϵ	0.48 ± 0.23	0.48 ± 0.23	0.26 ± 0.14	0.23 ± 0.14
minPts	25 ± 14	25 ± 14	31 ± 12	31 ± 12
Clusters (min. max)	3 ± 10 (0. 285)	2 ± 6 (0. 293)	3 ± 1 (0. 25)	3 ± 1 (0. 20)
Outliers	169 ± 256	107 ± 202	224 ± 126	209 ± 132
(min. max)	(0. 1274)	(0. 1195)	(3. 938)	(3. 963)
Silhouette score	0.31 ± 0.20	0.36 ± 0.19	0.27 ± 0.14	0.30 ± 0.15
Davies-Bouldin score	3.38 ± 6.99	3.35 ± 8.22	3.03 ± 5.11	2.8 ± 1.6
Execution time in s/it	8.3	7.3	6.5	5.6

In short, for DBSCAN ϵ is on average between 0.25 and 0.71, minPts between 11 and 39 regardless of which scaling was chosen. For HDBSCAN ϵ ranged on average between 0.12 and 0.4 for the standard scaling and between 0.09 and 0.37 for the robust scaling, and minPts varied between 19 and 43.

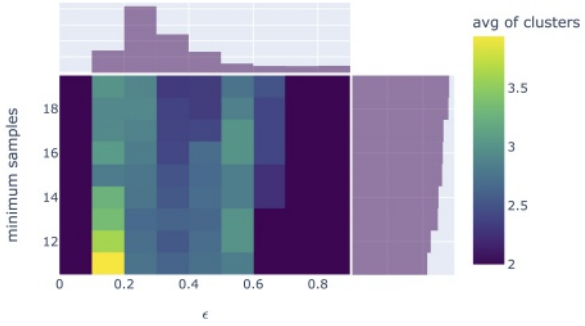
The average number of estimated clusters ranged from 2 to 4 for HDBSCAN regardless of which scaling was chosen. It is possible that no cluster is detected and the maximal numbers of detected clusters are 25 for the standard and 20 for the robust scaling. In DBSCAN, the variation of cluster estimation was much larger, maximal numbers reached up to 293 in the robust and 285 in the standard scaling cases. This result reflects the improved performance in cluster detection by the hierarchical feature of the cluster detection algorithm. Including the hierarchical component as given by HDBSCAN reduces the influence of the noise in cluster detection and the results get more robust. The about 10 times larger maximal number of detected clusters by DBSCAN shows that the noise in the data is likely to lead to too fuzzy PDFs of the density and therefore to a very high number of mini-clusters that fail to match visible comparison with the data. Therefore, HDBSCAN is the algorithm of

choice in cases of large noisy datasets.

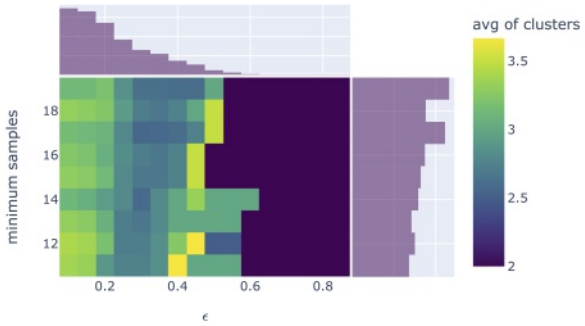
The silhouette scores are low for both algorithms and scalings indicate that the clusters detected by the two algorithms are located rather near each other over all the monthly data provided. HDBSCAN has obtained the lower silhouette score for both scaling scenarios and would be the algorithm of choice due to these criteria as well. Also, the Davies-Bouldin score was found lower for HDBSCAN and would suggest using this algorithm preferably on the SMEAR Estonia data. The algorithms' execution speed (numbers depend on hardware) was found consistently lower for HDBSCAN, also suggesting it as the best choice for the data given in the context of computational efficiency.

Figures 1a and 1b visualize the hyper-parameter space for ϵ and minPts for the robust scaling. Figure 1a suggests DBSCAN according to the marginal distributions for which the fitting hyper-parameters are $\epsilon = 0.3$ and minPts $\in [17, 20]$. We would obtain about three clusters. In the case of HDBSCAN (Figure 1b) the cluster detection is finer-granulated even though we used exactly the same steps during calculations. In line with the results in Table 1 we have the highest probability with small $\epsilon \approx 0.2$ and also minPts $\in [17, 20]$ and the number of clusters detected are between 3 and 4 mostly in that parameter range.

Interestingly we found a certain pattern in both algorithms' parameter spaces. For $\epsilon \approx 0.2$ and 0.5 the number of clusters is higher. If minPts is in the smaller range the largest number of clusters that can be found here is up to 4. In the range of $\epsilon \approx 0.3$ to 0.5 and minPts ≤ 15 an area with a lower number of clusters detected was found. This seemed to be more prominent for the DBSCAN algorithm.



(a) Parameter space for DBSCAN clustering. A smaller ϵ and minimum samples lead to more clusters on average.



(b) Parameter space for HDBSCAN clustering. A smaller ϵ and minimum samples lead also here to more clusters on average, but the highest number of clusters is found with lower minimum samples and a mid-range ϵ (0.4 to 0.5).

Figure 1. We filtered the calculated results to match the number of clusters ranging between 2 and 6, a silhouette score above 0.3, the minimum samples (minPts) in the range of 10 to 20, and the Davies-Bouldin score below 3.5. Marginal distributions of the hyper-parameters are shown atop and right.

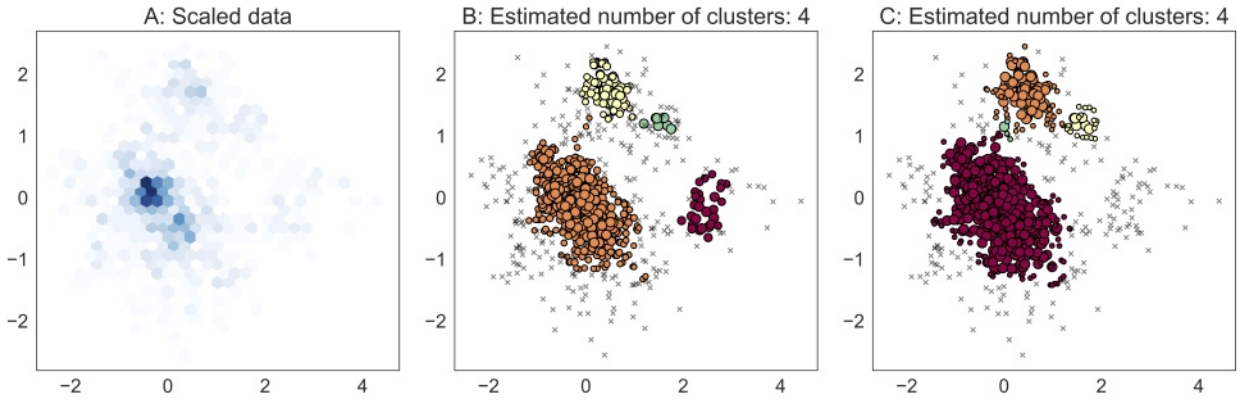


Figure 2. Example of clusters detected among data for July 2015 with parameters $\epsilon = 0.3$ and minPts = 15. Panel A shows the data scaled by the RobustScaler function and the color denotes the number of points per grid cell with a darker color for higher frequency. Panel B shows the estimated clusters using HDBSCAN, and C by DBSCAN. Note: The order in cluster labelling is different for the algorithms and therefore the coloring differs.

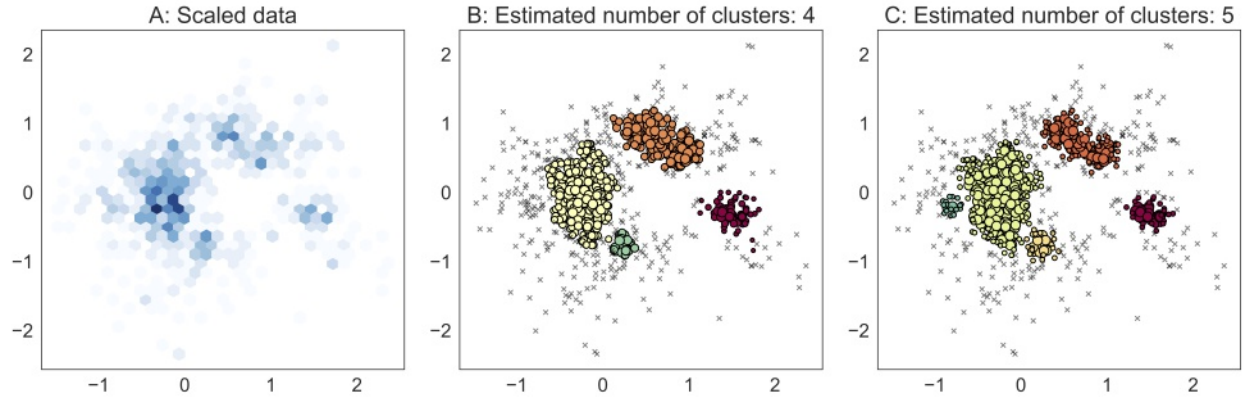


Figure 3. Example of clusters detected among data for July 2017 with parameters ϵ = 0.15 and minPts = 17. The panels follow the logic given in Figure 2.

Examples for cluster detections using similar hyper-parameters for both algorithms are shown in Figures 2 and 3. In Figure 2, when comparing the input data, we found that both algorithms – HDBSCAN and DBSCAN – capture the most dense clusters in an almost similar way but DBSCAN misses the less connected cluster centered at about [2.5, 0]. Instead, DBSCAN locates a very small cluster between the two most dense regions. HDBSCAN classifies this region with a lower probability as part of the largest cluster it detected. In the case of Figure 3, both algorithms were able to detect the clusters visible in the scaled data. In this case, HDBSCAN omitted one potential small cluster that was detected by DBSCAN.

Conclusion

Both algorithms resulted in preferable intervals for the hyper-parameters. On average, we found about three clusters for the maximal signal of ecosystem flux exchange on a monthly basis when we used all input data. Table 1 and Figures 1a and 1b confirm this finding. In that sense, both algorithms are capable of detecting clusters in the spatial data provided from the SMEAR Estonia station. The HDBSCAN algorithm led to a finer-grained detection (Figure 1), had a better retrieval of clusters which is seen by the lower Davies-Bouldin score, got the more consolidated number in maximal detected clusters and generally gave a smaller range of outliers.

We can conclude that the choice of scaling does not introduce very large differences in cluster detection for the most cases in our data. Usually, scaling prevents data from exhibiting large outliers in the features and eliminates distortions between the features (Chanal *et al.*, 2021). The major effect in our case is the re-scaling of the Estonian plane coordinate system's large numbers into a smaller domain ranging between [-3, 3] in both axes avoiding very large numbers in the calculation. Since the x and y coordinates show minimal scale distortion, the StandardScaler transformation – mean removal and scaling to unit variance – yields results that are very similar to those obtained with RobustScaler, which centers data by the median and scales by the interquartile range (Pedregosa *et al.*, 2011). The latter scaling algorithm is less affected by outliers and the choice of the robust scaling algorithm is safe for all types of data used in our work.

Calculating over a range of possible parameter values allows to decide which parameter pairs will be promising to detect clusters by unsupervised learning from the data. The choice of similar parameter pairs for both algorithms yielded very similar results (Figures 2 and 3) which means choosing either algorithm would give reasonable results. Depending on our results the HDBSCAN algorithm would be preferable in terms of finer granularity and lower computing times.

In summary, we would encourage the use of both techniques tested in this study to assess clusters in eddy covariance flux data. The safer choice is according to our results the use of HDBSCAN with the (RobustScaler). Since CO_2 and H_2O fluxes have strong dependence on soil and vegetation types, the determination of clusters of maximum flux contribution can help to understand the carbon and water dynamics of the forest ecosystem-atmosphere. The effects of different species and forest management practices within the flux tower footprint can be linked to spatially explicit areas, enabling comparisons and experiments to determine which measures improve carbon uptake and support climate mitigation strategies.

Acknowledgments. This study was funded by the Estonian Research Council (grant PRG 1674) and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. The authors are grateful to Erasmus+ Estonian University of Life Sciences (EMÜ) staff mobility program and the Network on Environmental Monitoring and Modeling (RESMA) project from the Federal University of Parana (UFPR) – Coordination for the Improvement of Higher Education Personnel (CAPES) – Institutional Internationalization Program (PRINT) for facilitating the exchange of researchers between Brazil and Estonia.

References

Baldocchi, D. 2014. Measuring fluxes of trace gases and energy between ecosystems and the atmosphere – the state and future of the eddy covariance method. – Global Change Biology, 20(12), 3600–3609. https://doi.org/https://doi.org/10.1111/gcb.12649.

Campello, R.J.G.B., Moulavi, D., Sander, J. 2013. Density-based clustering based on hierarchical density estimates. – Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds.). Advances in Knowledge Discovery and Data Mining. Berlin, Heidelberg, Springer, 160–172.

Chanal, D., Steiner, N.Y., Chamagne, D., Pera, M.-C. 2021. Impact of standardization applied to the diagnosis of LT-PEMFC by Fuzzy C-Means clustering. – Proceedings of the IEEE Vehicle Power and Propulsion Conference (VPPC), Spain, 25-28 October 2021. Gijon, 6 pp. https://doi.org/10.1109/VPPC53923.2021.9699234.

Davies, D.L., Bouldin, D.W. 1979. A cluster separation measure. – IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1 (2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. 1996. 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. Portland, Oregon, AAAI Press. 226–231.

Franz, D., Acosta, M., Altimir, N., Arriga, N., Arrouays, D., Aubinet, M., Aurela, M., Ayres, E., López-Ballesteros, A., Barbaste, M., Berveiller, D., Biraud, S., Boukir, H., Brown, T., Brümmer, C., Buchmann, N., Burba, G., Carrara, A., Cescatti, A., Ceschia, E., Clement, R., Cremonese, E., Crill, P., Darenova, E., Dengel, S., D'Odorico, P., Filippa, G., Fleck, S., Fratini, G., Fuß, R., Gielen, B., Gogo, S., Grace, J., Graf, A., Grelle, A., Gross, P., Grünwald, T., Haapanala, S., Hehn, M., Heinesch, B., Heiskanen, J., Herbst, M., Herschlein, C., Hörtnagl, L., Hufkens, K., Ibrom, A., Jolivet, C., Joly, L., Jones, M., Kiese, R., Klemedtsson, L., Kljun, N., Klumpp, K., Kolari, P., Kolle, O., Kowalski, A., Kutsch, W., Laurila, T., de Ligne, A., Linder, S., Lindroth, A., Lohila, A., Longdoz, B., Mammarella, I., Manise, T., Jiménez, S.M., Matteucci, G., Mauder, M., Meier, P., Merbold, L., Mereu, S., Metzger, S., Migliavacca, M., Mölder, M., Montagnani, L., Moureaux, C., Nelson, D., Nemitz, E., Nicolini, G., Nilsson, M.B., de Beeck, M.O., Osborne, B., Löfvenius, M.O., Pavelka, M., Peichl, M., Peltola, O., Pihlatie, M., Pitacco, A., Pokorný, R., Pumpanen, J., Ratié, C., Rebmann, C., Roland, M., Sabbatini, S., Saby, N.P.A., Saunders, M., Schmid, H.P., Schrumpf, M., Sedlák, P., Ortiz, P.S., Siebicke, L., Šigut, L., Silvennoinen, H., Simioni, G., Skiba, U., Sonnentag, O., Soudani, K., Soulé, P., Steinbrecher, R., Tallec, T., Thimonier, A., Tuittila, E.-S., Tuovinen, J.-P., Vestin, P., Vincent, G., Vincke, C., Vitale, D., Waldner, P., Weslien, P., Wingate, L., Wohlfahrt, G., Zahniser, M., Vesala, T. 2018. Towards long-term standardised carbon and greenhouse gas observations for monitoring Europe's terrestrial ecosystems: a review. – International Agrophysics, 32(4), 439–455. https://doi.org/10.1515/intag-2017-0039.

Hari, P., Kulmala, M. 2005. Station for Measuring Ecosystem-Atmosphere Relations (SMEAR II). – Boreal Environment Research, 10(5), 315–322.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E. 2020. Array programming with NumPy. – Nature, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Hsieh, C.-I., Katul, G., Chi, T. 2000. An approximate analytical model for footprint estimation of scalar fluxes in thermally stratified atmospheric flows. – Advances in Water Resources, 23(7), 765–772. https://doi.org/10.1016/S0309-1708(99)00042-1.

Hunter, J.D. 2007. Matplotlib: A 2D graphics environment. – Computing in Science & Engineering, 9(3), 90–95.

Keenan, T.F., Moore, D.J.P., Desai, A. 2019. Growth and opportunities in networked synthesis through AmeriFlux. – New Phytologist, 222(4), 1685–1687. https://doi.org/https://doi.org/10.1111/nph.15835.

Keenan, T., Moore, D., Kimberly N. 2021. The FLUXNET Coordination Project: A new initiative to support global network-enabled science. – Abstract of the AGU Fall Meeting, USA, 13-17 December 2021. New Orleans, B14D–01.

Kljun, N., Calanca, P., Rotach, M.W., Schmid, H.P. 2015. A simple two-dimensional parameterisation for Flux Footprint Prediction (FFP). – Geoscientific Model Development, 8(11), 3695–3713. https://doi.org/10.5194/gmd-8-3695-2015.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C. 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. – Loizides, F., Schmidt, B. (eds.). Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing. Amsterdam, Berlin, Washington, DC, IOS Press, 87 – 90.

Kollo, J., Padari, A., Krasnova, A., Kangur, A., Noe, S.M. 2023. Development of a footprint description tool utilizing SMEAR Estonia eddy-covariance data and footprint modelling in combination with remote sensed forest species and land cover data. – Forestry Studies / Metsanduslikud Uurimused, 79, 90–104. https://doi.org/doi:10.2478/fsmu-2023-0014.

Kormann, R., Meixner, F.X. 2001. An analytical footprint model for non-neutral stratification. – Boundary-Layer Meteorology, 99, 207–224.

Kowalski, A.S., Serrano-Ortiz, P. 2007. On the relationship between the eddy covariance, the turbulent flux, and surface exchange for a trace gas such as CO_2. – Boundary-Layer Meteorology, 124, 129–141. https://doi.org/10.1007/s10546-007-9171-z.

Kulmala, M. 2018. Build a global Earth observatory. – Nature, 553, 21–23. https://doi.org/10.1038/d41586-017-08967-y.

Maa- ja Ruumiamet. 2019. Estonian Land and Spatial Development Board. Geodetic System. [WWW document]. – URL https://geoportaal.maaamet.ee/geoportaal.maaamet.ee/eng/spatial-data/geodetic-data/geodetic-system-p668.html. [Accessed 21 May 2024].

McInnes, L., Healy, J. 2017. Accelerated hierarchical density based clustering. – Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW), USA, 18-21 November 2017. New Orleans, 33–42. https://doi.org/10.1109/ICDMW.2017.12.

McKinney, W. 2010. Data structures for statistical computing in python. – Proceedings of the 9th Python in Science Conference, USA, 28 June-3 July 2010. Austin, Texas, 56–61.

Mercuri, E.G.F., Tamm, T., Noe, S.M. 2023. Water and carbon balances in a hemi-boreal forest. – Forestry Studies / Metsanduslikud Uurimused, 78, 72–90.

Noe, S.M., Krasnova, A., Krasnov, D., Cordey, H.P.E., Kangur, A. 2021. Facilitating long-term 3D sonic anemometer measurements in hemiboreal forest ecosystems. – Forestry Studies / Metsanduslikud Uurimused, 75, 140–49. https://doi.org/10.2478/fsmu-2021-0016.

Noe, S.M., Niinemets, Ü., Krasnova, A., Krasnov, D., Motallebi, A., Kängsepp, V., Jõgiste, K., Hõrrak, U., Komsaare, K., Mirme, S., Vana, M., Tammet, H., Bäck, J., Vesala, T., Kulmala, M., Petäjä, T., Kangur, A. 2015. SMEAR Estonia: Perspectives of a large-scale forest ecosystem–atmosphere research infrastructure. – Forestry Studies / Metsanduslikud Uurimused, 63, 56–84. https://doi.org/10.1515/fsmu-2015-0009.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É. 2011. Scikit-learn: Machine learning in Python. – The Journal of Machine Learning Research, 12, 2825–2830.

Rannik, Ü., Keronen, P., Hari, P., Vesala, T. 2004. Estimation of forest–atmosphere CO_2 exchange by eddy covariance and profile techniques. – Agricultural and Forest Meteorology, 126(1–2), 141–155. https://doi.org/10.1016/j.agrformet.2004.06.010.

Rey-Sanchez, C., Arias-Ortiz, A., Kasak, K., Chu, H., Szutu, D., Verfaillie, J., Baldocchi, D. 2022. Detecting hot spots of methane flux using footprint-weighted flux maps. – Journal of Geophysical Research: Biogeosciences, 127(8), e2022JG006977. https://doi.org/https://doi.org/10.1029/2022JG006977.

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

Schmid, H.P. 2002. Footprint modeling for vegetation atmosphere exchange studies: a review and perspective. – Agricultural and Forest Meteorology, 113(1–4), 159–183. https://doi.org/https://doi.org/10.1016/S0168-1923(02)00107-7.

Van Rossum, G., Drake, F.L. 2009. Python 3 Reference Manual. Scotts Valley, CA, USA, CreateSpace. 242 pp.

Wu, J. 2012. Advances in K-means Clustering: A Data Mining Thinking. Berlin, Heidelberg, Springer Verlag. 180 pp.

Yu, M., Wu, B., Zeng, H., Xing, Q., Zhu, W. 2018. The impacts of vegetation and meteorological factors on aerodynamic roughness length at different time scales. – Atmosphere, 9(4), 149. https://doi.org/10.3390/atmos9040149.

Submitted May 4, 2025, accepted January 18, 2025

The journal is financially supported by the Estonian Environmental Investment Centre and the Institute of Forestry and Engineering of the Estonian University of Life Sciences