

Faculté des bioingénieurs

Assessing Wasp Biodiversity Potential of German Ecosystems

**Establishing a baseline for Ecosystem Health
Evaluation**

Author	Marieke Scheel
Supervisors	Prof. Pierre Defourny Dr. Julien Radoux
Jurors	Prof. Yannick Agnan Dr. Sophie Bontemps
Academic year	2023-2024
Programme	MSc Erasmus+ GEM - Geo-information Science and Earth Observation for Environmental Modelling and Management

Acknowledgements

I would like to express my deepest appreciation to my supervisors Prof. Pierre Defourny and Dr. Julien Radoux for their invaluable guidance at every stage of this master thesis. Your insightful feedback has significantly shaped this thesis.

Additionally, I would like to extend my gratitude to Prof. Yannick Agnan and Dr. Sophie Bontemps, for investing their time and expertise in evaluating my thesis. I look forward to the opportunity to discuss it further during the defence.

A sincere thank you to the GEM coordinators at UCLouvain, the University of Tartu and the University of Twente for their constant support. I am also extremely grateful for the generous support from the DAAD, who financed my studies.

Special thanks to my classmates and friends at the University of Tartu and UCLouvain, our late-night sessions and your moral support kept me going. I could not have reached this goal without the unwavering support by family, friends and loved ones, thank you.

Table of Contents

1. Introduction	1
2. Literature Review	3
2.1 Ecosystem definitions	3
2.2 Delineation of Subunits in the Landscape	3
2.3 Ecosystem Health	4
2.3.1 Resilience	5
2.3.2 Vigour	6
2.3.3 Organisational Structure	7
2.4 Ecosystem Health Metrics	7
2.4.1 Indicators	7
2.4.2 Key Species	8
2.4.3 Biodiversity Carrying Capacity	9
2.5 Clustering Algorithms	10
2.5.1 Partitioning Clustering Algorithms	10
2.5.2 Hierarchical Clustering Algorithms	11
2.5.3 Density-based Clustering Algorithms	12
2.5.4 Model-based Clustering Algorithms	12
2.6 Summary	13
3. Research Objectives	14
4. Study Area and Data	15
4.1 Study Area	15
4.2 Data	18
4.2.1 LifeWatch Dataset	19
4.2.2 EUNIS Habitat Dataset	20
4.2.3 GBIF Taxon Dataset	22
4.2.4 Natura 2000 Dataset	23
5. Methods	24
5.1 Overview Workflow	25
5.2 Unsupervised Clustering	27
5.2.1 Applying Quality Control to LifeWatch Data	27
5.2.2 K-means Clustering Algorithm	28
5.2.3 Genie Clustering Algorithm	29
5.2.4 GMM Clustering Algorithm	29
5.3 Cluster Validation	30
5.3.1 Intrinsic Metrics Assessing Cluster Quality	30
5.3.2 Extrinsic Metrics Assessing Clusters Quality	32

5.3.2.1 Applying Quality control to the EUNIS Habitat Data	32
5.3.2.2 Assess clusters with most accurate habitat representation	32
5.4 Establishing Baseline for Ecosystem Health	33
5.4.1 Applying Quality Control to GBIF Data	33
5.4.2 Validation of the Wasp Biodiversity Potential Assessment	34
5.4.3 Application of Wasp Biodiversity Potential for Ecosystem Health Assessment.....	34
6. Results	35
6.1 Clustering Results.....	35
6.1.1 K-means Clusters	35
6.1.2 Genie Clusters	37
6.1.3 GMM Clusters	39
6.1.3.1 Summary statistics of the GMM training data	41
6.1.4 Comparison of Cluster Algorithms	41
6.2 Clustering Validation.....	43
6.2.1 Intrinsic Metrics	43
6.2.2 Extrinsic Metric	44
6.2.2.1 Quality-controlled EUNIS Habitat Data	44
6.2.2.2 Assessment of Cluster Purity with EUNIS Habitat Data	45
6.2.2.3 Cluster Labelling with EUNIS	46
6.3 Calculation of Baseline for Ecosystem Health Assessment	49
6.3.1 Quality-Controlled GBIF Taxon Data	49
6.3.2 Wasp Biodiversity Potential	50
6.3.2.1 Establishing Wasp Biodiversity Potential	50
6.3.2.2 Validation and quality control of the Wasp Biodiversity Potential	52
6.3.3 Application of Wasp Biodiversity Potential as a Baseline	55
7. Discussion	60
7.1 Clustering Algorithms and Intrinsic Validation	60
7.2 External Clustering Validation	63
7.3 Wasp Biodiversity Potential	64
7.4 Recommendations	67
8. Conclusion.....	69
9. References	71
10. Appendix	80
10.1 Figures	80
10.2 Tables.....	87

Table of Figures

Fig. 1. Conceptualisation of ecosystem health by Tett et al. (2013). Ecosystem health is here defined as “the ability to maintain systems integrity” (Tett et al., 2013).....	5
Fig. 2. Overview of the main types of clustering algorithms, divided into four groups. Source: Gao et al. (2023).....	10
Fig. 3. K-means clustering schematic. Source: Gao et al. (2023)	11
Fig. 4. Description of Hierarchical Clustering methods and popular distance metrics. Source: Gao et al. (2023).....	12
Fig. 5. CORINE land cover (CLC) in the study area of Germany and location of 15 major cities. CLC dataset stems from the European Environment Agency (EEA).	16
Fig. 6. Major biogeographical regions in Germany shown in colours, and more detailed region names are highlighted with labels. The data stems from the German Federal Agency for Nature Conservation (Bundesamt für Naturschutz) and was manually translated from German to English by the author of this study.....	17
Fig. 7. Major rivers and mountain ranges in Germany. Source: Behr et al. (2021).....	18
Fig. 8. Ecotopes of LifeWatch data in Germany shown as individual red polygons outlined in white. Source: LifeWatch ERIC (2017).....	20
Fig. 9. Spatial distribution of the raw EUNIS data in Germany.	21
Fig. 10. Spatial distribution of the raw GBIF data in Germany.	23
Fig. 11. Natura 2000 habitats in Germany. Source: EEA (2022).....	24
Fig. 12. Overview workflow describing the major steps of the analysis conducted for this study.	26
Fig. 13. Description of the method to detect outliers using the interquartile range. Source: Vinutha et al. (2018).....	34
Fig. 14. The IDs of the clusters created using the K-means algorithm are plotted spatially. ..	36
Fig. 15. Cluster area frequency of the clusters created using the K-means algorithm. The bin size is 50. The blue line corresponds to the Kernel Density Estimate curve.	37
Fig. 16. The IDs of the clusters created using the Genie algorithm are plotted spatially.	38
Fig. 17. Cluster area frequency of the clusters created using the Genie algorithm. The bin size is 50. The red line corresponds to the Kernel Density Estimate curve.	39
Fig. 18. The IDs of the clusters created using the GMM algorithm are plotted spatially.	40
Fig. 19. Cluster area frequency of the clusters created using the GMM algorithm. The bin size is 50. The purple line corresponds to the Kernel Density Estimate curve.	41

Fig. 20. Distribution of the area of the clusters created with the K-means, Genie and GMM algorithms.....	42
Fig. 21. Comparison of the distribution of the intra-cluster distances of the clusters created with the K-means, Genie and GMM algorithms.	43
Fig. 22. Silhouette score, Davies-Bouldin Index and Calinski-Harabasz Index plotted for the clusters created using the K-means, Genie and GMM algorithms.....	43
Fig. 23. Spatial distribution of the original EUNIS data points (left), and spatial distribution of the EUNIS data points after quality control (right).....	44
Fig. 24. Frequency of the original (red) and quality-controlled (blue) habitat classes of the EUNIS data in Germany. The EUNIS habitat classes were aggregated to level 1.	45
Fig. 25. Distribution of the proportion of the majority EUNIS class within the clusters created using K-means, Genie and GMM. The number in white in the boxplots corresponds to the median of the distributions.....	46
Fig. 26. Frequency of the proportion of the majority of EUNIS classes within the clusters created using the K-means (blue), Genie (red) and GMM (purple) algorithms. The bin size is 0.075. The lines correspond to the Kernel Density Estimate curve.	46
Fig. 27. Labelling of the Genie clusters using EUNIS habitat samples a) by the majority of the EUNIS samples within clusters and b) by the majority of the weighted EUNIS samples within clusters. To the right, certain regions are highlighted for comparison.....	48
Fig. 28. Percentages of the different wasp species in the original GBIF dataset (green), the GBIF dataset with removed outliers (red) and the fully quality-controlled GBIF dataset (blue). Species names in light grey were removed by the quality control.	49
Fig. 29. Spatial distribution of the original GBIF data points (left) and spatial distribution of the GBIF data points after quality control (right).	50
Fig. 30. Frequency of unique wasp species counts within the selected clusters. The bin size is 0.7. The blue line corresponds to the Kernel Density Estimate curve.	51
Fig. 31. Number of unique wasp species plotted for the 42 selected clusters, each with > 100 GBIF samples.	52
Fig. 32. Total number of samples (red) and number of unique GBIF wasp species (blue) per cluster.	53
Fig. 33. Scatterplot for number of unique wasp species and the total area covered by clusters (in km ²) with >100 GBIF samples (top) and scatterplot for number of unique wasp species and number of GBIF samples per cluster with >100 GBIF samples (bottom). The black line corresponds to a linear regression line and the dotted to a LOESS (locally estimated scatterplot	

smoothing) line. They are given for guidance. The Spearman tests indicate significant correlation ($p < 0.05$).....	53
Fig. 34. Comparison of the percentage of the area by unique wasp species counts for the whole area of the Genie clusters with >100 GBIF samples (blue), for the Genie clusters within Natura 2000 areas (red) and for the Genie clusters not within Natura 2000 areas (purple).	54
Fig. 35. Spearman correlation coefficients between the biophysical variables used for clustering and the unique wasp species count. Variables with correlation < 0.2 and > -0.2 are masked out in white.	55
Fig. 36. Calculation of wasp biodiversity potential in Hamburg, Cologne, Frankfurt, Berlin and Munich, the five largest cities in Germany (by population). Natura 2000 areas within the city perimeters are highlighted in yellow, and grey areas indicate no data (≤ 100 GBIF samples).	56
Fig. 37. Area in % of unique wasp species count intersecting with Natura 200 areas within Munich (blue), Berlin (purple), Hamburg (green), Frankfurt am Main (red) and Cologne (orange).	57
Fig. 38. Ratio between unique species per ecotope and wasp biodiversity potential per cluster in Hamburg, Cologne, Frankfurt, Berlin and Munich, the five most populated cities in Germany. Natura 2000 areas within the city perimeters are highlighted in yellow, and grey areas indicate no data (≤ 100 GBIF samples).	58
Fig. 39. Ratio between unique species per ecotope and wasp biodiversity potential of cluster 181 (Genie). A demonstration of how the unique number of species per cluster can be applied.	59
Fig. A1. GBIF data points buffered by their respective coordinate uncertainties before the outlier removal (left) and GBIF data points after outlier removal (right).	80
Fig. A2. Areas of the 200 K-means clusters. The dotted line shows the median cluster area.	80
Fig. A3. Areas of the 200 Genie clusters. The dotted line shows the median cluster area.	81
Fig. A4. Areas of the 200 GMM clusters. The dotted line shows the median cluster area.	81
Fig. A5. Number of samples of the different EUNIS habitat classes before quality control.	82
Fig. A6. Number of samples of the different EUNIS habitat classes after quality control.	82
Fig. A7. The number of EUNIS samples within the different clusters before and after quality control.	82
Fig. A8. The cluster IDs of all Genie clusters with >100 GBIF samples.	83

Fig. A9. Comparison of the area in km ² by unique wasp species count for the whole area of the Genie clusters with >100 GBIF samples (blue), for the Genie clusters within Natura 2000 areas (red) and for the Genie clusters not within Natura 2000 areas (purple).....	84
Fig. A10. Number of unique wasp species plotted for the 158 remaining clusters, each with ≤ 100 GBIF samples.	85
Fig. A11. The cluster IDs of all Genie clusters with ≤ 100 GBIF samples.....	86

Table of Tables

Table 1. Summary of key species and measurement indicators by ecosystem type by Lu et al. (2015).	7
Table 2. Biophysical variables	19
Table 3. The 38 biophysical variables of the LifeWatch dataset removed due to a VIF factor > 10.	28
Table 4. Summary statistics of the quality-controlled LifeWatch dataset (original) and of the quality-controlled LifeWatch dataset used for training of the GMM algorithm (training).	41
Table A1. Factors that determine the quality for an indicator group according to Brown (1991). Source: Brown (1991)	87
Table A2. LifeWatch biophysical variable names and their explanation. Source: LifeWatch database, non-validated, pre-released version, validation will be provided in September 2024	87
Table A3. LifeWatch variable groups and their data sources. Source: LifeWatch database, non-validated, pre-released version, validation will be provided in September 2024.	89
Table A4. Biophysical variables from the LifeWatch dataset that were removed due to a VIF > 10 with the respective VIF values.	90
Table A5. Biophysical variables of the LifeWatch dataset remaining after removing 38 variables using VIF.	91
Table A6. EUNIS habitat names and their corresponding codes.	91

List of acronyms

BIRCH	Balance Iterative Reducing and Clustering using Hierarchies
CHI	Calinski-Harabasz Index
CLC	CORINE Land Cover
DBI	Davies-Bouldin Index
DBSCAN	Density-based algorithm for discovering clusters in large spatial databases with noise
EEA	European Environment Agency
EM	Expectation Maximisation algorithm
EUNIS	European Nature Information System
GBIF	Global Biodiversity Information Faculty
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
HPC	High Performance Computing
IQR	Interquartile range
KDE	Kernel density estimate
KS-test	Kolmogorov-Smirnov test
LifeWatch ERIC	LifeWatch European Research Infrastructure Consortium
LOESS	Locally estimated scatterplot smoothing
NDVI	Normalised Difference Vegetation Index
PCA	Principal Component Analysis
PNV	Potential Natural Vegetation
RAM	Random-Access Memory
SLURM	Simple Linux Utility for Resource Management

VIF	Variance Inflation Factor
WWF	World Wide Fund For Nature

1. Introduction

For the last 50 years, we have witnessed a biodiversity loss at a speed that is unprecedented (Lindqvist et al., 2016). On average, 52 species are moving closer to extinction each year, and this decline in biodiversity is projected to continue throughout the 21st century (Hoffmann et al., 2010). Already in 2010, about 20% of vertebrates, 17% of mammals and 11% of birds globally were threatened by extinction. Most of the declines in biodiversity would be reversible; however, in 13% of species, it has led to extinction. While the trend in biodiversity decline persists, it is also essential to acknowledge that the loss rate would be at least 20% higher if it were not for conservation efforts (Hoffmann et al., 2010). Nonetheless, these efforts remain inadequate as they have only slowed down the current loss of biodiversity (Watson et al., 2014).

Biodiversity is one of the most important factors affecting the resilience of an ecosystem. This is of particular relevance in light of the expected increase of perturbations in ecosystems due to climate change (Oliver et al., 2015). A lack of resilience can easily lead to a collapse of ecosystems if they are below a certain minimum threshold for too long (Oliver et al., 2015). While more resilient ecosystems can withstand being below a minimum threshold for a specific period, less resilient ecosystems are more sensitive. Hence, as for instance, droughts are becoming more frequent in Europe (Archambeau et al., 2020), less resilient ecosystems become more likely to collapse. Besides these ecosystems being of great importance for flora and fauna, they also have a tremendous socioeconomic relevance (Rapport et al., 1998). For instance, stresses on aquatic ecosystems can lead to declines in the availability of commercial fish stocks, leading to significant economic losses. This emphasises that the conservation of biodiversity should be of paramount importance for humankind. As conservation efforts should be tailored towards need, there is an increasing need for an accurate assessment of ecosystem health.

The term health regarding ecosystems has emerged in the last decades as the state of ecosystems has declined notably (Rapport et al., 1998). The health of ecosystems has also been described as quality, status, integrity, sustainability or protection, but health will be used as the term in this study (O'Brien et al., 2016). A healthy ecosystem has been characterised as balanced and sustainable, i.e. the system is resilient to perturbations and maintains its vigour as well as organisational structures (Rapport et al., 1998).

To assess the health of ecosystems, one can, for instance, use varying biophysical variables, including the Normalised Difference Vegetation Index (NDVI), which can be used as both an indicator for the vigour and amount of biomass (Siddig et al., 2016). The latter are parts of what constitutes ecosystem health (as described above). Secondly, community metrics like diversity, abundance or composition can be used to estimate the status of an ecosystem since those metrics can then be compared to a hypothetical maximum for the biodiversity potential (biodiversity carrying capacity) in a given region (Siddig et al., 2016). Additionally, indicator species are widely used to assess ecosystem health (Bakker, 2008) and can also be compared to a hypothetical maximum. It can be distinguished between composition indicators and condition indicators (of ecosystem, habitat or community), with the former being a comparison between different systems and the latter a comparison within a system (Zacharias & Roff, 2001).

Indicator species are organisms that mirror or predict the conditions of the ecosystem they live in, which can be monitored easily (Siddig et al., 2016). This is based on the hypothesis that changes in the immediate environment are reflected by a population's demographic parameters (like abundance, dispersal, density, growth and reproduction, etc.) (Siddig et al., 2016).

While studies to assess ecosystem health for a particular region or a specific type of ecosystem are relatively common (Birkhofer et al., 2018; Hasani et al., 2023; O'Brien et al., 2016; Peleg et al., 2023), health estimations applied to a diverse range of ecosystems or a more extensive region are found more rarely (Mengist et al., 2020). On the one hand, more focused studies might provide more detailed and accurate estimates. However, such methods can be very time and work-intensive and might only be valid and applicable locally. In comparison, ecosystem health estimations for various ecosystems and over a larger region could prove to be more cost- and time-effective. This could also aid in gaining more insight into the state of less studied ecosystems or ecosystems that are of greater interest. Endangered or particularly rapidly changing ecosystems could be of interest to assess, as well as those that provide vital ecosystem services (Mengist et al., 2020). Thus, a method that can assess the health of different types of ecosystems over a larger area would be particularly desirable to bridge a gap in a currently understudied field and assess the state of ecosystems more easily to implement measures to protect those effectively.

2. Literature Review

In the following section, methodologies, terms, and concepts that are relevant to this study will be introduced. The definition and estimation of ecosystem health as well as the delineation of subunits in the landscape will be explored in detail.

2.1 Ecosystem definitions

The term ecosystem was originally defined by Tansley (1935). He introduced the ecosystem concept as the integration of the biotic community with its physical environment within a hierarchical spatial system from atom- to universe-sized (Schowalter, 2016). This was then followed by Lindeman (1942) who introduced the modern definition of ecosystems by showcasing the inseparability of biotic communities and their environments.

For administrative, research and management purposes, it is of great importance to be able to delineate distinct spatial boundaries of ecosystems. Unfortunately, there is a wide range of different and often overlapping definitions for subdivisions in the landscape, based on biotic and/or abiotic factors as well as scale (Baptist et al., 2019). Klijn and de Haes (1994) proposed a hierarchical approach to classify and map ecosystems at different spatial scales. It differentiated (from large to small) between ecozone, ecoprovince, ecoregion, ecodistrict, ecosection, ecoseries, ecotope and eco-element. While ecozones were defined to have a minimum mapping unit of $> 62.500 \text{ km}^2$, eco-elements were defined to be $< 0.25 \text{ ha}$.

Meanwhile, together with the ecosystem concept, Tansley (1935) also introduced the term ecotope and defined it to be a small part of an ecosystem. Nowadays, the term ecotope is often confused with similar concepts, e.g. habitat. While the term habitat initially solely referred to the physical environment of a given species, its modern definition refers to the physical environment of a given community (Baptist et al., 2019). The modern definition of habitat is further solidified by its use in widely used datasets and maps e.g. by the European Habitats Directive and the European Nature Information System (EUNIS) (Baptist et al., 2019).

For the purpose of this study, an ecotope will be defined as the physical environment of communities, making its use interchangeable with the modern definition of the term habitat. Additionally, an ecotope will be defined as a subunit of an ecosystem of a given size.

2.2 Delineation of Subunits in the Landscape

For research, administrative and management purposes, subdividing a landscape into smaller areas with homogeneous characteristics is often essential. Such subdivision can be understood as an attempt to simplify and categorise a complex landscape. Clustering of landscape, i.e., the aggregation of subunits within the landscape that are more similar to each other, has often been attempted over time (e.g. Krohn et al. (1999)). Some early efforts to subpartition the landscape

focused on land use, including anthropogenic activities, for instance by Marschner (1950), as he assumed the biophysical environment characteristics to be the determinant of human activities. In contrast, Küchler (1964) distinguished ecoregions by identifying the Potential Natural Vegetation (PNV), which he hypothesised also reflects biophysical landscape components like topography, soil and climate. PNV is a theoretical concept that can be defined as “the expected state of mature vegetation in the absence of human intervention” (Chiarucci et al., 2010). A nowadays widely used ecoregion dataset was compiled by Olson et al. (2001) in association with the World Wide Fund For Nature (WWF). It was developed by combining and modifying existing delineations of biogeographic regions and biomes with the aim of the formed ecoregions representing distinct biotas (Olson et al., 2001).

An important aspect to be considered when delineating subdivisions in the landscape is the concept of different scales and the definition of the systems at those scales (Omernik, 2004). As introduced earlier, ecoregions are ecosystems at a larger spatial scale that group ecological and geographical characteristics. Hence, ecoregions can also be understood to be an aggregation of ecosystems (Omernik & Bailey, 1997) intended to “illustrate patterns in the capacities and potentials of ecological systems” (Omernik, 2004). However, they do not attempt to reflect any aspects of the ecological systems within the ecoregion perfectly (Omernik, 2004). Alleviating part of the ambiguity surrounding the scale when delineating ecosystems, it was found that as long as ecosystems of relative homogeneity are grouped, the findings for delineating ecoregions can be applied to any scale, from regional to local (Bryce et al., 1999; Omernik & Bailey, 1997).

However, there is no agreed-on framework on how to delineate ecoregions (Omernik, 2004). Among others, this ambiguity can be attributed to a general disagreement on how to define ecosystems and the general complexity of ecoregions, as well as the concept of boundaries in natural systems (Omernik, 2004). For instance, a core aspect of ecoregions can be considered to be biodiversity, so that ecoregions could be delineated based on their capacity to fulfil the habitat requirements of species (Loveland & Merchant, 2004). Nonetheless, ecoregions are most commonly defined using climatic and topographical parameters, e.g. slope, elevation, temperature, precipitation (Krohn et al., 1999).

As biotic and abiotic factors are correlated in local and regional landscapes, it was found that when delineating ecoregions, the abiotic factors are reliable in reflecting the biotic conditions (Krohn et al., 1999). Hence, it was suggested that when delineating ecoregions, a stronger emphasis should be placed on physical parameters that are more permanent (e.g. elevation, slope) compared to more transitory parameters like biological communities (Krohn et al., 1999).

2.3 Ecosystem Health

The term and definition of ecosystem health has been widely discussed since the 1980s (O’Brien et al., 2016). Among others, the following definitions of ecosystem health have been discussed in the past:

- Absence of disease, with disease being the ecosystem that fails to function (Schaeffer et al., 1988)
- Ecosystems as organisms with health as property (Suter II, 1993)

However, nowadays, the most widely accepted definition of ecosystem health states that it is mainly characterised by resilience, vigour and organisational structure (Constanza & Mageau, 1999; Rapport et al., 1998; Tett et al., 2013). Those reflect the ecosystems' ability to recover from perturbations, as well as reflecting its function and structure (Lu et al., 2015). Importantly, those are the major indicators on an ecosystem level, with differing indicators for community, population and individual level (Lu et al., 2015). Tett et al. (2013) showed that the vigour and organisation of the ecosystem are what both determine and are maintained by the resilience of an ecosystem, while the resilience buffers the ability of the ecosystem to be able to provide ecosystem functions against external pressures that are driven by humans (Fig. 1). In addition to this definition, ecosystem health is also closely linked to sustainability, which means, among others, establishing generational justice so that the current and future needs of the society are met in regard to ecosystem services (Lu et al., 2015). Hence, the definition of ecosystem health can also consider future generations, which has also been found to be strongly linked to the resilience of ecosystems (Cote & Darling, 2010).

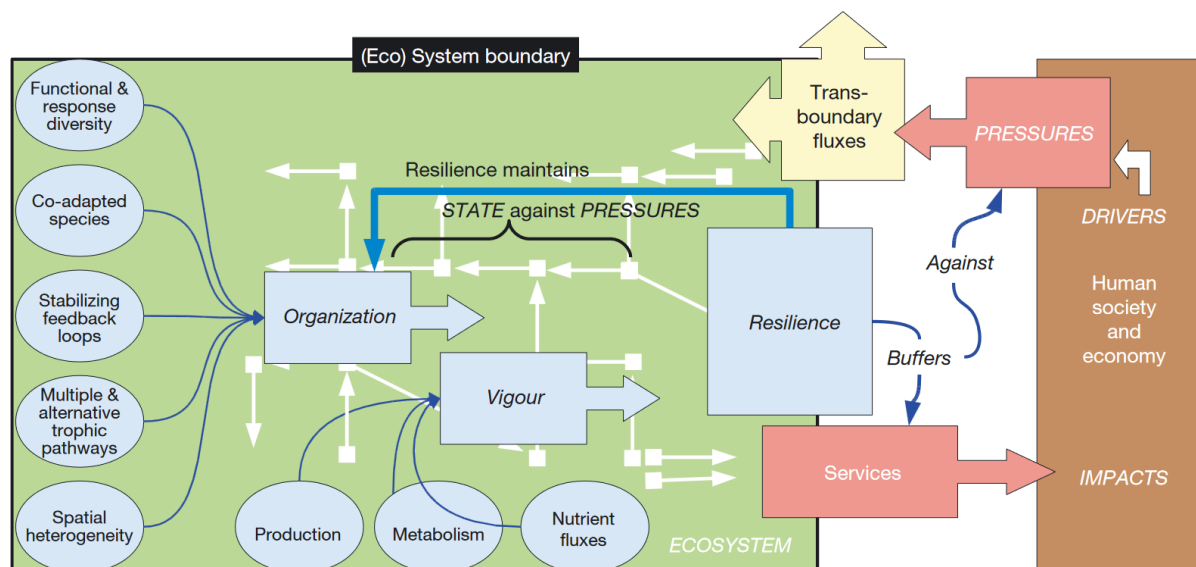


Fig. 1. Conceptualisation of ecosystem health by Tett et al. (2013). Ecosystem health is here defined as "the ability to maintain systems integrity" (Tett et al., 2013).

2.3.1 Resilience

Ecosystem resilience is commonly defined as withstanding perturbations without a regime shift, leading to losing ecosystem services and functions (Cote & Darling, 2010; Tett et al., 2013). Perturbations can be natural or human stresses on the ecosystem and can dampen productivity, disrupt nutrient cycling, lead to biodiversity loss and other dysfunctions of the ecosystem (Rapport et al., 1998). Human and natural stresses can also interact to degrade an ecosystem further; e.g., anthropogenic stress initially weakening an ecosystem can make it more

vulnerable to natural stresses, further degrading the state of the ecosystem (Lu et al., 2015). The term can be divided into two components: resistance and recovery of an ecosystem (Lu et al., 2015). Resistance refers to the size of the perturbation that causes the ecosystem to change its structure. In contrast, recovery refers to the speed at which it returns to its original organisational structure (Cote & Darling, 2010).

Regarding ecosystem resistance, it is commonly believed that a reduction in local stressors leads to a decrease in the impact of climate change (Peleg et al., 2023; Pereira et al., 2010). Cote and Darling (2010) suggest that this assumption could be flawed and propose a more differentiated view. A wide range of functional traits of species, with group-specific responses to disturbances, was suggested to be of critical importance for the resilience of ecosystems (Chillo et al., 2011). Aligned with this, Cote and Darling (2010) found that “selective culling” of perturbation-sensitive species by stresses could explain why more intact coral reef systems are not necessarily more resilient. Naturally, the initial diversity of an ecosystem is an essential factor, as a high diversity increases the chance of an ecosystem hosting more resilient species. Then, the more an ecosystem degrades, the more it will be composed of more disturbance-tolerant species. Thus, this might shift an ecosystem’s tipping point (climate change response) towards a higher climate disturbance being needed to trigger it and consequently make it more resistant. However, it was widely found that protected or more healthy ecosystems are recovering more quickly (Cote & Darling, 2010; Seidl et al., 2016).

The opposition between the two components of resilience raises the question of which one needs to be prioritised. Cote and Darling (2010) argue that resistance of ecosystems should be the focus, as the frequency of perturbations is expected to increase (Seidl et al., 2016). Since some of those perturbations will not be acute (e.g. ocean acidification, temperature increase) (Abram et al., 2016), they would leave no room for recovery. Mumby and Steneck (2008) found similar results and thus recommend concentrating on protection in local areas that are expected to be exposed to less disturbance in the future.

2.3.2 Vigour

As previously discussed, the vigour of an ecosystem mainly refers to it being able to function (Rapport et al., 1998). Simply put, vigour is a function of production, metabolism and nutrient fluxes (Tett et al., 2013) (Fig. 1). This includes the gross primary productivity of an ecosystem (fluxes of energy and material), its carrying capacity (Basset et al., 2013) as well as its ability to recover from perturbation through growth and recolonisation (Tett et al., 2007). This highlights again how intertwined resilience (recovery) and vigour are. All the processes determining ecosystem vigour are dependent on primary production. However, the relationship between ecosystem health and primary production was found to not be linear and is defined as the balance between too few and an excess of nutrients (in a marine ecosystem) (Tett et al., 2007).

2.3.3 Organisational Structure

Tett et al. (2013) define the organisational structure of ecosystems as dependent on five factors, which include - among others - species diversity, response and functional heterogeneity, co-adaptation and trophic pathways (Fig. 1). Other structural facets of ecosystems include species richness and distribution patterns (Cannicci et al., 2008). Generally, the idea of ecosystem structure is that it was formed by individual species adapting to each other and the abiotic environment (Kroes, 1977). However, the set of taxa in an ecosystem does not necessarily have to be unchangeable, and minor fluctuations in the species composition of a given system are likely to leave the system as a whole intact (Kroes, 1977).

2.4 Ecosystem Health Metrics

There are varying methods to estimate ecosystem health. While some attempt to take a holistic approach to estimate all aspects together, others try to estimate ecosystem health by assessing some of its different aspects, namely resilience, vigour and organisational structure. Additionally, one has to consider what kind of ecosystem one is assessing (e.g. marine or terrestrial). Lu et al. (2015) summarised some findings by Jørgensen et al. (2011) and showed the differences in key species used for ecosystem health estimation for e.g. terrestrial and aquatic ecosystems, but also for human-coupled systems like agricultural systems (Table 1). Moreover, they presented measurement indicators that apply to the different systems.

Table 1. Summary of key species and measurement indicators by ecosystem type by Lu et al. (2015).

Ecosystem types	Key species	Measurement indicators
Natural ecosystem		
Forests	indicator plants, sensitive plants, specific plants, specific animals, birds, soil animals, soil microbes	composition of flora and fauna; species diversity, biomass; primary production; proportion of exotic species; quantity and quality of habitats; ecosystem-level indices; ecosystem services
Wetlands	indicator plants, sensitive plants, specific plants	same as forests
Rivers	plankton, high-trophic-level fish	species diversity; biomass; primary production; proportion of exotic species; vertical structure and horizontal distribution of species; reproduction and regeneration of species; quantity and quality of habitats; ecosystem level indices; ecosystem services
Lakes	plankton, high-trophic-level fish	same as rivers
Estuaries and coastal zones	plankton, benthic invertebrates, high-trophic-level fish	same as rivers
Marine ecosystems	plankton, benthic invertebrates, high-trophic-level fish	same as rivers
Human-coupled ecosystem		
Agricultural ecosystem	soil animals, soil microbes, plant weed species	species diversity, primary production, ecosystem services
Urban ecosystem	indicator plants, sensitive plants, birds	species diversity, quantity and quality of habitats, ecosystem services

2.4.1 Indicators

The measurement indicators by Lu et al. (2015) can be classed as “descriptive” and “functional” indicators (Rombouts et al., 2013). Descriptive indicators relate to the structure of an ecosystem (e.g. species richness, composition and diversity), while functional indicators assess the

activities (e.g. primary productivity, nutrient cycling) (Rombouts et al., 2013). Those indicators often complement each other, which should be considered when selecting them. On the other hand, O'Brien et al. (2016) distinguished measurement indicators as biological, physical and chemical (the latter usually only relevant to aquatic ecosystems). However, they similarly recommend using a minimum of one biological indicator species/group in addition to chemical and/or physical indicators to holistically mirror the primary connections within the ecosystem and obtain a more reliable assessment of ecosystem health. Nonetheless, it is essential to note that the assessment of ecosystem health can still be unreliable despite the use of multiple diverse indicators if the justification for the selection of indicators is insufficient (Lindenmayer & Likens, 2010), i.e. if “clear predictions about how they [the indicators] represent ecosystem health.” (O'Brien et al., 2016) is lacking.

An ideal ecosystem health indicator should satisfy seven criteria according to Noss (1990). It has to be “(1) sufficiently sensitive to provide an early warning of change, (2) distributed over a broad geographical area or otherwise widely applicable, (3) capable of providing a continuous assessment over a wide range of stress, (4) relatively independent of sample size, (5) easy and cost-effective to measure, collect and calculate, (6) able to differentiate natural cycles or trends from those induced by anthropogenic stress, and (7) relevant to ecologically significant phenomena” (Noss, 1990). After satisfying those criteria, the choice of an indicator is also dependent on the spatial level that it has to represent, as indicators can range from a cellular to a community and regional landscape level (Rombouts et al., 2013). Therefore, indicators need to be carefully selected for each case (and according to the desired spatial extent), with a combination of multiple indicators representing different aspects of ecosystem health being preferable (Lu et al., 2015; Rombouts et al., 2013).

2.4.2 Key Species

As introduced in the previous section, indicator species are also commonly used as indicators for the overall health of an ecosystem (Siddig et al., 2016). For broader relevance, indicator species would need to be sampled at several sites, which is rarely done (Bakker, 2008). Several indicator taxa were found to be of low applicability as they are very spatially-temporally heterogeneous. However, dung beetles were shown to be reliable indicators. Additionally, terrestrial plants are suggested to be reliable due to a high consistency (Bakker, 2008). Also, certain taxa can be used to reliably reflect certain aspects of ecosystems, like lichen that are often used to assess and monitor air quality (Yang et al., 2023). It was found that indicator species analysis can be simplified by using binary presence/absence data compared to classically abundance data (Bakker, 2008). Besides the advantages of using indicator species that are seen by many (e.g. parameters of an individual population are relatively easy to measure), there are several limitations, among them:

- the use of a single taxon or population commonly cannot reflect the complexity of an environment (Siddig et al., 2016),
- unclear aims of monitoring and connection between indicator selection and its role in the environment (O'Brien et al., 2016)

- limited consideration of further biological interactions in communities (like predation or symbiosis) (Siddig et al., 2016)
- sampling difficulties (Siddig et al., 2016)

Most importantly, studies should not rely on a single indicator metric but rather apply multiple metrics (O'Brien et al., 2016; Siddig et al., 2016).

2.4.3 Biodiversity Carrying Capacity

The state of an ecosystem can also be assessed by comparing it to a hypothetical maximum health, i.e. a perfectly healthy ecosystem (Peleg et al., 2023; Siddig et al., 2016). However, in reality, such pristine ecosystems that are commonly seen as perfectly healthy are very hard to find (Lackey, 2011; Siddig et al., 2016). Moreover, one has to consider that in all ecosystems – including pristine ecosystems – there is no natural state of equilibrium but rather random fluctuations over time (Lackey, 2011). Thus, since reference data for healthy ecosystems is challenging to obtain, it can be modelled with a “state space” method (Tett et al., 2013). For this, one needs to set reference points for the indicators (Peleg et al., 2023). Romero et al. (2007) measured values of their selected indicators and averaged the highest values to estimate the values of these indicators in a “healthy ecosystem”, thus setting their reference points to which they compared all other indicator values. In comparison, Tett et al. (2013) set the reference points randomly to normalise and compare the indicator values. In either case, the distance to the reference points can be compared to estimate how healthy an ecosystem is, with values closer to the reference points, meaning the ecosystem is healthier (Tett et al., 2013). The state space method holds the advantage that the number of indicators used is not limited, thus being able to describe many different aspects of an ecosystem (Peleg et al., 2023). Thereby, it can be a multivariate approach that is more likely to be able to reflect the health of an ecosystem accurately (Peleg et al., 2023).

Similar to the state space methods, the concept of biodiversity carrying capacity can be used to compare ecosystems and assess their relative health (Storch et al., 2022). Some taxa exhibit certain traits that make them well-suited as indicators for ecological and biogeographical analysis (Brown, 1991). Hence, an assessment of the abundance of unique species of a suitable taxon in different ecosystems could provide a similar baseline to compare ecotopes in that ecosystem against and assess their health status (Storch et al., 2022). This baseline is likely to be an unattainable ideal status. “Biodiversity carrying capacity” suggests a fixed capacity of ecosystems for a particular species richness. On the contrary, it was suggested to be a local equilibrium around which the species richness oscillates over time (Storch et al., 2022). This equilibrium was hypothesised to emerge as a function of biodiversity-dependent extinction and speciation rates that are modified by available resources, the ability of the communities to use the resources, temperature, topography and environmental stability (Storch et al., 2022). Thus, the biodiversity carrying capacity remains a valuable concept even when an ecological system is largely non-equilibrium as it allows gauging the non-equilibrium state compared to an ideal but likely unattainable equilibrium state (Storch et al., 2022). The concept will also be referred to as biodiversity potential in this study.

2.5 Clustering Algorithms

Clustering algorithms are essential tools of data mining that are only increasing in relevance in times of big data. These algorithms can be used to group subsets of a dataset to create clusters that are similar based on a given set of criteria within themselves while being as different as possible from other clusters (Saemi et al., 2018). Among others, they are commonly applied to delineate subunits in the landscape based on abiotic and/or biotic variables (e.g. Oldeland et al. (2010), Soto and Pintó (2010)). For unsupervised clustering algorithms, the data is initially unlabelled, and the clustering is grouping the data solely based on its attributes (Swarndeeep Saket & Pandya, 2016). Following Gao et al. (2023), unsupervised clustering algorithms can be grouped into four major groups: centre-based partitioning clustering, hierarchical clustering, density-based clustering and model-based clustering (Fig. 2).

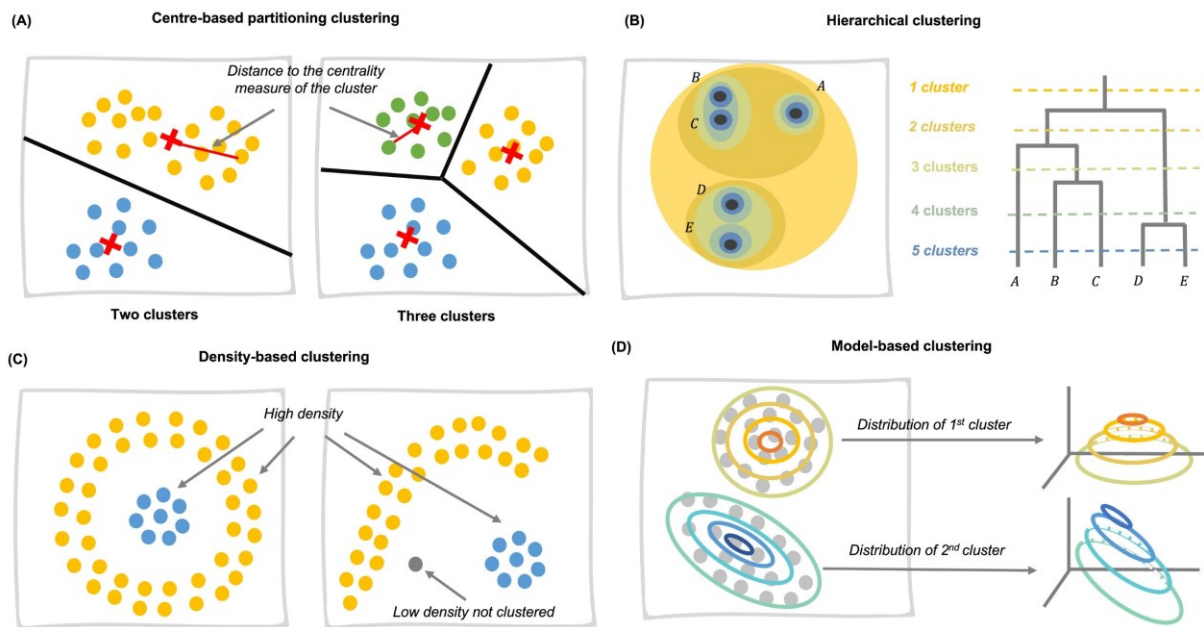


Fig. 2. Overview of the main types of clustering algorithms, divided into four groups. Source: Gao et al. (2023)

2.5.1 Partitioning Clustering Algorithms

Partition-based clustering algorithms divide data into k partitions, with each partition representing a cluster and k being required as input by the user (Jin & Han, 2010). The algorithm aims to minimise an objective error function iteratively.

The most popular partitioning clustering algorithm is the K-means algorithm (MacQueen, 1967), as it is easy to implement and relatively fast. It randomly places k centroids in the data, and each data point is assigned to the closest centroid. As shown in Fig. 3, the centroids are recalculated to be in the centre of the assigned data samples, and the data is repartitioned based on the distance to the nearest centroid (Gao et al., 2023). This process is repeated until convergence. While the speed of K-means and its ability to handle large datasets are a big advantage, the algorithm also has some inherent limitations. It can be relatively sensitive to outliers in the data and is at risk of getting stuck in local minima when repartitioning based on

random starting points. Additionally, it assumes all clusters to be spherical and approximately of equal size (Gao et al., 2023).

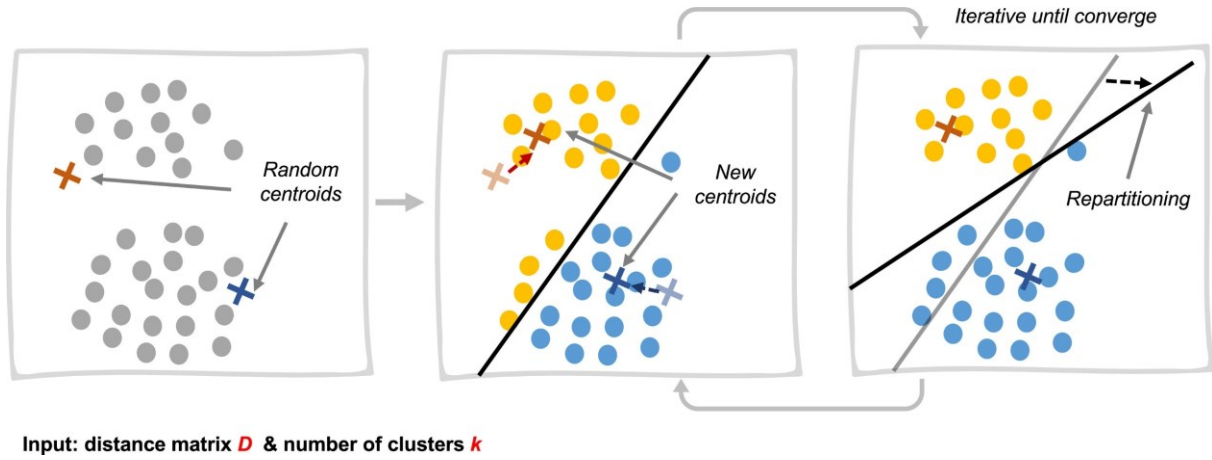


Fig. 3. K-means clustering schematic. Source: Gao et al. (2023)

2.5.2 Hierarchical Clustering Algorithms

There are two types of hierarchical clustering algorithms: agglomerative and divisive ones (Swarndee Saket & Pandya, 2016). Agglomerative algorithms are initiated by treating each sample as an individual cluster and then iteratively identifying the clusters closest to each other. The nearest (most similar) cluster pairs are then merged. This approach is also known as the bottom-up approach and is the more commonly used one (Krohn et al., 1999). The divisive algorithms (top-down approach) work the other way around in that they initially assume all samples to belong to one cluster and then iteratively split the cluster based on dissimilarity (Swarndee Saket & Pandya, 2016).

Popular examples of hierarchical clustering algorithms are single linkage, complete linkage, average linkage, Ward's method and BIRCH (Balance Iterative Reducing and Clustering using Hierarchies) (Gao et al., 2023). The linkage and Ward's methods are explained in Fig. 4. BIRCH is an algorithm developed to handle larger datasets, as most hierarchical clustering algorithms require extensive computing resources and time (Zhang et al., 1996). The single linkage is an exception to that; however, it often tends to form chains of small clusters (Gao et al., 2023). Ward's method, on the other hand, has the limitation of yielding clusters of similar sizes (Gao et al., 2023).

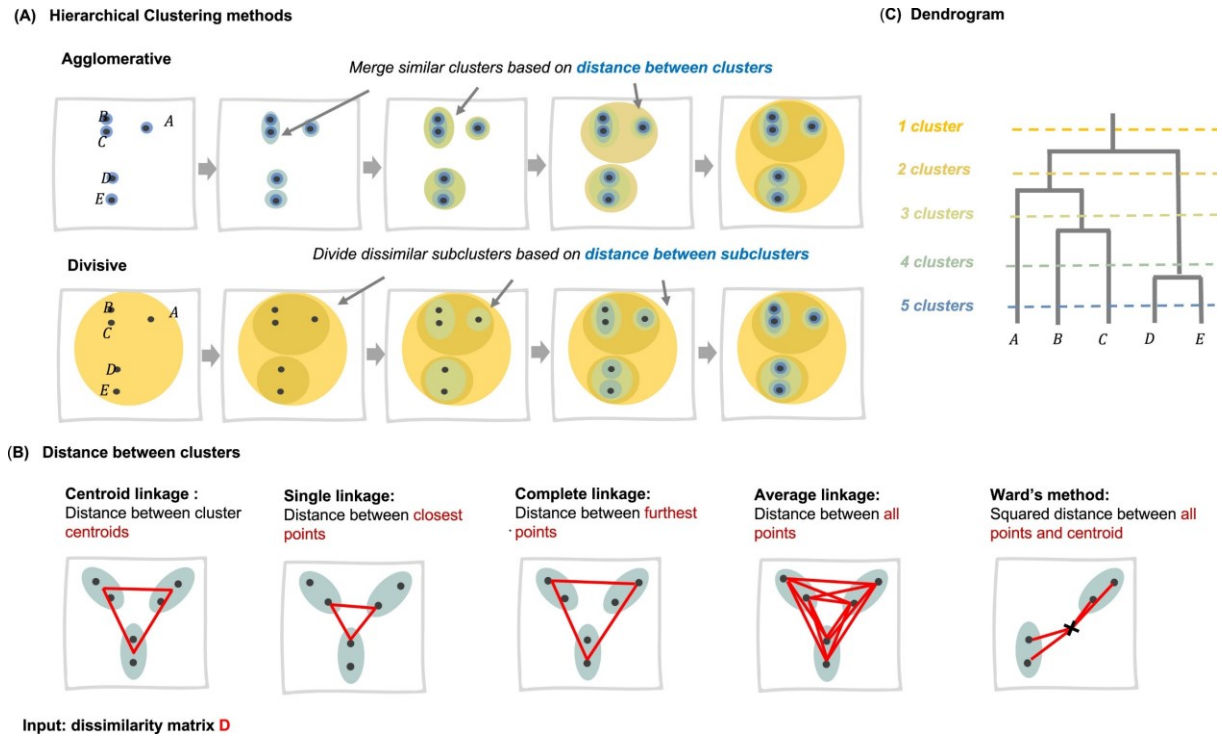


Fig. 4. Description of Hierarchical Clustering methods and popular distance metrics. Source: Gao et al. (2023)

2.5.3 Density-based Clustering Algorithms

Density-based algorithms can identify clusters with arbitrary shapes and sizes. They identify clusters as areas of higher density when compared to the data space around it (Bhattacharjee & Mitra, 2020). DBSCAN (Density-based algorithm for discovering clusters in large spatial databases with noise) is an example for a popular density-based algorithms. It works well for datasets with areas of higher density that are well-distinguished but can struggle with large datasets with high dimensionality, as well as for datasets with variable levels of density (Bhattacharjee & Mitra, 2020). Hence, they are unlikely to perform well for ecotope clustering.

2.5.4 Model-based Clustering Algorithms

Model-based clustering algorithms form clusters based on a probability distribution (Bouveyron & Brunet-Saumard, 2014). That means that the multivariate data will be assigned to clusters based on the assumption that the data was generated from a finite mixture of component models (Banerjee & Shan, 2010). For instance, the Gaussian Mixture Model (GMM) is one of the most popular models, and it assumes a Gaussian distribution for each component (Banerjee & Shan, 2010). Model-based algorithms generally generate good clusters for data with outliers and varying densities. Additionally, it can, unlike e.g. K-means, identify elliptical clusters as well as overlapping clusters (Gao et al., 2023). However, these algorithms do not perform well with arbitrarily shaped clusters and are at risk of converging to local maxima (Gao et al., 2023).

2.6 Summary

To summarise, the physical environment of communities will be referred to as either habitat or ecotope in this study. An ecotope is defined as a subunit of an ecosystem of undefined size. Additionally, spatial ecosystem units can be defined according to their ability to meet species habitat requirements but are more commonly delineated based on physical, more permanent parameters. The latter was also suggested to be more reliable in reflecting the biotic conditions (Krohn et al., 1999).

Ecosystem health is nowadays mostly defined as being a state that is characterised by resilience, vigour and organisational structure (Constanza & Mageau, 1999; Tett et al., 2013), since these reflect an ecosystems' capability to recover from disturbances and its structure and function. Indicators are an important and commonly applied method to estimate ecosystem health, however, they need to be carefully selected, with intention of representing certain aspects of ecosystem health. Also, a combination of multiple indicators representing different aspects of ecosystem health would be preferable (Lu et al., 2015; Rombouts et al., 2013).

A rather efficient method to assess ecosystem health for larger areas and different ecosystems was proposed to be the comparison of an ecotope to a hypothetical maximum health, i.e. a perfectly healthy ecosystem (Peleg et al., 2023; Siddig et al., 2016). The biodiversity carrying capacity (also biodiversity potential) concept could provide such a comparison by gauging the ecotope biodiversity in comparison to the total biodiversity potential of its ecosystem cluster. This biodiversity potential would ideally be based on several taxa but could also be computed for a single taxon.

To initially group ecotopes into ecosystem clusters based on ecotope variables, unsupervised clustering algorithms can be applied. Since density-based clustering algorithms are less likely to perform well for such a task, partitioning, hierarchical or model-based clustering algorithms would be preferred. While hierarchical and model-based algorithms would be expected to perform better, partitioning algorithms are highly advantageous in terms of time and computing-resource efficiency.

3. Research Objectives

While the assessment of the health status of ecosystems is a topic that is particularly relevant in light of climate change and preservation efforts (Hoffmann et al., 2010), many studies primarily focus on assessing the status of one particular ecosystem. Even though this will likely yield the most accurate results, this methodology is very time-intensive, and the results might only be valid locally or regionally. This approach increases the difficulty for policymakers and managers in assessing the general state of a type of ecosystem in a region, which is necessary to implement measures to protect ecosystems effectively.

Hence, this study aims to establish a methodology to assess the biodiversity potential for different ecosystems in Germany, in order to serve as a baseline for future ecosystem health studies. The biodiversity potential assessment will be applied to ecosystem clusters created using different unsupervised clustering algorithms. Based on the literature review (section 2.5), a partitioning, hierarchical and model-based clustering algorithm will be selected and compared in performance. The clustering algorithms will be trained on a dataset describing German ecotopes by biophysical variables, as those were found to be reliable in reflecting biotic variables. Grouped together, the ecotopes will form ecosystem clusters. The algorithm that forms the most suitable ecosystem clusters will be selected for further use. Then, the biodiversity potential will be assessed for all ecosystem clusters with sufficient taxon data samples. Even though several taxa would be preferable to reflect the complexity of ecosystems, only one taxon will be used for the computation of biodiversity potential in this study due to time constraints. This limitation will be considered in the interpretation of the results of this study. To demonstrate the use of the biodiversity potential, it will be applied to assess the health of ecotopes in given ecosystem clusters. Specifically, this study aims to answer the following research questions:

1. Which clustering method is the most suitable for the clustering of ecotopes in Germany?
2. Are existing datasets suitable as a basis for the development of this method?
3. Can the quality of the ecosystem clusters and biodiversity potential be validated with existing datasets?
4. Can regional differences in ecosystem clusters and biodiversity potential in Germany be captured as expected?

4. Study Area and Data

In the following the study area and its characteristics, as well as datasets used for training and validation will be introduced.

4.1 Study Area

Germany was selected as study area since it is a large country with a range of diverse ecosystems. Moreover, it is a country that is well-documented with an abundance of openly available data and the familiarity of this study's author with the country's biogeography will likely aid in the analysis of the results.

In the North, Germany is dominated by non-irrigated arable land, interspersed with coniferous and broad-leaved forests in the northeastern part (Fig. 5). Additionally, the region close to the North Sea is mainly comprised of pastures. In contrast, the southern part of Germany is dominated by broad-leaved and coniferous forests, regionally interrupted by arable land and urban fabric. The large urban centres stand out notably, with Hamburg in the north, Berlin in the northeast and Munich and Stuttgart in the southeast and southwest, respectively. Additionally, the heavily industrialised Ruhr District in the west stands out with five larger cities.

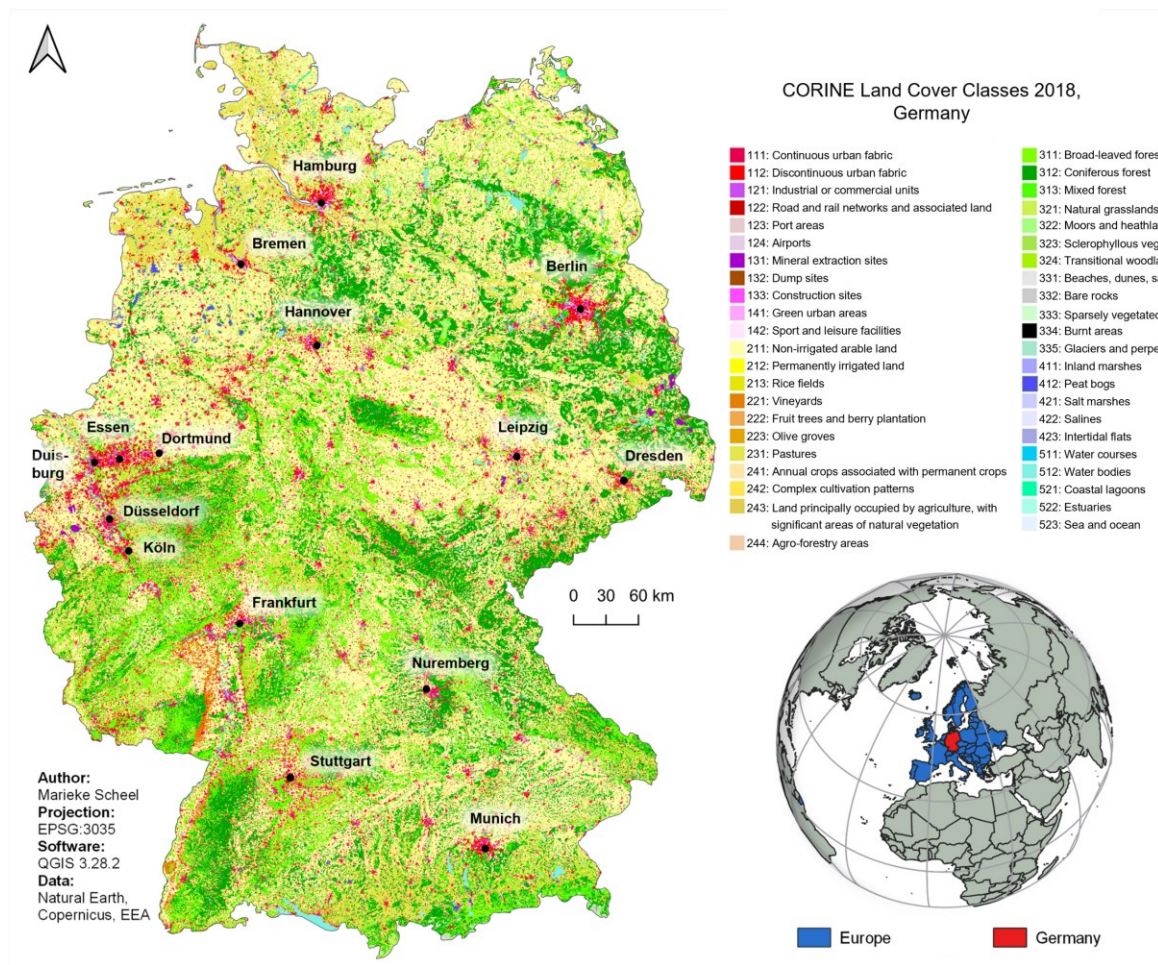


Fig. 5. CORINE land cover (CLC) in the study area of Germany and location of 15 major cities. CLC dataset stems from the European Environment Agency (EEA).

Germany's biogeography is characterised by adjacency to the Baltic and North Sea in the Northern part and by the Alps in the Southern part (Fig. 6). While the northern part of the country is notably flat (Fig. 7) and divided into the North-East and North-West German Plain, the elevation slowly rises southwards with central to southern Germany being divided into the low mountain regions. Finally, in the very southeast, the foothills of the Alps and a small part of the Northern Calcareous Alps are visible.

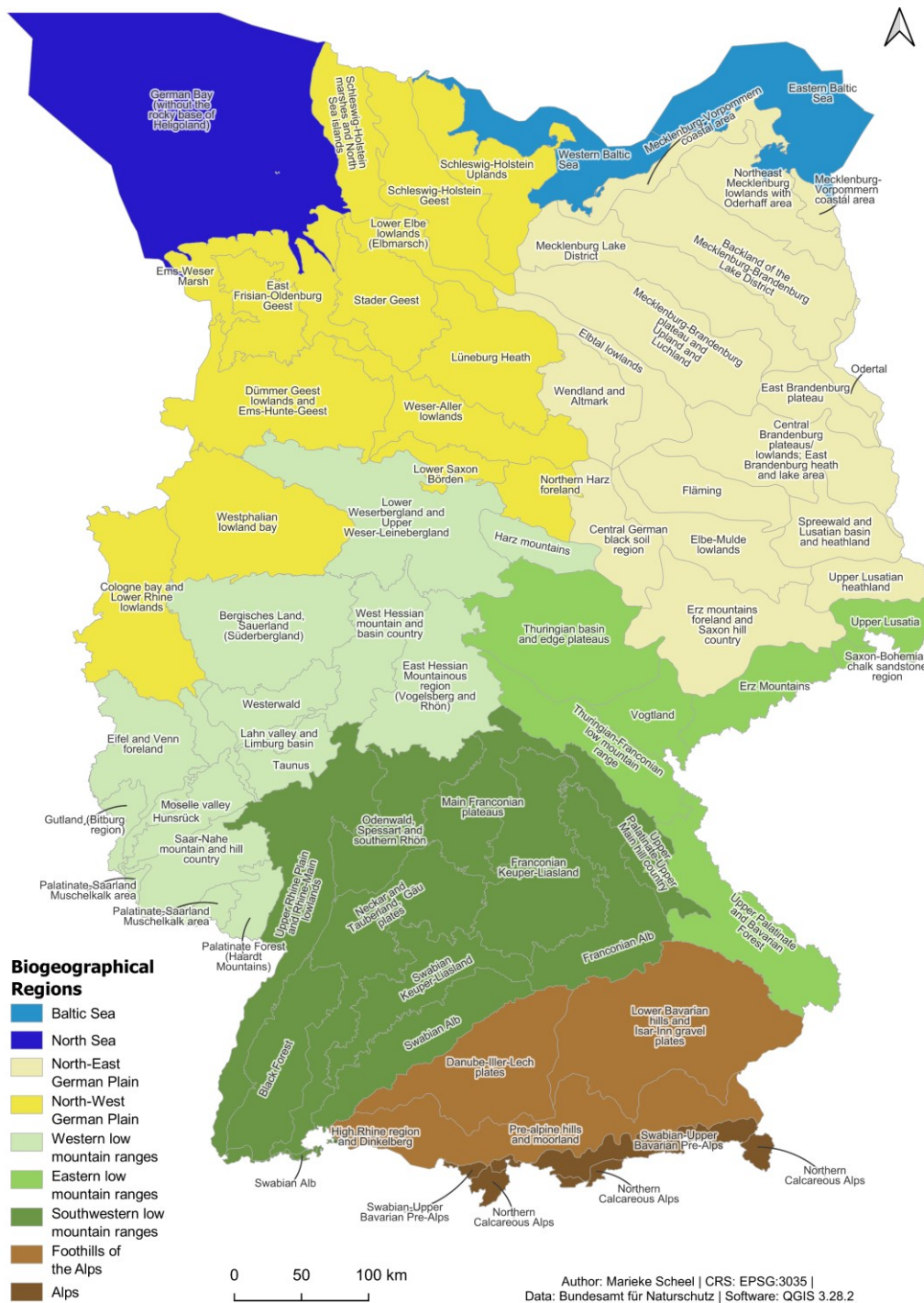


Fig. 6. Major biogeographical regions in Germany shown in colours, and more detailed region names are highlighted with labels. The data stems from the German Federal Agency for Nature Conservation (Bundesamt für Naturschutz) and was manually translated from German to English by the author of this study.

The Elbe, Weser, Rhine, Main and Danube are some of the biggest rivers in Germany (Fig. 7). Most of the mountains can be found in the south of Germany. Nonetheless, the Harz Mountains, Westerwald, Thuringian Forest, Ore Mountains (or Erz Mountains) and Eifel are low mountain ranges in central Germany. While the Northern (Bavarian) Alps dominate the southeast, the southwest is characterised by the Black Forest and the Swabian Alps (or Swabian Alb).

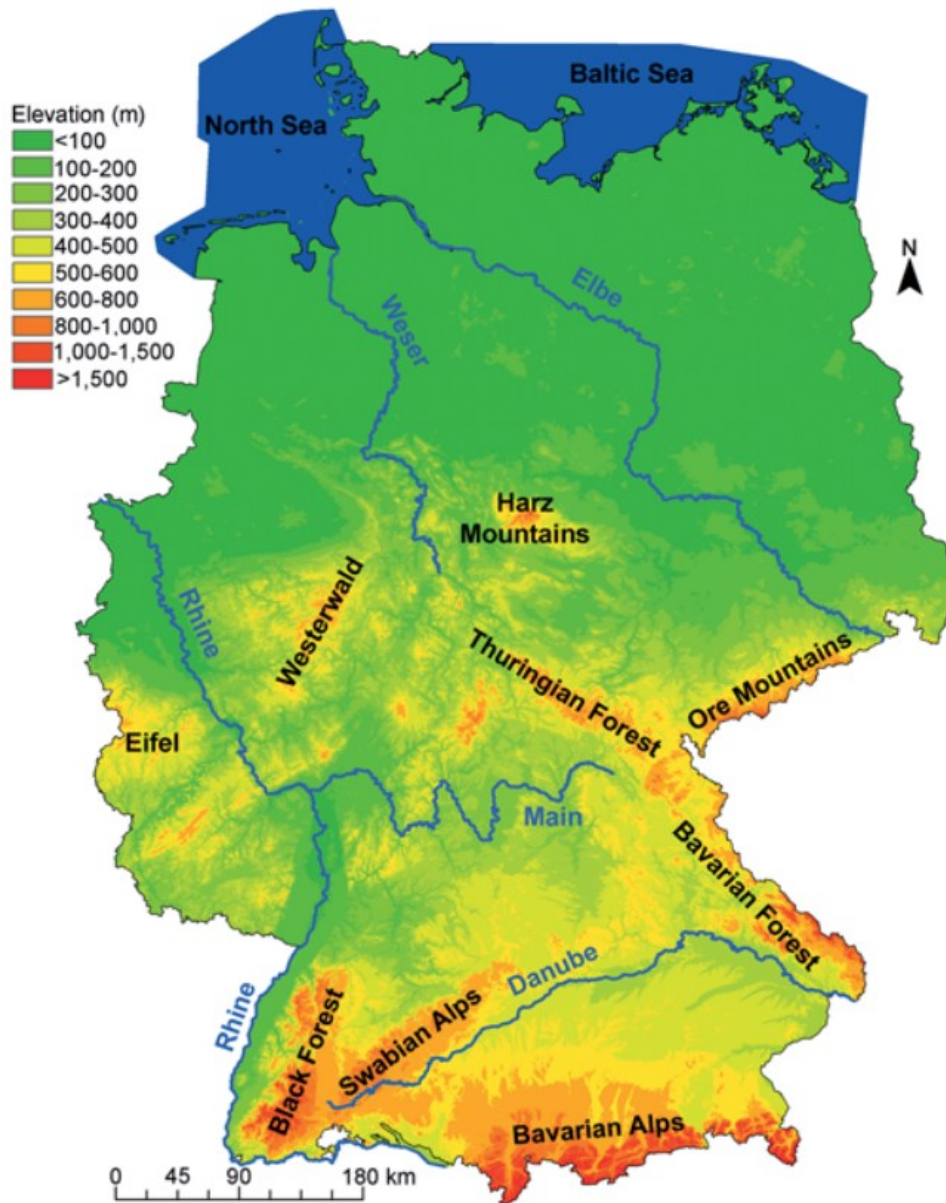


Fig. 7. Major rivers and mountain ranges in Germany. Source: Behr et al. (2021)

4.2 Data

The following sections will introduce the four datasets that were used for training and validation in this study. The LifeWatch dataset was used to train the clustering algorithms to form ecosystem clusters. The EUNIS habitat dataset was used to validate the clustering results and choose the most suitable clusters for further analysis. It should be noted that the EUNIS dataset includes data of potential habitats, as part of its samples are from the 19th and 20th centuries. Meanwhile, the LifeWatch dataset includes variables representing land cover in 2018. Nonetheless, both datasets are considered the most appropriate to use for this study, in light of their abundance of integrated data. Similarly, GBIF integrates an abundance of open-access data from several organisations. GBIF taxon data was used for the computation of the

biodiversity potential of the different ecosystem clusters. Finally, the Natura 2000 dataset was used to validate the results of biodiversity potential.

4.2.1 LifeWatch Dataset

The ecotope dataset for Germany stems from the LifeWatch European Research Infrastructure Consortium (ERIC). LifeWatch provides open access to it. The dataset consists of nearly 1 million ecotope patches that are each defined by 119 biophysical variables (Table 2). The dataset used in this study is a pre-released version and the accuracy assessment of the data will only be available in September 2024. The variables are explained in more detail in Table A2. They include climatic, topographic, landcover, soil property and distance variables. Spatial and temporal resolution vary among variables and can be explored in more detail in Table A2 and Table A3. The variables were derived from different sources, among them remote sensing, the LifeWatch landcover classification and historical climate data (Table A3).

Table 2. Biophysical variables

Biophysical variables					
Water	lczDv	MGramin_2km	AnVar_P	dmi_Sealed	dme_NeedLV
SnowIce	Water_1km	NOpen_2km	AnVar_SolR	dma_Sealed	dmi_NeedLV
Bare	SnowIce_1km	WOpen_2km	AnM_SolR	dme_Building	dma_NeedLV
Sealed	Bare_1km	Sparse_2km	GDD_5deg	dmi_Building	dme_BroadLV
Building	Sealed_1km	NeedLV_2km	elev_mean	dma_Building	dmi_BroadLV
Plowed	Building_1km	BroadLV_2km	elev_min	dme_Plowed	dma_BroadLV
MGramin	Plowed_1km	Shrub_2km	elev_max	dmi_Plowed	dme_Shrub
NOpen	MGramin_1km	SunSpring	slp_mean	dma_Plowed	dmi_Shrub
WOpen	NOpen_1km	TPI_250m	slp_min	dme_MGramin	dma_Shrub
Sparse	WOpen_1km	NightLi	slp_max	dmi_MGramin	TPI_1km
NeedLV	Sparse_1km	AnM_T	dme_Water	dma_MGramin	clay_5
BroadLV	NeedLV_1km	AnRg_T	dmi_Water	dme_NOpen	sand_5
Shrub	BroadLV_1km	AnVar_T	dma_Water	dmi_NOpen	silt_5
Majority	Shrub_1km	MaxWarmM_T	dme_SnowIce	dma_NOpen	soc_5
LCCSb	Water_2km	MinColdM_T	dmi_SnowIce	dme_WOpen	cec_5
ecosystem31Mj	SnowIce_2km	Snow_len	dma_SnowIce	dmi_WOpen	bdod_5
ecosystem31Dv	Bare_2km	AnTot_P	dme_Bare	dma_WOpen	cfvo_5
crotypeMj	Sealed_2km	WetM_P	dmi_Bare	dme_Sparse	nitrog_5
crotypeDv	Building_2km	DryM_P	dma_Bare	dmi_Sparse	phh2o
lczMj	Plowed_2km	An_PotETP	dme_Sealed	dma_Sparse	

The ecotope patches have different area sizes, with patches in Northern Germany and regions with water being generally larger (Fig. 8). Along the entire coastline, the patches are much larger than on land. Also, the patches in central Germany appear to be the smallest.

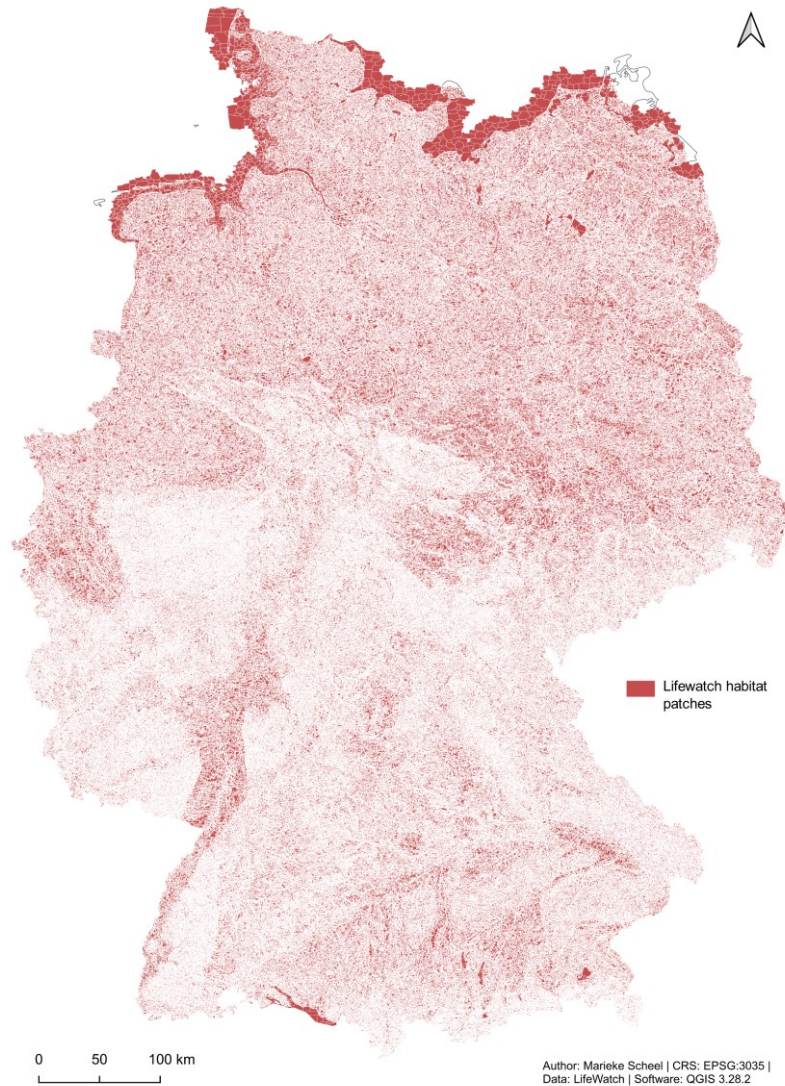


Fig. 8. Ecotopes of LifeWatch data in Germany shown as individual red polygons outlined in white. Source: LifeWatch ERIC (2017)

4.2.2 EUNIS Habitat Dataset

The European Nature Information System (EUNIS) aims to compile and distribute European data about types of species, protected sites and habitats (EEA, 2019). Details about the habitats are provided by the EUNIS habitat classification (Table A6). The data includes habitat classifications from the 19th, 20th and 21st centuries. The classification is hierarchical and has up to three levels per class. The habitat types were harmonised to provide a detailed and harmonised pan-European description of the different types of habitats. In 2012, the EUNIS habitat classification was extensively reviewed, and the habitat groups revised (EEA, 2019). The maintenance and collection of the EUNIS data is managed by the European Topic Centre on Biological Diversity, the European Environmental Information Observation Network and the European Environment Agency (EEA). For the habitat determination, a comprehensive key was developed that guides the user in the form of a series of questions towards the habitat

determination (EEA, 2019). The habitats are classified at a sample size between 1 and 100 m² and all represented by a single coordinate pair (EEA, 2019).

This study uses the currently most up-to-date EUNIS version that consists of the marine version 2022 and the terrestrial version 2021. However, some habitat groups are still in review and were thus unavailable for this study, including Inland water, Constructed, industrial and other artificial habitats and Complexes. The following habitat classes were included in this study: coastal habitats, grasslands, heathlands, forest and other wooded lands, sparsely vegetated habitats, vegetated man-made habitats and wetlands.

In Germany, there are 111 unique EUNIS classes and 120,000 sample points. The EUNIS points are distributed relatively well all over Germany. However, there are areas with notably fewer points and areas of higher concentration (Fig. 9). There is a higher concentration of points in the west of Germany and the northeast.

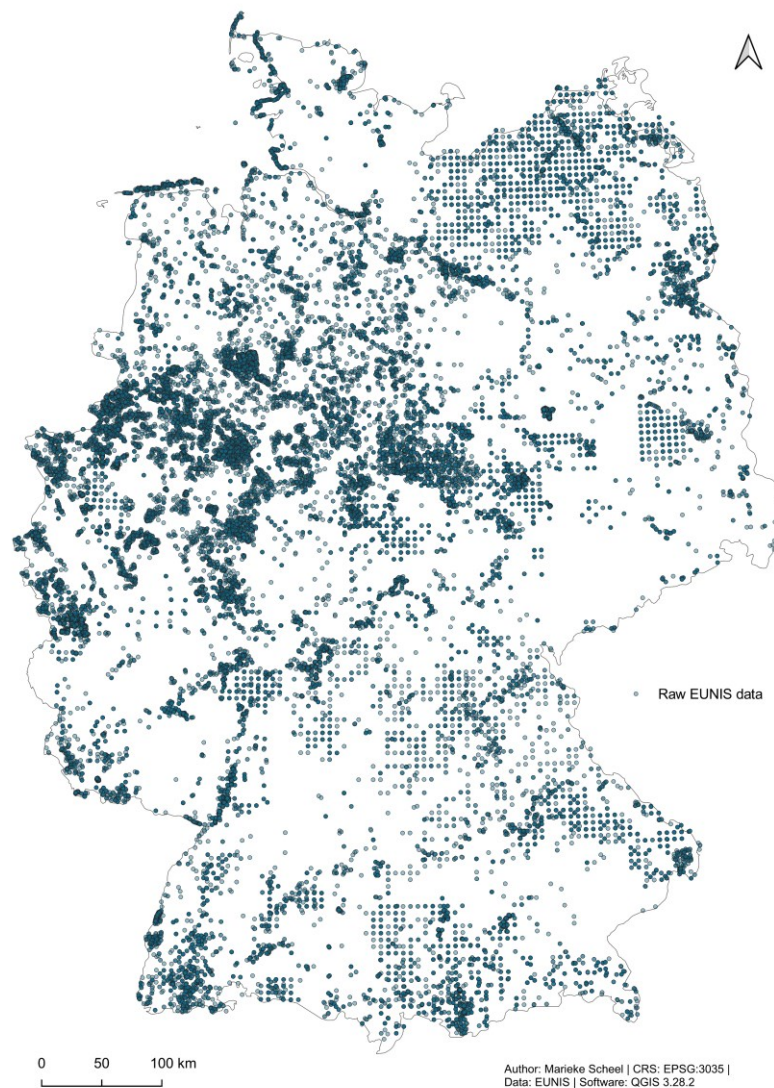


Fig. 9. Spatial distribution of the raw EUNIS data in Germany.

4.2.3 GBIF Taxon Dataset

The Global Biodiversity Information Faculty (GBIF) is a network and a data infrastructure that operates globally. Their objective is to provide open access data about global flora and fauna (GBIF, 2024). By implementing common practices and standards, they aim to provide data that can be used independently from the country of origin and those who collected them. Additionally, only organizations can publish their data on the website. They have varying data sources, from museum collections from the 18th century to recent smartphone pictures. Overall, they have collected hundreds of millions of species data, mostly occurrence data from field observations by researchers or citizen science efforts but also from e.g. camera traps and remote sensing.

Otegui et al. (2013) analysed species data of the Spain node of the GBIF network and found that they generally have a low error level. However, they also showed that the database includes a strong spatial bias due to national publishing so subsets lead to notably better model predictions. Globally, the completeness of inventories differs significantly, with them being generally the most complete in European countries (Meyer et al., 2015). These results support the selection of Germany as study area, since it has a high level of completeness in the GBIF inventory (Meyer et al., 2015) and choosing one country decreases spatial biases created by national publishing.

Brown (1991) classified various insect groups regarding their traits that would make them desirable as “indicators for biogeographical and ecological analysis and monitoring” (Brown, 1991). He distinguished between 12 desirable traits that were necessary for a good indicator species (Table A1). Overall, species should be “relatively immobile, have detectable trends, and baseline studies examining tolerance levels and correlation of the taxa to changes in the ecosystem” (Brown, 1991) to be used as an indicator. The total value as an indicator was then assessed out of 24. The *Vespidae* taxon (wasps in English) received one of the higher scores, 18. Moreover, GBIF had more than 71,000 samples of *Vespidae* available within Germany, and the data samples were collected all over the country (Fig. 10). As the data is mainly collected through citizen science efforts (in this case, ca. 47,000 samples stem from NABU (2024) and another at least 10,000 samples from other citizen science collection), it inherently includes uncertainties.

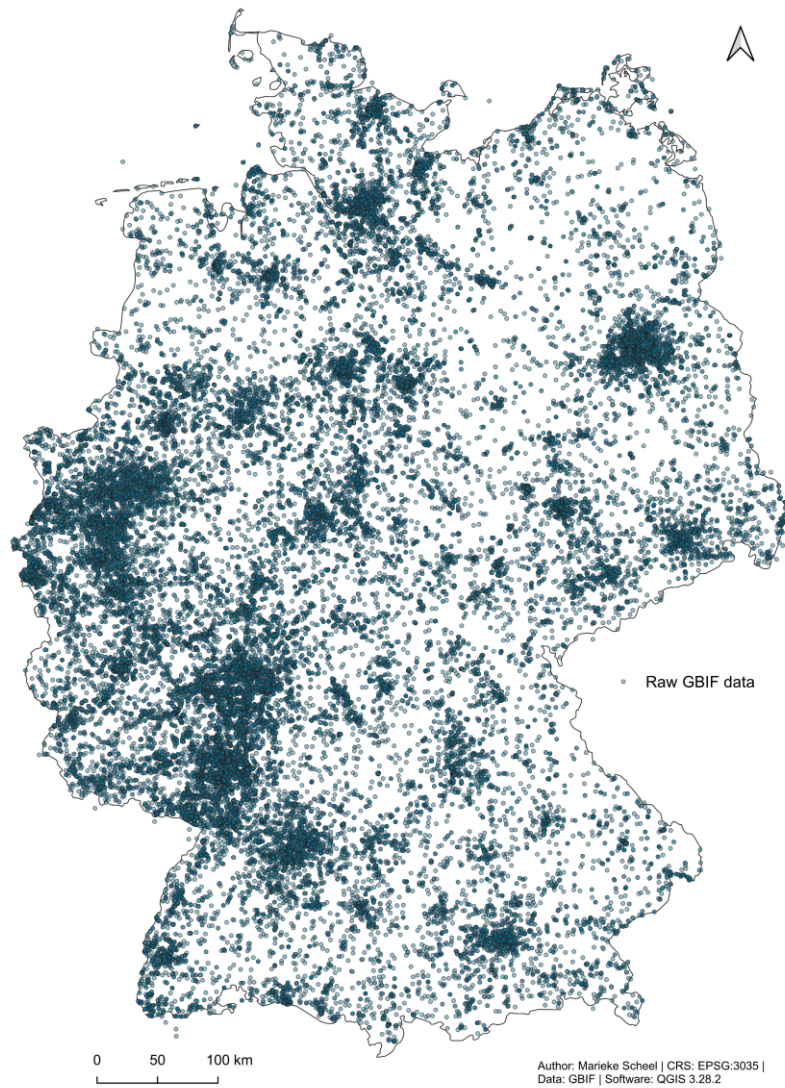


Fig. 10. Spatial distribution of the raw GBIF data in Germany.

4.2.4 Natura 2000 Dataset

Natura 2000 is a network of an EU-wide protection area that was established in 1992 (BMUV, 2023). It aims to preserve endangered and/or valuable species and habitats. The protection areas are designated under the Birds and Habitat directive (Special Areas of Conservation and Special Protected Areas) and selected according to EU-standards. There are currently more than 4,500 Special Areas of Conservation and 742 Special Protected Areas in Germany. Also, 15.5 % of Germany's area on land and ca. 45% of its marine area were protected under Natura 2000 in 2023. Humans can still use the protected areas, however, that use is constrained to aid in the conservation of the habitat and species (BMUV, 2023).

The Natura 2000 habitats are relatively evenly distributed over Germany (Fig. 11). However, the size of the habitats is notably bigger in northeastern, central and southwestern Germany.

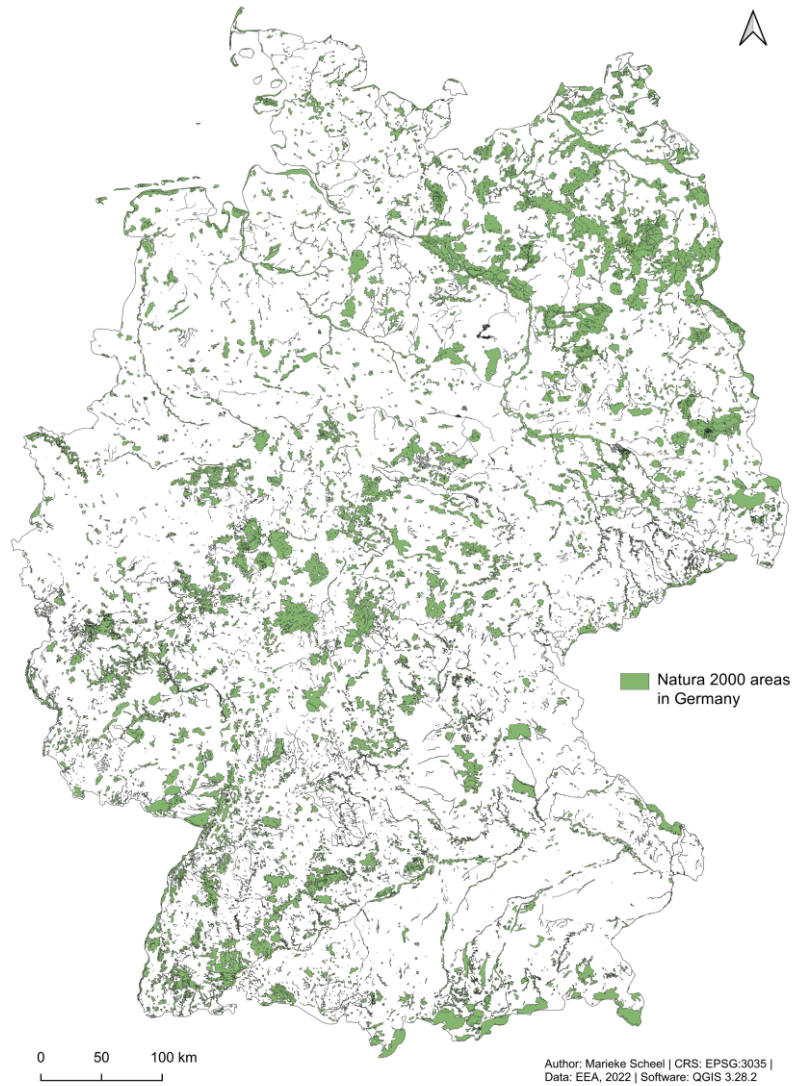


Fig. 11. Natura 2000 habitats in Germany. Source: EEA (2022)

5. Methods

Unless otherwise specified, the analysis was completed utilising Python 3.9.15 and QGIS 3.28.2. Additionally, the high-performance computing (HPC) cluster Rocket by the University of Tartu (<https://hpc.ut.ee/services/HPC-services/Rocket>) was used to run the clustering operations. The Rocket HPC is heterogeneous, consisting of ca. 64 computing nodes with almost 13,000 cores, as well as 64 Graphics Processing Units (GPUs). The HPC can be accessed using the Simple Linux Utility for Resource Management (SLURM) system, which is an open-source management and job scheduling system for Linux HPCs (Yoo et al., 2003). For each clustering task, a RAM (Random-Access Memory) of 400 GB was allocated.

5.1 Overview Workflow

A workflow diagram illustrates the most important processes, inputs and outputs that were conducted to answer the research questions (Fig. 12). Initially, the LifeWatch ecotope dataset was quality controlled and three different unsupervised clustering algorithms were used to group the ecotopes to ecosystem clusters. Based on the literature review, one partitional (K-means), one hierarchical (Genie) and one model-based clustering (GMM) algorithm were selected. The choice of 200 clusters will be explained in the following. Then, the most suitable cluster set was chosen by computing a purity score of the EUNIS habitat classes within the clusters. Hence, it was assumed that the EUNIS habitat classes are comparable to the LifeWatch ecotope patches. The implications and potential shortcomings of this assumption will be discussed in-depth in the discussion section. For additional validation, the clusters created by the Genie algorithm (highest purity score) were also labelled using the EUNIS habitat dataset. Next, the Genie clusters and GBIF wasp dataset were intersected to calculate the biodiversity potential for all clusters with more than 100 GBIF samples. The wasp biodiversity potential was then used for an arbitrary, relatively compact cluster to demonstrate its use for a health assessment. Additionally, the five most populated cities in Germany were selected, so that similar ecosystems that were already spatially delineated could be compared in regard to their biodiversity potential and ecotope health. The health assessment for the cities as well as the nation-wide wasp biodiversity potential were compared to Natura 2000 data for validation, with the assumption that the protected Natura 2000 zones should coincide with a higher biodiversity potential.

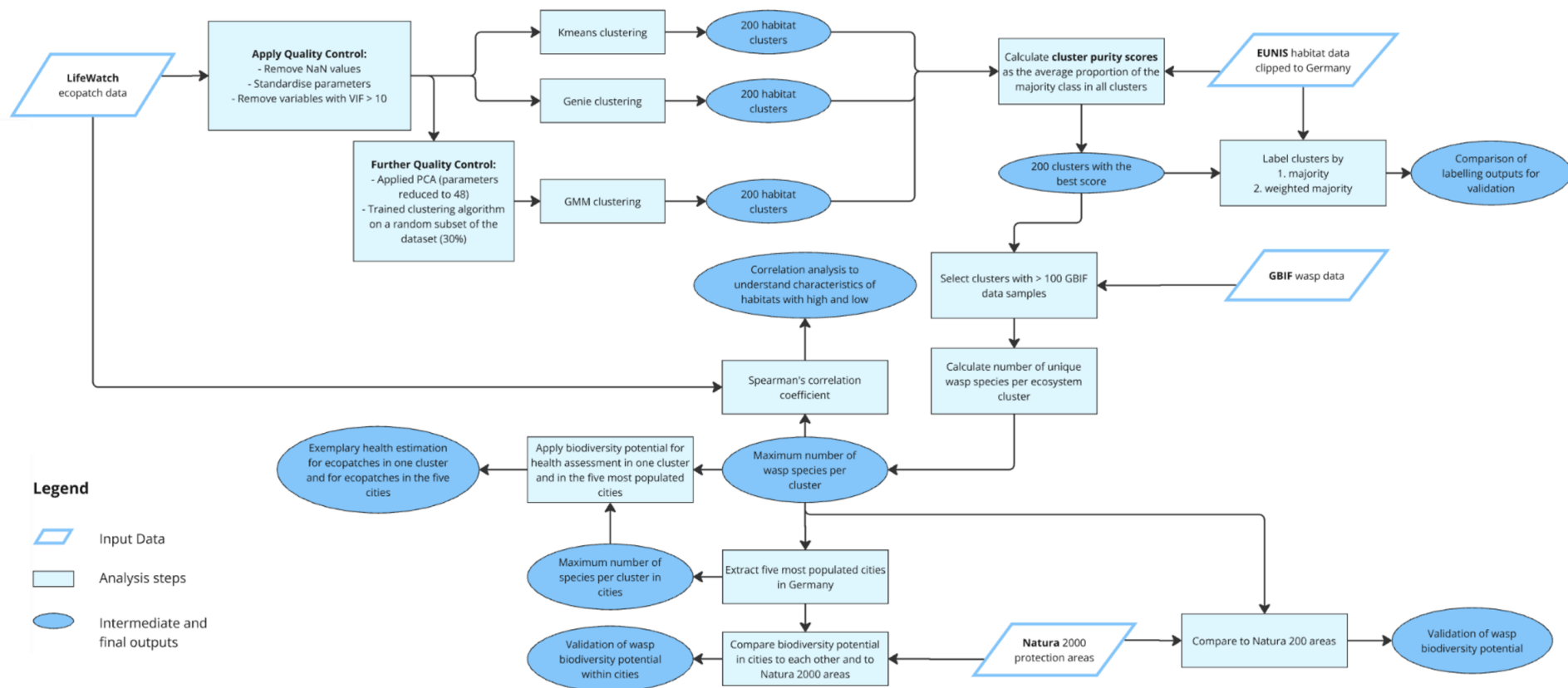


Fig. 12. Overview workflow describing the major steps of the analysis conducted for this study.

5.2 Unsupervised Clustering

5.2.1 Applying Quality Control to LifeWatch Data

The clustering was based on the dataset by LifeWatch, which defined 925,436 ecotopes in Germany that were each distinguished by 119 biophysical variables. Initially, the dataset was cleaned, and all ecotopes with NaN values were dropped to avoid introducing bias in the clustering.

In a next step, all biophysical variables were standardised by “removing the mean and scaling to unit variance” (Scikit learn, 2024d). The standard value z of the sample X is then calculated as

$$z = \frac{(X - u)}{s} \quad \text{Eq. 1}$$

where u is the mean of X and s is the standard deviation of X (Scikit learn, 2024d). Since clustering models are distance-based algorithms that measure the similarity between individual observations using a distance metric, standardisation is a necessary step in the quality control process to prevent parameters with high value ranges from exhibiting a larger influence on the clustering (Ismaili et al., 2014).

Then, the dataset was examined for correlated variables. Since more than two variables can be correlated with each other, this study tested for multicollinearity, which occurs if several of the biophysical variables are highly correlated. As clustering is based on distance metrics, correlated variables would assign more weight to the dependent variables (Kyriazos, 2023). This might not be a significant problem for the clustering, as it does not rely on the assumption of linearity (Kanungo et al., 2002). However, as the LifeWatch data is a very big dataset, a reduction in dimensionality is vital to run the clustering algorithms time and resource efficiently. A Variance Inflation Factor (VIF) analysis was conducted to determine and remove highly correlated variables was preferred over applying the also popular Principal Component Analysis (PCA), as the former preserves the interpretability of the data (Kyriazos, 2023). The VIF factor can be used to assess the degree to which the variance of one parameter is inflated due to multicollinearity (Neter et al., 1983). It is calculated as follows,

$$VIF_i = \frac{1}{(1 - R_i^2)} \quad \text{Eq. 2}$$

where R_i^2 is the R^2 -value attained from regressing the predictor i on the remaining predictors (Neter et al., 1983). A VIF factor of 1 corresponds to an R^2 -value of 0, implicating that this parameter has no multicollinearity, while a factor around 4 might indicate a moderate correlation. Typically, a parameter with a VIF value above 10 is considered to have significant multicollinearity (Kyriazos, 2023). This study calculated the VIF factor for all 119 biophysical variables on a random subsample of 30% of the dataset due to high computing power needs and

time constraints. When a VIF factor was determined for each variable, the variable with the highest VIF value was removed, and the VIF factor was recomputed for every variable. This iterative computation was conducted to account for the change in relationships (Neter et al., 1983). This iteration was continued until no parameter had a VIF factor above 10, removing 38 parameters (Table 3, Table A4) from the LifeWatch dataset. The quality-controlled LifeWatch dataset thus consists of 81 variables (Table A5).

Table 3. The 38 biophysical variables of the LifeWatch dataset removed due to a VIF factor > 10.

Removed biophysical variables		
Plowed	dme_NOpen	WetM_P
Plowed_1km	AnTot_P	elev_min
dme_Sparse	dma_Plowed	Sealed_2km
Plowed_2km	An_PotETP	dma_NeedLV
MinColdM_T	elev_max	NeedLV_1km
dme_WOpen	dme_MGramin	dma_NOpen
dma_Sparse	dme_SnowIce	Bare_1km
elev_mean	dme_BroadLV	NOpen_1km
dme_Plowed	dme_Water	BroadLV_1km
dme_Shrub	dme_Building	AnRg_T
dmi_WOpen	sand_5	slp_mean
dme_NeedLV	dme_Bare	Shrub_1km
MaxWarmM_T	dma_Shrub	

Three cluster algorithms were successfully applied to the quality-controlled LifeWatch dataset, including K-means, Genie and Gaussian Mean Mixture (GMM). Additionally, hierarchical clustering methods, namely BIRCH and Ward, and Spectral clustering, were attempted. However, this large dataset's required computing power and time for these clustering algorithms could not be provided. Hence, the analysis was only conducted with the other three algorithms. All clustering algorithms aimed to distinguish between 200 clusters, as that is a number that likely leads to over- than undersegmentation of clusters; for comparison, 111 unique EUNIS habitat classes were found in Germany. An oversegmentation is more desirable as it increases the probability that all the important boundaries were extracted (Troya-Galvis et al., 2015). This usually happens at the cost of extracting many less significant boundaries (Troya-Galvis et al., 2015). This could be rectified in a post-hoc processing step through informed merging of clusters. In the following, the clustering algorithms that were successfully applied will be explained in more detail.

5.2.2 K-means Clustering Algorithm

Lloyd's K-means clustering algorithm is arguably considered to be one of the most popular unsupervised machine learning algorithms (Kanungo et al., 2002). It is a partitional clustering algorithm that aims to increase the similarity of the points within a cluster, while trying to increase the dissimilarity to data points within other clusters (Nanda & Panda, 2014). Each cluster in K-means is represented by a centroid, which is calculated as the arithmetic mean of

all points within a cluster (Nanda & Panda, 2014). Iteratively, the points are assigned to the centroid closest to them; then the centroid is recomputed, taking the average of all the points in the cluster, following which the points are reassigned to the nearest centroid (Kanungo et al., 2002). This process is repeated until the sum of the distances between a centroid and its corresponding data points is minimised, and the centroid values remain unchanged or run through a given maximum number of iterations. The initial locations of the 200 cluster centroids were determined through the *kmeans++* method (Arthur & Vassilvitskii, 2007) that selects the seeds “based on an empirical probability distribution of the points’ contribution to the overall inertia [(i.e. variance)]” (Scikit learn, 2024b). This method increases the convergence speeds of clusters compared to a random distribution of initial clusters (Arthur & Vassilvitskii, 2007). Additionally, the K-means algorithm is at risk of getting caught in local minima (Kanungo et al., 2002). Hence, it was initiated ten times with different seed initialisations, selecting the result with the lowest variance. The maximum number of iterations of the algorithm for a single run was set to 5000, a high number chosen to ensure that convergence was reached before the maximum number of iterations was reached. This study used the K-means python package implemented by *scikit learn* (Scikit learn, 2024b).

5.2.3 Genie Clustering Algorithm

The Genie clustering algorithm was proposed by Gagolewski et al. (2016). An agglomerative hierarchical clustering algorithm offers a new, fast-computing linkage criterion. The Genie linkage criterion was developed to improve the single linkage criterion by addressing its weaknesses while retaining its advantages regarding required computing power. While clusters determined by applying the single linkage criterion often differ significantly in size, the Genie linkage criterion applies an inequity measure (namely the normalised Gini-index) to retain a given level of inequity (Gagolewski et al., 2016). Gagolewski et al. (2016) found that the Genie clustering algorithm yielded better overall results than the hierarchical clustering with the single, average or Ward’s linkage criterion, as well as clustering with the K-means and BIRCH algorithms. Additionally, they estimated a reduction in computing power compared to many other popular hierarchical clustering linkage criteria, making its performance comparable to the single linkage criterion. In this study, the python package *genieclust* (Gagolewski, 2024) with the Genie++ hierarchical clustering algorithm was used (faster reimplementation of the original algorithm), and the default Gini-index threshold value of 0.3 was applied (the smaller the threshold, the more the creation of small clusters is penalised). The pairwise distances are computed using the Euclidean distance metric.

5.2.4 GMM Clustering Algorithm

The GMM is a probabilistic clustering algorithm that assumes that the data generally follows a multivariate normal distribution (McLachlan & Peel, 2000). Each cluster is assigned a probability distribution by modifying a multivariate Gaussian distribution to fit the components of the subpopulation (Wang et al., 2019). In contrast to K-means, GMM allows the cluster components to have a unique orientation, shape and size (Agarwal et al., 2021). This study used the Gaussian Mixture python package implemented by *scikit learn* (Scikit learn, 2024a). The

GMM will estimate the membership of samples in a cluster with an iterative Expectation Maximisation (EM) algorithm that will run until the clusters converge. Generally, one could describe the GMM as behaving similarly to the K-means algorithm; besides that, it assesses the probability density compared to a deterministic classification of data points based on distance (Agarwal et al., 2021). Unfortunately, the GMM algorithm has high computational demands for high-dimensional data, with the number of model parameters rising quadratically with the number of dimensions. Hence, additional pre-processing, namely a PCA, was applied to reduce data dimensions and computing time. The number of components for the PCA was determined by applying the common practice of calculating the cumulative explained variance and finding the number of components that explain 95% of it (Groth et al., 2013). This was found to be 48, effectively reducing the number of components by more than a third from a previous total of 81. Lastly, as the algorithm was still too computationally expensive, the GMM model was trained on a random subset representing 30% of the data. Summary statistics were compared to ensure that the subset was representative of the entire population, and a Kolmogorov-Smirnov (KS) test was conducted (Massey, 1951). The KS-test tests whether the training data distribution follows the entire dataset's distribution (Berger & Zhou, 2014) and can thus provide more information about the random subset.

5.3 Cluster Validation

The clustering quality of the three algorithms was later assessed using extrinsic and intrinsic measures. Additionally, they were compared in terms of total area covered by clusters and geographical intra-cluster distances. The former was visually compared in a histogram plotting the frequency of cluster areas and a kernel density estimate (KDE) plot. Simplified, a KDE plot is a smoothed-out, continuous histogram plot that represents the data using a continuous probability density curve (Chen, 2017). The geographical intra-cluster distances were computed by calculating each cluster's Euclidean distance between the two points furthest apart. The geographical intra-cluster distance is a measure of compactness and cohesion of the clusters and could thus provide insights into whether an algorithm has created more local or regional ecosystem clusters.

5.3.1 Intrinsic Metrics Assessing Cluster Quality

The intrinsic cluster quality metrics estimate the quality by assessing the intra- and inter-cluster distance, with an ideal clustering minimising intra-cluster distances and maximising inter-cluster distances (Hassani & Seidl, 2017). Three intrinsic measures were calculated. Firstly, the silhouette score, which determines the quality as

$$\text{Silhouette score} = \frac{b - a}{\max(a, b)} \quad \text{Eq. 3}$$

where a is the average intra-cluster distance, and b is the average inter-cluster distance (Rousseeuw, 1987). The resulting values will range from -1 to 1, with 1 indicating an ideal

clustering, 0 indicating that the inter-cluster distances are not significant, and -1 indicating that the means of the clusters are wrongly assigned (Rousseeuw, 1987).

Secondly, the Davies-Bouldin Index (DBI) was computed, which measures the average similarity between each cluster and the cluster it is most similar to (Davies & Bouldin, 1979). Hence, clusters with less dispersion that are well-separated will receive a better score using this measure (Hassani & Seidl, 2017). It is calculated as

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} \left\{ \frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)} \right\} \quad \text{Eq. 4}$$

with $\delta(X_i, X_j)$ representing the inter-cluster distance and $\Delta(X_k)$ being the intra-cluster distance within the cluster k (Davies & Bouldin, 1979). The intra-cluster distances are thus calculated between each point belonging to a cluster and the centroid of a given cluster, while the inter-cluster distances are computed between the centroids of the respective clusters. Unlike the silhouette score, which assumes clusters to be roughly spherical, the DBI does not assume a cluster shape and can thus be better suited for more complex clusters (Karo et al., 2017). A lower DBI indicates a higher cluster separation (Hassani & Seidl, 2017).

The third intrinsic cluster quality assessment was the Calinski-Harabasz Index (CHI), which calculates the ratio of the sum of the inter-cluster variance to the sum of the intra-cluster variance (Harabasz & Karoński, 1974). The CHI for K clusters for a dataset $X = [x_1, x_2, x_3, \dots, x_N]$ is defined as

$$CHI = \frac{\left[\frac{\sum_{k=1}^K n_k \|c_k - c\|^2}{K - 1} \right]}{\left[\frac{\sum_{k=1}^K \sum_{i=1}^{n_k} \|x_i - c_k\|^2}{N - K} \right]} \quad \text{Eq. 5}$$

with c_k and n_k representing the number of points and the centroid of the k^{th} cluster (Harabasz & Karoński, 1974). A higher CHI value indicates well-separated and compact clusters (Hassani & Seidl, 2017). The value is dependent on the number of clusters as well as the dispersion of the data points (Harabasz & Karoński, 1974).

Overall, the CHI indicates cluster quality in terms of separation and compactness and can aid in determining the optimal number of clusters. At the same time, the silhouette score is more simplistic and intuitive and aids in assessing the consistency of the clusters. The DBI can be computed more quickly and does not assume cluster shapes but can be easily influenced by outliers.

5.3.2 Extrinsic Metrics Assessing Clusters Quality

5.3.2.1 Applying Quality control to the EUNIS Habitat Data

The EUNIS dataset was utilised for an extrinsic clustering quality measure. Initially, all EUNIS habitat datasets were merged and clipped to Germany. Then, the data was cleaned to remove any points with the same label, location, and year, which were likely generated due to coordinate inaccuracies or imprecisions. Additionally, points with different labels but the same location and year were removed, as they are ambiguous, and it cannot be determined which points were incorrectly assigned. Lastly, any points with different labels in the same location in a different year were kept as it can be assumed that the habitat has changed and that this location has the potential for both habitats. Generally, EUNIS data samples from all years were kept as the clustering of the LifeWatch data aggregated the data into clusters of potential habitats according to the environmental parameters rather than actual land use. Hence, an EUNIS sample from e.g. 1950 is assumed to still reflect the potential habitat. Assuming classes with too few samples were not meaningful enough, all classes with less than 30 samples were removed, leaving 68 EUNIS classes out of the original 111.

The quality-controlled data was then compared to the original dataset to estimate if and how the previous steps changed the proportion of classes or spatial distribution. The proportion of classes in the data was compared by plotting the number of samples per class (grouped to the first level). Meanwhile, the spatial distribution was compared visually, and the Gini coefficient was computed for both datasets. The Gini coefficient is a measure of inequality (Catalano et al., 2009) and can be applied to spatial data by assessing the differences in the points count in a grid generated for Germany (10x10 km). The Gini coefficient of inequality is calculated according to

$$G = \frac{\sum_{i=1}^n (2i - n - 1)x_i}{n \sum_{i=1}^n x_i} \quad \text{Eq. 6}$$

where n is the number of observations (number of grid cells), i is the rank of the observation from smallest to largest, and x_i is the count of points of the i^{th} observation sorted from smallest to largest (Moyes, 2007). The resulting G indicates all measurements being equal at 0 and the highest degree of inequality at 1 (Moyes, 2007).

5.3.2.2 Assess clusters with most accurate habitat representation

After the quality control, the EUNIS data points were overlayed with all three clustering results to compute a simple cluster purity score, as conducted e.g. by Amigó et al. (2009). The proportion of the majority class was assessed for each cluster, and from this, an average proportion, i.e. purity score, for each clustering algorithm could be computed.

As the data was not normally distributed, a non-parametric Kruskal-Wallis test was additionally conducted to test for significant differences between the purity scores. The Kruskal-Wallis test is rank-based and used to test for statistically significant differences between two or more

independent variables and a dependent variable (Kruskal & Wallis, 1952). Then, since the test showed significant differences, the Dunn's test was computed. The Dunn's test is a non-parametric post-hoc test that tests pairwise for statistically significant differences between groups and is typically applied after a Kruskal-Wallis test (Dunn, 1964). The Dunn's test was Bonferroni corrected, which means that the increased chance of a false positive due to performing several tests was accounted for (Emerson, 2020).

The extrinsic metric results were prioritised over the intrinsic metric results when selecting the best clustering result to assess the biodiversity potential. Since this study used several clusters that promote oversegmentation, this might be reflected in the intrinsic metric results due to overlapping clusters, making the extrinsic results likely more reliable. Nonetheless, the intrinsic metrics will provide a good insight into the differences between the clustering algorithms.

In addition to the purity assessment, the clusters of the algorithm with the highest purity scores will be labelled based on the EUNIS data. As the EUNIS data has a vastly different sample size for different habitat classes, two methods were applied and compared. Due to time constraints, no technique was developed to combine the two approaches, which would have allowed clusters to be merged based on the labelling and resume the following analysis with labelled and merged clusters. The first approach was to label the clusters by the majority EUNIS class among the samples within each cluster. The second method calculated the weight of one sample point for each EUNIS class based on the total amount of samples for each class (i.e. a point of a class with 30 members had a higher weight than a point of a class with 1000 members). Then, the clusters were labelled by assigning the EUNIS class with the highest weighted majority per cluster.

5.4 Establishing Baseline for Ecosystem Health

5.4.1 Applying Quality Control to GBIF Data

A rigorous quality control was applied to the GBIF data to ensure that only the most reliable data points were used for assessing biodiversity carrying capacity. The dataset initially contained about 60,000 points within Germany. Then, all samples that did not contain information regarding coordinates, species names or coordinate uncertainty were discarded. Additionally, the data was clipped to the outlines of Germany, effectively removing any points with incorrect coordinate values. Subsequently, data samples with a very high coordinate uncertainty were removed by applying the interquartile range (IQR) method. The IQR is a measure of variability for data with outliers or generally skewed distributions (Vinutha et al., 2018). It commonly considers a sample's lower (q_1) and upper (q_3) percentile to correspond to the 25th and 75th percentile, while the IQR is the interval between q_1 and q_3 . Thus, the outliers are defined as being $1.5 * IQR$ above q_3 or below q_1 (Fig. 13). Since it was applied to coordinate uncertainties of the points in this study, only the samples with values $1.5 * IQR$ above q_3 were considered. The remaining sample points were then buffered by their coordinate

uncertainties (Fig. A1) and overlaid with the clusters to retain only the points that are within a given cluster with high confidence.

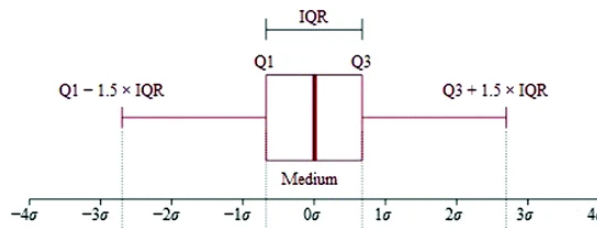


Fig. 13. Description of the method to detect outliers using the interquartile range. Source: Vinutha et al. (2018)

The quality-controlled data was compared to the original dataset to estimate if and how the previous steps changed the proportion of classes or spatial distribution. Firstly, the proportion of species classes in the original and quality-controlled data was compared by plotting the number of samples per class. Then, the spatial distribution of the data was assessed visually and by comparing the Gini coefficient of inequality for the original and quality-controlled data (Eq. 6).

Further, the clusters intersecting with more than 100 EUNIS data samples were extracted. That minimum threshold yielded 42 clusters and was chosen arbitrarily with the presumption that it would increase the significance of the results of the biodiversity carrying capacity.

5.4.2 Validation of the Wasp Biodiversity Potential Assessment

A Spearman's Rank correlation was conducted between the total area covered and the unique species per cluster, as well as the total number of GBIF sample and the unique species per cluster to test for dependencies. The Spearman's correlation measures the statistical dependence between a dependent and an independent variable without assuming the data to be normally distributed (Akoglu, 2018). The coefficient will show a perfect positive correlation as 1 and a perfect negative correlation as -1 (Akoglu, 2018). A dependency between those variables can be expected as larger clusters are likely more diverse and can thus be expected to provide shelter for a larger variety of wasps.

Then, the wasp biodiversity potential assessment was validated using the Natura 2000 dataset. The cluster areas were intersected with the Natura 2000 areas in Germany to compare the biodiversity potential within and outside of the Natura 2000 areas. Additionally, a Spearman's Rank correlation was conducted between the LifeWatch variables of the clusters and their biodiversity potential. The correlation results will aid in assessing the type of ecotopes that were calculated to have a higher or lower wasp biodiversity potential.

5.4.3 Application of Wasp Biodiversity Potential for Ecosystem Health Assessment

Furthermore, the biodiversity potential of the selected clusters was calculated by intersecting the clusters and quality-controlled GBIF points and calculating the unique number of wasp species per cluster. This result could then be used as a baseline for ecosystem health, and its functionality for assessing ecosystem health was demonstrated using a relatively compact and

small cluster. The number of unique species per ecotope polygon in that cluster was computed through an intersect operation. Then, the number of unique species per ecotope polygon was compared to the biodiversity potential of its cluster, thus indicating the health status of the different parts of the ecosystem cluster.

In addition to the application on a cluster, a health assessment was also conducted for five urban centres in Germany. The five largest cities by population in Germany (Hamburg, Cologne, Frankfurt, Berlin, Munich) were selected to compare their wasp biodiversity potential and conduct a health assessment as described above. The cities were extracted from an official German administrative WFS dataset (https://sgx.geodatenzentrum.de/wfs_vg5000_0101). Both the wasp biodiversity potential within the city clusters, as well as the ratio between ecotope species count and cluster biodiversity potential, were visually and quantitatively compared to the Natura 2000 areas within the cities. Both health assessments were computed for demonstration purposes, as the EUNIS dataset does not provide enough data points to apply this method for a health assessment with confidence.

6. Results

In the following, the data before and after quality control, as well as intermediate and final results will be presented in detail. Firstly, the clustering results will be compared visually and in regard to geographic cluster size and compactness. Then, the validation analysis will be showcased with the intrinsic and extrinsic metrics. The effect of the quality control on the EUNIS data for the extrinsic validation will additionally be presented. Then, the quality controlled GBIF wasp dataset will be presented and compared to the original dataset. The wasp biodiversity potential will then be shown and explored in detail. Its dependency on total area covered by clusters or GBIF sample size will be assessed. Also, it will be compared to Natura 2000 areas. Finally, the application of the biodiversity potential will be demonstrated for two use cases.

6.1 Clustering Results

6.1.1 K-means Clusters

The K-means clustering results yielded clusters that, while partially visually cohesive, seem to be mostly very fragmented spatially (Fig. 14). The coastline in the north of Germany was mainly grouped into three coastal ecosystem clusters, with the turquoise-coloured cluster being consistently the shallowest, closest to land. Interestingly, the large Lake Constance in the south of Germany appears to have been assigned the same clusters. Additionally, the area in the very northwest of Germany (Schleswig-Holstein) appears to distinguish well between three major clusters that follow the major biogeographical regions of Schleswig-Holstein marshes and North Sea islands, Schleswig-Holstein Geest and Schleswig-Holstein Uplands. The three

clusters are interspersed with other clusters to varying degrees. The northeast of Germany appears to be mainly assigned to three clusters (orange, light orange and purple) interspersed among each other and by several smaller clusters. The clusters in the central area of Germany look visually more compact and closer together. Nonetheless, many of them are still fragmented and interspersed with other clusters. The central-eastern and central-western regions appear to be mainly dominated by clusters with small total areas covered or ones that are very fragmented. In contrast, the southern part of Germany, particularly around the foothills of the Alps, is dominated by clusters that seem the more compact and contiguous, with a minimum of 7 clusters that can be clearly distinguished visually. Some of those clusters correspond notably well to the spatial borders of e.g. the Black Forest, the Swabian Alb, and the Upper Palatine-Bavarian Forest.

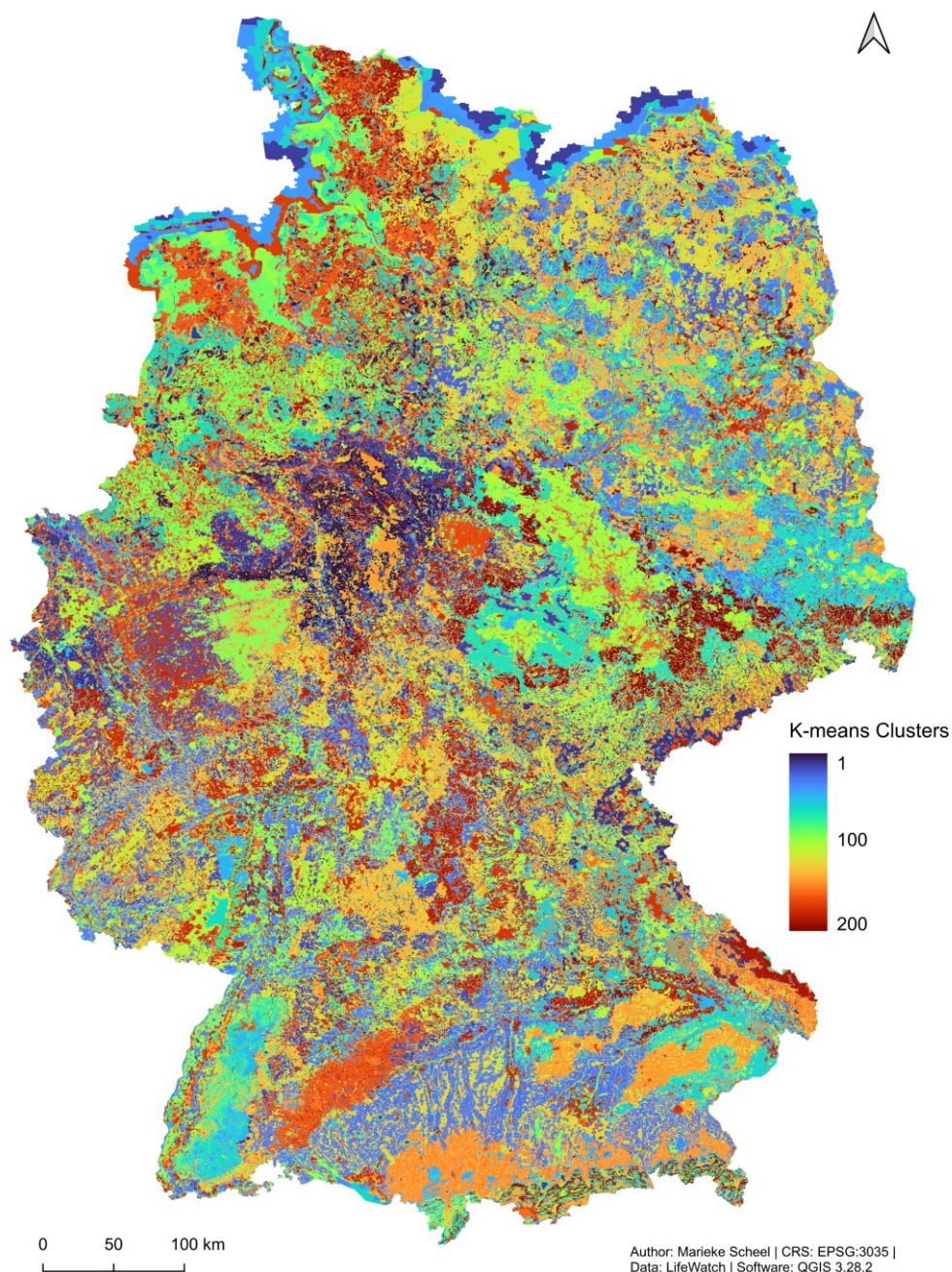


Fig. 14. The IDs of the clusters created using the K-means algorithm are plotted spatially.

Most of the clusters created using the K-means algorithm are relatively small in their total area covered, with areas below 500 km² being the most common (Fig. 15, Fig. A2). The frequency decreases slowly from a high frequency of clusters with a small total area covered (highest frequency around 700 km²) to a low frequency of around less than 1 cluster per bin at a total area covered of ca. 5000 km². Generally, the curve is positively skewed with a large dispersion. Due to the distribution being very spread out, the maximum frequency per bin is relatively low, with around 23 clusters per bin. Clusters with total area covered larger than 6000 km² appear to be the exception, with only about 10 occurrences. Also, there are no clusters with a total area covered larger than ca. 9000 km², which is about half of the size of the German Federal State of Saxony-Anhalt.

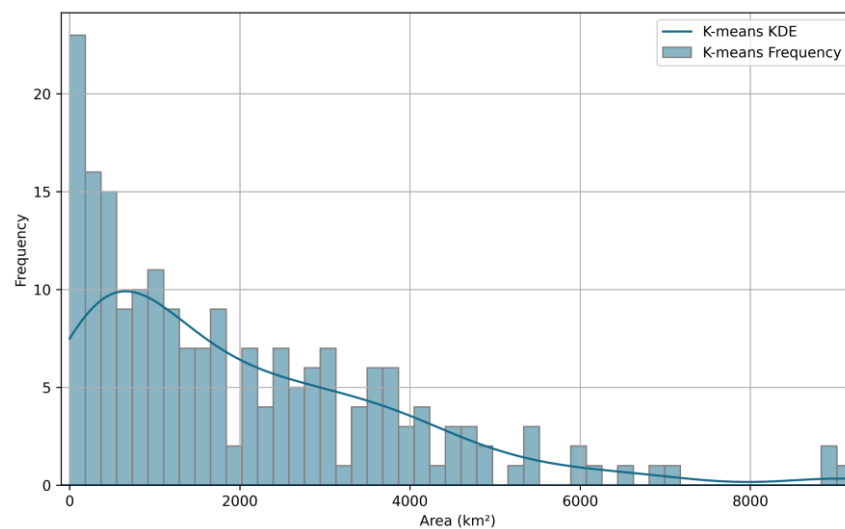


Fig. 15. Cluster area frequency of the clusters created using the K-means algorithm. The bin size is 50. The blue line corresponds to the Kernel Density Estimate curve.

6.1.2 Genie Clusters

The clusters that were created using the Genie algorithm look generally more contiguous than those that were created using the K-means algorithm (Fig. 14, Fig. 16). Additionally, the area of Germany seems to be dominated by several clusters with a very large total area covered, e.g. in middle blue in northern Germany and in green-turquoise in the east- and central-south. However, similar to the K-means clusters, one can observe several clusters that distinctly outline certain regions. Most of the coastal areas were, for instance, assigned to one cluster (green), with only some of the shallower areas being assigned to one other cluster (blue). Interestingly, Lake Constance was again assigned to the same cluster as the shallow coastal areas, as was, in this case, the river Elbe that flows into Hamburg. Also, the Genie clustering appears to distinguish between the different biogeographical regions along the North and the Baltic Sea. Albeit the border between those regions is less well-defined than in the K-means clusters (Fig. 14, Fig. 16). On the other hand, the outlines of other biogeographical regions are more discernible in the Genie clusters. For example, the Erz Mountains and “Oberlausitz” (Upper Lusatia) in Saxony in orange or the Black Forest in the southwest. Although the outlines of the latter are visible, the region appears to be assigned to several clusters. The same applies

to the Upper Rhine Plain in the southwest, which outline is clearly visible but assigned to several clusters. Also, the Harz Mountains and Thuringian-Franconian Highlands are well-defined in central Germany. Interestingly, the region south of the river Danube is well-distinguished from the region north of it. Also, the foothills of the Alps and the Alps in the very south are more well-distinguished than in the K-means clusters.

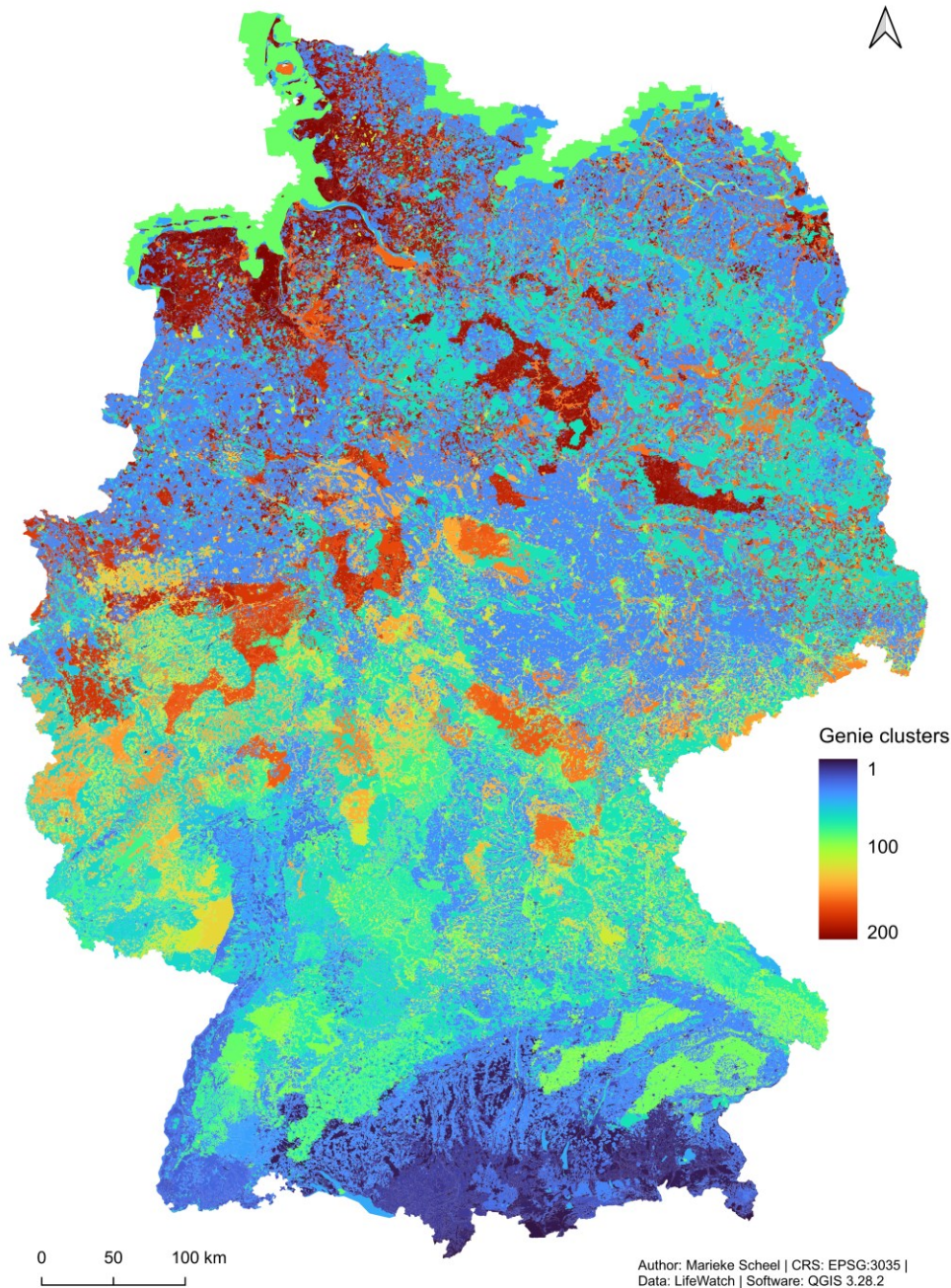


Fig. 16. The IDs of the clusters created using the Genie algorithm are plotted spatially.

The total area covered of the clusters created using the Genie algorithm is less widely distributed with the majority of the clusters having a total area covered of less than 2000 km² (Fig. 17, Fig. A3). The frequency peaks around a total area covered of 700 km² with a maximum number of 50 clusters per bin. The curve is positively skewed, with cluster with total areas covered beyond 4000 km² appearing to be outliers. About 10 clusters have a total area covered

larger than 4000 km², with the three largest clusters having a total area covered around 30,000 km² and 100,000 km², respectively. While the total area covered by the second and third largest clusters each already roughly corresponds to the size of the third largest Federal State in Germany (Baden-Württemberg), the largest cluster incorporates about a third of the total area of Germany.

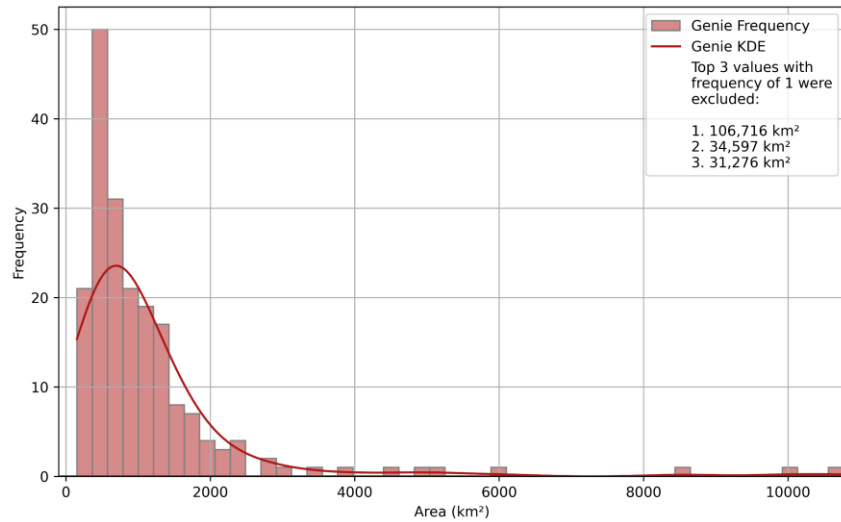


Fig. 17. Cluster area frequency of the clusters created using the Genie algorithm. The bin size is 50. The red line corresponds to the Kernel Density Estimate curve.

6.1.3 GMM Clusters

The clusters created by the GMM algorithm look more fragmented and interspersed than the clusters created by the other two algorithms (Fig. 18). In most regions, the clusters appear visually to have a very small total area covered and interspersed with each other without forming recognisable outlines around biogeographical regions. However, some regions seem distinguishable. The Upper Rhine Plain can be visually distinguished from the surrounding areas, albeit consisting of several clusters. Additionally, the coastline is visibly distinguishable from the land but is made up of more than 5 different clusters. The outline of Lake Constance is similarly visible; however, the lake is also assigned to several clusters. Interestingly, parts of the Rhine River, as well as the path of the Danube River and parts of its tributaries, are very visible.

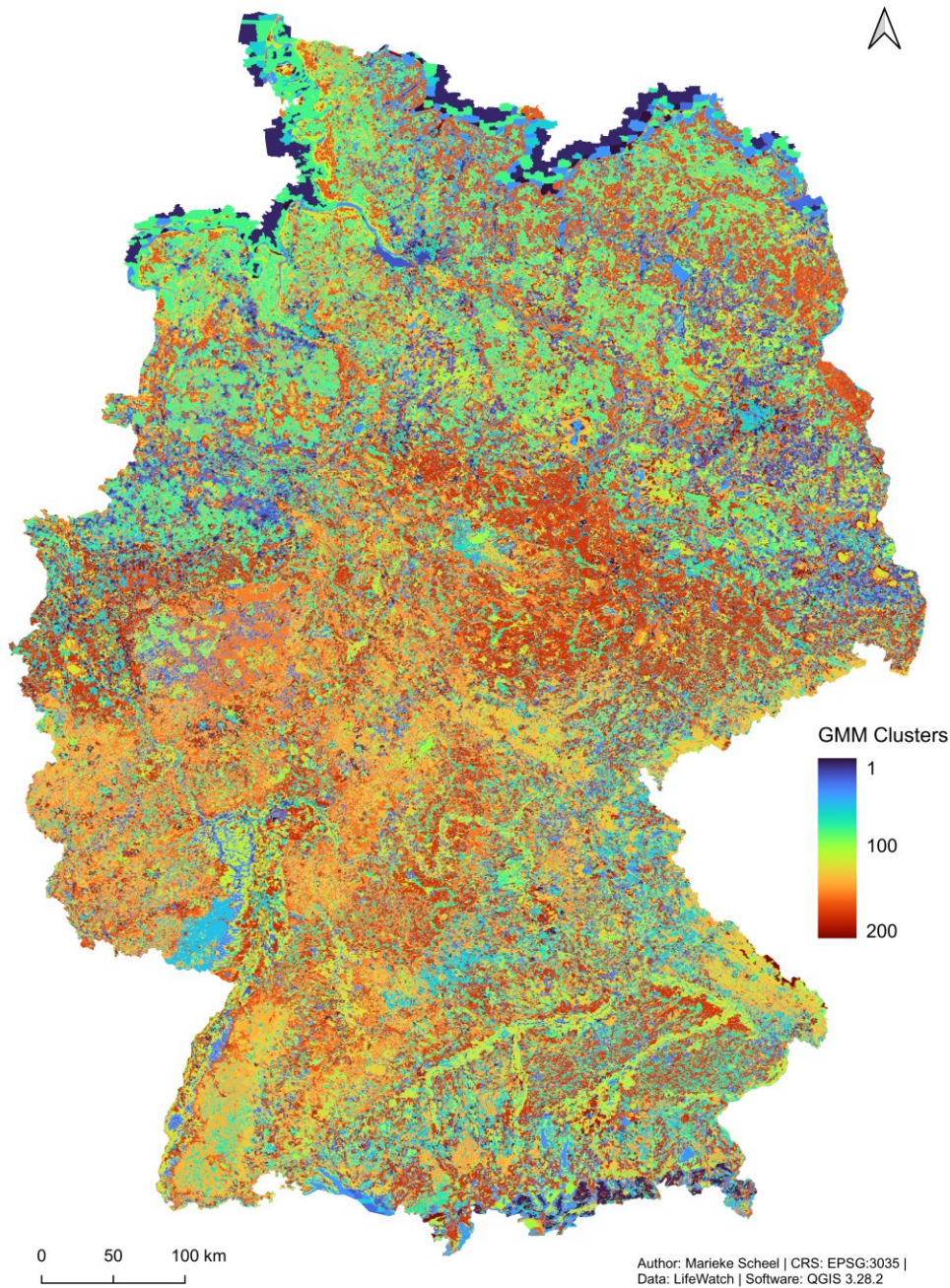


Fig. 18. The IDs of the clusters created using the GMM algorithm are plotted spatially.

The area of the clusters created by the GMM algorithm is again positively skewed with a relatively widely spread distribution (Fig. 19, Fig. A4). The majority of the clusters are relatively small, with the distribution peaking at an area around 600 km². The maximum frequency per bin is similar to Genie, with around 45 clusters per bin. After the peak in cluster area at ca. 700 km², the distribution decreases steadily until an area of ca. 3000 km²; then it plateaus with a stable frequency of ca. 1.5 clusters per bin until a size of ca. 6000 km². Beyond that size, the clusters visually seem to be outliers. About 10 clusters are larger than 6000 km², with three clusters being of a size greater than 10,000 km² and the largest cluster having an area of 30,000 km². Thus, the latter makes up about the area of Germany's third largest Federal State, Baden-Württemberg.

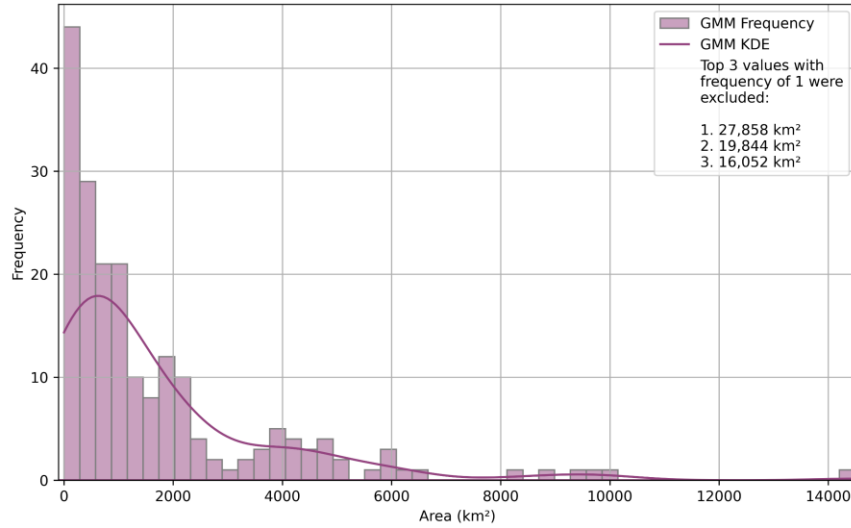


Fig. 19. Cluster area frequency of the clusters created using the GMM algorithm. The bin size is 50. The purple line corresponds to the Kernel Density Estimate curve.

6.1.3.1 Summary statistics of the GMM training data

The characteristics of the subset on which the GMM was trained correspond largely to the characteristics of the total dataset (Table 4). While the mean of the original dataset is roughly 0, the mean of the training dataset is only slightly elevated, with a mean of around 0.0002. Also, the variance of both datasets is ca. 1. The range of the training dataset is, additionally, slightly smaller, around 385, compared to the range of 390 of the original dataset. The minima of the dataset do not show large differences with values around -23, while the maxima are somewhat differing, with a value of 362 in the training dataset and a value of ca. 377 in the original dataset.

Table 4. Summary statistics of the quality-controlled LifeWatch dataset (original) and of the quality-controlled LifeWatch dataset used for training of the GMM algorithm (training).

	Training dataset	Original dataset
Mean	0.0002	-2.9E-17
Variance	0.9967881	1.0000012
Range	385.5	390.4
Min	-23.5	-23.6
Max	362.0	366.8

Furthermore, the KS test failed to reject the H_0 ($\alpha=0.05$) that hypothesised that the training data follows the same distribution as the original dataset, with an average $p=0.12$.

6.1.4 Comparison of Cluster Algorithms

Overall, the clusters created by the three different algorithms show notable differences in distribution of total area covered (Fig. 20). The K-means and GMM cluster area distributions appear to be the most similar. The IQR of both is of comparable size, with the median area of

K-means being slightly larger with an area of ca. 1500 km² compared to the GMM cluster area median of ca. 1000 km². However, the lower quartile q1 of the clusters created using GMM is bigger than the q1 of the clusters created using K-means. While the former extends to areas smaller than 0.1 km², the q1 of the latter extends only to an area of ca. 0.2 km². Additionally, the area of the GMM clusters has notably more outliers than that of the K-means clusters, with several clusters larger than 10,000 and 20,000 km². In contrast, the area of the Genie clusters has a markedly smaller spread, centring tightly around 1000 km². Also, the lower and upper quartiles of the Genie cluster area distribution are relatively small compared to that of the cluster area distribution of the other two algorithms. However, the Genie cluster area distribution has the largest number and the largest size of outliers, with areas up to 100,000 km².

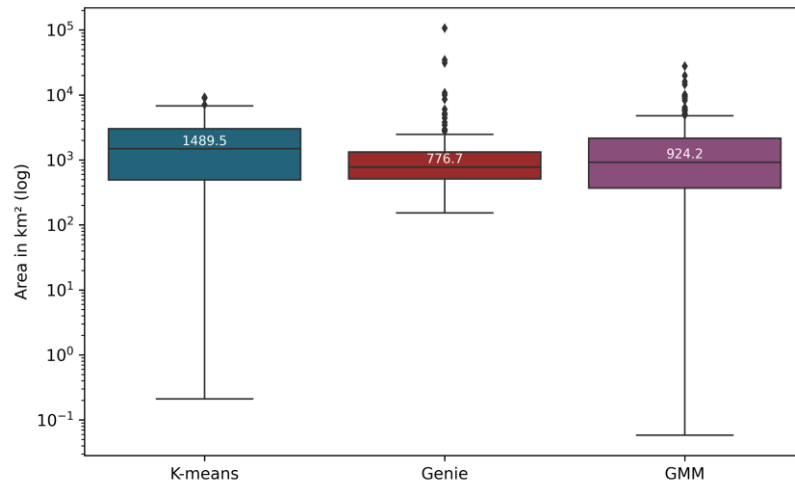


Fig. 20. Distribution of the total area covered of the clusters created with the K-means, Genie and GMM algorithms.

The geographic distance within the clusters generated by each algorithm can be compared with a measure of the furthest distance between two points that are members of a given cluster, i.e. the geographic intra-cluster distance (Fig. 21). Thus, it can be assessed whether an algorithm created rather local or regional clusters. Notably, the distance within clusters is the smallest for the Genie algorithm, with an IQR between ca. 300 and 600 km and a median of around 450 km. For reference, that median geographic distance roughly corresponds to half the north-south extension of Germany. The clusters created with K-means display a larger average geographic intra-cluster distance with an IQR between ca. 500 and 700 km and a median distance of ca. 600 km. Moreover, it includes several clusters with geographic distances smaller than 200 km within. The clusters created using GMM have the largest spread in terms of internal geographic distance distribution. Correspondingly, the IQR is very large between ca. 400 km and 800 km intra-cluster distance. Also, the median geographic distance is the largest among the three algorithms, at about 650 km. For all clustering algorithms, the largest geographic intra-cluster distance is about the largest distance possible in Germany (around 900 km).

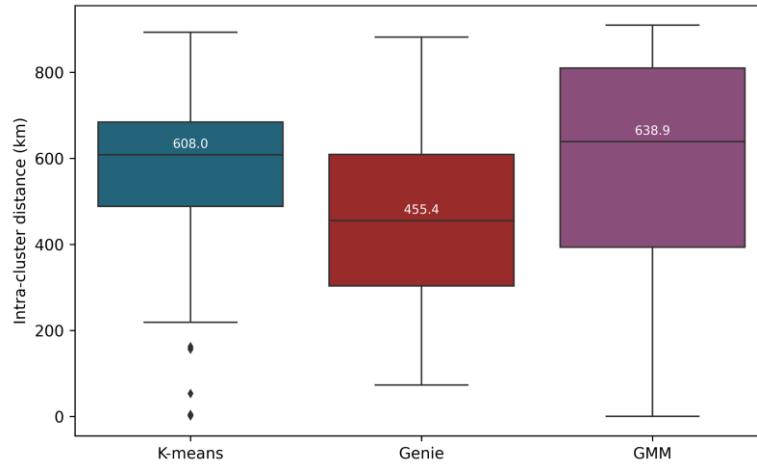


Fig. 21. Comparison of the distribution of the intra-cluster distances of the clusters created with the K-means, Genie and GMM algorithms.

6.2 Clustering Validation

6.2.1 Intrinsic Metrics

The intrinsic metrics reflect the cluster quality regarding the intra-cluster distances being minimised and the inter-cluster distances being maximised. As shown in Fig. 22, the clusters created using K-means are receiving the best score in that regard. The Silhouette score gives scores between -1 and 1. The K-means clusters receive a score of 0.07, which is only marginally above a score of 0. A score of 0 indicates no differences in the distance between clusters. Similarly, the GMM and Genie clusters receive a slightly negative score of -0.05 and -0.07, respectively. The DBI indicates greater cluster quality for smaller values. The K-means clusters receive a notably smaller score than the Genie and GMM clusters, with a value of around 2 compared to around 4. Since a greater CHI represents well-distinguished clusters, K-means is indicated here to have the clusters with the largest inter-cluster distance and the smallest intra-cluster distance. In contrast to the K-means CHI score of ca. 10,300, the Genie clusters receive only a CHI score of ca. 3500. The GMM clusters are given a slightly higher score of 4,500.

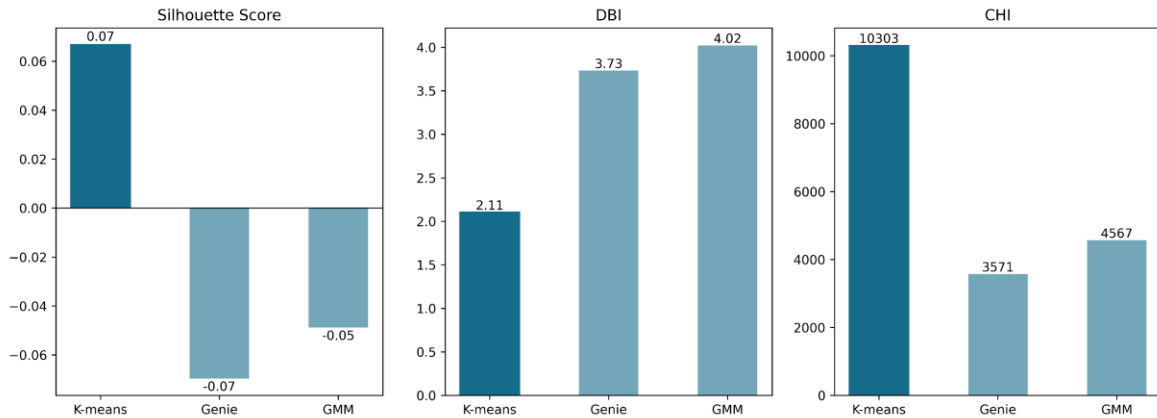


Fig. 22. Silhouette score, Davies-Bouldin Index and Calinski-Harabasz Index plotted for the clusters created using the K-means, Genie and GMM algorithms.

6.2.2 Extrinsic Metric

6.2.2.1 Quality-controlled EUNIS Habitat Data

While the EUNIS points are distributed relatively well all over Germany, there are areas with notably fewer points and areas of higher concentration (Fig. 23). After the quality control, the distribution of sample points does not change visually, with a higher concentration of points in the west of Germany and the northeast. While the dataset initially included ca. 120,000 samples and 111 classes, the quality control reduces it to about 35,000 samples with 68 classes (Fig. A7).

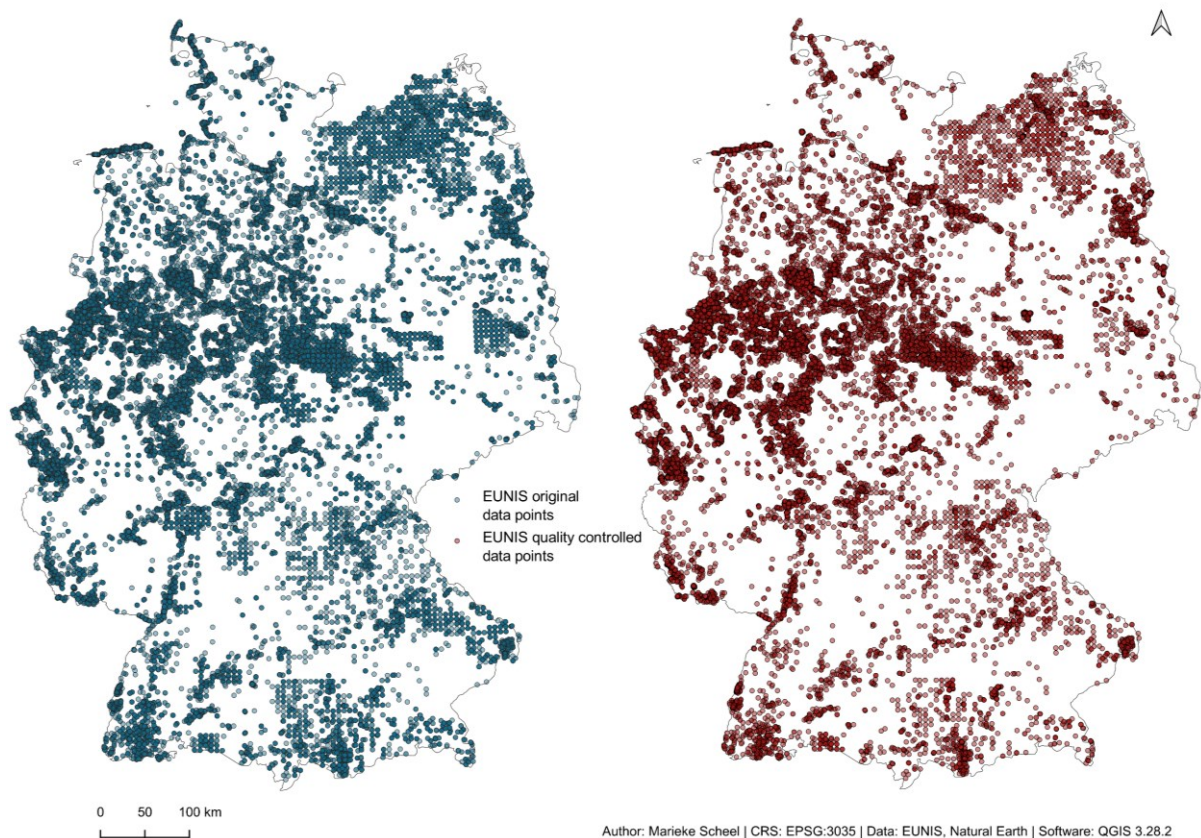


Fig. 23. Spatial distribution of the original EUNIS data points (left), and spatial distribution of the EUNIS data points after quality control (right).

The majority of the data samples are part of the habitat group of grasslands both before and after quality control (Fig. 24, Fig. A5, Fig. A6). After the latter, about half of the data samples were members of this class. Another 8,000 data samples are made up of members of the forest class, while marine benthic and coastal habitats, contained each less than 500 samples. Wetlands, vegetated man-made habitats and heathlands made up ca. 3,600, 2,700 and 1,300 samples after quality control, respectively. The size distribution between the classes remains similar before and after quality control. However, particularly inland habitats and coastal habitats consist of fewer point samples afterwards.

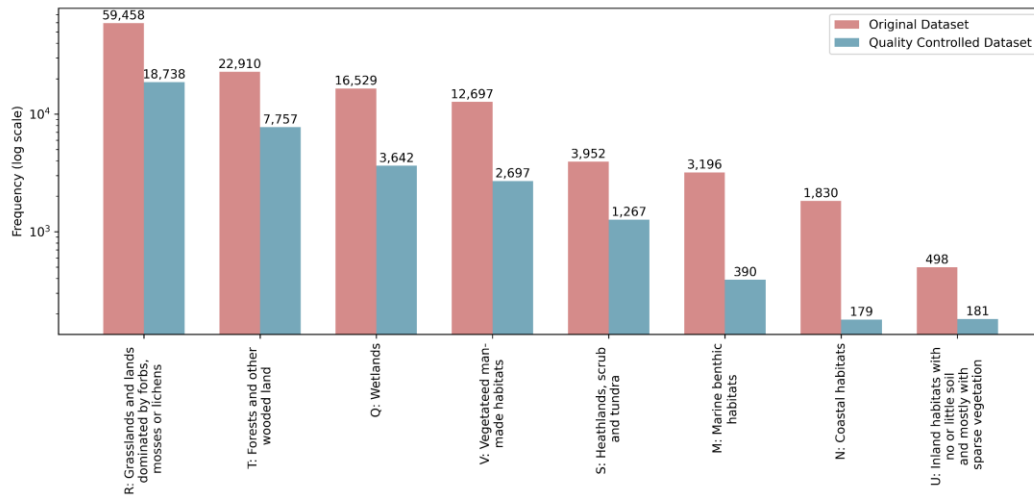


Fig. 24. Frequency of the original (red) and quality-controlled (blue) habitat classes of the EUNIS data in Germany. The EUNIS habitat classes were aggregated to level 1.

The Gini coefficient can reflect the equality of point distribution in space, which indicates an equal spatial distribution as 0 and the opposite as 1. The initial dataset received a coefficient of 0.843, while the quality-controlled dataset received a very similar one of 0.846.

6.2.2.2 Assessment of Cluster Purity with EUNIS Habitat Data

The specificity of the clusters can be assessed for the different algorithms by calculating a purity score that indicates the proportion of the majority EUNIS habitat class in each cluster. The average purity for Genie clusters was the highest, at 0.34. In comparison, the average purity for GMM and K-means clusters was lower, with a value of 0.29. As shown in Fig. 25, the IQR of the proportion of the majority class is additionally the highest in Genie clusters, with purity scores between 0.2 and 0.45 and a median of 0.3. Moreover, the upper quartile extends notably higher than the other two metrics, up to around 0.75. The GMM and K-means purity score IQR only extends between ca. 0.2 and 0.35, with a median of 0.26. Both have an upper quartile reaching up to purity scores of ca. 0.6. Also, they have visibly more outliers with larger purity scores compared to the Genie clusters (Fig. 26).

The Kruskal-Wallis test was conducted to compare the distribution of purity scores across the K-means, Genie and GMM clusters. The test results led to the rejection of the null hypothesis, with a significance level of $\alpha=0.05$, at $H=8.92$ and $p=0.02$. Hence, significant differences in the purity scores between the clustering methods were found. The post-hoc Dunn's test was thus conducted and yielded the following p-values: Genie vs. GMM: $p=0.012$, Genie vs. K-means: $p=0.106$ and K-means vs. GMM: $p=1.0$.

These findings suggest that there is a statistically significant difference between the purity score distributions of GMM and Genie ($p<0.05$). However, no significant difference could be found between the purity score distributions of GMM and K-means or Genie and K-means.

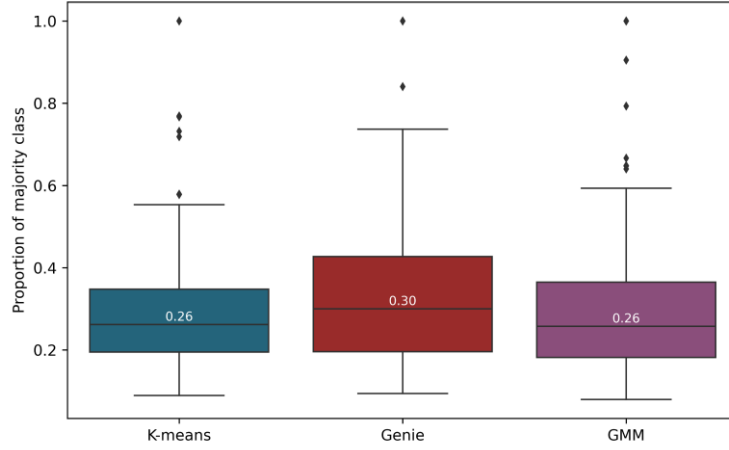


Fig. 25. Distribution of the proportion of the majority EUNIS class within the clusters created using K-means, Genie and GMM. The number in white in the boxplots corresponds to the median of the distributions.

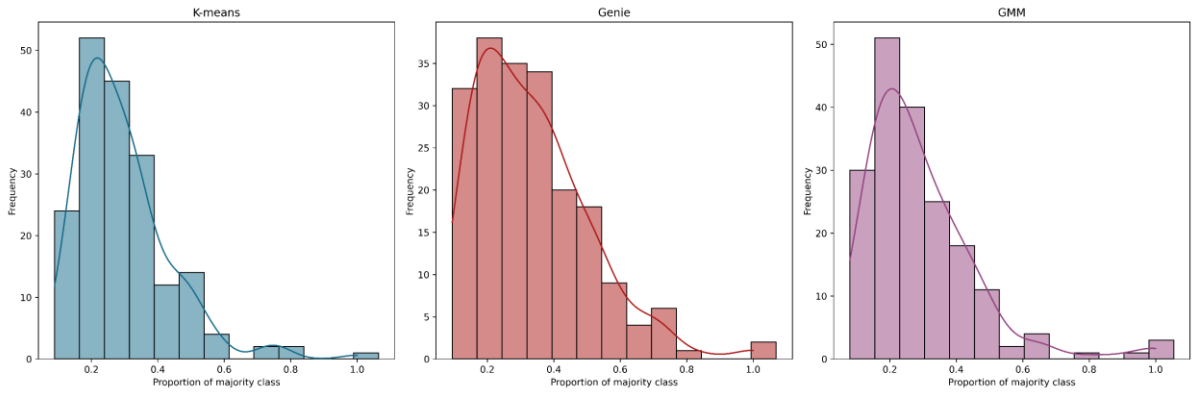


Fig. 26. Frequency of the proportion of the majority of EUNIS classes within the clusters created using the K-means (blue), Genie (red) and GMM (purple) algorithms. The bin size is 0.075. The lines correspond to the Kernel Density Estimate curve.

Despite the lack of significant differences between the purity score distributions of K-means and Genie, the Genie clusters were selected for further analysis. This decision was based on the higher average and median purity score of the Genie clusters (which was a significant difference compared to GMM).

6.2.2.3 Cluster Labelling with EUNIS

The Genie clusters were labelled using two methods. With method a) each cluster was labelled by assigning the majority class of the EUNIS samples within each cluster. In contrast, with method b), each cluster was labelled by assigning the weighted majority class of the EUNIS samples within each cluster (see section 5.3.2).

The labelling with both methods appears to have yielded many inconsistent class assignments. While method a) led to large and more homogeneous ecosystem classes, method b) yielded smaller and more diverse ecosystems. Additionally, vegetated man-made classes, forests and inland habitats are not represented at all. Moreover, as artificial classes and inland waters are missing from the EUNIS dataset, correct labelling of urban areas and aquatic non-coastal waters

could not be expected (see section 4.2.2). Generally, one could have expected coastal, marine and agricultural and lowland labels in the northern part of Germany. While the north-east has more wetlands and lakes, the south-east and south have more mountain ranges, forests, grasslands and scrubs (Fig. 5, Fig. 6, Fig. 7).

Method a) labelled marine benthic habitat classes just off the German coastline with a high consistency and accuracy (Fig. 27). However, only one marine benthic was assigned with method a), in contrast to five with method b). Additionally, part of the Alps in the south of Germany were labelled as Alpine grassland classes with method a) (Fig. 27, 4a). Also, the Harz lower mountain range in central Germany was assigned to alpine and subalpine ericoid heath, the small helophyte bed and temperate and boreal moist or wet oligotrophic grasslands (Fig. 27, 3a). Similarly, other lower mountain ranges were classified as alpine and subalpine ericoid heath. Hence, the labelling of mountainous areas appears to be relatively consistent with method a), despite some mountainous areas not being labelled as such.

Nonetheless, overall, the labelling with method a) appears to be quite inaccurate. Over 75% of Germany were assigned to wetland classes, which is surprising given that the majority of EUNIS samples were grassland samples (Fig. 24). The north of Germany was mainly labelled as calcareous quaking mire, as well as large parts of the south of Germany. The very industrial Ruhr district was nearly completely assigned to be wetland (Fig. 27, 1a). The city Berlin, in contrast, was nearly fully labelled as heathland classes (Fig. 27, 2a). Meanwhile, patches in Central Northern Germany were labelled as Atlantic and Baltic coastal dune forests and Atlantic, Baltic and Arctic sand beach. Particularly, the latter seems to be inaccurate. Moreover, a zone at the foothills of the Alps was assigned to a coastal habitat class (Fig. 27, 4a), further supporting that the coastal habitat labels seem to be inconsistently assigned.

Method b) appears to have yielded generally more heterogeneous results (so more in line with what could be expected given Germany's diversity). However, the labelling seems to be less consistent over the entire study area and highly inaccurate for some parts (Fig. 27). Generally, the labelling yields a more realistic north-south division than method a). However, the majority of Northern Germany was again labelled as wetland. In contrast, the majority of Southern Germany was labelled as grasslands. Also, the marine benthic labels were inconsistently assigned. For instance, the sea off the coast of Germany, was labelled as several marine benthic areas, also differentiating between the shallower zones in the Baltic Sea compared to those in the North Sea. However, this also led to parts of the islands being assigned a marine benthic label. Additionally, a large part of Eastern and Central Northern Germany were labelled as Marine benthic habitat. Similarly, coastal habitats were not assigned to coastal areas in Northern Germany but rather to areas in the Ruhr district and at the foothills of the Alps (Fig. 27, 1b and 4b). While more mountainous areas appear to be labelled as Alpine and subalpine (heath) by method b) compared to method a) (e.g. Fig. 27, 3b and central-western Germany), this result was inconsistent and did not apply to all mountainous areas (e.g. mountain range in south-east or foothills of Alps). The outlines of the cities in the Ruhr area were more visible for method b) (Fig. 27, 1b), and cities were generally more consistently assigned the same class labels in method b), i.e. heathland, scrub and tundra (e.g. Berlin in Fig. 27, 2b).

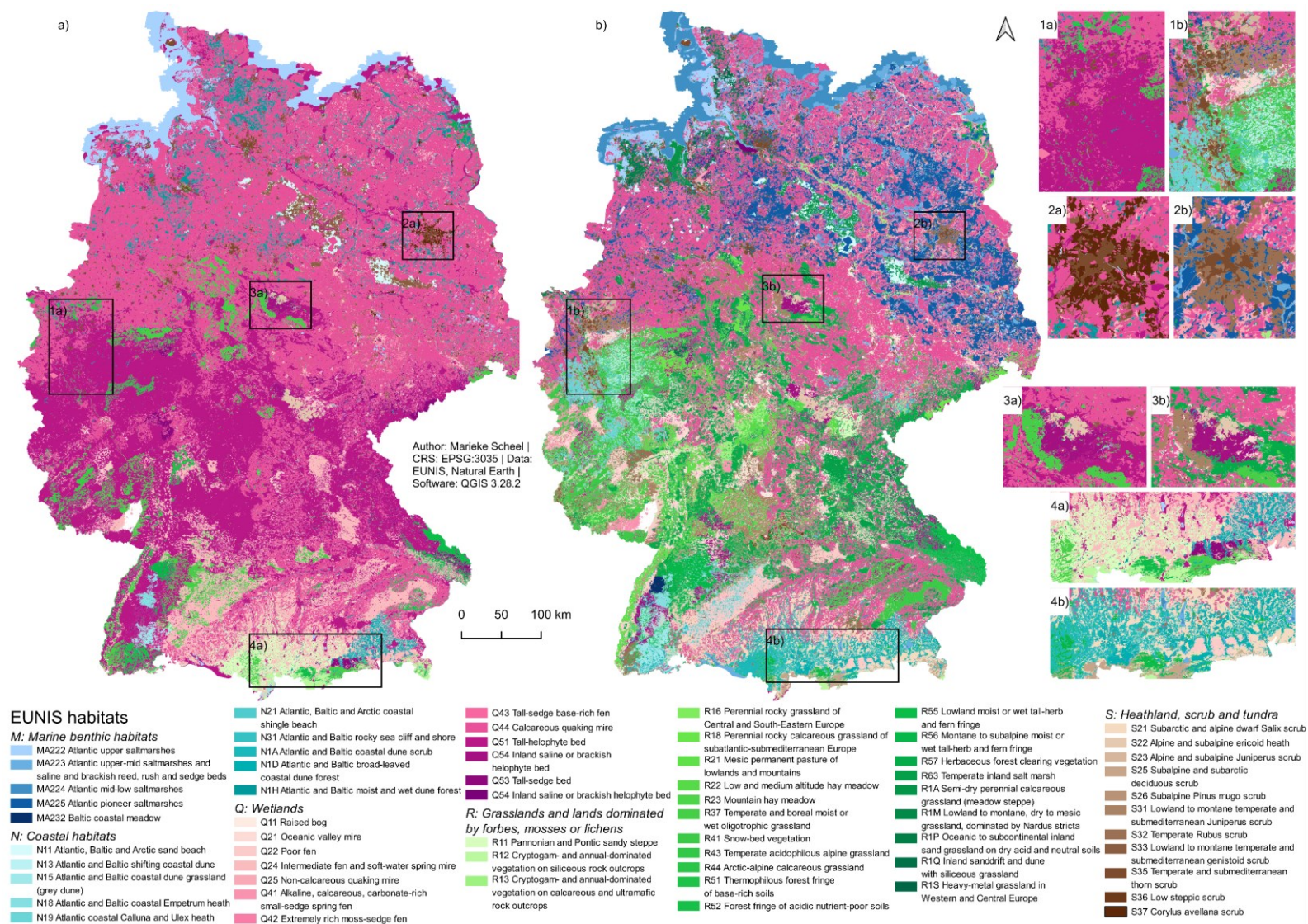


Fig. 27. Labelling of the Genie clusters using EUNIS habitat samples a) by the majority of the EUNIS samples within clusters and b) by the majority of the weighted EUNIS samples within clusters. To the right, certain regions are highlighted for comparison.

6.3 Calculation of Baseline for Ecosystem Health Assessment

6.3.1 Quality-Controlled GBIF Taxon Data

The relative proportion of wasp species samples has not changed substantially due to the quality control (Fig. 28). More than 60% of the samples are *Vespa crabro*, around 18% are *Vespula vulgaris* and ca. 10% are *Vespula germanica*. Another 5% are samples of *Vespa velutina*. The remaining species only comprise a marginal proportion of the data samples before and after the quality control. Nonetheless, the quality control reduced the total number of unique species from 46 to 36.

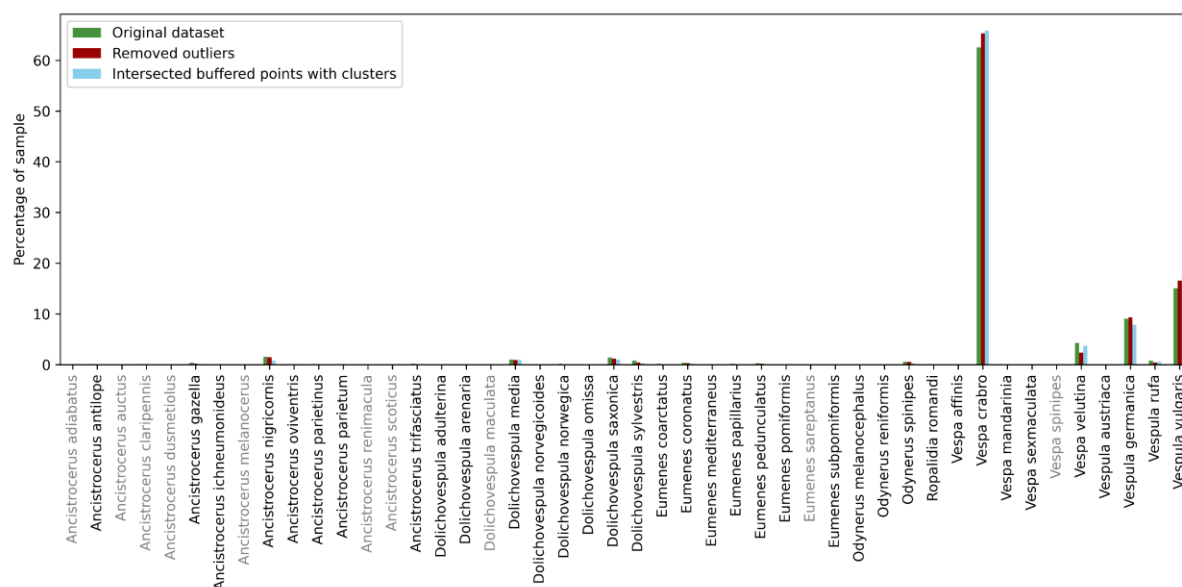


Fig. 28. Percentages of the different wasp species in the original GBIF dataset (green), the GBIF dataset with removed outliers (red) and the fully quality-controlled GBIF dataset (blue). Species names in light grey were removed by the quality control.

The spatial distribution of the GBIF samples appears to be relatively equal in space (Fig. 29). While there is a concentration of data points in the west of Germany and around major cities like Hamburg, Berlin, Munich and Stuttgart, especially the northeastern region north of Berlin seems to have more sparse data points. The quality control does not appear to change the distribution pattern. This is consistent with the results of the Gini coefficient assessing the equality of distribution in space. The Gini coefficient is 0.91 before and after the quality control of the data.

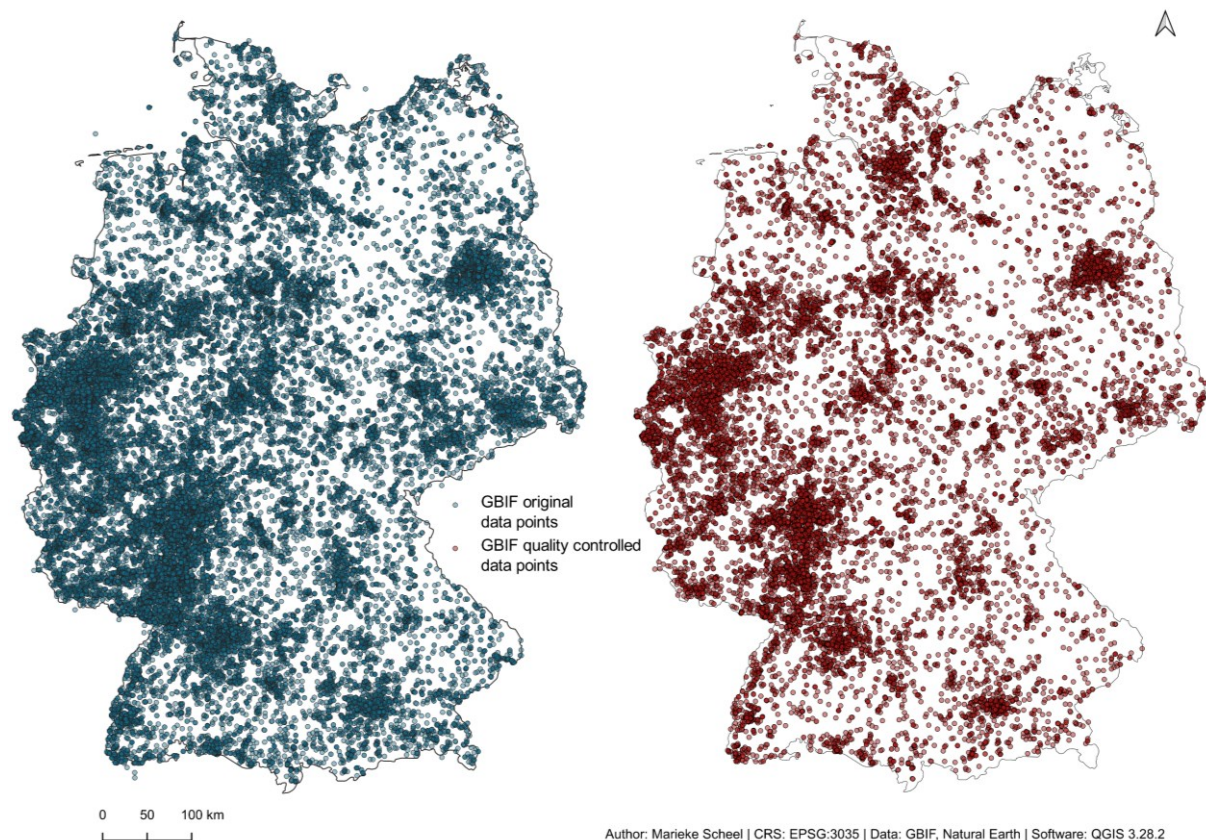


Fig. 29. Spatial distribution of the original GBIF data points (left) and spatial distribution of the GBIF data points after quality control (right).

6.3.2 Wasp Biodiversity Potential

6.3.2.1 Establishing Wasp Biodiversity Potential

The selected clusters with a sample size above the threshold of 100 GBIF samples (42 clusters) show very different potentials regarding species diversity (Fig. 30). Most clusters have a relatively low count of unique species. About 7 to 8 different wasp species in one ecosystem cluster occur the most frequently. The distribution curve of species diversity in the clusters is positively skewed with a large variance and a range between ca. 5 and 18 unique species per cluster.

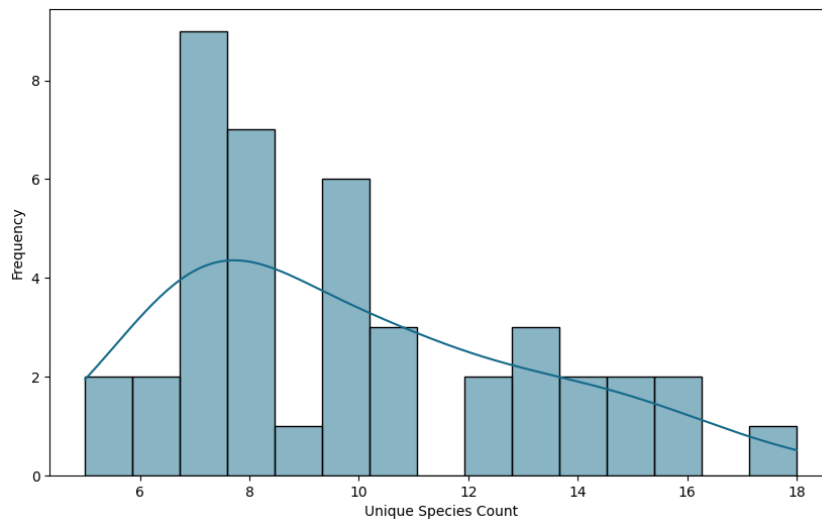


Fig. 30. Frequency of unique wasp species counts within the selected clusters. The bin size is 0.7. The blue line corresponds to the Kernel Density Estimate curve.

Regional differences in biodiversity potential could be found (Fig. 31). Firstly, the 42 clusters with a total species count above 100 encompass about two thirds of the total area of Germany (Fig. 31, Fig. A8). This provides an interesting insight, as none of the clusters with an area larger than 10,000 km² are included in the 42 selected clusters (Fig. 31, Fig. A3). Generally, the selected clusters cover more area in the northern and eastern parts of Germany (Fig. 31). In the northern part, most clusters have about 14 unique species (effectively the biodiversity carrying capacity of these clusters), with some clusters in the east having higher values of around 16 unique wasp species. Additionally, one or several clusters in the North appear to follow the rivers Elbe (flows through Hamburg) and Havel. Those clusters have fewer unique species than the surrounding clusters, with about 8 wasp species. Moreover, along the North Sea coastline of Germany, the cluster/clusters have a slightly lower number of unique species than the surrounding ecosystems with ca. 11 wasp species. The cluster in central-southern Germany has the highest count of unique species, with 18 species. Interestingly, this pattern also persisted in the clusters with ≤ 100 GBIF samples, where the clusters in Central Germany had the highest count of unique species with 12 species (Fig. A10, Fig. A11). At the foothills of the Alps, the majority of the ecosystem clusters have a species count of about 13 species. Additionally, the clusters in the Upper Rhine Plain have a notably lower number of unique wasp species with values between 5 and 10 species. Lastly, the Ruhr District region also has an outstandingly lower number of wasp species than the surrounding area, with numbers between 6 and 10 species. To distinguish between clusters beyond their number of unique species, the cluster IDs are additionally visualised in Fig. A8.

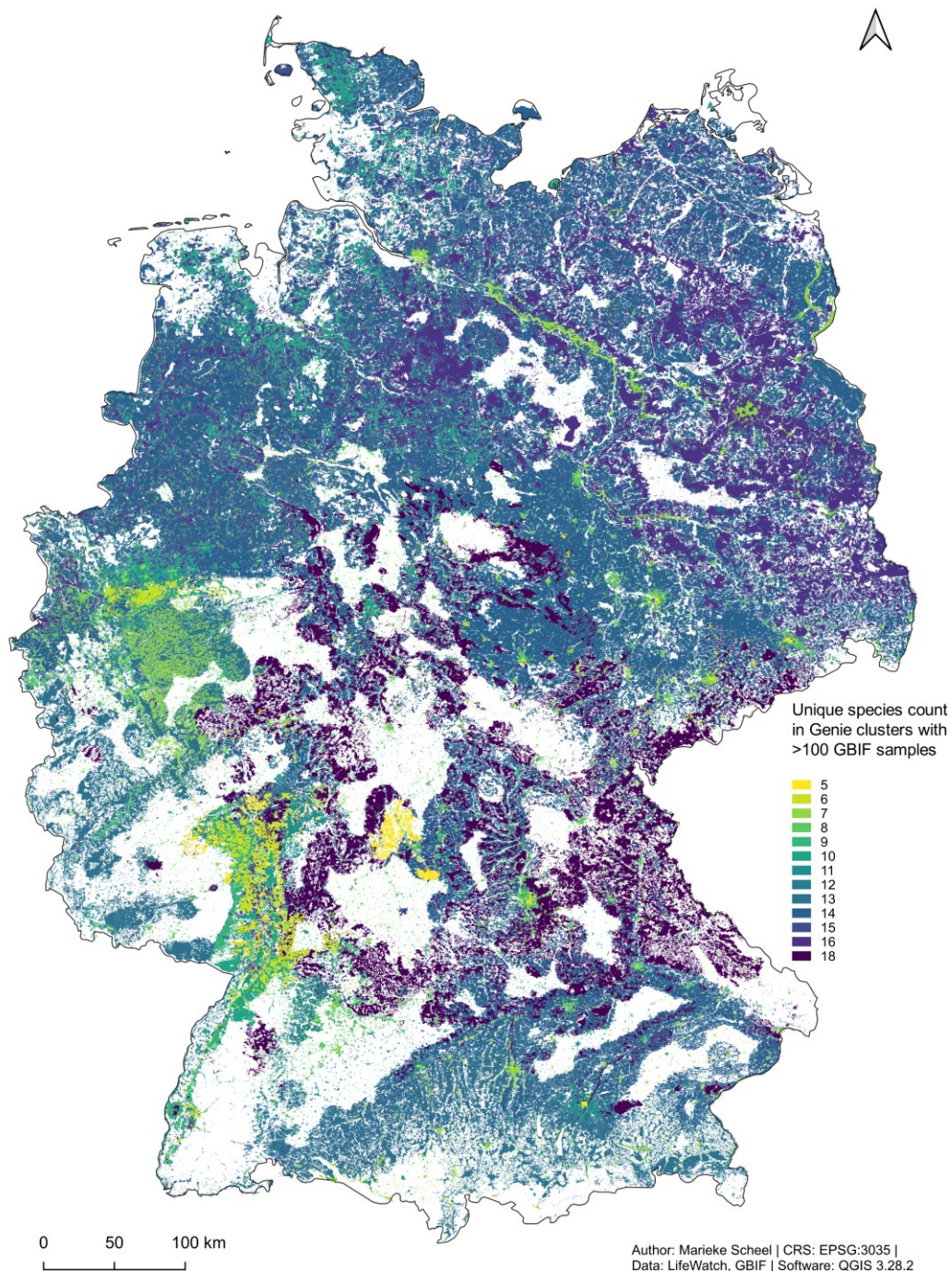


Fig. 31. Number of unique wasp species plotted for the 42 selected clusters, each with > 100 GBIF samples.

6.3.2.2 Validation and quality control of the Wasp Biodiversity Potential

Overall, the total number of wasp samples per cluster appears to be an important factor for the total number of unique species per cluster (Fig. 32). Six clusters with a total sample size around or below 250 have a count of unique species that is equal to or higher than 10 (i.e. clusters 3, 58, 59, 68, 187 and 189). However, most clusters with a high total species count also have a relatively high count of unique wasp species (e.g. clusters 21, 35, 60, 64, 66, 124, 152, 155, 187). Thus, 15 clusters with more than 250 samples had a unique species count that was larger or equal to 10.

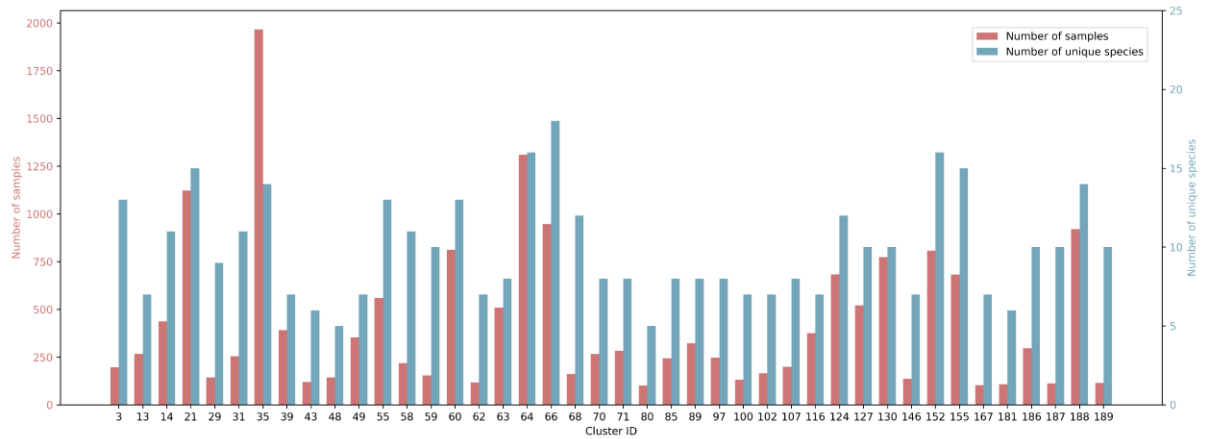


Fig. 32. Total number of samples (red) and number of unique GBIF wasp species (blue) per cluster.

Additionally, the number of unique species was shown to be significantly correlated with both the total area covered by a cluster and the number of GBIF samples within a cluster (Spearman's test, $p < 0.05$). As shown in Fig. 33, both variables are moderately to strongly positively correlated with the number of unique species.

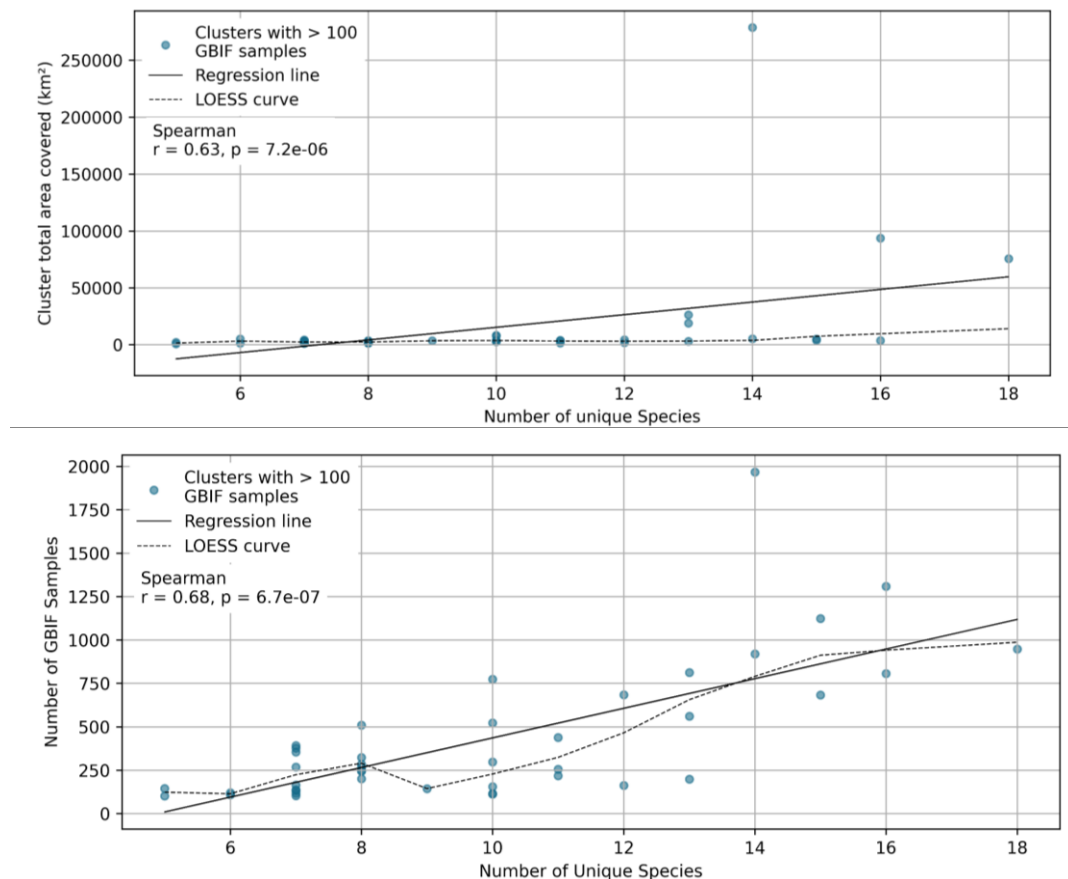


Fig. 33. Scatterplot for number of unique wasp species and the total area covered by clusters (in km²) with >100 GBIF samples (top) and scatterplot for number of unique wasp species and number of GBIF samples per cluster with >100 GBIF samples (bottom). The black line corresponds to a linear regression line and the dotted to a LOESS (locally estimated scatterplot smoothing) line. They are given for guidance. The Spearman tests indicate significant correlation ($p < 0.05$).

For comparison, the distribution of the area percentage per unique wasp species count was plotted for a) all of the Genie clusters with > 100 GBIF samples, b) the clusters within the Natura 2000 areas in Germany as well as c) the clusters not within the Natura 2000 areas in Germany (Fig. 34). For a) the area percentage for the areas with <13 unique wasp species is less than 5%. Areas with 13 unique wasp species take up about 10% of the total area, and areas with 14 species take up ca. 45%, which is the largest share for this group. Areas with 16 and 18 species comprise about 15% of the total area. For b) the area percentages are slightly more skewed towards a higher unique species count (Fig. 34). Areas with <13 unique wasp species similarly only take up ca. 5% of the total area each, while areas with 13 unique wasp species take up ca. 12%, with 14 species ca. 22%, with 16 species ca. 30% and with 18 species ca. 25%. Hence, most of the clusters within Natura 2000 areas have many unique wasp species. In contrast, for c) areas not within Natura 2000 areas, more than 80% of the areas have a unique species count of 11. Areas with a unique species count of 5 are about 6%, while the share of all other areas seems negligible with <5%. The total areas per unique wasp species count are additionally shown in Fig. A9.

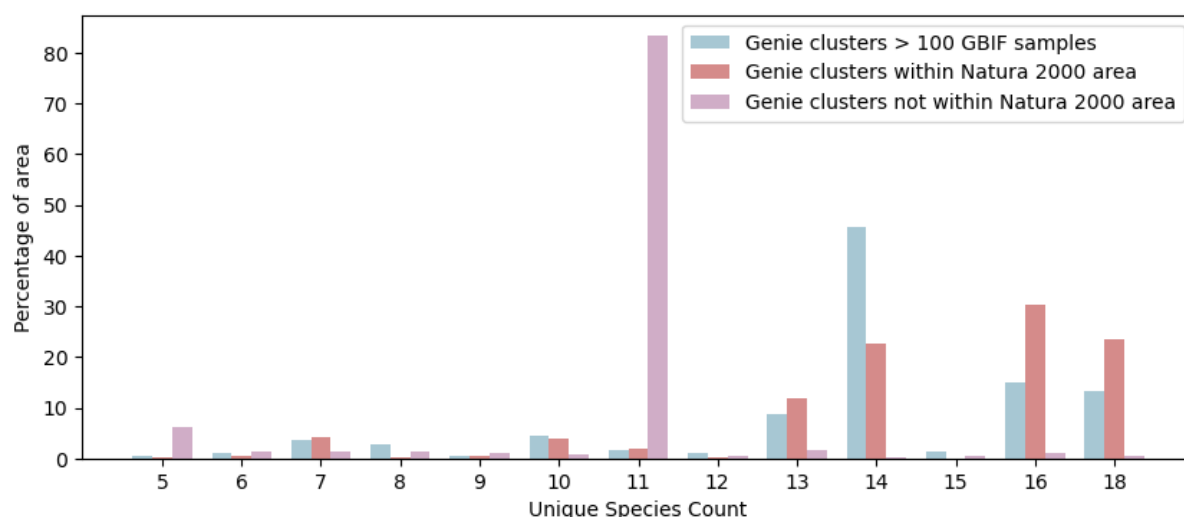


Fig. 34. Comparison of the percentage of the area by unique wasp species counts for the whole area of the Genie clusters with >100 GBIF samples (blue), for the Genie clusters within Natura 2000 areas (red) and for the Genie clusters not within Natura 2000 areas (purple).

The Spearman correlation coefficient indicates perfect negative or positive correlation as -1 and 1, respectively. Coefficients with values > |0.3| are considered to have a moderate and coefficients with values > |0.2| a weak correlation following the assessment by the Political Science Department at Quinnipiac University (Akoglu, 2018). Hence, it becomes visible that most of the unique wasp species have a moderately negative correlation with built-up areas as well as with “monospecific graminoid cover”, which is often found in gardens, leisure facilities and other intensive land uses (Fig. 35, Table A2). Additionally, the number of unique wasp species is moderately positively correlated with broadleaved forests and weakly positively correlated with coniferous trees. Similarly, the minimal distance to coniferous trees and sparse vegetation is weakly negatively correlated, and mean, minimum and maximum distances to sealed areas are moderately positively correlated to the number of unique wasp species.

Additionally, the amount of night light shows a moderate negative correlation. Interestingly, the mean annual temperature shows a moderate negative correlation, while the yearly mean variation in temperature shows a moderate positive correlation with the number of wasp species. Interestingly, some of the categorical variables also exhibit a weak to moderate correlation with the number of unique species.

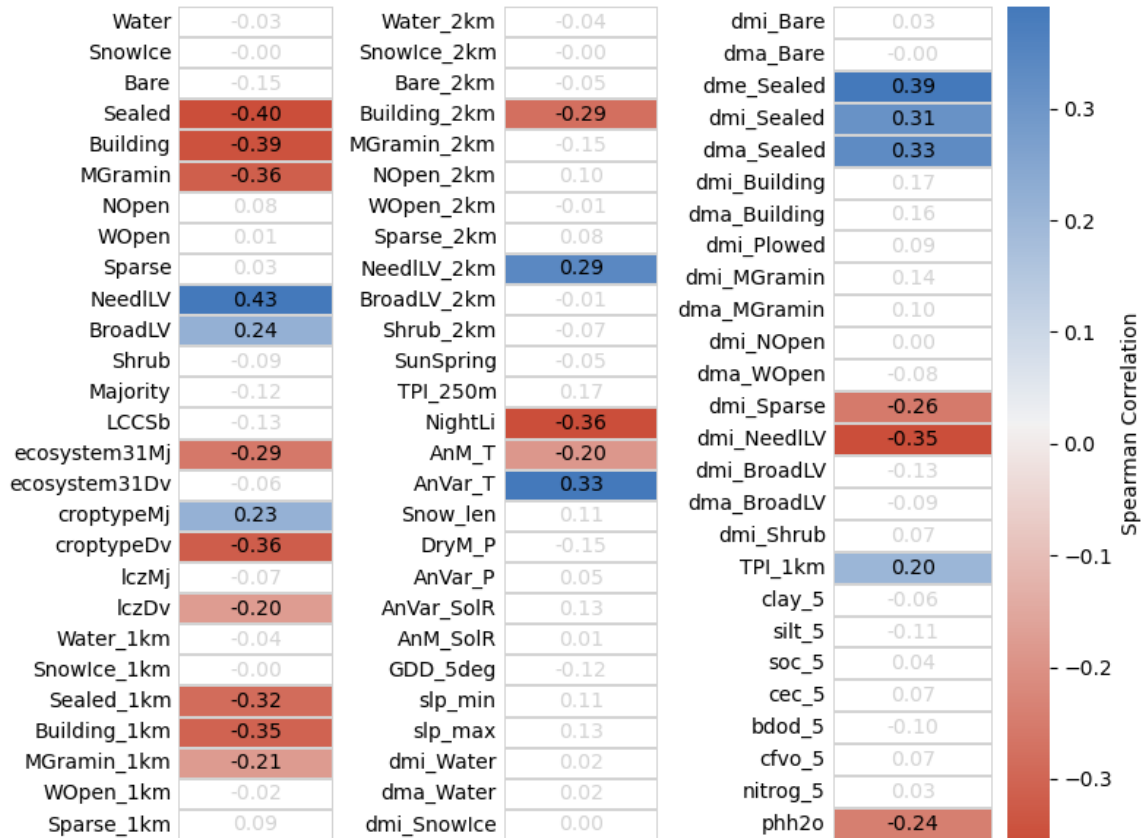


Fig. 35. Spearman correlation coefficients between the biophysical variables used for clustering and the unique wasp species count. Variables with correlation < 0.2 and > -0.2 are masked out in white.

6.3.3 Application of Wasp Biodiversity Potential as a Baseline

To compare the wasp biodiversity potential, the five largest cities by population were selected (Fig. 36). The average numbers of unique species per city (weighted by area) were:

- Frankfurt 17.3 species/m²
- Munich 16.0 species/m²
- Berlin 15.9 species/m²
- Hamburg 14.5 species/m²
- Cologne 13.0 species/m²

Besides Cologne, all cities exhibit the pattern of having lower wasp biodiversity potential within the city centres, with values around 5 species (Munich, Frankfurt) and 7 species (Hamburg, Berlin). The size of the city centres in Munich and Frankfurt is additionally smaller

than in Berlin and Hamburg. In Cologne, the city centre is not distinguishable from the surrounding areas, with values around 12 and 13 species most prevalent on the west and northeast of the river. In the very southeast of the city, a Natura 2000 area also has a higher wasp biodiversity potential with 13 species and small patches of 16 species. In Berlin, most of the city area outside of the centre has a relatively high biodiversity potential with 16 species. The Natura 2000 areas are within the same zones. The majority of the area around the city centre in Munich has a biodiversity potential of 11 species, with some small and non-connected patches having a higher potential of 18 species. The majority of those patches are within Natura 2000 areas. In Frankfurt, the area north of the river has mainly a biodiversity potential of 14 species. Most of the areas have a biodiversity potential of 18 species south of the river, with Natura 2000 areas being within those zones. Around the city centre in Hamburg, the most prevalent biodiversity potential is 14 species. The Natura 2000 areas in the city are mainly in areas with a potential of 16 species.

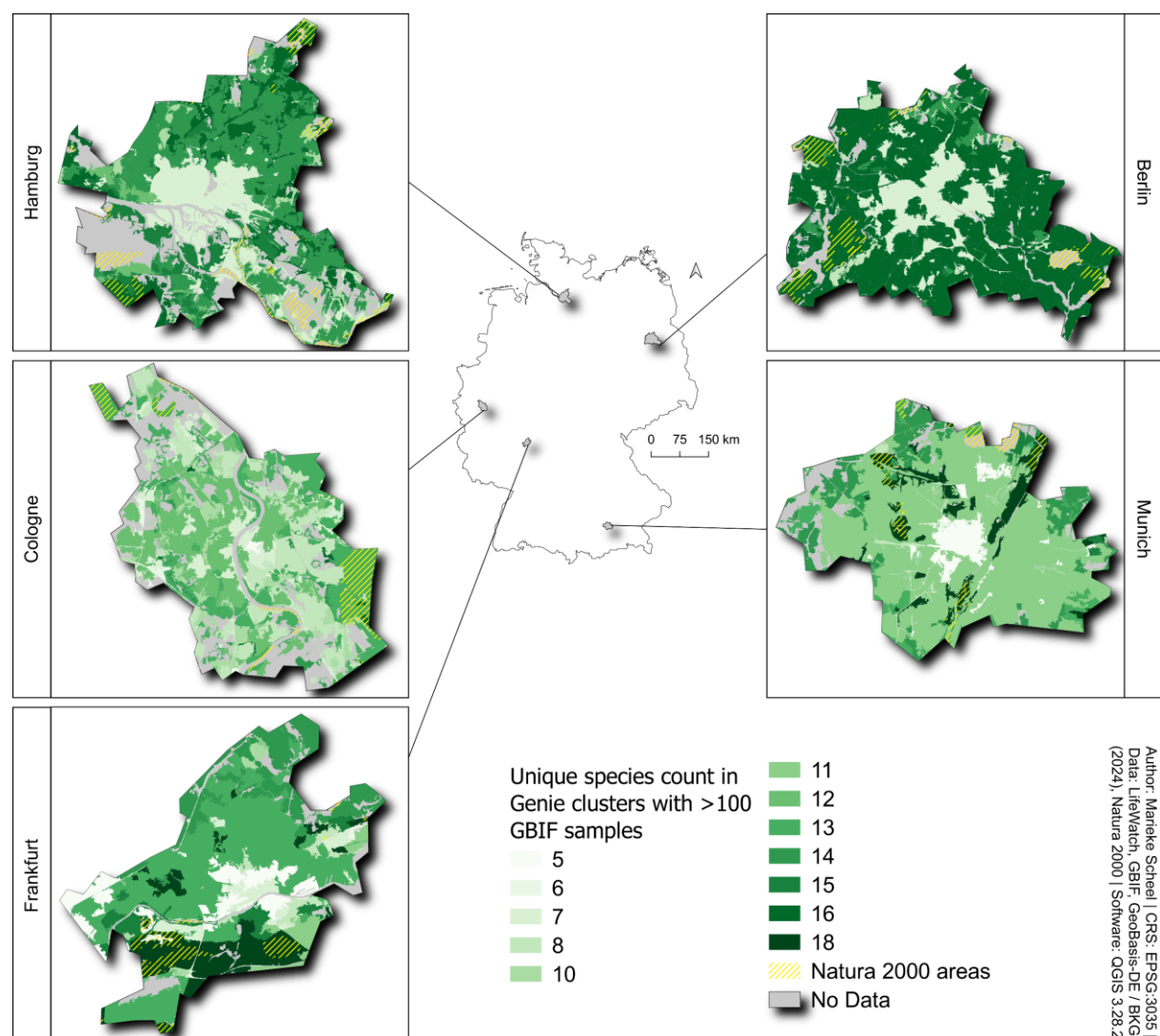


Fig. 36. Calculation of wasp biodiversity potential in Hamburg, Cologne, Frankfurt, Berlin and Munich, the five largest cities in Germany (by population). Natura 2000 areas within the city perimeters are highlighted in yellow, and grey areas indicate no data (≤ 100 GBIF samples).

The unique species count of areas intersecting with Natura 2000 zones is very high for all cities but Cologne (Fig. 37). In Frankfurt, more than 90% of the Natura 2000 zones were calculated to have a wasp biodiversity potential of 18 species. Also, nearly 70% of the Natura 2000 zones in Munich have a potential of 18 species, while ca. 25% have a potential of 13 species and ca. 15% have a potential of 11 species. About 95% of the Natura 2000 areas in Berlin and ca. 65% of the Natura 2000 zones in Hamburg have a wasp biodiversity potential of 16 species. About 10% and 15% of the Natura 2000 zones in Hamburg have a potential of 10 and 14 species, respectively. About 90% of the Natura 2000 zones in Cologne have a wasp biodiversity potential of 13 species.

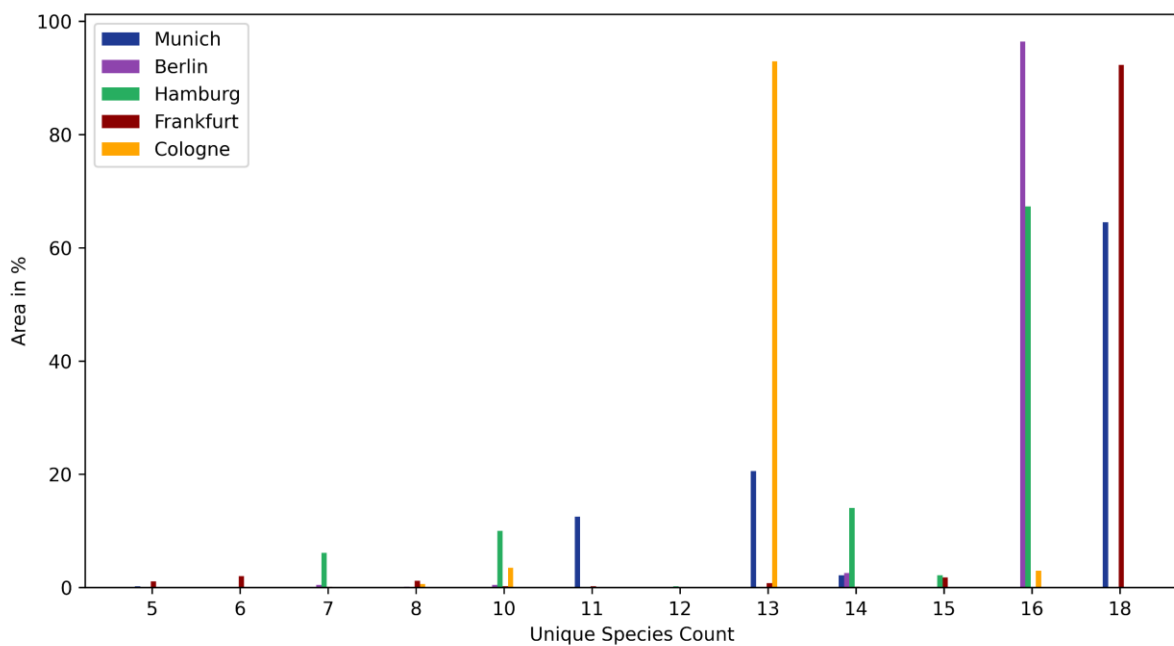


Fig. 37. Area in % of unique wasp species count intersecting with Natura 200 areas within Munich (blue), Berlin (purple), Hamburg (green), Frankfurt am Main (red) and Cologne (orange).

The ratio between the number of unique wasp species per ecotope and wasp biodiversity potential per ecosystem cluster can serve as an indicator of the health of a given ecotope. As the GBIF dataset has not been collected systematically (e.g. stratified), the assessment in Fig. 38 has no claim to correctness but rather highlights the application of the wasp biodiversity potential to estimate ecosystem health. Interestingly, the ratio of unique ecotope samples to cluster biodiversity potential appears to be the highest in all cities except Munich's city centres. The latter shows a high ratio between 0.53 and 0.7 within the city centre, while the majority of the rest of the city shows a ratio > 0.7 . The Natura 2000 zones show a very small ratio with ≤ 0.17 . This can also be found in all other cities, where the Natura 2000 have the lowest ratio relative to the rest of the cities. Also, the majority of Berlin has a very high ratio, as well as the south of Frankfurt. In contrast, most of Cologne has a rather low ratio, with the majority being between 0.17 and 0.35.

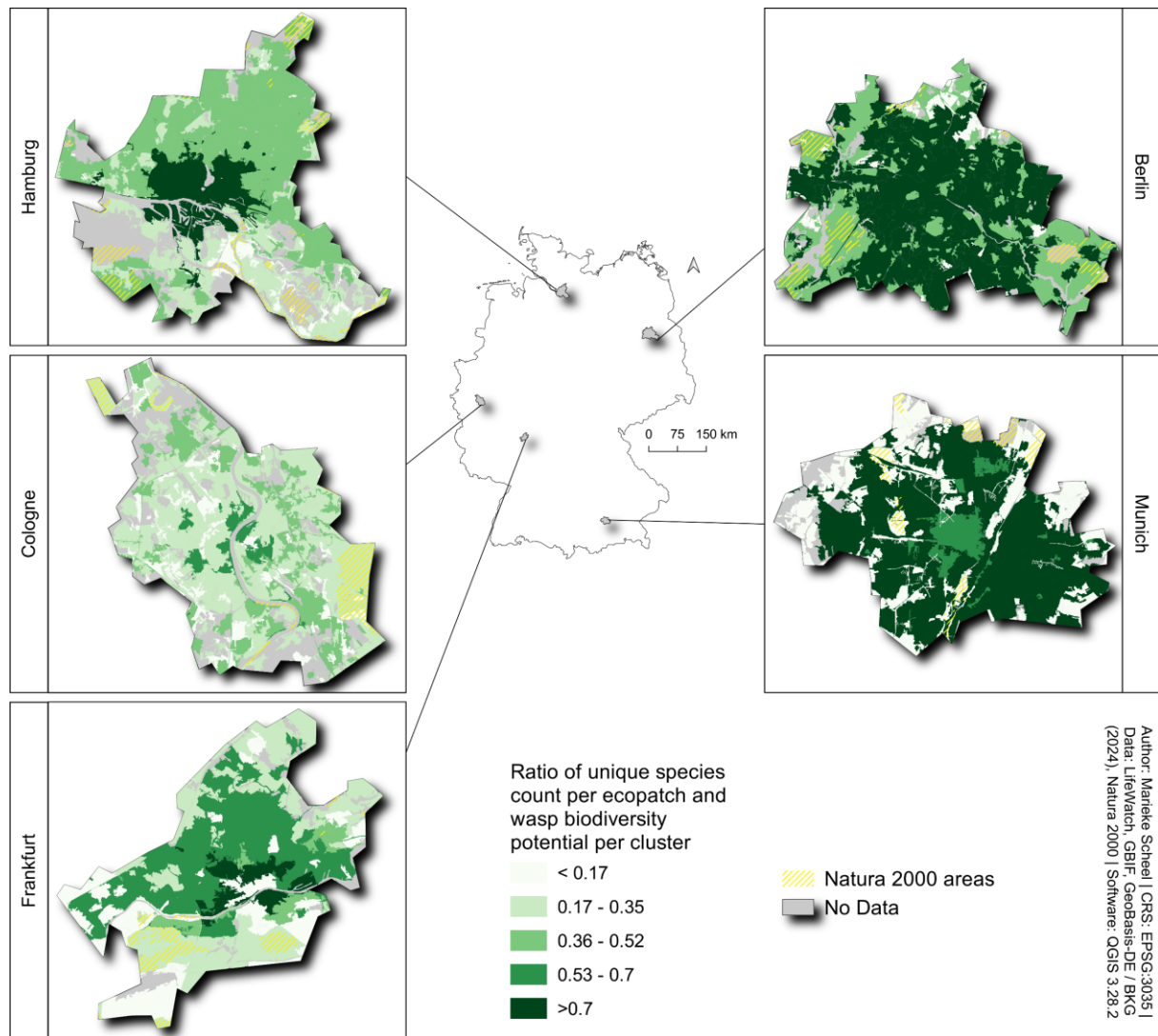


Fig. 38. Ratio between unique species per ecotope and wasp biodiversity potential per cluster in Hamburg, Cologne, Frankfurt, Berlin and Munich, the five most populated cities in Germany. Natura 2000 areas within the city perimeters are highlighted in yellow, and grey areas indicate no data (≤ 100 GBIF samples).

As previously shown, the clusters' wasp biodiversity potential can be used to assess the health status of the different ecosystems. However, there are not enough data samples for the ecosystem patches to assess the differences in health between them in this study. Nonetheless, Fig. 39 illustrates how such an assessment could look but has no claim to correctness. While most ecotope patches belonging to the selected ecosystem cluster do not have data samples, differences in biodiversity carrying capacity are shown, with some patches having only samples with one unique species and others having up to three unique species. While the biodiversity carrying capacity of this cluster was found to be 6 unique species, the different numbers of unique species per patch allow for comparison between the different patches that belong to this ecosystem cluster.

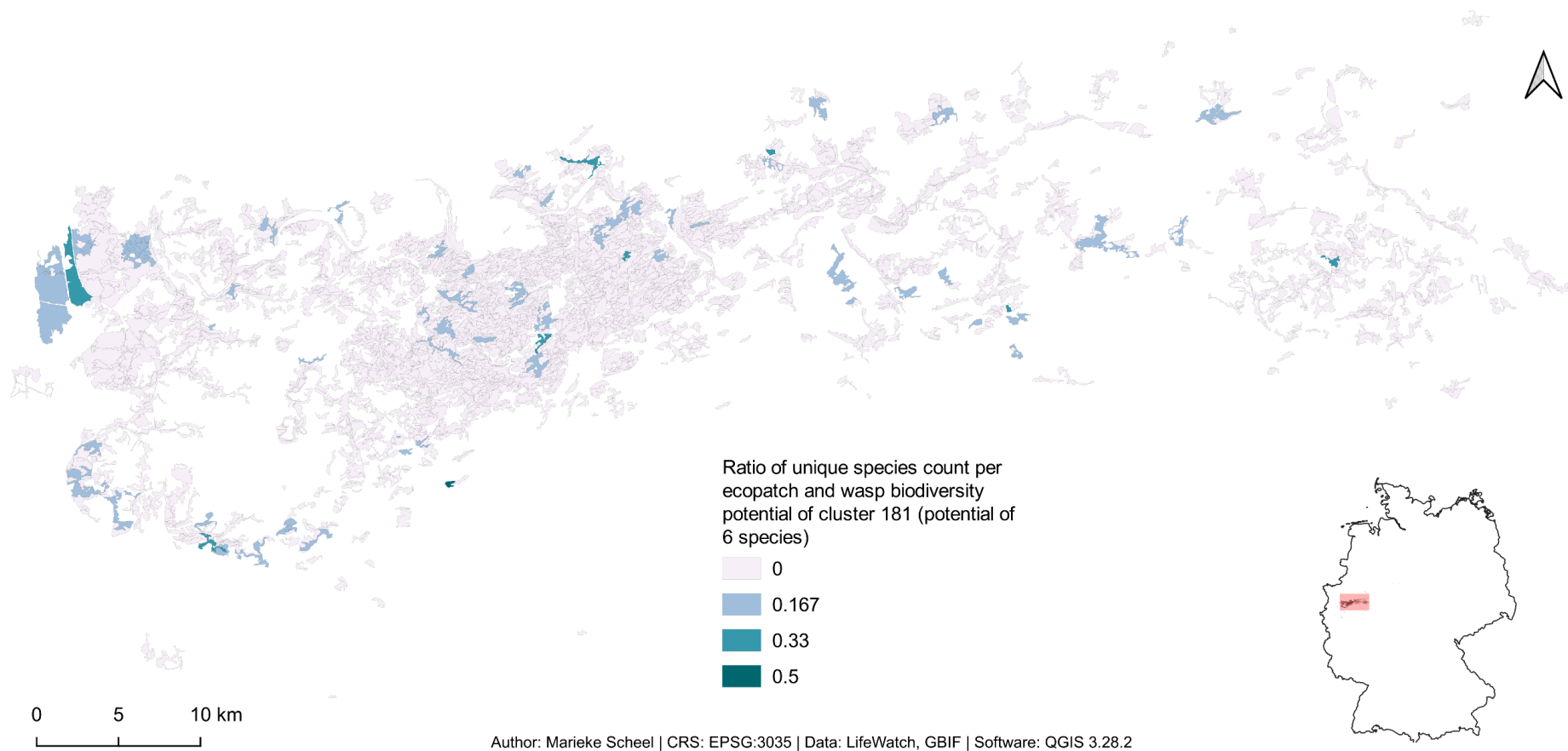


Fig. 39. Ratio between unique species per ecotope and wasp biodiversity potential of cluster 181 (Genie). A demonstration of how the unique number of species per cluster can be applied.

7. Discussion

This study demonstrates that the Genie clustering algorithm was able to form clusters that have a higher agreement with external habitat data than the clusters calculated using the K-means and GMM algorithm (Fig. 34). Interestingly, the clusters formed using Genie were indicated by the intrinsic quality metrics to be less well-separated than the other ones (Fig. 22). However, none of the three clustering algorithms achieved a high agreement, with Genie clusters receiving a mean score of 0.34. Additionally, when labelling the clusters, with both methods, the results highlighted areas that were mislabelled, which can both be discussed regarding the accuracy of the cluster delineation and the EUNIS habitat data (Fig. 27). Particularly the latter indicates a mismatch between the EUNIS habitat and LifeWatch biophysical data and highlight a need for a more rigorous quality control and/or pre-processing to make the datasets suitable for this purpose. The wasp biodiversity potential displayed regional differences that seemed to be generally dependent on sample sizes (Fig. 32 Fig. 34). Central and central-eastern Germany was identified as the region with the highest wasp biodiversity potential, while central-western Germany, particularly in the Ruhr District and Upper Rhin Plain, had notably the lowest wasp biodiversity potential (Fig. 31). Areas with a high wasp biodiversity potential exhibited a good agreement with Natura 2000 protection areas (Fig. 11, Fig. 34). Additionally, the wasp biodiversity potential demonstrated a negative correlation with biophysical variables that are typically indicating a less suitable environment for high biodiversity, e.g. artificial/urban environments. Vice versa, a positive correlation of wasp biodiversity potentially was found with biophysical variables that typically indicate a more suitable environment, e.g. natural environments (Fig. 35). The comparison between the wasp biodiversity potential in different cities revealed interesting variations in, among others, the maximum biodiversity potential and diversity in ecosystems (Fig. 36, Fig. 38).

7.1 Clustering Algorithms and Intrinsic Validation

The results of the Genie clustering algorithm outperforming the other two algorithms align with findings by Gagolewski et al. (2016), the developer of this algorithm. Benchmarks comparing the performance of the Genie algorithm with several established agglomerative hierarchical clustering methods revealed that the algorithm performed better on average. Furthermore, Moss et al. (2023) found that the Genie algorithm was able to delineate the highest number of optimal and distinct clusters compared to K-means and a hierarchical clustering algorithm. Moreover, they highlighted Genie's ability to distinguish outlier clusters and capture complex patterns. In the study by Moss et al. (2023), it was shown that the Genie algorithm was able to find clusters with lower concentrations or larger sizes, which compares well to the results of this study, where Genie delineated several very large clusters compared to the other two algorithms (Fig. 17, Fig. 20). Moreover, Moss et al. (2023) demonstrated that the Genie algorithm, while performing the best, also produced the most overlapping clusters, which could explain the results of the intrinsic clustering metrics (Fig. 22). All three metrics, the Silhouette Score, DBI

and CHI showed the Genie clusters to have the “worst” score. The Silhouette Score, which is around 0, expresses that the distance between the clusters is insignificant, which is backed by the high DBI and low CHI. Hence, it is likely that the arbitrary number of 200 clusters is not a good fit for the Genie algorithm to produce compact and well-distinguished clusters based on the LifeWatch data. However, this possibility was accounted for when determining the creation of a higher number of clusters (200) and risk over- rather than undersegmentation. Undersegmentation, while likely leading to more well-distinguished clusters, would have probably led to fewer optimal and distinct clusters (Moss et al., 2023). Thus, it is of greater importance for the further application of the clusters to distinguish between two different ecosystems, e.g. a heathland and a forest, than not to divide one ecosystem unnecessarily. The latter could be assumed to not necessarily be detrimental for assessing wasp biodiversity potential when assuming that the unique wasp species of given ecosystems are distributed evenly throughout the area. However, this is unlikely, especially given that the data was collected by citizen science, introducing spatial bias. Determining the optimal number of clusters before the clustering might have reduced this problem, e.g. using intrinsic clustering metrics like DBI, CHI and the Silhouette Score for different numbers of clusters. Nonetheless, the automated assessment for optimal clusters has been criticised for increasing the risk of aggregating information that would be reflected with a higher number of clusters (Moss et al., 2023).

The K-means clusters were indicated to be the best regarding distinctness and compactness (Fig. 22). However, while the DBI and CHI scores indicated the clusters to be better than those formed by the other two algorithms, the Silhouette Score was also close to 0. This was likely also caused by the intentional oversegmentation discussed above. Nonetheless, K-means is a fast algorithm that is easy to implement, making it suitable for large datasets such as the LifeWatch data (Arthur & Vassilvitskii, 2007). However, it was outperformed by the Genie algorithm when compared to external validation data. This could possibly be explained by the underlying assumptions of K-means, that clusters are roughly spherical and similar in size (Nanda & Panda, 2014). These assumptions indicate that the algorithm does not cope well with more complex data. Additionally, ecosystems are likely to vary greatly in size, with some being very small and local and others found over larger regions (Loveland & Merchant, 2004). Fig. 15 and Fig. 21 highlight the consequences of these initial K-means assumptions about size and shape, as the K-means clusters were shown to have the lowest area range out of the clusters by the three algorithms, with particularly no very large clusters. While K-means delineated a high number of small clusters, it is noteworthy that most of these small clusters are likely to have a higher dispersal, with K-means having the highest q1 bound for geographical intra-cluster distance (Fig. 21). In summary, these findings and the low purity score point to K-means having a rather low fit for the LifeWatch dataset.

The GMM algorithm was shown to perform well when spatially clustering geological data (Silva & Deutsch, 2018). Hence, it could have been expected to perform well for a similarly complex spatial delineation as in this study. While it was evaluated as better than Genie for the intrinsic cluster quality metrics (Fig. 22), the low Silhouette Score indicated overlapping and/or

more complex cluster shapes. Interestingly, the GMM clusters looked visually rather fragmented (Fig. 18) while having a similar size distribution to the K-means clusters, besides several very large clusters (Fig. 20). Together with the finding that the GMM clusters had the highest geographical intra-cluster distances (Fig. 21), these findings underline a low cohesion of the GMM clusters and imply that the algorithm does not capture the complexity of the data in this case. Meanwhile, it has to be considered that the GMM algorithm was only trained on 30% of the total LifeWatch data due to becoming computationally expensive with larger data. While the subset appeared to be representative of the dataset (Table 4), naturally, one can assume that details got lost when training the algorithm on the subset, and smaller and more local clusters are less likely to be recognised. This might also explain the rather large geographical intra-cluster distances of the GMM clusters. While uncertain, a constraint on the shapes of the GMM clusters might have introduced less bias while also significantly increasing the computational speed (Bouveyron & Brunet-Saumard, 2014). Moreover, the PCA analysis further reduced the details of the dataset, also likely at the cost of small and/or local clusters. Lastly, a bias might have been introduced by the data. The GMM algorithm assumes that the data follows a multivariate normal distribution, yet no data normalisation was conducted beyond standardisation (Eq. 1). A normal score transformation, as applied by Silva and Deutsch (2018), might have avoided such a bias.

It would have been interesting to compare the clustering results with the performance of hierarchical clustering algorithms. However, their high computational costs for large datasets could not be fulfilled in this study. Hierarchical clustering algorithms were found to be preferable for ecosystem delineation (Amigó et al., 2009; Krohn et al., 1999), since they allow for clusters of varying sizes and shapes. Their sensitivity to outliers and high computational demands could then potentially be remedied by further quality control, such as outlier removal and autoencoder (Moss et al., 2023). Particularly, the latter was shown to increase both clustering performance and computational speed and was interestingly also found to be particularly successful in combination with the Genie algorithm (Moss et al., 2023). As it also removes correlated variables, it could pose a good alternative to the VIF approach. However, similar to a PCA, the autoencoder alters the cluster attributes when reducing the dimensionality of the data (Moss et al., 2023).

The LifeWatch data quality control might have also introduced and/or preserved biases in the data. The standardisation of the data is an important step that was conducted in this study. However, the scikit StandardScaler (Scikit learn, 2024d) can be sensitive to outliers in the data (Moss et al., 2023). Thus, the scikit RobustScaler transformation (a log transform) (Scikit learn, 2024c) might have been a more suitable tool, as it removes the median of the data and scales it according to the IQR, effectively making it more outlier resistant (Moss et al., 2023).

Moreover, a qualitative LifeWatch variable screening could have been beneficial instead of or in addition to the VIF analysis. However, in this study, the time constraints did not allow for it. For instance, the variables describing distance metrics, proportions inside ecotopes, categorical land cover classes and contextual land cover proportion (Table A3) are likely to be highly dependent and might thus introduce noise and a weight towards land cover variables. Generally,

since the EUNIS habitat dataset was used as validation, land cover classes might have been preferable to be excluded from the training dataset.

7.2 External Clustering Validation

The purity scores yielded an average cluster purity of 0.34 for the Genie clusters and 0.29 for the GMM and K-means clusters. However, it should be noted that the differences in the purity scores of the Genie and K-means clusters were not statistically significant. Thus, further studies should apply further validation metrics (see next paragraph) to confirm these findings. Nonetheless, these results indicate that Genie found a higher number of optimal clusters than the other two algorithms, even though their q_1 lower bound for the proportion of the majority classes was similar to the bound of the distribution by GMM and K-means (Fig. 25). This finding, as well as the intrinsic metric results of Genie, imply that while Genie found the most optimal clusters, they additionally also found many overlapping and not well-delineated clusters. Generally, the purity scores are all rather low, with values <0.35 for all algorithms. Several factors could have additionally contributed to these results. Firstly, the EUNIS habitat data samples are not evenly sampled over the entire country, which introduces a spatial bias towards clusters in areas with a high EUNIS sample density (central-west, northeast and along the borders) in the validation (Fig. 9, Fig. 23). How this may or may not has influenced the validation scores is hard to gauge. Additionally, the sample sizes for different habitat classes differ vastly (Fig. 24). This could, on the one hand, be attributed to simply a higher occurrence of certain habitat classes. However, it is again difficult to validate this, and some habitat types in the EUNIS data likely have a sample size disproportionate to their actual occurrence in Germany, which would introduce a bias for certain ecosystem types in the validation.

Furthermore, using the purity score metric also comes with limitations. Due to its simplicity, it is one of the most popular measures for cluster evaluation (Amigó et al., 2009) and was also because of time constraints, the preferred metric for this study. Nonetheless, while it penalises noisy data in a cluster, it does not reward successful grouping of samples of the same type (Amigó et al., 2009). The latter might have been preferable for cases where clustering algorithms have e.g. successfully delineated two very similar ecosystems. Particularly since clusters will vary in their internal diversity. A simple measure for further studies to account for this would be to additionally compute the inverse purity score, which rewards the successful grouping of similar ecotopes but does not penalise mixing different sample types (Amigó et al., 2009). However, it also needs to be considered that the purity score is biased towards smaller clusters, as theoretically, a cluster with one EUNIS sample would have a purity score of 1. This is important to highlight since Genie had a slightly lower median cluster size than the other two clustering algorithms (Fig. 20). Hence, validation scores without a cluster size bias should be considered in further studies. For instance, the *BCubed* metric was shown by Amigó et al. (2009) to be a suitable validation metric for clusters.

Lastly, it is important to point out a potential mismatch between EUNIS habitat data and the LifeWatch biophysical variables since the majority of the clusters had very low purity scores (Fig. 25) independent of the clustering algorithms. The EUNIS habitat classes include old and recent samples; hence, they reflect both current and potential habitat classes (potential due to presumed (partially artificial) habitat changes). While the LifeWatch data also includes potential and categorical variables (Table A2, Table A3), the combination of both types of data in EUNIS and LifeWatch might differ and might make a comparison difficult. Moreover, the LifeWatch data is currently not validated, and no variable screening was conducted pre-analysis. Thus, it is unclear if all necessary variables were included to correctly identify habitat types (or if vice versa, too many variables were included that were not relevant and introduced noise). Subsequently, the cluster labelling also did not return a well-distinguished and consistent ecosystem delineation (Fig. 27). If time allowed, it would have been interesting to develop a method to combine the results from the two labelling approaches, as approach a) tended to omit smaller clusters and includes a bias towards the larger EUNIS sample classes and approach b) might have given too much weight to very small sample classes. Such a consistent method could have been used to merge clusters according to their assigned labels, which would have been beneficial in computing the wasp biodiversity potential more accurately. Additionally, it could have been interesting to use the labelling as a further validation metric between the three clustering algorithms, which was not conducted in this study due to time constraints.

7.3 Wasp Biodiversity Potential

The wasp biodiversity potential reached a maximum of 18 and a minimum of 5 species per cluster (Fig. 30). Additionally, clusters within Natura 2000 areas showed a higher wasp biodiversity potential than clusters outside of Natura 2000 areas (Fig. 34). Moreover, the location of the larger Natura 2000 areas in northeastern Germany and central Germany, correspond visually well to the occurrence of clusters with higher wasp biodiversity potential (Fig. 11, Fig. 31). They are also located in areas that are less fragmented, which in addition to heterogeneous natural habitats is an important determinant of high wasp biodiversity (Loyola & Martins, 2008; Steffan-Dewenter, 2003; Szczepko et al., 2020). While the northeastern region of Germany is the least fragmented (LiKi, 2018) and has a high percentage of Natura 2000 areas (Fig. 11), they have with 16 species a slightly lower potential than central Germany (18 species). Perhaps the relatively low sampling density in this region could have contributed to this result (Fig. 29). Interestingly, larger Natura 2000 areas in southwestern Germany fall within clusters with a lower wasp biodiversity potential (Fig. 11, Fig. 31). As the GBIF data has a high sample concentration in central-western and southwestern Germany, it is surprising that those areas have the lowest wasp biodiversity potential. This was also surprising as the sample size in clusters was found to be positively correlated with biodiversity potential (Fig. 33). Several factors could possibly contribute to these findings. While a regional sampling bias or other potential errors might have influenced this result, they could also indicate that the habitats in those ecosystem clusters are inappropriate for many wasp species. The latter would

thus emphasise a need for more than one taxon for a biodiversity potential calculation that is applicable to different ecosystems and capable to reflect their health status. As explained in section 2.4.3, the use of several taxa is also highly recommended in literature.

The central-southern region (roughly Ruhr district) with a relatively low wasp biodiversity potential is highly industrialised and has a high density of road and city infrastructure (LiKi, 2018). Since natural habitats with a high heterogeneity as well as the connectedness of those are the most important predictors for a high diversity of wasp species (Steffan-Dewenter, 2003), a low wasp biodiversity potential for the Ruhr district is unsurprising. The Upper Rhine Plain also showed similarly low levels of wasp biodiversity potential (Fig. 31). Given that there are several Natura 2000 areas within and around this area (Fig. 11), this result was a bit more surprising. Additionally, parts of the Upper Rhine Plain were declared biodiversity hotspots both by the responsible municipality and the federal government (Bundesamt für Naturschutz, 2024). Thus, it could be deduced that the mild climate and protected ecosystems should lead to a high wasp biodiversity potential in that area. However, it was found that the intensive land use, the already dense and still increasing infrastructure density as well as the vulnerability of the Upper Rhine Plain to climate change have led to increasing losses in biodiversity, particularly of insects (Regionalverband Ruhr, 2022). These factors could also be the cause of the relatively low wasp biodiversity potential in this region. Thus, the wasp biodiversity potential might already be able to reflect a deteriorated health of ecosystem clusters, if the potential of ecosystem clusters in a given region is lower than that of similar ecosystems in other regions.

The positive correlation between GBIF sample size and biodiversity potential as well as total area covered by clusters and biodiversity potential might be explained by several different factors. A large part of this result can likely be attributed to geographically larger clusters being more heterogenous and thus providing suitable conditions for a wider range of wasp species. However, it should also be noted that these results might indicate a spatial bias introduced by varying sample densities in different regions. To confirm this, a more even data distribution would be required.

The correlation analysis between the biophysical variables and biodiversity potential supported the results of the wasp biodiversity potential distribution. High numbers of unique wasp species were positively correlated with LifeWatch parameters that described natural ecotopes (e.g. forests, sparse vegetation, distance to sealed surfaces) and negatively correlated with LifeWatch parameters that described artificial ecotopes (e.g. buildings, night light, sealed surfaces). While the main determinants of a high wasp biodiversity potential, namely ecotope heterogeneity and connectivity (Steffan-Dewenter, 2003), were not assessed with this analysis, ecotope heterogeneity is typically higher in natural than artificial systems (Young & Jarvis, 2001). An ecosystem fragmentation analysis for areas with higher and lower wasp biodiversity potential could thus be a valuable further validation metric. Wang et al. (2014) suggest for instance the Connectance Index to be suitable. Yet, due to the dependency of such an analysis on the clustering algorithms such an analysis was deemed to be out of the scope of this study.

Most Natura 2000 areas in the five most populated cities in Germany intersected were within areas with high wasp biodiversity potential (Fig. 37). Interestingly, Cologne had the comparatively lowest potential with a maximum of 13 species in the areas with the highest potential (and in the Natura 2000 areas; Fig. 36). Since the city is located in central-western Germany (Fig. 5), the region that was found to have a generally lower wasp biodiversity potential (Fig. 31), this result is not unexpected. However, visually, it also appeared to be the most fragmented among the five cities (i.e. more and smaller clusters; Fig. 36), which could have added to a lower calculated potential. Berlin, Hamburg and Frankfurt were the cities with the largest wasp biodiversity potential, with the majority of their area having a potential of around 14 to 16 species (Fig. 36). On the other hand, most of the area of Munich had a medium potential (around 11 species). It is notable, however, that Munich, Berlin and Frankfurt have visually larger, more congruent and thus fewer clusters. In that light, Hamburg's high wasp biodiversity potential stands out as a more reliable result. This finding is also supported by Hamburg being the metropolis in Germany with the highest percentage of green area (ca. 71%) (Berliner Morgenpost, 2016). Interestingly, all other four cities are in the lower half of the ranking of the 14 German cities with more than 500,000 inhabitants regarding their green area percentage, with percentages between 50% and 60% of their area (Berliner Morgenpost, 2016). It should be noted, however, that all cities besides Berlin were commended on their well-implemented biodiversity projects (Deutscher Städtetag Berlin und Köln, 2021).

The application of the wasp biodiversity potential in the cities for an ecosystem health assessment was an example rather than a finding due to the lack of reliable data for such an application. Nonetheless, the results are of interest in assessing this methodology's use and potential pitfalls. The ratio of unique wasp species count per ecotope, and wasp biodiversity potential per cluster was consistently the highest in the city centres and the lowest in the Natura 2000 areas for all cities (Fig. 38). Due to the nature of the GBIF data, this is only somewhat surprising. The selected GBIF wasp data for Germany is mainly (min. 84% see section 4.2.3) based on citizen science data. People record more samples in places adjacent to residential areas, recreational areas as well as biodiversity hotspots (Geldmann et al., 2016). Thus, those areas are likely to have a higher number of samples. However, Geldmann et al. (2016) still found that forests and grasslands are still more likely to be undersampled in citizen science data collection. This might partly explain the low sample count in the Natura 2000 areas, and additionally highlights the need for more extensive data when applying the wasp biodiversity potential for a health assessment of ecotopes.

Additionally, when comparing the actual count in a given ecosystem patch to the calculated wasp biodiversity potential for that type of ecosystem, ecosystems with a higher potential will likely have a lower chance of getting a high proportion than ecosystems with low potential. Hence, it is more advisable only to compare ecotopes that are part of clusters with a high potential among each other (and vice versa for ecotopes part of a cluster with a low potential), as the ecosystem clusters vary in size, internal diversity and in this case also ecosystem specific sample sizes. Thus, a ratio of unique wasp species count per ecotope and wasp biodiversity

potential per cluster of e.g. 0.3 has to be assessed in the context of its cluster and the ratio value of other ecotopes of this cluster to assign it a rather high or low health status.

Lastly, the GBIF data initially included occurrence data of 46 wasp species, and after the quality control, it included 36 species. While the rigorous quality control removed more than half of the data points of the raw dataset, it aimed to reduce the uncertainties inherent in data based on citizen science (Jiménez et al., 2019). The time constraint of this study did not allow to assess the impact of the quality control. However, comparing the given results with results based on less rigorously quality-controlled data could have been interesting. Nonetheless, the spatial distribution of the data and the species composition in the data were not notably changed (Fig. 28, Fig. 29). Given that the occurrence samples were collected by citizens rather than experts, it is unsurprising that the dataset included only 46 wasp species, even though (Völkl & Blick, 2004) reported there to be 81 *Vespidae* species in Germany. Additionally, it is important to note that the lack of expert knowledge for species identification will have likely introduced a bias towards more common species. Moreover, one could have expected a bias due to sample size, i.e. areas with a higher concentration of samples will have more unique wasp species. However, such a bias was not observable, but rather the opposite could be found. The areas with the highest sample concentration were the Ruhr district and the Upper Rhine Plain, which were the regions that were calculated to have rather low wasp biodiversity potential (Fig. 29, Fig. 31). This finding emphasises a certain robustness of the results despite the uncertainties and limitations discussed.

7.4 Recommendations

An initial qualitative screening of the variables is recommended to refine the initial clustering of the biophysical variables. This could increase the accuracy of the results, particularly if the variables were selected to reflect the traits of the EUNIS habitat classification more closely. Additionally, while the Genie algorithm was found to perform the best, this result should be confirmed by improving the application of all algorithms with a pre-processing that includes a log normalisation of the data and an autoencoder. Also, more computing power and time could significantly improve the output of the GMM algorithm (Bouveyron & Brunet-Saumard, 2014). While oversegmentation was preferred over undersegmentation, a more thorough and time-intensive analysis could decrease the number of overlapping clusters by finding a more suitable number of clusters per algorithm. However, the risk of less optimal clusters for fewer total clusters should be considered (Moss et al., 2023). As the computing power and time (despite using the Rocket HPC cluster) for the clustering posed a general difficulty in this analysis, using a smaller area would also be recommended. It is advisable to develop a robust methodology on a smaller area and then scale it for a larger region. The additional computing power and an environmental gradient in the data should be accounted for when upscaling this methodology.

Also, it is recommended to use further external validation metrics in addition to the purity score, as it includes a bias for smaller clusters and does not reward grouping of similar data (Amigó et al., 2009). The inverse purity score could be a simple and time-efficient solution to add the

latter to the analysis. However, if permissible, the *BCubed* metric was found to be the most suitable (Amigó et al., 2009).

Also, there appears to be a mismatch between the EUNIS and LifeWatch data. While the latter includes several variables for artificial areas (Table A2, Table A3), the EUNIS data consists mainly of data for natural habitats (Table A6), and thus tends to mislabel the former (Fig. 27). Hence, it is recommended to either use a different dataset for validation or adapt the biophysical variables to match the EUNIS data. The latter is likely more difficult. Additionally, it is recommended that a suitable methodology be developed to label the clusters. The methodology would have to account for different sample sizes and should thus both consider total majorities and weighted majorities of habitat samples within a cluster.

Finally, this study demonstrated that a baseline for a biodiversity assessment can be established relatively easily, inexpensively, and in a timely manner. This method can be used for all types of ecosystems and can be applied to a large area. Further studies are recommended to apply the computed wasp biodiversity potential for a health assessment and compare the latter to an in-situ health analysis. Additionally, it would be relevant to compute additional biodiversity potential for further species, such as *Formicidae* (ants), *Heliconiini* and *Ithomiinae* (butterfly species) or *Isoptera* (termites) (Brown, 1991). Then, these could be compared to find which taxa are best suited for a general ecosystem health analysis. Additionally, such an analysis would be interesting to investigate whether certain taxa are more suitable for certain ecosystem types.

8. Conclusion

An efficient and fast tool to assess the health status of ecosystems is becoming increasingly important in light of rapidly changing and deteriorating ecosystems due to climate change and human land use. While many studies establish methods for assessing ecosystem health on a local or ecosystem specific scale, this study lays a foundation to establish a methodology for a tool that can be applied to a wide range of ecosystems.

It was demonstrated that the Genie clustering algorithm was the best for clustering biophysical variables to delineate ecosystems compared to GMM and K-means. However, this finding should be further validated due to the difference in purity scores not being significant between Genie and K-means. Moreover, the GMM and Genie algorithms were suggested by literature to be suitable for ecosystem delineation. Also, the external validation metric, the purity score, is biased towards smaller clusters. Additional validation metrics, such as the inverse purity score, are suggested to gain further insights into the cluster quality. Furthermore, the importance of in-depth quality control, pre-processing and variable screening was highlighted in this study, as those could have significantly improved clustering results, which were lower than expected with purity score values < 0.35 for all algorithms. This also highlighted that the clustering training data and the extrinsic validation data should be evaluated carefully in regard to their variables and what part of the ecosystem they reflect. Thus, the LifeWatch and EUNIS datasets would likely yield a higher agreement if a more rigorous quality control of variables and/or pre-processing was applied. However, a more in-depth data pre-processing and set-up of the algorithms was outside the scope of this study.

While literature suggested cluster oversegmentation to be preferred to find the most optimal clusters, this study did not find a suitable method to merge the resulting overlapping clusters due to inconsistencies and qualitative concerns of the labelling and time constraints. As intrinsic validation metrics indicated many overlapping clusters, this has likely influenced the computed wasp biodiversity potential baseline. By restricting that computation to the clusters with a minimum number of GBIF taxon samples, the smaller clusters were excluded, and the impact of the overlapping clusters was attempted to be minimised. In further studies, it is recommended to establish a methodology to merge overlapping clusters or, if this is not feasible, to compare different numbers of clusters and find the most optimal one.

Additionally, when labelling the clusters, with both methods, the results highlighted areas that were mislabelled, which can both be discussed regarding the accuracy of the cluster delineation and the EUNIS habitat data (Fig. 27). Particularly the latter indicates a mismatch between the EUNIS habitat and LifeWatch biophysical data and highlight a need for a more rigorous quality control and/or pre-processing to make the datasets suitable for this purpose. The wasp b

The computed wasp biodiversity potential was higher in the protected Natura 2000 areas than outside of them, thus supporting the results. Also, the wasp biodiversity potential tended to be higher in areas with a low infrastructure density. Importantly, the taxon data was collected through citizen science and thus includes an inherent bias. However, the wasp biodiversity

potential was found to be the lowest in the regions with the highest sample concentration in Germany. Applying the wasp biodiversity potential emphasised the need for evenly distributed and a larger number of data samples to accurately assess ecotopes. Additionally, it should be noted that the health assessment should not be compared for patches that belong to different ecosystems, especially when the same number of samples is available for e.g. an ecosystem with a potential of 3 species and another with a potential of 20 species.

Finally, it can be summarised that this study introduces an inexpensive and very fast method to establish a baseline for an ecosystem health assessment. It can be applied to various ecosystems and is thus optimal for policymakers and researchers to quickly gain knowledge about the state of the ecosystems in a given area. While the application of the biodiversity potential as a baseline requires further work, this study lays an important foundation. It is recommended that the taxon choice be investigated to determine if it introduces a bias for particular ecosystems.

9. References

- Abram, N. J., McGregor, H. V., Tierney, J. E., Evans, M. N., McKay, N. P., Kaufman, D. S., & Consortium, P. k. (2016). Early onset of industrial-era warming across the oceans and continents. *Nature*, 536(7617), 411-418. <https://doi.org/10.1038/nature19082>
- Agarwal, M., Rao, K. K., Vaidya, K., & Bhattacharya, S. (2021). ML-MOC: Machine Learning (kNN and GMM) based Membership determination for Open Clusters. *Monthly Notices of the Royal Astronomical Society*, 502(2), 2582-2599. <https://doi.org/10.1093/mnras/stab118>
- Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3), 91-93. <https://doi.org/https://doi.org/10.1016/j.tjem.2018.08.001>
- Amigó, E., Gonzalo, J., Artiles, J., & Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4), 461-486. <https://doi.org/10.1007/s10791-008-9066-8>
- Archambeau, J., Ruiz-Benito, P., Ratcliffe, S., Frejaville, T., Changenet, A., Castaneda, J. M. M., Lehtonen, A., Dahlgren, J., Zavala, M. A., & Garzon, M. B. (2020). Similar patterns of background mortality across Europe are mostly driven by drought in European beech and a combination of drought and competition in Scots pine. *Agricultural and Forest Meteorology*, 280, Article 107772. <https://doi.org/10.1016/j.agrformet.2019.107772>
- Arthur, D., & Vassilvitskii, S. (2007). *K-Means++: The Advantages of Careful Seeding* (Vol. 8). <https://doi.org/10.1145/1283383.1283494>
- Bakker, J. D. (2008). Increasing the utility of Indicator Species Analysis. *Journal of Applied Ecology*, 45(6), 1829-1835. <https://doi.org/10.1111/j.1365-2664.2008.01571.x>
- Banerjee, A., & Shan, H. (2010). Model-Based Clustering. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 686-689). Springer US. https://doi.org/10.1007/978-0-387-30164-8_554
- Baptist, M. J., van der Wal, J. T., Folmer, E. O., Gräwe, U., & Elschot, K. (2019). An ecotope map of the trilateral Wadden Sea. *Journal of Sea Research*, 152, 101761. <https://doi.org/https://doi.org/10.1016/j.seares.2019.05.003>
- Basset, A., Barbone, E., Elliott, M., Li, B.-L., Jorgensen, S. E., Lucena-Moya, P., Pardo, I., & Mouillot, D. (2013). A unifying approach to understanding transitional waters: Fundamental properties emerging from ecotone ecosystems. *Estuarine, Coastal and Shelf Science*, 132, 5-16. <https://doi.org/10.1016/j.ecss.2012.04.012>
- Behr, H. D., Jung, C., Trentmann, J., & Schindler, D. (2021). Using satellite data for assessing spatiotemporal variability and complementarity of solar resources—A case study from Germany. *Meteorol. Z*, 30, 515-532. <https://doi.org/10.1127/metz/2021/1081>
- Berger, V. W., & Zhou, Y. (2014). Kolmogorov–Smirnov Test: Overview. In *Wiley StatsRef: Statistics Reference Online*. <https://doi.org/https://doi.org/10.1002/9781118445112.stat06558>
- Berliner Morgenpost. (2016). *Das sind Deutschlands grünste Großstädte*. <https://interaktiv.morgenpost.de/gruenste-staedte-deutschlands/>
- Bhattacharjee, P., & Mitra, P. (2020). A survey of density based clustering algorithms. *Frontiers of Computer Science*, 15(1), 151308. <https://doi.org/10.1007/s11704-019-9059-3>

- Birkhofer, K., Rusch, A., Andersson, G. K. S., Bommarco, R., Dänhardt, J., Ekbom, B., Jönsson, A., Lindborg, R., Olsson, O., Rader, R., Stjernman, M., Williams, A., Hedlund, K., & Smith, H. G. (2018). A framework to identify indicator species for ecosystem services in agricultural landscapes. *Ecological Indicators*, 91, 278-286. <https://doi.org/10.1016/j.ecolind.2018.04.018>
- BMUV. (2023). *Natura 2000*. <https://www.bmu.de/themen/naturschutz/gebietsschutz-und-vernetzung/natura-2000/schutzgebietsnetz-natura-2000>
- Bouveyron, C., & Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71, 52-78. <https://doi.org/10.1016/j.csda.2012.12.008>
- Brown, K. S. (1991). 14 - Conservation of Neotropical Environments: Insects as Indicators. In N. M. Collins & J. A. Thomas (Eds.), *The Conservation of Insects and their Habitats* (pp. 349-404). Academic Press. <https://doi.org/10.1016/B978-0-12-181370-3.50020-8>
- Bryce, S. A., Omernik, J. M., & Larsen, D. P. (1999). Environmental Review: Ecoregions: A Geographic Framework to Guide Risk Characterization and Ecosystem Management. *Environmental Practice*, 1(3), 141-155. <https://doi.org/10.1017/S1466046600000582>
- Bundesamt für Naturschutz. (2024). *Förderschwerpunkt Hotspots der biologischen Vielfalt*. <https://www.bfn.de/bpbv-hotspots>
- Cannicci, S., Burrows, D., Fratini, S., Smith, T. J., Offenberg, J., & Dahdouh-Guebas, F. (2008). Faunal impact on vegetation structure and ecosystem function in mangrove forests: A review. *Aquatic Botany*, 89(2), 186-200. <https://doi.org/10.1016/j.aquabot.2008.01.009>
- Catalano, M., Leise, T., & Pfaff, T. (2009). Measuring Resource Inequality: The Gini Coefficient. *Numeracy*, 2. <https://doi.org/10.5038/1936-4660.2.2.4>
- Chen, Y.-C. (2017). A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1), 161-187. <https://doi.org/10.1080/24709360.2017.1396742>
- Chiarucci, A., Araújo, M., Decocq, G., Beierkuhnlein, C., & Fernández-Palacios, J. (2010). The concept of Potential Natural Vegetation. *Journal of Vegetation Science*, 21, 1172-1178. <https://doi.org/10.1111/j.1654-1103.2010.01218.x>
- Chillo, V., Anand, M., & Ojeda, R. A. (2011). Assessing the Use of Functional Diversity as a Measure of Ecological Resilience in Arid Rangelands. *Ecosystems*, 14(7), 1168-1177. <https://doi.org/10.1007/s10021-011-9475-1>
- Constanza, R., & Mageau, M. (1999). What is a healthy ecosystem? *Aquatic Ecology*, 33, 105-115.
- Cote, I. M., & Darling, E. S. (2010). Rethinking ecosystem resilience in the face of climate change. *PLoS Biol*, 8(7), e1000438. <https://doi.org/10.1371/journal.pbio.1000438>
- Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*(2), 224-227.
- Deutscher Städtetag Berlin und Köln. (2021). *Biodiversität: Diskussionspapier des Deutschen Städtetages*. Deutscher Städtetag: Die Stimme der Städte, <https://www.staedtetag.de/files/dst/docs/Publikationen/Positionspapiere/2021/diskussionspapier-biodiversitaet-2021.pdf>
- Dunn, O. J. (1964). Multiple Comparisons Using Rank Sums. *Technometrics*, 6(3), 241-252. <https://doi.org/10.1080/00401706.1964.10490181>

- EEA. (2019). *About the European Nature Information System, EUNIS*. <https://eunis.eea.europa.eu/about>
- EEA. (2022). *Natura 2000 - Spatial data*. https://www.eea.europa.eu/ds_resolveuid/ec69e703615c4c1489720e67d714ea04
- Emerson, R. W. (2020). Bonferroni Correction and Type I Error. *Journal of Visual Impairment & Blindness*, 114(1), 77-78. <https://doi.org/10.1177/0145482x20901378>
- Gagolewski, M. (2024). *genieclust: Fast and Robust Hierarchical Clustering with Noise Point Detection*. <https://genieclust.gagolewski.com/index.html>
- Gagolewski, M., Bartoszek, M., & Cena, A. (2016). Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm. *Information Sciences*, 363, 8-23. <https://doi.org/https://doi.org/10.1016/j.ins.2016.05.003>
- Gao, C. X., Dwyer, D., Zhu, Y., Smith, C. L., Du, L., Filia, K. M., Bayer, J., Mensink, J. M., Wang, T., Bergmeir, C., Wood, S., & Cotton, S. M. (2023). An overview of clustering methods with guidelines for application in mental health research. *Psychiatry Research*, 327, 115265. <https://doi.org/https://doi.org/10.1016/j.psychres.2023.115265>
- GBIF. (2024). *Global Biodiversity Information Facility: Free and open access to biodiversity data* <https://www.gbif.org/>
- Geldmann, J., Heilmann-Clausen, J., Holm, T. E., Levinsky, I., Markussen, B., Olsen, K., Rahbek, C., & Tøttrup, A. P. (2016). What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, 22(11), 1139-1149. <https://doi.org/https://doi.org/10.1111/ddi.12477>
- Groth, D., Hartmann, S., Klie, S., & Selbig, J. (2013). Principal Components Analysis. In B. Reisfeld & A. N. Mayeno (Eds.), *Computational Toxicology: Volume II* (pp. 527-547). Humana Press. https://doi.org/10.1007/978-1-62703-059-5_22
- Harabasz, C. T., & Karoński, M. (1974). A dendrite method for cluster analysis. In *Communications in Statistics* (Vol. 3, pp. 1-27).
- Hasani, M., Pieleśiak, I., Mahiny, A. S., & Mikaeili, A. (2023). Regional ecosystem health assessment based on landscape patterns and ecosystem services approach. *Acta Ecologica Sinica*, 43(2), 333-342. <https://doi.org/10.1016/j.chnaes.2021.11.004>
- Hassani, M., & Seidl, T. (2017). Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. *Vietnam Journal of Computer Science*, 4(3), 171-183. <https://doi.org/10.1007/s40595-016-0086-9>
- Hoffmann, M., Hilton-Taylor, C., Angulo, A., Böhm, M., Brooks, T. M., Butchart, S. H. M., Carpenter, K. E., Chanson, J., Collen, B., Cox, N. A., Darwall, W. R. T., Dulvy, N. K., Harrison, L. R., Katariya, V., Pollock, C. M., Quader, S., Richman, N. I., Rodrigues, A. S. L., Tognelli, M. F., Vié, J.-C., Aguiar, J. M., Allen, D. J., Allen, G. R., Amori, G., Ananjeva, N. B., Andreone, F., Andrew, P., Ortiz, A. L. A., Baillie, J. E. M., Baldi, R., Bell, B. D., Biju, S. D., Bird, J. P., Black-Decima, P., Blanc, J. J., Bolaños, F., Bolivar-G., W., Burfield, I. J., Burton, J. A., Capper, D. R., Castro, F., Catullo, G., Cavanagh, R. D., Channing, A., Chao, N. L., Chenery, A. M., Chiozza, F., Clausnitzer, V., Collar, N. J., Collett, L. C., Collette, B. B., Fernandez, C. F. C., Craig, M. T., Crosby, M. J., Cumberlidge, N., Cuttelod, A., Derocher, A. E., Diesmos, A. C., Donaldson, J. S., Duckworth, J. W., Dutson, G., Dutta, S. K., Emslie, R. H., Farjon, A., Fowler, S., Freyhof, J., Garshelis, D. L., Gerlach, J., Gower, D. J., Grant, T. D., Hammerson, G. A., Harris, R. B., Heaney, L. R., Hedges, S. B., Hero, J.-M., Hughes, B., Hussain, S. A.,

- Icochea M., J., Inger, R. F., Ishii, N., Iskandar, D. T., Jenkins, R. K. B., Kaneko, Y., Kottelat, M., Kovacs, K. M., Kuzmin, S. L., La Marca, E., Lamoreux, J. F., Lau, M. W. N., Lavilla, E. O., Leus, K., Lewison, R. L., Lichtenstein, G., Livingstone, S. R., Lukoschek, V., Mallon, D. P., McGowan, P. J. K., McIvor, A., Moehlman, P. D., Molur, S., Alonso, A. M., Musick, J. A., Nowell, K., Nussbaum, R. A., Olech, W., Orlov, N. L., Papenfuss, T. J., Parra-Olea, G., Perrin, W. F., Polidoro, B. A., Pourkazemi, M., Racey, P. A., Ragle, J. S., Ram, M., Rathbun, G., Reynolds, R. P., Rhodin, A. G. J., Richards, S. J., Rodríguez, L. O., Ron, S. R., Rondinini, C., Rylands, A. B., Sadovy de Mitcheson, Y., Sanciangco, J. C., Sanders, K. L., Santos-Barrera, G., Schipper, J., Self-Sullivan, C., Shi, Y., Shoemaker, A., Short, F. T., Sillero-Zubiri, C., Silvano, D. L., Smith, K. G., Smith, A. T., Snoeks, J., Stattersfield, A. J., Symes, A. J., Taber, A. B., Talukdar, B. K., Temple, H. J., Timmins, R., Tobias, J. A., Tsytulina, K., Tweddle, D., Ubeda, C., Valenti, S. V., Paul van Dijk, P., Veiga, L. M., Veloso, A., Wege, D. C., Wilkinson, M., Williamson, E. A., Xie, F., Young, B. E., Akçakaya, H. R., Bennun, L., Blackburn, T. M., Boitani, L., Dublin, H. T., da Fonseca, G. A. B., Gascon, C., Lacher, T. E., Mace, G. M., Mainka, S. A., McNeely, J. A., Mittermeier, R. A., Reid, G. M., Rodriguez, J. P., Rosenberg, A. A., Samways, M. J., Smart, J., Stein, B. A., & Stuart, S. N. (2010). The Impact of Conservation on the Status of the World's Vertebrates. *Science*, 330(6010), 1503-1509. <https://doi.org/doi:10.1126/science.1194442>
- Ismaili, O. A., Lemaire, V., & Cornuéjols, A. (2014, 2014//). A Supervised Methodology to Measure the Variables Contribution to a Clustering. *Neural Information Processing*, Cham.
- Jiménez, M., Triguero, I., & John, R. (2019). Handling uncertainty in citizen science data: Towards an improved amateur-based large-scale classification. *Information Sciences*, 479, 301-320. <https://doi.org/https://doi.org/10.1016/j.ins.2018.12.011>
- Jin, X., & Han, J. (2010). Partitional Clustering. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 766-766). Springer US. https://doi.org/10.1007/978-0-387-30164-8_631
- Jørgensen, S. E., Xu, F., Marques, J., & Salas, F. (2011). Application of indicators for the assessment of ecosystem health. In S. Jørgensen (Ed.), *Handbook of ecological indicators for assessment of ecosystem health* (Second ed., pp. 9–76). CRC Press.
- Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 881-892. <https://doi.org/10.1109/TPAMI.2002.1017616>
- Karo, I. M. K., MaulanaAdhinugraha, K., & Huda, A. F. (2017, 1-3 Nov. 2017). A cluster validity for spatial clustering based on davies bouldin index and Polygon Dissimilarity function. 2017 Second International Conference on Informatics and Computing (ICIC),
- Klijn, F., & de Haes, H. A. U. (1994). A hierarchical approach to ecosystems and its implications for ecological land classification. *Landscape Ecology*, 9(2), 89-104. <https://doi.org/10.1007/BF00124376>
- Kroes, H. W. (1977). The niche structure of ecosystems. *Journal of Theoretical Biology*, 65(2), 317-326. [https://doi.org/https://doi.org/10.1016/0022-5193\(77\)90327-7](https://doi.org/https://doi.org/10.1016/0022-5193(77)90327-7)

- Krohn, W. B., Boone, R. B., & Painton, S. L. (1999). Quantitative Delineation and Characterization of Hierarchical Biophysical Regions of Maine. *Northeastern Naturalist*, 6(2), 139-164. <https://doi.org/10.2307/3858346>
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583-621.
- Küchler, A. W. (1964). Potential natural vegetation of the conterminous United States (map scale—1,500,000). *The national atlas of the United States of America*, US Geological Survey, 89–92.
- Kyriazos, T. a. P., M. (2023). Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions. *Open Journal of Statistics*, 13, 404-424. <https://doi.org/10.4236/ojs.2023.133020>
- Lackey, R. T. (2011). Values, Policy, and Ecosystem Health: Options for resolving the many ecological policy issues we face depend on the concept of ecosystem health, but ecosystem health is based on controversial, value-based assumptions that masquerade as science *Bioscience*, 52(6), 437–443. [https://doi.org/https://doi.org/10.1641/0006-3568\(2001\)051\[0437:VPAEH\]2.0.CO;2](https://doi.org/https://doi.org/10.1641/0006-3568(2001)051[0437:VPAEH]2.0.CO;2)
- LifeWatch ERIC. (2017). *Resources & Services*. Retrieved 14.07.2024 from <https://www.lifewatch.eu/>
- LiKi. (2018). *Unzerschnittene verkehrsarme Räume (UZVR) > 100 km² in Deutschland*. Länderinitiative Kernindikatoren; Landesamt für Natur, Umwelt und Verbraucherschutz. https://www.liki.nrw.de/fileadmin/liki/Dokumente/Literatur/BfN_Artikel_UZVR.pdf
- Lindeman, R. L. (1942). The trophic-dynamic aspect of ecology. *Ecology*, 23(4), 399-417.
- Lindenmayer, D., & Likens, G. E. (2010). *Effective ecological monitoring*. CSIRO Publishing Collingwood, Vic.
- Lindqvist, M., Palme, U., & Lindner, J. P. (2016). A comparison of two different biodiversity assessment methods in LCA—a case study of Swedish spruce forest. *The International Journal of Life Cycle Assessment*, 21(2), 190-201. <https://doi.org/10.1007/s11367-015-1012-6>
- Loveland, T. R., & Merchant, J. M. (2004). Ecoregions and Ecoregionalization: Geographical and Ecological Perspectives. *Environmental Management*, 34(1), S1-S13. <https://doi.org/10.1007/s00267-003-5181-x>
- Loyola, R. D., & Martins, R. P. (2008). Habitat structure components are effective predictors of trap-nesting Hymenoptera diversity. *Basic and Applied Ecology*, 9(6), 735-742. <https://doi.org/https://doi.org/10.1016/j.baae.2007.06.016>
- Lu, Y., Wang, R., Zhang, Y., Su, H., Wang, P., Jenkins, A., Ferrier, R. C., Bailey, M., & Squire, G. (2015). Ecosystem health towards sustainability. *Ecosystem Health and Sustainability*, 1(1), 1-15. <https://doi.org/10.1890/ehs14-0013.1>
- MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. *University of California Press*, 1, 281-297.
- Marschner, F. J. (1950). Major land uses in the United States (map scale 1:5,000,000). *USDA Agricultural Research Service Washington*.

- Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253), 68-78. <https://doi.org/10.2307/2280095>
- McLachlan, G. J., & Peel, D. (2000). Finite Mixture Models. Wiley. <https://doi.org/10.1002/0471721182>
- Mengist, W., Soromessa, T., & Feyisa, G. L. (2020). A global view of regulatory ecosystem services: existed knowledge, trends, and research gaps. *Ecological Processes*, 9(1), 40. <https://doi.org/10.1186/s13717-020-00241-w>
- Meyer, C., Kreft, H., Guralnick, R., & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications*, 6(1), 8221. <https://doi.org/10.1038/ncomms9221>
- Moss, M. A. N., Hughes, D. D., Crawford, I., Gallagher, M. W., Flynn, M. J., & Topping, D. O. (2023). Comparative Analysis of Traditional and Advanced Clustering Techniques in Bioaerosol Data: Evaluating the Efficacy of K-Means, HCA, and GenieClust with and without Autoencoder Integration. *Atmosphere*, 14(9). https://mdpi-res.com/d_attachment/atmosphere/atmosphere-14-01416/article_deploy/atmosphere-14-01416-v2.pdf?version=1694431333
- Moyes, P. (2007). An extended Gini approach to inequality measurement. *The Journal of Economic Inequality*, 5(3), 279-303. <https://doi.org/10.1007/s10888-006-9051-7>
- Mumby, P. J., & Steneck, R. S. (2008). Coral reef management and conservation in light of rapidly evolving ecological paradigms. *Trends Ecol Evol*, 23(10), 555-563. <https://doi.org/10.1016/j.tree.2008.06.011>
- NABU. (2024). *Naturgucker*. <https://www.naturgucker.de/natur.dll/A~oflVvDzsa1SzPfQ8A91T6UCii/>
- Nanda, S. J., & Panda, G. (2014). A survey on nature inspired metaheuristic algorithms for partitional clustering. *Swarm and Evolutionary Computation*, 16, 1-18. <https://doi.org/https://doi.org/10.1016/j.swevo.2013.11.003>
- Neter, J., Wasserman, W., & Kutner, M. H. (1983). Applied linear regression models [Books]. xv, 546p. : tables.
- Noss, R. F. (1990). Indicators for Monitoring Biodiversity: A Hierarchical Approach. *Conservation Biology*, 4(4), 355-364. <https://doi.org/https://doi.org/10.1111/j.1523-1739.1990.tb00309.x>
- O'Brien, A., Townsend, K., Hale, R., Sharley, D., & Pettigrove, V. (2016). How is ecosystem health defined and measured? A critical review of freshwater and estuarine studies. *Ecological Indicators*, 69, 722-729. <https://doi.org/10.1016/j.ecolind.2016.05.004>
- Oldeland, J., Dorigo, W., Lieckfeld, L., Lucieer, A., & Jürgens, N. (2010). Combining vegetation indices, constrained ordination and fuzzy classification for mapping semi-natural vegetation units from hyperspectral imagery. *Remote Sensing of Environment*, 114(6), 1155-1166. <https://doi.org/https://doi.org/10.1016/j.rse.2010.01.003>
- Oliver, T. H., Heard, M. S., Isaac, N. J. B., Roy, D. B., Procter, D., Eigenbrod, F., Freckleton, R., Hector, A., Orme, C. D. L., Petchey, O. L., Proença, V., Raffaelli, D., Suttle, K. B., Mace, G. M., Martín-López, B., Woodcock, B. A., & Bullock, J. M. (2015). Biodiversity and Resilience of Ecosystem Functions. *Trends in Ecology & Evolution*, 30(11), 673-684. <https://doi.org/10.1016/j.tree.2015.08.009>
- Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C.

- J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., & Kassem, K. R. (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *Bioscience*, 51(11), 933-938. [https://doi.org/10.1641/0006-3568\(2001\)051\[0933:Teotwa\]2.0.Co;2](https://doi.org/10.1641/0006-3568(2001)051[0933:Teotwa]2.0.Co;2)
- Omernik, J. M. (2004). Perspectives on the Nature and Definition of Ecological Regions. *Environmental Management*, 34(1), S27-S38. <https://doi.org/10.1007/s00267-003-5197-2>
- Omernik, J. M., & Bailey, R. G. (1997). DISTINGUISHING BETWEEN WATERSHEDS AND ECOREGIONS1. *JAWRA Journal of the American Water Resources Association*, 33(5), 935-949. <https://doi.org/https://doi.org/10.1111/j.1752-1688.1997.tb04115.x>
- Otegui, J., Ariño, A. H., Encinas, M. A., & Pando, F. (2013). Assessing the Primary Data Hosted by the Spanish Node of the Global Biodiversity Information Facility (GBIF). *PLoS One*, 8(1), e55144. <https://doi.org/10.1371/journal.pone.0055144>
- Peleg, O., Blain, C., & Shears, N. (2023). Multi-indicator 'state space' approach to assessing changes in shallow urban reef ecosystem health. *Mar Environ Res*, 186, 105895. <https://doi.org/10.1016/j.marenvres.2023.105895>
- Pereira, H. M., Leadley, P. W., Proença, V., Alkemade, R., Scharlemann, J. P. W., Fernandez-Manjarrés, J. F., Araújo, M. B., Balvanera, P., Biggs, R., Cheung, W. W. L., Chini, L., Cooper, H. D., Gilman, E. L., Guénette, S., Hurtt, G. C., Huntington, H. P., Mace, G. M., Oberdorff, T., Revenga, C., Rodrigues, P., Scholes, R. J., Sumaila, U. R., & Walpole, M. (2010). Scenarios for Global Biodiversity in the 21st Century. *Science*, 330(6010), 1496-1501. <https://doi.org/doi:10.1126/science.1196624>
- Rapport, D. J., Costanza, R., & McMichael, A. J. (1998). Assessing ecosystem health. *Trends in Ecology & Evolution*, 13(10), 397-402. [https://doi.org/10.1016/S0169-5347\(98\)01449-9](https://doi.org/10.1016/S0169-5347(98)01449-9)
- Regionalverband Ruhr. (2022). *Regionale Biodiversitäts-Strategie Ruhrgebiet: Offensive Grüne Infrastruktur 2030* Konferenz zur regionalen Biodiversitätsstrategie Ruhrgebiet am 02.06.2022, <https://www.rvr.ruhr/themen/oekologie-umwelt/gruene-infrastruktur/news-gruene-infrastruktur/news/regionale-biodiversitaets-strategie-ruhrgebiet-veroeffentlicht/>
- Rombouts, I., Beaugrand, G., Artigas, L. F., Dauvin, J. C., Gevaert, F., Goberville, E., Kopp, D., Lefebvre, S., Luczak, C., Spilmont, N., Travers-Trolet, M., Villanueva, M. C., & Kirby, R. R. (2013). Evaluating marine ecosystem health: Case studies of indicators using direct observations and modelling methods. *Ecological Indicators*, 24, 353-365. <https://doi.org/https://doi.org/10.1016/j.ecolind.2012.07.001>
- Romero, J., Martinez-Crego, B., Alcoverro, T., & Perez, M. (2007). A multivariate index based on the seagrass *Posidonia oceanica* (POMI) to assess ecological status of coastal waters under the water framework directive (WFD). *Mar Pollut Bull*, 55(1-6), 196-204. <https://doi.org/10.1016/j.marpolbul.2006.08.032>
- 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.
- Saemi, B., Hosseinabadi, A. A. R., Kardgar, M., Balas, V. E., & Ebadi, H. (2018, 2018/). Nature Inspired Partitioning Clustering Algorithms: A Review and Analysis. *Soft Computing Applications*, Cham.

- Schaeffer, D. J., Herricks, E. E., & Kerster, H. W. (1988). Ecosystem Health: I. Measuring Ecosystem Health. *Environmental Management*, 12(4), 445-455.
- Schowalter, T. D. (2016). Chapter 11 - Ecosystem Structure and Function. In T. D. Schowalter (Ed.), *Insect Ecology (Fourth Edition)* (pp. 367-404). Academic Press. <https://doi.org/10.1016/B978-0-12-803033-2.00011-X>
- Scikit learn. (2024a). *GaussianMixture*. <https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html>
- Scikit learn. (2024b). *KMeans*. <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>
- Scikit learn. (2024c). *RobustScaler*. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html>
- Scikit learn. (2024d). *StandardScaler*. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
- Seidl, R., Spies, T. A., Peterson, D. L., Stephens, S. L., & Hicke, J. A. (2016). Searching for resilience: addressing the impacts of changing disturbance regimes on forest ecosystem services. *J Appl Ecol*, 53(1), 120-129. <https://doi.org/10.1111/1365-2664.12511>
- Siddig, A. A. H., Ellison, A. M., Ochs, A., Villar-Leeman, C., & Lau, M. K. (2016). How do ecologists select and use indicator species to monitor ecological change? Insights from 14 years of publication in Ecological Indicators. *Ecological Indicators*, 60, 223-230. <https://doi.org/10.1016/j.ecolind.2015.06.036>
- Silva, D. S. F., & Deutsch, C. V. (2018). Multivariate data imputation using Gaussian mixture models. *Spatial Statistics*, 27, 74-90. <https://doi.org/10.1016/j.spasta.2016.11.002>
- Soto, S., & Pintó, J. (2010). Delineation of natural landscape units for Puerto Rico. *Applied Geography*, 30(4), 720-730. <https://doi.org/10.1016/j.apgeog.2010.01.010>
- Steffan-Dewenter, I. (2003). Importance of Habitat Area and Landscape Context for Species Richness of Bees and Wasps in Fragmented Orchard Meadows. *Conservation Biology*, 17(4), 1036-1044. <https://doi.org/10.1046/j.1523-1739.2003.01575.x>
- Storch, D., Šimová, I., Smyčka, J., Bohdalková, E., Toszogyova, A., & Okie, J. G. (2022). Biodiversity dynamics in the Anthropocene: how human activities change equilibria of species richness. *Ecography*, 2022(4). <https://doi.org/10.1111/ecog.05778>
- Suter II, G. W. (1993). A critique of ecosystem health concepts and indexes. *Environmental Toxicology and Chemistry*, 12(9), 1533-1539. <https://doi.org/10.1002/etc.5620120903>
- Swarndeeep Saket, J., & Pandya, S. (2016). An overview of partitioning algorithms in clustering techniques. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 5(6), 1943-1946. <https://pubs.rsc.org/en/content/articlepdf/2022/nh/d1nh00621e>
- Szczepko, K., Kruk, A., & Wiśniowski, B. (2020). Local habitat conditions shaping the assemblages of vespine wasps (Hymenoptera: Vespidae) in a post-agricultural landscape of the Kampinos National Park in Poland. *Scientific Reports*, 10(1), 1424. <https://doi.org/10.1038/s41598-020-57426-8>

<https://www.nature.com/articles/s41598-020-57426-8.pdf>

- Tansley, A. G. (1935). The use and abuse of vegetational concepts and terms. *Ecology*, 16(3), 284-307.
- Tett, P., Gowen, R., Mills, D., Fernandes, T., Gilpin, L., Huxham, M., Kennington, K., Read, P., Service, M., Wilkinson, M., & Malcolm, S. (2007). Defining and detecting undesirable disturbance in the context of marine eutrophication. *Mar Pollut Bull*, 55(1-6), 282-297. <https://doi.org/10.1016/j.marpolbul.2006.08.028>
- Tett, P., Gowen, R. J., Painting, S. J., Elliott, M., Forster, R., Mills, D. K., Bresnan, E., Capuzzo, E., Fernandes, T. F., Foden, J., Geider, R. J., Gilpin, L. C., Huxham, M., McQuatters-Gollop, A. L., Malcolm, S. J., Saux-Picart, S., Platt, T., Racault, M. F., Sathyendranath, S., van der Molen, J., & Wilkinson, M. (2013). Framework for understanding marine ecosystem health. *Marine Ecology Progress Series*, 494, 1-27. <https://doi.org/10.3354/meps10539>
- Troya-Galvis, A., Gańczarski, P., Passat, N., & Berti-Équille, L. (2015). Unsupervised Quantification of Under- and Over-Segmentation for Object-Based Remote Sensing Image Analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(5), 1936-1945. <https://doi.org/10.1109/JSTARS.2015.2424457>
- Vinutha, H. P., Poornima, B., & Sagar, B. M. (2018). Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset. Information and Decision Sciences, Singapore.
- Völkl, W., & Blick, T. (2004). Die quantitative Erfassung der rezenten Fauna von Deutschland. *BfN-Skripten*, 117, 1-85.
- Wang, X., Blanchet, F. G., & Koper, N. (2014). Measuring habitat fragmentation: An evaluation of landscape pattern metrics. *Methods in Ecology and Evolution*, 5(7), 634-646. <https://doi.org/https://doi.org/10.1111/2041-210X.12198>
- Wang, Z., Da Cunha, C., Ritou, M., & Furet, B. (2019). Comparison of K-means and GMM methods for contextual clustering in HSM. *Procedia Manufacturing*, 28, 154-159. <https://doi.org/https://doi.org/10.1016/j.promfg.2018.12.025>
- Watson, J. E., Dudley, N., Segan, D. B., & Hockings, M. (2014). The performance and potential of protected areas. *Nature*, 515(7525), 67-73. <https://doi.org/10.1038/nature13947>
- Yang, J., Oh, S.-O., & Hur, J.-S. (2023). Lichen as bioindicators: assessing their response to heavy metal pollution in their native ecosystem. *Mycobiology*, 51(5), 343-353.
- Yoo, A. B., Jette, M. A., & Grondona, M. (2003). SLURM: Simple Linux Utility for Resource Management. In D. Feitelson, L. Rudolph, & U. Schwiegelshohn, *Job Scheduling Strategies for Parallel Processing* Berlin, Heidelberg.
- Young, C. H., & Jarvis, P. J. (2001). Assessing the structural heterogeneity of urban areas: an example from the Black Country (UK). *Urban Ecosystems*, 5(1), 49-69. <https://doi.org/10.1023/A:1021877618584>
- Zacharias, M. A., & Roff, J. C. (2001). Use of focal species in marine conservation and management: a review and critique. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 11, 59-76.
- Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2), 103-114.

10. Appendix

10.1 Figures

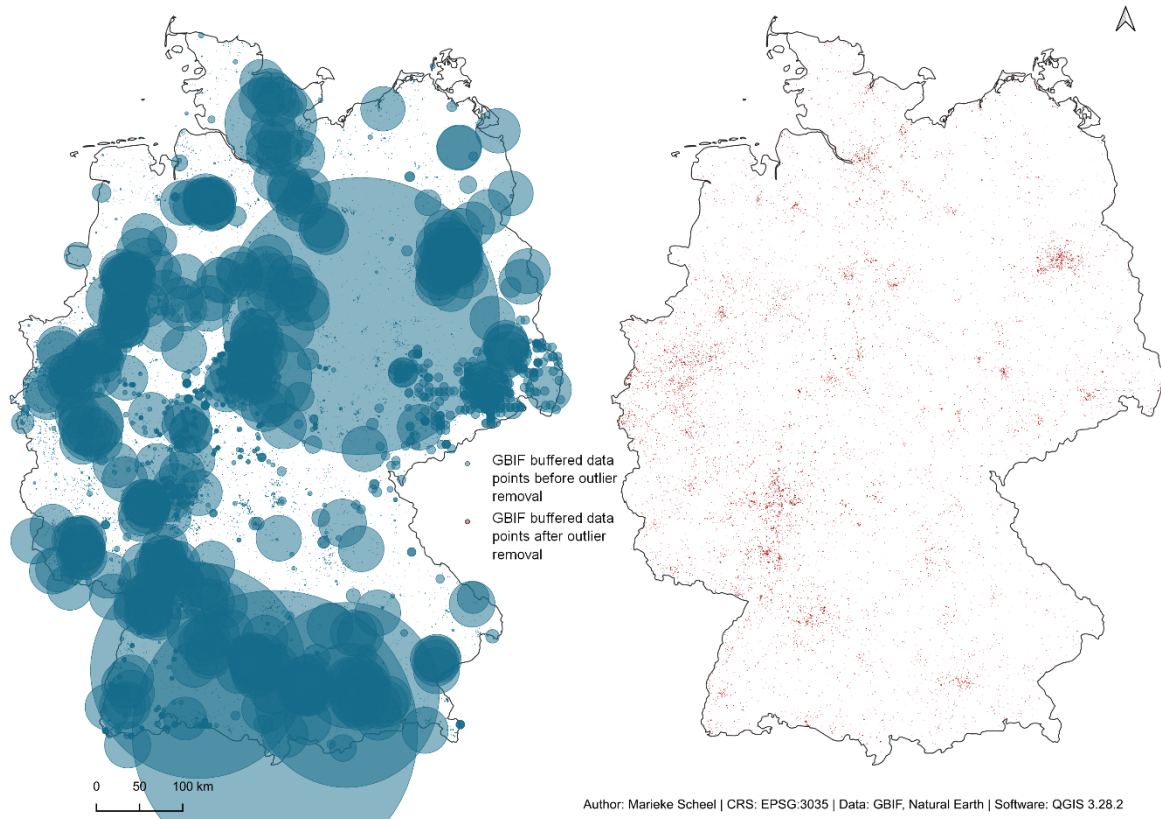


Fig. A1. GBIF data points buffered by their respective coordinate uncertainties before the outlier removal (left) and GBIF data points after outlier removal (right).

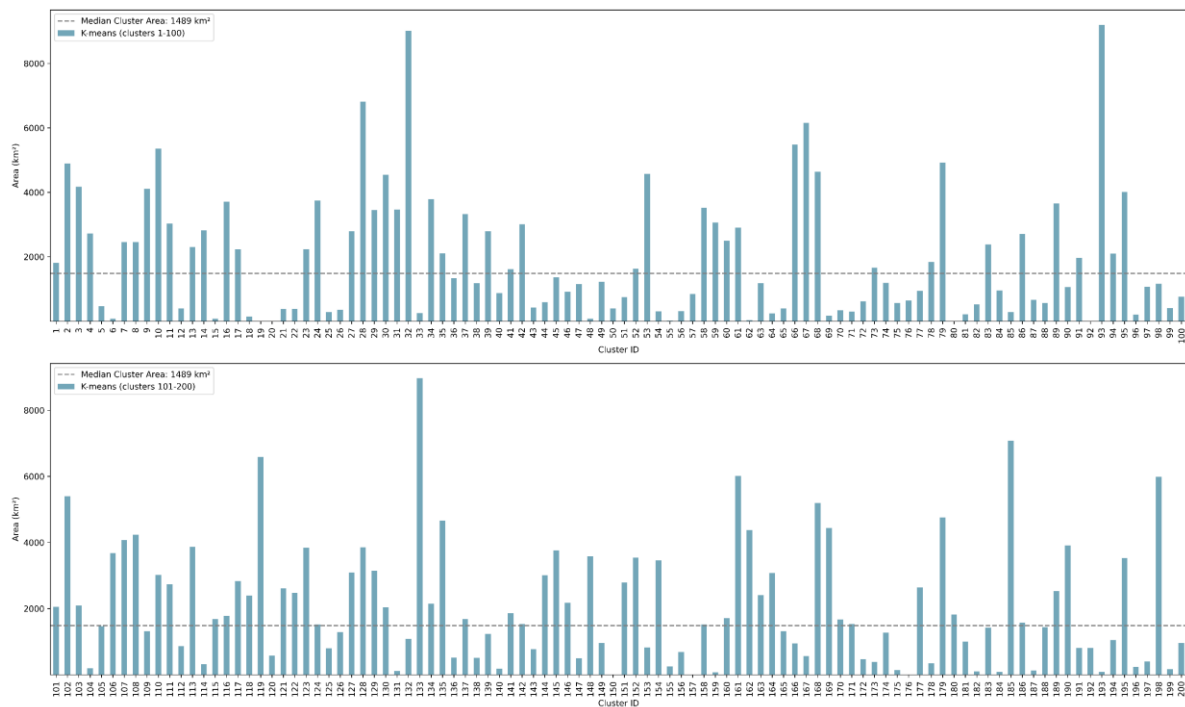


Fig. A2. Areas of the 200 K-means clusters. The dotted line shows the median cluster area.

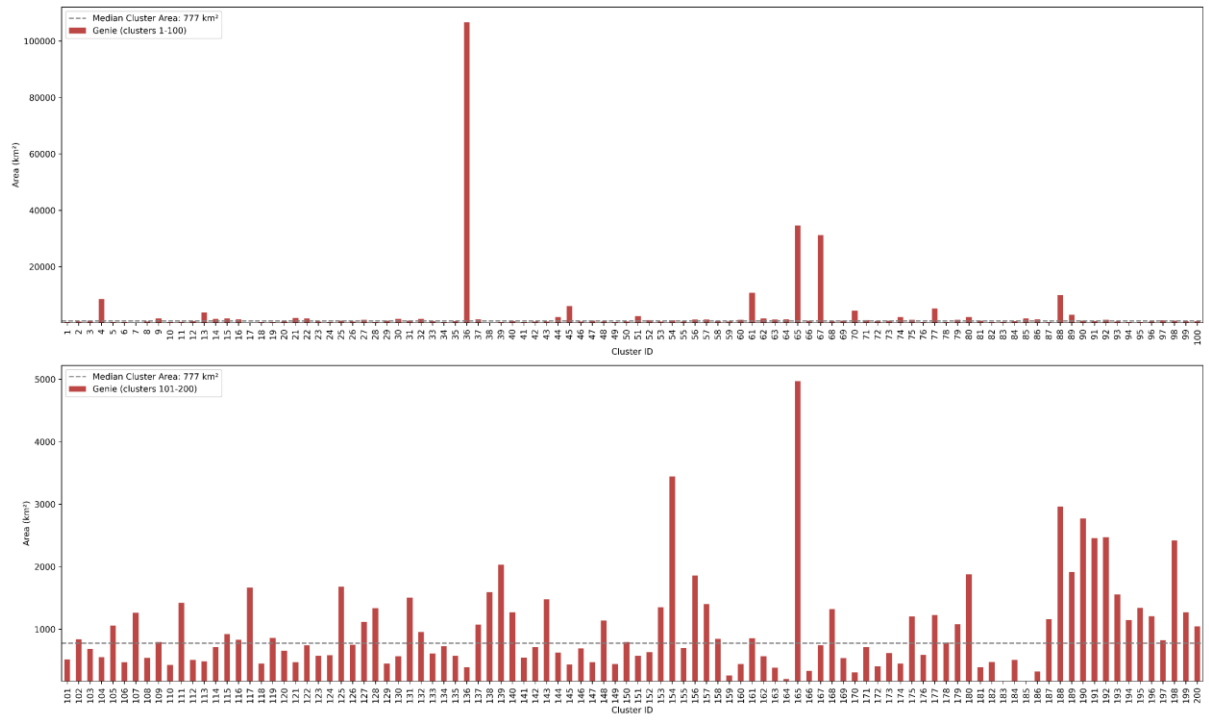


Fig. A3. Areas of the 200 Genie clusters. The dotted line shows the median cluster area.

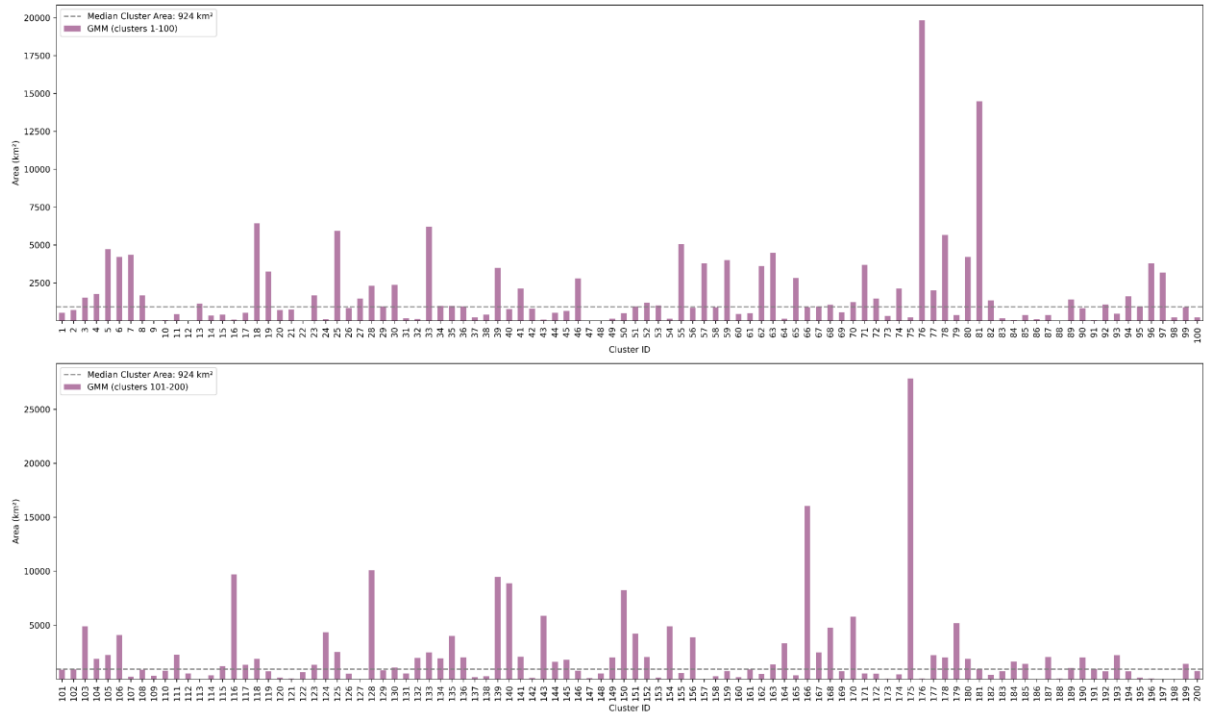


Fig. A4. Areas of the 200 GMM clusters. The dotted line shows the median cluster area.

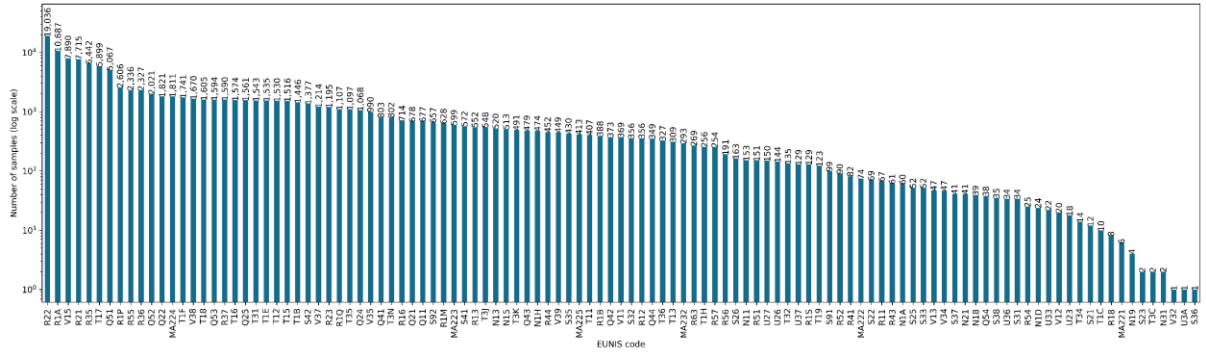


Fig. A5. Number of samples of the different EUNIS habitat classes before quality control.

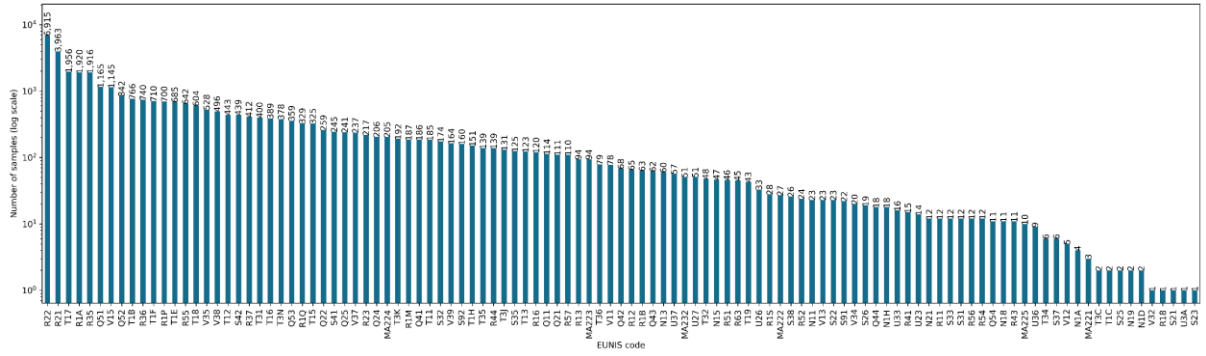


Fig. A6. Number of samples of the different EUNIS habitat classes after quality control.

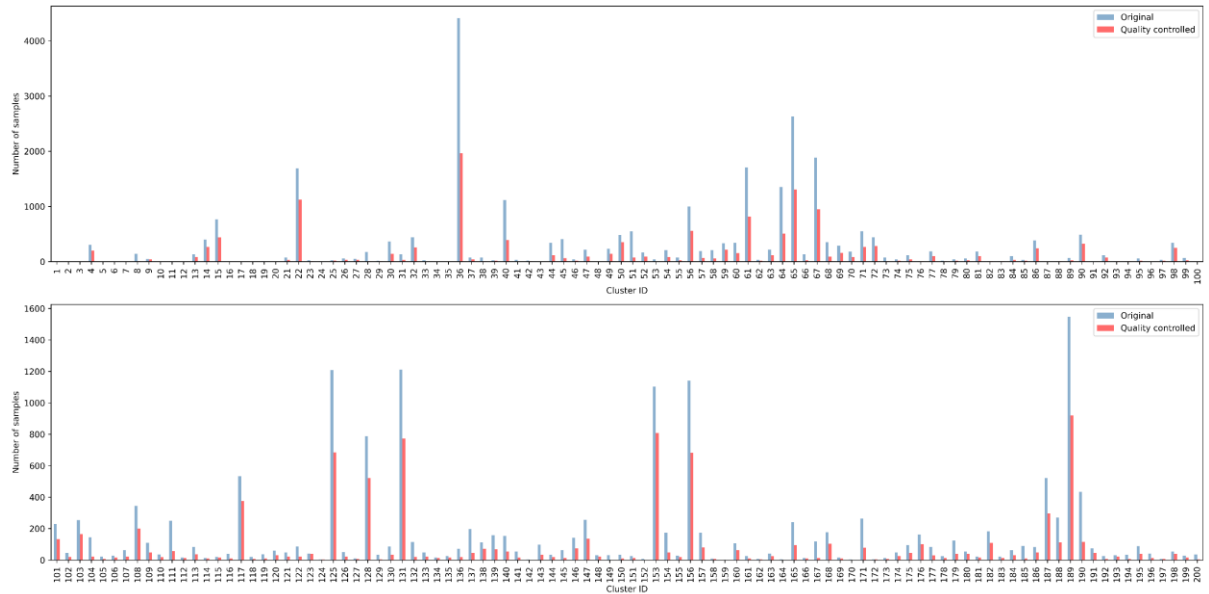


Fig. A7. The number of EUNIS samples within the different clusters before and after quality control.

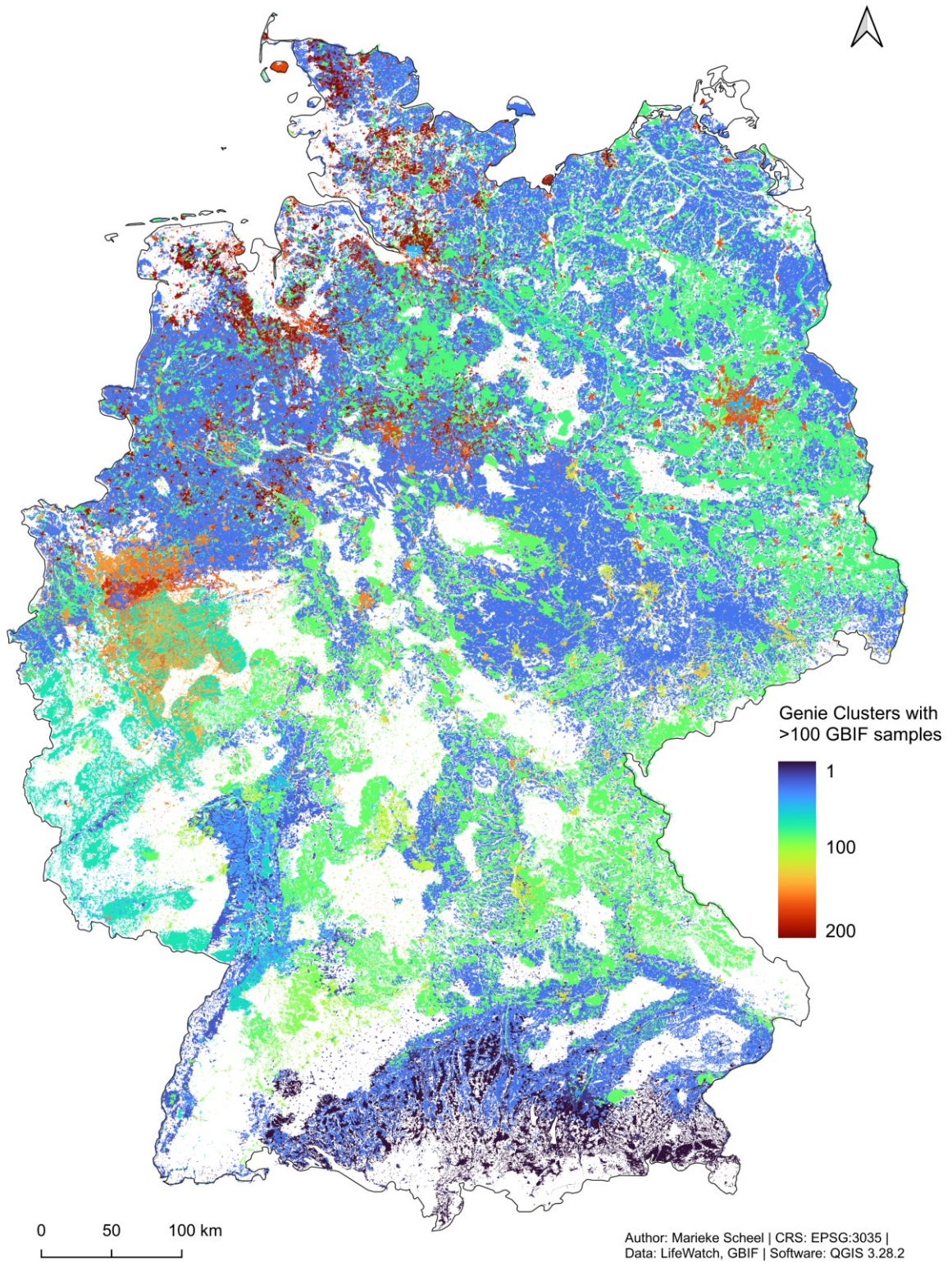


Fig. A8. The cluster IDs of all Genie clusters with >100 GBIF samples.

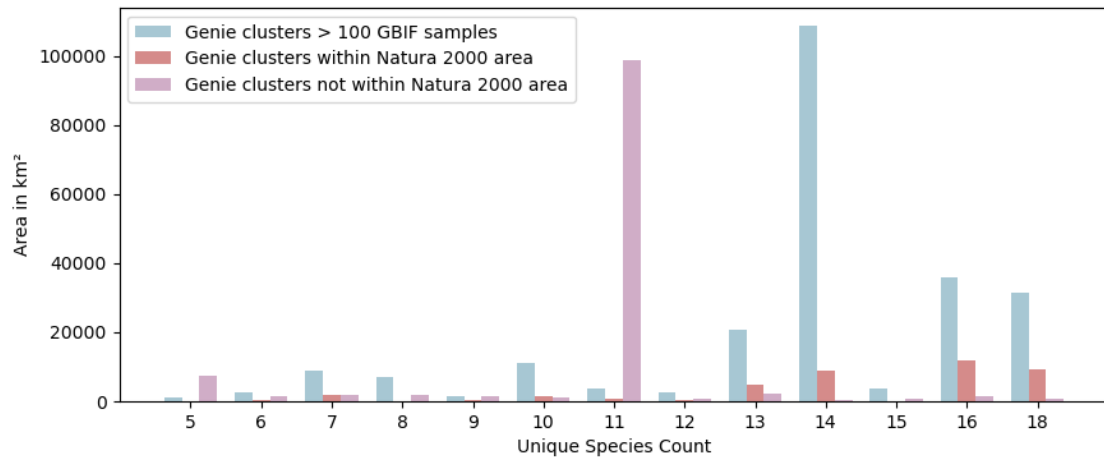


Fig. A9. Comparison of the area in km² by unique wasp species count for the whole area of the Genie clusters with >100 GBIF samples (blue), for the Genie clusters within Natura 2000 areas (red) and for the Genie clusters not within Natura 2000 areas (purple).

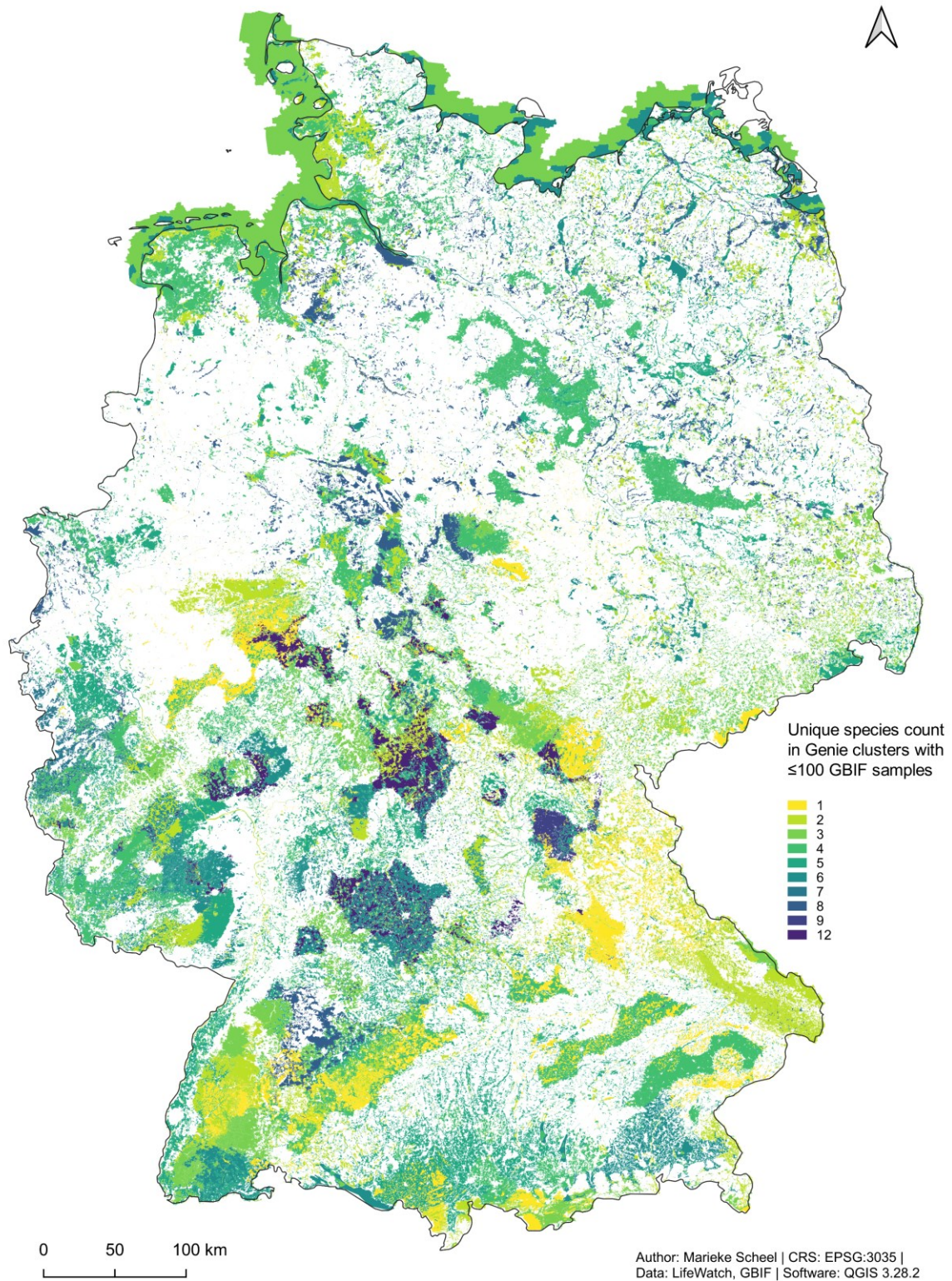


Fig. A10. Number of unique wasp species plotted for the 158 remaining clusters, each with ≤ 100 GBIF samples.

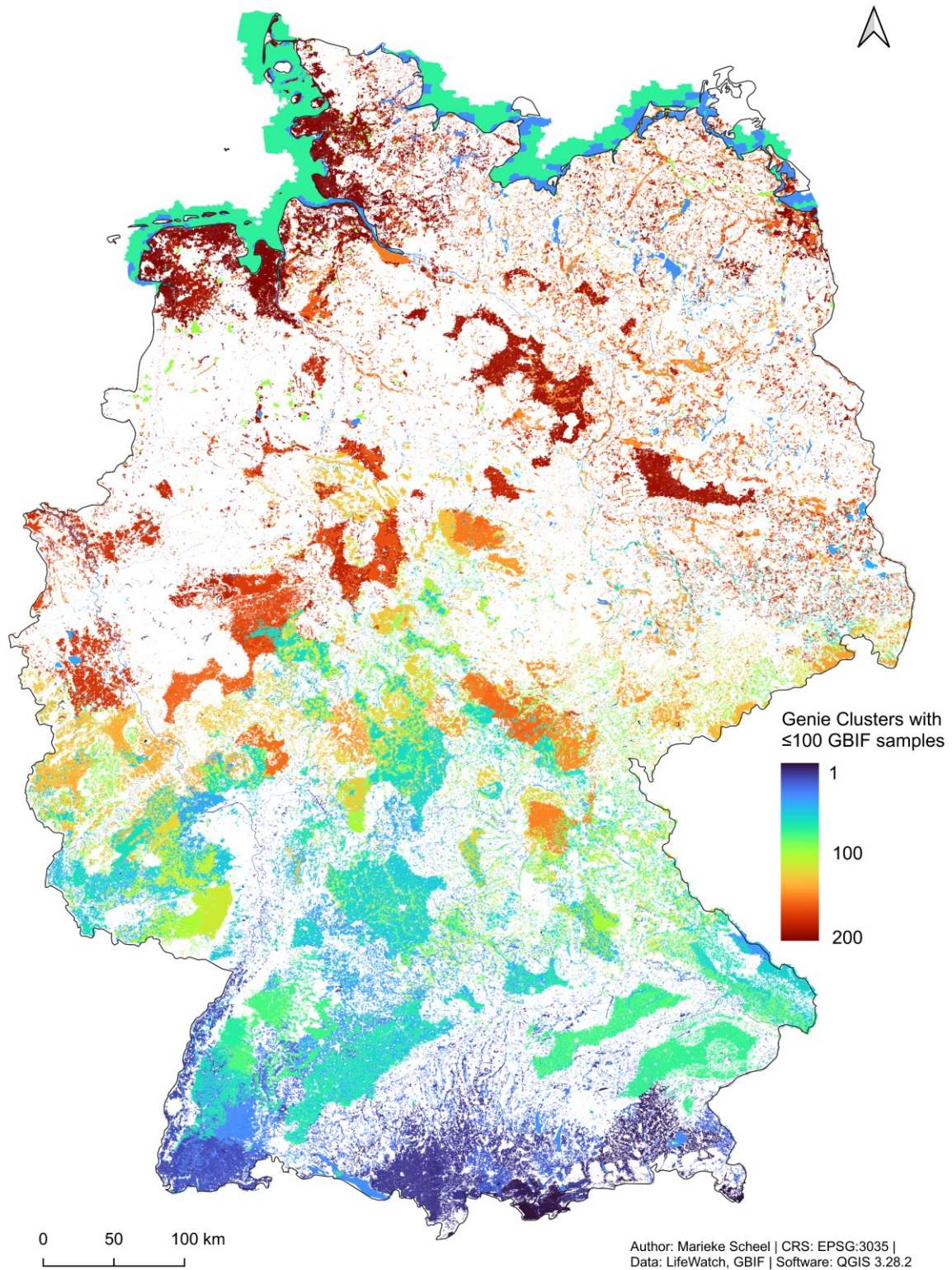


Fig. A11. The cluster IDs of all Genie clusters with ≤ 100 GBIF samples.

10.2 Tables

Table A1. Factors that determine the quality for an indicator group according to Brown (1991). Source: Brown (1991)

Desirable quality for an indicator group in ecology and biogeography
Taxonomically and ecologically highly diversified (many species in each locale or system)
Species have high ecological fidelity
Relatively sedentary
Species narrowly endemic, or if widespread, well differentiated (local or regional)
Taxonomically well known, easy to identify
Well studied (genetics, behaviour, biochemistry, ecology, biogeography)
Abundant, non-furtive, easy to find in field
Damped fluctuations (always present)
Easy to obtain large random samples of species and variation
Functionally important in ecosystem
Response to disturbance predictable, rapid, sensitive, analysable and linear
Associates closely with and indicates other species and specific resources
Total value as indicator (in 24)

Table A2. LifeWatch biophysical variable names and their explanation. Source: LifeWatch database, non-validated, pre-released version, validation will be provided in September 2024

Variable Names	Explanation
Climatic Variables	
AnM_T	Mean annual temperature in °C
AnVar_T	Annual variation of the temperature in °C (standard deviation)
AnRg_T	Temperature annual range in °C
MaxWarmM_T	Maximum temperature of the warmest month in °C
MinColdM_T	Minimum temperature of the coldest month in °C
WetM_P	Precipitation of the wettest month in mm
DryM_P	Precipitation of the driest month in mm
AnTot_P	Sum of annual rainfall in mm
AnVar_P	Annual variation of the precipitation (coefficient of variation)
GDD_5deg	Mean annual cumulated growing degree days (above 5°C)
An_PotETP	Annual mean potential evapotranspiration in mm day ⁻¹
AnM_SolR	Annual mean solar radiation in W m ⁻²
AnVar_SolR	Annual variation of the solar radiation in W m ⁻²
Snow_len	Average snow duration, in weeks
Topographic variables	
Elev_mean	Average of the elevation in the ecotope in m
Elev_min	Minimum of the elevation in the ecotope in m

Elev_max	Maximum of the elevation in the ecotope in m
SunSpring	Potential incident light energy in Wm^{-2} for the first day of spring, measure at the centre of the polygon. This variable integrates the clear sky sun energy reaching the ground during 24h (measure every hour) on March 21. Slope is measured in the centre of 4 pixels. Topographic shadows are taken into account.
Slp_mean	Mean percentage of slope of the ecotope. Slope is derived from 1-m or 5-m LIDAR data resampled at 10 m with Lancsoz method
Slp_max	Maximum percentage of slope of the ecotope. Slope is derived from 1-m or 5-m LIDAR data resampled at 10 m with Lancsoz method
TPI_250	Topographic position index, that is the mean relative position of the ecotope in a 250 m radius. This value range is negative when the ecotope is in a valley and positive when it is on a crest
TPI_1km	Topographic position index, that is the mean relative position of the ecotope in a 1000 m radius. This value range is negative when the ecotope is in a valley and positive when it is on a crest
TPI_2km	Topographic position index, that is the mean relative position of the ecotope in a 2000 m radius. This value range is negative when the ecotope is in a valley and positive when it is on a crest
Proportions inside ecotopes	
BroadLV (Alias Broadleaved trees)	Broadleaved trees (angiosperms), located in forests or other land use (parks, orchards...) with a height above 3m
NeedLV (Alias needleleaved trees)	Coniferous trees (gymnosperms) located in forests or other land use (hedges, gardens...) with a height above 3m
Sparse	Proportion of sparse vegetation
Shrub	Proportion of shrubs
Plowed (Alias ploughed herbaceous cover)	Arable lands (annual crops and temporary herbaceous cover)
MGramin (Alias permanent monospecific graminoid cover)	Permanent monospecific graminoid cover resulting from intensive land uses such as intensive farming, gardens, leisure
NOpen (Alias open area with relatively dry soils)	Permanent herbaceous cover mixed with other non-ligneous vegetation. This class covers a large number of potential biodiversity hotspots such as heathlands or extensive grasslands
WOpen (Alias open area with humid soils)	Flooded herbaceous cover mixed with other non-ligneous vegetation. This class covers a large number of potential biodiversity hotspots such wetlands and peatlands
Water (Alias permanent water bodies)	Permanent water bodies
Bare (Alias bare soils)	Bare soils or soils sparsely covered by vegetation (<15%), mainly quarries in Wallonia
Sealed (Alias sealed surface)	Artificial sealed surfaces at ground level
Building (Alias sealed surface)	Above ground buildings and other man-made constructions
SnowIce	Proportions of permanent snow or ice
Contextual land cover proportions	
BroadLV_1km or 2km	Broadleaved trees (angiosperms)
NeedLV_1km or 2km	Coniferous (gymnosperms)
Plowed_1km or 2km	Land being ploughed during the year
MGramin_1km or 2km	Permanent monospecific herbaceous cover

NOpen_1km or 2km	Permanent mix of herbaceous cover and other non-ligneous vegetation or small shrub
WOpen_1km or 2km	Permanent inundated mixture of herbaceous cover and other non-ligneous vegetation or small shrub
Water_1km or 2km	Open water bodies
Bare_250 or 500 or 1km	Permanent bare soil
Sealed_1km or 2km	Artificial sealed surfaces at ground level
Building_1km or 2km	Above ground buildings and other man-made constructions
SnowIce_1km or 2km	Proportions of permanent snow or ice
Soil properties	
cec_5	Cation Exchange Capacity
soc_5	Soil Organic Content
nitrog_5	Soil nitrogen
silt_5	Silt percentage
sand_5	Sand percentage
clay_5	Clay percentage
bdod_5	Bulk density of fine earth fraction
cfvo_5	Volumetric fraction of coarse fragments
phh2o	pH of the soil solution
Distance metrics	
dme_*	Mean distance to landcover class
dmi_*	Minimum distance to landcover class
dma_*	Maximum distance to landcover class
Categorical variables	
Majority	Majority land cover based on the 10 m LifeWatch land cover (http://maps.elie.ucl.ac.be/download/ecotopes_documentation_en_v4-16.pdf , chapter 4.4) based on the analysis and Sentinel-2 data (composite and NDVI time series) fused with ancillary information from several existing data sources and models.
LCCsb	Adaptation of the CCI legend, which is based on the land cover classification system (http://maps.elie.ucl.ac.be/download/ecotopes_documentation_en_v4-16.pdf , chapter 4.4)
ecosystem31Mj	Categorical variable based on crop type map of JRC 2018 (ecosystem map from EEA and local climate zones), Mj indicates majority
ecosystem31Dv	Categorical variable based on crop type map of JRC 2018 (ecosystem map from EEA and local climate zones), Dv indicates diversity
croptypeMj	Categorical variable based on crop type map of JRC 2018 (ecosystem map from EEA and local climate zones), Mj indicates majority
croptypeDv	Categorical variable based on crop type map of JRC 2018 (ecosystem map from EEA and local climate zones), Dv indicates diversity
IczMj	Categorical variable based on crop type map of JRC 2018 (ecosystem map from EEA and local climate zones), Mj indicates majority
IczDv	Categorical variable based on crop type map of JRC 2018 (ecosystem map from EEA and local climate zones), Dv indicates diversity
Other	
NightLi	Artificial light during the night from remote sensing (rescaled in 0-255)

Table A3. LifeWatch variable groups and their data sources. Source: LifeWatch database, non-validated, pre-released version, validation will be provided in September 2024

Variable Groups	Data source
Climatic Variables	Bioclimatic variables were extracted from the high-resolution future climate data for species distribution models in Europe (https://zenodo.org/record/3694065). They represent historical climate between 1971 and 2005.
Topographic variables	The source of elevation information for computing the topographic variables is the 2013-2014 LIDAR dataset of the Walloon region (0.8 pts/m ²). It was smoothed and resampled at 10 m resolution, then combined with the 2015 LIDAR dataset of the Flemish region in order to cover the full extent of Belgium. A linear combination of the two DEM was used in the overlap region in order to avoid sharp transitions.
Proportions inside ecotopes	The ecotopes are automatically delineated based on the land cover class probabilities and topographic indices. The land cover proportion is then computed based on the 10 m LifeWatch land cover layer described in http://maps.elie.ucl.ac.be/download/ecotopes_documentation_en_v4-16.pdf , chapter 4.4.
Contextual land cover proportions	Contextual information is based on the 10-m land cover information from 2018. Two circular neighbourhoods are used: the first has a radius of 100 pixels, the second has a radius of 200 pixels. The size of the radius is mentioned in the field name (1km for 100 pixels and 2km for 200 pixels). The average of the proportion of each land cover is computed for each ecopatch and rescaled between 0 and 1000.
Soil properties	Soil properties are derived by averaging the values of the soil-GRID dataset for the topsoil (0-5 cm) inside each ecopatch. More information about the variables is available on (https://soilgrids.org/).
Distance metrics	Distances are computed to all land cover classes from the 10 m LifeWatch Landcover 2018 layer, then the statistics are computed inside each polygon.

Table A4. Biophysical variables from the LifeWatch dataset that were removed due to a VIF > 10 with the respective VIF values.

Removed biophysical variables	Variance Inflation Factor	Removed biophysical variables	Variance Inflation Factor	Removed biophysical variables	Variance Inflation Factor
Plowed	210462	dme_NOpen	167	WetM_P	38
Plowed_1km	17972	AnTot_P	147	elev_min	34
dme_Sparse	17583	dma_Plowed	107	Sealed_2km	27
Plowed_2km	11848	An_PotETP	103	dma_NeedILV	24
MinColdM_T	2816	elev_max	84	NeedILV_1km	21
dme_WOpen	2765	dme_MGramin	80	dma_NOpen	19
dma_Sparse	1575	dme_SnowIce	73	Bare_1km	17
elev_mean	1272	dme_BroadLV	69	NOpen_1km	15
dme_Plowed	1211	dme_Water	58	BroadLV_1km	12
dme_Shrub	353	dme_Building	53	AnRg_T	12
dmi_WOpen	283	sand_5	49	slp_mean	12

dme_NeedLV	257	dme_Bare	45	Shrub_1km	11
MaxWarmM_T	216	dma_Shrub	39		

Table A5. Biophysical variables of the LifeWatch dataset remaining after removing 38 variables using VIF.

Remaining variables					
Water	ecosystem31Mj	SnowIce_2km	AnVar_T	dma_Bare	dma_BroadLV
SnowIce	ecosystem31Dv	Bare_2km	Snow_len	dme_Sealed	dmi_Shrub
Bare	croptypeMj	Building_2km	DryM_P	dmi_Sealed	TPI_1km
Sealed	croptypeDv	MGramin_2km	AnVar_P	dma_Sealed	clay_5
Building	lczMj	NOpen_2km	AnVar_SolR	dmi_Building	silt_5
MGramin	lczDv	WOpen_2km	AnM_SolR	dma_Building	soc_5
NOpen	Water_1km	Sparse_2km	GDD_5deg	dmi_Plowed	cec_5
WOpen	SnowIce_1km	NeedLV_2km	slp_min	dmi_MGramin	bdod_5
Sparse	Sealed_1km	BroadLV_2km	slp_max	dma_MGramin	cfvo_5
NeedLV	Building_1km	Shrub_2km	dmi_Water	dmi_NOpen	nitrog_5
BroadLV	MGramin_1km	SunSpring	dma_Water	dma_WOpen	phh2o
Shrub	WOpen_1km	TPI_250m	dmi_SnowIce	dmi_Sparse	
Majority	Sparse_1km	NightLi	dma_SnowIce	dmi_NeedLV	
LCCSb	Water_2km	AnM_T	dmi_Bare	dmi_BroadLV	

Table A6. EUNIS habitat names and their corresponding codes.

Code	Habitat	Code	Habitat
MA221	Atlantic saltmarsh driftlines	S21	Subarctic and alpine dwarf Salix scrub
MA222	Atlantic upper saltmarshes	S22	Alpine and subalpine ericoid heath
MA223	Atlantic upper-mid saltmarshes and saline and brackish reed, rush and sedge beds	S23	Alpine and subalpine Juniperus scrub
MA224	Atlantic mid-low saltmarshes	S25	Subalpine and subarctic deciduous scrub
MA225	Atlantic pioneer saltmarshes	S26	Subalpine Pinus mugo scrub
MA232	Baltic coastal meadow	S31	Lowland to montane temperate and submediterranean Juniperus scrub
N11	Atlantic, Baltic and Arctic sand beach	S32	Temperate Rubus scrub
N13	Atlantic and Baltic shifting coastal dune	S33	Lowland to montane temperate and submediterranean genistoid scrub
N15	Atlantic and Baltic coastal dune grassland (grey dune)	S35	Temperate and submediterranean thorn scrub
N18	Atlantic and Baltic coastal Empetrum heath	S36	Low steppic scrub
N19	Atlantic coastal Calluna and Ulex heath	S37	Corylus avellana scrub
N1A	Atlantic and Baltic coastal dune scrub	S38	Temperate forest clearing scrub
N1D	Atlantic and Baltic broad-leaved coastal dune forest	S41	Wet heath
N1H	Atlantic and Baltic moist and wet dune slack	S42	Dry heath
N21	Atlantic, Baltic and Arctic coastal shingle beach	S91	Temperate riparian scrub
N31	Atlantic and Baltic rocky sea cliff and shore	S92	Salix fen scrub

Q11	Raised bog	T11	Temperate Salix and Populus riparian forest
Q21	Oceanic valley mire	T12	Alnus glutinosa-Alnus incana forest on riparian and mineral soils
Q22	Poor fen	T13	Temperate hardwood riparian forest
Q24	Intermediate fen and soft-water spring mire	T15	Broadleaved swamp forest on non-acid peat
Q25	Non-calcareous quaking mire	T16	Broadleaved mire forest on acid peat
Q41	Alkaline, calcareous, carbonate-rich small-sedge spring fen	T17	Fagus forest on non-acid soils
Q42	Extremely rich moss-sedge fen	T18	Fagus forest on acid soils
Q43	Tall-sedge base-rich fen	T19	Temperate and submediterranean thermophilous deciduous forest
Q44	Calcareous quaking mire	T1B	Acidophilous Quercus forest
Q51	Tall-helophyte bed	T1C	Temperate and boreal mountain Betula and Populus tremula forest on mineral soils
Q52	Small-helophyte bed	T1E	Carpinus and Quercus mesic deciduous forest
Q53	Tall-sedge bed	T1F	Ravine forest
Q54	Inland saline or brackish helophyte bed	T1H	Broadleaved deciduous plantation of non site-native trees
R11	Pannonian and Pontic sandy steppe	T31	Temperate mountain Picea forest
R12	Cryptogam- and annual-dominated vegetation on siliceous rock outcrops	T32	Temperate mountain Abies forest
R13	Cryptogam- and annual-dominated vegetation on calcareous and ultramafic rock outcrops	T34	Temperate subalpine Larix, Pinus cembra and Pinus uncinata forest
R16	Perennial rocky grassland of Central and South-Eastern Europe	T35	Temperate continental Pinus sylvestris forest
R18	Perennial rocky calcareous grassland of subatlantic-submediterranean Europe	T36	Temperate and submediterranean montane Pinus sylvestris-Pinus nigra forest
R1A	Semi-dry perennial calcareous grassland (meadow steppe)	T3C	Taxus baccata forest
R1B	Continental dry grassland (true steppe)	T3J	Pinus and Larix mire forest
R1M	Lowland to montane, dry to mesic grassland usually dominated by Nardus stricta	T3K	Picea mire forest
R1P	Oceanic to subcontinental inland sand grassland on dry acid and neutral soils	T3N	Coniferous plantation of site-native trees
R1Q	Inland sanddrift and dune with siliceous grassland	U23	Temperate, lowland to montane siliceous scree
R1S	Heavy-metal grassland in Western and Central Europe	U26	Temperate high-mountain base-rich scree and moraine
R21	Mesic permanent pasture of lowlands and mountains	U27	Temperate, lowland to montane base-rich scree
R22	Low and medium altitude hay meadow	U33	Temperate, lowland to montane siliceous inland cliff
R23	Mountain hay meadow	U36	Temperate high-mountain base-rich inland cliff
R35	Moist or wet mesotrophic to eutrophic hay meadow	U37	Temperate, lowland to montane base-rich inland cliff
R36	Moist or wet mesotrophic to eutrophic pasture	U3A	Temperate ultramafic inland cliff
R37	Temperate and boreal moist or wet oligotrophic grassland	V11	Intensive unmixed crops

R41	Snow-bed vegetation	V12	Mixed crops of market gardens and horticulture
R43	Temperate acidophilous alpine grassland	V13	Arable land with unmixed crops grown by low-intensity agricultural methods
R44	Arctic-alpine calcareous grassland	V15	Bare tilled, fallow or recently abandoned arable land
R51	Thermophilous forest fringe of base-rich soils	V32	Mediterranean subnitrophilous annual grassland
R52	Forest fringe of acidic nutrient-poor soils	V34	Trampled xeric grassland with annuals
R54	Pteridium aquilinum vegetation	V35	Trampled mesophilous grassland with annuals
R55	Lowland moist or wet tall-herb and fern fringe	V37	Annual anthropogenic herbaceous vegetation
R56	Montane to subalpine moist or wet tall-herb and fern fringe	V38	Dry perennial anthropogenic herbaceous vegetation
R57	Herbaceous forest clearing vegetation	V39	Mesic perennial anthropogenic herbaceous vegetation
R63	Temperate inland salt marsh		

Abstract

The biodiversity is rapidly declining worldwide, and conservation and management projects are subsequently increasing in importance. Since perturbations of ecosystems are expected to increase in the future, it is of paramount interest to increase ecosystem resilience, which is largely affected by biodiversity. Therefore, this study aimed to assess the wasp biodiversity potential for different terrestrial ecosystems in Germany to establish a baseline for an ecosystem health evaluation. The biodiversity potential was computed for 200 ecosystem clusters created based on biophysical LifeWatch ecotope data using three different clustering algorithms, namely K-means, GMM and Genie. The clustering algorithms were evaluated against EUNIS habitat data, and the algorithm with the highest average purity score was selected for further analysis. Then, the number of unique wasp species was calculated for any cluster with >100 GBIF samples. Further, a health status was calculated for five cities in Germany to demonstrate the application of the biodiversity potential to assess ecosystem health. This study showed that the Genie clustering algorithm could form clusters with the highest agreement with the EUNIS habitat data. Interestingly, the clusters formed using Genie were indicated by the intrinsic quality metrics to be the least well-distinguished. Central and central-eastern Germany were identified as the regions with the highest wasp biodiversity potential, while central-western Germany had the lowest wasp biodiversity potential. Also, areas with a high wasp biodiversity potential exhibited a good agreement with Natura 2000 protection areas both within the five cities and within all of Germany, supporting the validity of the findings. Finally, this study introduces an inexpensive and fast method to establish a baseline for ecosystem health assessments. It can be applied to a wide range of ecosystems and is thus optimal for policymakers and researchers to quickly gain knowledge about the state of the ecosystems in a larger area. While the application of the biodiversity potential as a baseline requires further work, this study lays an important foundation.

Keywords: Biodiversity Potential, Unsupervised Clustering, Ecosystem Clustering, Genie, GMM, K-Means, GBIF, LifeWatch, Ecosystem Health