

Article

DBSCAN and GIE, Two Density-Based “Grid-Free” Methods for Finding Areas of Endemism: A Case Study of Flea Beetles (Coleoptera, Chrysomelidae) in the Afrotropical Region

Maurizio Biondi, Paola D’Alessandro, Walter De Simone  and Mattia Iannella * 

Department of Life, Health & Environmental Sciences, University of L’Aquila, Via Vetoio Coppito, 67100 L’Aquila, Italy; maurizio.biondi@univaq.it (M.B.); paola.dalessandro@univaq.it (P.D.); walter.desimone@graduate.univaq.it (W.D.S.)

* Correspondence: mattia.iannella@univaq.it

Simple Summary: Areas of endemism (AoEs) are one of the most important topics discussed in biogeography, considering that the analysis of areas of sympatry between endemic species is essential to understand species distribution patterns, reconstruct evolutionary events, regionalize biogeographical areas, and assess regions of high conservation concern. Here, we propose a workflow based on the application of a clustering-based algorithm to identify AoEs and compare it to another method, the Geographical Interpolation of Endemism, based on a kernel density approach. We apply this framework to the flea beetles of the whole sub-Saharan Africa, identifying several AoEs through both methods, but with differences in their delimitation, number and features of characteristic species, and surface. Considering that our proposed workflow can be applied to any territorial context and sets of endemic species, we also provide a GIS tool that implements all the steps into one single toolbox. The identification of AoEs, possibly facilitated by our approach, can provide useful spatial information when dealing with several biodiversity-related issues, even applied to practical conservation measures, such as protected areas management and landscape planning.

Abstract: Areas of endemism (AoEs) are a central area of research in biogeography. Different methods have been proposed for their identification in the literature. In this paper, a “grid-free” method based on the “Density-based spatial clustering of applications with noise” (DBSCAN) is here used for the first time to locate areas of endemism for species belonging to the beetle tribe Chrysomelidae, Galerucinae, Alticini in the Afrotropical Region. The DBSCAN is compared with the “Geographic Interpolation of Endemism” (GIE), another “grid-free” method based on a kernel density approach. DBSCAN and GIE both return largely overlapping results, detecting the same geographical locations for the AoEs, but with different delimitations, surfaces, and number of detected sinendemisms. The consensus maps obtained by GIE are in general less clearly delimited than the maps obtained by DBSCAN, but nevertheless allow us to evaluate the core of the AoEs more precisely, representing of the percentage levels of the overlap of the centroids. DBSCAN, on the other hand, appears to be faster and more sensitive in identifying the AoEs. To facilitate implementing the delimitation of the AoEs through the procedure proposed by us, a new tool named “CLUENDA” (specifically developed in GIS environment) is also made available.

Keywords: areas of endemism; density-based clustering; DBSCAN; GIE; Chrysomelidae; Afrotropical region; GIS analysis; ArcGIS Pro; Model Builder



Citation: Biondi, M.; D’Alessandro, P.; De Simone, W.; Iannella, M. DBSCAN and GIE, Two Density-Based “Grid-Free” Methods for Finding Areas of Endemism: A Case Study of Flea Beetles (Coleoptera, Chrysomelidae) in the Afrotropical Region. *Insects* **2021**, *12*, 1115. <https://doi.org/10.3390/insects12121115>

Academic Editor: Stewart B. Peck

Received: 17 November 2021

Accepted: 10 December 2021

Published: 13 December 2021

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Endemisms are one of the most important features in the distribution of biodiversity on Earth, and their identification is essential to define the biological value of an area and its intrinsic conservation requirements [1–3]. The endemism of a species is a result of both ecological and historical factors [4,5]; while ecology explains how biotic and abiotic

factors can shape a species' range, a historical reconstruction can uncover which geological and evolutionary events contributed to confine a species to its current distribution [6]. Areas of endemism are a widely explored concept in biogeography, as they are identified by the spatial overlapping of two or more endemic species [7–11]. In fact, a non-random distributional congruence between taxa can identify common evolutionary biogeographical processes [12]. Throughout geological time, an assemblage of endemic species sharing a common space might have responded differently to the same ecological factors: this is the reason why each area of endemism usually has fuzzy edges [11,13], making it more difficult for biogeographers to define its exact borders.

There is still a high level of disagreement over what the areas of endemism actually represent, and what the correct way to identify them is [7,14]. Despite this, all authors agree that these spatial units are dynamic entities, representing a current snapshot of the evolution of species, or groups of species, sharing a common history [15]. Therefore, it is essential to identify areas of endemism not only to infer the history of biogeographical units [16,17], but also to lay the groundwork for suitable conservation plans within a specific study area [3,18]. Different methods have been proposed to detect areas of endemism: Parsimony Analysis of Endemicity (PAE) [8,19,20], Cladistic Analysis of Distributions and Endemism (CADE) [21], Endemicity Analysis with Optimality Criterion [22–24], and Network Analysis [25] are only a few examples of the numerous techniques suggested throughout the years to identify and analyse these historical and ecological units.

In this paper, two different approaches to infer areas of endemism independent of grid cells ("grid-free") are compared and discussed: the Density-Based Spatial Clustering of Application with Noise (DBSCAN), applied here for the first time to identify areas of endemism, and the Geographic Interpolation of Endemism (GIE), recently proposed by Oliveira et al. [26]. Both these methods, instead of dividing the study area into grid cells, use the species' distribution points as raw data to detect areas of endemism. DBSCAN, proposed by Ester et al. [27], is a density-based clustering algorithm that works on the assumption that clusters are higher-density regions in space separated by regions of lower density (noise). The DBSCAN measure of density uses the distances between data points and applies local or global density criteria to separate out clusters from noises, represented by the not allocated points. This approach does not rely on a pre-defined number of clusters, such as the several variants of K-Means Clustering or Fuzzy C-means and can be made flexible for different shapes and varying densities of points.

GIE, instead, is based on a modified kernel interpolation function [26]. The kernel estimator has been widely used in ecology to estimate the density of species occurrence and consequently draw species range probability maps [28]. With this method, an area of influence is defined for each species range, and its radius is used to categorize species into different classes sharing areas of influence of similar size. A Gaussian function is applied for each occurrence, obtaining a homogeneous circular area of influence in which the function values decrease when moving away from the occurrence record (which has the highest value). Then, a kernel density function (an algorithm that expresses the spatial density of features in a geographic space) calculates the occurrence of endemism by identifying the overlap between different areas of influence, weighted by the degree of overlapping [29].

DBSCAN and GIE are here applied to a database of flea beetle species (Chrysomelidae, Galerucinae, Alticini) endemic to the Afrotropical region. Due to their high number of species and habitat specificity, often entailing a restricted geographic distribution, the insects are generally considered adequate candidates for biogeographical studies on endemism [30]. The leaf beetles (Coleoptera Chrysomelidae) represent a large portion of the herbivorous insect fauna for many habitats [31–33] and are considered a useful tool to analyse insect community structure [34–36]. In particular, the Alticini are a tribe of Chrysomelidae comprising over 534 genera and about 8000 species [37,38], occurring all over the world. Members of this tribe are commonly defined as flea beetles because of the presence of a metafemoral extensor tendon that enables them to jump [39,40]. Adult and larval stages mainly feed on stems, leaves, or roots, although rarely on flowers, of

almost all the higher plant families, generally with high levels of specialization and in different environments [31,41–43]. The Afrotropical flea beetle fauna includes about 1600 known species in 103 genera, of which over 80% are endemic to sub-Saharan Africa and/or Madagascar [44–46].

2. Materials and Methods

2.1. Study Area and Species Database

The study area (Figure 1) comprises the Afrotropical region as defined by Udvardy and Udvardy [47]. The analyses were performed on a dataset including 337 well-known species of flea beetles (Chrysomelidae, Galerucinae, Alticini) occurring in the Afrotropical region, with a total of 3296 records of presence. Distribution data (database available from the authors upon request) were collected from peer-reviewed literature (Supplementary File S1), and the species identification of every specimen was confirmed by the authors (MB and PD). Geographic coordinates possibly missing were retrieved using Google Earth.

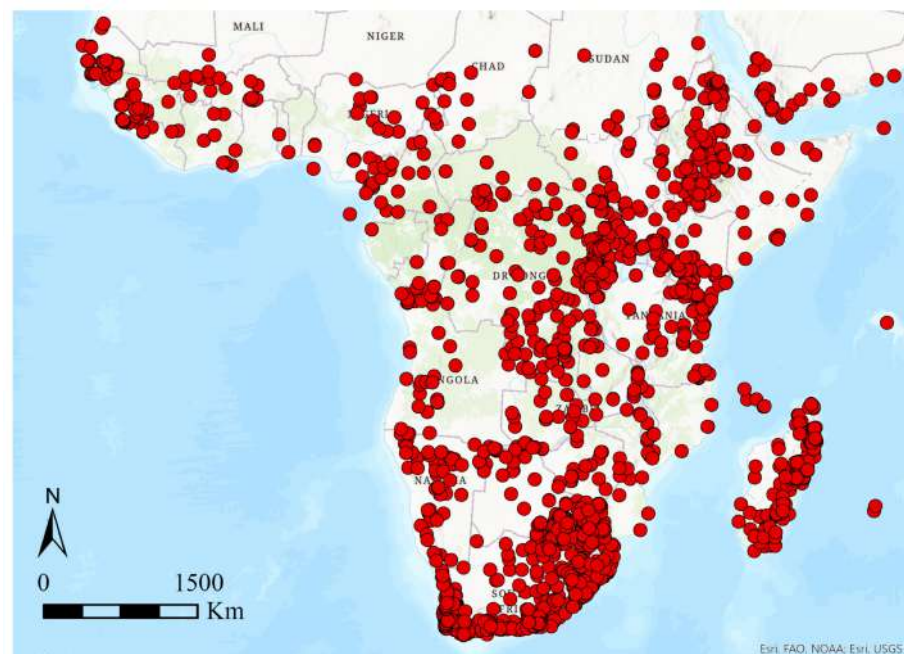


Figure 1. Study area and occurrence localities used for analysis (see text).

2.2. Abbreviations Used

AMM: Amber Mountain; ANR: Antananarivo region; AoE: area of endemism; BER: Betsiboka region; DBSCAN: Density-Based Spatial Clustering of Application with Noise; dc: maximum distance between the centroids; dd: maximum diameter of the distribution of the species; DKM: Drakensberg Mountains; GIE: Geographical Interpolation of Endemism; KAR: Katanga region; KIL: Kilimanjaro region; KLR: Kivu Lake region; KWN: KwaZulu-Natal; LI: Limpopo; MP: Mpumalanga; WCP: Western Cape Province.

2.3. Geographical Interpolation of Endemism (GIE)

This analysis was performed using the tool “GIE” in the BioDinamica package [48]. The maximum value of centroid distance was chosen as a parameter to define species classes. To compare the results of GIE with those of DBSCAN, the species database was divided into 3 classes, with the first class (Class 1) having a radius of the distance between centroids equal to or less than 100 km, the second one (Class 2) equal to or less than 150 km, and the third one (Class 3) equal to or less than 200 km. The analyses of the three classes considered (Class 1, Class 2, and Class 3) were performed separately due to unclear results obtained whether a single analysis including the three classes altogether is performed.

2.4. Density-Based Spatial Clustering of Application with Noise (DBSCAN)

As also stated in the Introduction, it is important to highlight that other existing clustering methods, such as K-means and its variants, or Fuzzy C-means, require setting an a priori number of clusters to carry out the analysis. Thus, these algorithms are not indicated in our case, considering that the number of endemic areas should be a result of an appropriate analysis, and not a parameter pre-determined by the operator.

The DBSCAN algorithm was preferred to HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [49], because this last one does not select clusters based on a global “epsilon” threshold (distance between centroids in our case) but creates a hierarchy for all possible epsilon values and thus only requires “minPts” (number of sinendemisms in our case) as single input parameter. The epsilon parameter used by DBSCAN is instead useful to discover all the clusters (AoEs) of variable densities with a definite value of epsilon (distance between centroids).

This analysis was performed using the tool Density-based Clustering implemented in the Spatial Statistics toolset of ArcGIS Pro 2.8. For ease of reference, in the workflow reported after, we name as “sp” (species name), “x” (longitude), and “y” (latitude) the fields of the input table hosting the species’ occurrence data; during the steps proposed, dummy names also will be assigned to intermediate files.

Step 1: Implementation of the dataset

Set the fields of a table in the order: sp, x, y; longitude and latitude are reported in decimal degrees. Then, load and display the occurrences (points) from the table in GIS environment.

Step 2: Selection of species

Use, as the first algorithm, the tool “Minimum Bounding Geometry” in ArcGIS Pro. It is necessary to implement the whole process, setting Geometry type = Circle, Group option = List and Group Field = sp (to group the output of this tool based on the species); furthermore, check the “Add geometry characteristics as attributes to output”. Then, the output features with a Minimum Bounding Geometry value (MBG_Diameter) equal to or less than the chosen diameter (in our case: ≤ 1.00 for species with a maximum distribution range width equal to or smaller than 100 km; ≤ 3.00 for species at 300 km; ≤ 5.00 for species at 500 km) were selected and exported to new files, named, for instance, “species_100 km”, “species_300 km”, or “species_500 km”, respectively. This step permits creating three selections of species with a distribution range width limited to 100, 300, or 500 km, respectively, avoiding the chance of including species with broad or disjunct distributions, which could cause a disturbing factor in the analysis.

Step 3: Calculation of the centroids

Calculate the centroids of the files obtained through the previous distribution range filtering (i.e., “species_100 km”) and convert this information, together with the corresponding species name, to a point-geometry file named, for instance, “centroids_100 km”. Apply the Density-based Clustering tool, using the different distance between centroids (“Defined Distance” = 100, 150, 200 km, similar to the distances used in GIE), with the option “Minimum Features” ≥ 2 , depending on the number of sinendemisms one wants to consider (five in our case). This parameter could also be used “backwards”, starting from a high number of sinendemism species to infer the highest number of sinendemisms by which the areas of endemism start to form. Then, use the Intersect tool between the DBSCAN file (i.e., DBSCAN_100 km_50 km_5 sin) and the corresponding centroid data used for the analysis (i.e., centroids_100 km), to obtain the information about the names of the species included in the clusters identified by DBSCAN.

Step 4: Developing Areas of endemism

As the last step, after having to discard in the attribute table all records reporting the –1 value (noise species not allocated into a specific cluster), use the tool “Create buffer” to definite the areas of endemism around the clusters obtained by the DBSCAN, setting the following parameters: Buffer Type = Distance, Buffer Distance = 80 km (in our case,

but this distance is depending by the scale of the referring area), Dissolve Option = List, Dissolve Fields = Cluster ID (to define the buffers based on the clusters).

2.5. The CLUENDA Tool

To implement the identification of areas of endemism using DBSCAN as described above, a new tool was developed in GIS environment through the Model Builder of ArcGIS Pro. The tool, named CLUENDA (Figure 2 and Supplementary File S2), aims to speed up data processing and ensure the comparability of maps of endemism areas obtained from species databases. The framework combines several geoprocessing tools in the ArcGIS Pro 2.8 environment to generate the necessary factors to identify the areas based on the clusters of sinendemisms. ArcGIS Pro Model Builder combines several GIS operations and runs these modules on a single spatial dataset (point features). Each model consists of three fundamental elements: the input parameters, the geoprocessing tools, and the output data. Model parameters are model-specific inputs that must be user-defined (e.g., species occurrences, to which a geographic projection should be applied; instead, chordal distances will be used). The geoprocessing tools produce output data in a defined sequence using the input datasets until areas with clustered sinendemism species are obtained. With the CLUENDA tool, we have mapped the areas of endemism inferred based on: (a) three different selections of species with maximum diameter of the distance to 100, 300 and 500 km, respectively; (b) three different distances between centroids up to 100, 150, and 200 km, respectively.

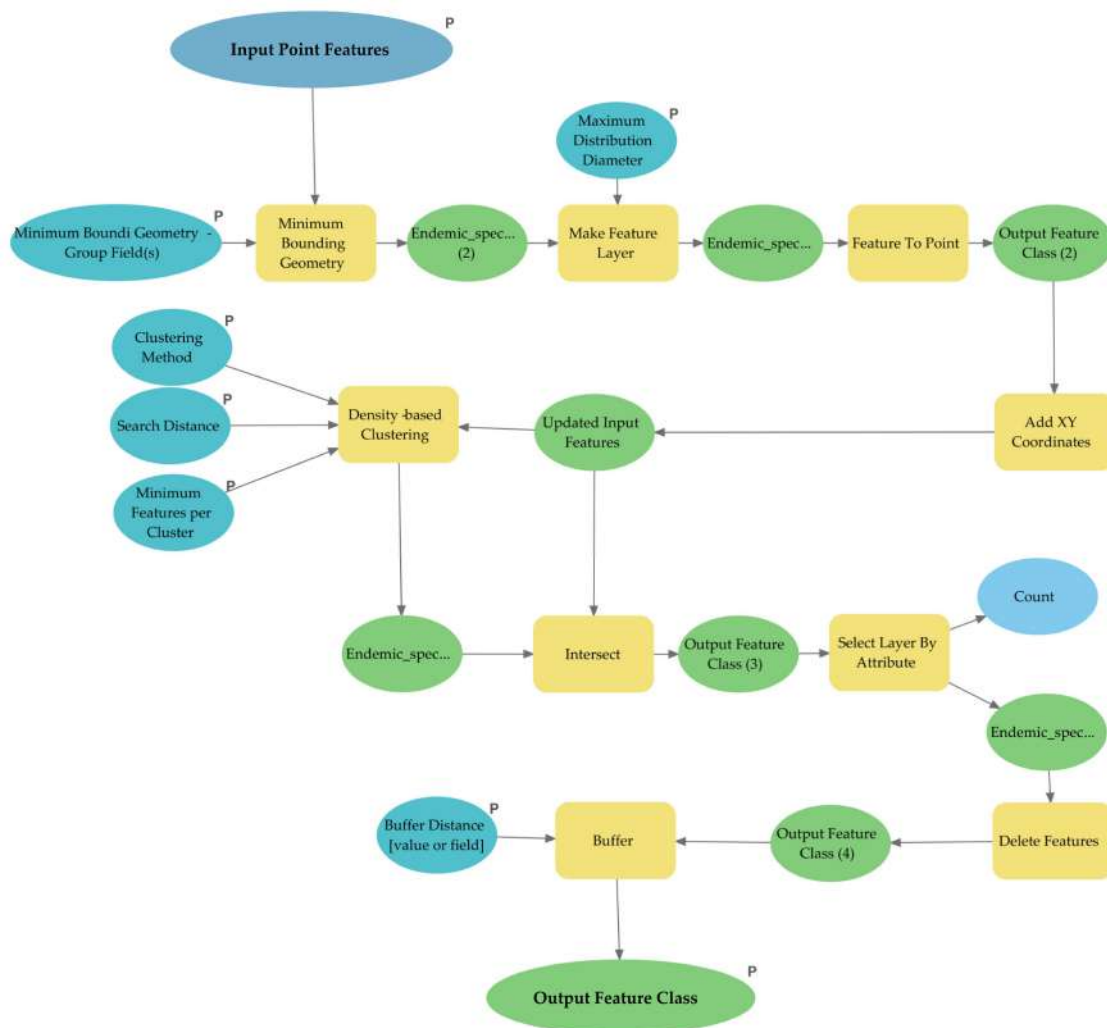


Figure 2. Scheme of the CLUENDA toolbox, developed in ArcGIS Pro 2.8.

To investigate possible differences in computation performances, execution times were measured for both GIE and CLUENDA. A 4-core Intel i7 processor (2.80 GHz), equipped with 32 GB of RAM (Windows 10) was used for this purpose.

The analyses were run with the dataset described above, as well as with a 10-times-bigger dummy dataset (generated based on the original dataset used, thus containing 32,960 points for the whole sub-Saharan Africa) to stress-test the two software used.

3. Results

3.1. Areas of Endemism (AoEs)

The execution times needed for the identification of the AoEs for the “regular” dataset were 75 s for GIE and 5 s for CLUENDA; the dummy “10-times bigger” dataset took 96 s and 6 s for the computation in GIE and CLUENDA, respectively.

The AoEs identified by both GIE and DBSCAN are listed below in alphabetical order:

- AMM (Amber Mountain): this refers to the Amber Mountain, a famous protected area in the Diana region of Northern Madagascar. AMM is well known for its endemic flora and fauna. AMM is a part of the “Madagascar and the Indian Ocean Islands” biodiversity hotspot [50].
- ANR (Antananarivo region): this AoE is located in Central Madagascar and mainly includes Anamalanga, Bongolava, Itasy, and Vakinankaratra regions. ANR is a part of the “Madagascar and the Indian Ocean Islands” biodiversity hotspot.
- BER (Betsiboka region): this AoE is located in Northern Madagascar and includes, in addition to the Betsiboka, the Sofia and Analanjirofo regions. BER is a part of the “Madagascar and the Indian Ocean Islands” biodiversity hotspot.
- DKM-KWN (Drakensberg Mountains–KwaZulu Natal): this AoE comprehends the Drakensberg Mountains, which are main mountain range in southern Africa, and the coastal areas of KwaZulu Natal. The Drakensberg Mountains rise to more than 3475 m a.s.l., extending roughly from northeast to southwest for over 1100 km parallel to the southeastern coast of South Africa. They are part of the Great Escarpment and separate the extensive highlands of the South African interior from the lower lands along the coast. In the Drakensberg Mountains, alpine grasslands and small pockets of Afromontane Forest are present. The coastal regions of KwaZulu-Natal typically have subtropical thickets and deeper ravines; steep slopes host small spots of Afromontane Forest. The midlands have moist grasslands. The northern area has a primarily moist savannah habitat. DKM-KWN is a part of the “Maputaland–Pondoland–Albany” biodiversity hotspot.
- KAR (Katanga region): this AoE is located in the Democratic Republic of Congo and mainly includes the ridges of the plateaus of Katanga (Shaba) province. They include Kundelungu (1600 m a.s.l.), Mitumba (1500 m), and Hakansson (1100 m) mountains. The Katanga plateaus reach as far north as the Lukuga River and contain the Manika Plateau, the Kibara and the Bia mountains, and the high plains of Marungu. Despite the high plant diversity present [51], KAR is not included in any biodiversity hotspot.
- KIL (Kilimanjaro region): this AoE has the Kilimanjaro region as its central area. Towards the southeast, it extends to the Tsavo West National Park in Kenya and Mkomazi National Park in Tanzania, while northwards it extends to the Amboseli National Park. KIL comprehends both montane and savannah areas. The Kilimanjaro area is a part of the “Eastern Afromontane” biodiversity hotspot.
- KLR (Kivu Lake region): this AoE is located in the area of the Kivu Lake in the Albertine Rift Valley and also includes the mountain areas of Birunga, Volcan Mikeno, and Volcan Karisimbi, between Uganda, Rwanda, and the Democratic Republic of Congo (Kivu Sud). KLR is a part of the “Eastern Afromontane” biodiversity hotspot.
- MP-LI (Mpumalanga–Limpopo): this AoE comprehends part of the Mpumalanga and Limpopo provinces. Mpumalanga is divided by the Drakensberg escarpment into a westerly half consisting mainly of high-altitude grassland called the Highveld and an eastern half situated in low-altitude subtropical Lowveld/Bushveld, mostly savannah

habitat. The Lowveld is relatively flat with interspersed rocky outcrops. The Lebombo Mountains form a low range in the far east, on the border with Mozambique. Limpopo contains much of the Waterberg Biosphere, a massif of approximately 15,000 km² which is the first region in the northern part of South Africa to be named a UNESCO Biosphere Reserve (<http://waterbergbiospherereserve.org/why-the-waterberg-is-a-biosphere.html>, accessed on 12 October 2021). MP-LI is partially included in the “Maputaland–Pondoland–Albany” biodiversity hotspot.

- WCP (Western Cape Province): this AoE is mainly restricted to South Africa’s Western Cape Province. Most of the region is covered with fynbos, a sclerophyllous shrubland occurring on acid sands or nutrient-poor soils derived from Table Mountain sandstones. This area covers the Mediterranean climate region of South Africa in the southwestern corner of the country and extends eastward into the Eastern Cape Province. WCP is a part of one of the world’s six floral kingdoms, the “Cape Floristic Region” biodiversity hotspot.

3.2. GIE

The weighted consensus map obtained by the GIE analysis, setting the “Minimum number of endemism” = 5, is shown in Figure 3a. In this map, kernels identify seven AoEs, five larger—WCP (32 sinendemisms), DKM-KWN (19), MP-LI (14), and KIL (12) in sub-Saharan Africa and AMM+BER+ANR (51) in Madagascar—and two smaller—KAR (7) and KLR (5) in Central Africa. There is no significant difference in the results between the weighted and the unweighted maps.

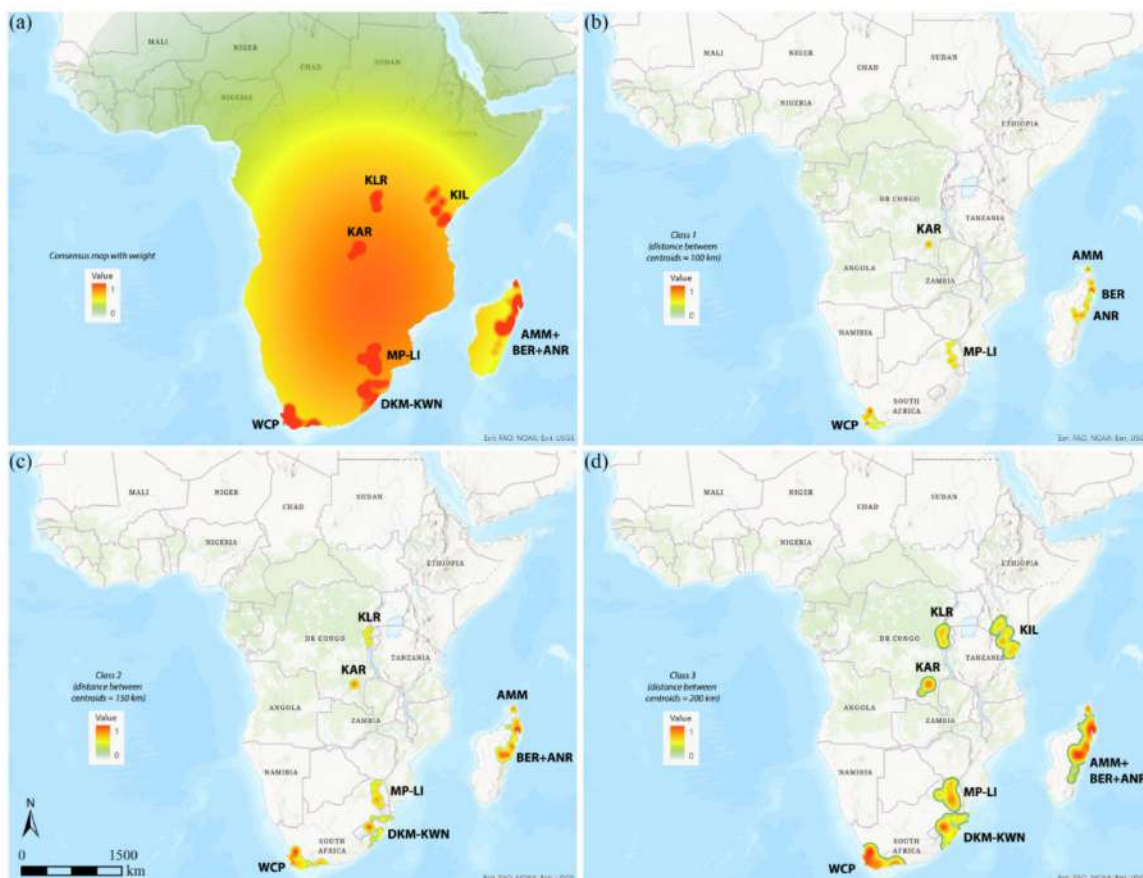


Figure 3. (a) Consensus maps of the areas of endemism detected by GIE. (b) Areas of endemism detected by GIE in Class 1 (100 km), (c) Class 2 (150 km) and (d) Class 3 (200 km).

As the intrinsic value of an endemic species is weighted by the width of its distribution range, the kernel maps of the three classes selected (Class 1, Class 2, Class 3) are displayed

individually (Figure 3b–d), and the endemic species included in each of them are listed in Supplementary File S3. The use of different classes was performed to identify areas of endemism with a higher degree of specificity based on the different distances of the centroids of the species range. In Class 1, which includes only the species whose distances between the centroids is equal to or smaller than 100 km, only small and fragmented areas were identified and attributed to the main five AoEs (Figure 3b; Table 1): western part of WCP (20 sinendemisms), MP-LI (13), and KAR (6) for sub-Saharan Africa and AMM (6) and BER + ANR (28) for Madagascar. Instead, considering the species with distances between the centroids equal to or smaller than 150 km (Class 2), the seven AoEs identified were located (Figure 3c; Table 1): in WCP (+11 sinendemisms), BER + ANR (+8), with larger areas, and in MP-LI, KAR, and AMM, unchanged; to these the areas DKM-KWN (17) and KLR (5) were added for sub-Saharan Africa. Finally, with distances between the centroids equal to or smaller than 200 km (Class 3), the definitive seven AoE (Figure 3d; Table 1) were detected similarly to those identified in the consensus maps. The correlation values among the maps were calculated for the three classes considered. The highest value ($r = 0.92525$) is between Class 2 and Class 3, while the lesser ($r = 0.73043$) is between Class 1 and Class 3.

3.3. DBSCAN

The results by DBSCAN analysis for identifying AoEs, setting the number of sinendemisms to 5, are reported in Figures 4–9, and the endemic species included in each of them are listed in Supplementary File S3. The extension, the number of sinendemisms, and the ratio “number of endemic species/total of species” for each area are reported in Tables 1 and 2 and in Figure 10.

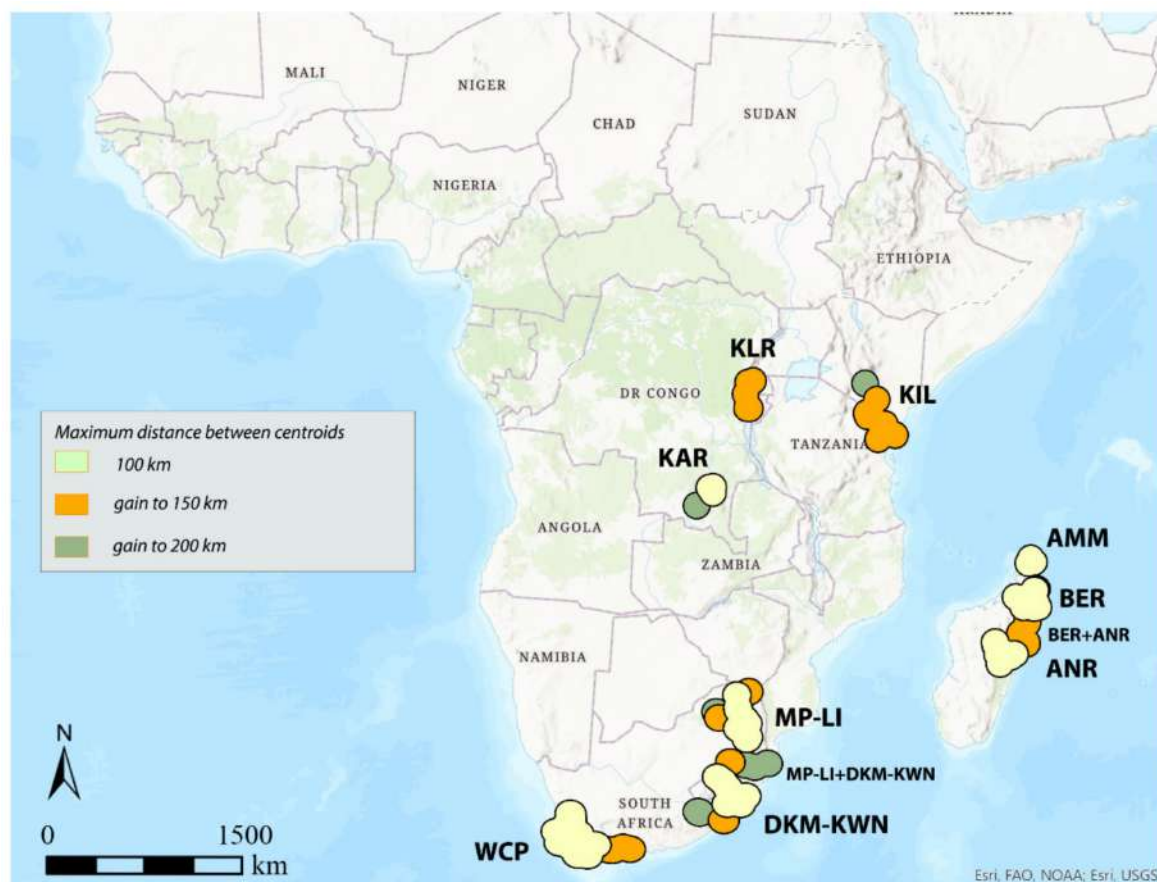


Figure 4. Areas of endemism detected by DBSCAN using species with distribution range up to 100 km and distance between centroids of 100, 150, and 200 km.

Table 1. Surface and number of sinendemisms for the AoEs identified by the DBSCAN analysis. dd = maximum diameter of the distribution of the species; dc = maximum distance between the centroids; ► indicates the merger of areas.

AoE →	KLR	Area	KIL	Area	KAR	Area	DKM-KWN	Area	MP-LI	Area	DKM-KWN + MP-LI	Area	WCP	Area	AMM	Area	BER	Area	ANR	Area	BER + ANR	Area
dd-dc ↓	syn	km ²	syn (tot)	km ²	syn (tot)	km ²	syn (tot)	km ²	syn (tot)	km ²	syn (tot)	km ²	syn (tot)	km ²	syn (tot)	km ²	syn (tot)	km ²	syn (tot)	km ²	syn (tot)	km ²
100–100	<5	/	<5	/	6 (26)	44,295	10 (53)	102,354	11 (55)	108,354	-	/	21 (57)	146,352	6 (12)	43,465	10 (27)	91,580	10 (34)	85,126	-	/
300–100	<5	/	<5	/	6 (26)	44,295	13 (53)	114,482	11 (55)	108,354	-	/	32 (67)	197,269	6 (12)	43,465	►		►		35 (60)	231,762
500–100	<5	/	<5	/	6 (26)	44,295	16 (54)	143,527	11 (55)	108,354	-	/	35 (67)	197,269	6 (12)	43,465	►		►		40 (61)	242,557
100–150	5 (31)	80,810	8 (35)	127,636	6 (26)	44,295	13 (58)	151,118	14 (59)	148,547	-	/	25 (67)	190,544	6 (12)	43,465	►		►		25 (60)	224,984
300–150	5 (31)	80,810	11 (36)	144,800	6 (26)	44,295	17 (58)	175,819	14 (59)	148,547	-	/	32 (67)	197,269	6 (12)	43,465	►		►		35 (60)	231,762
500–150	6 (31)	82,179	17 (39)	221,626	6 (26)	44,295	►		►		38 (94)	379,608	35 (67)	197,269	6 (12)	43,465	►		►		40 (61)	242,557
100–200	5 (31)	80,810	9 (38)	154,709	7 (28)	71,240	►		►		34 (97)	389,194	25 (67)	190,544	6 (12)	43,465	►		►		25 (60)	224,984
300–200	7 (42)	112,281	15 (39)	218,954	7 (28)	71,240	►		►		37 (97)	393,318	32 (67)	197,269	6 (12)	43,465	►		►		35 (60)	231,762
500–200	7 (42)	112,281	17 (39)	221,626	7 (28)	71,240	►		►		41 (97)	419,098	35 (67)	197,269	6 (12)	43,465	►		►		40 (61)	242,557

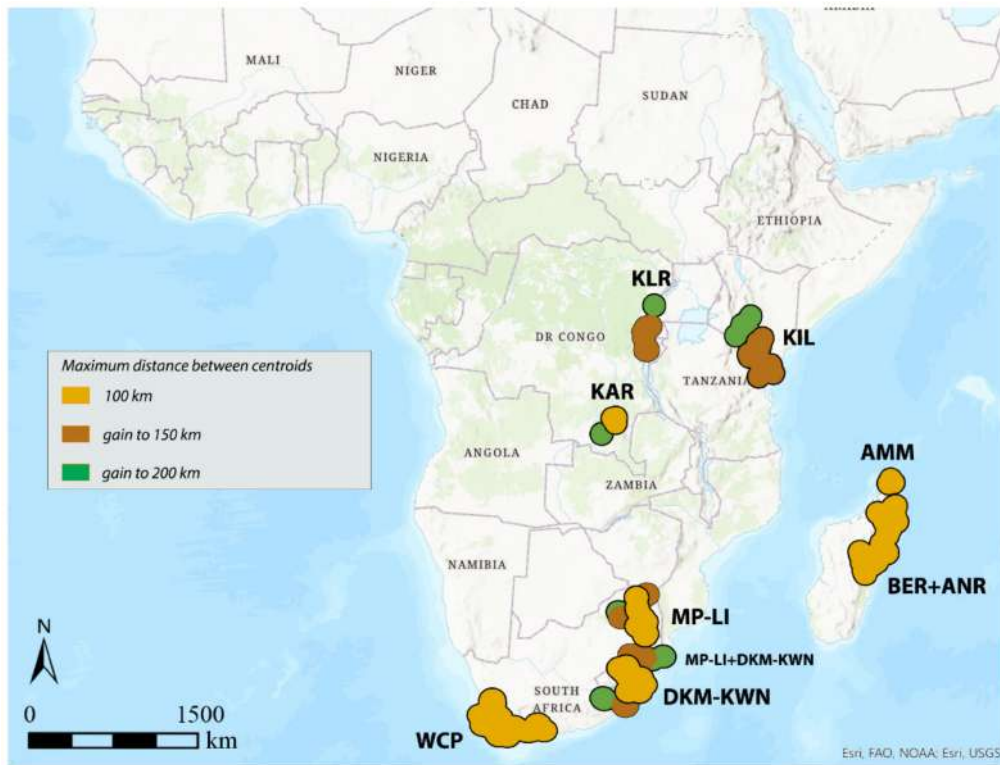


Figure 5. Areas of endemism detected by DBSCAN using species with distribution range up to 300 km and distance between centroids to 100, 150, and 200 km.

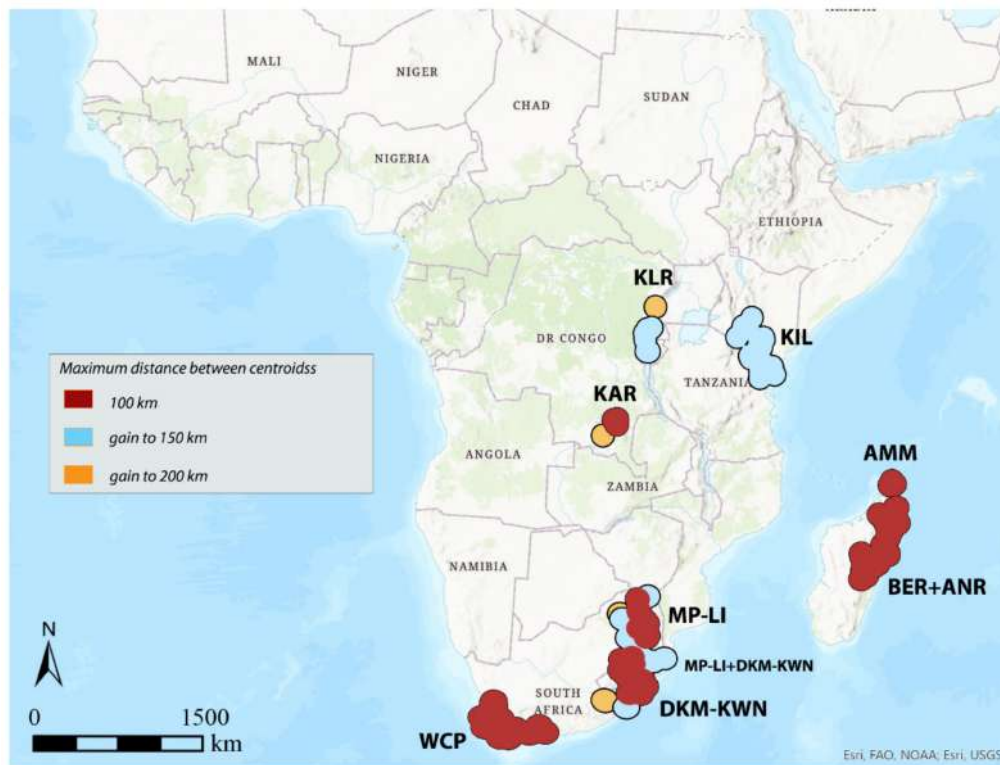


Figure 6. Areas of endemism detected by DBSCAN using species with distribution range up to 500 km and distance between centroids to 100, 150, and 200 km.

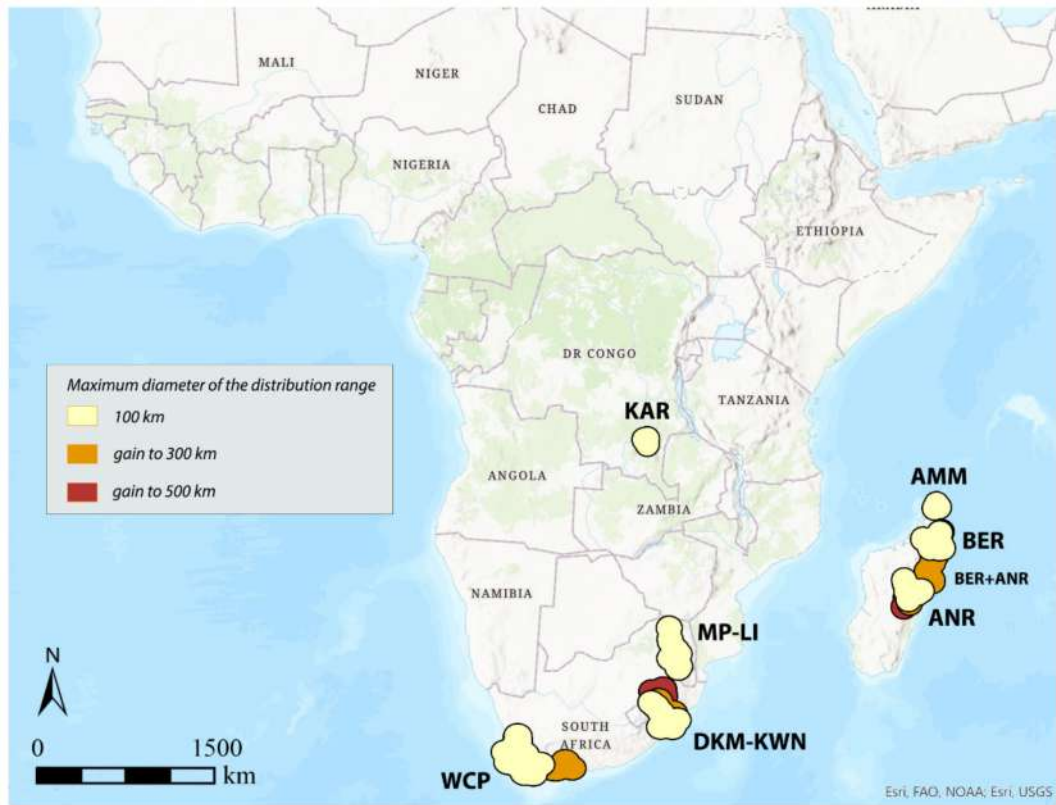


Figure 7. Areas of endemism detected by DBSCAN using the distance between centroids to 100 km and species with distribution range up to 100, 300, and 500 km.

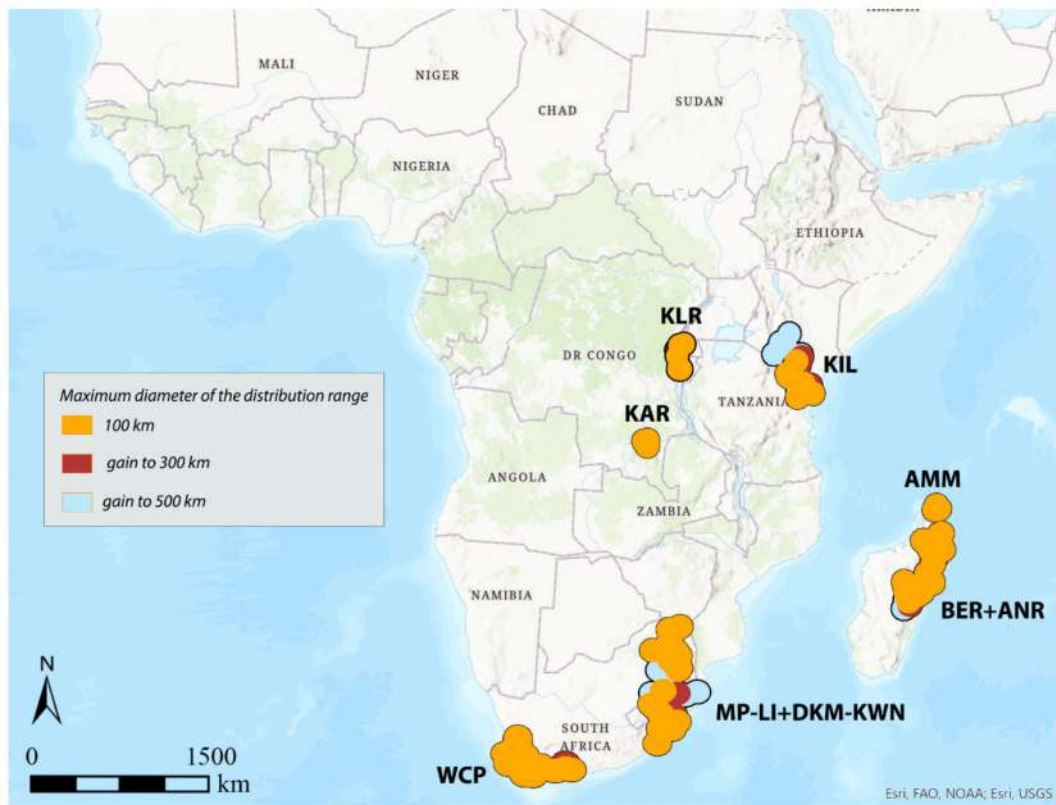


Figure 8. Areas of endemism detected by DBSCAN using the distance between centroids to 150 km and species with distribution range up to 100, 300, and 500 km.

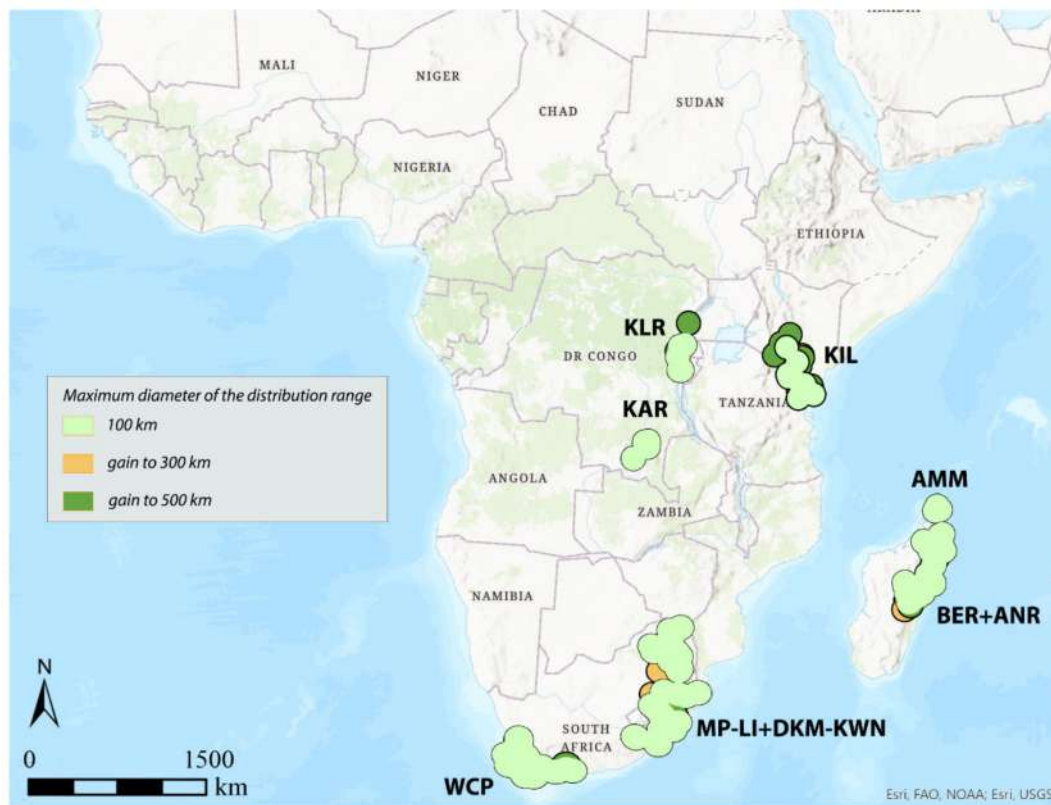


Figure 9. Areas of endemism detected by DBSCAN using the distance between centroids to 200 km and species with distribution range up to 100, 300, and 500 km.

Table 2. Maximum number of sinendemisms for AoE by GIE and DBSCAN. Please refer to paragraph 2.2 for the names in full of Areas of Endemism reported in the first row.

	KLR	KIL	KAR	DKM-KWN	MP-LI	DKM-KWN + MP-LI	WCP	AMM	BER + ANR	AMM + BER + ANR
GIE	syn	syn	syn	syn	syn	syn	syn	syn	syn	syn
100 km	/	/	6	/	13	/	20	6	28	/
150 km	5	/	6	17	13	/	31	6	35	/
200 km	5	12	7	19	14	/	32	/	/	51
DBSCAN										
100 km	/	/	6	16	11	/	35	6	40	/
150 km	6	17	6	/	/	38	35	6	40	/
200 km	7	17	7	/	/	41	35	6	40	/

In general, it can be observed that as the distribution interval of the species considered increases, respectively, with a maximum diameter of 100, 300, and 500 km, the identified AoEs are, at the same distance between the centroids, homogeneous in their geographical location, although they may be more extensive or merged into larger areas (Figures 4–6).

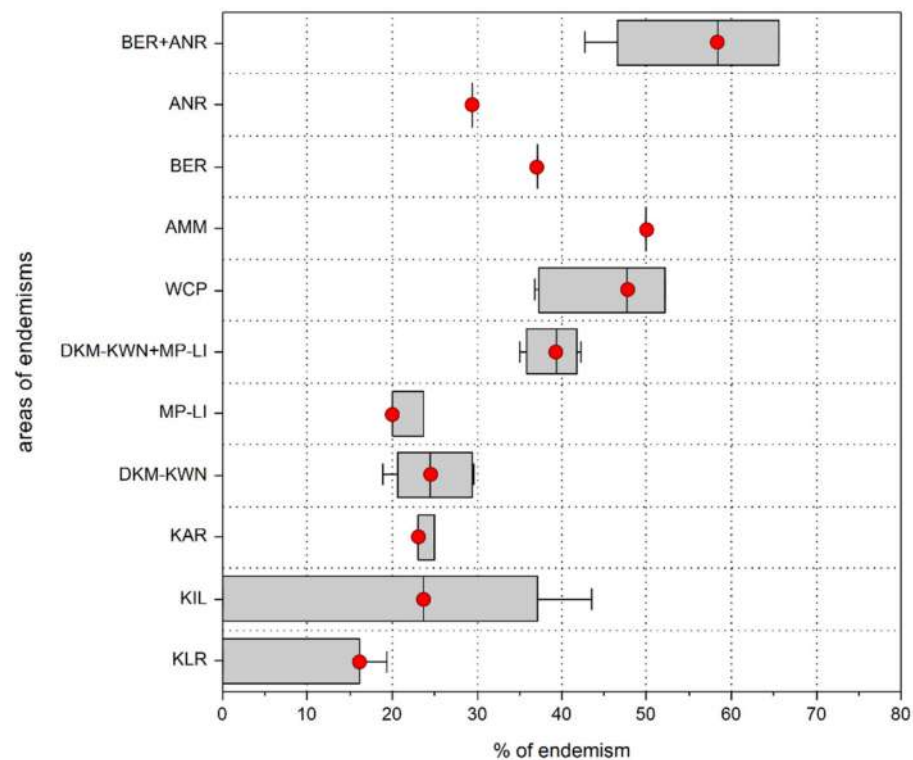


Figure 10. Box plots of the ratio “number of endemic species/total of species” for each of the AoEs identified by the DBSCAN analysis. Red spot = median.

Considering 100 km as the maximum distance between the centroids of the selected species (“defined distance” = 100 km), with the species with distribution up to 100 km (150 species out of 337 total), the analysis identifies the following seven AoEs (Figure 4, Table 1): KAR (6 sinendemisms) in Central Africa; DKM-KWN (10), MP-LI (11), and WCP (21) in Southern Africa; AMM (5), BER (11), and ANR (10) in Madagascar. With species up to 300 (181 out of 337) and 500 km (197 out of 337) (Table 1), a general increase in surface is observed for the DKM-KWN (+3 sinendemisms) and WCP (+11) areas, as well as a merger of the BER and ANR areas (35). On the other side, the areas KAR, MP-LI, and AMM remain unchanged. By increasing the “defined distance” in DBSCAN to 150 and 200 km, the results are quite different. In the first case, the analysis identifies two new AoEs in Eastern Africa—KIL, with 8, 11, and 17 sinendemisms, respectively, and KLR, with 5, 7, and 7 sinendemisms—while an area increase is observed for DKM-KWN and MP-LI, which merge into a unique area when the range of the species considered is up to 500 km (Figure 5, Table 1). By increasing the distance of the centroids to 200 km (“defined distance” = 200 km), in all three cases (species up to 100, 300, and 500 km), we observe the merger of the DKM-KWN and MP-LI areas and the gradual increase, in terms of surface, of KLR and KAR (Figure 6, Table 1). The correlation values (Kendall index) among the maps obtained by the different combinations of the widths of the distribution ranges and the distance between centroids are reported in Table 3. The correlation index decreases quickly as the distance between the centroids increases (100, 150, 200 km) rather than with the increase in the width of the distribution ranges (100, 300, 500 km). The non-parametric Kendall index was preferred because it uses pairs of observations and determines the strength of association based on the pattern of concordance and discordance between the pairs, which best fits our analysis.

Table 3. Kendall rank correlation index between the maps obtained by the DBSCAN analysis. dd = maximum diameter of the distribution of the species; dc = distance between the centroids.

dd-dc	DBSCAN								
	100 km/ 100 km	300 km/ 100 km	500 km/ 100 km	100 km/ 150 km	300 km/ 150 km	500 km/ 150 km	100 km/ 200 km	300 km/ 200 km	500 km/ 200 km
100 km/ 100 km	1	0.522	0.459	1.000	0.522	0.459	1.000	0.522	0.459
300 km/ 100 km	0.522	1	0.920	0.522	1.000	0.920	0.522	1.000	0.920
500 km/ 100 km	0.459	0.920	1	0.459	0.920	1.000	0.459	0.920	1.000
100 km/ 150 km	1.000	0.522	0.459	1	0.522	0.459	1.000	0.522	0.459
300 km/ 150 km	0.522	1.000	0.920	0.522	1	0.920	0.522	1.000	0.920
500 km/ 150 km	0.459	0.920	1.000	0.459	0.920	1	0.459	0.920	1.000
100 km/ 200 km	1.000	0.522	0.459	1.000	0.522	0.459	1	0.522	0.459
300 km/ 200 km	0.522	1.000	0.920	0.522	1.000	0.920	0.522	1	0.920
500 km/ 200 km	0.459	0.920	1.000	0.459	0.920	1.000	0.459	0.920	1

4. Discussion

DBSCAN and GIE both return largely overlapping results, detecting the same geographical locations for the AoEs, even if their results as a whole result in generally different delimitation and surface and the number of the detected sinendemisms (Table 2), also resulting in very different computation times. The consensus maps obtained by GIE are in general less clearly delimited than the maps obtained by DBSCAN, but nevertheless allow us to evaluate the core of the AoEs more precisely, as they can be displayed with a “stretched” symbology representing of the percentage levels of the overlap of the centroids, keeping a sort of hierarchy between the nested areas. DBSCAN, on the other hand, appears to be more sensitive in identifying areas of endemism. This method, in fact, detects the same areas as GIE but with comparatively shorter distances between the centroids. For example, with a “defined distance” of 150 km, regardless of the extent of the distribution of the species used in the analysis (100, 300, or 500 km in diameter), DBSCAN identifies the position of all the possible areas, including KIL and KLR, while with GIE, the KLR area is identified only in Class 2 (radius 150 km) and the KIL area in Class 3 (200 km). The same is also true for the DKM-KWN area, identified by DBSCAN with a distance between centroids of 100 km and not by GIE. Creating datasets of species with different widths of the distribution ranges, such as those up to 100, 300, and 500 km here proposed, allows DBSCAN to identify more precise and better-defined areas, with a higher number of identified sinendemisms (Table 2). Generally, in DBSCAN, the increase in the width of the distribution range of the species selected for the analysis does not allow the identification of new AoEs or the disappearance of them, but only larger areas, or the merge of areas previously identified using species with more limited distributions. New areas are detected only with the increase in the distance between centroids.

From the conservation/landscape planning point of view, it is important to highlight the main difference between the two approaches: the DBSCAN produces discrete territorial units, while GIE gives a continuous output. Researchers, stakeholders, and decision-makers should bear this in mind when searching for AoEs to manage or protect them. For instance, if GIE is applied and a specific territory has to be managed, a threshold value must be chosen to discretize the GIE’s output. This may introduce subjectivity, but also strengthen the results if this choice is supervised by experts. On the other hand, DBSCAN was allowed

to identify comparable AoEs with shorter distances between centroids (with respect to GIE), also detecting a higher number of sinendemisms for each of them. Thus, the application of DBSCAN may be suggested when more conservative solutions are needed.

5. Conclusions

GIE and DBSCAN consider the degree of overlap between species ranges as a pivotal parameter. This requirement stems from the necessity to bypass the geometric constraints and spatial scale bias imposed by grid cell use. In fact, some authors have previously argued that areas of endemism should be modelled by using more “natural” geographical units instead of overlapping geometric-sized grid cells onto the study area [52]. Additionally, areas of endemism should have irregular edges, and quadrats traditionally used in methods grid-based commonly fail to fully describe the dynamic structure of species distributions [15]. Therefore, the methodologies here discussed are a promising way to implement the research of areas of endemism by taking into account all the aforementioned observations.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/insects12121115/s1>, File S1: Supplementary references used for the building of the Afrotropical flea beetle database; File S2: CLUENDA, an ArcGIS Pro toolbox for identification of areas of endemism by the DBSCAN method; File S3: List of the sinendemisms in each AoE identified by the DBSCAN and GIE methods.

Author Contributions: Conceptualization, M.B. and M.I.; methodology, M.B.; software, W.D.S.; validation, P.D. and M.I.; formal analysis, M.B. and M.I.; investigation, P.D. and W.D.S.; resources, M.I. and P.D.; data curation, M.I. and P.D.; writing—original draft preparation, M.B., M.I., P.D. and W.D.S.; visualization, M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministero dell’Istruzione, dell’Università e della Ricerca-A.I.M. Project PON R & I 2014-2020, grant number 1870582.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: All other data not available in Supplementary Materials can be requested to the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Prado, J.R.; Brennard, P.G.G.; Godoy, L.P.; Libardi, G.S.; Abreu-Júnior, E.F.; Roth, P.R.O.; Chiquito, E.A.; Percequillo, A.R. Species richness and areas of endemism of oryzomyine rodents (Cricetidae, Sigmodontinae) in South America: An NDM/VNDM approach. *J. Biogeogr.* **2015**, *42*, 540–551. [\[CrossRef\]](#)
2. DaSilva, M.B.; Pinto-da-Rocha, R.; de Carvalho, C.; Almeida, E.A.B. *Biogeografia Da América Do Sul: Padrões E Processos*; Editora Roca: São Paulo, Brazil, 2011.
3. Da Cardoso Silva, J.M.; de Cardoso Sousa, M.; Castelletti, C.H. Areas of endemism for passerine birds in the Atlantic forest, South America. *Glob. Ecol. Biogeogr.* **2004**, *13*, 85–92. [\[CrossRef\]](#)
4. Hurdu, B.-I.; Escalante, T.; Puşcaş, M.; Novikoff, A.; Bartha, L.; Zimmermann, N.E. Exploring the different facets of plant endemism in the South-Eastern Carpathians: A manifold approach for the determination of biotic elements, centres and areas of endemism. *Biol. J. Linn. Soc.* **2016**, *119*, 649–672. [\[CrossRef\]](#)
5. Brown, J.H.; Stevens, G.C.; Kaufman, D.M. The geographic range: Size, shape, boundaries, and internal structure. *Annu. Rev. Ecol. Syst.* **1996**, *27*, 597–623. [\[CrossRef\]](#)
6. Morrone, J.J. *Evolutionary Biogeography: An Integrative Approach with Case Studies*; Columbia University Press: New York, NY, USA, 2008; ISBN 0-231-51283-X.
7. Fattorini, S. Endemism in historical biogeography and conservation biology: Concepts and implications. *Biogeogr. J. Integr. Biogeogr.* **2017**, *32*, 47–75. [\[CrossRef\]](#)
8. Morrone, J.J.; Crisci, J.V. Historical biogeography: Introduction to methods. *Annu. Rev. Ecol. Syst.* **1995**, *26*, 373–401. [\[CrossRef\]](#)
9. Platnick, N.I. On areas of endemism. *Aust. Syst. Bot.* **1991**, *4*, 11–12.
10. Risp, M.D.; Linder, H.P.; Weston, P.H. Cladistic biogeography of plants in Australia and New Guinea: Congruent pattern reveals two endemic tropical tracks. *Syst. Biol.* **1995**, *44*, 457–473. [\[CrossRef\]](#)
11. Harold, A.S.; Mooi, R.D. Areas of endemism: Definition and recognition criteria. *Syst. Biol.* **1994**, *43*, 261–266. [\[CrossRef\]](#)
12. Cracraft, J. Species diversity, biogeography, and the evolution of biotas. *Am. Zool.* **1994**, *34*, 33–47. [\[CrossRef\]](#)

13. Anderson, S. Area and endemism. *Q. Rev. Biol.* **1994**, *69*, 451–471. [CrossRef]
14. DaSilva, M.B.; Pinto-da-Rocha, R.; DeSouza, A.M. A protocol for the delimitation of areas of endemism and the historical regionalization of the Brazilian Atlantic Rain Forest using harvestmen distribution data. *Cladistics* **2015**, *31*, 692–705. [CrossRef] [PubMed]
15. Crother, B.I.; Murray, C.M. *Ontology of Areas of Endemism*; Wiley Online Library: Hoboken, NJ, USA, 2011; ISBN 0305-0270.
16. Linder, H.P. On areas of endemism, with an example from the African Restionaceae. *Syst. Biol.* **2001**, *50*, 892–912. [CrossRef]
17. Henderson, I.M. Biogeography without area? *Aust. Syst. Bot.* **1991**, *4*, 59–71. [CrossRef]
18. Sigrist, M.S.; de Carvalho, C.J.B. Detection of areas of endemism on two spatial scales using Parsimony Analysis of Endemicity (PAE): The Neotropical region and the Atlantic Forest. *Biota Neotrop.* **2008**, *8*, 33–42. [CrossRef]
19. Morrone, J.J. Parsimony analysis of endemism (PAE) revisited. *J. Biogeogr.* **2014**, *41*, 842–854. [CrossRef]
20. Morrone, J.J. On the identification of areas of endemism. *Syst. Biol.* **1994**, *43*, 438–441. [CrossRef]
21. Cracraft, J. Patterns of diversification within continental biotas: Hierarchical congruence among the areas of endemism of Australian vertebrates. *Aust. Syst. Bot.* **1991**, *4*, 211–227. [CrossRef]
22. Aagesen, L.; Szumik, C.; Goloboff, P. *Consensus in the Search for Areas of Endemism*; Wiley Online Library: Hoboken, NJ, USA, 2013; ISBN 0305-0270.
23. Szumik, C.A.; Goloboff, P.A. Areas of endemism: An improved optimality criterion. *Syst. Biol.* **2004**, *53*, 968–977. [CrossRef]
24. Szumik, C.A.; Cuezco, F.; Goloboff, P.A.; Chalup, A.E. An optimality criterion to determine areas of endemism. *Syst. Biol.* **2002**, *51*, 806–816. [CrossRef] [PubMed]
25. Torres-Miranda, A.; Luna-Vega, L.; Oyama, K. New approaches to the biogeography and areas of endemism of red oaks (*Quercus L.*, Section *Lobatae*). *Syst. Biol.* **2013**, *62*, 555–573. [CrossRef]
26. Oliveira, U.; Brescovit, A.D.; Santos, A.J. Delimiting areas of endemism through kernel interpolation. *PLoS ONE* **2015**, *10*, e0116673.
27. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining (KDD)*; University of Munich: Munich, Germany, 1996; Volume 96, pp. 226–231.
28. Fleming, C.H.; Calabrese, J.M. A new kernel density estimator for accurate home-range and species-range area estimation. *Methods Ecol. Evol.* **2017**, *8*, 571–579. [CrossRef]
29. Bailey, T.C.; Gatrell, A.C. *Interactive Spatial Data Analysis*; Longman Scientific & Technical Essex: New York, NY, USA, 1995; Volume 413.
30. Navarro, F.R.; Cuezco, F.D.C.; Goloboff, P.A.; Szumik, C.A.; Lizarralde, M.S.; Quintana, M.G. Can insect data be used to infer areas of endemism? An example from the Yungas of Argentina. *Soc. Biol. Chile* **2009**, *82*, 507–522. [CrossRef]
31. Biondi, M.; Urbani, F.; D’Alessandro, P. Relationships between the geographic distribution of phytophagous insects and different types of vegetation: A case study of the flea beetle genus *Chaetocnema* (Coleoptera: Chrysomelidae) in the Afrotropical region. *Eur. J. Entomol.* **2015**, *112*, 311–327. [CrossRef]
32. Andrew, N.R.; Hughes, L. Species diversity and structure of phytophagous beetle assemblages along a latitudinal gradient: Predicting the potential impacts of climate change. *Ecol. Entomol.* **2004**, *29*, 527–542. [CrossRef]
33. Basset, Y. Invertebrates in the canopy of tropical rain forests How much do we really know? *Plant Ecol.* **2001**, *153*, 87–107. [CrossRef]
34. Thormann, B. Biodiversity of Leaf Beetles (Coleoptera: Chrysomelidae) in a Tropical Montane Rainforest Ecosystem Assessed with DNA Barcoding. Ph.D. Thesis, University of Bonn, Bonn, Germany, 2016.
35. Flowers, R.W.; Hanson, P.E. Leaf beetle (Coleoptera: Chrysomelidae) diversity in eight Costa Rican habitats. In *Special Topics in Leaf Beetle Biology, Proceedings of the 5th International Symposium on the Chrysomelidae, Foz do Iguaçu, Brazil, 25–27 August 2000*; Pensoft Publishers: Sofia, Bulgaria, 2003; pp. 25–51.
36. Ødegaard, F.; Diserud, O.H.; Engen, S.; Aagaard, K. The magnitude of local host specificity for phytophagous insects and its implications for estimates of global species richness. *Conserv. Biol.* **2000**, *14*, 1182–1186. [CrossRef]
37. Nadein, K.S.; Beždek, J. Galerucinae Latreille, 1802. *Morphol. Syst. Phytophaga Gruyter Berl.* **2014**, *3*, 251–259.
38. Nadein, K.S. Catalogue of Alticini genera of the World (Coleoptera: Chrysomelidae). Available online: <http://www.zin.ru/Animalia/Coleoptera/eng/alticinw.htm> (accessed on 20 October 2021).
39. Ruan, Y.; Konstantinov, A.S.; Shi, G.; Tao, Y.; Li, Y.; Johnson, A.J.; Luo, X.; Zhang, X.; Zhang, M.; Wu, J.; et al. The jumping mechanism of flea beetles (Coleoptera, Chrysomelidae, Alticini), its application to bionics and preliminary design for a robotic jumping leg. *ZooKeys* **2020**, *915*, 87–105. [CrossRef] [PubMed]
40. Nadein, K.; Betz, O. Jumping mechanisms and performance in beetles. I. Flea beetles (Coleoptera: Chrysomelidae: Alticini). *J. Exp. Biol.* **2016**, *219*, 2015–2027. [CrossRef]
41. Iannella, M.; D’Alessandro, P.; de Simone, W.; Biondi, M. Habitat specificity, host plants and areas of endemism for the genera-group *Blepharida* sl in the afrotropical region (Coleoptera, Chrysomelidae, Galerucinae, Alticini). *Insects* **2021**, *12*, 299. [CrossRef]
42. D’Alessandro, P.; Iannella, M.; Frasca, R.; Biondi, M. Distribution patterns and habitat preference for the genera-group *Blepharida* sl in Sub-Saharan Africa (Coleoptera: Chrysomelidae: Galerucinae: Alticini). *Zool. Anz.* **2018**, *277*, 23–32. [CrossRef]
43. Jolivet, P.; Verma, K.K. *Biology of Leaf Beetles*; Intercept Limited: Andover, UK, 2002; ISBN 1-898298-86-6.

44. Biondi, M.; Iannella, M.; D'Alessandro, P. *Adamastoraltica humicola*, new genus and new species: The first example of possible moss-inhabiting flea beetle genus from sub-Saharan Africa (Coleoptera, Chrysomelidae, Galerucinae). *Zootaxa* **2020**, *4763*, 99–108. [[CrossRef](#)] [[PubMed](#)]
45. Biondi, M.; D'Alessandro, P. *Ntaolaltica* and *Pseudophygasia*, two new flea beetle genera from Madagascar (Coleoptera: Chrysomelidae: Galerucinae: Alticini). *Insect Syst. Evol.* **2013**, *44*, 93–106. [[CrossRef](#)]
46. Biondi, M.; D'Alessandro, P. Afrotropical flea beetle genera: A key to their identification, updated catalogue and biogeographical analysis (Coleoptera, Chrysomelidae, Galerucinae, Alticini). *ZooKeys* **2012**, *253*, 1. [[CrossRef](#)] [[PubMed](#)]
47. Udvardy, M.D.; Udvardy, M.D.F. *A Classification of the Biogeographical Provinces of the World*; International Union for Conservation of Nature and Natural Resources: Morges, Switzerland, 1975; Volume 8.
48. Oliveira, U.; Soares-Filho, B.; Leitão, R.F.M.; Rodrigues, H.O. BioDinamica: A toolkit for analyses of biodiversity and biogeography on the Dinamica-EGO modelling platform. *PeerJ* **2019**, *7*, e7213. [[CrossRef](#)] [[PubMed](#)]
49. Campello, R.J.G.B.; Moulavi, D.; Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*; Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 160–172.
50. Habel, J.C.; Rasche, L.; Schneider, U.A.; Engler, J.O.; Schmid, E.; Rödder, D.; Meyer, S.T.; Trapp, N.; del Sos Diego, R.; Eggermont, H. Final countdown for biodiversity hotspots. *Conserv. Lett.* **2019**, *12*, e12668. [[CrossRef](#)]
51. Küper, W.; Sommer, J.H.; Lovett, J.C.; Mutke, J.; Linder, H.P.; Beentje, H.J.; van Rompaey, R.S.A.R.; Chatelain, C.; Sosef, M.; Barthlott, W. Africa's hotspots of biodiversity redefined. *Ann. Mo. Bot. Gard.* **2004**, *1*, 525–535.
52. Morrone, J.J.; Escalante, T. Parsimony analysis of endemism (PAE) of Mexican terrestrial mammals at different area units: When size matters. *J. Biogeogr.* **2002**, *29*, 1095–1104. [[CrossRef](#)]