Unsupervised learning of satellite images enhances discovery of late Miocene fossil sites in the Urema Rift, Gorongosa, Mozambique

João d'Oliveira Coelho[1,2], Robert L. Anemone[3] and Susana Carvalho[1,2,4,5]

[1] University of Oxford, Primate Models for Behavioural Evolution Lab, Institute of Human Sciences, Oxford, United Kingdom
[2] Universidade de Coimbra, Centre for Functional Ecology (CFE), Coimbra, Portugal
[3] University of North Carolina at Greensboro, Department of Anthropology, Greensboro, North Carolina, United States of America
[4] Universidade do Algarve, Interdisciplinary Centre for Archaeology and Evolution of Human Behaviour (ICArEHB), Faro, Portugal
[5] Gorongosa National Park, Sofala, Mozambique

ABSTRACT

Background: Paleoanthropological research focus still devotes most resources to areas generally known to be fossil rich instead of a strategy that first maps and identifies possible fossil sites in a given region. This leads to the paradoxical task of planning paleontological campaigns without knowing the true extent and likely potential of each fossil site and, hence, how to optimize the investment of time and resources. Yet to answer key questions in hominin evolution, paleoanthropologists must engage in fieldwork that targets substantial temporal and geographical gaps in the fossil record. How can the risk of potentially unsuccessful surveys be minimized, while maximizing the potential for successful surveys?

Methods: Here we present a simple and effective solution for finding fossil sites based on clustering by unsupervised learning of satellite images with the *k*-means algorithm and pioneer its testing in the Urema Rift, the southern termination of the East African Rift System (EARS). We focus on a relatively unknown time period critical for understanding African apes and early hominin evolution, the early part of the late Miocene, in an overlooked area of southeastern Africa, in Gorongosa National Park, Mozambique. This clustering approach highlighted priority targets for prospecting that represented only 4.49% of the total area analysed.

Results: Applying this method, four new fossil sites were discovered in the area, and results show an 85% accuracy in a binary classification. This indicates the high potential of a remote sensing tool for exploratory paleontological surveys by enhancing the discovery of productive fossiliferous deposits. The relative importance of spectral bands for clustering was also determined using the random forest algorithm, and near-infrared was the most important variable for fossil site detection, followed by other infrared variables. Bands in the visible spectrum performed the worst and are not likely indicators of fossil sites.

Discussion: We show that unsupervised learning is a useful tool for locating new fossil sites in relatively unexplored regions. Additionally, it can be used to target

Submitted 3 July 2020
Accepted 18 May 2021
Published 8 June 2021

Corresponding author
João d'Oliveira Coelho,
joao.coelho@anthro.ox.ac.uk

Academic editor
Bruce Lieberman

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj.11573**

specific gaps in the fossil record and to increase the sample of fossil sites.
In Gorongosa, the discovery of the first estuarine coastal forests of the EARS fills an important paleobiogeographic gap of Africa. These new sites will be key for testing hypotheses of primate evolution in such environmental settings.

Subjects Anthropology, Evolutionary Studies, Paleontology, Data Mining and Machine Learning, Spatial and Geographic Information Science
Keywords Geospatial Paleontology, Southeast Africa, Late Miocene, Remote Sensing, Unsupervised Learning

INTRODUCTION

Paleontological and molecular evidence indicate that *Homo* shared a most recent common ancestor (MRCA) with the *Pan* lineage during the late Miocene (11.6–5.3 Ma) in Africa (*Montagnon, 2013*; *Moorjani et al., 2016*; *Barba-Montoya, Dos Reis & Yang, 2017*; *Dos Reis & Yang, 2019*). This makes this period critical to understand the origins of our clade, and yet the geographic distribution of African fossil sites during this time is surprisingly sparse and strongly biased (Figs. 1A, 1B). Currently only three regions contain lithostratigraphic units that have yielded possible early hominins from the late Miocene: Djourab Desert, Chad (*Sahelanthropus tchadensis Brunet et al., 2002*); Tugen Hills, Kenya *(Orrorin tugenensis Senut et al., 2001*); and Middle Awash, Ethiopia (*Ardipithecus kadabba Haile-Selassie, 2001*).

If we expand this sample to include fossils of the African great apes during the same time period, the record is likewise sparse (*Andrews, 2019*), being limited to the gorilla-like *Chororapithecus abyssinicus* from Ethiopia (*Suwa et al., 2007*), *Samburupithecus kiptalami* (*Ishida & Pickford, 1997*) and the basal hominid *Nakalipithecus nakayamai* (*Kunimatsu et al., 2007*) from Kenya, plus some taxonomically uncertain dental and mandibular finds from Kenya and Niger (Table 1). Furthermore, there is still no consensus regarding the phylogenetic relationships among late Miocene African hominids or how they relate to the MRCA (*Senut, 2010*, *2015*; *Kunimatsu et al., 2016*). Even less understood are the biogeographic distributions and temporal spans of these taxa that existed within or around the estimated time range for the *Pan/Homo* divergence event. For the great apes, clear fossil evidence only reappears much more recently, in the Pleistocene, demonstrating the temporal extension of these gaps in the fossil record (*McBrearty & Jablonski, 2005*).

Fundamentally, many of these uncertainties originate from significant gaps in the fossil record. As a discipline, paleoanthropology seeks to answer key questions regarding origins and adaptations of our and other related lineages, but the available data are often too sparse to provide satisfactory answers. Overcoming these limitations may depend largely on developing new strategies for field work, targeting areas and time periods that are undersampled and poorly understood. This may involve considerable risk for paleoanthropological campaigns that most field workers and funding institutions might not be willing to take (*Emerson & Anemone, 2012*). Moreover, finding hominins is not a

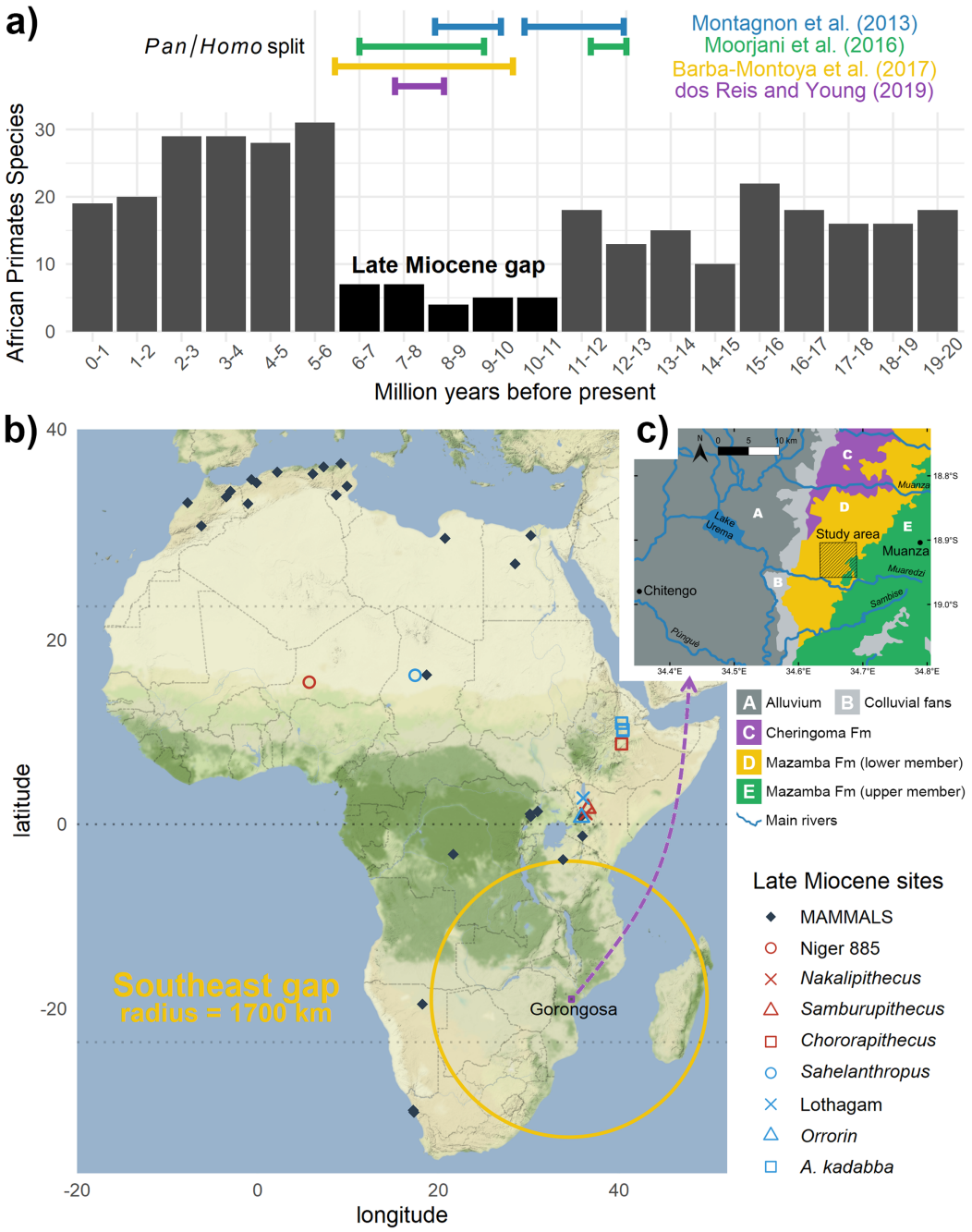

Figure 1 **The great gaps of the African late Miocene:** (A) Time gap**:** during this key period the African fossil record of primates is very incomplete (evaluated through species richness); but notice the split estimates from genomics; (B) spatial gap: virtually no fossils of this age are known in southeastern Africa, notice the strategic location of Gorongosa. Data extracted from paleobiodb.org, map adapted from *Bobe et al. (2018)*; (C) study area for *k*-means within the geological context of Gorongosa, adapted from *Habermann et al. (2019)*. Full-size DOI: 10.7717/peerj.11573/fig-1

straightforward endeavour. For instance, it took the Leakeys 33 years of work at Olduvai Gorge to find their first hominin, the holotype skull of *Paranthropus boisei* (*Leakey, 1959*). Therefore, innovative risk-aversion methods that can maximize information and

Table 1 All hominids described for Africa during the Late Miocene (11.6–5.3 Ma).

Age (Ma)	Fossil site(s)		Taxon	Reference(s)
ca. 5.4	ESC2;3;8, Gona, Ethiopia		*Ardipithecus kadabba*	(*Simpson et al., 2015*)
5.6–5.2*	Amba East, Ethiopia		*Ardipithecus* cf. *kadabba*	(*Haile-Selassie, 2001*; *Haile-Selassie, Suwa & White, 2004*)
5.77–5.54	ALA; ASK; DID; STD†		*Ardipithecus kadabba*	
5.8–5.7	Tugen Hills, Kenya	Kapcheberek	*Orrorin tugenensis*	(*Senut et al., 2001*; *Sawada et al., 2002*; *Pickford & Senut, 2005*; *Senut, 2010*, *2015*; *Senut, Pickford & Gommery, 2018*)
5.9–5.8		Kapsomin	*Orrorin tugenensis*	
			Incertae sedis (early *Gorilla?*)	
ca. 6.1		Cheboit	*Orrorin tugenensis*	
			Incertae sedis (early *Pan*?)	
		Aragai	*Orrorin tugenensis*	
6–5	Nkondo, Uganda		cf. Gorillini	(*Pickford et al., 1988*)
6.3	ABD1, Gona, Ethiopia		cf. *Ardipithecus kadabba*	(*Simpson et al., 2015*)
6.5–5‡	Lothagam, Kenya		Homininae indet.	(*Leakey & Walker, 2003*)
7.34-7.1§	Toros-Menalla, Chad		*Sahelanthropus tchadensis*	(*Brunet et al., 2002*, *2005*)
8.0	Ch'orora, Ethiopia		*Chororapithecus abyssinicus*	(*Suwa et al., 2007*)
9.5	Samburu Hills, Kenya		*Samburupithecus kiptalami*	(*Ishida & Pickford, 1997*)
9.9–9.8	NA39, Nakali, Kenya		*Nakalipithecus nakayamai*	(*Kunimatsu et al., 2007*)
			Hominidea indet.	(*Kunimatsu et al., 2016*)
11–8¶	N 885, Niger		*Incertae sedis* (early *Pan*?)	(*Pickford et al., 2008*, *2009*)

Notes:
* Probably closer to 5.2 Ma than 5.6, but the Kuserale Mb of the Sagantole Fm is bracketed as in *Renne et al. (1999)*.
† Fossiliferous localities in the Asa Koma Mb of the Adu Asa Fm (Middle Awash, Ethiopia). ALA = Alayla (ALA-VP-2), ASK = Asa Koma (ASK-VP-3), DID = Digiba Dora (DID-VP-1); STD = Saitune Dora (STD-VP-2).
‡ Not the KNM-LT 329 mandible, but two teeth (KNM-LT 22930 M_3; KNM-LT 25935 I_1) from older deposits (Upper Nawata).
§ Cosmogenic nuclide dating (*Lebatard et al., 2010*).
¶ Only biostratigraphic dating available, a more conservative range would be 11.6–5.3 Ma (Late Miocene).

improve current surveying approaches are urgently needed (*Njau & Hlusko, 2010*; *Anemone, Emerson & Conroy, 2011*).

How do we know where to look for fossils in unexplored and remote regions? When an area has been extensively surveyed, we can focus on previous work to guide our efforts. If that is not the case, we can use the geology and topography to enhance surveying efforts (*Asfaw et al., 1990*). However, that is not always possible if the modern topographic and ecological conditions for looking for fossil are not ideal. In places like the miombo woodlands of Gorongosa (Fig. 2), where nine late Miocene paleontological localities have recently been described, the surrounding dense vegetation cover can make reading the landscape for clues to fossiliferous deposits extremely difficult (*Habermann et al., 2019*). The fossil sites are found just north of the Muaredzi river and east of Lake Urema, between the villages of Chitengo and Muanza (Fig. 1C). The terrain of the study area can be described as flat to slightly undulating, with active incision and upstream erosion of stream sources. Topographic relief is also dotted by island thickets on termitaria hills. The vegetation in the study area is dominated by closed canopy *Brachystegia* (miombo) woodlands, with riverine forests occurring along the drainage network (Fig. 3). The grass layer density can vary greatly spatially and across the seasons, from widely spaced grass

Figure 2 The miombo woodland and the challenges it presents to fossil prospecting: Gorongosa Paleontological Locality 1 (GPL-1). (A) GPL-1 outcrops, notice how the surrounding vegetation is far more dense and extensive than in typical fossil sites from the EARS; (B) GPL-1 in high-resolution satellite image, extracted from bing.com, shows a reduction of vegetation, but outcrops are barely noticeable; (C) GPL-1, in a black rectangle, appears brighter than surrounding areas, when being mapped by lower resolution Landsat 8 false colour (infrared) image, and the same happens with other fossil sites, suggesting that infrared bands might be a useful indicator of fossiliferous deposits.
Full-size DOI: 10.7717/peerj.11573/fig-2

Figure 3 Surveying for fossils in a densely vegetated context. (A) Despite the ground foliage and dense vegetation, in situ and surface evidence of fossils abound in the gully valley connecting GPL-12 to GPL-12B; (B) Systematic mapping and collecting of surface fossil finds by students of the field school; (C) Side gully (~3 m deep) exposure and shovel test pit at GPL-12. Photographs are from the Paleo-Primate Project Gorongosa archive.
Full-size DOI: 10.7717/peerj.11573/fig-3

tufts in sandier substrates, to medium and tall (>2 m height) grass layers that are subject to regular, typically annual, fires (*Tinley, 1977*).

Modern geospatial technologies such as remote sensing, handheld GPS devices and geographic information systems (GIS) software have been applied to different paleoanthropological research questions with spatial components (*Conroy, 2006*; *Egeland, Nicholson & Gasparian, 2010*; *Anemone & Conroy, 2018*). Surveys aimed at discovering

new fossil assemblages with hominins have also been guided by military, geological and topographic maps, aerial photography and satellite images (*Asfaw et al., 1990*; *Malakhov, Dyke & King, 2009*; *Hlusko, 2018*; *Habermann et al., 2019*). With recent advances in artificial intelligence, cheaply available computer power, and free access to satellite imagery of reasonable resolution, paleoanthropologists and GIS technicians have been implementing machine learning techniques to automate the demanding visual analysis of remote fossil site detection (e.g., *Anemone, Emerson & Conroy, 2011*; *Conroy et al., 2012*, *2018*; *Conroy, 2014*; *Emerson et al., 2015*; *Block et al., 2016*; *Wills, Choiniere & Barrett, 2018*).

Here, we pioneer the application of such automated computational approaches for remote fossil site detection within the EARS (East African Rift System), specifically at its youngest and southernmost subsection, the Urema Rift (where rifting initiated at ~3 ± 1 Ma), in Gorongosa National Park, Mozambique (*Böhme et al., 2006*; *Steinbruch, 2010*; *Fonseca et al., 2014*; *Macgregor, 2015*). The southern part of the EARS is far less explored than its northern counterpart, which is rich in lacustrine and fluvial paleontological settings (*Bobe et al., in press*). The first fossil sites of the Urema Rift and their geological features have been recently described by *Habermann et al. (2019)*. All the sites described are part of the lower member of the Mazamba Formation which has been attributed to a Miocene or Mio-Pliocene age (*Real, 1966*; *Grantham et al., 2011*; *Pickford, 2013*; *Habermann et al., 2019*). Preliminary authigenic $^{10}Be/^{9}Be$ dating suggests the fossil sites fit into the earliest part of the late Miocene interval (*Bobe et al., 2021*). The lower Mazamba successions record coastal nearshore conditions that formed in a shallow ramp setting by marine transgression that occurred prior to rifting. North-eastern sites tend to be richer in fossils of marine fauna and are dominated by sand, while the southern sites are dominated by basal conglomeratic and sandy units overlain by clayey sandstones and sandy marl- and claystone facies (*Habermann et al., 2019*). The late Miocene paleoenvironments of Gorongosa were reconstructed as estuarine coastal forests and woodlands (*Habermann et al., 2019*)—a unique context in the EARS—that is both promising for the presence of primates (*Nowak, Barnett & Matsuda, 2019*) and crucial for testing biogeographic hypothesis of early hominin evolution in estuarine or deltaic wetlands (*Wrangham, 2005*) and in coastal forest biomes of eastern Africa (*Kingdon, 2003*; *Joordens et al., 2009*, *2019*; *Bobe, Martínez & Carvalho, 2020*).

In order to find new fossil sites in Gorongosa, a decision support system based on unsupervised learning of satellite images was created to guide prospecting during the 2018 field season. This was achieved by applying a simple clustering algorithm to satellite images of the field area. There are only two other examples in the literature that applied unsupervised clustering for detecting fossil sites, one in the Uinta Basin, Utah (*Conroy, 2014*) and the other in the Bighorn and Great Divide Basins, Wyoming (*Conroy et al., 2018*). Cluster analysis does not use categories that tag objects with prior identifiers (class labels). The absence of categorical information distinguishes data clustering (unsupervised learning) from classification or discriminant analysis (supervised learning). The aim of clustering is to find structure in data and it is therefore exploratory in nature (*MacQueen, 1967*). One of the most popular and simple clustering algorithms, *k*-means, was first

conceptualized in the 1950s (*Steinhaus, 1956*) and while thousands of different clustering algorithms have been published since then, *k*-means is still widely used because of its effectiveness, speed and simple implementation (*Bock, 2008*). While *k*-means is a standard technique for many remote sensing applications, it has never been used for fossil site detection. Here, for the first time, this algorithm has been adopted for the purpose of finding fossil sites using satellite imagery.

The goal of this pilot study was to determine if *k*-means, a simple unsupervised computer vision technique for processing satellite images, could improve our ability to identify fossil sites in a limited region of interest in Gorongosa with no a priori knowledge of the geology, stratigraphy, topography and land cover. In addition, we sought to characterize the lower Mazamba Formation (see *Habermann et al., 2019*) through extensive ground surveys to catalogue the presence and distribution of paleontological resources. These new fossil discoveries will help us to better understand an ancient ecosystem that existed during a crucial period of African great ape diversification and hominin origins. The insights gained from our fieldwork, in combination with the analyses of satellite imagery, allow us to gauge the accuracy of our model and to refine future iterations of these statistical modelling approaches for fossil site detection.

MATERIALS & METHODS

Field experiments were approved by Dr. Marc Stalmans, Director of the Department of Scientific Services of Gorongosa National Park in Mozambique (project number: PNG/DSCi/C019/2018). The code for the complete protocol and input files are available in an open access repository at https://github.com/Delvis/kmeansGorongosa/. The dataset consists of a freely available atmospherically corrected scene with 30 × 30 m resolution of Landsat 8 OLI (*Roy et al., 2014*) from July 28th, 2017 (id: LC08_L1TP_167073_20170628_20170714_01_T1), covering a portion of central Mozambique. Spectral data from Landsat and other medium-resolution satellites is updated daily and can be accessed at the USGS Earth Explorer website (https://earthexplorer.usgs.gov/). After downloading the satellite image, we applied a cropping filter to restrict it to an area of interest known to contain late Miocene to early Pliocene deposits (*Real, 1966*; *Grantham et al., 2011*), corresponding to approximately 36 km^2, in the lower Mazamba Formation within Gorongosa National Park (Fig. 1C). A multidimensional matrix containing the brightness values for seven spectral bands (two short-wave infrared, one near-infrared, three visible 'RGB', and one ultrablue) of Landsat 8 in UTM/WGS 84 coordinate system was processed through *k*-means, an unsupervised learning algorithm that can split the satellite image into different *k* clusters based on the spectral pattern of each pixel (a 30 × 30 m point in space). The clusters were then compared to geocoordinates of known late Miocene fossil sites in Gorongosa (available in *Habermann et al., 2019*) in order to determine whether there was a consistent visual pattern matching cluster(s) with fossil sites, as in *Conroy's (2014)* 'walking back the cat' approach. In other words, to algorithmically retrace our steps to identify other potential fossiliferous outcrops. This is achieved by generating clusters in an unsupervised manner, and then selecting the cluster(s) containing most fossil sites as the target cluster(s) for the field season (Fig. 4).

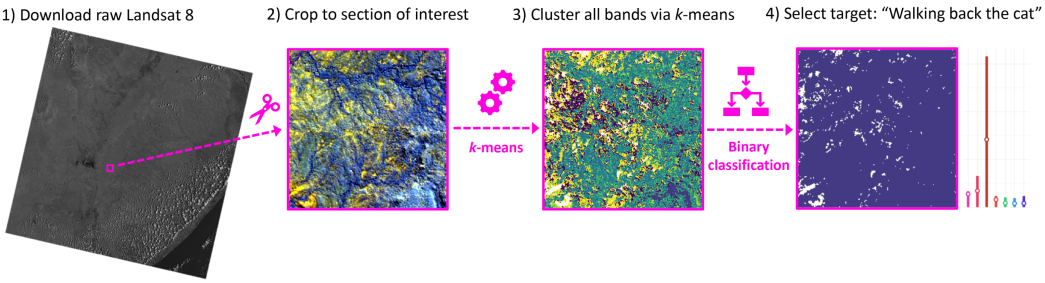

Figure 4 Flowchart of the algorithmic pipeline used for remote fossil site detection. (1) Example of one of the seven spectral bands satellite images used in this study; (2) false colour map based on the infrared bands, after cropping to study area; (3) results of clustering using all seven spectral bands; (4) Binarize clusters for classification by selecting the cluster that contains most fossil sites as the target class ("walking back the cat") versus all other clusters combined into a single class.

Full-size DOI: 10.7717/peerj.11573/fig-4

The advantages of such approach is that it works: (a) when we only have very few fossil sites known in the region; (b) in the absence of any prior knowledge of landscape cover over the proposed survey area; and (c) without pre-processing or training any landcover classes (*Conroy, 2014*; *Conroy et al., 2018*). For data analysis purposes, we define "fossil site" as the 30 × 30 m pixel in the Landsat grid that includes the geographic centroid corresponding to a locality in the landscape with exposed outcrops containing either fossilized vertebrates, invertebrates and/or wood reported in the study area (*Pickford, 2012*, *2013*; *Habermann et al., 2019*).

Applying the *k*-means algorithm to satellite images

Partitional clustering is a family of unsupervised learning techniques for grouping a set of data points (instances) into k disjoint groups, better known as clusters. The goal is to increase intra-cluster similarity (in this case, Euclidean distance between grouped instances) while decreasing inter-cluster similarity. More specifically, the k-means algorithm approximates the best division of n data points in k groups, so that the total distance between each grouped instance $x_i, \quad i \in \{1, \ldots, n\}$ and its corresponding centroid $\mu_j, \quad j \in \{1, \ldots, k\}$ is minimized. Formally, this partition occurs through minimization of the within-cluster sum of squares (WCSS), defined as

$$WCSS = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^j - \mu_j||^2 \tag{1}$$

where $||x^j{}_i - \mu_j||$ calculates the distance between an instance and a cluster's centroid.

With a rich history starting in the 1950s, the k-means algorithm was independently discovered in various scientific fields (*Steinhaus, 1956*; *Ball & Hall, 1965*; *Forgy, 1965*; *MacQueen, 1967*; *Lloyd, 1982*). Its most standard implementation can be broken down into two stages: initialization, where k cluster centroids $(\mu_1, \ldots \mu_k)$ are randomly selected (*Forgy, 1965*); and a second, iterative or repeat stage, following Lloyd's algorithm (*Lloyd, 1982*). Lloyd's iterative refinement technique can be further broken down into the assignment and update steps: first, each instance is assigned to the cluster with the closest

centroid μ_j; then the set of centroids is updated, by recalculating each centroid with the new instances attached to the clusters. This repetition should run iteratively until cluster membership for the entities converges to a stable solution. Computationally, if *C* and *C*' are the set of centroids obtained at consecutive Lloyd's iterations, then the algorithm stops when

$$|WCSS(C) - WCSS(C')| \leq \varepsilon, \text{ for a fixed threshold } \varepsilon \ll 1 \quad (2)$$

The greatest advantage of *k*-means is scalability, as only the centroid coordinates are stored in memory, it can deal with very large datasets. Moreover, every step can be parallelized, which increases computation performance (*Wu et al., 2008*; *Zhao, Ma & He, 2009*). Nevertheless, it can be slow to compute, since each instance might be processed many times throughout the iterations. Another limitation is that the results are dependent on the initial random allocation of the centroids (*Äyrämö & Kärkkäinen, 2006*).

In the specific case of satellite images, each instance is a geolocated point/pixel (30 × 30 m) with associated brightness values for all the seven spectral bands. Therefore, by applying the *k*-means algorithm to such data, each instance is assigned to one of the *k* clusters, and instances with similar spectral characteristics will cluster together. Thus *k*-means has many uses, including the ability to: (a) explore and mine hidden patterns in the satellite data; (b) partition a dataset into different groups that might correspond to real types of land cover; and (c) feature learning (a data science technique for encoding new features from raw data), since the output clusters can be used as new variables for subsequent modelling approaches (*Jain, 2010*).

Clusters as survey guides for fossil site discovery

If we allow a specific cluster, say cluster 1, to be a predictor for fossil sites, after applying a *k*-means where k = 8, we can consider from a statistical classification paradigm: True Positives (TP) = fossil sites on cluster 1; False Positives (FP) = non-localities on cluster 1; False Negatives (FN) = fossil sites on clusters 2 to 8; True Negatives (TN) = non-localities on clusters 2 to 8. The clustering results were validated digitally using geolocated coordinates for fossil sites and non-localities collected through ground surveys during the 2016 and 17 field seasons. Then, the region was revisited during the first two weeks of August 2018 to survey unexplored areas on foot and ground-truth the model. Localities that could not be thoroughly observed due to obstacles such as high vegetation density, dry foliage, difficult access, or other factors, were recorded as NA (not applicable) and were removed from the current analysis. All other surveyed localities were evaluated as either TP, FP, FN or TN and analysed through measures of statistical performance to assess the efficiency of the implemented approach.

To clarify the mechanisms responsible for generating different clusters, a supervised random forest algorithm for classification (*Breiman, 2001*) was built on top of the new learned features (cluster 1 to 8). Variable importance metrics, like the permuted mean square-error (%IncMSE), can be extracted from the random forest model. Calculating % IncMSE is achieved by internal out-of-bag estimates of error rate, and then verified by reruns excluding each predictor variable (i.e. permutation). For all the trees in the model,

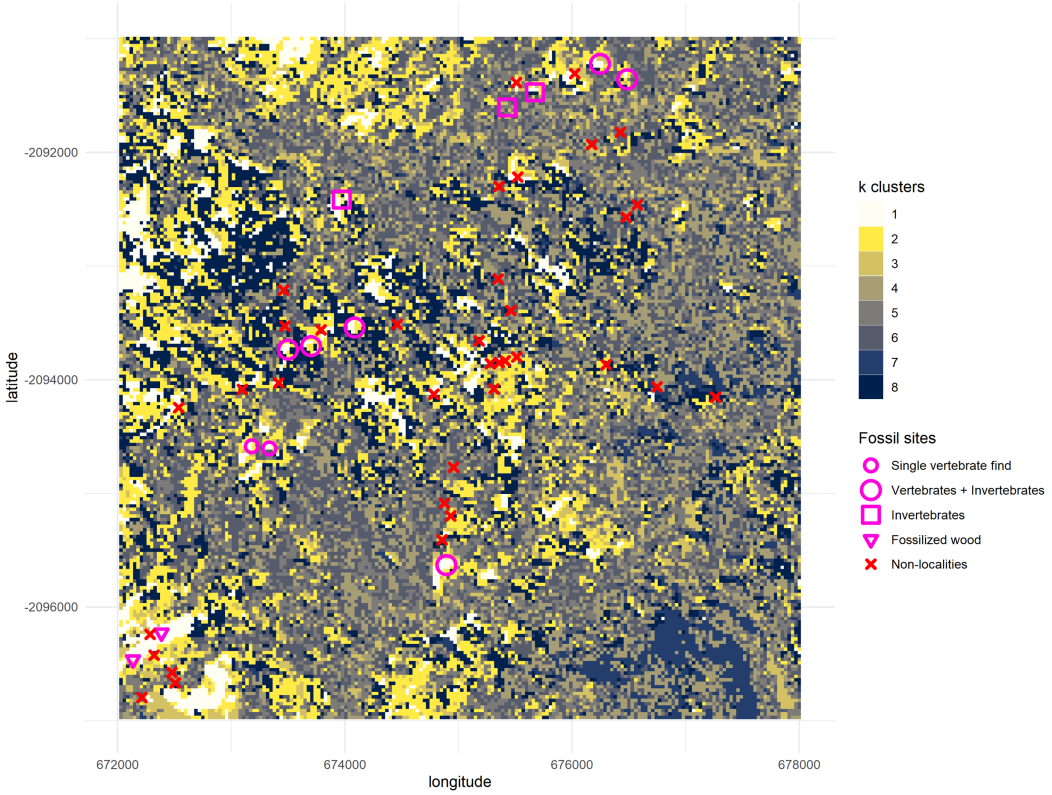

Figure 5 Output from the *k*-means algorithm for data mining. All geolocations "Vertebrates + Invertebrates" and "Invertebrates" recorded by the PaleoPrimate Project Gorongosa team during 2016 and 2017 (*Habermann et al., 2019*) are plotted over the clusters, as well as the "Single vertebrate find" and "Fossilized wood" localities reported by *Pickford (2012, 2013)*. You can see the cluster 1 (white) tends to be represented in locations with fossil vertebrates, indicating that it has some potential as a new feature/variable for finding new fossil sites. The map displays a 6 by 6 km square; axes scales are in meters.
Full-size DOI: 10.7717/peerj.11573/fig-5

the difference between error rates is averaged, and normalized by the standard deviation of the differences, thus calculating a metric of variable importance (*Liaw & Wiener, 2002*).

RESULTS

Digital validation

A plot with the clustering results of the satellite image was rendered and superimposed with a layer representing all the nine known Gorongosa Paleontological Localities (GPL) at the time, provided by *Habermann et al. (2019)*, and two possible vertebrate sites with single-fossil finds of large mammals, plus two sites with fossil wood from earlier work by *Pickford (2012, 2013)*. Additionally, all areas previously surveyed in 2016 and 2017 but without any fossiliferous material discovered ('non-localities') were also mapped. Clusters resulting from the *k*-means analysis can thus be compared with all the coordinates to uncover the patterns in the satellite images, such as any combination of spectral wavelength values having more likelihood of remotely detecting the fossil sites (Fig. 5).

All fossil sites containing vertebrate remains recorded in the area overlapped cluster 1 (Fig. 5), as well as the fossil wood sites and two out of three invertebrate sites (only GPL-9

did not match the pattern). However, the non-fossiliferous localities spread randomly over all clusters. Consequently, we decided to consider cluster 1 as a statistical classifier (i.e. predictive cluster) of potential new fossil sites to guide us during survey and thus increase the likelihood of finding sites. As a statistical classifier, cluster 1 cells that do not overlap with a fossil site are considered FP, which is required information to evaluate accuracy and other performance metrics. It should be noted that before being targeted a considerable range of this same cluster remained mostly unexplored, representing only 4.49% of the total area of the cropped image analysed (1,795 pixels out of 40,000). This severe reduction in total space to prospect—from 36 km^2 to roughly 1.6 km^2—is one of the main advantages of geospatial paleontology approaches. Prior work has demonstrated that the likelihood of success in locating fossil localities can be greatly increased by highlighting priority targets for prospecting (*Oheim, 2007*; *Egeland, Nicholson & Gasparian, 2010*; *Anemone, Emerson & Conroy, 2011*; *Conroy et al., 2018*). The spectral range of each cluster generated by *k*-means was also compared with the spectral range of the fossil sites. This again shows us very clearly that cluster 1 matches best the pattern exhibited by the fossiliferous deposits at the lower Mazamba Fm, especially in the NIR, SWIR1, and SWIR2 wavelength regions of the spectral band (Fig. 6).

Ground-truthing

During the 2018 season, four new fossil localities were discovered (GPL-10, 11, 12 and 12B), 3 out of 4 completely overlapped with cluster 1, the cluster identified as more likely to enable fossil discovery (Fig. 7). However, the only "miss" (GPL-12, a false negative, since it was not detected by the algorithm), was less than 90 m from a high concentration of white pixels (cluster 1), and thus it was highly likely to be detected following this survey protocol, since it was in the same gully system as GPL-12B—which is detected by *k*-means, and thus was extensively surveyed. Therefore, it can be argued that all the new sites were ultimately discovered by using the *k*-means surveying workflow, including the only one that does not meet the stricter criteria to be considered a true positive, since its discovery was a by-product of surveying guided by the algorithm to that particular area containing a high density of white pixels in proximity, and where GPL-12B was found. There are two possible interpretations, either (1) the buffer areas around a high concentration of cluster 1 pixels might also be more likely to be fossiliferous or (2) the algorithm simply failed to detect this site as the spectral signature differed from other fossil sites. Considering the expressly dense canopy at GPL-12 (Fig. 3), the second is likely the most parsimonious explanation. To be explicit, GPL-9 and -12 were considered false negatives (Table 2A) since cluster 1 of the *k*-means algorithm was defined as the predictive cluster, and it fails to detect them.

The new outcrops GPL-10 and GPL-11 were located at two river cut-banks (ca. 500 m apart) exposing a series of sandstone beds and they are both rich in invertebrate remains, but vertebrates were also found at the latter. They have been interpreted as representing a coastal delta-plain and fluvio-deltaic to marginal marine conditions, respectively. GPL-12 and GPL-12B are two sites in a gully system that might be continuously exposing fossiliferous sediments from at least three different sandstone layers for more than 100 m

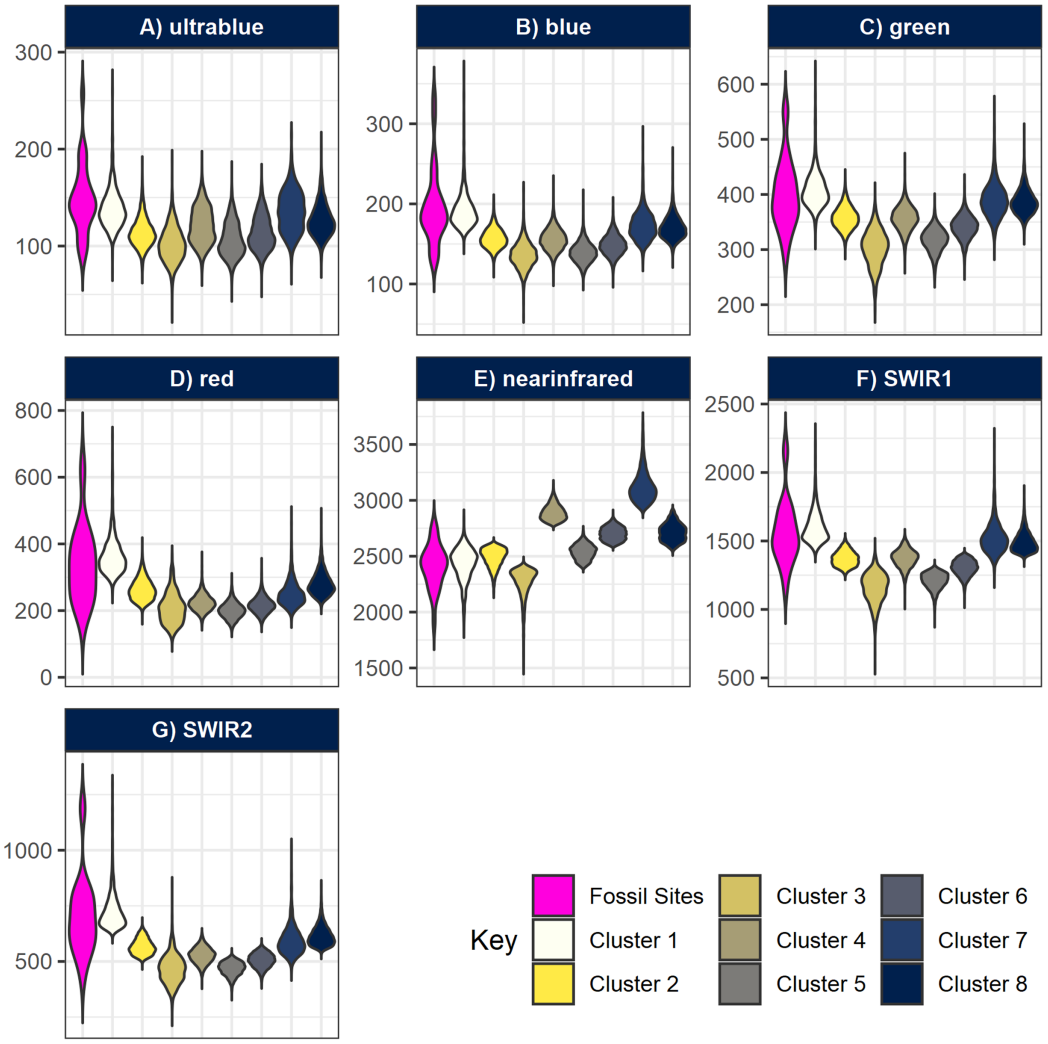

Figure 6 Violin-boxplots comparing sample distribution of spectral bands between clusters. Spectral bands (A) ultrablue; (B) blue; (C) green; (D) red; (E) infrared; (F) short-wavelength infrared 1; and (G) short-wavelength infrared 2 are represented with nine bars comparing the range of spectral bands values at the geocoordinates of the fossil sites, plus the eight clusters generated by *k*-means. Notice how overall, cluster 1 tends to approximate better the true spectral range of known fossil sites in Gorongosa.
Full-size DOI: 10.7717/peerj.11573/fig-6

(from north to south). These localities yield a remarkably well-preserved fossil record, including in situ and surface finds. Hundreds of identifiable fossil mammals, as well as crocodiles, turtles, and fish, have been recovered from this new area, and it has been tentatively inferred to be a river-dominated estuarine context (*Bobe & Carvalho, 2019a*, *2019b*; *Bobe, Martínez & Carvalho, 2020*).

Our approach had a high overall accuracy of 84.6%. Statistical measures of performance for model evaluation (defined in Table 2B) such as the negative predictive value (NPV = 96.88%) indicate that the model was able to detect almost perfectly the vast majority of TN 'non-localities'. In contrast, a low 55.56% precision shows the model overestimated false positives. On the other hand, the model demonstrated a high level of

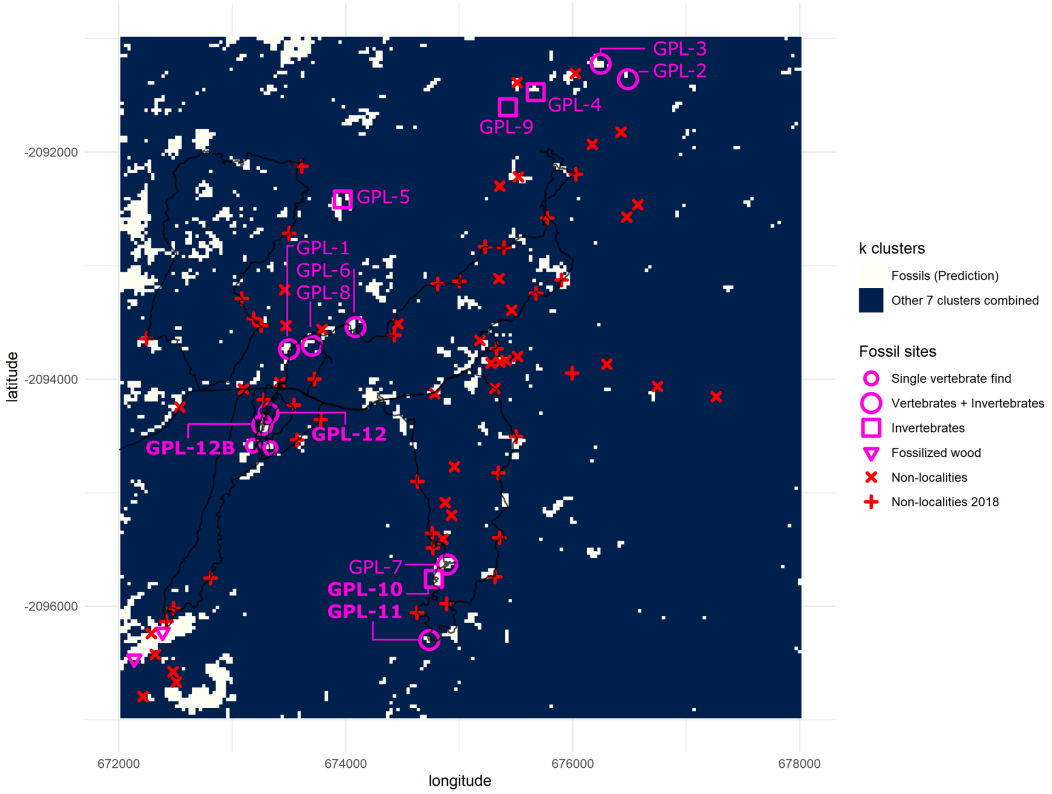

Figure 7 Binarized classification plotting cluster 1 versus all other clusters. New fossil sites GPL-10, 11, 12 and 12B are documented here for the first time. Trackways of surveys during 2018 are drawn in black. Clusters 2–8 are merged into a single cluster and compared against cluster 1 (predictive cluster). Total area = 36 km^2. One grid square = 1 km^2. One pixel-cell = 900 m^2.

Full-size DOI: 10.7717/peerj.11573/fig-7

sensitivity (88.24%), because of the very low number of false negatives generated, since only two fossil sites from the database did not match cluster 1 (see Fig. 7). Other metrics like specificity performed well (83.78%), albeit slightly lower than sensitivity, again due to the high number of FP detections, and this performance metric fared much better than the precision metric because it takes into account the model's ability to correctly detect TN (see Table 2).

The supervised random forest model (*Breiman, 2001*; *Liaw & Wiener, 2002*) was built to assess the relative importance (%IncMSE) of spectral variables in the clustering analysis (Fig. 8). It can thus be demonstrated that for this region of the lower Mazamba Formation the bands in the visible spectrum (0.45–0.67 µm) are the worst predictors of different clusters, while the near-infrared (NIR) is the variable that most contributed to the clustering. This suggests that visual inspection of satellite images (usually depicted in RGB bands) is far from being ideal for this particular task (finding fossil sites). In addition, the bands sampling longer wavelengths in the electromagnetic spectrum (0.85–2.29 µm), from NIR to short-wave infrared 2 (SWIR 2), are notably important at improving detection of cluster 1, and thus are also likely indicators of fossil sites (Fig. 8).

Table 2 Model goodness-of-fit (A) confusion matrix; (B) performance metrics.

A)

	Real +	Real −
Predicted +	TP=15	FP=12
Predicted −	FN=2	FP=12

B)

Measure	Derivations	Value
Sensitivity	$TPR = TP/(TP + FN)$	0.8824
Specificity	$SPC = TN/(FP + TN)$	0.8378
Precision	$PPV = TP/(TP + FP)$	0.5556
Negative predictive value	$NPV = TN/(TN + FN)$	0.9688
Accuracy	$ACC = (TP + TN)/(P + N)$	0.8462

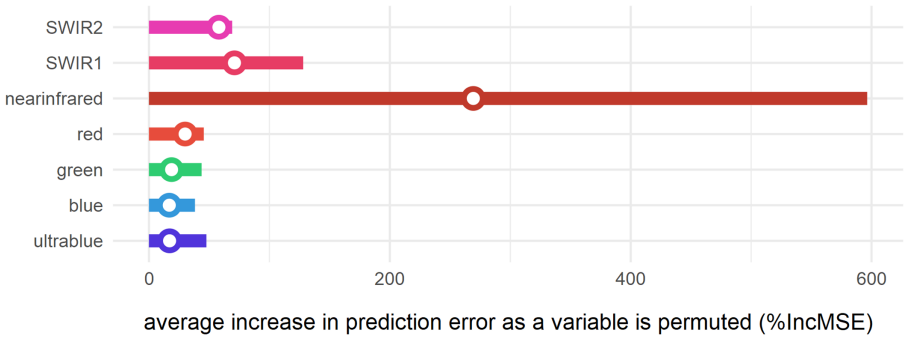

Figure 8 **Variable importance of spectral bands for clustering.** Bars represent relative importance of the spectral predictors for optimally classifying all clusters as calculated by a supervised random forest algorithm (*Breiman, 2001*). Specific variable importance for detecting cluster 1 is shown with open circles. Full-size DOI: 10.7717/peerj.11573/fig-8

DISCUSSION

Four new fossil sites from the late Miocene of southerneast Africa have been discovered by remote unsupervised learning. This analysis has also succeeded in pointing out the highest priority regions for future fossil prospecting in the lower Mazamba Formation of Gorongosa. In terms of overall accuracy, our unsupervised approach yielded similar results (84.6%) when compared with other implementations for remote fossil site detection, using supervised neural networks (84.21% with an hold-out sample in *Anemone, Emerson & Conroy, 2011*; *Emerson & Anemone, 2012*) and object-based binary classifier (73.1% with an hold-out sample in *Emerson et al., 2015*). Notice that in an unsupervised pipeline like *k*-means, by definition, all the sample is hold-out as it has not been used to train the algorithm. However, overestimation of false positives (precision = 55.56%) needs to be improved, this can be done in future iterations using different model architectures that also learn from data collected in this campaign. The non-localities are particularly helpful as they can be used as negative weights in models to improve precision.

A limitation of unsupervised methods is that the results are not dependent on any target, since a cluster might not fully match a true cover class. Therefore, a cluster is a subjective entity whose significance and interpretation requires domain knowledge, in this case, the field researchers comparing the different areas targeted by the cluster of interest to understand what they have in common (*Jain, 2010*). By visually inspecting satellite imagery of higher resolution and having visited multiple points of interest, we suggest that cluster 1 is detecting outcrops with soil erosion, or slightly less vegetated areas that occasionally contain sediment exposures (sometimes hidden under forest canopy). In Gorongosa, the sediment exposures are usually along river valleys or gullies subject to fluvial incision, and the most productive paleontological localities were found at the upstream terminations of tributary channels. In a modern miombo woodland context, like much of Gorongosa, knowing before-hand where to go and find such rare contexts might be sufficient to increase considerably the chances of finding fossiliferous material, simply as a consequence of improved visibility for surveying in the points highlighted remotely through machine learning approaches.

Our variable importance analysis using the random forest algorithm (*Breiman, 2001*; *Liaw & Wiener, 2002*) showed the key importance of longer wavelengths for detecting these promising spots within a forest to woodland setting. Infrared reflectance has been used for more than a century to analyse rocks and minerals (e.g. *Coblentz, 1906*) and its applications in geology are well established (*Lyon & Burns, 1963*). More recently, near-infrared remote sensing has been applied in soil sciences for multiple purposes, including: mapping proportions of sand, silt, and clay content (*Saleh, Belal & Arafat, 2013*; *Silva et al., 2016*); determination of soil salinity (*Feyziyev et al., 2016*); monitoring soil moisture and capacity of water absorption (*Ben-Dor et al., 2002*; *Whiting, Li & Ustin, 2004*); estimate organic carbon content (*Hu, Chau & Si, 2015*; *Viscarra Rossel & Hicks, 2015*); differentiate types of clay minerals (*Surech, Sreenivas & Sivasamy, 2014*); and assessing soil contamination and detection of heavy metals (*Mohamed et al., 2016*, *2018*). Yet, until now, machine learning approaches in geospatial paleontology tended to achieve good results but as "black-boxes" that did not reveal how exactly they were interpreting a "fossiliferous signal", and thus the relative importance of specific spectral bands for remote fossil site detection has not previously been demonstrated. Here, we show another application of NIR in remote sensing: increasing probability of finding fossils by clustering spectral reflectance values (a proxy for soil properties) that are likely to enhance fossil discovery. While the specific properties are hard to discern at this level of resolution (both spatial and spectral) the most likely candidates are reduced biomass (low foliage and vegetation, are proxies for visibility), and detection by clustering of similar soil types and mineral contents (*Mohamed et al., 2018*). SWIR1 and SWIR2 variables were also important for the model and higher values are associated with higher soil moisture, and thus are a proxy for water erosion of the soils. This result makes sense considering that most fossil sites in Gorongosa are close to the upstream terminations of the modern drainage system of the region. Erosion proxies have been used before in remote sensing models for fossil site discovery (e.g. *Block et al., 2016*; *Wills, Choiniere & Barrett, 2018*). This suggests

that the predictive cluster is remotely identifying the same environmental features that Habermann and colleagues observed in the field regarding the "rock exposures most commonly provided by incised gullies at head regions of small tributary channels or by channel flanks, are highly localized due to dense ground vegetation" (2019:727).

Our approach is likely reliable and replicable in other areas of dense vegetation, but it might have limited potential when applied in other areas of the Rift that are drier and have much less vegetation cover and consequently, more exposure. However, previous implementations of unsupervised clustering approaches in modern deserts of North America have been successful at finding Eocene deposits with early primate fauna (*Conroy, 2014*; *Conroy et al., 2018*), thus indicating that similar approaches are likely to be useful throughout the EARS. One last issue with *k*-means is that it creates equiprobable clusters, and thus might fail to properly represent the spectral signatures of the landcovers in certain surveying areas, so other unsupervised learning models without this characteristic might also improve the current results (*Schubert et al., 2017*).

CONCLUSIONS

The *k*-means approach performed well with high overall accuracy, contributing to the discovery of four new fossil sites, albeit with some limitations (see precision metric). To address this constraint, in future field seasons a set of different modelling approaches will be implemented using different architectures and other features besides the spectral bands, such as geo-ecological variables (elevation, slope, aspect, vegetation index, etc.). Importantly, the 2018 field season yielded several new sites as well as numerous negative coordinates (non-localities), and these data are vital to implement supervised approaches, which are more robust and may achieve better results than the method presented here. Considering NIR is usually interpreted as a proxy to biomass content in the soils (*Hu, Chau & Si, 2015*; *Viscarra Rossel & Hicks, 2015*), ground-truthing future implementations of different machine learning models should be done in the peak of the dry season in Gorongosa and/or after the wildfires, to further enhance visibility of topographic features, outcrops, and exposures of sedimentary rock. It is still too soon to fully understand how much remote sensing and unsupervised learning methods can be generalized to the problem of remote fossil site detection, but with an accuracy of 84.6% and new fossil sites detected the potential is clear.

What we do know, and what we can know about evolutionary processes is constrained by the fossil record. Indeed, paleontological inference is dependent—as are all historical disciplines—on the contexts that produce the data. The way to avoid biased inferences is thus to find more contexts to sample, so that our data do not increase only in quantity but also in representativeness of the natural and historic processes intricately linked to evolution. Early hominin occupation of humid and watered habitats such as riparian woodlands next to lake margins is well-documented across the basins of the EARS (*Cerling et al., 2011*; *Behrensmeyer & Reed, 2013*). But Gorongosa is the first of its kind, a rift valley estuarine and littoral site, partially covered by marine deposits (*Habermann et al., 2019*). Finding new fossil sites in under-studied African paleo-biomes of the late Miocene,

such as coastal forests in estuaries, is crucial in order to test critical hypotheses related to early hominin evolution in such contexts (*Kingdon, 2003*; *Wrangham, 2005*; *Joordens et al., 2019*; *Bobe, Martínez & Carvalho, 2020*). Future studies should use unsupervised learning algorithms coupled with ground-truthing as these are likely to become key tools to identify areas worth prospecting, leading to discovery of fossil-bearing localities. As these and other computer vision approaches improve, they will directly tackle some of the main limiting factors for paleontological studies: sample size, geographic gaps, and temporal biases, and thereby allowing substantial progress in the pace and content of paleontological and paleoanthropological discoveries.

ACKNOWLEDGEMENTS

We want to thank Dr. René Bobe and Dr. Thomas Püschel for insightful comments on early drafts of the manuscript. We are very grateful to the reviewers, Dr. Anna Behrensmeyer and Dr. Jonah Choiniere, for their insightful comments which helped greatly improve this article. Jd'OC thanks Dr. Jörg Habermann and Mr. Nhampoca João for all the incredible help during the 2018 field season. Our work is only possible due to the visionary approach of Greg Carr and the dedicated staff from Gorongosa National Park, guided by Dr. Mateus Mutemba and Pedro Muagara. We are very grateful to all the park rangers, our students, and colleagues across all our institutions who have been very enthusiastic about this project.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Portuguese Foundation for Science and Technology (FCT)—Grant SFRH/BD/122306/2016—and the field work was supported by The Boise Trust Fund. The Paleo-Primate Project Gorongosa received support from the Gorongosa Restoration Project, the National Geographic Society, the John Fell Fund Oxford, and the Leverhulme Trust. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Portuguese Foundation for Science and Technology (FCT): SFRH/BD/122306/2016.
The Boise Trust Fund.
Gorongosa Restoration Project.
National Geographic Society.
John Fell Fund Oxford.
Leverhulme Trust.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- João d'Oliveira Coelho conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Robert L. Anemone conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Susana Carvalho conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

Field Study Permissions

The following information was supplied relating to field study approvals (i.e., approving body and any reference numbers):

Field experiments were approved by Dr Marc Stalmans, Director of the Department of Scientific Services of Gorongosa National Park in Mozambique (project number: PNG/DSCi/C019/2018).

Data Availability

The following information was supplied regarding data availability:

Files to replicate the kmeans+randomforest pipeline for detecting fossil sites in Gorongosa are available at GitHub: https://github.com/Delvis/kmeansGorongosa.

Spectral data from Landsat and other medium-resolution satellites is updated daily and can be accessed at the USGS Earth Explorer website (https://earthexplorer.usgs.gov/).

REFERENCES

Andrews P. 2019. Last common ancestor of apes and humans: morphology and environment. *Folia Primatologica* **91(2)**:1–27 DOI 10.1159/000501557.

Anemone RL, Conroy GC. 2018. Geospatial anthropology: integrating remote sensing and geographic information sciences into anthropological fieldwork and analysis. In: Anemone RL, Conroy GC, eds. *New Geospatial Approaches to the Anthropological Sciences*. Albuquerque, NM: New Mexico Press, 1–20.

Anemone RL, Emerson CW, Conroy GC. 2011. Finding fossils in new ways: an artificial neural network approach to predicting the location of productive fossil localities. *Evolutionary Anthropology* **20(5)**:169–180 DOI 10.1002/evan.20324.

Asfaw B, Ebinger C, Harding D, White TD, Woldegabriel G. 1990. Space based imagery in paleoanthropological research: an Ethiopian example. *National Geographic Research* **6**:418–434.

Äyrämö S, Kärkkäinen T. 2006. Introduction to partitioning-based clustering methods with a robust example. In: *Reports of the Department of Mathematical Information Technology—Series C: Software Engineering and Computational Intelligence*. Jyväskylä: University of Jyväskylä.

Ball GH, Hall DJ. 1965. *ISODATA: a novel method of data analysis and pattern classification*. Menlo Park, CA: Stanford Research Institute.

Barba-Montoya J, Dos Reis M, Yang Z. 2017. Comparison of different strategies for using fossil calibrations to generate the time prior in Bayesian molecular clock dating. *Molecular Phylogenetics and Evolution* **114**:386–400 DOI 10.1016/j.ympev.2017.07.005.

Behrensmeyer AK, Reed KE. 2013. Reconstructing the habitats of australopithecus: paleoenvironments, site taphonomy, and faunas. In: Reed KE, Fleagle JG, Leakey RE, eds. *The*

Paleobiology of Australopithecus—Vertebrate Paleobiology and Paleoanthropology. Dordrecht: Springer Netherlands, 41–60.

Ben-Dor E, Patkin K, Banin A, Karnieli A. 2002. Mapping of several soil properties using DAIS-7915 hyperspectral scanner data—a case study over clayey soils in Israel. *International Journal of Remote Sensing* **23(6)**:1043–1062 DOI 10.1080/01431160010006962.

Block S, Saltré F, Rodríguez-Rey M, Fordham DA, Unkel I, Bradshaw CJA. 2016. Where to dig for fossils: combining climate-envelope, taphonomy and discovery models. *PLOS ONE* **11(3)**:e0151090 DOI 10.1371/journal.pone.0151090.

Bobe R, Aldeias V, Alemseged Z, Archer W, Aumaître G, Bamford MK, Biro D, Bourlès DL, Braun DR, Capelli C, d'Oliveira Coelho J, Habermann J, Keddadouche K, Kupczik K, Lebatard A-E, Lüdecke T, Macôa A, Martínez FI, Mathe J, Mendes C, Paulo LM, Pinto M, Püschel TA, Regala FT, Sier M, Silva MJF, Stalmans M, Carvalho S. 2021. A new window on the evolution of Africa's ancient ecosystems: Fossil vertebrates and paleoenvironments from Gorongosa National Park, Mozambique. Manuscript in preparation.

Bobe R, Alemseged Z, Bamford MK, Carvalho S. 2018. New Mio-Pliocene fossil sites from Gorongosa National Park and the biogeography of hominin origins. In: *Abstracts of the International Primatological Society Congress*. Event 713. Nairobi.

Bobe R, Carvalho S. 2019a. Gorongosa National Park: a new window on the late Miocene at the southern end of the African Rift Valley. In: *Proceedings of the European Society for the study of Human Evolution*. 20.

Bobe R, Carvalho S. 2019b. Late Miocene primates and the biogeography of hominin origins: a role for the unknown south? In: *European Federation of Primatology/Primatological Society of Great Britain*. Oxford.

Bobe R, d'Oliveira Coelho J, Carvalho S, Leakey M. Fauna and paleoenvironments of the Koobi Fora Formation, Kenya. In: Reynolds SC, Bobe R, eds. *African Paleoecology and Human Evolution*. Cambridge: Cambridge University Press. (in press).

Bobe R, Martínez FI, Carvalho S. 2020. Primate adaptations and evolution in the Southern African Rift Valley. *Evolutionary Anthropology: Issues, News, and Reviews* **evan.21826(3)**:94–101 DOI 10.1002/evan.21826.

Bock H-H. 2008. Origins and extensions of the k-means algorithm in cluster analysis. *Electronic Journal for History of Probability and Statistics* **4**:1–18.

Böhme B, Steinbruch F, Gloaguen R, Heilmeier H, Merkel B. 2006. Geomorphology, hydrology, and ecology of Lake Urema, central Mozambique, with focus on lake extent changes. *Physics and Chemistry of the Earth, Parts A/B/C* **31(15–16)**:745–752 DOI 10.1016/j.pce.2006.08.010.

Breiman L. 2001. Random forests. *Machine Learning* **45(1)**:5–32 DOI 10.1023/A:1010933404324.

Brunet M, Guy F, Pilbeam D, Lieberman DE, Likius A, Mackaye HT, Ponce de León MS, Zollikofer CPE, Vignaud P. 2005. New material of the earliest hominid from the Upper Miocene of Chad. *Nature* **434(7034)**:752–755 DOI 10.1038/nature03392.

Brunet M, Guy F, Pilbeam D, Mackaye HT, Likius A, Ahounta D, Beauvilain A, Blondel C, Bocherens H, Boisserie J-R, Bonis LD, Coppens Y, Dejax J, Denys C, Duringer P, Eisenmann V, Fanone G, Fronty P, Geraads D, Lehmann T, Lihoreau F, Louchart A, Mahamat A, Merceron G, Mouchelin G, Otero O, Campomanes PP, Leon MPD, Rage J-C, Sapanet M, Schuster M, Sudre J, Tassy P, Valentin X, Vignaud P, Viriot L, Zazzo A, Zollikofer C. 2002. A new hominid from the Upper Miocene of Chad, Central Africa. *Nature* **418(6894)**:145–151 DOI 10.1038/nature00879.

Cerling TE, Wynn JG, Andanje SA, Bird MI, Korir DK, Levin NE, Mace W, Macharia AN, Quade J, Remien CH. 2011. Woody cover and hominin environments in the past 6 million years. *Nature* **476(7358)**:51–56 DOI 10.1038/nature10306.

Coblentz WW. 1906. Radiometric investigations of infra-red absorption and reflection spectra. *Bulletin of the Bureau of Standards* **2(3)**:457–462 DOI 10.6028/bulletin.048.

Conroy GC. 2006. Creating, displaying, and querying interactive paleoanthropological maps using GIS: an example from the Uinta Basin, Utah. *Evolutionary Anthropology: Issues, News, and Reviews* **15(6)**:217–223 DOI 10.1002/evan.20111.

Conroy GC. 2014. Walking back the cat: Unsupervised classification as an aid in "remote" fossil prospecting. *Evolutionary Anthropology: Issues, News, and Reviews* **23(5)**:172–176 DOI 10.1002/evan.21422.

Conroy GC, Chew A, Rose KD, Bown TM, Anemone RL, Gunnell GF. 2018. Assessing unsupervised image classification as an aid in paleoanthropological explorations. In: Anemone RL, Conroy GC, eds. *New Geospatial Approaches to the Anthropological Sciences*. Albuquerque, NM: New Mexico Press, 59–80.

Conroy GC, Emerson CW, Anemone RL, Townsend KEB. 2012. Let your fingers do the walking: a simple spectral signature model for "remote" fossil prospecting. *Journal of Human Evolution* **63(1)**:79–84 DOI 10.1016/j.jhevol.2012.04.002.

Egeland C, Nicholson C, Gasparian B. 2010. Using GIS and ecological variables to identify high potential areas for paleoanthropological survey: an example from Northern Armenia. *Journal of Ecological Anthropology* **14(1)**:89–98 DOI 10.5038/2162-4593.14.1.8.

Emerson CW, Anemone RL. 2012. An artificial neural network-based approach to identifying mammalian fossil localities in the Great Divide Basin, Wyoming. *Remote Sensing Letters* **3(5)**:453–460 DOI 10.1080/01431161.2011.621463.

Emerson CW, Bommersbach B, Nachman B, Anemone RL. 2015. An object-oriented approach to extracting productive fossil localities from remotely sensed imagery. *Remote Sensing* **7(12)**:16555–16570 DOI 10.3390/rs71215848.

Feyziyev F, Babayev M, Priori S, L'Abate G. 2016. Using visible-near infrared spectroscopy to predict soil properties of Mugan Plain, Azerbaijan. *Open Journal of Soil Science* **6(03)**:52–58 DOI 10.4236/ojss.2016.63006.

Fonseca JFBD, Chamussa J, Domingues A, Helffrich G, Antunes E, Van Aswegen G, Pinto LV, Custódio S, Manhiça VJ. 2014. MOZART: a seismological investigation of the East African Rift in Central Mozambique. *Seismological Research Letters* **85(1)**:108–116 DOI 10.1785/0220130082.

Forgy EW. 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* **21**:768–769.

Grantham GH, Marques JM, Wilson MGC, Manhiça V, Hartzer FJ. 2011. *Explanation of the Geological Map of Mozambique, 1:1,000,000*. Maputo: National Directorate of Geology.

Habermann JM, Alberti M, Aldeias V, Alemseged Z, Archer W, Bamford M, Biro D, Braun DR, Capelli C, Cunha E, Da Silva MF, Lüdecke T, Madiquida H, Martinez FI, Mathe J, Negash E, Paulo LM, Pinto M, Stalmans M, Regala FT, Wynn JG, Bobe R, Carvalho S. 2019. Gorongosa by the sea: first miocene fossil sites from the Urema Rift, central Mozambique, and their coastal paleoenvironmental and paleoecological contexts. *Palaeogeography, Palaeoclimatology, Palaeoecology* **514**:723–738 DOI 10.1016/j.palaeo.2018.09.032.

Haile-Selassie Y. 2001. Late Miocene hominids from the Middle Awash, Ethiopia. *Nature* **412(6843)**:178–181 DOI 10.1038/35084063.

Haile-Selassie Y, Suwa G, White TD. 2004. Late Miocene teeth from Middle Awash, Ethiopia, and early hominid dental evolution. *Science* **303(5663)**:1503–1505 DOI 10.1126/science.1092978.

Hlusko LJ. 2018. Geospatial approaches to hominid paleontology in Africa. In: Anemone RL, Conroy GC, eds. *New Geospatial Approaches to the Anthropological Sciences*. Albuquerque, NM: New Mexico Press, 39–58.

Hu W, Chau HW, Si BC. 2015. Vis-near IR reflectance spectroscopy for soil organic carbon content measurement in the Canadian Prairies. *CLEAN—Soil, Air, Water* **43(8)**:1215–1223 DOI 10.1002/clen.201400400.

Ishida H, Pickford M. 1997. A new Late Miocene hominoid from Kenya: *Samburupithecus kiptalami* gen. et sp. nov. Comptes Rendus de l'Académie des sciences—series IIA. *Earth and Planetary Science* **325(10)**:823–829 DOI 10.1016/S1251-8050(97)82762-0.

Jain AK. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* **31(8)**:651–666 DOI 10.1016/j.patrec.2009.09.011.

Joordens JCA, Feibel CS, Vonhof HB, Schulp AS, Kroon D. 2019. Relevance of the eastern African coastal forest for early hominin biogeography. *Journal of Human Evolution* **131(5760)**:176–202 DOI 10.1016/j.jhevol.2019.03.012.

Joordens JCA, Wesselingh FP, De Vos J, Vonhof HB, Kroon D. 2009. Relevance of aquatic environments for hominins: a case study from Trinil (Java, Indonesia). *Journal of Human Evolution* **57(6)**:656–671 DOI 10.1016/j.jhevol.2009.06.003.

Kingdon J. 2003. *Lowly origin: where, when, and why our ancestors first stood up*. Princeton: Princeton University Press.

Kunimatsu Y, Nakatsukasa M, Sawada Y, Sakai T, Hyodo M, Hyodo H, Itaya T, Nakaya H, Saegusa H, Mazurier A, Saneyoshi M, Tsujikawa H, Yamamoto A, Mbua E. 2007. A new Late Miocene great ape from Kenya and its implications for the origins of African great apes and humans. *Proceedings of the National Academy of Sciences of the United States of America* **104(49)**:19220–19225 DOI 10.1073/pnas.0706190104.

Kunimatsu Y, Nakatsukasa M, Sawada Y, Sakai T, Saneyoshi M, Nakaya H, Yamamoto A, Mbua E. 2016. A second hominoid species in the early Late Miocene fauna of Nakali (Kenya). *Anthropological Science* **124(2)**:75–83 DOI 10.1537/ase.160331.

Leakey LSB. 1959. A new fossil skull from olduvai. *Nature* **184(4685)**:491–493 DOI 10.1038/184491a0.

Leakey MG, Walker AC. 2003. 6.2. The Lothagam Hominids. In: Leakey MG, Harris JM, eds. *Lothagam: The Dawn of Humanity in Eastern Africa*. New York Chichester, West Sussex: Columbia University Press, 249–257.

Lebatard A-E, Bourlès DL, Braucher R, Arnold M, Duringer P, Jolivet M, Moussa A, Deschamps P, Roquin C, Carcaillet J, Schuster M, Lihoreau F, Likius A, Mackaye HT, Vignaud P, Brunet M. 2010. Application of the authigenic 10Be/9Be dating method to continental sediments: reconstruction of the Mio-Pleistocene sedimentary sequence in the early hominid fossiliferous areas of the northern Chad Basin. *Earth and Planetary Science Letters* **297(1–2)**:57–70 DOI 10.1016/j.epsl.2010.06.003.

Liaw A, Wiener M. 2002. Classification and regression by randomForest. *R News* **2**:18–22.

Lloyd S. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28(2)**:129–137 DOI 10.1109/TIT.1982.1056489.

Lyon RJP, Burns EA. 1963. Analysis of rocks and minerals by reflected infrared radiation. *Economic Geology* **58(2)**:274–284 DOI 10.2113/gsecongeo.58.2.274.

Macgregor D. 2015. History of the development of the East African Rift system: a series of interpreted maps through time. *Journal of African Earth Sciences* **101(2/3/4/)**:232–252 DOI 10.1016/j.jafrearsci.2014.09.016.

MacQueen J. 1967. *Some methods for classification and analysis of multivariate observations*. California: The Regents of the University of California.

Malakhov DV, Dyke GJ, King C. 2009. Remote sensing applied to paleontology: exploration of Upper Cretaceous sediments in Kazakhstan for potential fossil sites. *Paleontologica Electronica* **12(3)**:12–13.

McBrearty S, Jablonski NG. 2005. First fossil chimpanzee. *Nature* **437(7055)**:105–108 DOI 10.1038/nature04008.

Mohamed ES, Ali AM, El Shirbeny MA, Abd El Razek AA, Savin IY. 2016. Near infrared spectroscopy techniques for soil contamination assessment in the Nile Delta. *Eurasian Soil Science* **49(6)**:632–639 DOI 10.1134/S1064229316060065.

Mohamed ES, Saleh AM, Belal AB, Gad A. 2018. Application of near-infrared reflectance for quantitative assessment of soil properties. *The Egyptian Journal of Remote Sensing and Space Science* **21(1)**:1–14 DOI 10.1016/j.ejrs.2017.02.001.

Montagnon D. 2013. Strepsirhine divergence dates estimated from mitochondrial gene sequences, and the status of Daubentonia madagascariensis. In: Masters J, Gamba M, Génin F, eds. *Leaping Ahead: Advances in Prosimian Biology. Developments in Primatology: Progress and Prospects*. New York, NY: Springer, 21–32.

Moorjani P, Amorim CEG, Arndt PF, Przeworski M. 2016. Variation in the molecular clock of primates. *Proceedings of the National Academy of Sciences of the United States of America* **113(38)**:10607–10612 DOI 10.1073/pnas.1600374113.

Njau JK, Hlusko LJ. 2010. Fine-tuning paleoanthropological reconnaissance with high-resolution satellite imagery: the discovery of 28 new sites in Tanzania. *Journal of Human Evolution* **59(6)**:680–684 DOI 10.1016/j.jhevol.2010.07.014.

Nowak K, Barnett AA, Matsuda I. 2019. *Primates in flooded habitats: ecology and conservation*. Cambridge: Cambridge University Press.

Oheim KB. 2007. Fossil site prediction using geographic information systems (GIS) and suitability analysis: the two medicine formation, MT, a test case. *Palaeogeography, Palaeoclimatology, Palaeoecology* **251(3–4)**:354–365 DOI 10.1016/j.palaeo.2007.04.005.

Pickford M. 2012. *Mozambique paleontology reconnaissance—November 2012*. Mozambique: Gorongosa National Park.

Pickford M. 2013. *Gorongosa palaeontology survey—November, 2013*. Mozambique: Gorongosa National Park.

Pickford M, Coppens Y, Senut B, Morales J, Braga J. 2009. Late Miocene hominoid from Niger. *Comptes Rendus Palevol* **8(4)**:413–425 DOI 10.1016/j.crpv.2008.11.003.

Pickford M, Senut B. 2005. Hominoid teeth with chimpanzee-and gorilla-like features from the Miocene of Kenya: implications for the chronology of ape-human divergence and biogeography of Miocene hominoids. *Anthropological Science* **113(1)**:95–102 DOI 10.1537/ase.04S014.

Pickford M, Senut B, Morales J, Braga J. 2008. First hominoid from the Late Miocene of Niger. *South African Journal of Science* **104**:337–339.

Pickford M, Senut B, Ssemmanda I, Elepu D, Obwona P. 1988. Premiers resultats de la mission de l'Uganda palaeontology expedition a Nkondo (Pliocene du bassin du lac Albert, Ouganda). *Comptes rendus de l'Académie des sciences—Série 2: Mécanique, Physique, Chimie, Sciences de l'univers, Sciences de la Terre* **306**:315–320.

Real F. 1966. *Geologia da bacia do rio Zambeze (Moçambique): características geológico-mineiras da bacia do rio Zambeze, em território Moçambicano*. Lisboa: Junta de Investigações do Ultramar.

Dos Reis M, Yang Z. 2019. Bayesian molecular clock dating using genome-scale datasets. In: Anisimova M, ed. *Evolutionary Genomics: Statistical and Computational Methods. Methods in Molecular Biology*. New York, NY: Springer, 309–330.

Renne PR, WoldeGabriel G, Hart WK, Heiken G, White TD. 1999. Chronostratigraphy of the Miocene-Pliocene Sagantole Formation, Middle Awash Valley, Afar rift, Ethiopia. *GSA Bulletin* **111**:869–885 DOI 10.1130/0016-7606(1999)111<0869:COTMPS>2.3.CO;2.

Roy DP, Wulder MA, Loveland TR, Ce W, Allen RG, Anderson MC, Helder D, Irons JR, Johnson DM, Kennedy R, Scambos TA, Schaaf CB, Schott JR, Sheng Y, Vermote EF, Belward AS, Bindschadler R, Cohen WB, Gao F, Hipple JD, Hostert P, Huntington J, Justice CO, Kilic A, Kovalskyy V, Lee ZP, Lymburner L, Masek JG, McCorkel J, Shuai Y, Trezza R, Vogelmann J, Wynne RH, Zhu Z. 2014. Landsat-8: science and product vision for terrestrial global change research. *Remote Sensing of Environment* **145(1)**:154–172 DOI 10.1016/j.rse.2014.02.001.

Saleh AM, Belal AB, Arafat SM. 2013. Identification and mapping of some soil types using field spectrometry and spectral mixture analyses: a case study of North Sinai, Egypt. *Arabian Journal of Geosciences* **6(6)**:1799–1806 DOI 10.1007/s12517-011-0501-6.

Sawada Y, Pickford M, Senut B, Itaya T, Hyodo M, Miura T, Kashine C, Chujo T, Fujii H. 2002. The age of Orrorin tugenensis, an early hominid from the Tugen Hills, Kenya. *Comptes Rendus Palevol* **1(5)**:293–303 DOI 10.1016/S1631-0683(02)00036-2.

Schubert E, Sander J, Ester M, Kriegel HP, Xu X. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems* **42**:19:1–19:21 DOI 10.1145/3068335.

Senut B. 2010. Upper Miocene hominoid distribution and the origin of hominids revisited. *Historical Biology* **22(1–3)**:260–267 DOI 10.1080/08912961003603175.

Senut B. 2015. The Miocene Hominoids and the Earliest Putative Hominids. In: Henke W, Tattersall I, eds. *Handbook of Paleoanthropology*. Berlin, Heidelberg: Springer.

Senut B, Pickford M, Gommery D. 2018. Dental anatomy of the early hominid, *Orrorin tugenensis*, from the Lukeino Formation, Tugen Hills, Kenya. *Revue de Paléobiologie* **37**:577–591.

Senut B, Pickford M, Gommery D, Mein P, Cheboi K, Coppens Y. 2001. First hominid from the Miocene (Lukeino Formation, Kenya). *Comptes Rendus de l'Académie des Sciences: Series IIA—Earth and Planetary Science* **332(2)**:137–144 DOI 10.1016/S1251-8050(01)01529-4.

Silva EB, Ten Caten A, Dalmolin RSD, Dotto AC, Silva WC, Giasson E. 2016. Estimating soil texture from a limited region of the visible/near-infrared spectrum. In: Hartemink AE, Minasny B, eds. *Digital Soil Morphometrics: Progress in Soil Science*. Cham: Springer International Publishing, 73–87.

Simpson SW, Kleinsasser L, Quade J, Levin NE, McIntosh WC, Dunbar N, Semaw S, Rogers MJ. 2015. Late Miocene hominin teeth from the Gona Paleoanthropological Research Project area, Afar, Ethiopia. *Journal of Human Evolution* **81(72)**:68–82 DOI 10.1016/j.jhevol.2014.07.004.

Steinbruch F. 2010. Geology and geomorphology of the Urema Graben with emphasis on the evolution of Lake Urema. *Journal of African Earth Sciences* **58(2)**:272–284 DOI 10.1016/j.jafrearsci.2010.03.007.

Steinhaus H. 1956. Sur la division des corp materiels en parties. *Bulletin De L Academie Polonaise Des Sciences: Serie Des Sciences Mathematiques Astronomiques Et Physiques* **1**:801–804.

Surech GJ, Sreenivas K, Sivasamy R. 2014. Hyperspectral analysis of clay minerals. *ISPRS—International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XL–8**:443–446 DOI 10.5194/isprsarchives-XL-8-443-2014.

Suwa G, Kono RT, Katoh S, Asfaw B, Beyene Y. 2007. A new species of great ape from the late Miocene epoch in Ethiopia. *Nature* **448(7156)**:921–924 DOI 10.1038/nature06113.

Tinley KL. 1977. Framework of Gorongosa ecosystem. Unpublished DSc thesis Thesis. South Africa: University of Pretoria.

Viscarra Rossel RA, Hicks WS. 2015. Soil organic carbon and its fractions estimated by visible-near infrared transfer functions. *European Journal of Soil Science* **66(3)**:438–450 DOI 10.1111/ejss.12237.

Whiting ML, Li L, Ustin SL. 2004. Predicting water content using Gaussian model on soil spectra. *Remote Sensing of Environment* **89(4)**:535–552 DOI 10.1016/j.rse.2003.11.009.

Wills S, Choiniere JN, Barrett PM. 2018. Predictive modelling of fossil-bearing locality distributions in the Elliot Formation (Upper Triassic–Lower Jurassic), South Africa, using a combined multivariate and spatial statistical analyses of present-day environmental data. *Palaeogeography, Palaeoclimatology, Palaeoecology* **489(3)**:186–197 DOI 10.1016/j.palaeo.2017.10.009.

Wrangham RW. 2005. The delta hypothesis: hominoid ecology and hominin origins. In: Lieberman DE, Smith RJ, Kelley J, eds. *Interpreting the Past: Essays on Human, Primate, and Mammal Evolution in Honor of David Pilbeam*. Boston: Brill Academic Publishers, 231–242.

Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D. 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems* **14(1)**:1–37 DOI 10.1007/s10115-007-0114-2.

Zhao W, Ma H, He Q. 2009. Parallel K-means clustering based on MapReduce. In: Jaatun MG, Zhao G, Rong C, eds. *Cloud Computing: Lecture Notes in Computer Science*. Berlin Heidelberg: Springer, 674–679.