XAI and geographic information: application to paleoenvironmental reconstructions

Matthieu Boussard, Zimmermann Bastien, Nicolas Boulbes, Sophie Grégoire

XAI and geographic information: application to paleoenvironmental reconstructions

Bastien Zimmermann[1], Matthieu Boussard[1], Nicolas Boulbes[2], Sophie Grégoire[3]

[1] craft.ai, France
[2] Institut de Paléontologie Humaine, Fondation Albert Ier, Paris, UMR 7194 HNHP, EPCC-CERP Tautavel, France
[3] UMR 7194 HNHP, EPCC-CERP Tautavel, France

Abstract. This work shows the contribution of explicability to the reconstruction of paleoenvironments. Integrating explanations to a predictive approach provides the domain expert additional information and trust. It aims at building a model allowing, from excavation data, to infer the environment corresponding to a given layer and a given paleolithic period. In this context, the prediction of the model alone has less value than the underlying explanations that allow archaeologists to question their assumptions. Due to the uncertainties in the data, this work focuses more on data-oriented explanations tools, such as data-Shapley. Finally, a contribution of this article is the use of a Geographic Information System, allowing us to exploit to the maximum the information we can obtain from the explainability tools.

Keywords: XAI, Explainable AI, Geographic Information System, Machine Learning, Animal Communities, Palaeoenvironments.

1 Introduction

Making Artificial Intelligence (AI) based systems reliable is one of the major challenges in the research and development of machine learning techniques. The reliability of a system is easily questioned when it works in an obscure way or if it is based on bad foundations, such as bad data. Additionally, the adoption of artificial intelligence approach in other fields is hindered by many issues such as reproducibility, data quality, or even metric choice [8]. To overcome this, it is essential to provide trust. Explainable artificial intelligence (XAI) is one of the pillars for achieving trusted AI. In particular, data-centric explanatory methods are part of an approach to refocusing attention on the data that is as important or more important than the model in the machine learning process. Indeed, corrupted data or data used in the wrong context lead to wrong conclusions. In this context, algorithms that evaluate the quality of data points can be used to more effectively direct the work of cleaning, increasing and improving data quality and thus build better performing models. Here we have used **Beta-Shapley** on archaeological excavation data, worked within the framework of the

ANR SCHOPPER program [6], in order to assign a value to our different data points.

The objective is to use a machine learning algorithm to predict a biome (a geographically positioned ecological unit) based on the characteristic animal species of that biome or the ecoregions that comprise it. Fossil remains of animals found in an archaeological layer are used to estimate, based on the principle of actualism, the associated biome and consequently the climatic conditions that prevailed at the time the level was established.

Beta-Shapley allows for an upstream analysis of this process and analysis of what is used to drive our model. In a first step, the tool's relevance was evaluated by identifying high and low-quality data. We show the consistency between the point value and its impact on the performance of a model. Then a more detailed analysis highlights the different information that this tool can provide. Finally, the explanations provided by these tools allow us to see our data in a new light and open up new perspectives, in particular via the exploitation of the spatial dimension. The growing use of geographic information systems coupled with species distribution models (Ecological Niche Modelling) for the study of paleoenvironments [14] justifies the development of explainable machine learning methods on this type of tool.

1.1 Context

In order to provide a concrete illustration of the contributions of XAI tools, and more specifically data-oriented XAI tools, we have chosen the example of paleoenvironmental reconstruction from an archaeological site. The aim is to determine the environment and climate corresponding to a given period by considering certain biological clues identified during archaeological excavations. The site used is the Palaeolithic cave of the *Caune de l'Arago* in Tautavel, in the south of France. It has benefited from 54 years of excavations and multidisciplinary studies of a 15-metre-thick stratigraphic sequence, developed between 690,000 and 90,000 years ago BP. [1]. This site has yielded nearly 600,000 artefacts in 55 archaeological levels. The richness of this Palaeolithic record, the quality of the conservation of the remains and the standardized data recording system from which it benefits, make it one of the best fields of application of the tools presented here for palaeoenvironmental reconstructions. Paleoenvironmental reconstructions are based on the principle of actualism. It is based on the assumption that biological systems in the past functioned in the same way as those that can be observed today. For example, if it is possible to observe the current distribution of different animal species for different regions of the globe, it is theoretically possible to determine the climate of the past from a fossil faunal assemblage thanks to the known ecological affinities of current animal species by transposing them to fossil animal communities. This principle makes it possible to identify the climate corresponding to a given archaeological layer, based on the remains of bones found in that layer. However, many risk factors can lead to an erroneous conclusion, such as:

– the quality of the current dataset
– the representativeness of the taxa found during excavations (conservation status and predation bias)
– the evolution of species (adaptation, migration, extinction)
– the principle of actualism.

Machine learning approaches for reconstructing paleoenvironments have been proposed in [12] from pollen data. We propose here to use models based on faunal data. More fundamentally, where the authors focus on the predictive aspect of the models, we are interested here in the explanations of predictions that machine learning can provide.

2 Data presentation

The data used in this article are of two kinds. One is current, used as a reference for the constitution of the models, and the other is archaeological, with the aim of classifying them using machine learning models and thus defining the associated environmental conditions.

2.1 Datasets on current environments

The actualistic dataset used is the *wildfinder dataset* [2], which is a biogeographical representation of terrestrial biodiversity. The basic unit is the ecoregion as represented in Fig.1 is established according to biogeographic criteria, defined as "an extended unit of land or water that contains a geographically distinct assemblage of species, natural communities and environmental conditions". Each

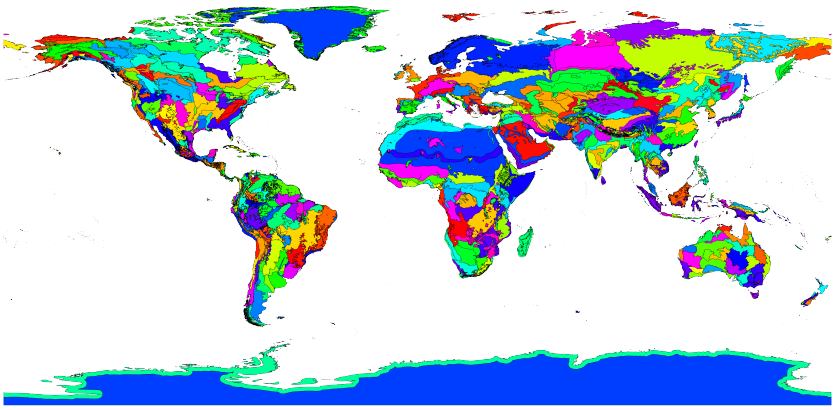


Fig. 1: The world's ecoregions

ecoregion contains a list of the more than 26,000 species present, as well as an

associated biome type. The latter is a set of ecosystems characteristic of a biogeographical area. Each of the 14 biomes is representative of a certain climate (Fig. 2). An even more general level of description exists, that of the ecozones, of which there are eight, representing the distribution of the current fauna on the planet. The two that we consider in this work are the Palearctic, corresponding to Europe, North Africa, the northern two-thirds of Asia and the Middle East (except Arabia), and the Nearctic, corresponding to most of North America, i.e. the ecozones of the Northern Hemisphere. The wealth of data is presented in Fig.3. The different biomes are not equally represented, there are more *Temperate Broadleaf and Mixed Forests* ecoregions than *Montane Grasslands and Shrublands*. In addition, some biomes have a lower species diversity, with fewer species present on average in the Tundra ecoregions than in the Temperate Coniferous Forests.

The first usable data explainability tool is data visualization. The figure Fig.4 allows for alternative representations. The use of PCA (Principal Component Analysis) and t-SNE (t-distributed Stochastic Neighbour Embedding) makes it possible to obtain a relevant representation of the data in two dimensions. Indeed, PCA makes it possible to transform the variables between space to keep only the two principal components, which are decorrelated from the others and explain the variance as well as possible. The t-SNE algorithm, on the other hand, has the characteristic of preserving the proximity of the points during the dimension-reducing transformation.

These two high-level representations allow us to see that certain clusters exist and to appreciate the Vapnik-Chervonenkis dimension of our data (the theory of the same name aims to explain learning from a statistical point of view). Indeed, it seems a priori possible to distinguish the different groupings of biomes.

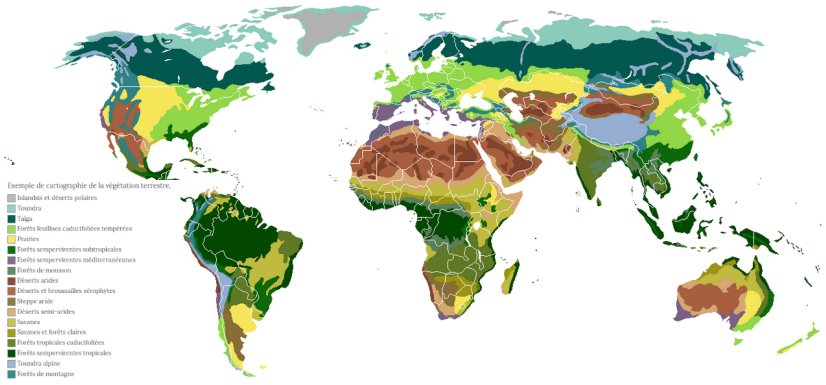


Fig. 2: World's biomes

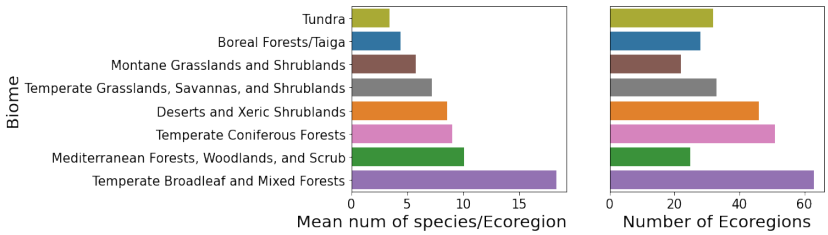


Fig. 3: Number of species by ecoregion and number of ecoregion per biome (restricted to the *Caune de l'Arago* species

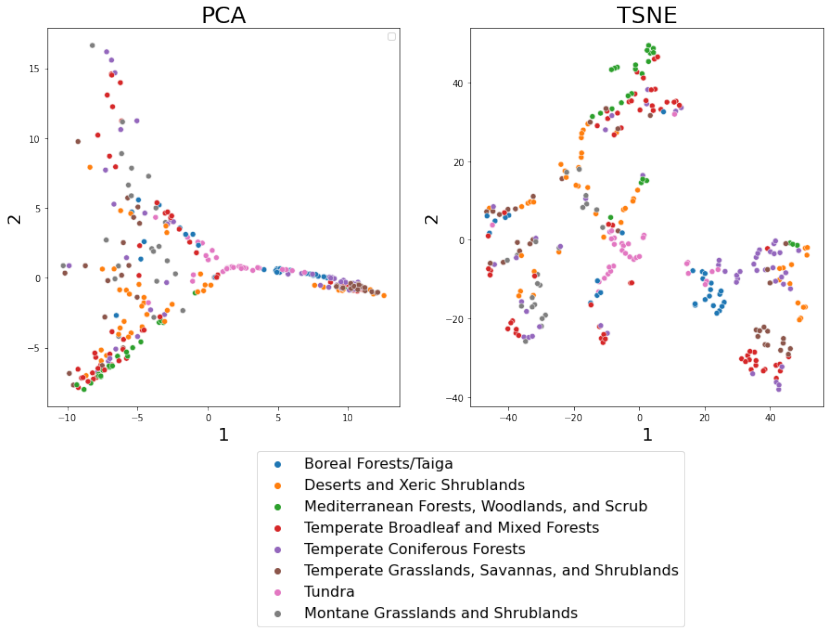


Fig. 4: 2D representation of the WWF dataset

2.2 The Paleolitic dataset

A dataset is constituted with the faunal species identified in all the archaeostratigraphic levels of the Arago cave. It gathers the species of large and small mammals, amphibians, reptiles and birds, determined from the fossil remains (bones, teeth) and corresponds to the taxonomic inventory of the vertebrate community present in each archaeological layer of the Arago Canyon. In total, the number of species, that represent the variables in the dataset, amounts to 144. The aim of this dataset is to identify the biomes and ecoregions represented in each archaeological layer in order to reconstruct the cave environment and to identify the type of landscape and climate that prevailed during each period of occupation of the site by human groups.

Once the two datasets are constructed, we restrict the WWF dataset [2]. Geographically, only the ecozones relevant to the Caune de l'Arago site are retained, namely the Palearctic and Nearctic ecozones. As the set of taxa found during the excavations only represents a small part of the taxa existing today, the WWF dataset had to be adapted. The starting points are as follows:

- Consideration of species presence/absence so that the discrepancy between current natural quantifications and those of the archaeological corpus (necessarily more limited) does not bias the results of the predictions.
- The choice not to use the criterion of taxon abundance also makes it possible to avoid the biases linked to differential archaeological conservation and to the selection of species by their predators (humans, carnivores, raptors).
- Consideration only of species that have been found at least once in at least one of the archaeological levels, in order not to take into account species that are too distant from the fossil assemblages.
- Replacement of some extinct species with the most ecologically similar current species. Species with no current equivalent (for instance, the grassland rhinoceros) are removed from the dataset.

The figure 5 shows the restricted dataset. It displays the distribution of the remaining 127 species according to the ecoregions to which they belong today. They are most numerous in two western European ecoregions (Northeastern Spain and Southern France Mediterranean forests, including the *Caune de l'Arago*, and Western European broadleaf forests) and their density decreases with distance.

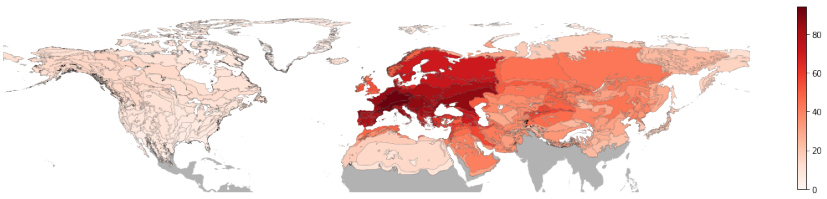


Fig. 5: Number of species found in Arago living in each ecoregion

2.3 Goal: Predict the Biome

These 127 taxa therefore allow us to train a model with the aim of predicting a biome from the list of species (presence/absence) present in the ecoregion. The main objective is to exploit this model by inferring from the archaeological data and predicting the climatic conditions that prevailed during the emplacement of the archaeological layers according to the species found. The multi-class model in Figure 6 identifies the probability of occurrence of the different biomes for each archaeological layer.

This approach takes on a new dimension through explainability tools, as the inference of the model alone is of little value to an expert. Accompanying a prediction with explanations allows it to be enriched and to provide new information. One example is the use of Shap [10] to provide explanations to accompany the inferences of a model as shown in Figure 7.

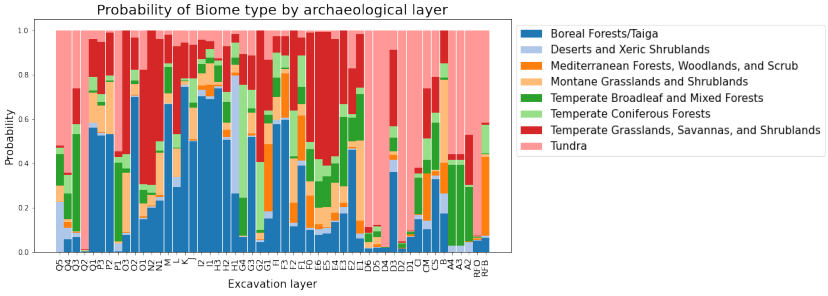


Fig. 6: Distribution of biome's in every archeostratigraphic layer

3 A data-oriented explainability tool: beta-Shapley

In order to assign a value to data points, A. Ghorbani et al. introduced the concept of data-Shapley [4]. This method is based on the concept of Shapley's game theory of values. Originally introduced by Lloyd Shapley, it proposes a fair method of payoff distribution. Thus, given a learning algorithm and a training dataset, data-Shapley is a metric that quantifies the value of each point in the training set relative to the performance of the predictor. This approach has many advantages, including that low value points capture outliers and corrupted points, high value points can inform us about what kind of new data could benefit our study [5].

The marginal contribution is defined as Δ_j as follows [9]:

Definition 1 (Marginal Contribution). *For a function h ; $j \in [\![1\ ;\ n]\!], n = |D|$ with D our dataset, we define the marginal contribution of a data point*

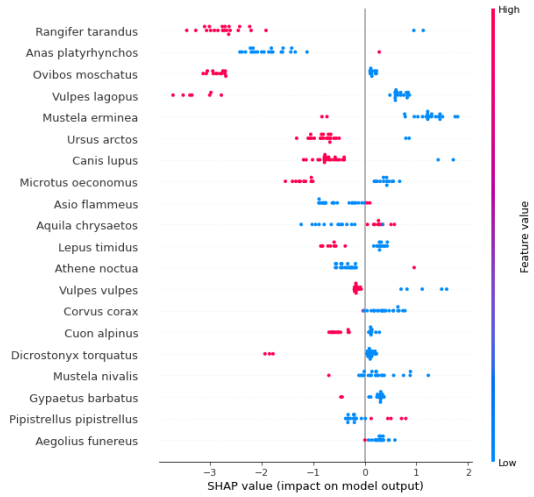


Fig. 7: Visualization of taxa's importance via Shapley values for biomes model's predictions

$z^* \in D$ with respect to $j-1$ points as:

$$\Delta_j(z^*; h; D) = \frac{1}{\binom{n-1}{j-1}} \sum_{S \in D_j^{\setminus\{z^*\}}} h(S \cup z^*) - h(S)$$

with $D_j^{\setminus z^*} = \{S \subseteq D \setminus \{z^*\} : |S| = j-1\}$

The calculation of the data-Shapley value for a data point is defined by [7]:

Definition 2 (Data Shapley). *The data-Shapley value of point $z^* \in D$*

$$\psi_{shap}(z^*; U; D) := \frac{1}{n} \sum_{j=1}^{n} \Delta_j(z^*; U; D)$$

with $|D| = n$; $U : \cup_{j=0}^{k} z^j \to \mathbb{R}$ a utility function representing the performance of a model trained on a dataset $\cup_{j=0}^{k} z^j$; $k \in \mathbb{N}$

Data-Shapley values uniquely satisfy the following properties [7]:

1. **Efficiency:** The sum of allocations equals the utility value of the entire dataset.
$$\forall U, \sum_{z \in D} \psi(z; U; D) = U(D) \tag{1}$$
2. **Symmetry:** $\forall U$ and any permutation π on D
$$\forall S \subseteq D, \psi(U(\pi(S))) = U(\pi(\psi U(S)))$$

3. **Null player:** A point z_i having no marginal contribution the allocation is 0.
$$U(S \cup \{z*\}) = U(S)\ \forall S \subseteq D \setminus z^*\ \psi(z*; U; D) = 0$$
4. **Linearity:** $\forall U_1, U_2$ utility functions, $\forall \alpha_1, \in \mathbb{R}$,

$$\begin{aligned}\psi(z^*; \alpha_1 U_1 + U_2; D) =& \alpha_1 \psi(z^*; U_1; D) + \\ & \psi(z^*; U_2; D)\end{aligned} \tag{2}$$

A generalization of this tool is presented through beta-Shapley defined in [9]. The so-called beta-Shapley values are derived from data-Shapley by relaxing the efficiency axiom (1) and adding two hyperparameters (α, β) deciding on added weights according to cardinality. A high value of (α) will put increased emphasis on sets of small cardinality and conversely (β) will put emphasis on sets of large cardinality. An illustration of the parameters is shown in Fig. 8.

Definition 3 (Beta Shapley). *The Beta Shapley value of point* $z^* \in D$

$$\psi_{beta}(z^*; U; D; w^{(n)}) := \frac{1}{n} \sum_{j=1}^{n} w^{(n)}(j) \Delta_j(z^*; U; D) \tag{3}$$

with $w^{(n)} : [n] \to \mathbf{R}$ *such as:*

$$n = \sum_{j=1}^{n} \binom{n-1}{j-1} w^{(n)}(j) \Delta_j(z^*; U; D)$$

Specifically we will use a weight scheme defined from the parameters $alpha, \beta$ defined as:

$$w^{(n)}_{\alpha,\beta}(j) = n \frac{Beta(j + \beta - 1, n - j + \alpha)}{Beta(\alpha, \beta)}$$

with: $Beta(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$; Γ the gamma function.

For our study we calculated the beta-Shapley values of each ecoregion. To do this, we calculated the average of the beta-Shapley values through a strategy of *stratified shuffle split*, which we can do thanks to the linearity property (2) of the beta-Shapley values. This split method allows a cross-validation by keeping the percentage of points of each class. This allows us to keep an unbiased performance evaluation while having a value for each point of our dataset.

Thus, for each split, the beta-Shapley values are approximated by a Monte-Carlo method whose convergence is supervised via the Gelman-Rubin statistic [3]. The model used is a gradient boosting model, lightGBM, and the utility metric is the multi-class accuracy on the test set corresponding to the training set provided by the *stratified shuffle split*.

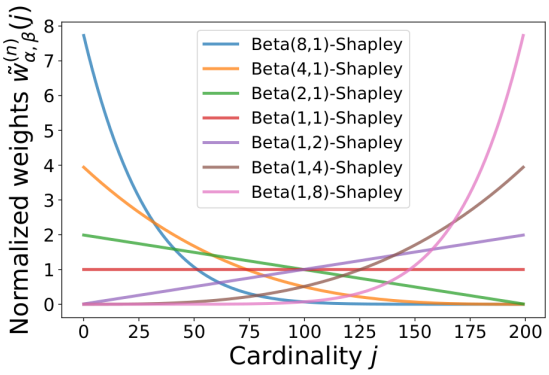


Fig. 8: Illustration of α , β roles

Figure 9 gives an overview of the results for different parameters α, β. The lower the value attributed to an ecoregion, the more red its colour will be, the higher its value, the more green it will be, the white colour corresponding to 0.

It is easy to see that a high β makes it possible to isolate precise points whereas a high α highlights groups of points. Indeed, the higher the α, the closer we get to a leave-one-out approach which corresponds to removing a point and measuring the difference in performance.

3.1 Relation between data quality and Shapley values

Our dataset is not perfect, as described above we have only kept some species, which results in some ecoregions appearing identical in their assemblage of animal species; this is problematic as they are however characterized by different biomes. These so-called ambiguous ecoregions are shown in Figure 10.

The beta-Shapley values of these ecoregions are low, as shown in Fig. 11. In fact, due to their defects, they do not allow a good training of the model, or even harm it, and therefore receive a low value as a result.

An alternative way of seeing how this tool works is an analysis of the correspondence between performance and Beta-Shapley value, as shown in Fig. 12.

Several heuristics for valuing data points are evaluated. Each heuristic proposes a ranking of ecoregions, from which points are removed one by one in order of importance. At each stage, the model is re-trained and its performance is evaluated. It can be seen that beta-Shapley is significantly more efficient than a random selection method. Indeed, heuristics based on beta-Shapley even allow a performance gain by removing harmful points.

3.2 Grouped tendencies inspection

Parameterized beta-Shapley such as $\alpha >> \beta$ for example $Beta(8:1)$ put large weights on small cardinalities and de-built large ones. As a result, the allocations

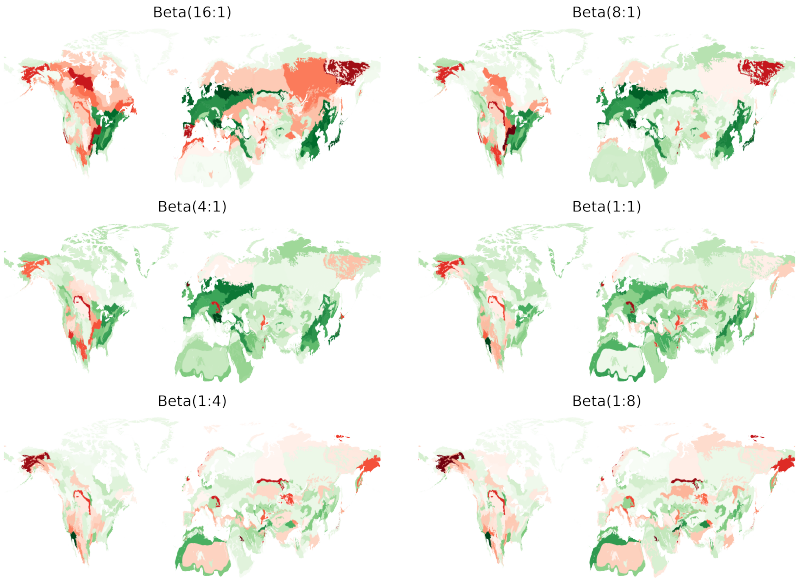


Fig. 9: α , β impact on data-Shapley values

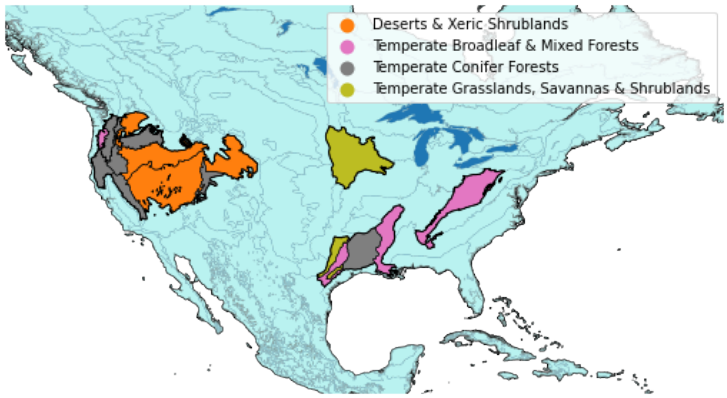


Fig. 10: Groups of ecoregions that are indistinguishable from each other based on the species present but labelled as different biomes

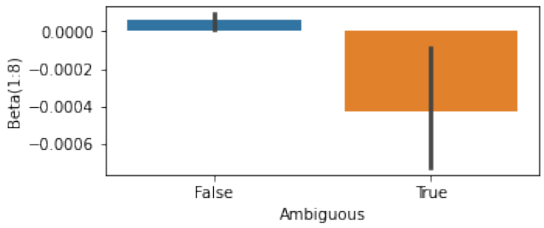


Fig. 11: Distribution of $Beta(1{:}8)$ for ambiguous ecoregions

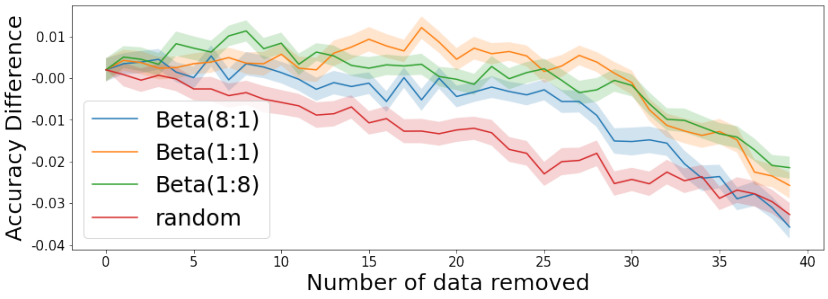


Fig. 12: Iterative suppression of training data points and impact on model's performances

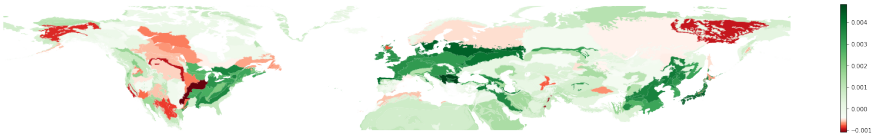


Fig. 13: Distribution of Beta(8:1) by ecoregion

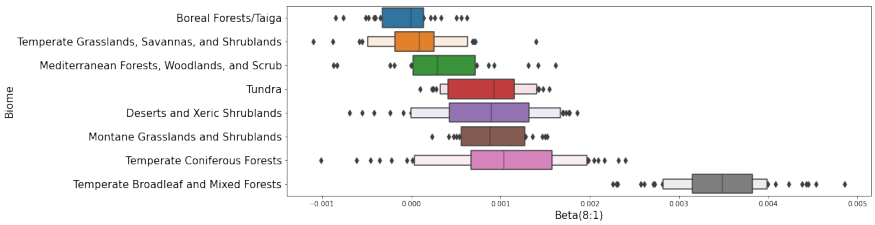


Fig. 14: Distribution of Beta(8:1) per Biome

are more homogeneous, and it is possible to distinguish the impact of groups of points on the performance of the model.

For example, if we group these beta-Shapley values by biome, an immediate distinction appears between the biome *Temperate Broadleaf and mixed Forests* and the others (Fig. 14). The ecoregions belonging to this biome contribute more to the performance of the model. This biome is both the most represented in the domain of our data (Fig 3) and also, the one with the greatest variety of species per ecoregion. It is possible that the multi-class accuracy, the utility function used, has increased these values in favour of the above biome. Indeed, this function is sensitive to the imbalance of the different classes, and does not measure well the performance of the less represented classes.

3.3 Inspection of influential species

Beta-Shapley gives us access to a ranking of ecoregions reflecting their marginal utility as a point in a training set. The rank of an ecoregion is thus considered to be the ranking of its value of *Beta(1:8)* ordered in ascending order. This rank makes it possible to highlight elements useful for training the model. For example, by looking at the ecoregions of the Biome and conditioning on the presence of *Buteo buteo* (Common Buzzard) a clear distinction appears 15. For a given ecoregion, the presence of this animal helps to determine whether the biome is Montane Grasslands or not. Ecoregions containing this animal were more informative for the determination of this biome and for our model than those that did not.

Fig. 15: Rank distribution of the *Buteo buteo* function of its presence in *Biome Montane Grasslands & Shrublands*

This information takes on a new dimension when put into context. Indeed, a visualization on a map, or in a geographic information system, gives an expert more elements to reach conclusions. For example, in this case *Buteo buteo* is not present in the Tibetan plateaus south of the Taklamakan desert, whereas the biome extends over a continuous area. In the context of our study, the ecoregion

of the western Karakoram in which the buzzard lives allows us to more easily identify the biome Montane Grasslands and Shrublands compared to ecoregions of the same biome in which this animal does not live.

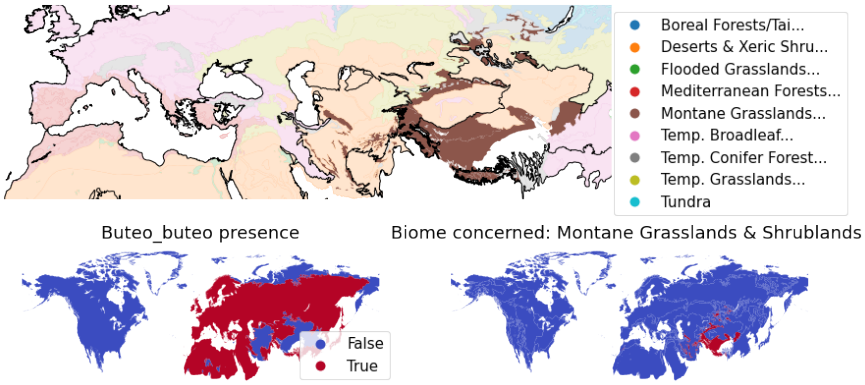


Fig. 16: Repartition map of the *Buteo buteo*

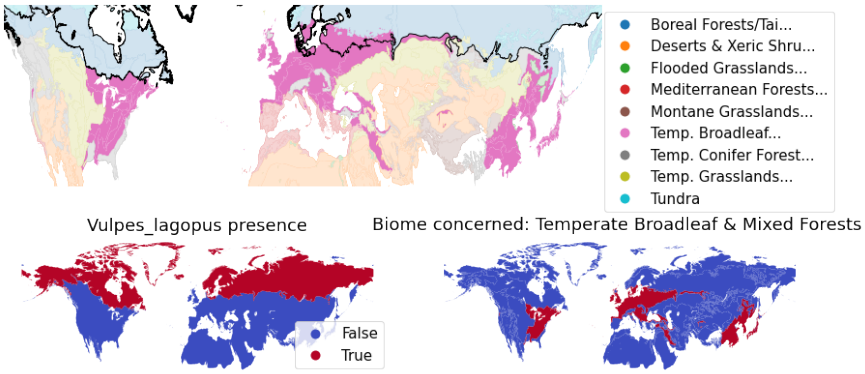


Fig. 17: Repartition map of the Vulpes lagopus

Similar information can be obtained from Vulpes lagopus (arctic fox) and the biome *Temperate broadleaf & mixed forests*. The ecoregions that seem relevant are: *Sarmatic mixed forests* et *Baltic mixed forests*.

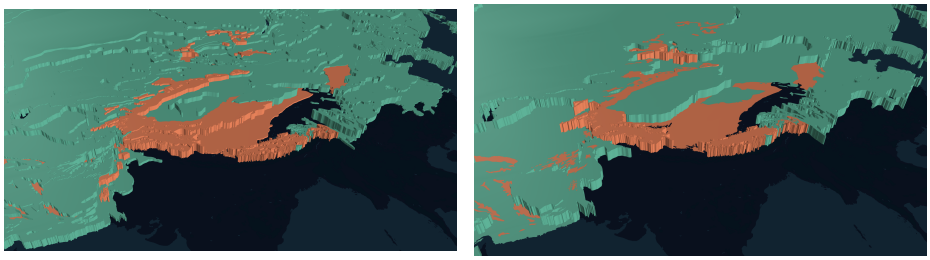
4 Geographic Information System integration

In order to make the explainability tools more usable, we have integrated the beta-Shapley values and other available attributes into the Kepler.gl tool [13]. Kepler.gl is an open source geospatial data visualization tool that provides an interactive, three-dimensional view of many layers. These two elements allow for contextualization of data as well as personalized exploration of the data.

The Beta(8:1) Shapley values are shown in figure 18. Each segmented area of uniform colour corresponds to an ecoregion, and the colour and height correspond to the beta Shapley value. The minimum corresponds to the darkest blue and zero height, the maximum to the most intense red. It is possible to highlight any ecoregion as a result of a click by the user, who can thus access the details of this data point. Beyond the multiple customisations possible (colouring, displayed variables, ...) the user is free to navigate the map and choose the point of view that suits him best in this three-dimensional space.

The figures 19a, 19b present the same information described in section 3.3. Thus, in orange, the biome of interest to us appears: *Montane Grasslands & Shrublands* for the determination of which the presence of the buzzard is informative. Thanks to these visualizations it is easy to determine that the ecoregion of *North Tibetan Plateau-Kunlun Mountains alpine desert* is the only one that changes significantly in height among the regions *Montane Grasslands & Shrublands*. The Buteo buteo appears in this ecoregion, but it has a low beta-Shapley value which contrasts with its neighbours. The high informatively of this ecoregion may be due to the presence or absence of another animal. With this visualization, a domain expert would have both a starting point for a critical look at the data and the new dimension of beta-Shapley values, that of the informativeness of a data point in relation to a machine learning model.

Fig. 18: Kepler.gl interface



(a) Presence - height representing the animal presence

(b) Rank - height representing the animal rank

Fig. 19: *Buteo buteo* information view on Kepler

5 Discussion and conclusion

The aim of this paper is to present the use and utility of a data-centric explainability tool in the concrete context of paleoenvironmental reconstruction. We have shown how the exploitation of geographical information contained in the data allows to better visualize the explanations of machine learning models. The elements highlighted are not conclusions, it is necessary that a palaeontologist, an expert in the field, makes this tool his own and contributes his knowledge. The reliability and usefulness of this tool, which has been described in the course of the experiments, need to be taken carefully. Indeed, the information delivered is complex, and the usefulness of a training data point for a machine learning model with respect to a score function is, to say the least nuanced. Despite the good properties of (section 3) Many different dimensions impact the beta-Shapley values. To draw valid conclusions, it is at least necessary to know both the characteristics of the type of learning model used and its limitations, and also to know the dataset on which it is trained.

One of the limitations of beta-Shapley is its algorithmic complexity. Indeed, the computation of exact values is exponentially more complex in terms of the number of points. Although some methods allow less expensive approximations, such as Monte Carlo methods [11], their use on large datasets becomes almost impossible. It is still possible to apply it to a sample of representative points, but the usefulness of this tool would be diminished. For example, for anomaly detection, which would be appropriate in the context of massive data, it would not be possible to use this process. However, it would be possible to use it for verification purposes.

Lastly, this work can be positioned in the field of trustworthy AI. Indeed, adapting machine learning based solution in other domains has posed many problems [8]. Even with proper reproducibility and good practices, it is still essential to improve modelling practices. Combining explanations to a classical predictive approach is one way to do so [7]. The added value of explanation has been highlighted throughout this paper and confirmed with domain experts. Adding explanations and combining them to model inferences provides more

information, context, insights, allowing more robust conclusions. Those qualities inherently foster trust in the AI usage.

References

1. Henry de Lumley. *La Caune de l'Arago Tome I, Tautavel-en-Roussillon, Pyrénées-Orientales, France.* Éd. du CNRS, Paris, 2014.
2. World Wildlife Fund. Wildfinder: online database of species distributions, 2006.
3. Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457 – 472, 1992.
4. Amirata Ghorbani and James Zou. Data Shapley: Equitable Valuation of Data for Machine Learning. *arXiv:1904.02868 [cs, stat]*, June 2019. arXiv: 1904.02868.
5. Amirata Ghorbani, James Zou, and Andre Esteva. Data Shapley Valuation for Efficient Batch Active Learning. *arXiv:2104.08312 [cs, stat]*, April 2021. arXiv: 2104.08312.
6. Sophie Grégoire, Nicolas Boulbes, Bernard Quinio, Matthieu Boussard, Caroline Chopinaud, Anne-Marie Moigne, Agnès Testu, Vincenzo Celiberti, Cédric Fontaneil, Christian Perrenoud, Anne-Sophie Lartigot Campin, Thibaud Saos, Tony Chevalier, Véronique Pois, Henry de Lumley, Marie-Antoinette de Lumley, Antoine Harfouche, Rolande Marciniack, Philippe Carrez, and Thierry Hervé. Innovative multidisciplinary method using Machine Learning to define human behaviors and environments during the Caune de l'Arago (Tautavel, France) Middle Pleistocene occupations. In Archaeopress, editor, *Big Data and Archaeology : Proceedings of the XVIII UISPP World Congress (4-9 June 2018, Paris, France), Sessions III-1 François Djindjian, (éd.) ; Paola Moscati, (éd.)*, Proceedings of the XVIII UISPP World Congress (4-9 June 2018, Paris, France), pages 28–47. 2021.
7. Duncan J. Watts Susan Athey Filiz Garip Thomas L. Griffiths Jon Kleinberg Helen Margetts et al. Hofman, Jake M. Integrating Explanation and Prediction in Computational Social Science. *Nature*, 2021.
8. Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in ml-based science, 2022.
9. Yongchan Kwon and James Zou. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. *CoRR*, abs/2110.14049, 2021.
10. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
11. Tomasz Pawel Michalak, Aadithya V. Karthik, Piotr L. Szczepanski, Balaraman Ravindran, and Nicholas R. Jennings. Efficient computation of the shapley value for game-theoretic network centrality. *CoRR*, abs/1402.0567, 2014.
12. Magdalena Sobol and S. Finkelstein. Predictive pollen-based biome modeling using machine learning. *PLOS ONE*, 13:e0202214, 08 2018.
13. Open source. Kepler.gl, a powerful open source geospatial analysis tool for large-scale data sets. Online. Accessed: 2022-04-12.
14. Christian Willmes, Kamil Niedziółka, Benjamin Serbe, Sonja Grimm, Daniel Groß, Andrea Miebach, Michael Maerker, Felix Henselowsky, Alexander Gamisch, Masoud Rostami, Ana Mateos, Jesús Rodríguez, Heiko Limberg, Isabell Schmidt, Martin Müller, Ericson Hölzchen, Michael Holthausen, Konstantin Klein, Christian Wegener, and Georg Bareth. State of the art in paleoenvironment mapping for

modeling applications in archeology-summary, conclusions, and future directions from the paleomaps workshop. *Quaternary*, 3:13, 05 2020.