BIODIVERSITAS
Volume 19, Number 5, September 2018
Pages: 1763-1770

ISSN: 1412-033X
E-ISSN: 2085-4722
DOI: 10.13057/biodiv/ d190525

Short Communication: Categorization models as a powerful tool in paleontological data analyses – the Phanerozoic bivalves

AHMED AWAD ABDELHADY[1,♥], MOHAMMED MASOUD ABDALLA[2]

[1]Department of Geology, Faculty of Science, Minia University. 61519 El-Minia, Egypt. ♥email: alhady2003@yahoo.com
[2]Department of Computer Science, Faculty of Science, Minia University. 61519 El-Minia, Egypt

Manuscript received: 16 November 2017. Revision accepted: 31 August 2018.

Abstract. *Abdelhady AA, Abdalla MM. 2018. Short Communication: Categorization models as a powerful tool in paleontological data analyses – the Phanerozoic bivalves. Biodiversitas 19: 1769-1776.* Predicting biotic responses to current and future global change can be acquired through understanding how biological and environmental traits shaped the past origination, dispersion and extinction patterns. A global dataset encompasses 161,357 taxon occurrences belonging to 2,378 bivalve genera from past and recent environments were analyzed based on the categorization model, a widely-used machine-learning analysis, using MS-SQL and Excel PowerView. The occurrence data was standardized using square-root transformation to downplay the effect of sampling effort. Thus, the examined traits are resulting from reliable ecological interactions. The results indicate that the biotic traits of the bivalve can be determined by the abiotic ones. Moreover, ecological traits such as life habit (i.e., infaunal vs. epifaunal), diet (suspension vs. deposit feeders, herbivores vs. carnivores), composition (aragonite vs. calcite), and locomotion (stationary vs. mobile) all exhibit significant relation to a specific environment. The results demonstrated that decision tree and association rules are primary powerful tools in analyzing huge biological data and in testing many useful bio-ecological hypotheses.

Keywords: Association rules, biotic traits. bivalvia, biodiversity, decision tree

INTRODUCTION

The efforts of the paleontologists in the last centuries have generating ultra-scale data sets of very high spatial resolution. These data were stored in many databases such as the Paleobiology Database (PBDB; http://paleobiodb.org/#/). However, the web-based system of the PBDB has not yet sufficient analytical functions for spatiotemporal analysis (http://paleobiodb.org/#/). Ecologists and climatologists are focusing their research now on understanding climate changes over broader range of time and space scales via the paleontological data from the fossil record (Alroy 2008; Nürnberg and Aberhan 2013; Foote 2014; Abdelhady and Fürsich 2015). According to Groth et al. (2012), insightful analysis of the life evolution on the earth is depending on applying different software tools to explore, manipulate, and visualize huge data sets. In addition, closer integration of geographic visualization and/or geo-computation is essential to address many environmental concerns (Varela et al. 2009; 2015). However, exploration tasks are usually complex and consume much time (Groth et al. 2012; Varela et al. 2015).

Advances in the field of information visualization offer a number of innovative and promising approaches (Kehrer et al. 2010). The applications of the visualization techniques has grown rapidly for different geologic purposes (Best and Lewis 2010; Gorricha and Lobo 2012; Romañach et al. 2012; Du et al. 2015). Effective computer tools, incorporate the Microsoft office package can provide intelligence from raw paleontological data by creating visual graphs, thus new concepts will be constructed to understand the paleo-ecosystems easily and efficiently. The set of utilities in MS-SQL and Excel PowerView enable preprocessing and visualization of large data in addition to plenty of statistical and numerical analysis. In addition, although categorization models are fundamental in decision-making and all kinds of environmental interaction, they are rarely used by paleontologists (e.g., Boyer 2010; Finnegan et al. 2012). In this paper, we implement the stratigraphic and geographic occurrence data of the bivalves, which is one of the best-preserved and documented fossil group through the Phanerozoic to illustrate how interactive exploration and visualization tools such as MS-PowerView and MS-SQL can successfully analyze the paleontological data. Therefore, we proposed to analyze the bivalve occurrence data and to test the ability of the categorization models (decision tree and association rules) to generate meaningful ecological results. Moreover, we aim to determine the main environmental factors that controlling the bivalve diversification (temporal) and distribution (spatial).

DATA AND METHODS

Dataset

The bivalve occurrence data is retrieved essentially from the PBDB. The PBDB was organized and operated by a multi-disciplinary international group of paleontological

researchers (Alroy et al. 2001). It provides a global, collection-based occurrence and taxonomic data for organisms of all ages. The PBDB encourages and enables addressing large-scale paleobiological and biological questions (Varela et al. 2015). Occurrence and range data of bivalve through the Phanerozoic as a whole were downloaded from the PANGAEA (for details see https://issues.pangaea.de/secure/attachment/97680/Appendix%20C.pdf.). The occurrence matrix includes 161,357 records belonging to 2,378 genera. The data were compiled into two tables, the range table, includes genera and their First Appearance Datum (FAD) and Last Appearance Datum (LAD) in addition to mean abundance longitude. While occurrence table include taxonomy (order, family, genus), life habit (infaunal, endobysate, epifaunal, boring, etc.), diet (suspension feeders, deposit feeders, and carnivores, etc.), locomotion (mobile, sessile, etc.), and shell composition (aragonite and calcite), in addition to age data (epoch, 2 MY bin, and 10 MY bin). Spatial data include country, paleo-latitude, and plaeo-longitude (Abdelhady 2015).

Data preparation and standardization

We introduce herein an easy, office-based approach that integrates exploration and visualization tools to different biologic and ecologic data sets. The implemented approach include data extraction, standardization, classification, and visualization of the results (Figure 1). Although the PBDB store a high number of paleontological data, from which many questions regarding the history of earth can be answered, uncertainties about these data clogged such interpretations. The results may be a sampling artifact. Increasing biodiversity from the Cambrian (540 MY) may represent the sampling efforts applied to the younger strata (the pull of the recent, Alroy et al. 2001). Therefore and to reduce potential errors, pre-processing of fossil data is often required before advanced analyses.

There are many standardization methods such as subsampling and residual methods. Although subsampling is one of the most frequently used for data standardization, its implementation has resulted always in reduction of the examined data, which is unfavorable for statistical analysis. Subsampling and residual methods were used to test reproducibility and consistency of the results of diversity calculations. In addition and according to Tomašových and Kidwell (2009), square-root transformation of the abundance data, downplay the impact of numerically abundant species and increase the effect of rare species (Abdelhady and Fürsich 2015).

For measuring sampling effort and for evaluating the latitudinal diversity gradient of the bivalve throughout the Phanerozoic, the occurrence date was square-root transformed for estimating a reliable diversity, origination, and extinction rates. Measuring the geographic dispersion was done by summing the number of genera in equal grids. Each grid include ten latitudinal degrees. The numbers then were normalized by the maximum number of occurrences (quantifying sampling effort). A computer code was developed in Mathematica Wolfram 10 by which standardizations were carried out.

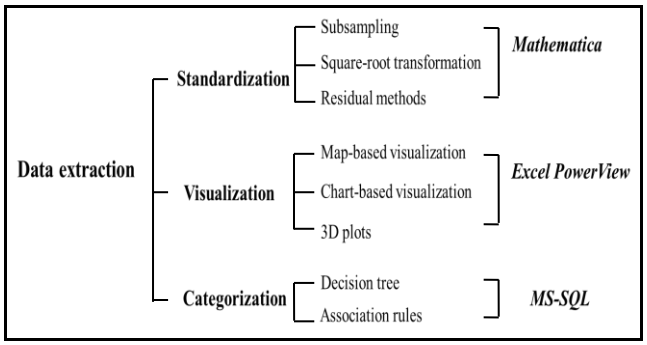

Figure 1. Schematic representation of the procedures applied in this study

Decision tree and association rules

The process in which ideas and objects are recognized, differentiated, and understood is known as 'Categorization' it implies that objects are grouped into categories for specific purpose, and thus, it illuminates a relationship between the subjects and objects of knowledge (Cohen and Lefebvre 2005). We used two different models herein, the association rules and decision tree. Decision tree is data analysis in the shape of extracting a model describing a considerable data classes (Han et al. 2011). For example, we can build a classification model, which predicts whether a taxon will be found in a given environment, or predicts which of five categories a new database item belongs to. Such analysis can assist in providing us with a better recognition of the data at large scale. Many classification methods have been developed regarding pattern recognition, machine learning, and statistical data classification, which is a two-step methodology consisting of a learning step (i.e., a classification model is built) and a classification step (i.e., the model is used to infer class labels for given data). The decision tree is a very popular classification technique characterized by quick training performance (for details see Han et al. 2011). A decision tree is a mathematical model, which help managers make decisions and it is rely on estimates and probabilities to calculate likely outcomes. Thus, it enables individual or organization to tack a decision based on costs and benefits (Quinlan 1987). Therefore, they can be used either to drive informal discussion or to map out an algorithm that predicts the best choice mathematically. For example, suppose that we have a set of N items, which fall into two categories, n have to Label 1 and m = N-n have Label 2. To get our data a bit more ordered, they well grouped by labels and two ratios will be calculated to estimate the Entropy (E):

$$ratio\ p = \frac{n}{N} and\ ratio\ q\frac{m}{N}.\ Entropy\ E = -plog2(p) - qlog2(q)$$

The decision tree graph can be read from left to right as follows: The rectangles, which are referred to as nodes, hold subsets of the data. The title on the node announces the defining characteristics of that subset; the leftmost node, titled 'All', depict the complete data set. All Following nodes represent subsets of the data. A decision tree contains many splits where the data diverges into

multiple sets depending on attributes. As for instance, the first split in the sample model divides the dataset into nine groups by 'Taxon Environment' The split immediately after the 'All' node is most important because it shows the primary condition that divides this dataset. Extra splits occur to the right, thus by analyzing different segments of the tree, we can learn which attributes have the most influence factor.

The Association rules are statements that help uncover relationships between seemingly unrelated data in a relational database or other information repositories. The association Rules find all sets of items (itemsets) that have supported greater than the minimum support and then using the large itemsets to generate the desired rules that have confidence greater than the minimum confidence. The lift of a rule is the ratio of the observed support to that expected if X and Y were independent:

$$Rule\ X \rightarrow Y:\ \text{Support} = \frac{frq(X,Y)}{N}, Confidence = \frac{frq(X,Y)}{frq(X)}, Lift = \frac{Support}{Supp(X) * Supp(Y)}$$

The data set was analyzed by two related software packages, MS-SQL and MS-PowerView. Microsoft SQL Server Data Mining is a collection of machine learning algorithms that explore your data for patterns. Once discovered, these patterns can be browsed for greater insight into your data, or they can be applied to new data to create "predictions" - which allow you to determine unknown facts about data based on data the algorithms have seen before. The MS-SQL analysis service is used for applying the classification and association rules techniques to the data. PowerView and PowerPivot, a feature of Microsoft Excel 2013, are used to perform powerful data analysis and create sophisticated data models. With PowerPivot, one can mash up large volumes of data from various sources, create a Data Model, a collection of tables with relationships, perform information analysis rapidly, and share insights easily. Where the PowerView plug-in within Excel 2013 is used to visualize the geographical information to enable the user to understand and interact with the presented knowledge easily and smoothly. PowerView is an interactive data exploration, visualization, and presentation experience that encourages intuitive ad-hoc reporting. Power Pivot and power view have been described as the most important new feature in Excel in 20 years (Winston 2014).

RESULT AND DISCUSSION

Bivalve occurrence

The final dataset encompasses 161,357 taxon occurrences belonging to 2,378 bivalve genera and 33 orders (Table 1). These taxa have variable ecological traits regarding shell composition, life-habit, diet, and mobility level. The distribution pattern of the bivalves may thus be linked to specific biotic or abiotic factor (see below). The occurrence of the bivalve orders throughout the geologic time scale is given in Table 1. In general, there is a steady increase in the bivalve occurrence from the Cambrian onward (Table 1).

Map-based visualization

Visual representations of the data allow an easy way of constructing knowledge (Varlea et al. 2015). Visualization of taxon distribution during specific ages using PowerView is shown in Figure 2. The PBDB provides a similar visualization technique (http://paleobiodb.org/#/). However, it lacks important filtering tools. Here we have included important auto-ecological filters, namely composition, diet, life habit, and locomotion. The advantage of such filter is to evaluate the influence of ecological parameters on the spatial distribution of the biota. Note that according to Fang et al. (2014) and Abdelhady and Fürsich (2014, 2015) and Abdelhady and Mohamed (2017), ecological aspects such as life-habit influence the geographic dispersion of the invertebrates. In addition, life-habit (e.g., benthic vs. planktic) of the invertebrates is influencing their temporal durations (Abdelhady et al. 2018). Pelagic fauna with planktonic larvae can be drift for longer distances than do benthic and non-planktonic larval taxa. The latter indicates how the filters-based visualization technique are useful in examining and answering very important ecological questions. In addition, the Map-based visualization techniques allow the mapped data to be changed interactively (Dykes 1997). Thus, it permits the users to change the appearance of the objects mapped and consequently define clusters. The occurrences map of the bivalve according to their life-habit shows that the infaunal and epifaunal taxa are the most dominated groups, (Figure 2), while other life habits are less abundant. Moreover, the distribution map indicate a strong latitudinal diversity trend. In general, the map visualization (Figure 2) is neat and the base map is certainly an aesthetic improvement over the base map provided by the PBDB. However, until now, there is no possibility to change the modern world map with paleomaps, which are more informative in paleontological data analyses.

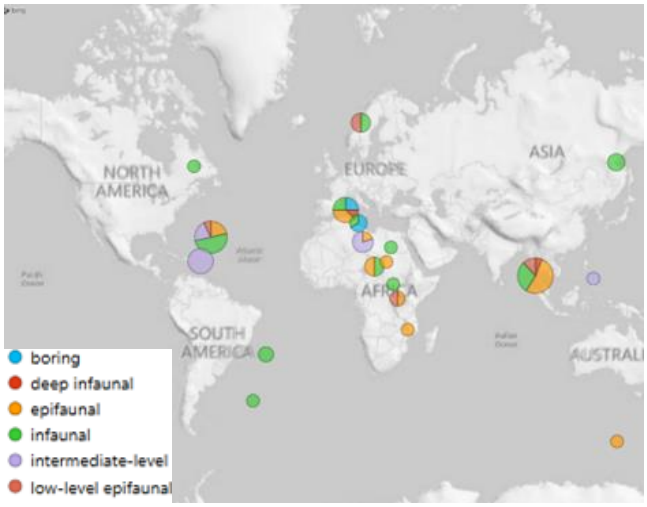

Figure 2. Distribution of the bivalves according to their life-habit

Table 1. Distribution of the bivalve orders throughout the Phanerozoic (Data compiled from the PBDB)

Order	Geologic time										
	Cambrian	Ordovician	Silurian	Devonian	Carboniferous	Permian	Triassic	Jurassic	Cretaceous	Cenozoic	Total
Actinodontida		18		25		1			2		60
Anomalodesmacea										4	4
Arcida		27	6	65		234	168	1270	2353	6530	11295
Cardiida		1	2	49	61	439	587	2143	5606	21263	32208
Carditida		8	27	364	61	285	289	1460	1543	4803	9623
Colpomyida		20			98						22
Cyrtodontida		238	462	168		11					1023
Fordillida					4						11
Hiatellida	11						3	15	358	849	1369
Hippuritida			3				14	72	3685		4520
Lucinida			14	252		18	164	717	782	3451	5699
Megalodontida				5	2	2	215	247			517
Modiomorphida		511	203	532		24	2	4			1441
Myalinida		401	124	248	18	690	444	363	3208	1	6045
Myoida					235			80	44	89	225
Mytilida		5	2	25		140	268	1617	1133	1811	5450
Nuculanida		160	106	1055	46	452	316	546	940	2664	7083
Nuculida		224	45	320	187	235	150	553	670	1426	4146
Nuculiformes					114						61
Ostreida	60	197	288	965		597	2644	4935	5005	3931	19932
Pandorida					233		13	91	186	191	509
Pectinida			15	260		2906	3917	7134	5384	7802	29577
Pholadida					581	302	220	1402	1276	3482	7242
Pholadomyida		60	7	69	4	411	82	1258	412	115	2734
Poromyida					124		3	24	443	349	859
Protobranchia		1									1
Pterioida		19	11	29		134	66	122	71	24	518
Solemyida		192	73	74	6	72	11	37	57	73	725
Solenida		2	12	48	14	9	1	3	176	603	902
Thraciida							11	221	92	447	829
Trigoniida			5	108		557	1371	1108	1209	30	4819
Tuarangiida					93						3
Unionida	3	2				4	198	23	51	90	414
Total	74	2251	1486	5058	1881	7864	11187	25446	34691	60113	161357

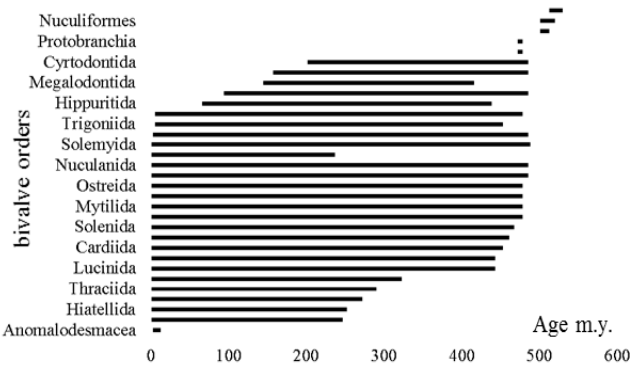

Figure 3. Range chart (temporal range of the taxa: species, genera, families, etc.) of the bivalve orders through the Phanerozoic using the stacked bar charts. Data figured include LAD and the duration of the taxa. The timescale is by millions of years before

Chart-based visualization

There are two forms of data visual representations in PowerView; maps and charts. Maps are frequently used for climate studies. Meanwhile, charts are needed where no country data are examined (i.e. latitudes instead of country). Histograms, bar chart, scatter, and bei-graph are the most used charts. Therefore, both are integrated into the analysis of the bivalve data to explore additional options. For example, fossils can be used for relative dating of rocks. To date rocks, range chart has to build. Almost all of the paleontological articles contain such chart. Using the stacked bar chart, temporal range of the taxa (species, genera, families, etc.) can be easily constructed. The required data are the LADs and the duration of the taxa (Figure 3).

One of the most important charts is that representing diversity across time to show the diversity dynamics (Figure 4) or across latitudes to show the pattern of the

latitudinal diversity gradient (Figure 5). Diversity (richness) can be determined based on different taxonomic ranks (species, genera, families, etc.). Moreover, the user can construct different filtering elements using the PowerPivot chart such as life habit, diet and locomotion exactly as done in map-based visualization. The five big mass-extinctions are clearly shown in Figure 4.

Again, implementing filters enable answering much debated ecological questions. According to Ros et al. (2011), the taxa originated under stress such as global crises event (the end Permian mass extinction) have longer ranges. To test this hypothesis the range of the bivalve taxa at the two biggest earth crises, namely the end of Permian and the end of Cretaceous were analyzed. The ranges of the taxa originated at the beginning of the crises (i.e, Induan ≈ 252.17 Ma, and Danian≈66 Ma) compared to those originated at normal stable conditions (i.e. the Carnian ≈227 Ma and the Maastrichtian≈ 72.1 Ma) using the range chart method described above. Just in few seconds, one can answer the question from a lock to the chart, which agrees to the hypothesis of Ros et al. (2011) in case of Danian and disagree in case of the Induan. Note that taxa originated under stress (Induan) have shorte age ranges (contrasting Ros et. 2011), while in Danian have longer age ranges agree with (Ros et al. 2011).

Similarly, FADs and LADs can be used to show number of originations and extinctions (Figure 5A). Group function of the pivot-chart was used to divide the Phanerozoic into 18-equal intervals (each = 30 MY). In addition, the temporal change of the epifaunal/infaunal proportional was constructed using the 100% stacked chart (Figure 5B). The information obtained from this chart could be used to analyze the diversity patterns at specific age or to follow replacement among taxa and their controlling factors (autoecological or eco-environmental).

The latitudinal distribution of the Phanerozoic bivalves shows greatest concentrations of occurrences between 10 and 50 N (Fig. 6). The steepness of the latitudinal pattern and the high similarity/correlation with the total number of collection may suggest a possible sampling artifact. The square-root transformation downplays the impact of sampling effort (Fig. 6B). Note that areas located in Europe and North America have extensive sampling efforts comparatively to those located in Africa or South America; Figure 6C).

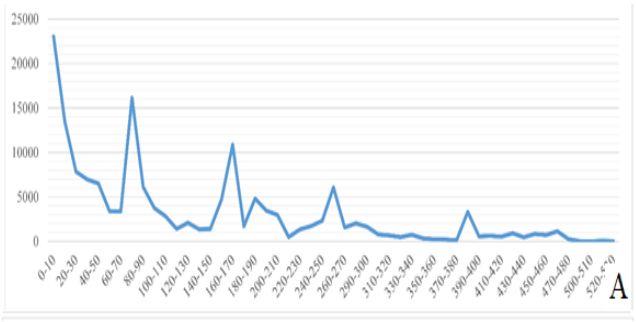

Figure 4. Diversity dynamics (represented number of genera) of the bivalve through the Phanerozoic. The beaks on the curve represent the 5-major mass extinction events

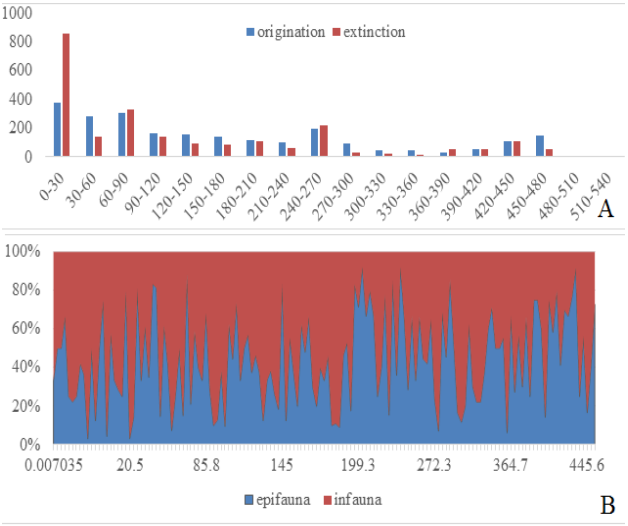

Figure 5. Origination and extinction of the bivalves (A) and the percentage of infaunal vs. epifaunal bivalves through the whole Phanerozoic (B). X-axis represents age in m.y. before present

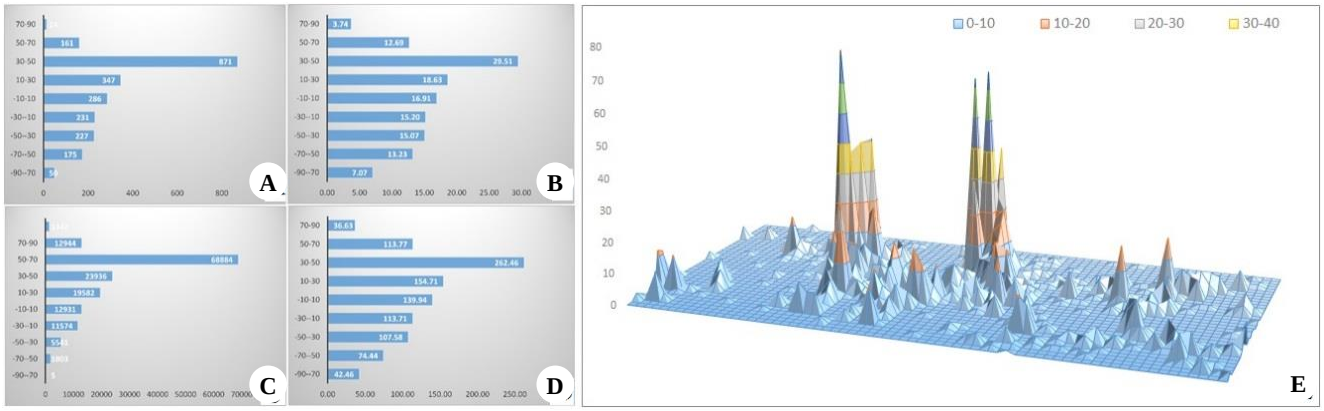

Figure 6. Bar chart shows the Latitudinal (y-axis) Diversity (x-axis, number of genera) Gradient: A. Raw data, B. Square-root data, C. 'Sampling effort' (number of collections) based on the raw data, D. 'Sampling effort' after square-root transformation of the raw data, E. 3D surface plot of the Latitudinal Diversity Gradient

Data model

The Data Model enables combining data that comes from SQL Server (Winston 2014). DISTINCT COUNT, a valuable function is activated when the excel table transformed to a data model. The function enable estimation of number of distinct genera from multiple occurrence record by counting the genus name only once and ignore duplications, hence the real number of the different taxonomic levels (i.e., genus, family, or order) can be estimated together with the normal counts of the records, which represent the density of a taxon (frequency in the rocks). The density of the collections can also be artifact represent the sampling efforts. In addition, relation among variable can be directly estimated quantitatively (Figure 7). The figure correlates between life mode and diet with other biologic or ecologic variables. The score in the figure refers to the importance/strength of each variable. From Figure 7, we can estimate that there is a strong relationship between the bivalve orders and the diet (i.e., for each bivalve order there is a specific diet mode for all genera included within this order (r = 0.76). in addition both environment (abiotic) and life-habit (biotic) can also determine the taxon diet (r = 0.26). In addition, for each bivalve order, there is a given life-habit (r = 0.6). Similarly, environment and shell composition have a considerable relation with the life-habit (r = 0. 34 and 0.22, respectively)

Decision tree

Table 2 summarize two cases of different rules. For each rule, the classification model determines number of cases and the probability of occurrence. Table 2 representing the main findings of the decision tree model elaborating the selected node (i.e. rule).

Association rules

From the itemsets herein, we can conclude that there is a complex pattern and association between biotic factors (such as life-habit, diet, locomotion) and the biotic ones (environment, age, and geographic occurrence (paleo-latitudes). The Rules tab in Figure 7 combines information about the itemsets and their relative value. Probability represents the portion of cases in the dataset that contain the targeted collection of items. Probability gives a hint of how likely the result of a rule is to occur. We can change the value of minimum probability in this pane to filter the rules that are exhibited. The value for minimum probability that we initially see is the threshold value that was used by the association rule algorithm when building the model. After the model is completed, we cannot reduce this value, but we can increase it to show only the higher probability items. Importance column is designed to measure the utility of a rule. A rule that is very common might have little information value. The greater the significance, the more valuable the rule is for predicting the outcome. Herein, we can summarize the following; carnivore bivalve are usually actively mobile taxa. In contrast, herbivore bivalves are usually passively mobile taxa.

In addition, deep infaunal bivalve are chemosymbiotic taxa dominating the reef environments characterized by marl deposition.

Table 2. Decision tree rules and their cases and probabilities

Rule	Diet value	No. of cases	Probability
Taxon Environment = ‘marine' and	Chemo-symbiotic	32	77.07%
Locomotion = 'actively mobile'	Deposit-feeder	9	21.83%
Taxon Environment = ‘inner shelf’ or Taxon Environment = ’outer’ or Taxon Environment = ’shelf’ or Taxon Environment = ’oceanic' and	Chemo-symbiotic	32	77.07%
LAD > = 103.200	Deposit-feeder	9	21.83%

Probability	Importance	Rule
1.000		Locomotion = passively mobile, Life-habit= boring --> Diet = herbivore
1.000		Locomotion = passively mobile, LAD < 64.4 --> Diet = herbivore
0.848		Locomotion = passively mobile,--> Diet = herbivore
0.979		Locomotion = passively mobile, Paleolat. = -7.3- 22..7 --> Diet = herbivore
0.879		Locomotion = passively mobile, Environment = Marine indet. --> Diet = herbivore
0.569		Life-habit = boring, Paleolat. = -7.3- 22..7 --> Diet = herbivore
0.406		Locomotion = actively mobile, Paleolat. >= 41.2 --> Diet = carnivore
0.434		Life-habit = deep infaunal, LAD < 64.4 --> Diet = chemosymbiotic
0.505		Environment = reef, buildup or bioherm, Life-habit = deep infaunal --> Diet = chemosymbiotic
0.417		Life-habit = deep infaunal, Lithology = marl, --> Diet = chemosymbiotic
0.812		Taxon environment = coastal, Life-habit = infaunal --> Diet = deposit feeder
0.763		Taxon environment = coastal, Locomotion = facultatively mobile --> Diet = deposit feeder
0.672		Taxon environment = coastal, --> Diet = deposit feeder
0.990		LAD = 279.91- 39.96, Taxon environment = coastal --> Diet = deposit feeder

Figure 7. The results of the association rules applied to portion of the raw dataset encompassing environment, locomotion, and life habit to predict the taxon diet. The figure shows the item, its probability, and its importance

Finally, we can clonclude that the application of interactive visual methods to analyze paleontological data is still hampered for paleontologists and paleo-climatologists, who are usually non-visualization experts (Groth et al. 2012). The goal here was not to describe in details a new software or the Phanerozoic history of bivalves but to illustrate how to implement the MS-SQL and PowerView software to analyze and visualize the paleontological data quickly and efficiently. Herein, we focused on performing some standard paleobiological analysis. In addition to a less-common machine-learning analysis. In fact, many of the basic analyses such as tabulating diversity, calculating extinction rates can be done with the PBDB using FossilWorks website (http://fossilworks.org). However, one has no chance for filtering such analyses based on bio-ecological traits. Furthermore, there are some specialized statistical analysis scripts such as PAST (Hammer et al. 2001) to achieve particular research purposes. Although PAST is one of the easiest statistics packages for paleontologists and biologists, it lakes exploratory data analysis or machine-learning algorithms (i.e. association rules or decision-making).

Integrating PowerView with MS-SQL here in has enabled the following: (i) Visualizing spatial and temporal data in a similar manner to that of PaleoDB package (Varela et al. 2015), (ii) Transform records into meaningful tables and/or matrices, and (iii) Search and find single piece of information, (iv) analyzing portion of the raw data based on multiple specific criteria as example it determine/draw the geographic range size of taxa that (a) are suspension-feeder, (b) belonging to a specific taxonomic rank such as family, and (c) that became extinct at the end of Cretaceous crises (66 Ma).

Some limitations in PowerView are present and should be highlighted herein. For example, the country map is limited to the modern continental configuration, which makes this visualization tool of limited utility to paleobiologists. Paleobiologists are more interested in where fossils were located in the geologic past than where they are found today. However, the approach implemented herein permits easily map/charts visualization in addition to many community analyses necessary for paleo-/ecological interpretation and ecosystem reconstructions. The results indicated that environmental type are the main factor controlling bivalve taxa (their diet, life-habit, etc.). In addition, there is a cyclic pattern among these attributes (i.e. they affect each other).

We concluded that decision tree and association rules may provide a valuable and advanced information for finding relationships among biological/ecological traits and their environmental parameters.

ACKNOWLEDGEMENTS

For fruitful discussions regarding mathematical and statistical issues, we would like to thank Dr. Abdelhafez Ahmed (Department OF Mathematics, Minia University, Egypt). This research was supported by the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH/Centre for International Migration and Development (CIM), Germany (Grant Number 41704). We thank all of the contributors to the Paleobiology Database. This is Paleobiology Database publication 320.

REFERENCES

Abdelhady AA, Fürsich FT. 2014. Macroinvertebrate palaeo-communities from the Jurassic succession of Gebel Maghara (Sinai, Egypt). J Afr Earth Sci 97: 173-193.

Abdelhady AA, Fürsich FT. 2015. Palaeobiogeography of the Bajocian-Oxfordian macrofauna of Gebel Maghara (North Sinai Egypt): Implications for eustacy and basin topography. Palaeogeogr Palaeoclimatol Palaeoecol 417: 261-273.

Abdelhady AA, Mohamed RSA. 2017. Paucispecific macroinvertebrate communities in the Upper Cretaceous of El Hassana Dome (Abu Roash, Egypt): Environmental controls vs adaptive strategies. Cretaceous Res 74: 120-136.

Abdelhady AA, Seuss B, El-Dawy MH, Obaidalla NH, Mahfouz AK, Abdel Wahed SA. 2018. The Unitary Association method in biochronology and its potential stratigraphic resolving power: A case study from Paleocene-Eocene strata of southern Egypt. Geobios. DOI: 10.1016/j.geobios.2018.06.005

Abdelhady AA. 2015. Occurrence and range data of bivalve through the Phanerozoic, with links to Excel files. Pangaea. DOI: 10.1594/PANGAEA.854072

Alroy J, Marshall CR, Bambach RK, Bezusko K, Foote M, Fürsich FT, Hansen TA, Holland SM, Ivany LC, Jablonski D, Jacobs DK, Jones DC, Kosnik MA, Lidgard S, Low S, Miller AI, Novack-Gottshall PM, Olszewski TD, Patzkowsky ME, Raup D, Roy M, Sepkoski KJ, Sommers MG, Wagner PJ, Webber A. 2001. Effects of sampling standardization on estimates of Phanerozoic marine diversification. Proc Natl Acad Sci USA 98: 6261-6266.

Alroy J. 2008. Dynamics of origination and extinction in the marine fossil record. Proc Natl Acad Sci USA 105: 11536-11542.

Best DM, Lewis RR. 2010. GWVis: A tool for comparative ground-water data visualization. Comput Geosci 36: 1436-1442.

Boyer A. G. 2010. Consistent Ecological Selectivity through Time in Pacific Island Avian Extinctions. Conserv Biol 24:511-519.

Cohen H, Lefebvre C. 2005. Handbook of Categorization in Cognitive Science, Elsevier.

Du Z, Fang L, Bai Y, Zhang F, Liu R. 2015. Spatio-temporal visualization of air-sea CO_2 flux and carbon budget using volume rendering. Comput Geosci 77: 77-86.

Dykes JA. 1997. Exploring spatial data representation with dynamic graphics. Comput Geosci 23 (4); 345-370.

Finnegan S, Heim NA, Peters SE, Fischer WW. 2012. Climate change and the selective signature of the Late Ordovician mass extinction. Proc Natl Acad Sci 109: 6829-6834.

Foote M. 2014. Environmental controls on geographic range size in marine animal genera. Paleobiology 40 (3):440-458.

Gorricha J, Lobo V. 2012. Improvements on the visualization of clusters in geo-referenced data using Self-Organizing Maps. Comput Geosci 43: 177-186.

Groth P, Frew J, Santos E, Koop D, Maxwell T, Doutriaux C, Ellqvist T, Potter G, Freire J, Williams D, Silva C. 2012. Designing a Provenance-Based Climate Data Analysis Application. In Provenance and Annotation of Data and Processes (eds). Springer, Berlin.

Hammer Ø, Harper DAT, Ryan PD. 2001. PAST: Paleontological statistics software package for education and data analysis. Palaeontologia Electronica 4: 1-9.

Han J, Kamber M, Pei J. 2011. Data Mining: Concepts and Techniques. Elsevier, Berlin.

Kehrer J, Filzmoser P, Hauser H. 2010. Brushing moments in interactive visual analysis. Computer Graphics Forum 29 (3): 813-822.

Nürnberg S. Aberhan M. 2013. Habitat breadth and geographic range predict diversity dynamics in marine Mesozoic bivalves. Paleobiology 39 (3): 360-372.

Quinlan JR. 1987. Simplifying decision trees. Intl J Machine Stud 27 (3): 221-234.

Romaňach SS, McKelvy M, Suir K, Conzelmann C. 2012. EverVIEW: A visualization platform for hydrologic and Earth science gridded data. Comput Geosci 76: 88-95.

Ros S, De Renzi M, Damborenea SE, Márquez-Aliaga A. 2011. Coping between crises: Early Triassic-early Jurassic bivalve diversity dynamics. Palaeogeogr Palaeoclimatol Palaeoecol 311: 184-199.

Tomašových A, Kidwell SM. 2009. Preservation of spatial and environmental gradients by death assemblages. Paleobiology 35: 122-148.

Varela S, González-Hernández J, Sgarbi LF, Marshall C, Uhen MD, Peters S, McClennen M. 2015. paleobioDB: an R package for downloading visualizing and processing data from the Paleobiology Database. Ecography 38: 419-425.

Varela S, Rodríguez J, Lobo J, M. 2009. Is current climatic equilibrium a guarantee for the transferability of distribution model predictions? A case study of the spotted hyena. J Biogeogr 36: 1645-1655.

Winston WL. 2014. Microsoft Excel 2013: Data Analysis and Business Modeling. O'Reilly Media Inc. California.