Geographical variation in morphometry, craniometry, and diet of a mammalian species (Stone marten, *Martes foina*) using data mining

Malamati A. PAPAKOSTA¹, Kyriaki KITIKIDOU²*, Dimitrios E. BAKALOUDIS³, Christos G. VLACHOS¹, Evangelos CHATZINIKOS⁴, Olga ALEXANDROU⁴, Anastasios SAKOULIS⁵

¹Laboratory of Wildlife & Freshwater Fisheries, Department of Forestry and Natural Environment, Aristotle University of Thessaloniki, Thessaloniki, Greece
²Laboratory of Forest Biometry - Biostatistics, Department of Forestry and Management of Environment & Natural Resources, Democritus University of Thrace, Orestiada, Greece
³Fourth Hunting Federation of Sterea Hellas, Athens, Greece
⁴Society for the Protection of Prespa, Agios Germanos, Prespa, Greece
⁵First Hunting Federation of Crete and Dodekanisa, Chania, Greece

Received: 14.11.2016  •  Accepted/Published Online: 12.10.2017  •  Final Version: 10.01.2018

Abstract: Ecologists use various data mining techniques to make predictions and estimations, to identify patterns in datasets and relationships between qualitative and quantitative variables, or to classify variables. The aim of this study was to investigate if the application of data mining could be used to study geographical variation in the morphometry, craniometry, and diet of a mammalian species (*Martes foina*), and to determine whether data mining can complement genetic analysis to recognize subspecies. Morphometric, craniometric, and dietary data were collected from three different geographical regions in Greece (mainland, Aegean islands, and the island of Crete), and data mining techniques were applied. Our results showed that there is no geographical differentiation between morphometry, craniometry, and diet of the species; therefore, the species cannot be separated into subspecies. Our results support preliminary results from a genetic study that annuls previous classification into three subspecies. Data mining techniques could be used to examine the geographical variation of a species to support separation not subspecies.

Key words: Biogeography, craniometric data, data mining, dietary data, geographical variation, morphometric data

1. Introduction

Due to fast growing computational technology, large amounts of experimental data for complex biological systems have become increasingly available. This provides opportunities and challenges on how to efficiently and effectively handle these data for novel discoveries. Data mining, which is the process of statistically analyzing data from different perspectives and summarizing them into useful information and patterns, is of great importance in bioinformatics. For example, it is used in automatic classification, regression, clustering, and future selection of biological data (Frank et al., 2004). With more and different sources of data, it requires sophisticated computational analyses to study them. One main obstacle is how to analyze large noisy and heterogeneous datasets quickly and precisely (Zhang, 2013). Data mining algorithms and tools can be used to undertake these challenging and interesting computational problems in biogeography.

Additionally, analysis of ecological data can be used to describe geographical variation of a species to subspecies, mainly by studying morphologic and genetic characters as well as diet patterns (Benton, 1980; Powell and King, 1997; Virgos et al., 1999; Monteiro et al., 2003; Yom-Tov and Geffen, 2006; Alexandri et al., 2012; Papakosta et al., 2012, 2014).

Many researchers use statistical approaches to analyze ecological data in order to investigate relationships between a response variable and a set of predictors, using hypothesis tests. Data mining is a way for a user to collect large amounts of data and analyze the data in order to compare or link variables to give a prediction; this method has been applied to ecological data, i.e. to determine relationships between environmental factors and species distribution (Su et al., 2004) or to analyze the genome of vertebrate and fish species (Glusman et al., 2000; Rise et al., 2004) in order to uncover patterns and relationships in datasets.
2. Materials and methods

2.1. Field procedures and data collection

Samples (dead animals) were collected from three different regions in Greece: the mainland, the Aegean islands, and the island of Crete; they were animals hunted throughout the year (from 2003 to 2011). Data collected from these dead animals include morphological measurements (body and skull) and diet composition of the stone marten. Eighty-seven skulls, which belonged to 52 males and 35 females, from the mainland (n = 38) and insular Greece (Aegean islands n = 39, Crete n = 10), were measured. Body size variables were measured from 215 individuals (109 males and 106 females; mainland n = 119, Aegean islands n = 69, Crete n = 27). Diet composition was determined from 292 stomachs (151 males and 141 females; mainland n = 194, Aegean islands n = 71, Crete n = 27).

The variables for each dataset were (Table 1):
- Morphometric data: weight, height, body length, tail length, ear length, front paw length, back paw length.
- Craniometric data: sex, weight, age class (juvenile, subadult, adult), 45 variables for dimensions including height, width of the skull (see Table 1 and Figure 1 for landmarks of the variables of dimensions that were statistically important).
- Dietary data: percentages in stomach contents (Pineda-Munoz and Alroy, 2014), categorized in seven food groups, using microscopes and stereoscopes (mammals, birds-eggs, reptiles-amphibians, snails-worms, arthropods, plants-fruit, others), measured with two different methods: barometric (dry matter of food consumed) and frequency of occurrence.

2.2. Statistical analysis

In this study, we were interested in descriptive mining, i.e. segmentation – clustering of records – for cases with common characteristics. In these analyses, we did not want to define independent and dependent variables (input-output data), but rather to examine whether there is any trend (pattern) of data from the same geographical region: in other words, to investigate if there is any geographical differentiation. The following techniques were applied, using IBM SPSS Modeler v.14.2 (IBM Corporation, 2011a, 2011b), for the three groups of data (body size, skull, diet): K-Means is a cluster analysis with unsupervised learning. The algorithm aims to create internally homogeneous groups (clusters) while maximizing the variation between groups.

Two-step is a grouping analysis in two steps: during the first step, a manageable set of subgroups is created, while in the second step homogeneous subgroups are merged into larger groups.

Kohonen is a type of artificial neural network (ANN). In the neural network, each variable is called a neuron and the links between them synapses. A weight corresponds to each synapsis (synaptic weight), expressing its importance. With respect to the Kohonen network type, there is no need for distinction between input and output, while a K-network (K-net) or else a self-organizing map (SOM) is created among neurons. The SOM is a two-dimensional...
### Table 1. Input importance of variables (predictors) to clustering (groups’ separation).

<table>
<thead>
<tr>
<th>Segmentation method</th>
<th>Most important variable (predictor) for clustering (normalized importance)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Body size data</td>
</tr>
<tr>
<td>K-means</td>
<td>Sex (1.0000)</td>
</tr>
<tr>
<td></td>
<td>Age class (1.0000)</td>
</tr>
<tr>
<td></td>
<td>Body length (1.0000)</td>
</tr>
<tr>
<td></td>
<td>Palate length (Figure 1, landmark 5) (0.4283)</td>
</tr>
<tr>
<td></td>
<td>Distance between angular and coronary apophyses (Figure 1, landmark 6) (0.4593)</td>
</tr>
<tr>
<td></td>
<td>Intra-ophthalmic width (Figure 1, landmark 7) (0.5927)</td>
</tr>
<tr>
<td></td>
<td>Age class (1.0000)</td>
</tr>
<tr>
<td>Two-step</td>
<td>Sex (1.0000)</td>
</tr>
<tr>
<td></td>
<td>Age class (1.0000)</td>
</tr>
<tr>
<td></td>
<td>Body length (1.0000)</td>
</tr>
<tr>
<td></td>
<td>Palate length (Figure 1, landmark 5) (0.4283)</td>
</tr>
<tr>
<td></td>
<td>Distance between angular and coronary apophyses (Figure 1, landmark 6) (0.4593)</td>
</tr>
<tr>
<td></td>
<td>Intra-ophthalmic width (Figure 1, landmark 7) (0.5927)</td>
</tr>
<tr>
<td></td>
<td>Age class (1.0000)</td>
</tr>
<tr>
<td>Kohonen</td>
<td>Sex (1.0000)</td>
</tr>
<tr>
<td></td>
<td>Age class (1.0000)</td>
</tr>
</tbody>
</table>

**Figure 1.** Landmarks of craniometrical variables (1: length of jaw, 2: distance between the mastoid apophyses, 3: nose width, 4: width of cheekbones, 5: palate length, 6: distance between angular and coronary apophyses, 7: intraophthalmic width, 8: face length, 9: condylobasal length).
grid of neurons, with no synapses among them. The algorithm runs under unsupervised learning also and aims at the creation of groups (clusters) as homogeneous as possible, which will differ as much as possible between them.

The silhouette measure indicates whether the formation of groups (clusters) is poor, fair, or good, with respect to the cohesion and the separation (Kaufman and Rousseeuw, 2005). A silhouette measure equal to −1 means that all entries are in the wrong group, a measure equal to 0 means that all entries have the same distance from the center of the group where they belong and from the centers of the other groups, and a measure equal to 1 means that all entries are in the correct group.

The importance of an independent variable is a measure of how much the mining model-predicted value changes for different values of the independent variables. A sensitivity analysis to compute the importance of each predictor is applied (Streiner and Cairney, 2007).

3. Results
When examining if there are any clusters (i.e. groups) in the three datasets (body size, skull, diet), by applying the three algorithms (K-means, two-step, Kohonen), the two-step segmentation gave poor clustering quality, i.e. poor group separation, (silhouette measure < 0.25), in contrast to the K-means and Kohonen techniques, which resulted in good clustering quality, for all three datasets (Figure 2).

In Figure 3 clusters and cluster sizes for each dataset and each data mining technique are illustrated. We observe that clusters resulting from data mining, considering all variables of each dataset as input, cannot coincide with the three geographical regions (mainland, Aegean islands, Crete) from which data were collected; in other words, there is no matching overlap of clusters from data mining and clusters from geographical regions.

Previously, we stated that the geographical variation of the stone marten cannot be confirmed, neither for morphology nor diet patterns, whatever the data mining technique used. However, we can find that there is sexual differentiation, based on morphometric data: sex is an important variable for clustering (groups’ separation) when the K-means and Kohonen techniques, which have high silhouette measures, are applied (Table 1).

Based on the P-values of the chi-square independence test between clusters, formed from data mining techniques and clusters-geographical regions, we can conclude that the two classifications are independent (P > 0.05), i.e. geographical variation for body size, craniometrical, and dietary data cannot be established (Table 2).

4. Discussion
The successful implementation of data mining in this case study revealed that the stone marten is a predator with similar feeding patterns between geographic regions. There is no geographic pattern according to body size, craniometrical, and dietary data; in other words, animals from the three regions (mainland, Aegean, and Crete) cannot be distinguished based on morphology and diet. The results arising from data mining, which involves the analysis of qualitative and quantitative data, agree with those from the use of other statistical approaches that analyzed the same datasets separately for the stone marten in the Mediterranean (Papakosta, 2013). In that study it was concluded that the species cannot be separated into subspecies between mainland and islands, as opposed to previous analyses (Kryštufek, 2004a, 2004b), which were based on limited samples. The present study comes to the same conclusion using data mining, since no geographical patterns in body size, craniometry, and diet were detected. The results in this study are consistent with the preliminary results of genetic analysis of stone marten populations in Greece, which do not support either the existence of subspecies or the existence of a polymorphism characteristic for a specific Greek stone marten population (Papakosta et al., 2012).

The use of data mining reinforces the existence of sexual dimorphism in mustelids (Erlinge, 1977; Moors, 1980). According to Hedrick and Temeles (1989), three major hypotheses have been advanced to explain sexual differences in size and morphology based upon (a) sexual selection, (b) intersexual food competition, and (c) reproductive role division.

The similar diet pattern, which has been documented in the three Mediterranean regions, is due to the flexible and opportunistic feeding behavior of stone marten (Papakosta et al., 2014), which adjusts its foraging strategy to alternate available food.

The absence of significant geographical variation in the morphology and diet of the species in the Mediterranean could be explained by the lack of diverse populations before the last glacial period in Europe (Hofreiter et al., 2004). At the end of the Quaternary period, insular ecosystems were substantially different from those of the mainland. Paleontologists and archaeologists have found differences in insular and mainland wildlife with lower biodiversity in islands (Sondaar, 1971; Azzaroli, 1977). However, since prehistoric times, the colonists of the islands of the Mediterranean changed the endemic fauna with the import of mammals from the mainland (Masseti, 1995). Such a scenario could be true for the stone marten, which was imported from the mainland. In addition, it should
Figure 2. a) Silhouette measure for body size data. b) Silhouette measure for craniometrical data. c) Silhouette measure for dietary data.

Table 2. Chi-square independence test for regional pattern recognition.

<table>
<thead>
<tr>
<th>Data</th>
<th>Chi-square</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-means</td>
<td>2.613</td>
<td>0.271</td>
</tr>
<tr>
<td>Two-step</td>
<td>0.071</td>
<td>0.965</td>
</tr>
<tr>
<td>Kohonen</td>
<td>2.613</td>
<td>0.271</td>
</tr>
<tr>
<td>Craniometrical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-means</td>
<td>11.814</td>
<td>0.160</td>
</tr>
<tr>
<td>Two-step</td>
<td>5.146</td>
<td>0.076</td>
</tr>
<tr>
<td>Kohonen</td>
<td>13.036</td>
<td>0.222</td>
</tr>
<tr>
<td>Dietary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-means</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Two-step</td>
<td>0.400</td>
<td>0.819</td>
</tr>
<tr>
<td>Kohonen</td>
<td>2.182</td>
<td>0.999</td>
</tr>
</tbody>
</table>

be taken into consideration that the species has a flexible and opportunistic feeding behavior (Tikhonov et al., 2008; Bakaloudis et al., 2012) and morphological differences are not expected (Moors, 1980).

In conclusion, our study demonstrates that data mining could be used successfully as an ecological data handling technique. Data mining can describe traits regarding the physical characteristics, ecological niche, and functional role of species within ecosystems. There are benefits arising from the adoption of these approaches, including more open collaboration and web-based data-sharing (Baird et al., 2008). In recent years, the use of data mining methods has become increasingly familiar as a tool for addressing biogeographical and taxonomic problems. Relevant databases are often large in size and complex in structure, and their study deserves a wider appreciation of some of the biases, gaps, and potential drawbacks common to them (Soberón et al., 2000; Bhugra, 2013). The successful implementation of data mining in this case study, which revealed that there is no separation of a mammalian species (*Martes foina*) to subspecies, encourages the use of data mining as a technique for analyzing ecological data.

Acknowledgments
This work was a part of a PhD thesis and was partially supported by the Research Committee of Aristotle University of Thessaloniki (grant number: 89033). We would like to thank the Ministry of Agriculture Development & Food for the permission to collect samples and the Greek Hunting Federation for sample provision. The authors declare that the experiments comply with Greek and EU laws.
Figure 3. a) Cluster sizes for body size data. b) Cluster sizes for craniometrical data. c) Cluster sizes for dietary data.
References


Moors P (1980). Sexual dimorphism in the body size of mustelids (*Carnivora*): the roles of food habits and breeding systems. Oikos 34: 147-158.

Papakosta M (2013). Diet diversity of stone marten (*Martes foina*) in Mediterranean ecosystems. PhD, Aristotle University of Thessaloniki, Thessaloniki, Greece.


Zhang W (2013). Data mining for biological data learning: algorithm and application. PhD, University of Notre Dame, Notre Dame, IN, USA.