Segmentation of paleoecological spatio-temporal count data

Kari Vasko[1], Hannu Toivonen[2] and Atte Korhola[3]

[1]CSC – Scientific Computing ltd., Finland, Kari.Vasko@csc.fi, Phone: +358-9-4572734,
Fax: +358-9-4572302
[2]Department of Computer Science, University of Helsinki, Finland,
Hannu.Toivonen@cs.helsinki.fi
[3]Department of Hydrobiology, University of Helsinki, Finland, Atte.Korhola@helsinki.fi

Key words: Spatio-temporal data analysis, segmentation, analysis of compositional data

Segmentation analysis addresses the following data analysis problem: given a time series, find a partitioning of the sequence to segments that are internally homogenous with respect to the desired pattern language and cost function. We will consider applications of segmentation analysis towards analysis of paleoecological spatio-temporal time series data. Our emphasis is both on computational and model building issues.

We outline a probabilistic framework for the spatio-temporal segmentation problems that occurs in paleoecology and discuss computational issues that arise in this setting. To this end, there has been no solid theoretic framework behind the zonation task. For instance, the current methods for numerical zonation of biostratigraphic sequences, e.g., broken stick, are limited since they do not fully specify local and global likelihoods of the data and, thus, they do not provide explicit assumptions concerning the data generating mechanism [1].

We introduce as an application Dirichlet-Multinomial Bayesian segmentation model for spatio-temporal count data that occurs frequently in paleoecological data analysis. A typical example of paleoecological time series count data is a sediment core data or a set of sediment cores that consists of abundances of species. The most probable segmentation model for species count data can be used to identify environmental changes if the species composition is known to be sensitive to the environmental changes. For instance, an organism called chironomid can be used to identify likely changes in air temperature, since chironomids are known to be sensitive with respect to the air temperature. As a simple example suppose we know that species A prefers warm conditions and species B prefers colder conditions. Further, suppose we collect the data represented in Table 1. The data indicates that it is more probable that 3000 years before present (BP) the environmental conditions have been clearly warmer than 1000 to 2000 years BP. A reasonable guess could now be that there are two zones in the data set illustrated in Table 1: one that covers time points 1000 and 2000 years BP and another one that covers 3000 years BP.

We will discuss computational issues that are related to the determination of the number segments using the probabilistic approach we adopt. We will introduce an efficient technique to compute an approximation of the marginal likelihood of the Dirichlet-Multinomial segmentation model for paleoecological spatio-temporal count data, which is needed in the determination of the number of segments.

The framework we introduce is capable of analyzing multiple data sets, e.g. data sets that consist of several time series that are possibly collected in different spatial locations, unlike existing methods used by paleoecologists. This feature can be used to identify local vs. global changes as follows. Suppose we have two sediment cores A and B that were collected from two different spatial locations. Further, suppose that the both cores consist of abundances of the same organism, which, in turn, is known to be sensitive to the changes of the environment. Assume that at some time point there is a probable segment boundary in the core A but which is not probable in the core B. Then it can be argued that the corresponding indirect indication of the environmental change is not probably global one.

Layer id	*Species A*	*Species B*	*Time before present*
1	24	76	1000
2	15	85	2000
3	78	22	3000

Table 1: *Example count data for a segmentation task.*

We demonstrate using synthetic data that the proposed approximation gives more accurate predictions than Bayesian information criteria (BIC) altough the proposed approximation has the same time complexity as BIC. We will also consider and demonstrate performance of the state-of-the-art frequentist techniques [2,3]. Our experiments indicate that the Bayesian approach to zonation analysis is more suitable than the frequentist one, in particular, when short zones exist. Finally, we give demonstrations using real data.

References

[1] Bennett, K. Determination of the number of zones in a biostratigraphical sequence. *New Phytologist*, 132, 155-170, 1996.

[2] Vasko, K. and Toivonen, H. Estimating the number of segments in time series data. *Proceedings of The 2002 IEEE International Conference on Data Mining (ICDM'02)* pages 466-473.

[3] Vasko, K. *Computational methods and models for paleoecology*, Department of Computer Science Series of Publications A, Report A-2004-3 (PhD Thesis).