School of Engineering and Applied Science (SEAS), Ahmedabad University

B.Tech(ICT) Semester IV: Probability and Random Processes (MAT 202)

- Group No : H_B27
- Group Members
  - Yashil Depani (AU1841005)
  - Vidish Joshi (AU1841019)
  - Prayag Savsani (AU1841035)

- Project Title:

Seriation in Paleontological Data Using Markov Chain Monte Carlo (MCMC) Methods

1 Introduction

1.1 Background

Fossils and fossil sites provide important information necessary to reconstruct the history of different life forms and civilizations of Earth. The chronological ordering of these fossil sites and the correlation of these sites to determine their interdependence over time are two of the fundamental and important problems of paleontology. This relative ordering helps to determine the relative ages of sites, the species of animals found there and allows us to create a chronology of how history might have panned out.

Fossils have been vastly used for seriation and correlation in modern paleontology [1]. Large databases of fossil sites and their corresponding occurrences of taxa are available today [2]. Conventional paleontological techniques such as biostratigraphy [3] cannot be easily applied in most datasets as the data regarding the distribution of taxa at a local level is absent. On the other hand, data regarding rock distributions and temporal age estimates based on the same is available but hard to make use of in a global context. Thus to tackle this problem at a global level with large datasets, we need to consider both the local taxa distribution and the supplementary stratigraphic information. An example of this is the fossils collected from the *Eurasian* region which belong to the *Cenozoic* era. In that case, the marine and the terrestrial fossils are more closely related compared to other periods. Such data collected outside of the rock-context is difficult to process through conventionals methods like Biostratigraphy [4, 5].

Many different methods of numerical analysis have already been implemented for the purpose of seriation and correlation. Different data-sets comprise of various sets of variables . Also different parameters can be

used to define the behaviour of a variable in a data-set. This results in many unique modelling of the problem such as the PAST program [6], the graph-theoretical unitary associations method [7, 8], Bayesian approaches [9], etc. Our base article explores the same problem with a probabilistic approach as the local taxa distribution may turn out to be biased in different ways as it may only cover a smaller or larger area or time interval. Moreover, some of the presences recorded might just be weakly found specimens and in the same way, some of the absences might just be missing data. These observations in the dataset point towards a probabilistic analysis of the paleontological data available.

1.2 Motivation

The main motivation is to solve the problem of seriation of paleontological sites through probabilistic approach. Previous works such as the PAST program achieve this using Principle Coordinates Analysis(PCA) of the matrix or by graphical methods used by Guex while this method solves this by straightforward approach of finding an ordering with maximum Bayesian likelihood by MCMC methods. The available dataset is represented as an occurrence matrix. Moreover, a fossil site in this context is regarded as a snapshot of the taxa that were alive at around the same time in these sites. Assuming that the rows of the matrix represent the sites in observation and the columns represent the taxa, a '1' at the cell (i, j) represents the presence of j^{th} taxa at i^{th} site. And a '0' represents otherwise. These occurrence matrix can be analysed through probabilistic method to order the sites. This snapshot of taxa is, however, biased for various reasons as the time interval over which it is considered might be too short or too long, or the area over which the taxa are considered might be too small or too large, thus giving inaccurate relative picture of the taxa that were actually present. Thus the values in the matrix do not necessarily have same weights. Some taxa might have been present rarely in very few sites while the others might have been present frequently in large numbers of sites, thus showing not all 1's in the matrix are same. Similarly some 0's in the matrix might just represent absence of taxa, while other 0's can be due to deeper natural reasons that might have made the taxa absent at the site. Thus, all 0's in the matrix are not same. Then, there is the possibility of error in data in the form of a missing 1 or a wrong 1. These observations clearly indicate the requirement of probabilistic modelling and analysis of the presence-absence data to understand it, comment on it and order it.

1.3 Problem Statement/ Case Study

As discussed above, the occurrence matrix is bounded by errors in the collected data. This leads to the need of probabilistic analysis to determine the likelihood of the collected data.

The task here is to propose a model based on parameters that best describe the behaviour of the data-set as well as maximize the likelihood of the collected data. Considering the data-set, the parameters affecting its behaviour can be identified as errors in the data,

- Ordering of the sites
- Lifespan of each taxa

- Probabilities of errors i.e. probability of false positive and that of false negative

Here, the lifespan of a taxa is defined by the first and last occurrence of the taxa in the matrix.

More formally, if we denote the occurrence matrix by X and the set of parameters by parameter vector θ, then we need to find the parameter set that maximizes the likelihood of X. Now, considering any random possible ordering of the sites and observed lifespans of taxa, the likelihood of X is, thus, dependent on the errors in the data. The modelling should be such that it maximizes this likelihood.

The performance metrics for this problem are the expected values of probabilities of errors and the expected correlation of the predicted ordering with pre-existing orderings viz. the MN (Mammal Neogene) ordering. The closer the values of probabilities of false positive and false negative to that in the original data, more accurate we can say is the newly sampled data. Higher the correlation of predicted ordering with the MN ordering for instance, better the accuracy of the model.

2 Data Acquisition

The implementation and testing of this model requires a dataset derived from fossil sites to run the model and verify it with observed behaviours.

This implementation of the model, same as that of the base article, uses the data available in NOW database (http://www.helsinki.fi/science/now). This dataset consists of original collected data of European Late Cenozoic large mammals. It consists of sites of Eurasian continent and islands in the Mediterranean sea. This dataset is also available in the website (http://www.cis.hut.fi/projects/patdis/paleo) of the authors of the base article.

This dataset considers 3 different ages of a site: database age, MN (Mammal neogene) age and geochronologic age. We have data of some sites which predates some other sites. These sites are called hard sites. The hard sites in the data set are determined by referencing the MN ages [5] of those sites whose relative ordering is then fixed.

This data is then represented as the occurrence matrix, upon which MCMC methods are applied to obtain the pair-order matrix.

3 Probabilistic Model Used/ PRP Concept Used

The probabilistic model employed should be able to replicate the collected data (represented by matrix occurence matrix X) for a particular set of input parameters. As discussed above, these parameters are the ordering of the sites (denoted by π), the lifespan of taxa, where lifespan of each taxon m is denoted by a_m (first occurrence) and b_m (last occurrence) in the matrix, and the probability of errors where c is the probability of false positive and d is the probability of false negative.

Here, $\pi(i) < \pi(j)$ represents that site i is older than j. A taxon m is considered to be present at site k if $a_m \leq \pi(k) \leq b_m$. A presence outside this range is assumed to be an error. We assume log-uniform priors

for c (false positive) and d (false negative) in the intervals $0.001 \leq c \leq 0.1$ and $0.1 \leq d \leq 0.8$, respectively.

Knowledge of hard sites provides necessary apriori information to order sites. Without that information, the probability of $Pr(\pi(i) < \pi(j)) \sim Pr(\pi(j) < \pi(i)) \sim 0.5$ thus making it impossible to decide the ordering. Introducing hard-sites provides information that makes the ordering of other sites meaningful.

This parameters together form a parameter vector θ, so

$$\theta = (\pi, \bar{a}, \bar{b}, c, d)$$

Given this parameter vector θ, we need to maximize the likelihood of the occurence matrix X, which is dependent on the probability of errors. So, probability of matrix X for parameter vector θ is,

$$Pr(X|\theta) = c^{\alpha}(1-c)^{\beta}d^{\gamma}(1-d)^{\delta} \tag{1}$$

Here, α the number of false ones, β the number of correct zeros, γ the number of false zeros and δ the number of correct ones.

Errors and inconsistencies in the data, and the sites in observation being of the same time period result in the inability to find the perfect ordering of sites where probability of relative ordering of each site can be 0 or 1. Thus, there are some sites whose ordering is uncertain. We have $N!$ (where N is the number of sites) possible orderings out of which we have to draw samples where the seriation is uncertain i.e. finding pairs of sites for which the probability of one site pre-dating another is close to one-half.

Therefore, we work in the Bayesian framework, and find a sample of parameter vectors where the probability of a vector is proportional to its posterior probability.

$$Pr(\theta|X) \propto P(X|\theta)P(\theta) \tag{2}$$

$$\therefore \Pr(\theta|X) = \frac{\Pr(X|\theta)\Pr(\theta)}{\int \Pr(X|\theta)\Pr(\theta)} \tag{3}$$

The above integral doesn't have an analytical solution given the nature of our probability distribution and so we need to numerically solve this by drawing samples from the probabilistic space formed by the parameters. As parameter vector θ comprises of multiple random variables, so samples cannot be drawn directly using Monte Carlo method generates samples in such cases. Also, the probability of parameter vector is proportional to its posterior probability. That and the fact that this is a multi-dimensional probability distribution points towrds using Markov Chain Monte Carlo Methods to generate samples.

Using MCMC also allows us to skip a few steps. As mentioned before, the performance metric for our problem are the expectation values which can be calculated as follows:

$$E_{\Pr(\theta|X)}\{f(\theta)\} = \int \mathrm{d}\theta \Pr(\theta|X) f(\theta) \tag{4}$$

Since MCMC gives us let's say T samples of the parameters θ^t that satisfy the posterior distribution i.e. $\theta^t \sim \Pr(\theta|X)$, the above equation can be approximated as,

$$E_{\Pr(\theta|X)}\{f(\theta)\} \approx \frac{1}{T}\sum_{i=1}^{T} f(\theta^t) \tag{5}$$

the MCMC implementation generates the possible sample space and draws the best permutation from it, thus the effects of small changes in data can be observed. Also, in this problem, we have apriori information of hard sites in the data. This works as additional information while ordering the sites and MCMC modelling allows to set this order of hard sites before running the algorithm. Effectively, restricting the MCMC method to only accept permutations that satisfy this constraint.

4 Pseudo Code/ Algorithm

Since MCMC is very sensitive to the initialization step, we run 100 chains in parallel and then select the chains with the expected log-likelihood ($E[\log \Pr(X|\theta)]$) within one standard deviation of the best chain into account. In our case, we select 8 chains of the 100 total chains started.

As far as the initialization step is concerned, the ordering is totally random with the order of hard sites maintained. The lifespan is calculated from that and the interval is set to have no false ones. The probability of false positive is initialized to 0.01 and that of false negative is initialized to 0.3.

To make sure the random walk algorithm covers most of the probabilistic space or at least the high density portion of it, we run the chain for a burn in period of 10000 samples. Those samples are ignored. The final state of this burn-in period is used to initialize the actual chain which is further used in analysis. That chain is also ran for 10000 iterations where only the 10th sample is saved to overcome the problem of correlation among near samples in MCMC.

Sampling procedures:

- Ordering of sites:

  The random walk algorithm used here is the Metropolis-Hastings algorithm. Due to the small value of acceptance probability in case of permutation of sites, we try multiple MH proposals.

  The first sampling method is moving a site n with time index i to time index j all the while shifting the intervening sites and limits accordingly. Here, i and j are chosen randomly.

  The next step is to randomly choose one interval out of $[i, j+1]$, $[i, j+1[$, $]i, j+1[$, $]i, j+1]$ and reversing the order of sites in that given interval. Next is to reverse the non-hard sites in the same interval. Last step is to swap the neighbouring sites. Five iterations of all these methods is performed for every sample.

- Lifespan of taxa:

  To sample the parameters a_m and b_m, we calculate the relative likelihood of the data for all $a_m \in [0, b_m]$. Those likelihoods are then normalized to unity and new values of a_m are sampled from Multinomial$(p; b_m + 1)$. The sampling for b_m proceeds analogously.

- Probabilities of errors:

  $\log c$ is updated by sampling the proposal from the normal distribution, $\log c' \sim N(\log c, 0.15)$. The

proposal is accepted with the MH probability $min(1, \frac{\Pr(X|\theta')}{\Pr(X|\theta)})$. The sampling for d, proceeds analogously.

At the end of the sampling, k chains are selected out of the 100 total chains. In our case, $k = 8$. An individual chain gives expectation of $f(\theta)$ as follows,

$$E_{\Pr(\theta|X)}\{f(\theta)\} \approx \hat{F}_k = \frac{1}{T}\sum_{i=1}^{T} f(\theta_k^t) \tag{6}$$

The expectation given by all k chains is given as,

$$E_{\Pr(\theta|X)}\{f(\theta)\} \approx \hat{F} = \frac{1}{K}\sum_{i=1}^{K} \hat{F}_k \tag{7}$$

Using this, the seriation given by all the chains can be visualized by a pair order matrix defined as,

$$O_{ij} = \frac{1}{K}\sum_{i=1}^{K} O_{ij}^k \tag{8}$$

$$O_{ij}^k = E_{\Pr(\theta|X)}\{b(\pi(i) < \pi(j))\} \approx \frac{1}{T}\sum_{i=1}^{T} b(\pi(i)_k^t < \pi(j)_k^t) \tag{9}$$

where b is a boolean function.

5 Coding and Simulation

5.1 Simulation Framework

As discussed before, 100 chains are run to sample data. To speed up the computations, we use multiprocessing functions to run multiple (6 in our case) chains in parallel. This way it only takes 15-20 minutes to complete the sampling iterations. Each chain saves 1000 samples and out of the 100 total chains, we select 8 of them having their expected log likelihood within one standard deviation of the best chain. The dataset that we're using has two parameters, n_t i.e. minimum no. of occurences of a genus and n_s i.e. minimum no. of genus per site. All of the figures shown below use the subset of dataset where $n_t = n_s = 10$.

5.2 Results

Table 2. Results on the Large Mammal Dataset

n_t	n_s	N	M	Chains	$E\{c\}$	$E\{d\}$	CORRMN
10	10	124	139	8	0.0113	0.518	0.951
5	5	273	202	2	0.0068	0.661	0.925
10	2	501	139	2	0.0093	0.686	0.704

Figure 1: Results from the base article

n_t	n_s	N	M	Chains	$E\{c\}$	$E\{d\}$	CORRMN
10	10	124	139	8	0.0119	0.5127	0.940
5	5	273	202	2	0.0066	0.6699	0.926
10	2	501	139	2	0.0093	0.6833	0.669

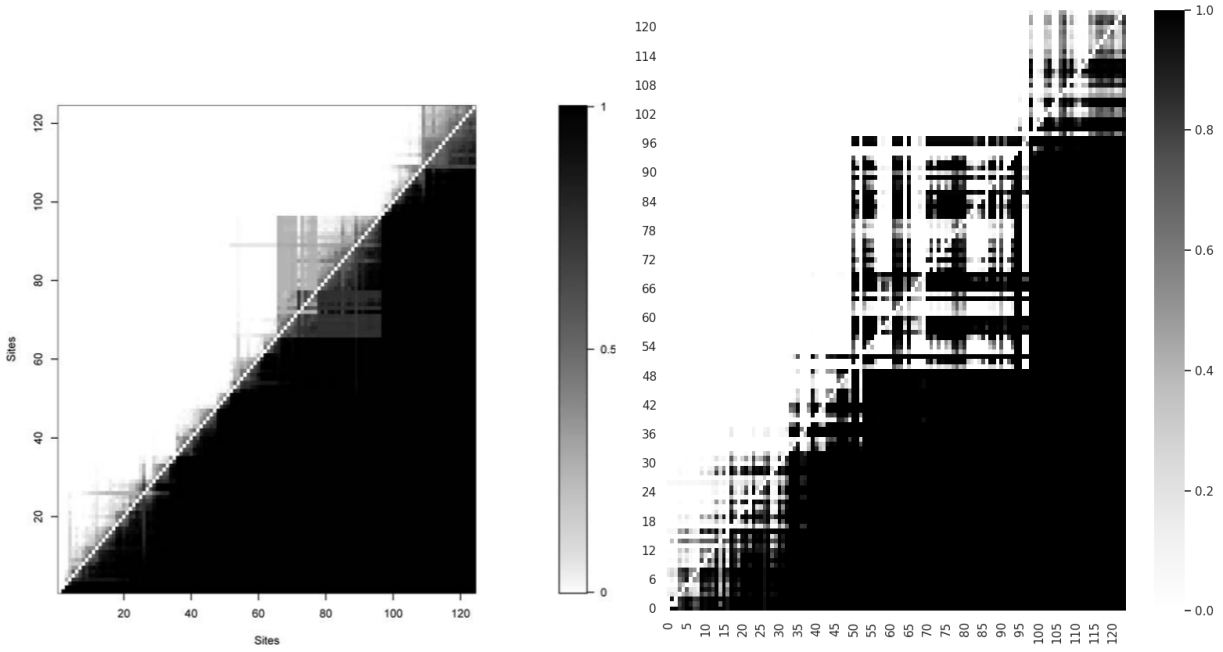
Table 1: Results reproduced

Here,

- N: No. of sites
- M: No. of taxa
- Chains: No. of chains selected
- $E\{c\}$: Expected probability of false positive
- $E\{d\}$: Expected probability of false negative
- CORRMN: $E\{corr(\pi, MN)\}$ i.e. the correlation between predicted seriation and seriation according to MN classification

5.3 Reproduced Figures

- Used Tools: Python, C
- Reproduced Figure-1

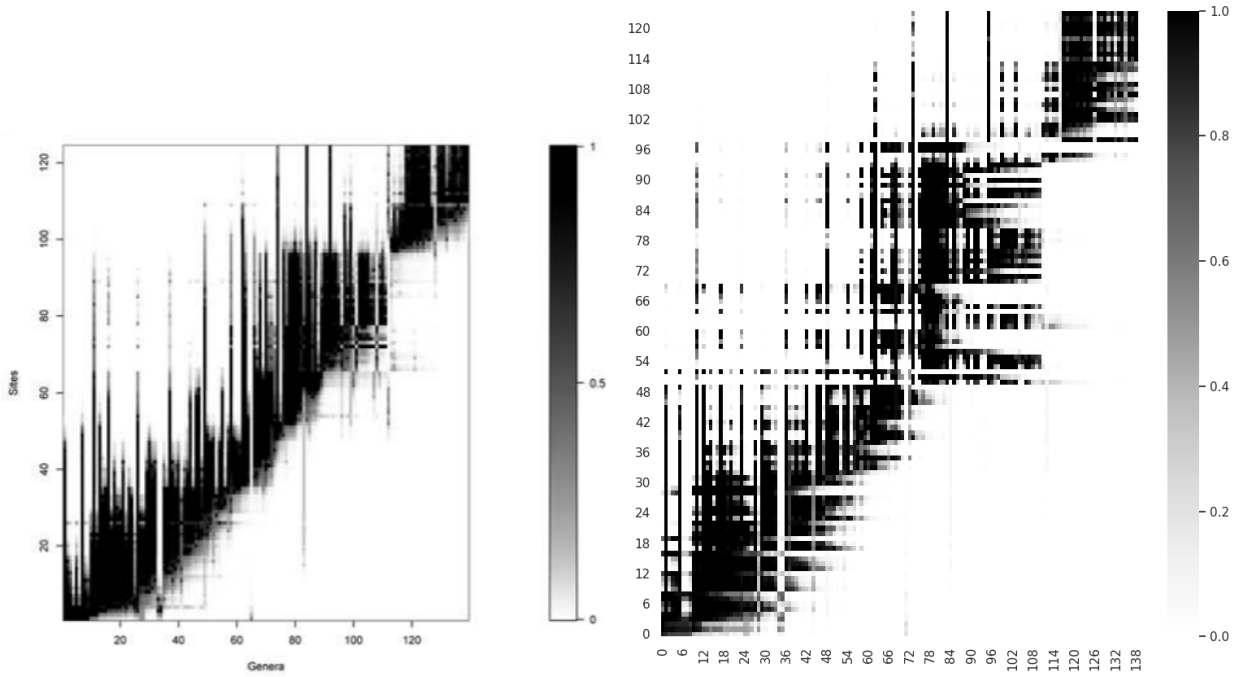


(a) Figure 1B (b) Figure 1R

Figure 2: Pair order matrix of sites

The inference to be derived here is dataset specific as the outliers and time intervals detected are the chief inference here. The probability pattern seen here is a straightforward paleontological interpretation based on the MN ages [10]. Starting from the lower left, the sites up to number 50 contain a sequence from the beginning of the Miocene at about 23 million years ago to the main faunal turnover event known in western Europe during this interval, the "Vallesian Crisis," at about 10 million years ago. Sites 51–63 represent mostly the first million years of post-crisis time, while the large block between sites 63 and 99 represents the relatively stable latest Miocene from 8 million to 5 million years ago, known as the Turolian. A major faunal turnover event separates the Turolian from the Pliocene sites 100–117, and sites 118–124 represent the beginning of the last epoch of the Cenozoic, the Pleistocene. Of the two blocks of localities that the model cannot order well internally, the first and largest thus corresponds to an interval during which little change happened in the mammal faunas. In contrast, the second block, spanning the later Pliocene and the early Pleistocene, corresponds to a time of rapid climatic and faunal change, characterized by the increasingly prominent alternation of cold and warm intervals and, especially in Europe, the cyclic alternation of their attendant "cold" and "warm" mammal assemblages.

- Reproduced Figure-2



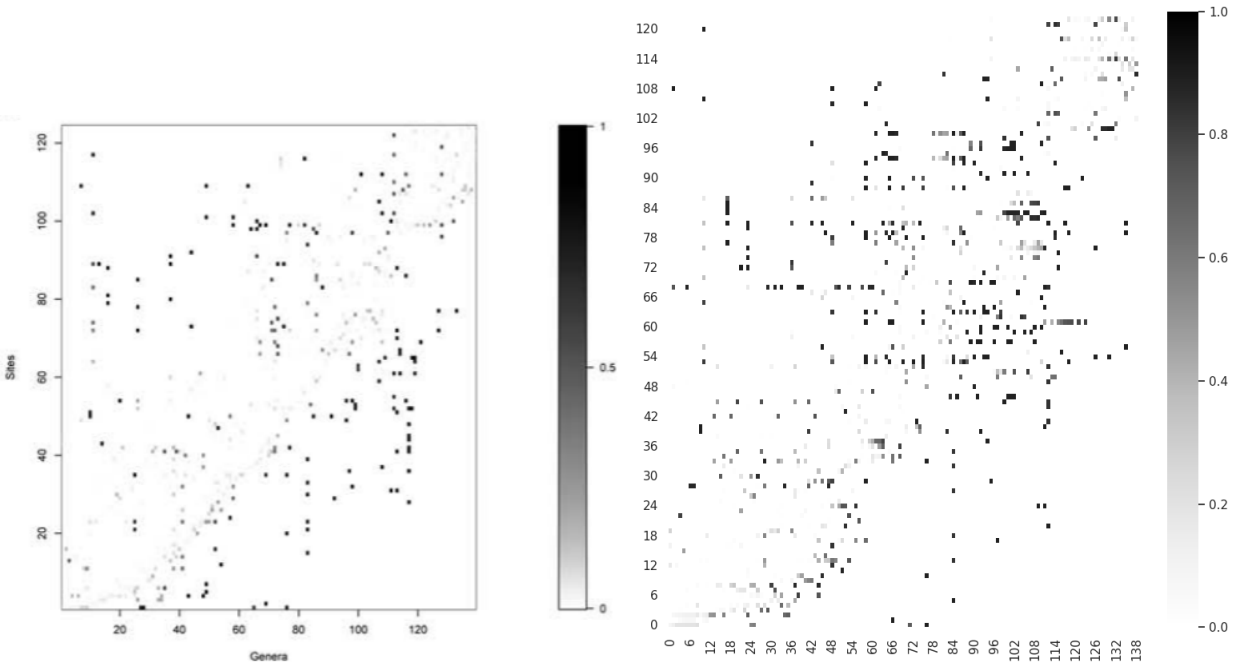
(a) Figure 1B

(b) Figure 1R

Figure 3: Newly ordered occurence matrix where sites are ordered by $E[\pi(n)]$ and taxa by $E[a_m]$

The structure just described also agress with the patterns of Figure 2, where the observed and estimated taxon ranges of the ordered sites trace a pattern of the three major faunal units separated by events where many genera go extinct and new, long-lived genera appear. Most marked of these is the turnover event associated with the Miocene–Pliocene transition at 5 million years ago, while the establishment of the two earlier faunal blocks appears more gradual. Figure 2 suggests that the model is especially skeptical of the early occurrences of genera, especially when their record is spotty, as for Tapirus (genus number 117 on the lower right).

- Reproduced Figure-3



(a) Figure 1B

(b) Figure 1R

Figure 4: Probability that an occurence is false

This figure shows that which occurence falls outside the lifespan interval (a_m, b_m) of a taxa. Some of the occurences that the model considers false are real errors in the data whereas others represent true outliers or are of genera with unusual species. For example, of the ten cases at the head of the dataset, some of them are true early occurrences, while some are isolated early European occurrences of a genus of African origin. Canis at Concud is a similar occurrence of a genus dispersing from North America, Plesiogulo at Pasalar is a missed hit from a published preliminary list, while Stephanorhinus at Belometchetskaya and Amphicyon at Stavropol are probable errors in the actual NOW database. Some apparent false occurrences reflect the biology of the animal in question. For example, the genus Tapirus, noted above for its spotty record, occurs eight times in the first 50 rows of probable errors.

6 Inference Analysis/ Comparison

The model described above has parameters having natural interpretation and the hard sites enable us to apriori knowledge. MCMC methods have the advantage of sampling various parts of the parameter space and the problem of convergence has been solved by sampling 100 chains and only taking the best ones from that. The pair order matrices generated by different chains are consistent with each other.

The results for late Cenozoic mammals (our dataset) show strong correlation with existing MN (Mammal Neogene) orderings ($CORRMN = 0.9413$). The model also indicates outliers with pretty good accuracy as the probabilities of errors ($E[c] = 0.0118$, $E[d] = 0.5150$) matches with that in the actual data. We can say that the model is a powerful tool to detect not only possible errors in the dataset but also detect genera with unusual distributional or ecological characteristics.

All the taxon and time period specific inferences are described above along with the respective figures.

7 Contribution of team members

7.1 Technical contribution of all team members

Tasks	Yashil Depani	Vidish Joshi	Prayag Savsani
Sampling procedures	✓	✓	✓
Simulation of figures			✓
Expected correlation and probabilities			✓
Inference analysis	✓	✓	✓

7.2 Non-Technical contribution of all team members

Tasks	Yashil Depani	Vidish Joshi	Prayag Savsani
Concept maps	✓	✓	✓
Abstract	✓	✓	✓
Report	✓	✓	✓

References

[1] M. J. OBrien and R. L. Lyman, *Seriation, stratigraphy, and index fossils: the backbone of archaeological dating.* Kluwer Academic Pub., 2002, pp. 59–60.

[2] Q. Schiermeier, "Paleobiology: Setting the record straight." *Nature 424*, pp. 482–483, 2003.

[3] H. D. Hedberg, *International stratigraphic guide: a guide to stratigraphic classification, terminology, and procedure.* Wiley-Interscience, 1976, p. 200.

[4] N. G. Lindsay, P. J. Haselock, and A. L. Harris, "The extent of grampian orogenic activity in the scottish highlands," *Journal of the Geological Society 146*, pp. 733 – 735, 1989.

[5] F. F. Steininger, W. A. Berggren, D. V. Kent, R. L. Bernor, S. Sen, and J. Agusti, "Circum-mediterranean neogene (miocene and pliocene) marine-continen- tal chronologic correlations of european mammal units." *New York: Columbia University Press*, pp. 7 – 46.

[6] Ø. Hammer, D. A. Harper, P. D. Ryan *et al.*, "Past: Paleontological statistics software package for education and data analysis," *Palaeontologia electronica*, vol. 4, no. 1, p. 9, 2001.

[7] J. Guex and E. Davaud, "Unitary associations method: use of graph theory and computer algorithm," *Computers & Geosciences*, vol. 10, no. 1, pp. 69–96, 1984.

[8] J. Savary and J. Guex, *Discrete biochronological scales and unitary associations: description of the BioGraph computer program.* Section des sciences de la terre, Institut de géologie et paléontologie . . . , 1999, no. 34.

[9] U. Halekoh and W. Vach, "A bayesian approach to seriation problems in archaeology," *Computational statistics & data analysis*, vol. 45, no. 3, pp. 651–673, 2004.

[10] R. Mittmann, R. Bernor, V. Fahlbusch, and H. Mittmann, *The Evolution of Western Eurasian Neogene Mammal Faunas.* Columbia University Press, 1996, p. 528. [Online]. Available: https://books.google.co.in/books?id=FEPBwAEACAAJ