BayLum - an R package for Bayesian Analysis of OSL Ages: an introduction.

Anne Philippe¹, Guillaume Guérin², Sebastian Kreutzer²

¹Laboratoire de mathématiques Jean Leray, Université de Nantes 2 rue de la Houssinière, BP 92208 44322 Nantes cedex 3, France

²IRAMAT-CRP2A, UMR 5060 CNRS - Université Bordeaux Montaigne, Maison de l’archéologie, Esplanade des Antilles, 33607 Pessac cedex, France

* Corresponding author: Anne.Philippe@univ-nantes.fr

Abstract We introduce a new R package ‘BayLum’ devoted to the analysis of OSL data and to construct chronological models in a Bayesian framework. Our contribution presents a summary of the statistical models, and introduces the main functions implemented in the package.

Highlights
- Introduction of a new R package
- Bayesian methodology and MCMC samplers
- Bayesian modelling of OSL ages
- Chronological modelling

Keywords
Age Model; Bayesian statistics; MCMC algorithm; Statistical software; Luminescence dating

1. INTRODUCTION

Bayesian methods provide powerful, alternative, analytical frameworks to process (luminescence) data. The beauty of Bayesian statistics comes with their explicit use of probability inferring quantities by adapting (unknown) model parameters, but leaving the original data untouched. During the last decade, it appears that Bayesian methods gained increasing attention by the luminescence dating community (to name a few, e.g., Rhodes et al., 2003; Millard 2006; Huntriss, 2008; Zink, 2015; Cunningham et al., 2015). More recent Bayesian models to analyse, in particular, Optically Stimulated Luminescence (OSL) data were presented by Combès et al. (2015) and Combès & Philippe (2017). On the flip side, however, mathematical models using Bayesian statistics easily become complex, demanding particular programming skills to be implemented. Mercier et al. (2016) implemented the ‘baSAR’ model by Combès et al. (2015) in the R (R Core Team, 2017) package ‘Luminescence’ (Kreutzer et al.,
Their work can be seen as a first step supporting a broader adaption of Bayesian methods in luminescence dating. Here, we considerably extend this work by combining the work by Combès et al. (2015) and Combès & Philippe (2017) in an own R package dedicated to Bayesian statistics called ‘BayLum’ (see CRAN release by Christophe et al. 2017a). We present the main functions included in the package ‘BayLum’ for estimating a sequence of ages based on OSL and dose rate measurements. Contrary to the other OSL data analysis softwares, ‘BayLum’ takes into account dosimetry related systematic and non-systematic errors due to the dosimetry. The statistical approach is based on a multivariate model with multiplicative Gaussian errors. Furthermore, the method is extended to a chronological model in order to combine measurements from other dating methods than luminescence dating, and to include additional information such as ordering constraints between ages of a given stratigraphic sequence. Note that we restrict the information derived from a stratigraphic sequence to the temporal ordering only i.e. we do not integrate information on the sedimentation process. To integrate an age-depth model, it would be necessary to also have information on the sediment accumulation rates. Theoretical questions on the construction of such model are still open.

The main interest of the Bayesian approach presented here is to provide a joint estimation of a sequence of ages. In our context, this joint distribution can be exploited to estimate archaeological, geological phases and/or to test, for instance, the presence of hiatuses between two dates or two groups of dates (see Philippe and Vibet, 2017 for details, and Philippe and Vibet, 2018, Guérin et al. 2017, Lahaye et al., 2018+ for applications).

In Section 2, we sketch the main ideas of Bayesian statistics, which we consider being a useful prerequisite when using R functions from the ‘BayLum’ package. We explain in particular the numerical and graphical outputs to support the application of ‘BayLum’ and help to interpret the results. We provide a first illustration through the estimation of the palaeodose of a sample from single aliquot regenerated (SAR, Murray and Wintle, 2000) dose measurements, with a palaeodose assumed to be common to all aliquots of the sample.

Section 3 is devoted to the description of the R functions related to the chronological modelling (see Table 1 for an exhaustive list of functions included in ‘BayLum’):

- To estimate a sequence of ages based on the previous central equivalent dose model, and including errors on dose rates.
- To estimate a sequence of ages including different dating methods (OSL, $^{14}$C) and prior information derived from stratigraphic constraints.

Internally, all the functions use JAGS (Just Another Gibbs Sampler), a program for analysing Bayesian hierarchical models using Markov Chain Monte Carlo (MCMC) simulations (see Plummer, 2003, 2016). JAGS is a separate program, which needs to be installed on top of R and the ‘BayLum’ package. Users can download JAGS from the web page of the mcmc-jags project at sourceforge (https://sourceforge.net/projects/mcmc-jags/files/).

At the beginning of an R session, first the package itself needs to be loaded.
library(BayLum)

### Loading required package: coda

In the following, we provide R code snippets to exemplify certain package functionalities. Moreover, all running examples are also provided in the so-called package ‘Vignette’, a document that comes with the R package itself, and can be easily kept up-to-date with new released versions.

## 2. SOME GENERALITIES ON THE BAYESIAN INFERENCE

### 2.1. Methodology

The Bayesian statistics (for an introduction see, e.g., Gelman et al., 2013) provide estimations in a parametric framework. The main interest of this approach is to combine the information coming from the observations with the prior knowledge available on the unknown parameters. In our context, the observations are the measurements \( M_1, \ldots, M_n \) provided by different dating techniques and the unknown parameters \( \theta \) are ages (or central doses in the case of a dose recovery experiment). This approach is well adapted to the construction of chronologies, because whenever working on a series of samples in a stratigraphic sequence, prior information is available on the ages; in particular the stratigraphic constraints define a temporal order.

The construction of a Bayesian model first requires the choice of a probability distribution \( L(\cdot | \theta) \) (the symbol \( ' | \theta \) means ‘depending on \( \theta \)’) for the observations \( M_1, \ldots, M_n \) and a translation of the prior information into a probability density \( \pi \) on the set of parameters \( \theta \). Then, applying the Bayes’ formula, the density of \( \theta \) is updated as follows:

\[
\pi(\theta | M_1, \ldots, M_n) = \frac{\pi(\theta)L(M_1, \ldots, M_n | \theta)}{\int_\Theta \pi(\theta)L(M_1, \ldots, M_n | \theta) d\theta}
\]

This density is called the posterior distribution, it contains all the information on the parameters (i.e. the prior information and the probability of the observations knowing the parameters). Statistical inferences may be based on this probability. As in the frequentist context, we can construct estimates of \( \theta \) by taking, for instance, the mean value, the median or the mode of the posterior density. The posterior variance is one summary of the spread of the posterior distribution. The larger the posterior variance, the more uncertainty we have in the estimation of the parameters.

The main idea of the Bayesian inference is to give a probabilistic estimation based on the determination of credible sets. A credible set is defined as being a subset of parameter space that refers conditionally to the observations, \( \theta \) belongs to this subset \( I \) with a desired probability \( 1 - \alpha (= 95 \% \text{ or } 68 \%) \):

\[
P(\theta \in I | M_1, \ldots, M_n) = 1 - \alpha.
\]

The set \( I \) can be constructed as the shortest credible interval or the region with the highest
posterior density (HPD), matching the condition (2). The HPD region is the most concise possible set limiting the desired probability. Note that these sets take into account the possible asymmetry of the probability distribution. Therefore, Bayesian statistics come up with a natural solution to the problem of summary statistics for skewed distributions.

Due to the complexity and the dimension of the considered models, the explicit form of the posterior distribution is not available. Thus, numerical approximation of the posterior density is required to approximate the Bayes’ estimates (mean value of the posterior distribution) and credible sets. A well-adapted solution is given by the Monte Carlo method using Markov Chains (called MCMC). The main idea is to simulate a stationary stochastic process, for which the stationary distribution is the posterior distribution. In other words, first the prior distribution is randomly sampled for each parameter; then, for each parameter the value is updated based on the observations (measured data). Convergence is reached once the distributions of each parameter do not vary any more (i.e. the parameters have reached a set of values which seem to reflect the data satisfactorily). Then, the posterior density is estimated by the histogram of the simulated sample (or a regularised version of the histogram). The empirical quantiles, mean, variance of the simulated sample allow us to approximate the Bayes’ estimate and the credible sets. The implementation of an MCMC sampler is standard, but the major difficulty comes from the fact that the method is not exact, but asymptotically precise only. Thus, it requires a step of convergence assessment. The number of MCMC iteration is crucial to obtain accurate estimates of characteristics of the posterior density.

We summarise the most commonly used diagnostics (see Plummer et al., 2006, for more details). These diagnostics require the simulation of multiple Markov chains. The idea is to accept the convergence when all the chains have the same stationary behaviour after a burn-in period. By increasing the number of chains we avoid the risk that a region of parameter space is visited by none of the chain. In practice we recommend at least three chains, but it is also important to increase this number with the complexity of the model (i.e. the number of analysed samples).

- The history or Trace Plots represent the values of the Markov chains recorded at each iteration. For cases where several Markov chains have been generated independently, when they all reach the same distribution, without any particular visible pattern, they are said to have reached their stationary distribution. In other words, the chains are well adapted if only noise like patterns are visible.

- The representation of marginal densities evaluated on independent Markov chains allows us to confirm that all simulated Markov chains reached the same stationary distribution. All the curves must be equal to accept the convergence of the MCMC algorithm.

To supplement the graphical analysis, convergence can be quantified using different criteria. The most popular is the Gelman-Rubin diagnostic. It requires the simulations of multiple Markov chains. This is a quantitative diagnostic, which combines two ways to estimate the variance of a stationary distribution: the within-chain and between-chain variances. The convergence of MCMC samplers is achieved when the value of the Gelman-Rubin statistic is close to 1 (see Gelman & Rubin, 1992). In practice, values below 1.05 may be deemed acceptable. To evaluate the precision of this estimation, additionally upper confidence limits are also calculated. The statistic must be smaller than this bound. This condition is necessary (but not sufficient) to conclude that all chains have converged.
The results of these diagnoses are given with each of the functions requiring an MCMC algorithm. Using the R package ‘coda’, the convergence assessment can be strengthened for instance by an analysis of Autocorrelation Plots. Such plots display the temporal dependence between the values of samples produced by the MCMC algorithms. When the autocorrelation is large at lag 1 and decays slowly as the lag increases, the Markov chain mixes slowly, and will converge slowly. It means that a very large number of iterations is required to obtain a precise and accurate estimation of the characteristics of the posterior distribution. The rate of subsampling for the output of MCMC algorithm can be increased in order to reduce the dependence between the values.

2.2. Illustration: Palaeodose_Computation()

This function allows evaluating a central dose from the observed luminescence signals of a set of aliquots. The applied model, described in detail by Combès et al. (2015), takes into account individual effects on a single aliquot level (which might be of single or multi-grain nature) in the sense that each aliquot is associated with a natural dose and a dose-response function. Regarding the latter, different parametric forms are included: the user may choose saturating exponential functions, with or without a linear component, and forcing (or not) the dose response curve through the origin.

The choice of a model for evaluating the central dose reflects the fact that all aliquots from one given sample can be described by one (central) dose. We consider individual palaeodoses \( D_j, j = 1, \ldots, n \) (where \( n \) is the number of aliquots). We assume that \( D_j = D + \epsilon_j \), where \( \epsilon_j \) represents the dispersion of \( D_j \) around \( D \). A scale parameter (the variance if it exists) parameterises the distribution of \( \epsilon_j \) in order to characterise the dispersion of the individual palaeodoses. Different distributions may be investigated for \( \epsilon_j \). In the original work by Combès et al. (2015), a Cauchy distribution (symmetric around zero) was chosen because it has the advantage of having infinite variance, ensuring robustness in terms of catching extreme values. Thanks to this robustness, it is not necessary to add a step to detect the presence of outliers (see Tribolo et al., 2017 for an application). Further distributions have been added in the package ‘BayLum’:

- the Gaussian distribution (with zero mean)
- the lognormal distribution with two parameter options: either the median or the mean of \( D_j \) is equal to \( D \).

The latter facilitates the asymmetry of the individual palaeodoses distribution (see Heydari and Guérin (2018), for tests of these models in laboratory-controlled experiments). By using the function Palaeodose_Computation(), we estimate the palaeodose \( D \) and the dispersion of \( D_j \) around \( D \).

Consider the following dataset available in the package:

data(DATA3)

# dataset from the article ‘Luminescence dating of the Palaeolithic sequence of La Ferrassie based on new excavations adjacent to the La Ferrassie 1 and 2 skeletons’ (Guérin et al., 2015)
The estimation of the palaeodose of this sample is obtained using the following code:

```r
P = Palaeodose_Computation(
  DATA = DATA3,
  Nb_sample = 1,
  SampleNames = c("FER1"),
  Iter = 1000)
```

To illustrate the behaviour of Markov chains having not reached a stationary distribution, we show in Figure 1 (top) the first iterations of the three chains. These values cannot be used to approximate the posterior distribution. Figure 1 (bottom) shows the MCMC output when the Markov Chains have converged. The estimations of the posterior density calculated on the three simulated chains give the same results. The Gelman-Rubin diagnostic confirms the convergence, since the statistic is equal to one (see the numerical output below). If we diagnose the convergence of MCMC algorithm, the rest of the numerical output can be analysed. It contains information about the inference on parameters:

- The mean value of the posterior distribution. This is the Bayes’ estimate of the parameter, which gives a central value for all the parameters - here the sample palaeodose and the dispersion of individual palaeodoses around it.
- The credible interval at level 95 % and 68 % for all the parameters.

## >> Results of the Gelman and Rubin criterion of convergence <<
## -----------------------------
## Sample name: FER1
## -----------------------------
## Point estimate Uppers confidence interval
## D_FER1  1  1
## sD_FER1  1  1

## >> Bayes estimates of Age, Palaeodose and its dispersion for each sample
## and credible interval <<
## -----------------------------
## Sample name: FER1
## -----------------------------
## Parameter  Bayes estimate  Credible interval
## D_FER1  73.119
## lower bound  upper bound
## at level 95%  64.29  83.36
## at level 68 %  67.31  76.55
##
## sD_FER1  9.183
## lower bound  upper bound
## at level 95%  2.33  18.06
## at level 68 %  3.99  11.23
3. AGE MODELS IN ‘BAYLUM’.

To date, three models are implemented in ‘BayLum’ corresponding to the functions

1. Age_Computation() and AgeS_Computation()
2. AgeC14_Computation()
3. Age_OSLC14()

We will not detail here the function Generate_DataFile() and Generate_DataFile_MG(), which allow generating the normalised OSL ratios and associated regenerative doses in Gy. While Generate_DataFile() is supposedly only to be used for single grain data, the latter should be used for multiple grain data.

3.1. AgeS_Computation()

The model

The function gives an estimate of the ages of a series of OSL samples, as described by Combès & Philippe (2017) (note: it is also possible to compute the age for one unique sample using the function Age_Computation(); Guérin et al. (2015) demonstrated the advantage of using such models for known-age samples). Observations are the natural and regenerated luminescence signals of each aliquot $j$ of each sample $i$ $L/T = (L/T)_{i,j}$, and the parameters of interest are the ages of each sample $A = (A^{(i)})_{i}$. The model is a hierarchical Bayesian model, where the palaeodoses of each sample $D = (D^{(i)})_{i}$ are included as unobserved variables (called latent variables). The hierarchical structure can be written in the form


The dose model (corresponding to $P(L/T|D)$) defines the relationship between the palaeodose of each sample and the natural luminescence signal of each aliquot. This model is the same as that implemented in the function Palaeodose_Computation() function to estimate the equivalent dose.

The age model (corresponding to $P(D|A)$) is based on the relation

$$D_{i} = A_{i}(d_{i} + e_{d_{i}} + \alpha_{i}e_{c}),$$

where the dose rate $d_{i}$ of sample $i$ is assumed to be known up to the error term. The originality of the modelling is the decomposition of the error term in two parts: the individual, random error (which is sample specific) $e_{d_{i}}$ and the systematic error (common for all samples) $e_{c}$, with the possibility to add an individual weight $\alpha_{i}$. The weight can be seen as a degree of systematic error contamination. All the error terms are assumed to be of Gaussian nature with
zero mean and a known variance. Equation (3) induces a dependence structure between the palaeodoses of each sample, the conditional distribution of $D$ given $A$ is a Gaussian vector $D|A \sim \mathcal{N}((A_i d_i), \Sigma)$ where the variance matrix $\Sigma$ is not diagonal. It is of the form

$$
\Sigma_{ii} = A_i \alpha_i \sigma_i^2 + \alpha_i^2 \sigma_c^2
$$

$$
\Sigma_{ij} = A_i A_j \alpha_i \alpha_j \sigma_c^2
$$

The values of $\alpha_i$, $\sigma_i$ and $\sigma_c$ are supposed to be known, hence these quantities must be provided by the laboratory in charge of the measurements (see the appendix in Combès et al, 2017, and Guérin et al., 2018, for a practical implementation of this matrix).

Finally, the prior distribution is assumed being non-informative; this is the Jeffreys prior, which is of the form

$$
\pi(A_1, \ldots, A_n) \propto \prod_{i=1}^{n} \frac{1}{A_i} 1_S(A_1, \ldots, A_n)
$$

where $S$ is a subset of the study period $[T_b, T_e]$ defining the temporal order between the ages.

For example,
- $S = \{A_1 < T_b < A_2 < \ldots < A_n < T_e\}$ if no stratigraphic constraint is known,
- $S = \{A_1, \ldots, A_n | T_b < A_1 < \ldots < A_n < T_e\}$ if the sequence of ages is supposed increasing.

We define by the same way the stratigraphic constraints for the three following functions: AgeS_Computation(), Age_OSLC14() and AgeC14_Computation(). Its construction is illustrated in the section devoted to Age_OSLC14().

**The study Period**

Using the most simple age model implemented in ‘Baylum’, we illustrate that it is important to pay attention to the choice of the study period. Indeed, this prior information cannot be corrected by the information coming from the measurements. For the dataset DATA3, we know that the true age is greater than 30 ka (see Guérin et al., 2015). We decide to choose a period, which does not contain this value, we fix the period [10; 20] (in ka) with the argument PriorAge.

The code is

```r
Age <- Age_Computation(
  DATA = DATA3,
  SampleName = "FER1",
  PriorAge = c(10, 20),
  Iter = 1000)
```

*Figure 2* shows the posterior distribution of the age. All the diagnostics done on the MCMC output indicate convergence of the chains. But as can be seen, the values of the simulated Markov Chain are concentrated on the upper bound of the study period and the credible interval is [13; 20] (in ka). The measurements ‘try’ to correct the period by forcing the posterior density to take high values in the proximity of 20 (in ka) but this bound cannot be
exceeded.

When the estimation is close to the bounds of the study period, it is recommended to check whether the results are still preserved when a larger period is fixed.

*Figure 3* shows the results with a larger study period \( \text{PriorAge} = c(10, 100) \). This confirms that the previous results were arbitrarily constrained by the chosen study period.

The numerical results associated with Figure 3 are given below. The calculated statistics are the same as previously. Here the list of parameters is the age, the palaeodose and the dispersion of individual palaeodoses.

```
##
## >> Sample name <<
## ---------------------------------------------
## FER1
##
## >> Results of the Gelman and Rubin criterion of convergence <<
## ---------------------------------------------
## Point estimate Uppers confidence interval
## A  1 1.01
## D  1 1.01
## sD 1 1
##
## ---------------------------------------------
## parameter  Bayes estimate  Credible interval
## ---------------------------------------------
## A  37.913
##   lower bound  upper bound
##      at level 95% 31.78 43.99
##      at level 68% 34.67 40.63
##---------------------------------------------
## D 73.044
##   lower bound  upper bound
##      at level 95% 63.61 84.06
##      at level 68% 67.59 76.87
##---------------------------------------------
## sD 9.131
##   lower bound  upper bound
##      at level 95% 1.65 17.36
##      at level 68% 4.29 11.17
```

From *Figure 1*, *Figure 3* and the numerical outputs, it can be seen that the posterior distributions of the palaeodose estimated by `Palaeodose_Computation()` or
Age_Computation() are the same. This result is reassuring because it shows that there is no error transfer between age and palaeodose in the age model.

3.2. AgeC14_Computation()

This is a function for calibrating radiocarbon dates/ages involving individual calibration curves. It is also possible to include temporal orders between the ages based on stratigraphic constraints. Conditionally to the calendar date \( A^* \), the radiocarbon age \( M^* \) follows a Gaussian distribution \( M^* \sim \mathcal{N}(g(A^*), s^2 + \sigma_g^2(A^*)) \) where \( g \) is the calibration curve and \( \sigma_g \) its error, and where \( s \) is the measurement error provided by the laboratory. Atmospheric South/North and Marine curves are included in the package, but it also allows including custom calibration curves via a CSV file.

Moreover, by fixing the argument Model = naive instead of Model = full (by default), we get the simplified model without error on the curve (i.e. \( \sigma_g^2(A^*) = 0 \) in the distribution of \( M^* \)). The temporal order is imposed via the prior distribution on the ages \( A_1^*,\ldots,A_n^* \). The prior distribution is the uniform law on \( S \), a subset of the study period \( [T_b,T_e] \) containing the temporal orders. An outlier model is also implemented obeying the same spirit as the models available in the software OxCal (Bronk Ramsey & Lee, 2013). We consider a mixture of two Gaussian components:

\[
M^* \sim (1-p)\mathcal{N}(g(A^*), s^2 + \sigma_g^2(A^*)) + p\mathcal{N}(g(A^*), \alpha(s^2 + \sigma_g^2(A^*))).
\]

The factor \( \alpha \) represents the uncertainty on the variance when the measurement is an outlier and \( p \) is the probability that \( M^* \) is an outlier. The parameter \( \alpha \) is assumed to be unknown and its prior distribution is an inverse gamma distribution (see Christen et al. (2009))

We illustrate this model on the dataset DATA_C14 available in ‘BayLum’. This R object contains four radiocarbon ages with the measurement errors. We assume that the study period is \([20,60] \) (in ka). The radiocarbon calibration curve is IntCal13 Northern Hemisphere atmospheric.

The code is the following

```r
library(BayLum)
#load example data
data(DATA_C14)

#14C age with error
C14Cal=DATA_C14$C14[1]
SigmaC14Cal=DATA_C14$C14[2]
```
To complete the numerical and graphical results already presented for previous functions, the posterior distributions of the multiple estimated ages can be summarised as shown in Figure 4. For each calibrated age, we display the mean value of the posterior distribution and the credible intervals. When an outlier is detected, a warning appears in the graphical representation of the posterior distribution. Moreover the numerical output of this function (the object Age) contains the list of outliers in the object Age$\text{outlier}$. Note that it is also possible to include stratigraphic constraints between $^{14}$C ages. The implementation is explained in the next section devoted to the function Age_OSLC14().

### 3.3. Age_OSLC14()

The function allows us to estimate a sequence of ages from measurements coming from both OSL and radiocarbon dating; stratigraphic constraints may also be taken into account. Conditionally to the ages, the OSL and radiocarbon measurements are supposed to be independent. Moreover, they satisfy the models implemented in Age_Computation() and AgeC14_Computation().

The prior distribution on the ages is given by

$$
\pi(A_1, \ldots, A_n, A'_1, \ldots, A'_m) \propto \prod_{i=1}^{n} A'_i - \prod_{i=1}^{m} A'_i,
$$

where $\mathbf{S}$ defines the temporal ordering on $A, A'$. 

The final example illustrates the joint estimation of OSL and radiocarbon ages. For this function in addition to the dataset, we have to specify the type of the dating method and the stratigraphic constraints. We exemplarily consider two OSL samples (see Tribolo et al. 2017) and a fictive radiocarbon age. We use the function Concat_DataFile() to concatenate the OSL samples (named GDB5 and GDB3) available in the package.

With the matrix samplenature we define the dating method for each sample. The values 1 on the first (resp. second) line indicate OSL (resp. $^{14}$C) samples. The column number identifies the sample through its position in the vector of sample names (SampleNames). So in our example, only the second sample (named MyAge) is a $^{14}$C age. With the matrix SC we define the
stratigraphic constraints. The matrix can be defined by using interactive function SCMatrix(). The answer is 1 if the sample i is younger than sample j.

We compare two situations:
1. The sample ages are consistent with the imposed stratigraphic constraints
   \[ \text{GDB5 < MyAge < GDB3} \, \]
2. the $^{14}$C age and the second OSL age are in stratigraphic inversion with respect to the model
   \[ \text{GDB5 < GDB3 < MyAge}. \]

For these examples, we run the function Age_OSLC14 with a matrix imposing the stratigraphic constraints corresponding to case 1 (i.e. the data in case 2 is inconsistent with the matrix): the three samples are totally ordered and supposed to be of increasing ages. In such a case, we can also use the function SC_Ordered(3) to define the matrix SC (see Table 1).

The code reads as follows (for the situation SC1):

```r
data(DATA1, envir = environment())
data(DATA2, envir = environment())
Data <- Concat_DataFile(DATA2, DATA1)

# 14C data sample called "myAge"
C14Cal <- 30500
SigmaC14Cal <- 100

# Prior Age
prior = rep(c(1, 60), 3)

# Definition of the dating methods
samplenature = matrix(data = c(1, 0, 1, 0, 1, 0), ncol = 3, nrow = 2, byrow = TRUE)
print(samplenature)

# [1,] 1 0 1
# [2,] 0 1 0

# Definition of the temporal ordering
SC <- SCMatrix(SampleNames = c("GDB5", "MyAge", "GDB3"), Nb_sample = 3)
print(SC)

# [1,] 1 1 1
# [2,] 0 1 1
# [3,] 0 0 1
# [4,] 0 0 0
```
## Age computation of samples

```r
Age <- Age_OSLC14(DATA=Data,Data_C14Cal=C14Cal,Data_SigmaC14Cal=SigmaC14Cal,
StratiConstraints = SC,
SampleNames=c("GDB5","MyAge","GDB3"),
Nb_sample=3,
SampleNature=samplenature,
PriorAge=prior,
Iter=10000,
Nb_chaines=3)
```

The numerical results are given below. The set of parameters in the output is limited to the sequence of ages. This function does not return statistical inference on the palaeodoses associated to the OSL ages.

**Model SC1**

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Bayes estimate</th>
<th>Credible interval:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_GDB5</td>
<td>7.176</td>
<td>lower bound</td>
</tr>
<tr>
<td></td>
<td></td>
<td>at level 95%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>at level 68%</td>
</tr>
</tbody>
</table>

**Model SC2**

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Bayes estimate</th>
<th>Credible interval:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_GDB5</td>
<td>7.149</td>
<td>lower bound</td>
</tr>
<tr>
<td></td>
<td></td>
<td>at level 95%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>at level 68%</td>
</tr>
</tbody>
</table>
Both models SC1 and SC2 give similar results for the samples GBD5 and MyAge. Figure 5 presents the comparison between the two choices of stratigraphic constraints. The condition SC2 leads to a stratigraphic inversion in the sense that the sample ages are not compatible with the temporal ordering imposed by the prior information. As shown in Figure 5, the radiocarbon age appears in this case as a bound for the OSL sample GDB3, whereas the age of the sample obtained by radiocarbon dating is still well estimated. The weight of the radiocarbon age in the construction of the chronology accounts for its precision.

4. CONCLUSION

We introduced a new R package ‘BayLum’ dedicated to Bayesian statistics. The package (current version: 0.1.1) provides a statistical solution for the analysis of OSL data by taking into account systematic and individual errors. Functions are also available to construct chronologies including radiocarbon dating and stratigraphic constraints. Future releases plan to include models
- for the systematic detection of outliers (currently limited to radiocarbon measurements), and
- for poorly bleached sediment samples using the model proposed by Christophe et al. (2017b).

The prior information taken into account is currently limited to the study period, and the temporal arrangements. A future challenge will be the custom integration of more diverse information such as information on the duration of a period or the depths in the stratigraphy for age-depth modelling. Theoretical questions on the construction of such prior distributions are still open, however.

5. REFERENCES


Heydari, M., Guérin, G. OSL signal saturation and dose rate variability: investigating the behaviour of different statistical models. Radiation measurements, this issue.


<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Modelling</strong></td>
<td></td>
</tr>
<tr>
<td>AgeC14_Computation</td>
<td>14C age estimation of multiple samples</td>
</tr>
<tr>
<td>AgeS_Computation</td>
<td>OSL age estimation of multiple samples</td>
</tr>
<tr>
<td>Age_Computation</td>
<td>Particular case of AgeS_Computation for one sample.</td>
</tr>
<tr>
<td>Age_OSLC14</td>
<td>Estimation of a sequence of ages included OSL and 14C ages.</td>
</tr>
<tr>
<td>Palaeodose_Computation</td>
<td>Palaeodose estimation of multiple OSL samples</td>
</tr>
<tr>
<td><strong>Graphical representation</strong></td>
<td></td>
</tr>
<tr>
<td>MCMC_plot</td>
<td>Trace Plots and representation of marginal densities evaluated on MCMC output Markov chains</td>
</tr>
<tr>
<td>ScatterSamples</td>
<td>Graphical representation of all pairwise posterior distributions.</td>
</tr>
<tr>
<td><strong>Data Manipulation</strong></td>
<td></td>
</tr>
<tr>
<td>Concat_DataFile</td>
<td>Merge two object provided by Generate_DataFile or Generate_DataFile_MG functions</td>
</tr>
<tr>
<td>Generate_DataFile</td>
<td>Generates, from BIN file(s) of Single-grain OSL measurements, the list of values required for the Bayesian estimation of ages implemented in AgeS_Computation, Age_Computation and Age_OSLC14</td>
</tr>
<tr>
<td>Generate_DataFile_MG</td>
<td>Generates, from BIN file(s) of multiple-grain OSL measurements, the list of values required for the Bayesian estimation of ages.</td>
</tr>
<tr>
<td>LT_RegenDose</td>
<td>Graphical representation of Lx/Tx values as a function of Regenerative Dose for selected grains.</td>
</tr>
<tr>
<td>SCMatrix</td>
<td>Interactive function to define the stratigraphic constraint matrix used in function AgeS_Computation, AgeC14_Computation and Age_OSLC14</td>
</tr>
<tr>
<td>SC_Ordered</td>
<td>Particular case of stratigraphic constraint matrix when the samples are totally ordered and supposed increasing.</td>
</tr>
</tbody>
</table>

Table 1. List of the functions implemented in the current version (0.1.1) of ‘BayLum’
Figure 1: Estimation of the palaeodose (D) and the dispersion (sD). Graphical diagnostic for MCMC output: history plots (left), and marginal densities (right) are shown for three independent Markov Chains. The top plots show Markov chains which are not stationary. Indeed, the beginning and the end of the trajectories do not behave similarly. For the bottom plots, convergence is achieved. Indeed, if two blocks are exchanged, the aspect of the trajectory is preserved.
Figure 2: Example for an age estimation of sample FER1 for the case of wrongly chosen prior information on the age. Representation of MCMC output and the marginal posterior density of age (A).
Figure 3: Age estimation of the sample FER1 when the study period is $[10,100]$ (in ka). The graphical output contains the MCMC output and the marginal posterior density of the following parameters: the age (A), the palaeodose (D) and the dispersion (sD).
Figure 4: Multiple calibration of four radiocarbon ages. The red point is the mean value of the posterior distribution and the blue lines represent the credible intervals at level 68% and 95%.
Figure 5: Estimation of a sequence of ages with stratigraphic constraints. The prior information imposes the temporal ordering GDB5 < MyAge < GDB3 [top] GDB5 < GDB3 < MyAge [bottom]. For each estimated age, the red point indicates mean value of the posterior distribution (Bayes’ estimate) and the blue lines represent the credible intervals at level 68% and 95%.