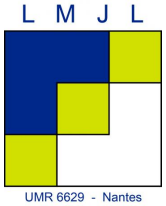


R et la statistique bayésienne appliquée à l'archéologie

Anne Philippe
Laboratoire de mathématiques Jean Leray
Université de Nantes, France
Anne.Philippe@univ-nantes.fr

Mars 2018, Meetup R Nantes

Plan

Contexte : Modèle paramétrique

On observe une réalisation d'un vecteur aléatoire $x_1, \ldots, x_n$

$$x = (x_1, \ldots, x_n) \sim f_\theta^{(n)}(x), \qquad \theta \in \Theta \text{ est inconnu}$$

On suppose que la famille de lois $\{f_\theta^{(n)};\ \theta \in \Theta\}$ est connue

Objectif : l'estimation du paramètre θ à partir

1. des observations $x_1, \ldots x_n$
2. des informations complémentaires $\rightsquigarrow$ information a priori

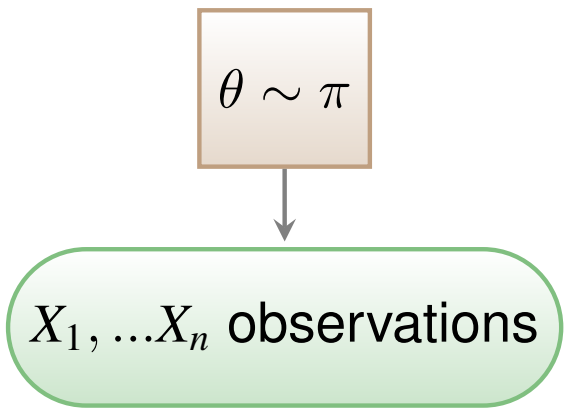
Approche bayésienne

- Connaissance à priori sur le paramètre θ est représentée par une probabilité π sur Θ.
- Le paramètre inconnu devient une variable aléatoire comme les observations

- On interprète la loi des observations f_θ comme la loi conditionnelle des observations sachant θ

$$f(x|\theta) = f_\theta(x)$$

Définition
π est la loi *a priori* sur θ.



Inférence Bayésienne

loi a priori
$\theta \sim \pi$

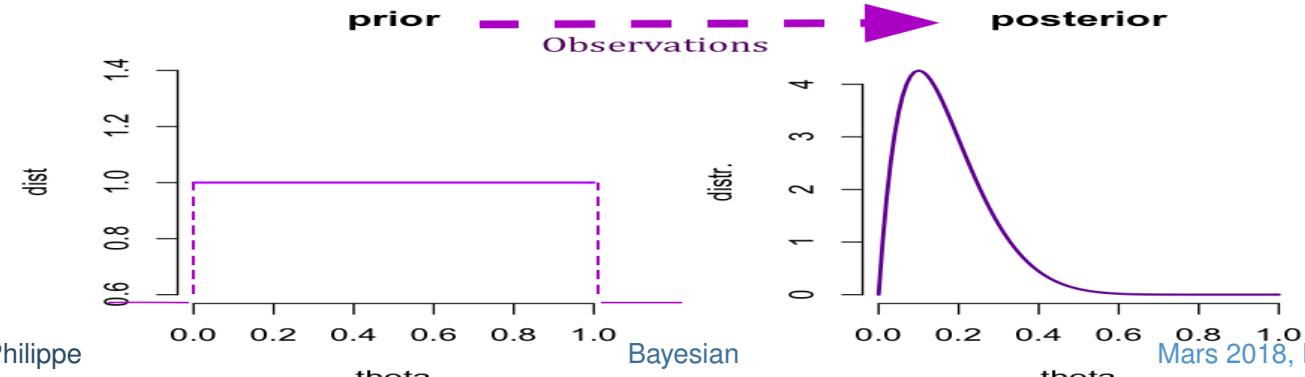
Observations
$x_1 \ldots x_n \sim f(x|\theta)$

On actualise la loi sur θ
àÂ partir des observations

$$\pi(\theta|x) = f(x|\theta)\frac{\pi(\theta)}{m(x)}.$$

Définition

La loi conditionnelle de θ sachant les observations x est appelée loi a posteriori



Echantillon Gaussien

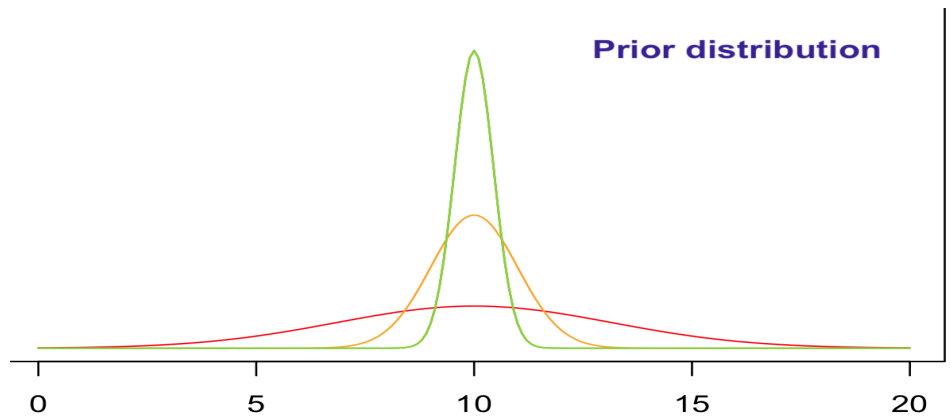
- Observations : n mesures du paramètre θ inconnu, avec une erreur gaussienne centrée et de variance s^2

$$M_i = \theta + \epsilon \qquad \epsilon \overset{iid}{\sim} \mathcal{N}\left(0, s^2\right)$$

- Information a priori sur θ : θ est proche de 10
- On traduit cette information en supposant que θ suit une loi gaussienne de moyenne 10.

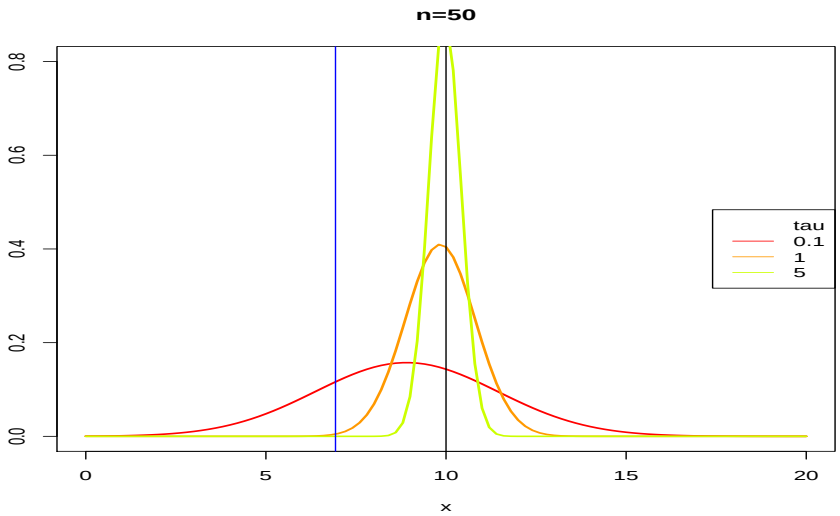
$$\theta \sim \mathcal{N}(10, \frac{1}{\tau})$$

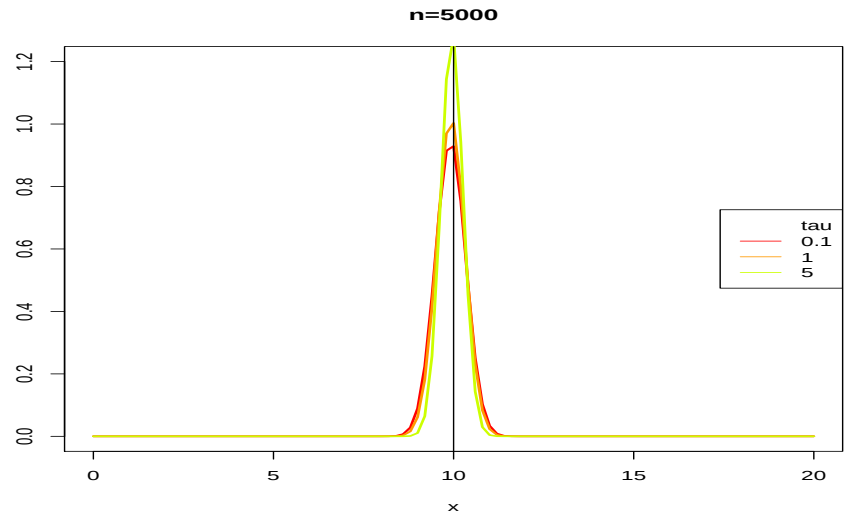
choix de la valeur de τ ?
Il dépend de la confiance que l'on va accorder à l'information a priori



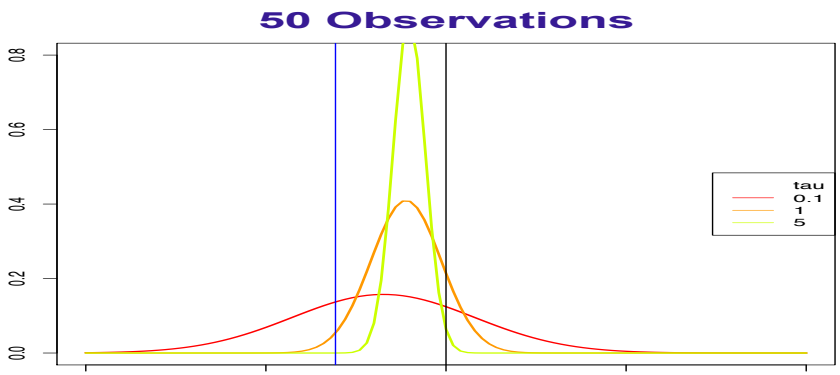
Comportement des lois a posteriori

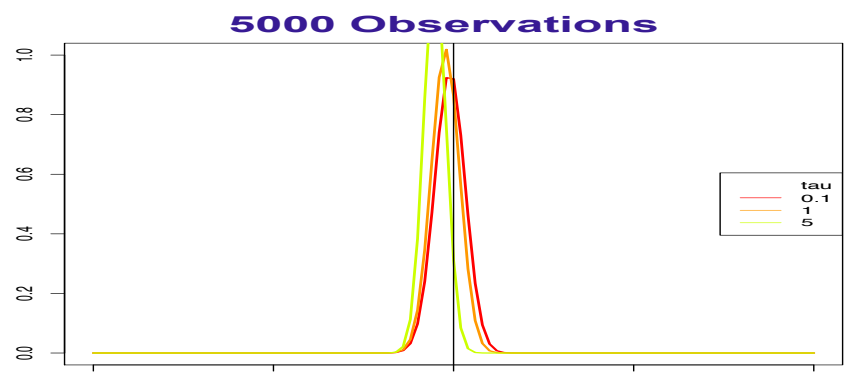
$\theta = 10$





$\theta \neq 10$





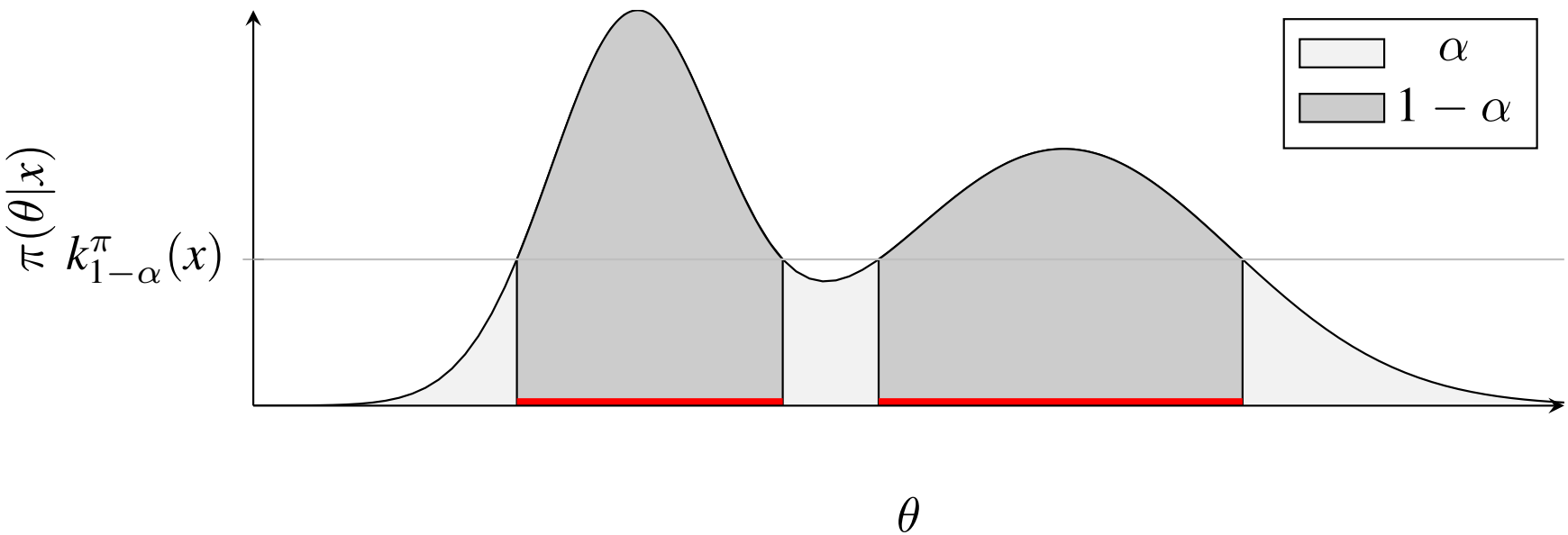
Région de de crédibilité

- On fixe $1-\alpha$ un niveau de confiance.
- On cherche les région de plus haute densité a posteriori. :

$$Q^{\pi}_{1-\alpha}(x) = \left\{\theta;\ \pi(\theta|x) \geq k^{\pi}_{1-\alpha}(x)\right\},$$

- $k^{\pi}_{1-\alpha}(x)$ est choisi tel que

$$\int_{Q^{\pi}_{1-\alpha}(x)} \pi(\theta|x)\mathrm{d}\theta = 1-\alpha$$



Estimateurs de Bayes

- La moyenne de la loi a posteriori
- La médiane de la loi a posteriori
- L'estimateur du maximum a posteriori (MAP) la valeur de θ qui maximise la densité de la loi a posteriori

$$\tilde{\theta}_n = \arg\max \pi(\theta|x_1, .., x_n)$$

Approximation de Monte Carlo

L'inférence est calculée à partir de la loi a posteriori.
Elle utilise

- sa densité
- sa fonction quantile
- sa constante de normalisation
- ses moments
- son mode

Ces quantités ne sont généralement pas connues explicitement et une approximation de Monte Carlo est nécessaire pour calculer

- les régions de crédibilité ,
- les estimateurs de Bayes

Description de la méthode

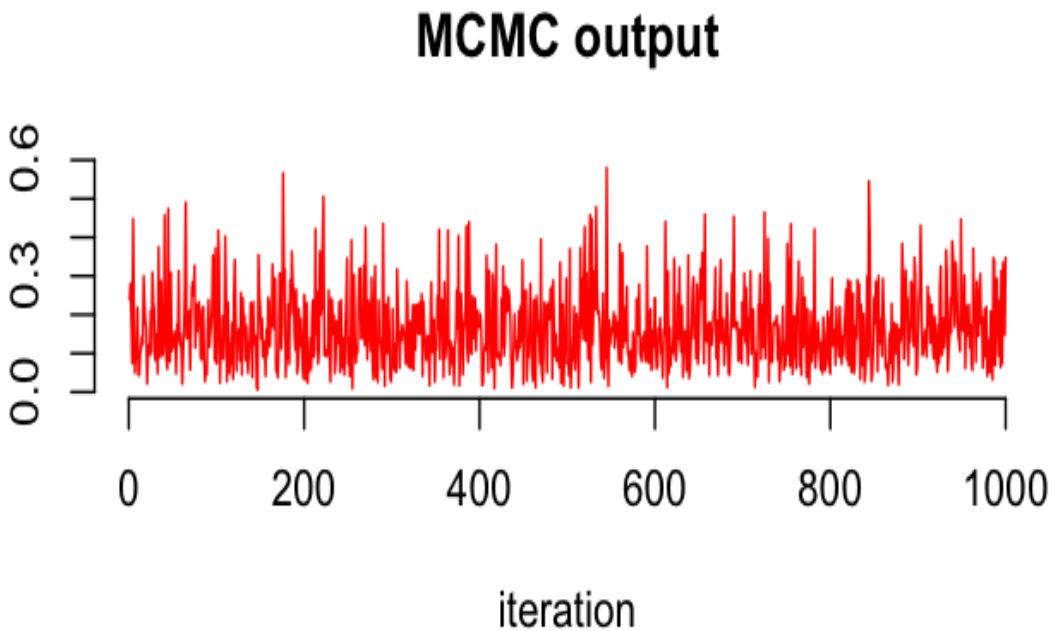
Simuler

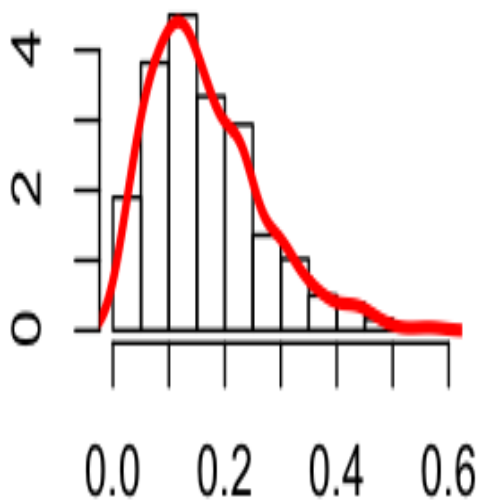
$\theta_1, \ldots, \theta_m \sim \pi(\theta|\, x)$ ou une loi qui approche $\pi(\theta|x)$

méthode exacte ou méthode (MCMC)

A partir de l'échantillon simulé, on estime

- la densité $\pi(\theta|x)$ par un estimateur classique (histogramme / estimateur à noyau) calculé sur $\theta_1, \ldots, \theta_m$
- la moyenne par la moyenne de l'échantillon
- etc





La solution R

- les logiciels STAN / JAGS /WINBUGS permettent de simuler un échantillon suivant la loi a posteriori (boite noire)
  - Input : la loi des observations & la loi a priori
  - Output : échantillon suivant la loi a posteriori

- Librairies R : `rjags` , `rstan`, `R2WinBUGS`
  - pour lancer les applications Ã partir de R
  - Intérêt : récupérer des objets R compatibles avec d'autres libraries
  - Exemple : `coda` pour contrôler la convergence de la méthode numérique

- Librairies pour des modèles spécifiques
  - `bmixture` modèle de mélange
  - `MCMCpack` regression linéaire , logit, linéaire généralisée etc
  - `BMS` regression etc... avec sélection de variables
  - `https://cran.r-project.org/web/views/Bayesian.html`

Plan

Bayesian approach to Interpreting Archaeological Data

The statistical modelling within the Bayesian framework is widely used by archaeologists :

- 1988 Naylor , J . C. and Smith, A. F. M.
- 1990 Buck C.E.
- 1994 Christen, J. A.
- etc

Examples

- Bayesian interpretation of 14C results , calibration of radiocarbon results.
- Constructing a calibration curve.
  to convert a measurement into calendar date
- Bayesian models for relative archaeological chronology building.

Observations

Each dating method provides a measurement M, which may represent :

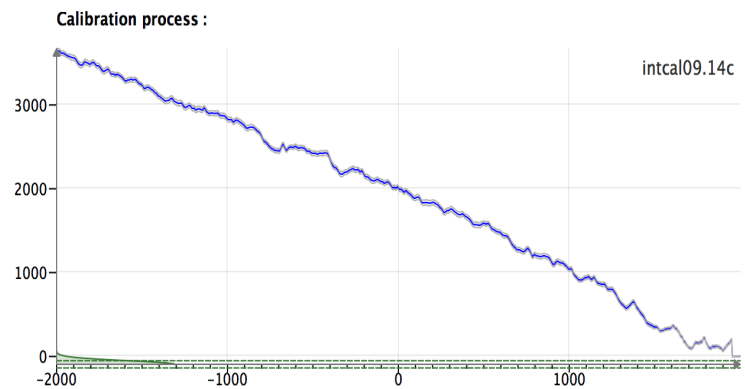
- a 14C age,
- a paleodose measurement in TL/OSL,
- an inclination, a declination or an intensity of the geomagnetic field

Relation with calendar date

$$M = g(\theta) + \epsilon$$

where

- θ is the calendar time
- g is a calibration function which relates the measurement to θ

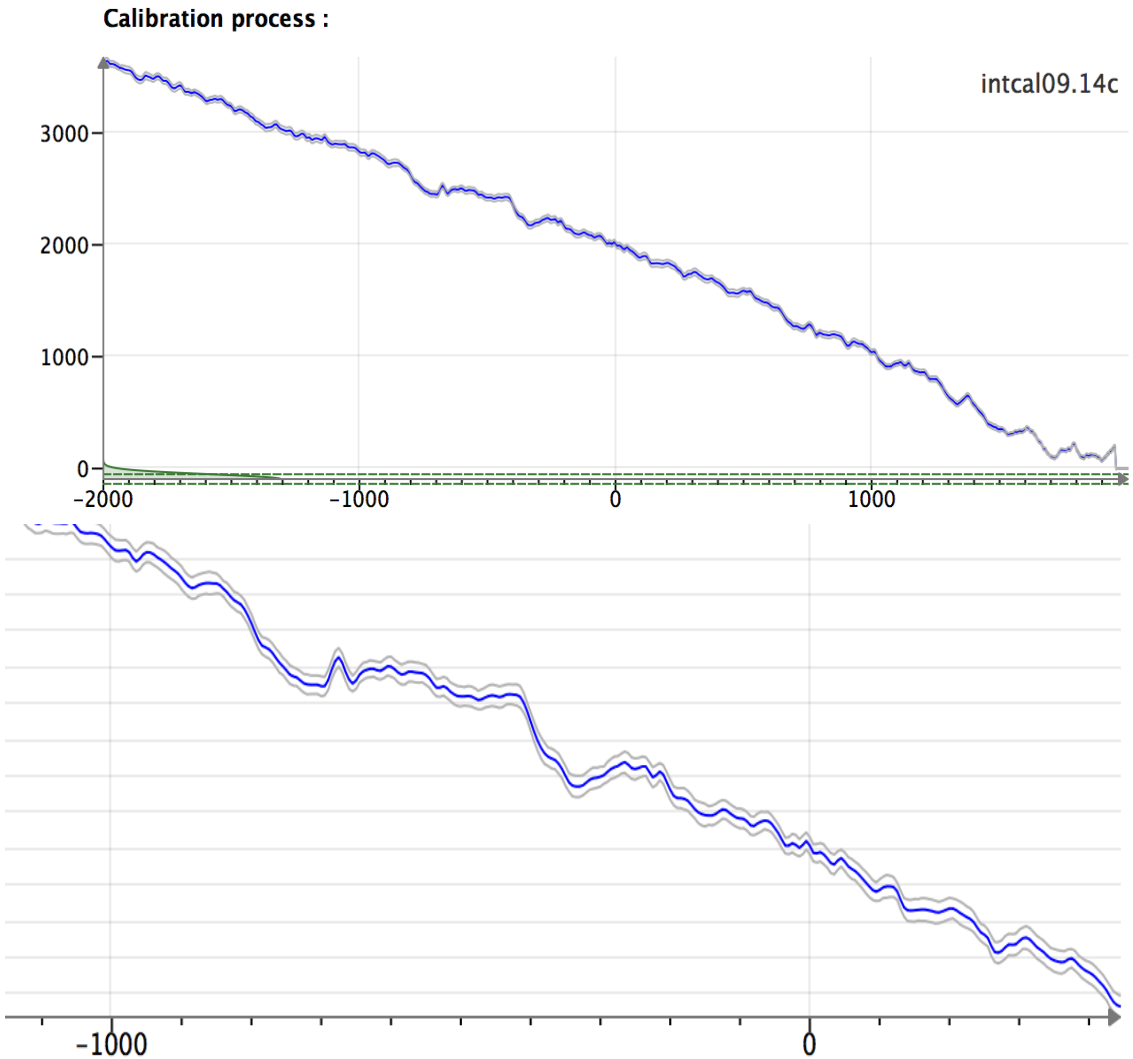


Radiocarbon *IntCal14*

Different calibration curves

1. In radiocarbon :

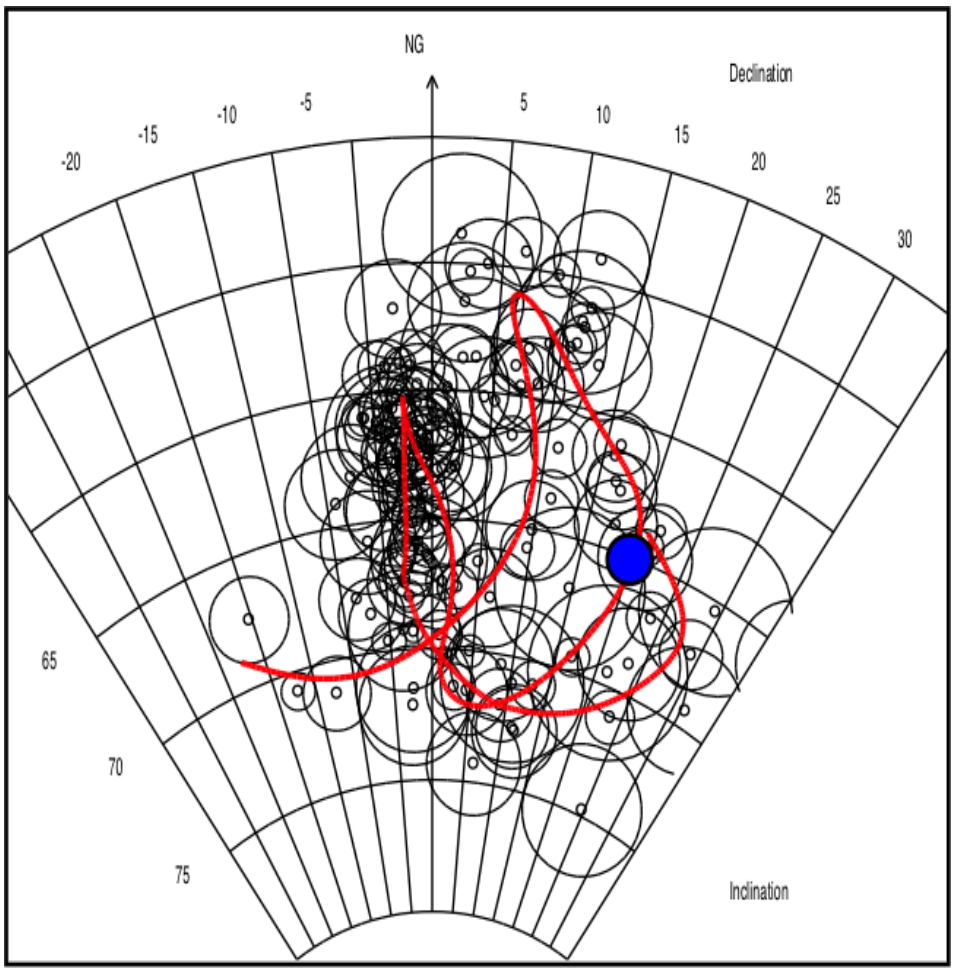
the curve *IntCal14* is used to convert an age measurement into calendar date for continental origin samples.



Different calibration curves

1. In radiocarbon :
2. In archaeomagnetism (AM),

the curve of secular variation of the geomagnetic field established for a given region are used to convert a measurement of inclination, declination or intensity into calendar dates.



Individual calibration

1. the error of measurement.

$$M = m + \epsilon, \quad \epsilon \sim \mathcal{N}(0, s^2)$$

2. error on the calibration curve

$$m = g(\theta) + \epsilon', \quad \epsilon \sim \mathcal{N}(0, \sigma_g^2(\theta))$$

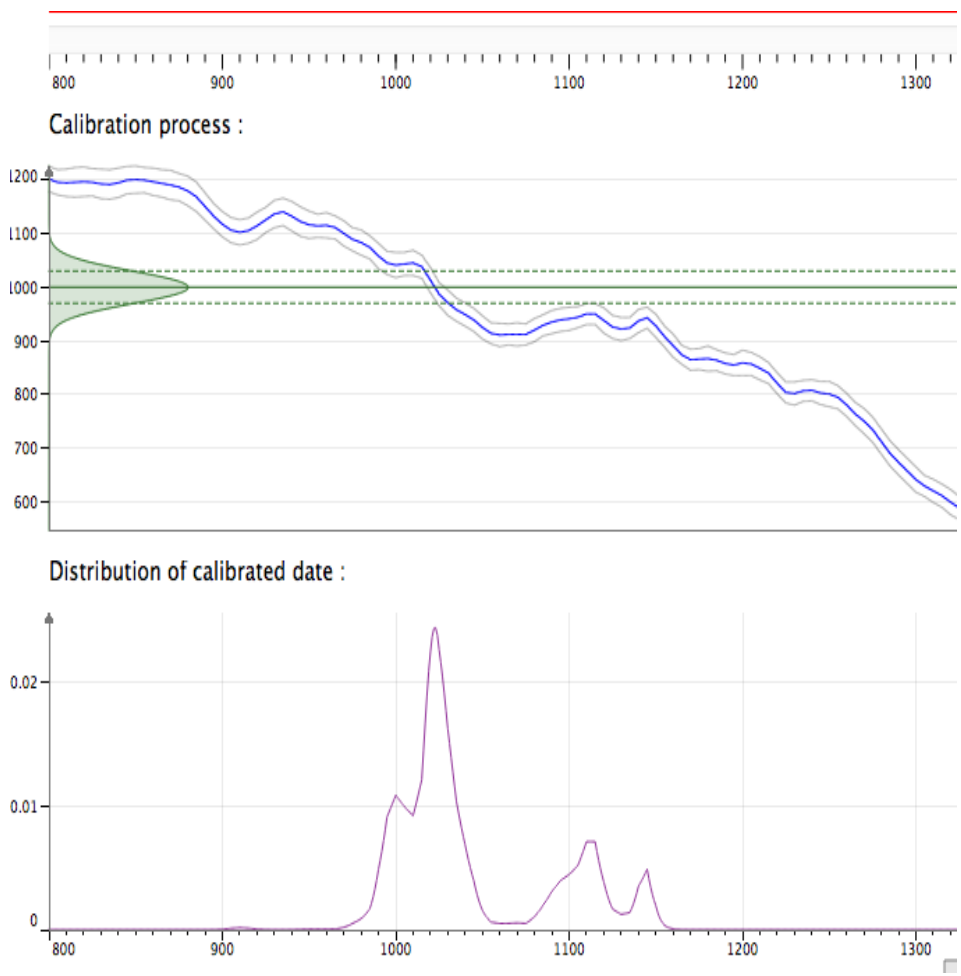
3. Prior dist. $\theta \sim$ Uniform

Posterior distribution :

$$p(\theta|M) \propto \frac{1}{S} \exp\left(\frac{-1}{2S^2}(M - g(\theta))^2\right) 1_T(\theta)$$

where

$$S^2 = s^2 + \sigma_g^2(\theta)$$



Converting an age 14C (= 1000 $\pm$ 30)

Archaeological information

After the archaeological excavations, prior information is available on the dates.

Examples :

- Dated archaeological artefacts are contemporary
- Stratigraphic Information which induces an order on the dates.
- the differences between two dates is known (possibly with an uncertainty).
- *Terminus Post Quem/ Terminus Ante Quem*
- etc

Solution logiciel pour la modélisation

Logiciel :

1. `BCal` is an on-line Bayesian radiocarbon calibration tool.
2. `Oxcal` provides radiocarbon calibration and analysis of archaeological and environmental chronological information.
3. `Chronomodel` a robust Bayesian tool for chronology building.

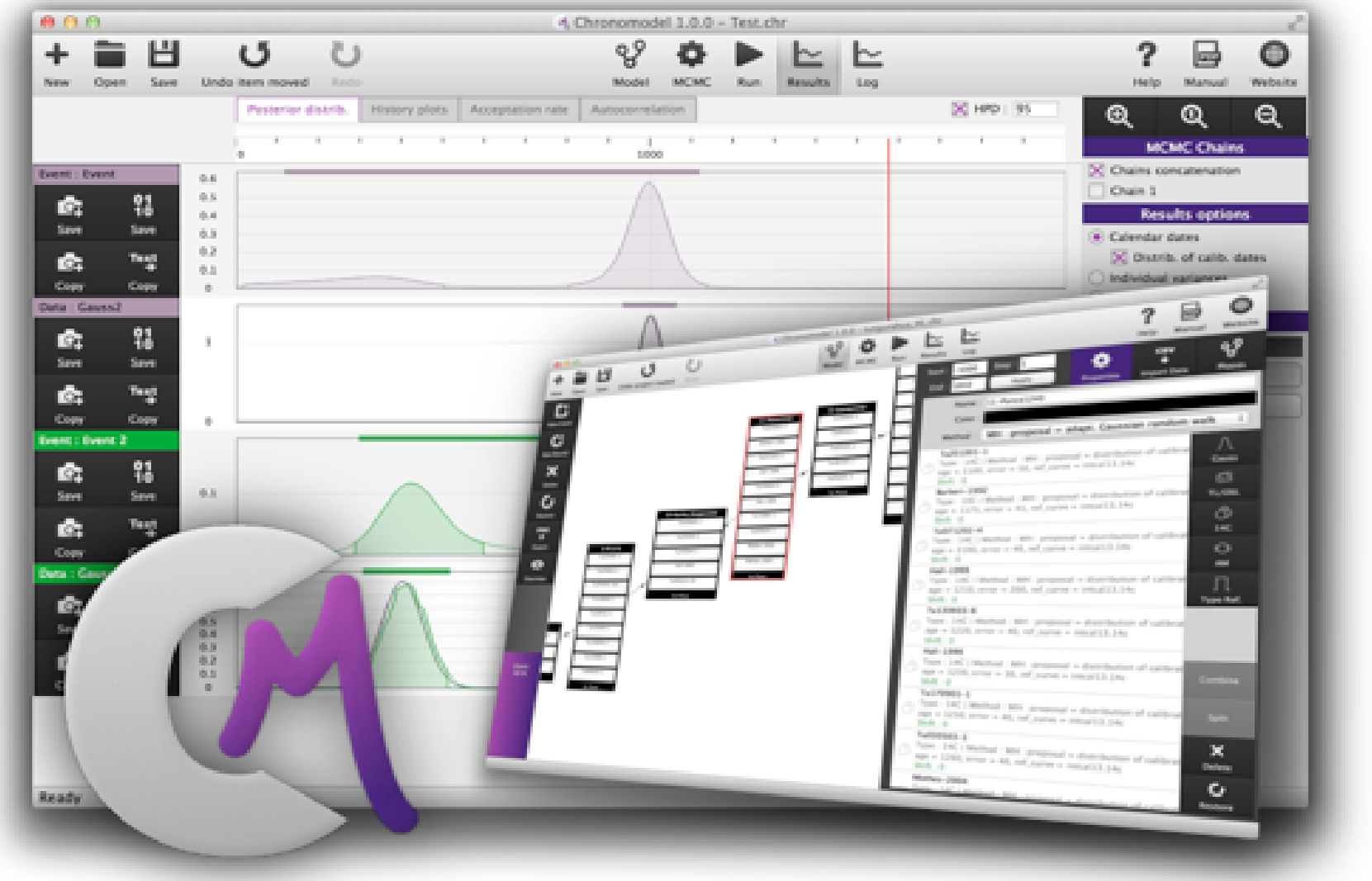
R software

1. `ArchaeoPhases` Post-Processing of the Markov Chain Simulated by 'ChronoModel', 'Oxcal' or 'BCal'
2. `BayLum`. Chronological Bayesian Models Integrating Optically Stimulated Luminescence and Radiocarbon Age Dating
3. `ArchaeoChron` Bayesian Modeling of Archaeological Chronologies
4. `Luminescence` Comprehensive Luminescence Dating Data Analysis
5. `rbacon` age-modelling ;
6. `Bchron` Radiocarbon Dating, Age-Depth Modelling

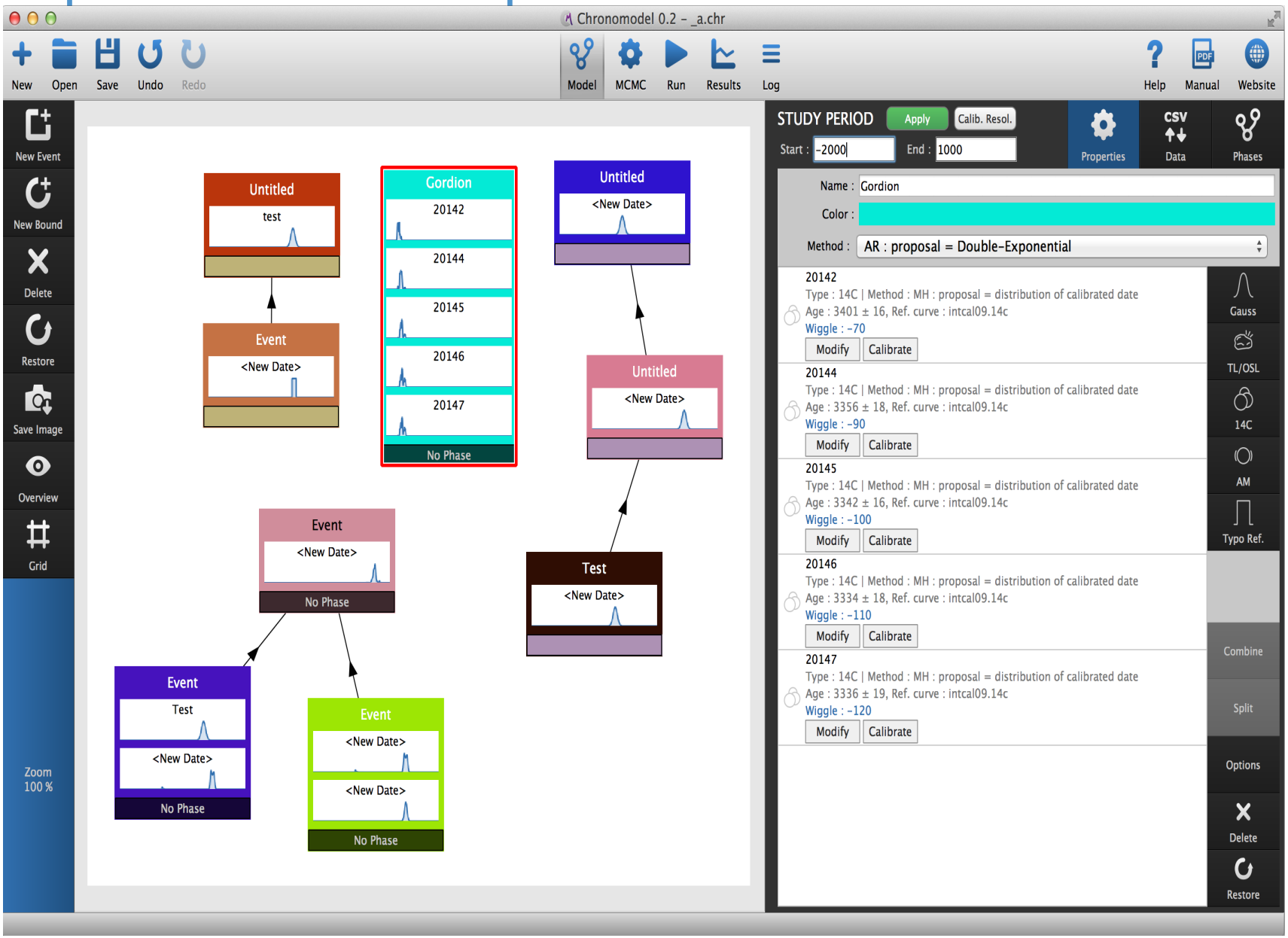
Plan

ChronoModel

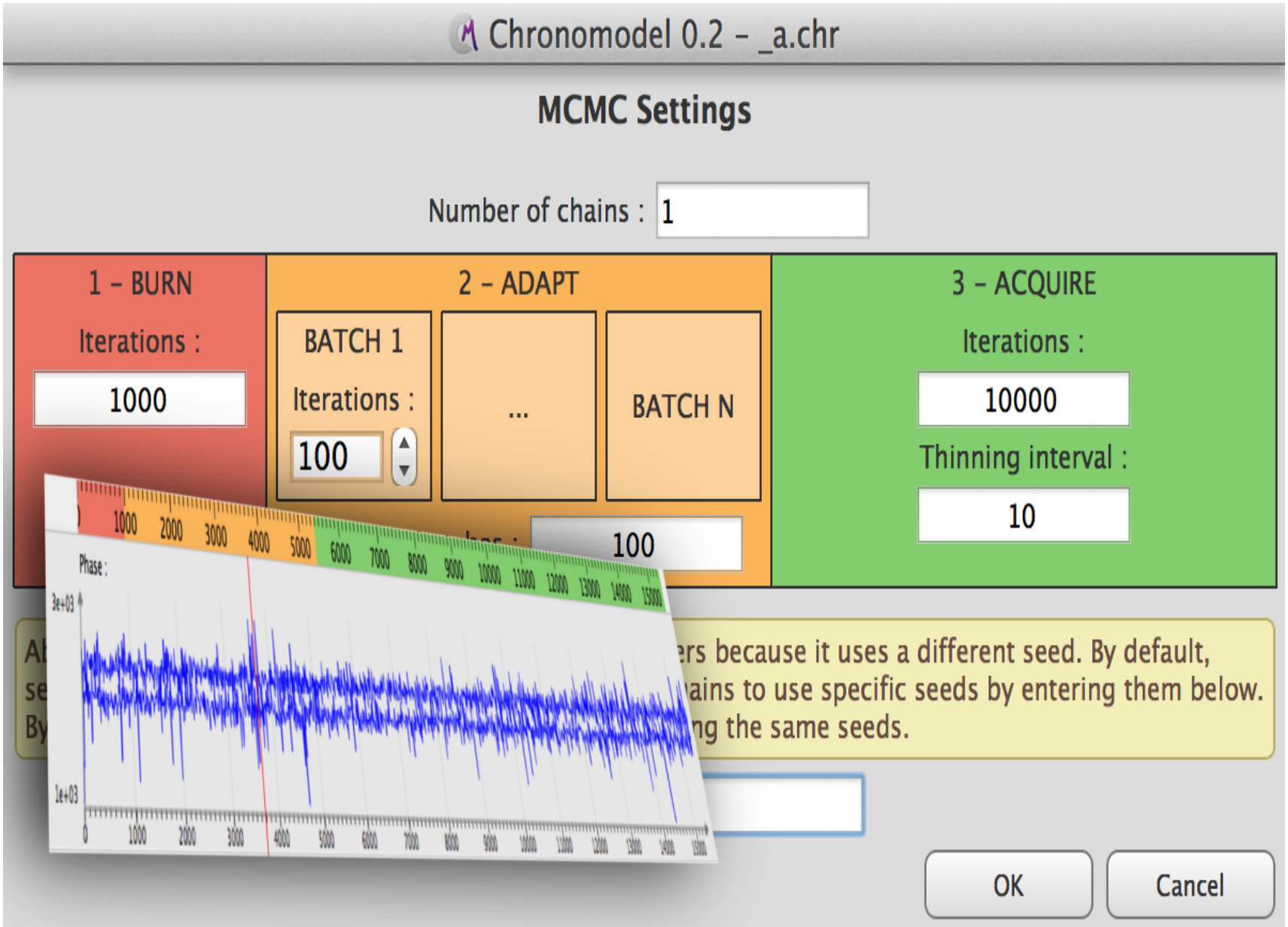
P. Lanos (Archéologue) & A. Philippe



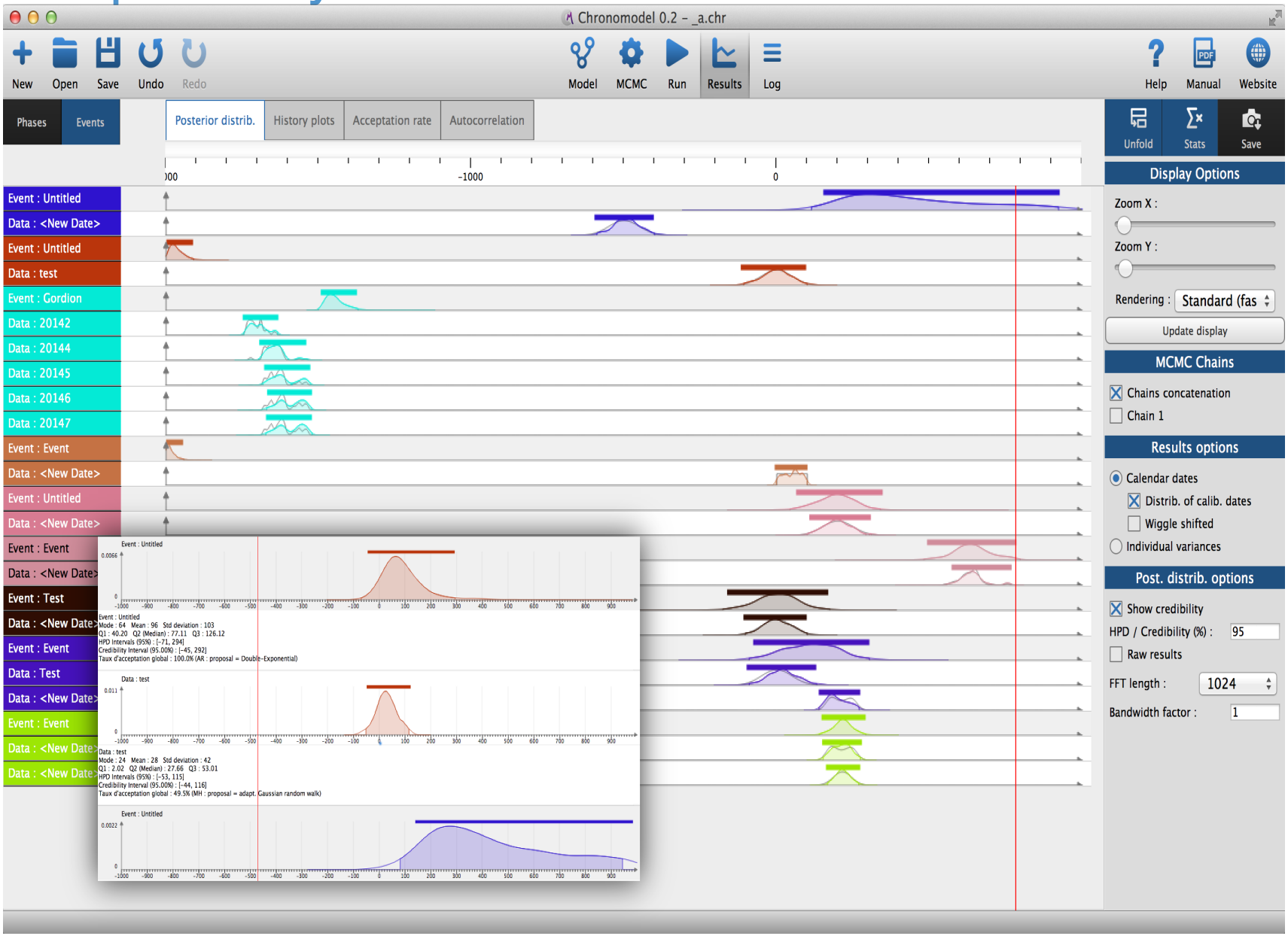
Inputs : datasets - prior information



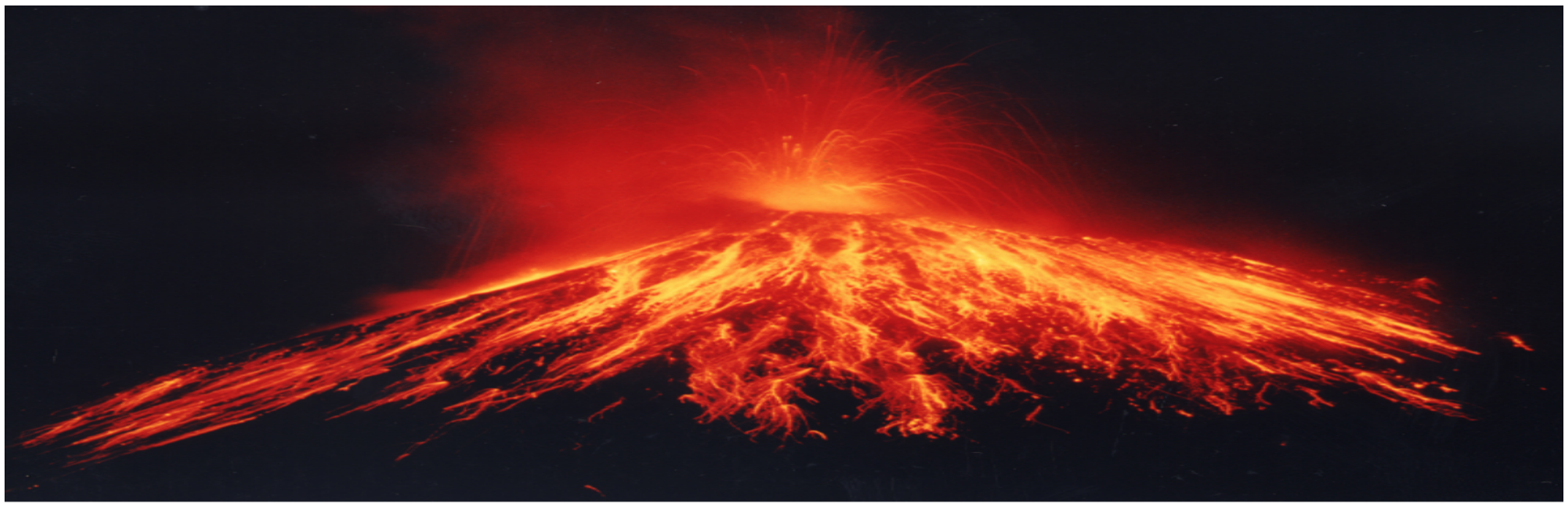
Monte Carlo method



Output : bayesian estimation of the dates



Volcanic eruptions



- Target Event : Eruptive period with flow deposits
- Dated events : organic samples found in a flow deposit are dated by 14C.

Medieval kiln of the potter's workshop in Lezoux

[1] Menessier-Jouannet *et al.* 1995

1. **Aim** : Dating the last firing of the kiln. This is any date between 0 and 2 000
2. **dated events** :
   - baked clays dated by
     AM $>$ *Estimation of the last time the temperature exceeded a critical point*
     TL $>$ *Estimation of the last firing*
   - bones
     14C $>$ *Estimation of the death of the animal*

Definition of the target Event

- we choose a group of dated events that are related the target event.

$\rightsquigarrow$ Characterize the date of a target event from the combination of the dates of contemporaneous dates.

1. n measurements : $M_1, ..., M_n$
   - For each $i = 1, ..., n$ the measurement M_i is done on material whose calendar date t_i is unknown.

$$M_i = g_i(t_i) + \epsilon_i$$

   - ϵ_i represents the experimental and calibration error
2. Relation with the date of the target event
   - For each $i = 1, ..., n$

$$t_i = \theta + \lambda_i$$

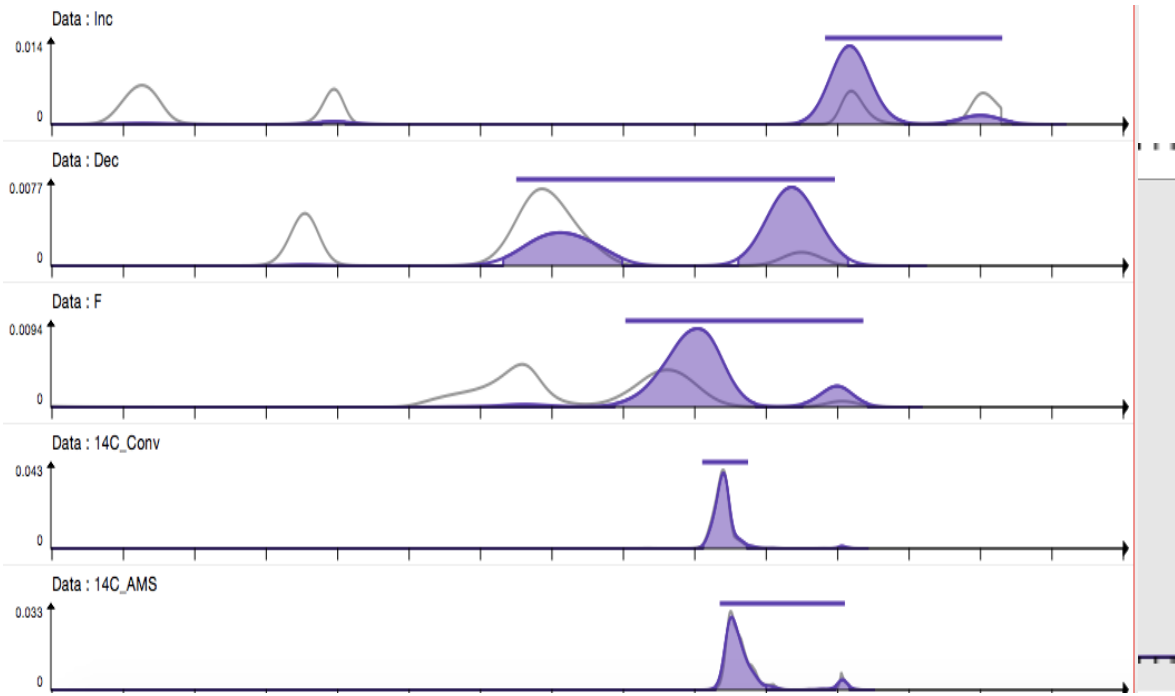
   - λ_i represents the difference between the date of artifacts t_i and the target event θ
3. The prior information is

   the date of the target event belongs to $T = [T_b; T_e]$ *: Uniform distribution*

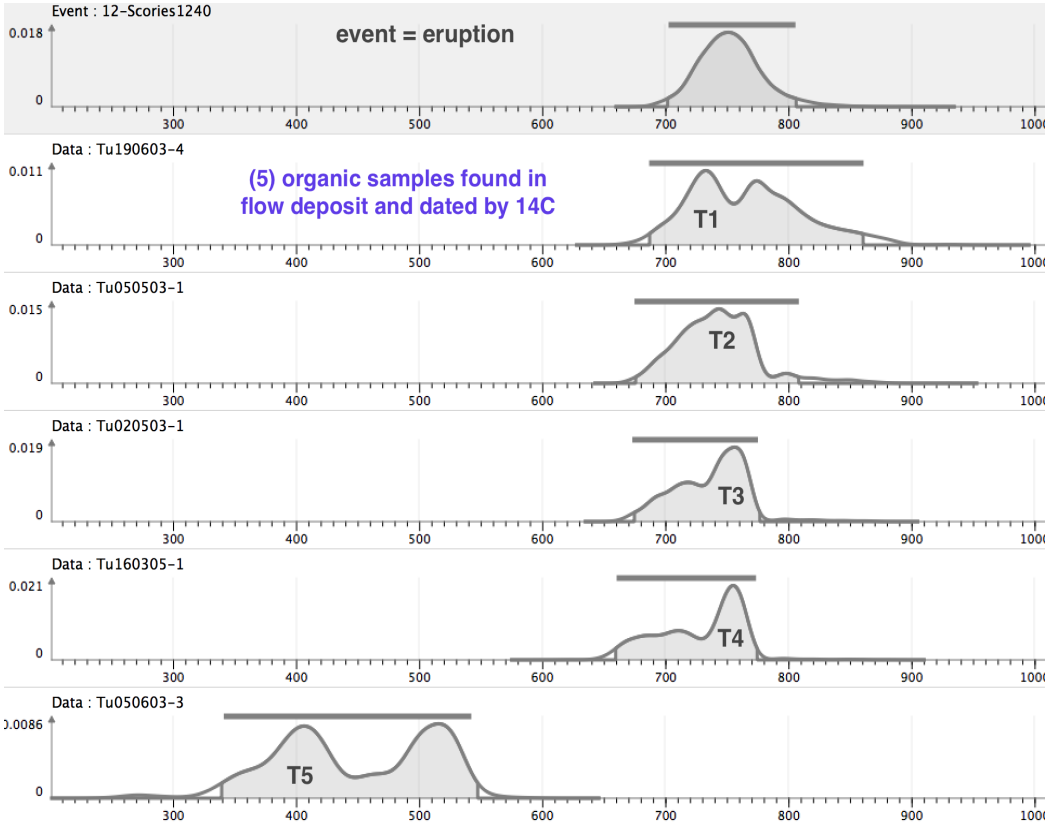
Medieval kiln of the potter's workshop in Lezoux

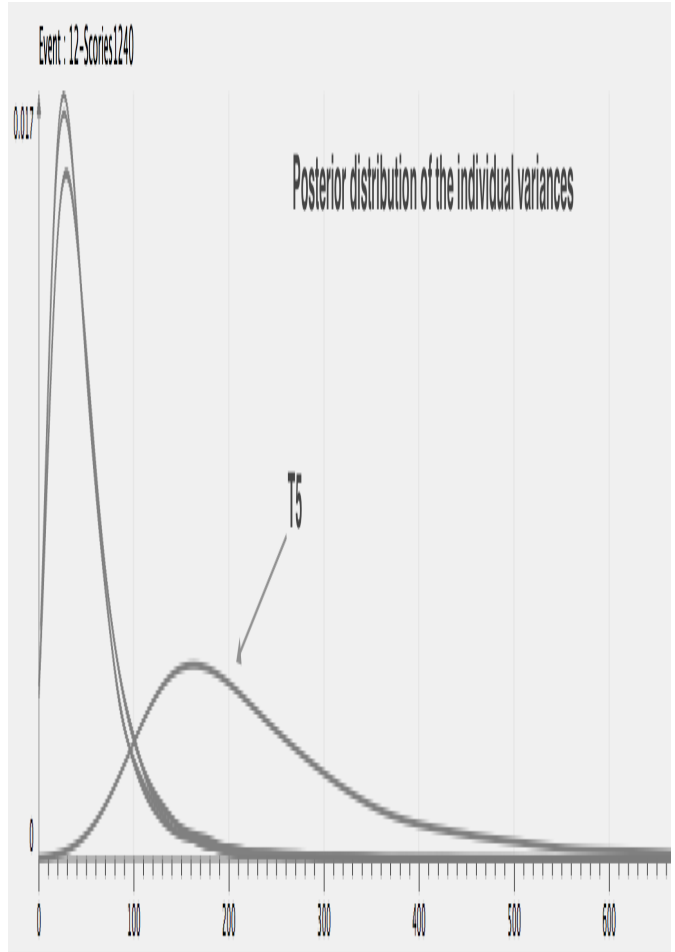
- **Measurements**
  (AM) Inclination : I = 69.2
  (AM) Declination : I = 69.2
  (TL) age 1170 +/- 140 years
  (TL) age 1280 +/- 170 years
  (14C) age 1370 +/- 50 BP

- **Prior information** We assume that the study period is [0 ; 2 000]



One pyroclastic flow : Detection of outlier





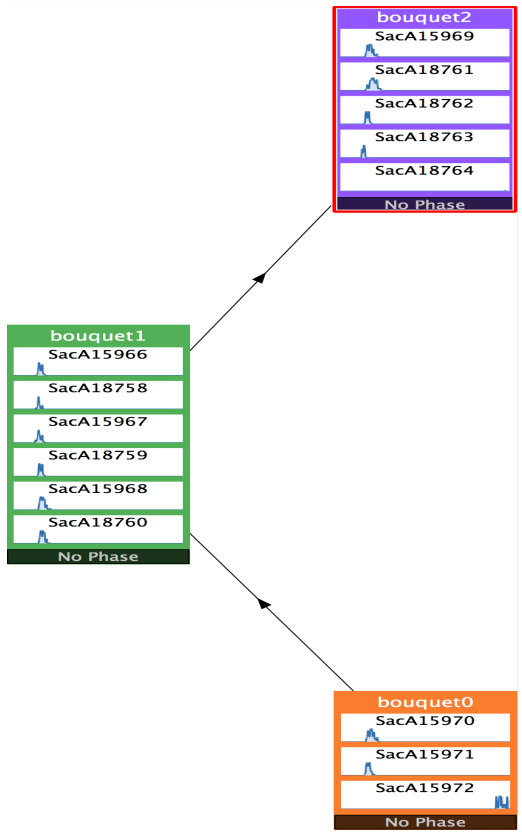
- the posterior density of date of the target Event remains almost insensitive to the outlier.
- In progress with Jean Michel Galharret (PhD student) : construction of tool for rejecting outlying data.

Sequence of target event

We consider Bayesian tools for constructing chronological scenarios.

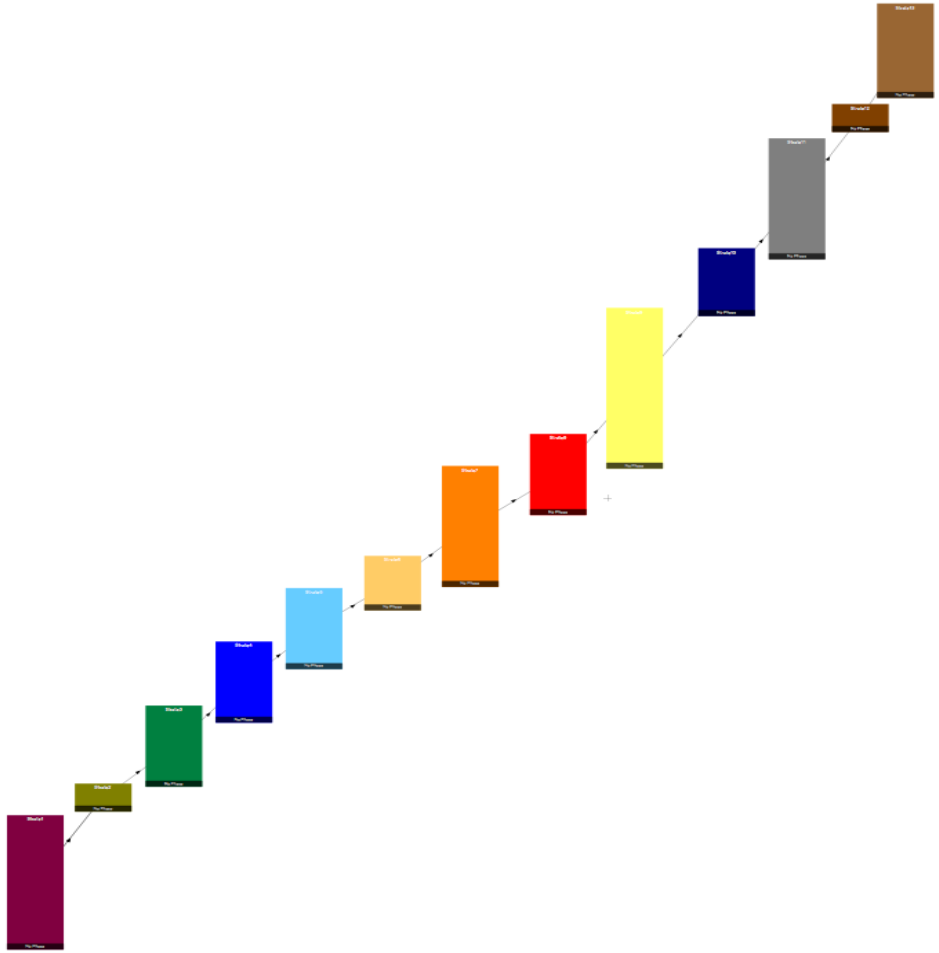
Main idea of the model implemented in `Chronemodel`

1. we define target event as a group of contemporaneous dated events.
2. We construct a chronology (= collection of dates) of target events taking into account temporal relationship between the dates of target events



Volcanic eruptions

- Target Event : Eruptive period with flow deposits
- Dated artefacts : organic samples found in a flow deposit are dated by 14C.
- Prior information Stratigraphic constraint on deposits



Restrictions

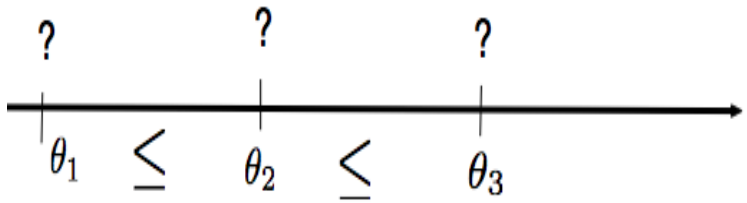
- Each event contains at least one measurrement.
- Each measurement is associated to one (and only one) target event.

Prior information on the dates of the target event

We want to estimate $\theta_1, ...\theta_K$ the calendar dates of target events.

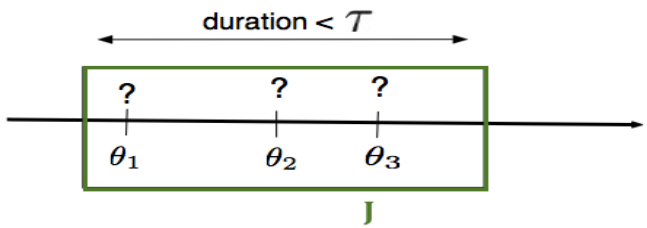
1. The stratigraphic constraints.

$\rightsquigarrow$ a partial order on $(\theta_1, ...\theta_K) := \vartheta \subset T^K$



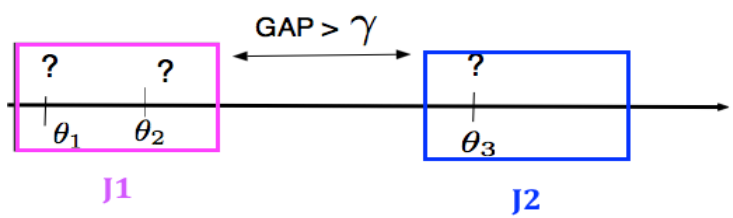
2. Duration information :

$\max_{j\in J} \theta_j - \min_{j\in J} \theta_j \leq \tau$ where τ is known



3. Hiatus information :

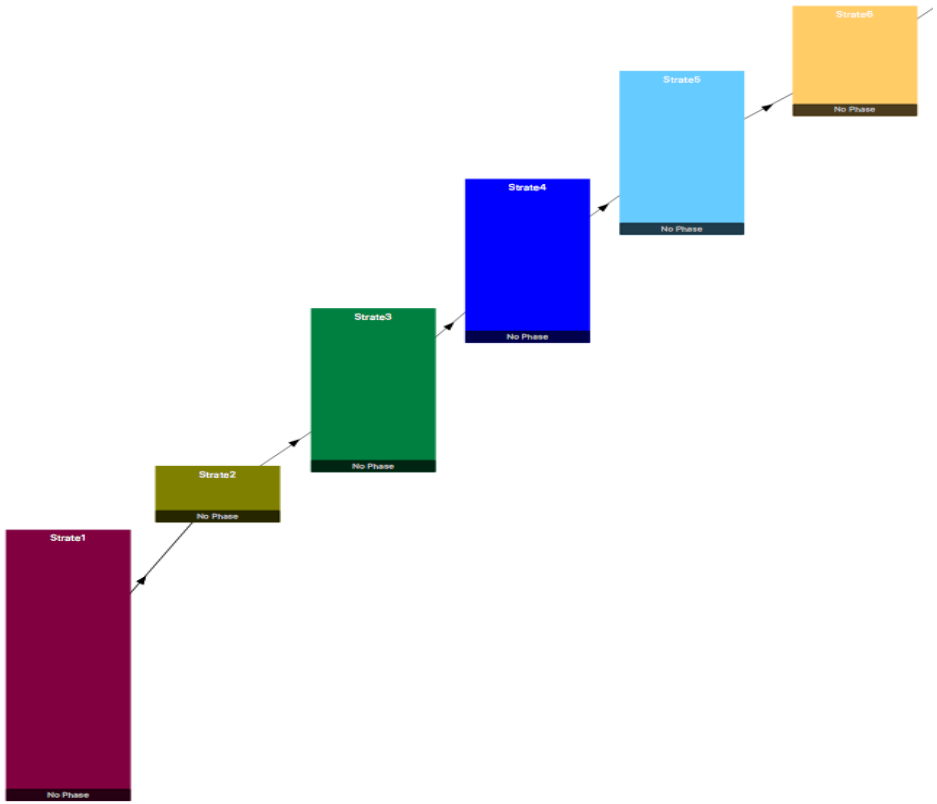
J_1, J_2 two groups, $\min_{j\in J_2} \theta_j - \max_{j\in J_1} \theta_j \geq \gamma$ where γ is known

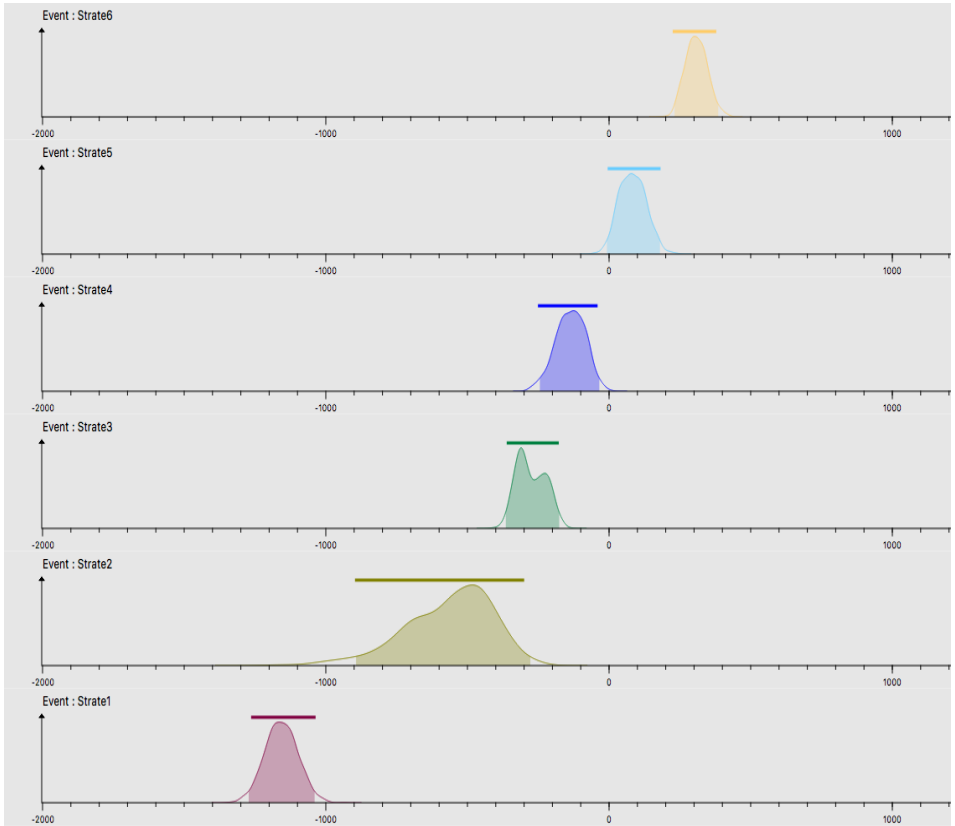


Chronology of Volcanic eruptions

6 pyroclastic flows from volcano dated by 14C $\rightsquigarrow$ 6 ordered target events

$S = \{\vartheta : \ \theta_1 \leq ... \leq \theta_6\}$





Maya city with information on occupation time

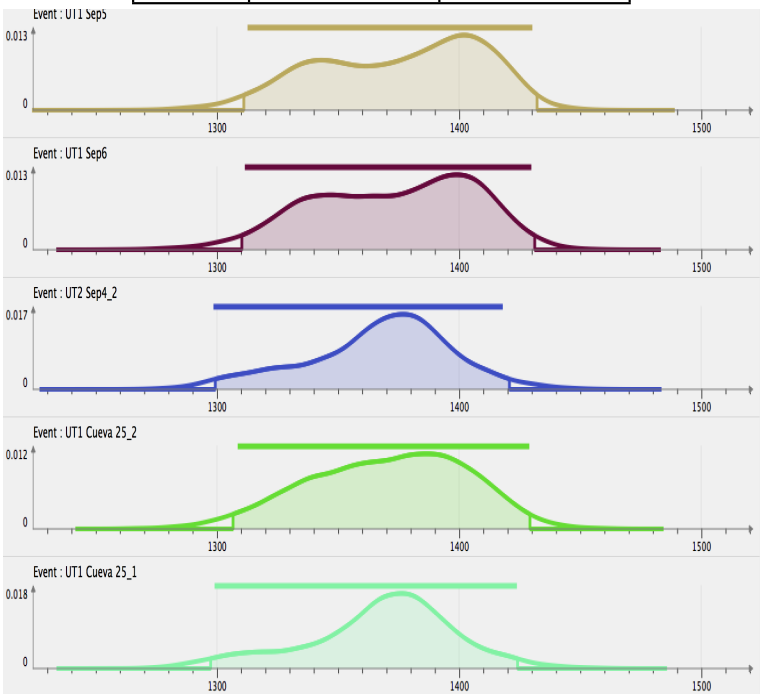
Prior information on the archaeological phase :

The occupation time is smaller than 50 years.

Comparison : HPD regions and posterior densites

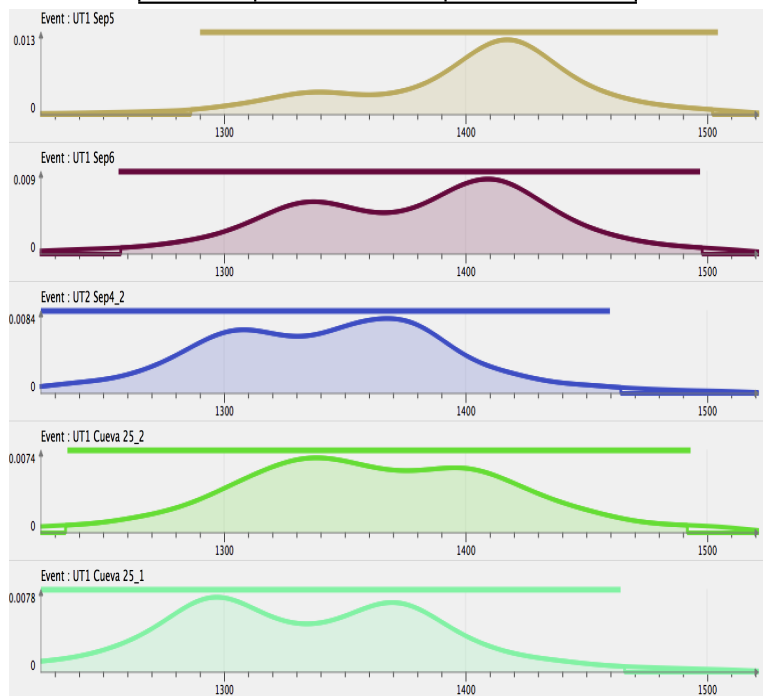
Prior information on the duration

θ_1	1309	1433
θ_2	1308	1430
θ_3	1299	1423
θ_4	1305	1429
θ_5	1297	1425



without prior information

θ_1	1284	1506
θ_2	1253	1502
θ_3	1213	1469
θ_4	1230	1497
θ_5	1192	1469



Plan

Description of the R package ArcheoPhase :

THis R package has its web interface

- Compatible with Oxcal or Chronomodel.
- The inputs are MCMC samples generated by both softwares.

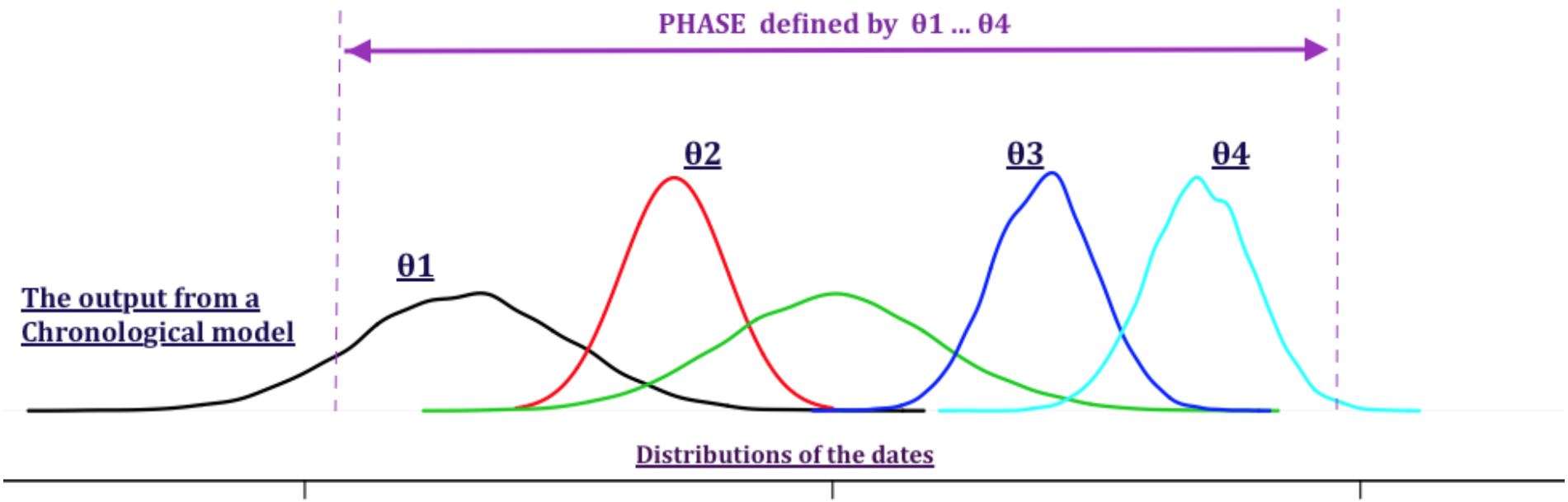
This package contains Statistical Tools for analysis the chronological modelling

Examples

1. Characterisation of a group of dates [begin / end /duration/ period]
2. Testing the presence of hiatus between two dates or two groups of dates.
3. Construction of tempo plot to evaluate the repartition in time
4. Prediction : Age - depth model

Phases : definition

A phase is a group of dates defined on the basis of objective criteria such as archaeological, geological or environmental criteria.



The collection of dates is estimated from a chronological model. [Chronomodel / Oxcal ...]

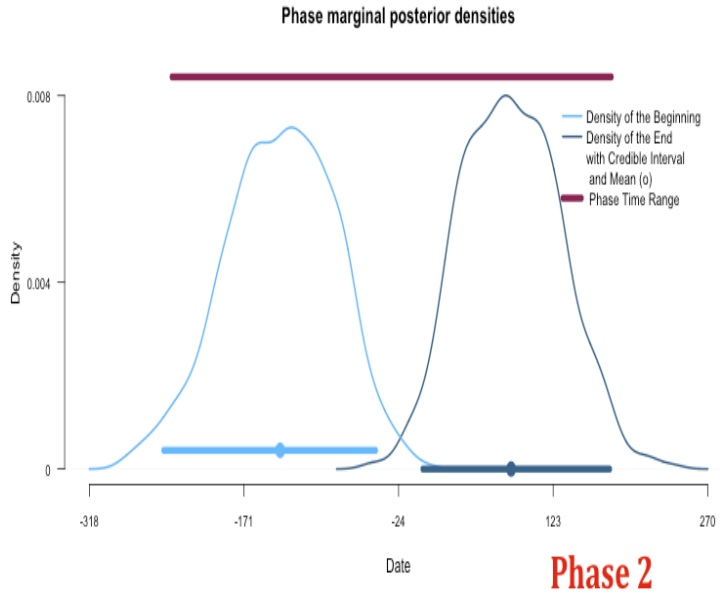
$$\text{Phase} = \{\theta_j,\ j \in J \subset \{1, ..., K\}\}$$

Estimation of the phase

$\text{Phase}_1 = \{\theta_j,\ j \in J \subset \{1, ..., K\}\}$.

1. posterior distribution of the minimum
   $\alpha = \min_{j \in J} \theta_j$
   $\rightsquigarrow$ Estimation of the beginning

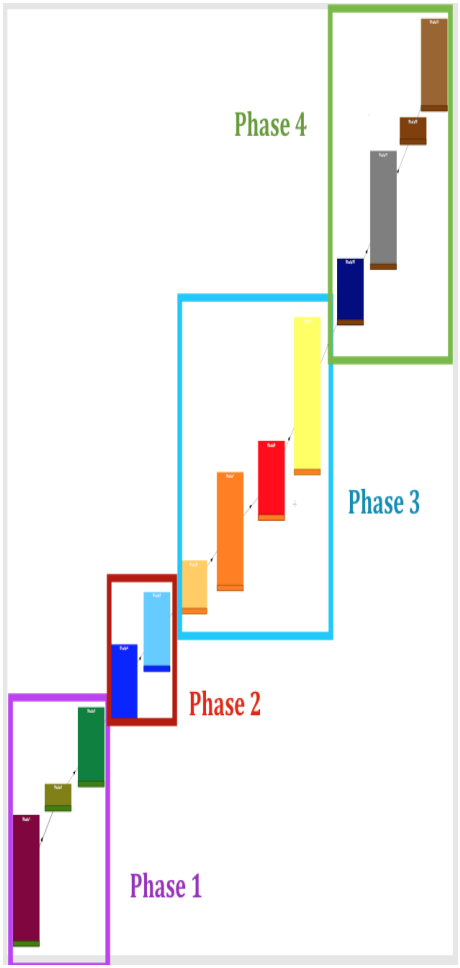
2. posterior distribution of maximum
   $\beta = \max_{j \in J} \theta_j \rightsquigarrow$ Estimation of the end

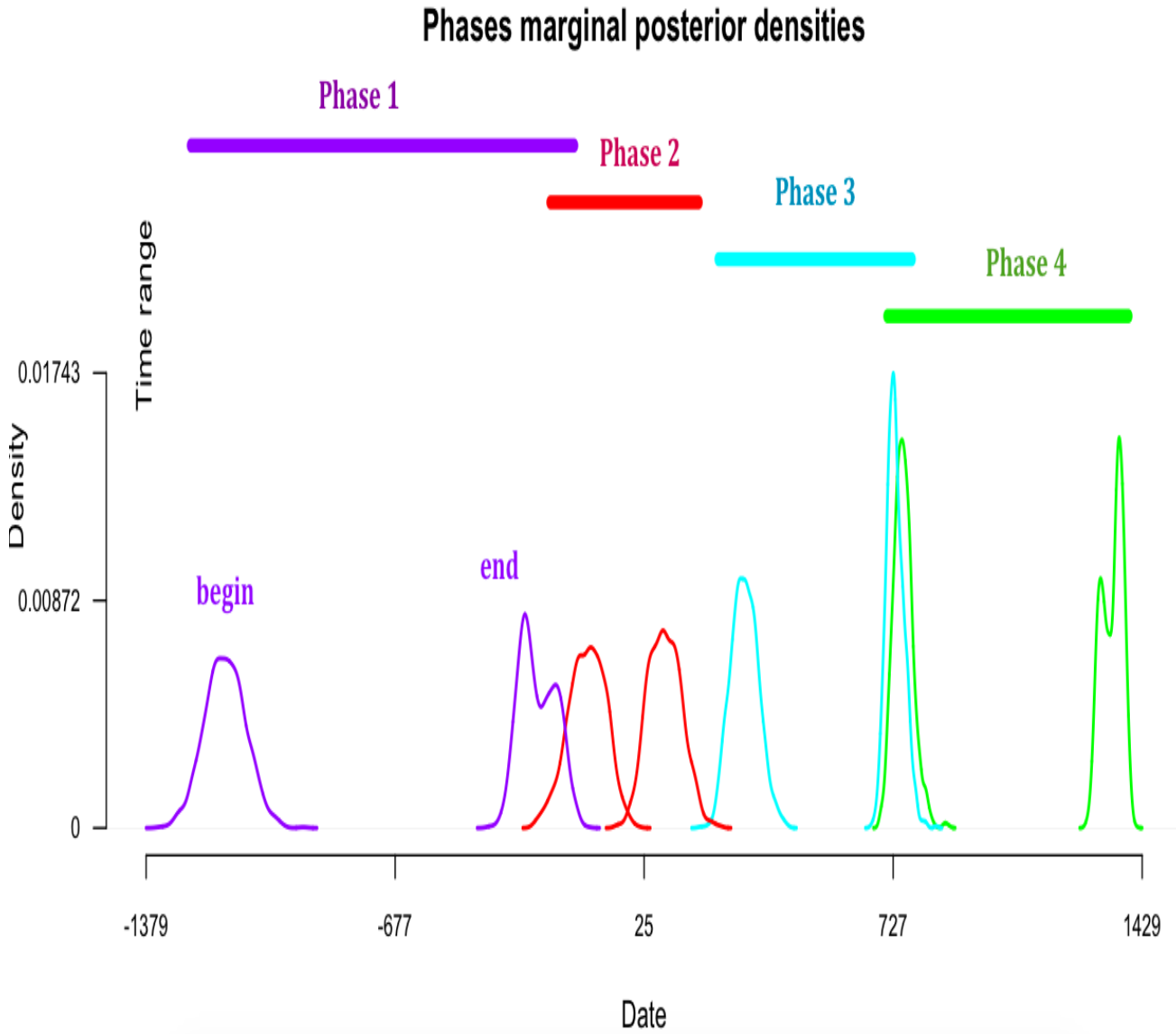


3. Phase time range The shortest interval that covers all the dates θ_j included in the phase at level 95%
   i.e. the shortest interval $[a, b] \subset T$ such that

$$P(\text{for all } j\ \theta_j \in [a, b] | M_1, ..., M_n) = P(a \leq \alpha \leq \beta \leq b | M_1, ..., M_n) = 95\%$$

Application to Volcanic eruptions [cont]

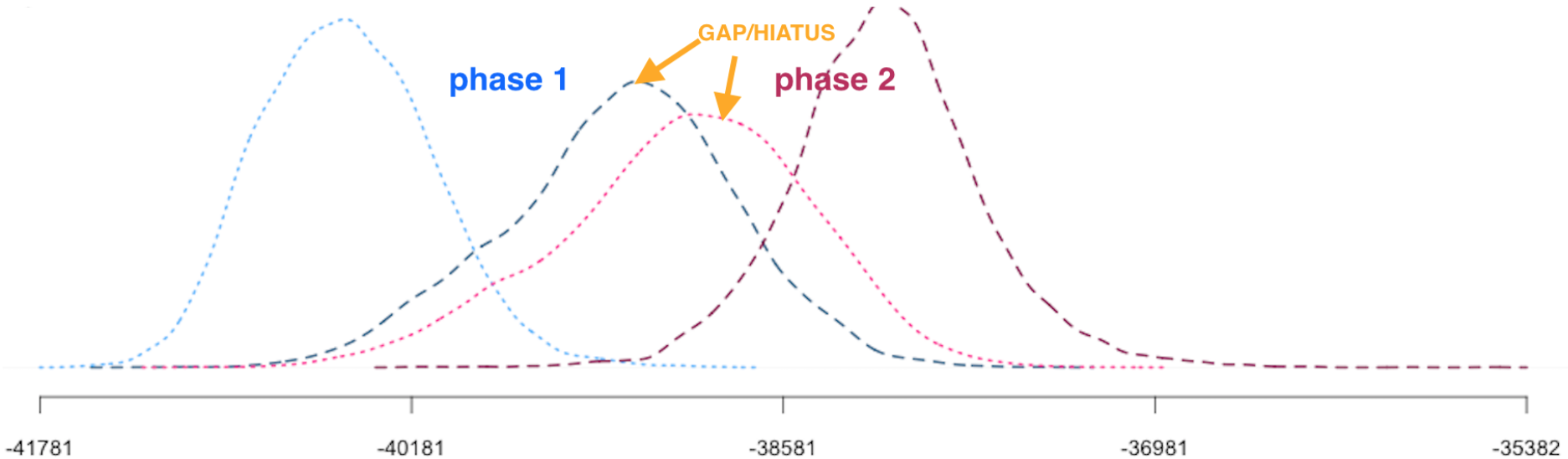




$$P_1 = \{\theta_1, \theta_2, \theta_3\}, \cdots\cdots$$

$$P_4 = \{\theta_{10}, \theta_{11}, \theta_{12}, \theta_{13}\}$$

Hiatus



Detection of a hiatus between two phases $\theta_j,\ j \in J_1$ and $\theta_j,\ j \in J_2$

1. $\beta_1 = \max_{j \in J_1} \theta_j$ and $\alpha_2 = \min_{j \in J_2} \theta_j$
2. Can we find $[c, d]$ such that

$$P(\beta_1 < c < d < \alpha_2 | M_1, ..., M_n) = 95\%?$$

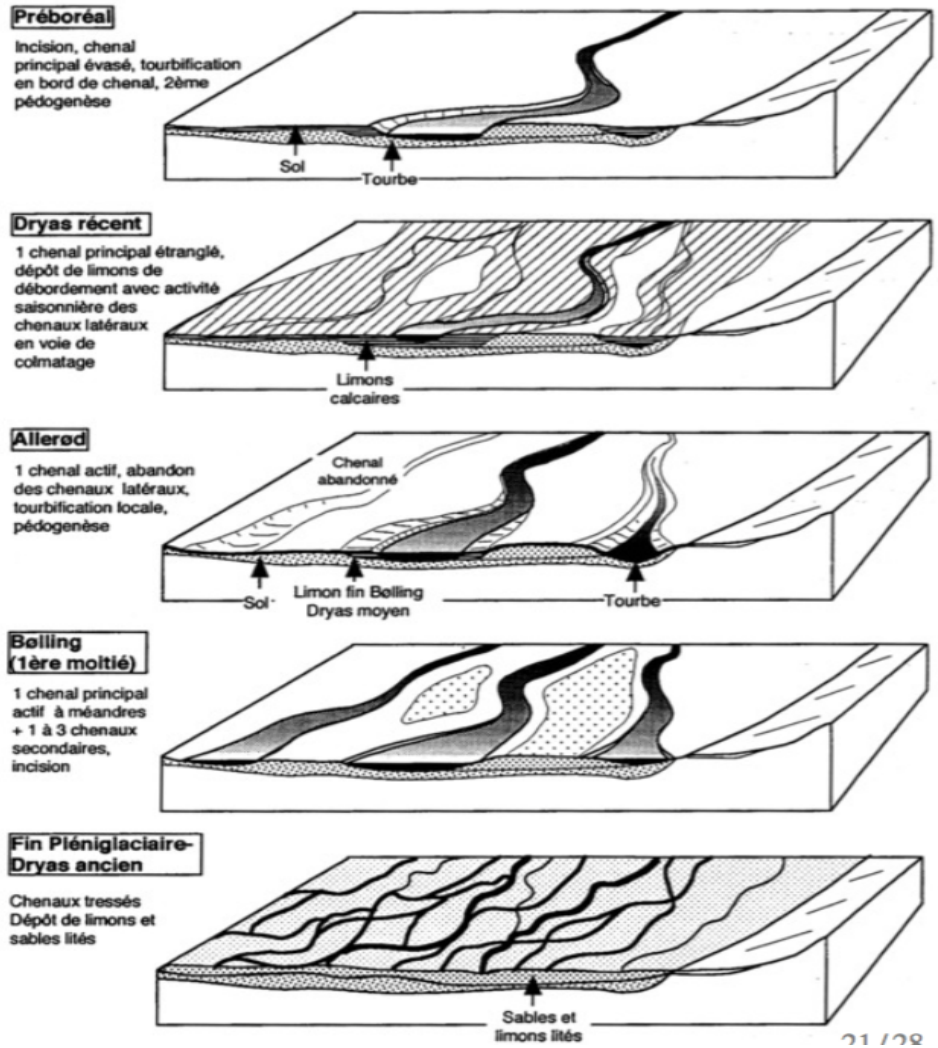
Applications : Palynozones

Lateglacial pollen zones in the Paris basin[4]

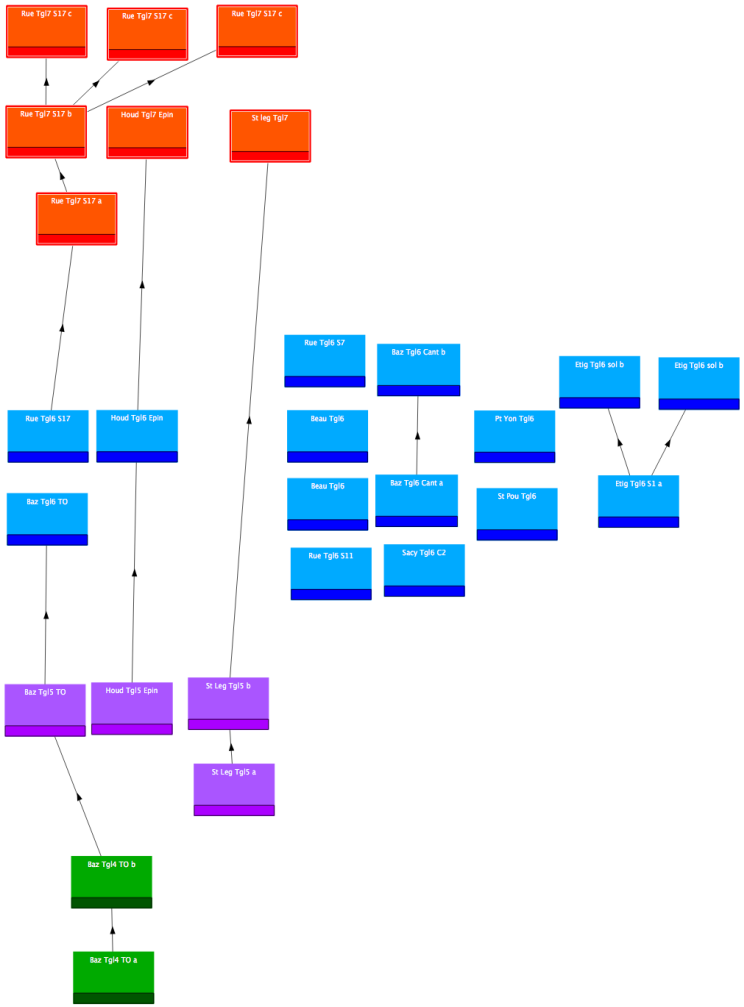
Aim : Defining chronological transitions between 4 phases

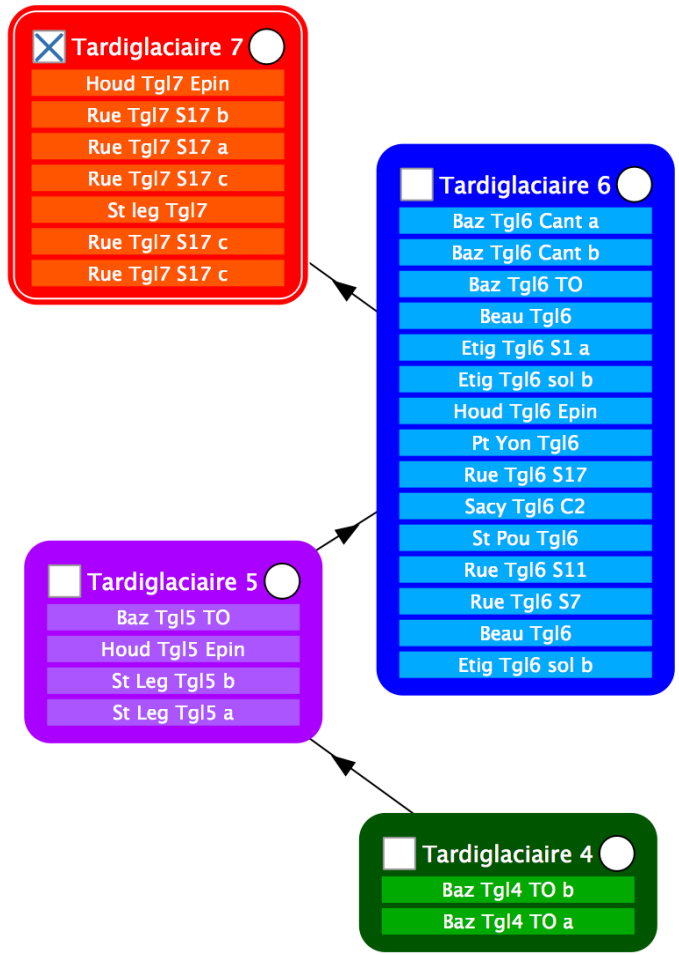
Tgl 7 : the younger Dryas
Tgl 6 : the second part of Allerød
Tgl 5 : the first part of Allerød
Tgl 4 : the older Dryas

[4] Leroyer *et al.* 2011, 2014

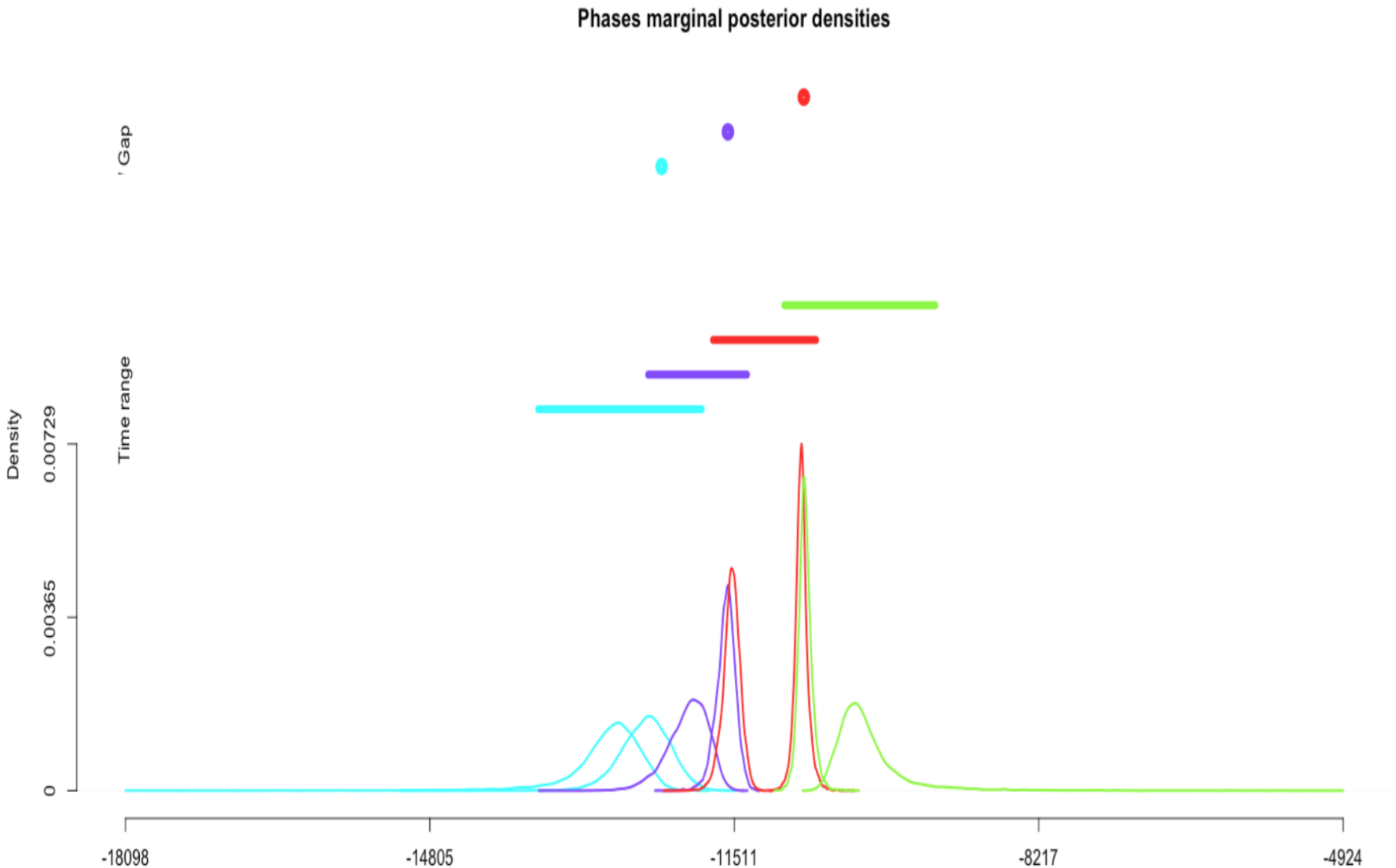


Tardiglaciare





Estimation and testing procedure



The chronology of Canímar Abajo in Cuba

(Rocksandic *et al.* 2015 Philippe & Vibet (2018) RadioCarbon.

The site has evidence for two episodes of burial activity separated by a shell midden layer.

- 12 AMS radiocarbon dates (human bones collagen and a charcoal) obtained from burial contexts
- 7 from the Older Cemetery (OC),
- 5 from the Younger Cemetery (YC))

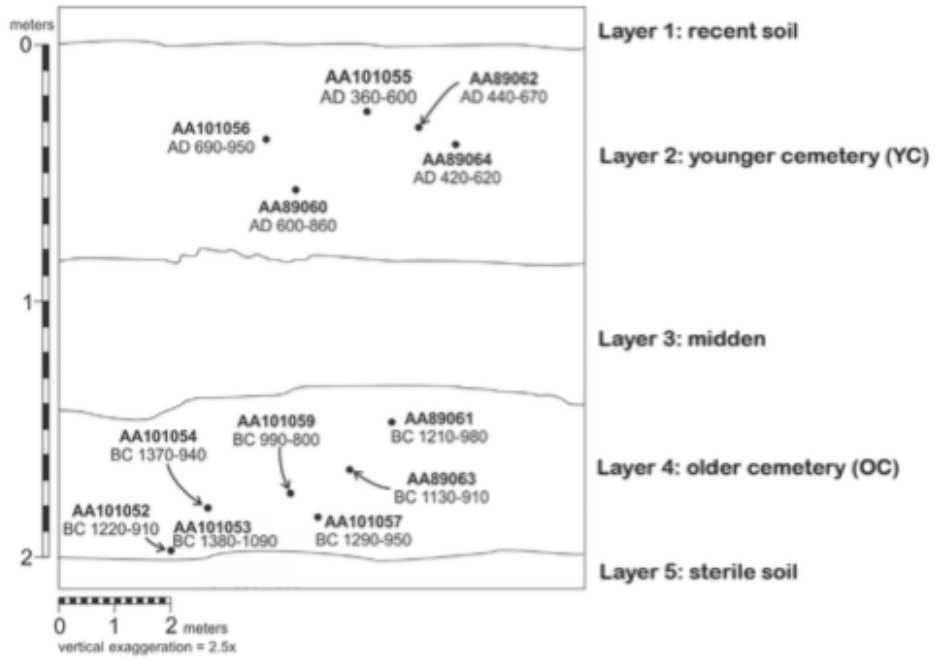
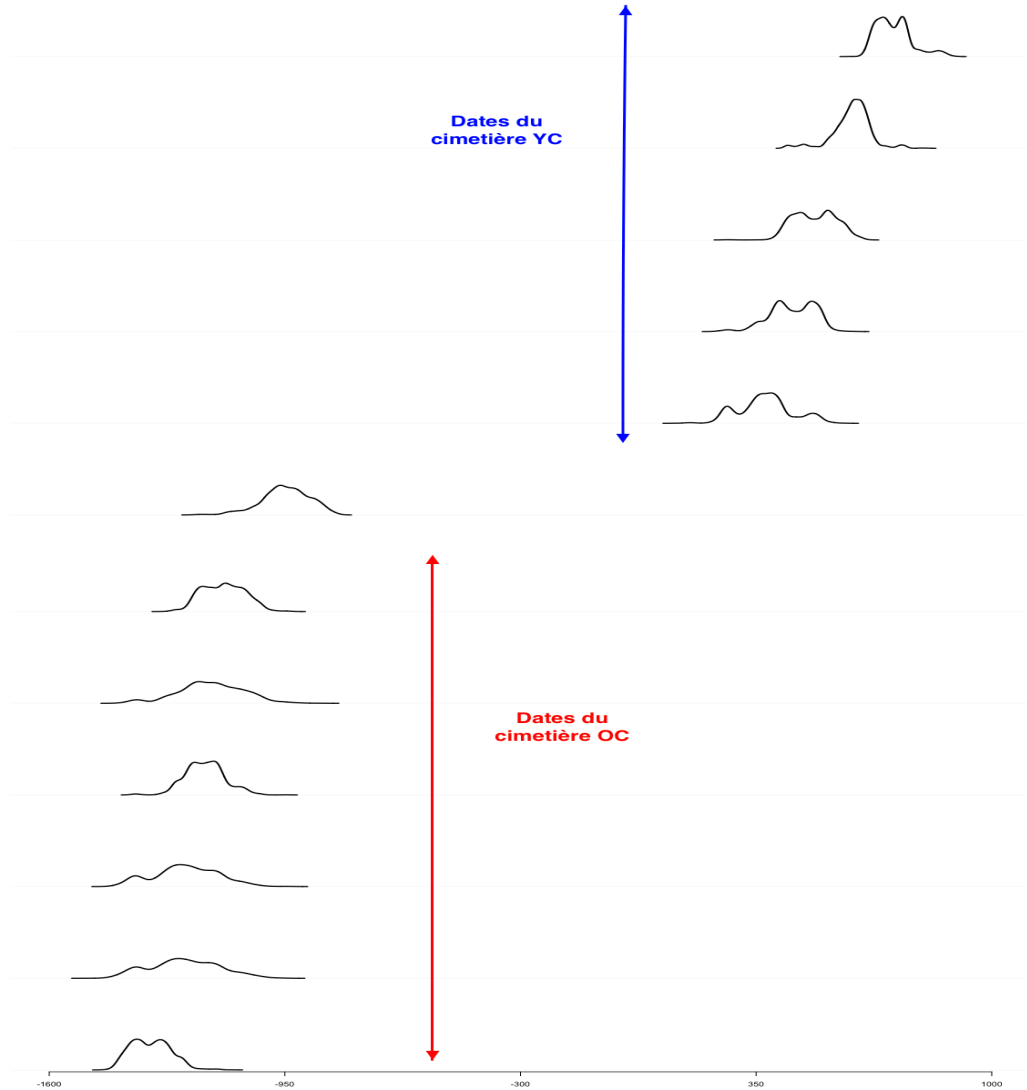


Figure 2 Stratigraphic profile indicating relative positions of samples for AMS ^{14}C dating

The aim : Bayesian model based on these 12 AMS radiocarbon dates in order to draw conclusions about

- the time of both mortuary activities
- the hiatus between them

The chronology

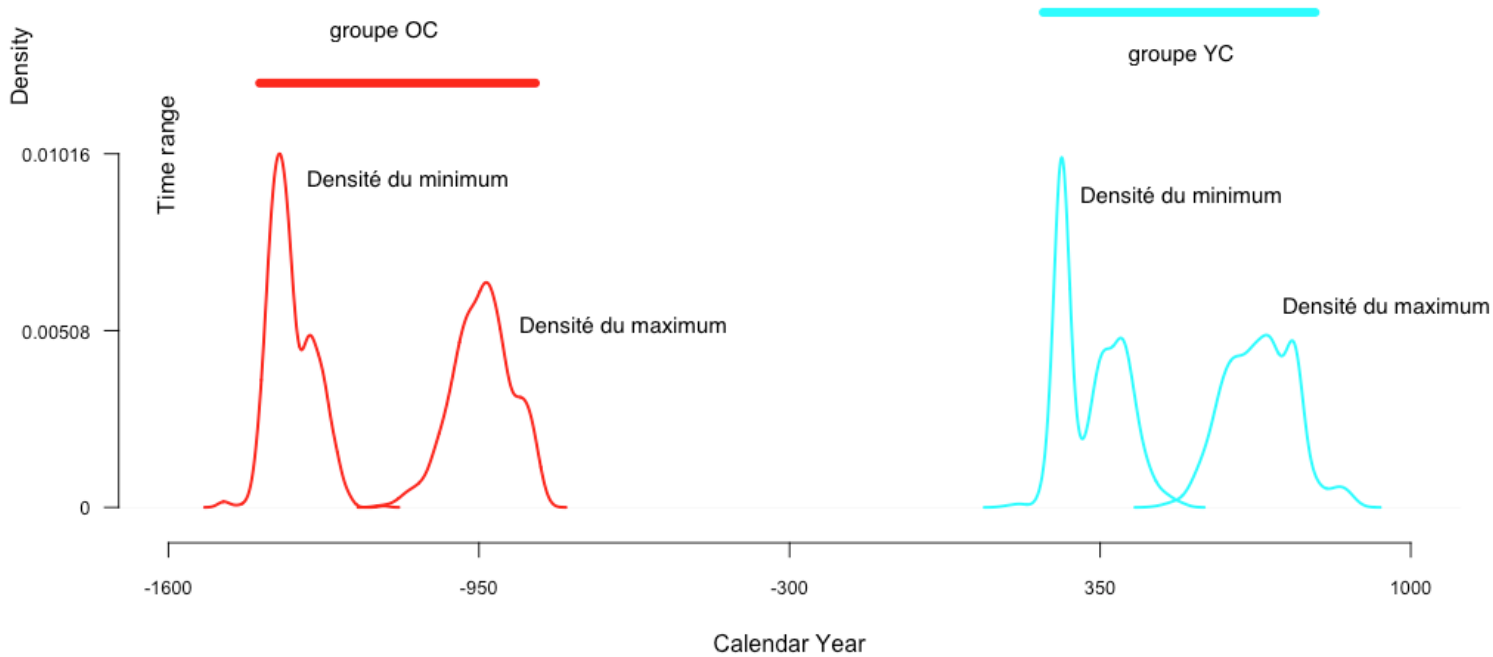


Estimation of the dates t

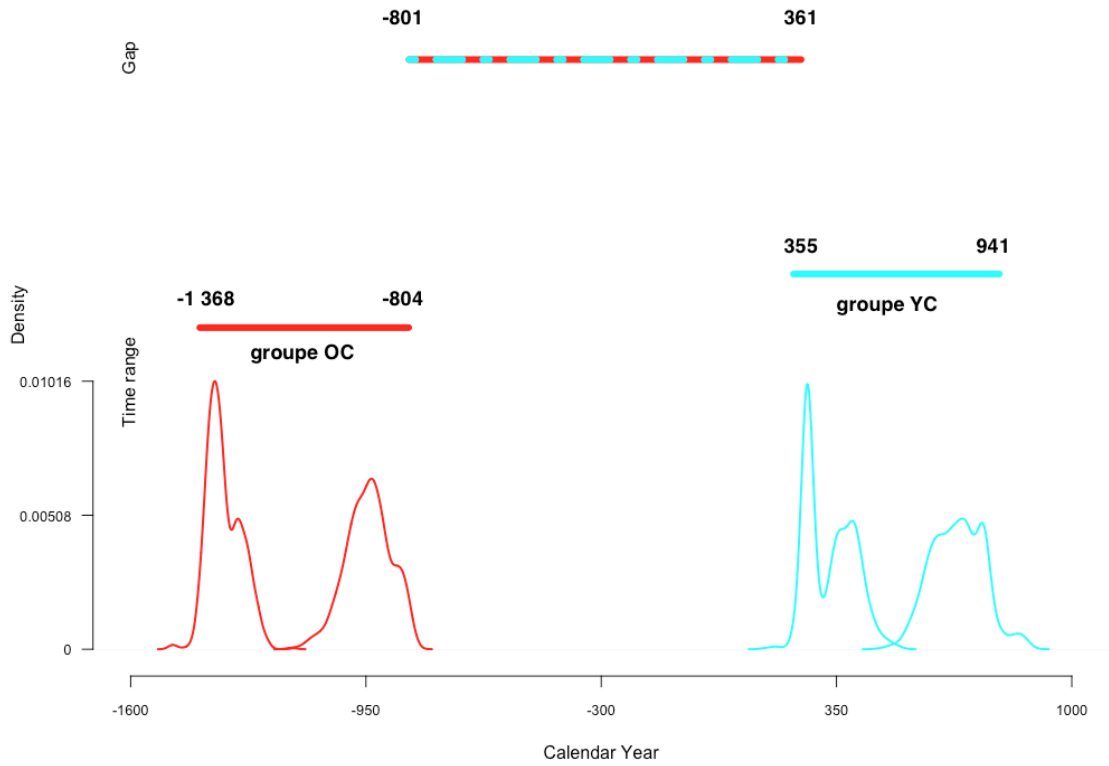
From the estimation of the sequence of dates $t_1, ..., t_{12}$ (using Bayesian model) we estimate

- the beginning and the end of the Older Cemetery
- the beginning and the end of the Younger Cemetery
- the gap between these two periods

Chronology of the activities in the site of Canimar Abajo.



Estimation of the gap



Tempo plot (see Dye 2016 and Philippe & Vibet 2017)

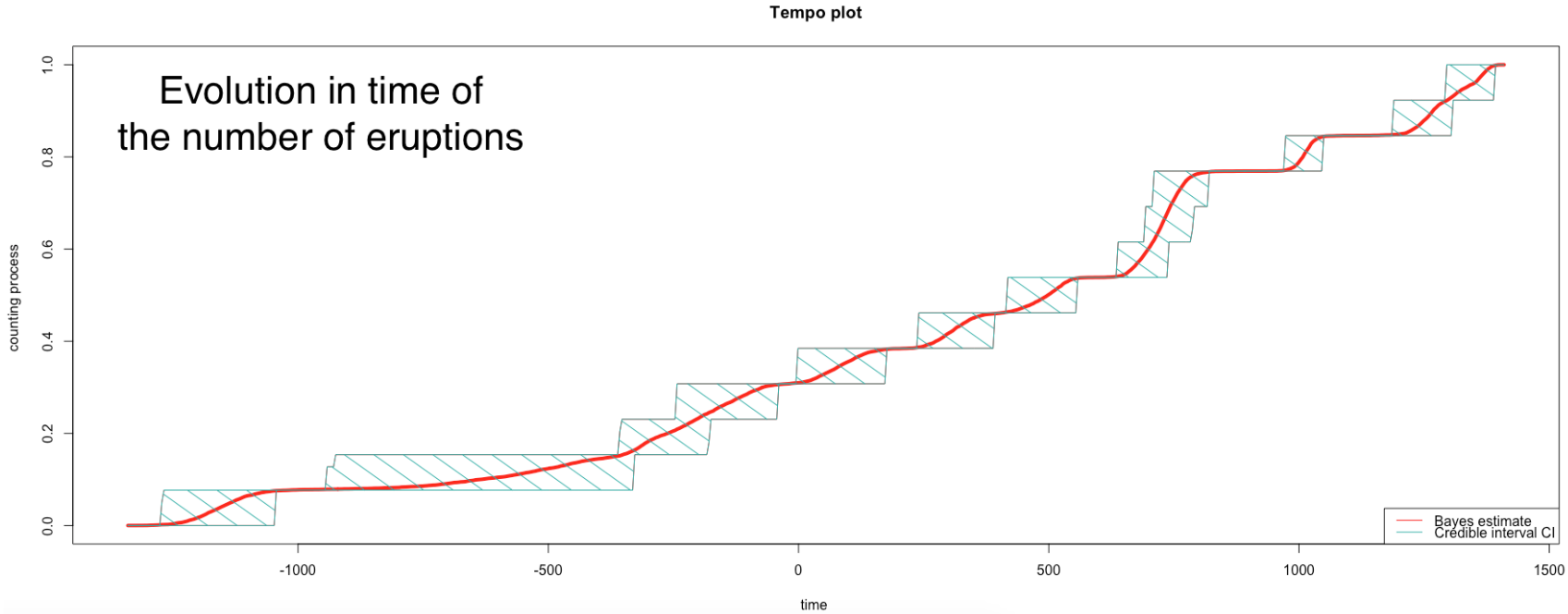
A statistical graphic designed for the study of rhythms.

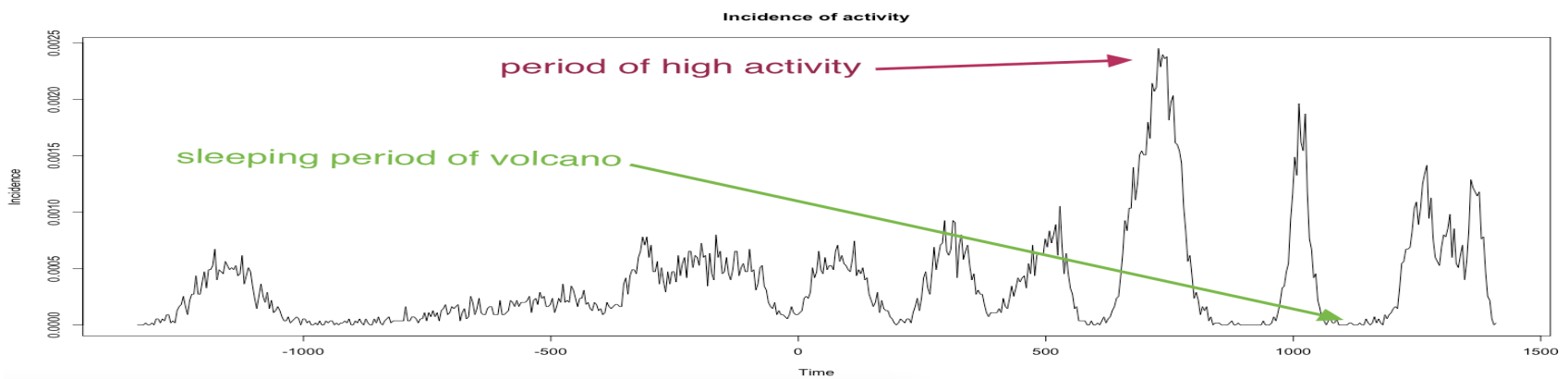
- The tempo plot measures change over time :
- For each date t, we estimate the number of events $N(t)$ which occurs before the date t, we have

$$N(t) = \sum_{i=1}^{n} \mathbb{I}_{]-\infty,t]}(\theta_i)$$

where θ_i, $i = 1, ..., n$ represents the unknown dates of the events.

Application : Evaluation of the activity of volcano





Age-depth model

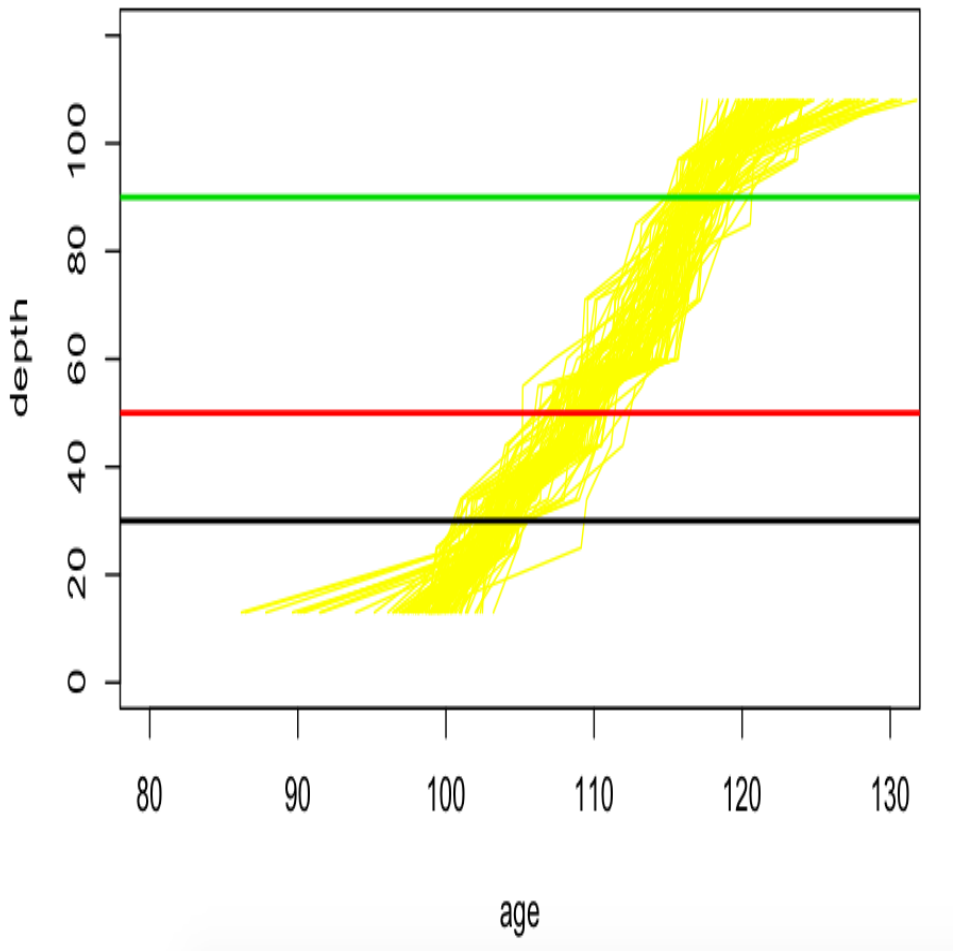
Additional information : the depth of the dated event.

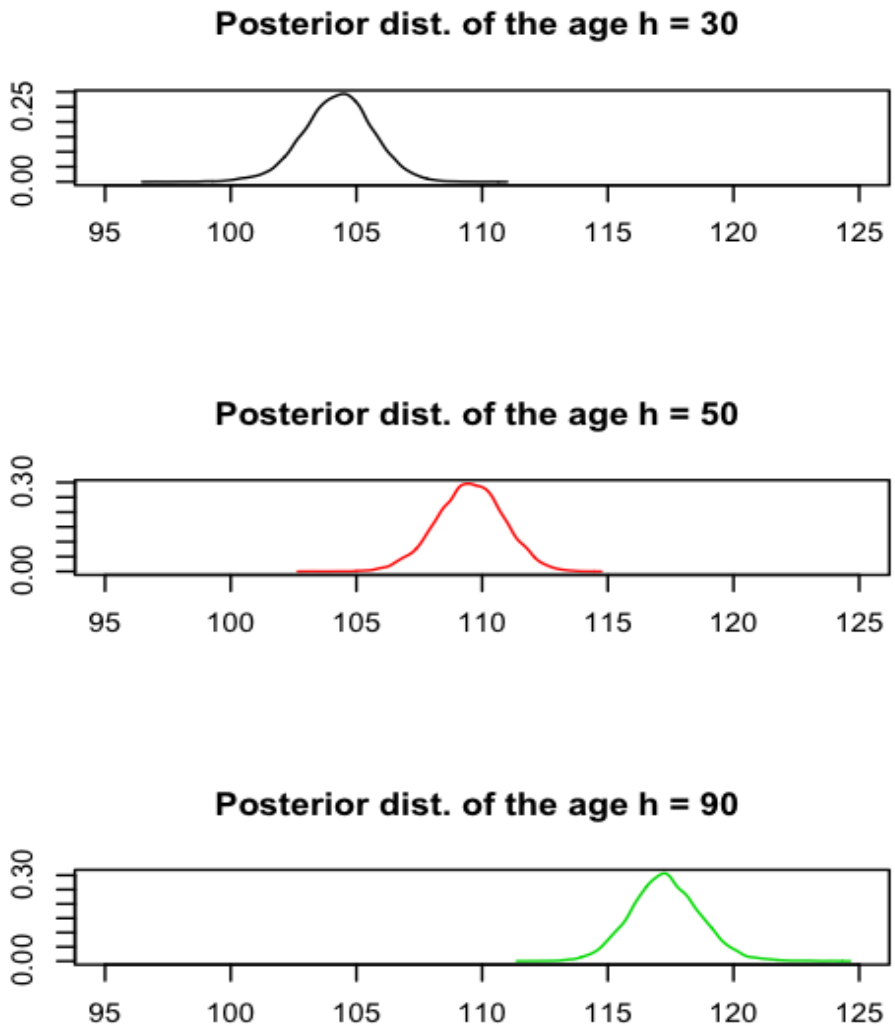
1. We estimate the relation f between the dates θ and the depth h

$$f(\theta) = h \qquad \text{age-depth curve}$$

   using
   - Posterior distribution of the sequence of dates (estimated by the Bayesin chronological model)
   - Non parametric regression method.
2. From the estimated curve, we predict the date as function of the depth.

Age -depth curve and forecasting





In progress R Package ArchaeoChron