IMATI-Milano Department

Main research topics:

- Data analysis
- Decision analysis
- Evaluation of seismic risk
- Industrial statistics
- Multivariate approximation
- Nonparametric Bayesian inference
- Reliability analysis
- Robustness of Bayesian analysis
- State-space modeling
- Stochastic models and parameter estimation in population dynamics

Stochastic analysis of natural hazards

Bruno Betrò, Antonella Bodini, Carla Brambilla,
Renata Rotondi, Elisa Varini

CNR - IMATI, Milano

Natural hazards in Italy

Natural phenomena of serious concern in Italy

- earthquakes
- landslides caused by extreme rainfalls
- Mathematical modelization requires the development of *stochastic models*

Earthquake occurrence analysis

Modelization of large earthquake sequences

- Renewal processes are considered appropriate models for sequences of large earthquakes, as one can assume that the stress accumulation process restarts after each event.
  The renewal model implies that the times between large seismic events can be considered as realizations of i.i.d. random variables $T_1, T_2, \ldots$.
- If F is the common distribution function of the T_i, the interest is in computing the occurrence probability at time t of an event in the next u years given the date t_{last} of the last event before t

$$\frac{F(t+u-t_{last}) \; - \; F(t-t_{last})}{1-F(t-t_{last})}$$

Most used distributions

- **exponential** distribution, hazard function
  $h(t) = \frac{f(t)}{1-F(t)} = \lambda, \quad f$ density function
- **gamma** distribution
  $$h(t) = \frac{b^a t^{a-1} e^{-bt}}{\Gamma(a) - \Gamma(a, bt)}$$ decreasing for $a < 1$, increasing for $a > 1$
- **lognormal** distribution
  $$h(t) = \frac{f(t)}{1 - \Phi\left(\frac{\log t - \xi}{\sigma}\right)}$$ initially increasing, then decreasing, $\longrightarrow 0$
- **Weibull** distribution
  $h(t) = c a^c \, t^{c-1}$ decreasing for $c < 1$, increasing if $c > 1$
- if $h(\cdot)$ is multimodal ?

Nonparametric estimation of F

- Renewal processes are clearly a simplification of the real physical process but they can lead to useful results if F is properly estimated.
- Nonparametric methods are adaptive to anomalous behaviour in the data set
- We do not make any assumption on the functional form of the distribution F of the inter-event times but consider this distribution as a random function modelled by a mixture of Polya trees (Lavine, *Ann. Statist.*, 20, 1225-1235 (1992))

The Bayesian approach

- X, Y absolutely continuous r.v.'s; *Bayes formula*:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{\int f_{Y|X}(y|u)f_X(u)\ du}$$

- in Bayesian inference, Bayes formula used for combining
  - *information from observations* x expressed by the *likelihood* $L(\theta; x) = f(x|\theta)$, $f(x|\theta)$ density of r.v. $X = (X_1, \ldots, X_n)$
  - a priori available information about unknown θ, *assumed summarizable in a density function* $\pi(\theta)$ *(a priori density)*;
- θ is seen as a r.v. with density $\pi(\theta)$, $f(x|\theta)$ is seen as a conditional density, $\Rightarrow$*a posteriori density* of θ

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|u)\pi(u)\ du}$$

The Bayesian nonparametric approach

- if we don't want to specify the form of $f(x|\theta)$ up to an unknown parameter, we can model f or the corresponding distribution function F as a *stochastic process* whose trajectories are densities or distribution functions *(random distribution function, random probability measure)*.
- e.g., if $Y(x), x \in R$ is a (right) continuous non-decreasing stochastic process such that $Y(-\infty) = 0$ and $Y(\infty) = \infty$, then

$$F(x) = 1 - exp(-Y(x))$$

  has trajectories satisfying conditions characterizing distribution functions.
- extending the parametric Bayesian approach to this situation makes it possible to obtain a posteriori information on F on the basis of a sample, e.g.

$$E(F(x)|x_1, \ldots, x_n)$$

Polya Trees

- $E^m = \{\epsilon : \epsilon$ *binary string of length* $m\}$, $\quad E^* = \bigcup_{m=0}^{\infty} E^m$, $E^0 = \emptyset$
- $\mathcal{X}$ separable measurable space (e.g. R^n)
  $\Pi = \{\pi_m; m = 0, 1, 2, \ldots\}$ nested partitions of $\mathcal{X}$
- $\pi_0 = \mathcal{X}$, $\pi_1 = \{B_0, B_1\}$, $B_0 \cap B_1 = \emptyset$, $B_0 \cup B_1 = \mathcal{X}$
- $\pi_2 = \{B_{00}, B_{01}, B_{10}, B_{11}\}$, $B_{00} \cap B_{01} = \emptyset$, $B_{00} \cup B_{01} = B_0$
  $B_{10} \cap B_{11} = \emptyset$, $B_{10} \cup B_{11} = B_1$
- $$\epsilon \in E^m, B_\epsilon \in \pi_m \begin{array}{c} \nearrow \\ \\ \searrow \end{array} \left.\begin{array}{c} B_{\epsilon 0} \\ \\ \\ B_{\epsilon 1} \end{array}\right\} \in \pi_{m+1}$$

Polya Trees (continued)

- A random probability measure $\mathcal{P}$ on $\mathcal{X}$ is said to have a Polya tree distribution with parameters $(\Pi, \mathcal{A})$, if there exist nonnegative numbers $\mathcal{A} = \{\alpha_\epsilon, \epsilon \in E^*\}$ and random variables $\mathcal{Y} = \{Y_\epsilon, \epsilon \in E^*\}$ s.t.
  - all the random variables in $\mathcal{Y}$ are independent
  - $\forall \epsilon \in E^*, Y_\epsilon$ has a Beta distribution with parameters $\alpha_{\epsilon,0}$ and $\alpha_{\epsilon,1}$
  - $\forall m = 1, 2, \ldots$ and $\epsilon \in E^m$
    $\mathcal{P}(B_{\epsilon_1,\ldots,\epsilon_m}) =$
    $\prod_{j=1;\epsilon_j=0}^{m} Y_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}} \times \prod_{j=1;\epsilon_j=1}^{m} \left(1 - Y_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}}\right)$

The role of $E(\mathcal{P})$

Define the probability measure $Q = E(\mathcal{P})$, by $Q(B) = E(\mathcal{P}(B))$ for any measurable set B

- it is easy to compute $Q(B_\epsilon)\ \forall B_\epsilon \in \bigcup_{m=0}^{\infty} \pi_m$
- Q can be extended to the measurable sets generated by $\bigcup_{m=0}^{\infty} \pi_m$
- if the r.v.'s $X_1, X_2, \ldots$ are a sample from $\mathcal{P}$, i.e. given $\mathcal{P}$, they are i.i.d. with distribution $\mathcal{P}$, then

$$\mathcal{P}(X_i \in B) = Q(B)$$

- Q is determined once $\Pi, \mathcal{A}$ are given; in the case $\mathcal{X} = R$, a distribution function $G(x)$ is given and the partition construction is lead by G;
- usual choices for $\alpha_{\epsilon_1, \ldots, \epsilon_m}$: m^2, 2^m, $k^m\ (k > 1)$

Predictive distribution

- $\mathcal{P}|X_1 = x_1$ has still a PT distribution; simple updating rule: it is enough to add 1 to every α_ϵ s.t. $x_1 \in B_\epsilon$
- exploiting the updating rule it is easy to compute $\mathcal{P}|X_2 = x_2, X_1 = x_1$ and so on
- if $\mathcal{X} = R$, then it is easy to compute

$$E(\mathcal{P}((-\infty, x))|x_1, \ldots, x_n) = E(\mathcal{F}(x)|x_1, \ldots, x_n)$$

i.e. a Bayesian estimate of the (unknown) distribution function of the observations $x_1, \ldots, x_n$

Choice of G: the Generalized gamma distribution

According to the information provided by the literature on the possible shape of the inter-event time distribution for strong earthquakes G is taken as a Generalized gamma distribution ($\mathcal{X} = R_+$)

- distribution with density

$$g(t; \eta, \xi, \rho) = \frac{\eta \xi^{\rho} t^{\rho\eta - 1} exp(-\xi t^{\eta})}{\Gamma(\rho)}, \eta, \xi, \rho > 0$$

- this class of distributions properly includes usual distributions

$\eta = \rho = 1$	exponential
$\eta = 1$	gamma
$\rho = 1$	Weibull
$\rho \to \infty$	lognormal

Mixtures of Polya Trees

Mixtures of PT instead of single PT's have the advantage of decreasing influence of the partition scheme and of the parameters of G

- Given a random variable U (index) with mixing distribution H s.t. for each u we have $\mathcal{P}|U = u \sim \mathrm{PT}\left(\Pi_u, \mathcal{A}_u\right)$
- the distribution of a random measure $\mathcal{P}$ is said to be a mixture of Polya Trees if, for any measurable set $\mathcal{S}$ of probability measures on $\mathcal{X}$

$$\Pr\left[\mathcal{P} \in \mathcal{S}\right] = \int \Pr\left[\mathcal{P} \in \mathcal{S}|u\right] \; H(du)$$

- parameters of G random vector $\mathbf{u} = (\eta, \xi, \rho)$, hierarchical structure:
  - $G(t|\mathbf{u})$
  - $H(\mathbf{u}) =$ Gamma$(\eta|\delta; g) \times$ Gamma$(\xi|\beta; a) \times \mathrm{Exp}(\rho; b)$
  - $\delta \sim$ Gamma(h, f) $\beta \sim$ Gamma(c, d)

Estimation via Markov Chain Monte Carlo (MCMC)

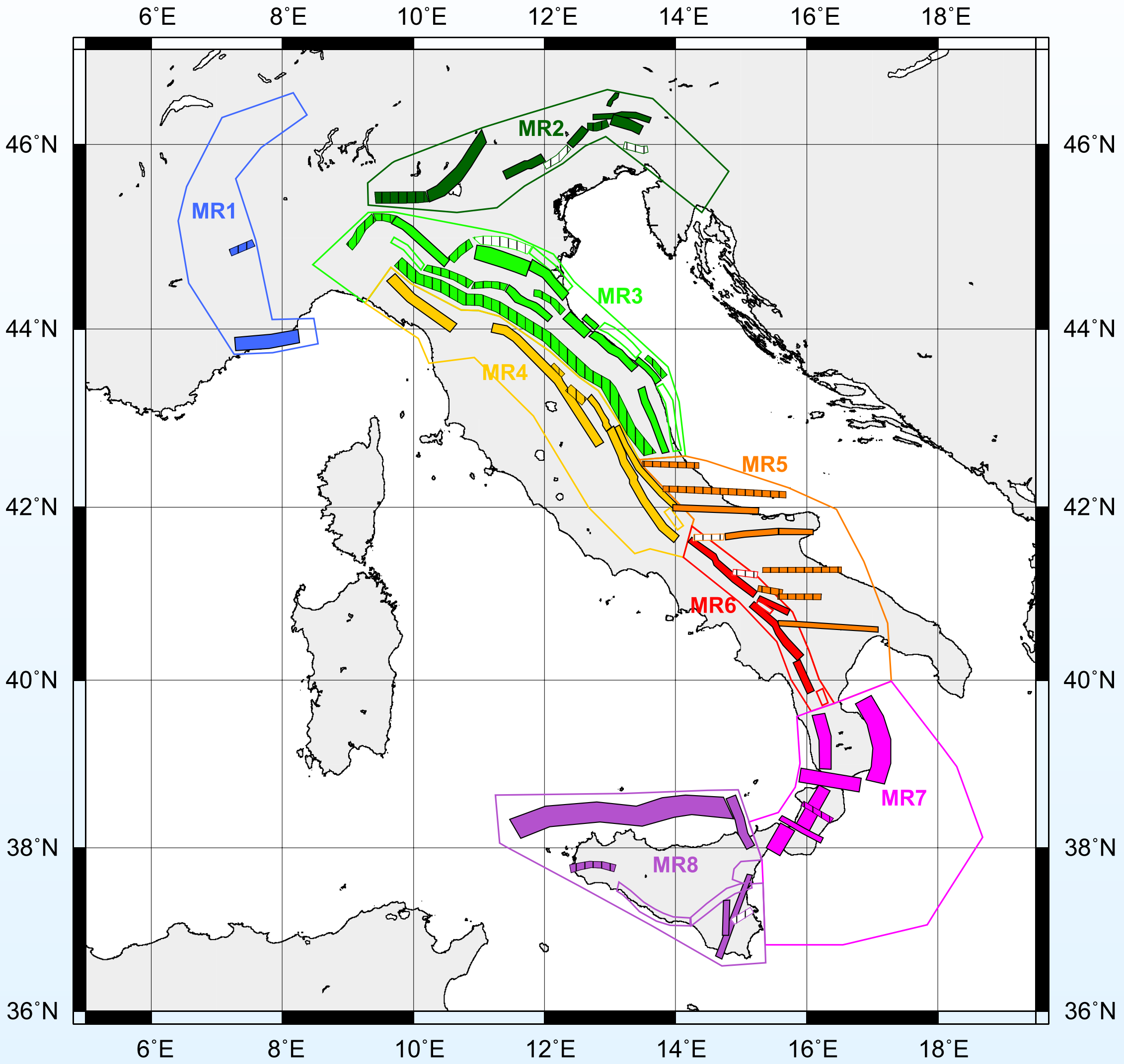
$\mathbf{t} = t_1, \ldots, t_n$ inter-event times

- update the current state $\mathbf{z} = (\eta, \xi, \rho, \delta, \beta)$, performing an iteration of MCMC computation using Metropolis-Hastings within Gibbs sampling such that $\pi(z_i | z_{-i}, \mathbf{t})$ is the equilibrium distribution of the Markov chain; repeat 500,000 times discarding the first 100,000 states (burn-in)
- every 50 iterations sample from the full conditional distribution of $\mathcal{P} | \mathbf{z}, \mathbf{t}$ by sampling from the Polya tree
  - generate values for the variates Y_ϵ's, and hence obtain probabilities $\{p_1, \ldots, p_{2^m}\}$ of belonging to the sets at the level m
  - draw samples of 50 inter-event times according to those probabilities
- use the simulated inter-event times to get a density estimate after a kernel smoothing

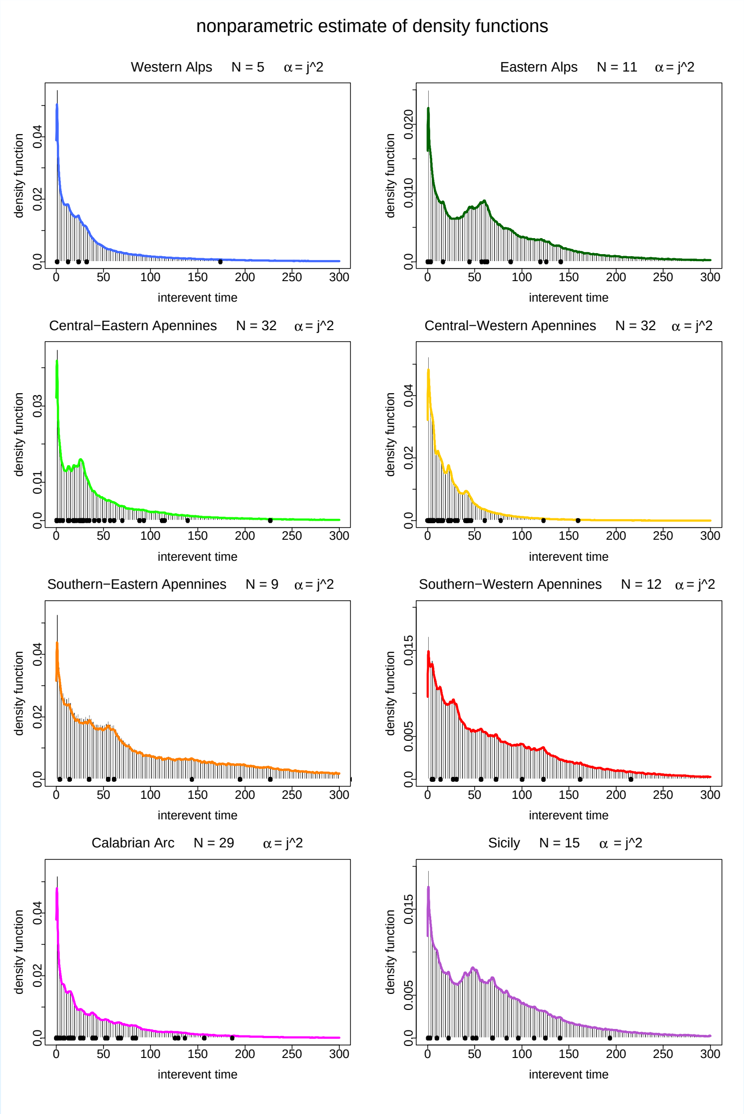
Hazard maps of Italy

- Data sets: earthquakes with M_w (Moment Magnitude) $\geqslant$ 5.3, occurred after 1600 up to 2002, drawn from the catalogue CPTI04; the inter-event times are calculated between shocks recorded in each of the seismogenic areas (DISS) belonging to the same tectonically homogeneous macroregion (MR)
  $\Longrightarrow$ we get 8 data sets used to estimate 8 density functions
- For each area the probability is obtained that an event occurs in the interval $(t, t+u)$ given the date t_0 of the last event
  $$\frac{F(t+u-t_0)-F(t-t_0)}{1-F(t-t_0)}$$
- for each MR parameters of $H(\mathbf{u})$ estimated through data from the other MR's and some information available in the seismic literature

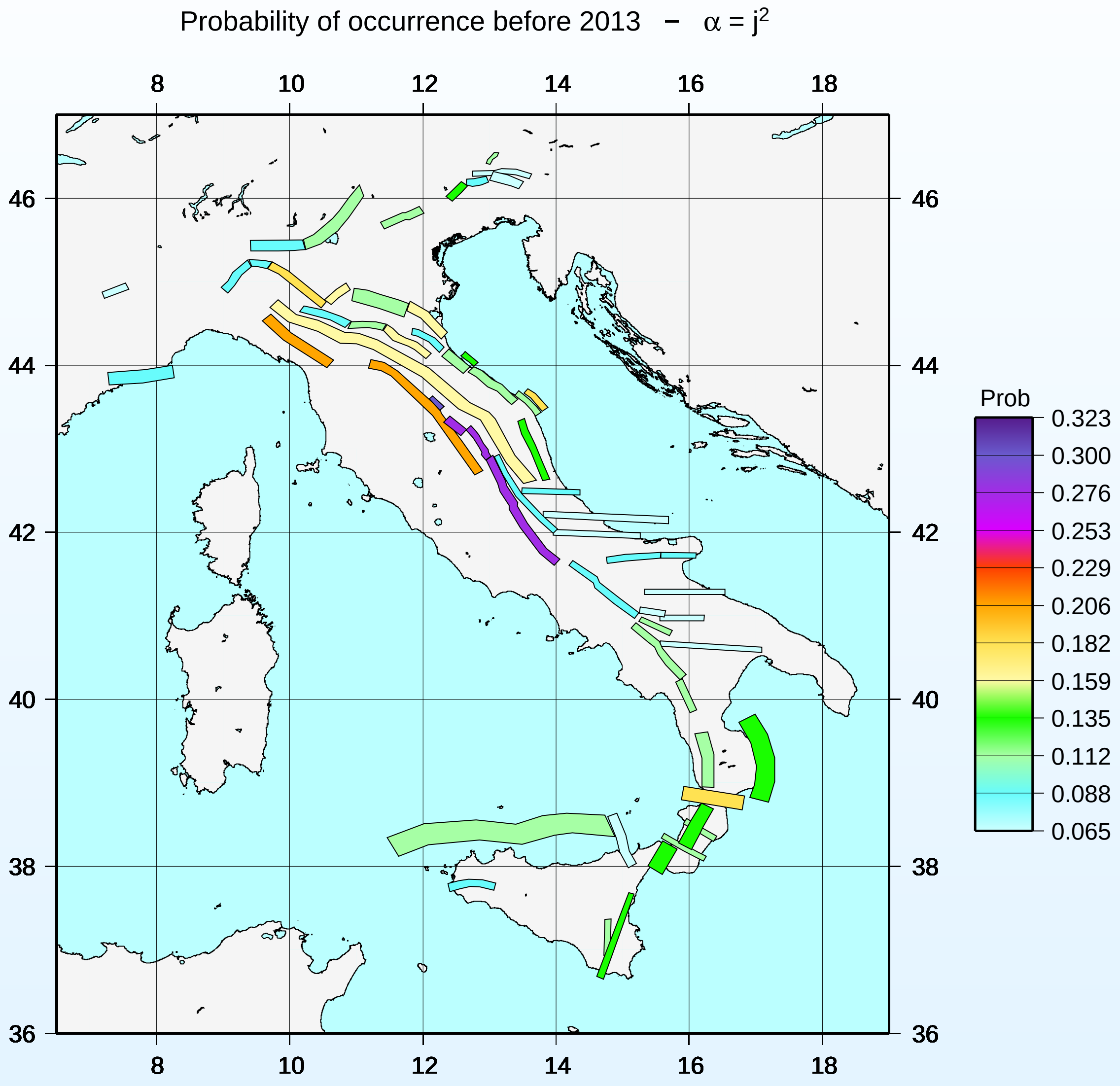
Homogeneous regions and seismogenic areas



estimated densities

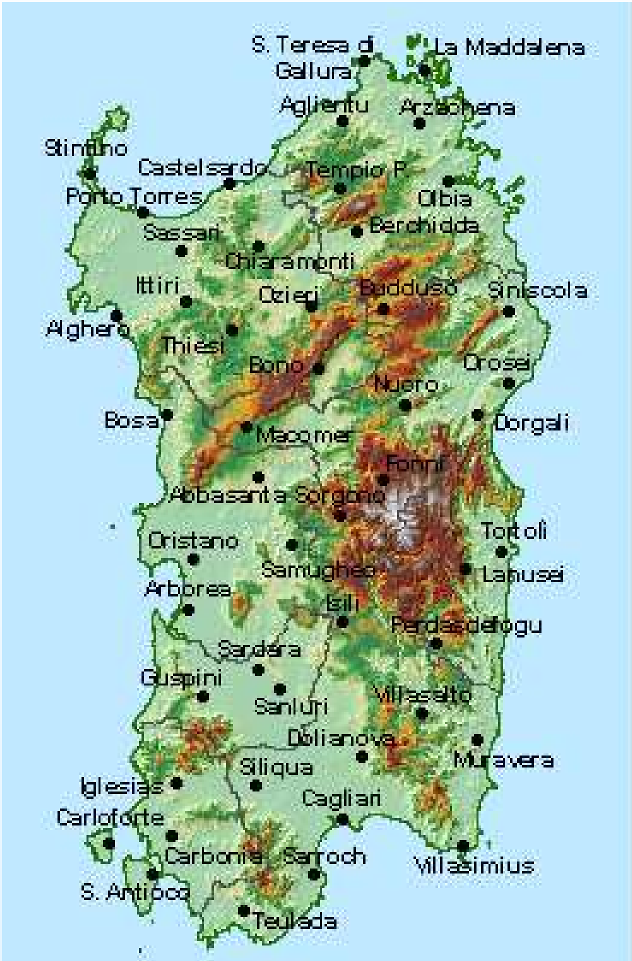


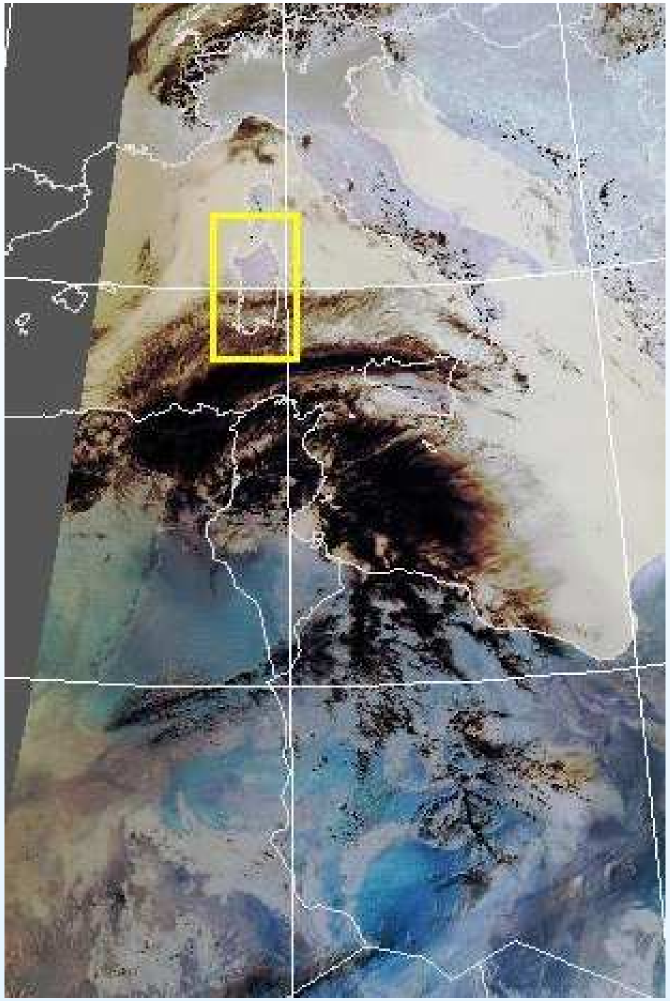
Occurrence probabilities



Modelization of rainfall patterns

The problem





6–9 Dec. 2004, Villagrande.

Peak of 500 mm in 12 h.

Aim of the work

Because of

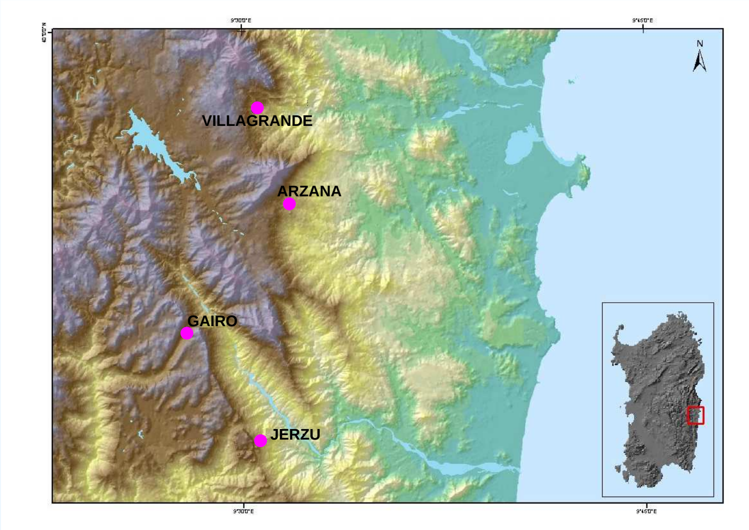
- high frequency of heavy or extreme rainfall events, which usually occur in a sudden way
- very local phenomenon $\Longrightarrow$ failure of GCM's

Then

- characterize the occurrence of extreme events in the seasonal rainfall path
- highlight a reasonable, although necessarily simplified, rainfall mechanism
- derive hydrogeological risk indexes, flash flood thresholds, ecc.

The study area

4 pluviometric stations of the Governmental Hydrographic Service:



- Daily rainfall data
- Standard Period from 1961–1990 (WMO)
- Season from September–January
- no other data are available
- many missing data from 1990–2000 and changes in the location of the pluviometric stations

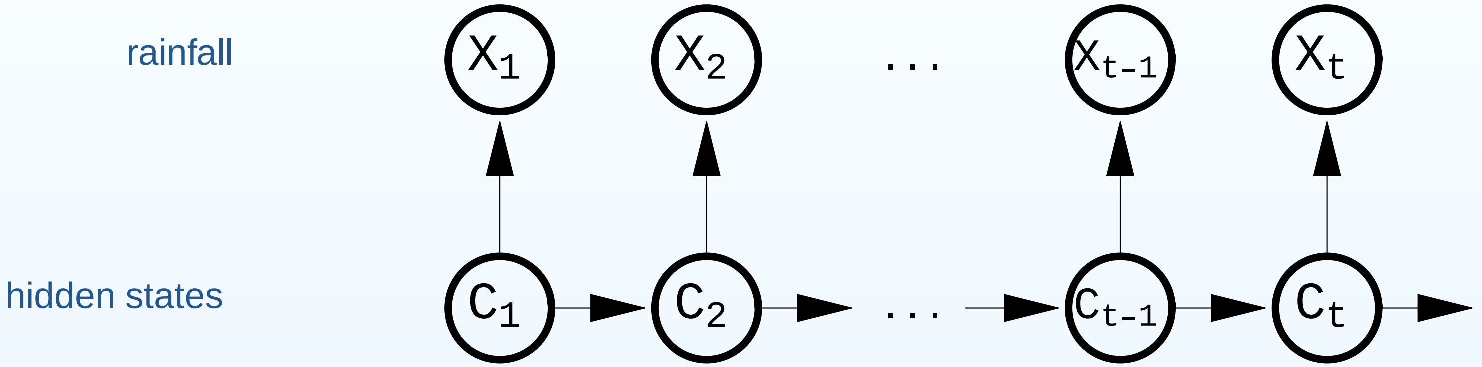
Data analysis

From September–January, 1961-1990

	ARZANA	GAIRO	JERZU	VILLAG.
Altitude (m a.s.l.)	674	784	550	679
mean n. of wet days (nr)	45.9	38.3	45.9	30.0
mean daily rainfall (mm) (μ)	3.86	3.47	3.26	3.55
Std. Deviation (mm) (σ)	14.15	12.74	10.89	13.81
mean maximum (mm)	120.1	100.5	92.3	133.4
mean Cumulate (mm)	591.0	531.1	499.2	542.5
mean n. of events $>$ 40mm (nr)	3.9	3.1	3.1	3.4
Outlier threshold (mm) ($\mu + 10\sigma$)	145.2	130.9	112.2	141.6
Total n. of events $\geqslant \mu + 10\sigma$ (nr)	9	9	7	7
Total n. of events $\geqslant$100 mm (nr)	22	16	9	15
complete records (nr)	22	24	23	16

Hidden Markov Models (HMMs): A graphical definition

$\{X_t\}_t$ rainfall process, $\{C_t\}$ *hidden* process



Correspondence between **hidden states and the concept of discrete weather state**

(Bárdossy & Plate, 1992)

- **In B & P: states defined a priori (GCM's output)**
- **In HMMs: states inferred from data**

HMMs: A formal definition

$X_t = (X_{t1}, \ldots, X_{tq})$ r.v., q rain stations; $x_{ti} \in \mathbb{R}_0^+$
$C_t \in \{1, \ldots, m\}$ hidden process

$X_{1:T} := (X_1, \ldots, X_T)$, $C_{1:T} := (C_1, \ldots, C_T)$
$\mathcal{L}(\cdot) \equiv$ distribution of $\cdot$

- $\{C_t\}$ homogeneous, first–order Markov Chain
- $\mathcal{L}(X_t | X_{1:t-1}, C_{1:t}) = \mathcal{L}(X_t | C_t) = \prod_i \mathcal{L}(X_{ti} | C_t)$
- $\mathcal{L}(X_t | C_t)$ does not depend on t
- $\mathcal{L}(X_{ti} | C_t = c) \ = w_{ic}\, \delta_0 + (1 - w_{ic}) F(\,\cdot\,| \theta_{ic})$

Charles *et al.* (1999)

The adopted model

$F =$ **mixture** of **Weibull** distributions:

$$W(x;\alpha,\beta) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} \exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right], x>0; \alpha>0,\ \beta>0$$

- Weibull distr. is an extreme value distribution
- is a transformation of the exponential distribution
- mixtures can capture different types of extreme values

$$\Rightarrow$$
$$\mathcal{L}(X_{ti}|C_t = c) =$$
$$w_{ic}\,\delta_0 + (1-w_{ic})\sum_{k=1}^{K}\gamma_k W\big(\,\cdot\,|(\alpha_{ic}^{k},\beta_{ic}^{k})\big)$$
$$\gamma_k > 0,\ \sum_{k=1}^{K}\gamma_k = 1,\ K = 1,2,3.$$

Estimation and diagnostics

- MVNHMM toolbox, Kirshner (2005, 2007), http://www.cs.ualberta.ca/~sergey/MVNHMM/ (EM algorithm)

$$X \sim W(\alpha, \beta) \Rightarrow \left(\frac{X}{\alpha}\right)^{\beta} \sim \exp(1)$$

- $\beta = 2$, by fitting a Weibull distribution to annual maxima in each station: JUST A TRICK!!
- BIC + cross validation, for model selection
- goodness–of–fit, by comparing empirical and estimated relevant quantities

Estimated model

- 6 hidden states
- $\mathcal{L}(X_{ti}|C_t = c) = w_{ic}\,\delta_0 + (1 - w_{ic})\sum_{k=1}^{2}\gamma_k W(\,\cdot\,|(\alpha_{ic}^{k}, 2))$

w_{ic}	C=1	C=2	C=3	C=4	C=5	C=6
Arzana	0.80	0.04	0.05	0.09	1.00	0.29
Gairo	0.79	0.05	0.17	0.35	1.00	0.49
Jerzu	0.69	0.02	0.04	0.13	1.00	0.16
Villagrande	0.93	0.06	0.10	0.35	1.00	0.76

Interpretation of states

In terms of *weather states*:

- 5 = high pressure system
- 2 = moist currents from South–East
- 3 = moist currents from South–East, less intense phenomena, except for Gairo & Villagrande
- 4 = moderate rainfall
- 6 = rainfall from absent to weak
- 1 = negligeable rainfall, apart from Gairo

Note that states appear to be well separated.