HIDDEN MARKOV MODELS

FOR RAINFALL MODELING

Antonella Bodini *et al.*

anto@mi.imati.cnr.it

CNR-IMATI, Milan

Summary

- Some definitions
- An application to data from Sardinia (Italy)
- Some comments and references

!!! Complementary summary !!!

- Spatial issues
- Bayesian Inference

Some definitions

$X_t = (X_{t1}, \dots, X_{tq})$ r.v., q rain stations:

$x_{ti} \in \{0, \dots, K\}$ or $x_{ti} \in \mathbb{R}^+$

$C_t \in \{1, \dots, m\}$ hidden process

$X_{1:T} := (X_1, \dots, X_T),\ C_{1:T} := (C_1, \dots, C_T)$

Some definitions

$X_t = (X_{t1}, \ldots, X_{tq})$ r.v., q rain stations:

$x_{ti} \in \{0, \ldots, K\}$ or $x_{ti} \in \mathbb{R}^+$

$C_t \in \{1, \ldots, m\}$ hidden process

$X_{1:T} := (X_1, \ldots, X_T),\ C_{1:T} := (C_1, \ldots, C_T)$

MacDonald and Zucchini (1997)

- $\mathcal{L}(X_t | X_{1:t-1}, C_{1:t}) = \mathcal{L}(X_t | C_t)$
- C_t homogeneous, first–order Markov Chain

Some definitions

$X_t = (X_{t1}, \ldots, X_{tq})$ r.v., q rain stations:

$x_{ti} \in \{0, \ldots, K\}$ or $x_{ti} \in \mathbb{R}^+$

$C_t \in \{1, \ldots, m\}$ hidden process

$X_{1:T} := (X_1, \ldots, X_T), \; C_{1:T} := (C_1, \ldots, C_T)$

- $\mathcal{L}(X_t | X_{1:t-1}, C_{1:t}) = \mathcal{L}(X_t | C_t)$
- C_t homogeneous, first–order Markov Chain
- $\mathcal{L}(X_t | C_t) = \prod_i \mathcal{L}(X_{ti} | C_t)$ and DOES NOT DEPEND ON t

  Zucchini and Guttorp (1991)

Interpretation

The main interest of HMMs lies in the underlying correspondence between the hidden states and the concept of discrete weather states. Instead of explicity defining the weather states, HMMs allow to define them according to observed data. Therefore, an explicit mechanism for simulating the phenomenon is provided.

Cases of interest

Rainfall occurrences:

$$X_{ti} = \begin{cases} 0 & \text{DRY day at station } i \\ 1 & \text{WET day } \ldots \end{cases}$$

Rainfall intensities:

$$X_{ti} = \begin{cases} 0 & \text{DRY day at station } i \\ 1 & \text{WEAK rainfall } \ldots \\ \vdots & \qquad \vdots \\ K & \text{VERY STRONG rainfall } \ldots \end{cases}$$

Rainfall amounts:

$$X_{ti} \geq 0$$

Cases of interest: distributions

Rainfall occurrences:

$$X_{ti} = \begin{cases} 0 & \text{DRY day} \\ 1 & \text{WET day} \end{cases} \Rightarrow \quad P(X_{ti} = 1 | C_t = c) = p_{ic}$$

Rainfall intensities:

$$X_{ti} = \begin{cases} 0 & \text{DRY day at station } i \\ 1 & \text{WEAK rainfall } \ldots \\ \vdots & \vdots \\ K & \text{VERY STRONG rainfall } \ldots \end{cases}$$

Rainfall amounts:

$$X_{ti} \geq 0$$

Cases of interest: distributions

Rainfall occurrences:

$$X_{ti} = \begin{cases} 0 & \text{DRY day} \\ 1 & \text{WET day} \end{cases} \quad \Rightarrow \quad P(X_{ti} = 1 | C_t = c) = p_{ic}$$

Rainfall intensities:

$$X_{ti} = \begin{cases} 0 & \text{DRY day at station } i \\ 1 & \text{WEAK rainfall } \ldots \\ \vdots & \qquad \vdots \\ K & \text{VERY STRONG rainfall } \ldots \end{cases}$$

Rainfall amounts:

$$X_{ti} \geq 0 \; \Rightarrow \; \mathcal{L}(X_{ti} | C_t = c) = w_{ic}\,\delta_0 + (1 - w_{ic}) F(\,\cdot\,| \theta_{ic})$$

The study area

see the map

Central–East Sardinia; 4 stations (Arzana, Gairo, Jerzu and Villagrande).

Data: standard 30 year period, season from September to January $\Rightarrow$ 4437 data.

Available data: daily rainfall and temperature.

Unfortunately temperature does not predict rainfall ...

Estimation and selection model

The numerical maximization of log–likelihood is essentially based on an EM algorithm. The MVNHMM toolbox (Kirshner, 2005) is available on line at the web site

http://www.datalab.uci.edu/software/mvhmm/

The Bayesian Information Criterion (BIC) can be used to determine the number of states. Cross–validation arguments can be used too.

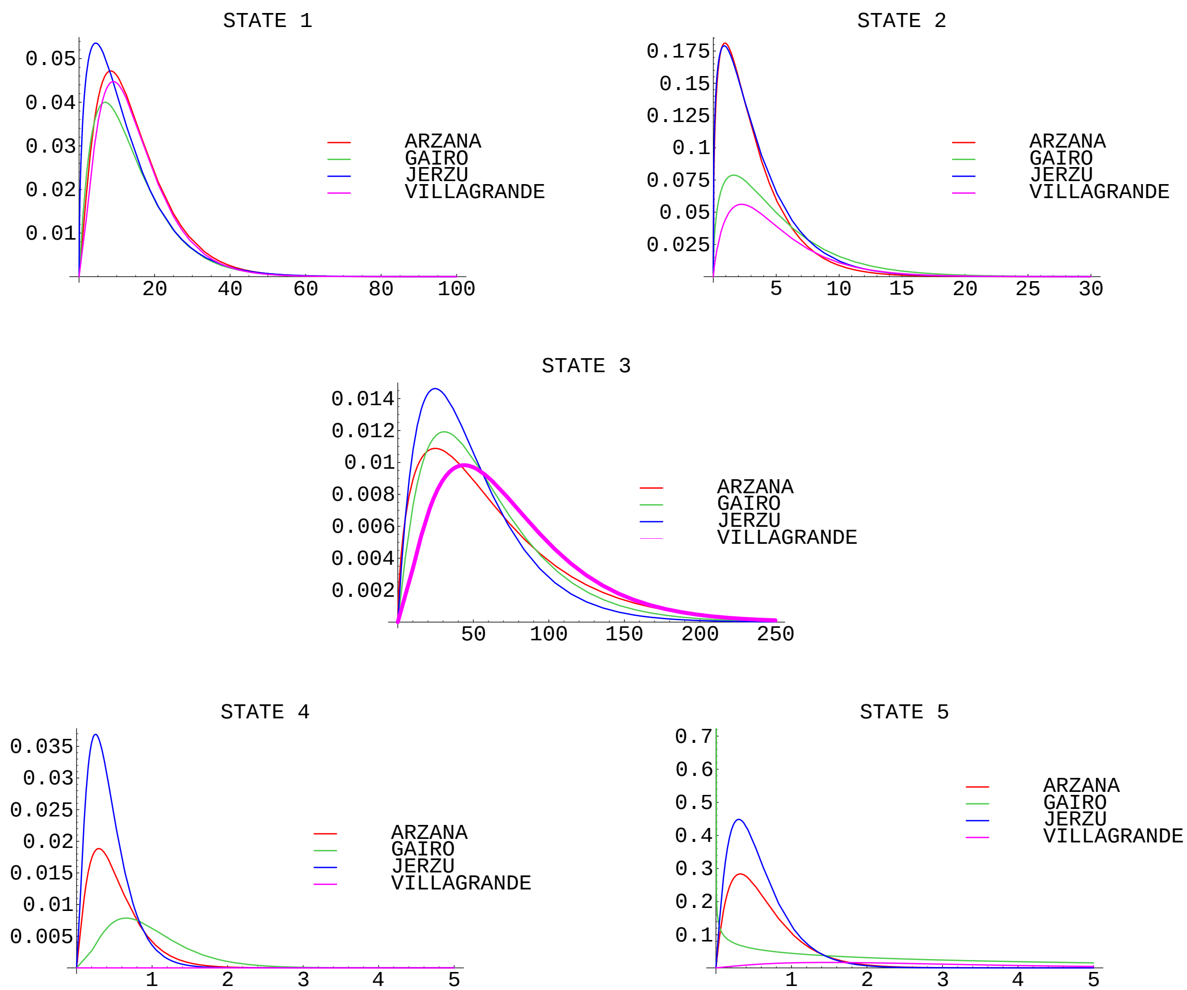
Estimated model, I

$$X_{ti}|C_t = c \sim w_{ic}\delta_0 + (1 - w_{ic})Gamma(\cdot|\alpha_{ic}, \beta_{ic})$$

Estimated Dirac's weights

stations	C=1	C=2	C=3	C=4	C=5
Arzana	0.08	0.21	0.04	0.99	0.75
Gairo	0.24	0.46	0.06	0.99	0.75
Jerzu	0.08	0.16	0.02	0.98	0.66
Villagrande	0.15	0.62	0.07	0.999	0.94
π	0.10	0.18	0.03	0.51	0.18

Estimated model, II



Estimated model, III

Estimated State Sequence (Viterbi's algorithm):

the most likely sequence of states associated with data.

	C=1	C=2	C=3	C=4	C=5
Frequencies	15.4	25.9	4.1	83.2	24.3
Mean daily rainfall	12.6	2.6	58.8	0.01	0.55

Mean daily rainfall conditioned to C=3

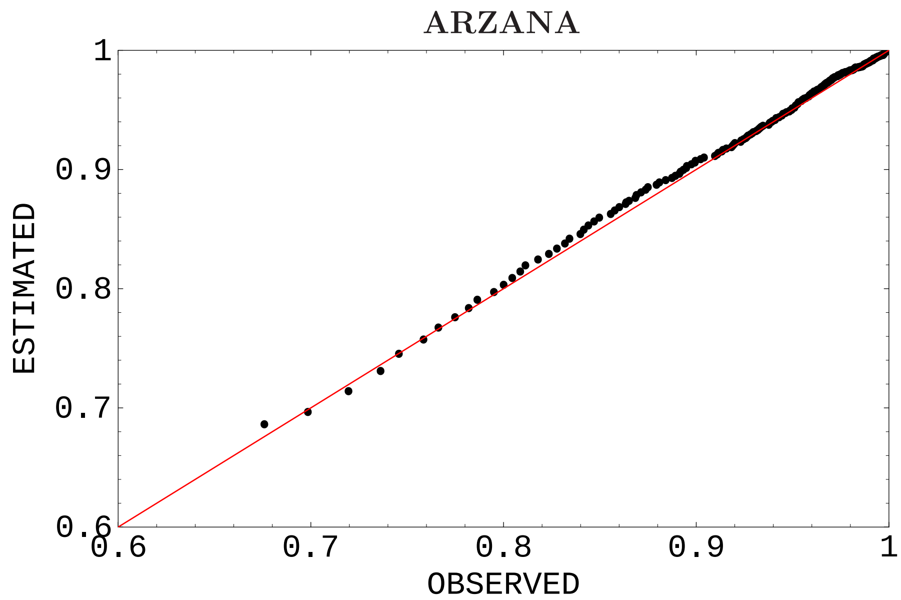
Arzana	Gairo	Jerzu	Villagrande
64.4	57.7	50.0	70.8

Goodness of fit

NB: empirical frequencies are usually matched by the corresponding estimates.

Goodness of fit

Comparison of empirical and estimated distribution function. Note that here observations are dependent (Altman, 2004).



Some comments

- Boostrap can be used for determining confidence intervals

Some comments

- Boostrap can be used for determining confidence intervals
- Spatial correlation has to be considered

Spatial correlation

- Spatial correlation has to be considered

(Hughes *et al.*, 1999): autologistic model

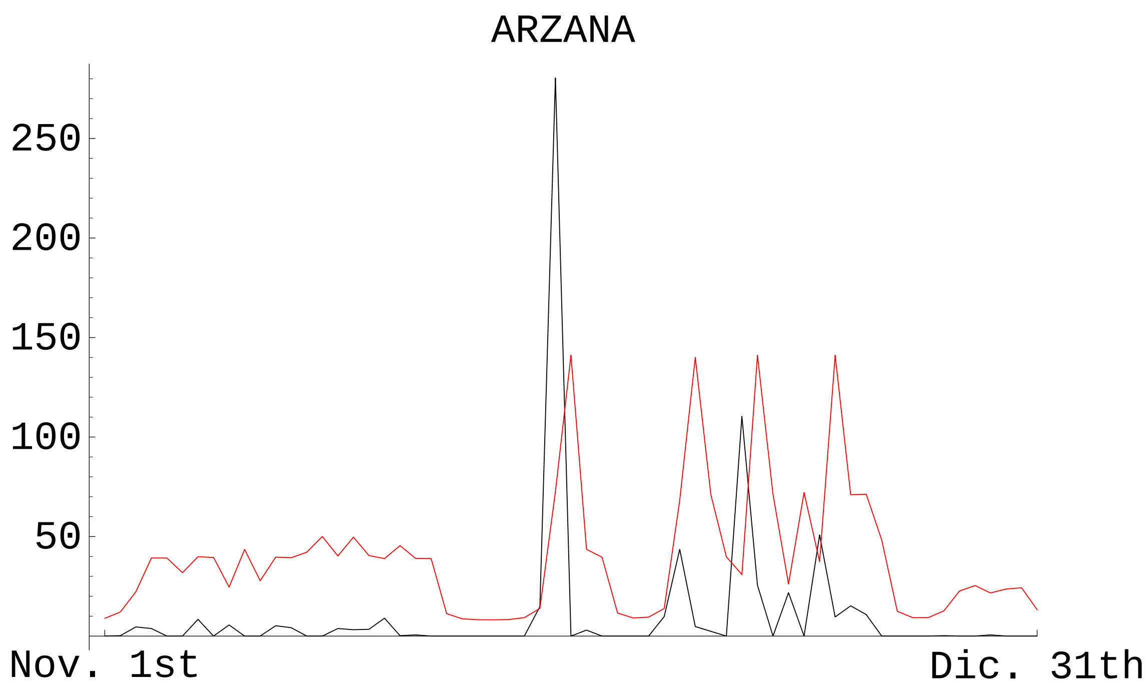
$$P(X_t|C_t = c) \propto \exp\left(\sum_{i=1}^{q} \alpha_{ci} x_{ti} + \sum_{i=1}^{q} \beta_{cij} x_{ti} x_{tj}\right)$$

Some comments

- Boostrap can be used for determining confidence intervals
- Spatial correlation has to be considered
- !!! The estimated model does not provide good predictions !!! $\Rightarrow$
  - other atmospheric data
  - downscaling (Hughes *et al.*, 1999)

Predictions

$$P(X_{t+1,Arzana} \leq \ red\ line | X_{1:t}) = 0.95$$



- !!! The estimated model does not provide good predictions !!! $\Rightarrow$
  - Downscaling of GCM

(Hughes *et al.*, 1999)

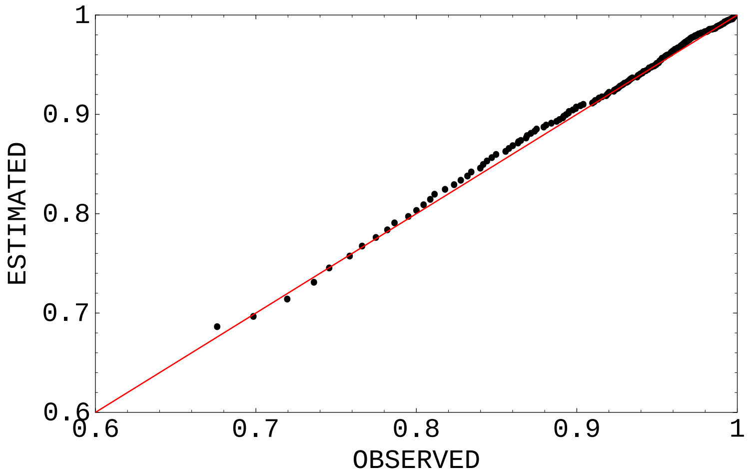
$$P(C_t = i | C_{t-1} = j, X_t) \propto$$

$$P(C_t = i | C_{t-1} = j) P(X_t | C_{t-1} = j, C_t = i) = \gamma_{ij} \mathcal{N}(\mu, V)$$

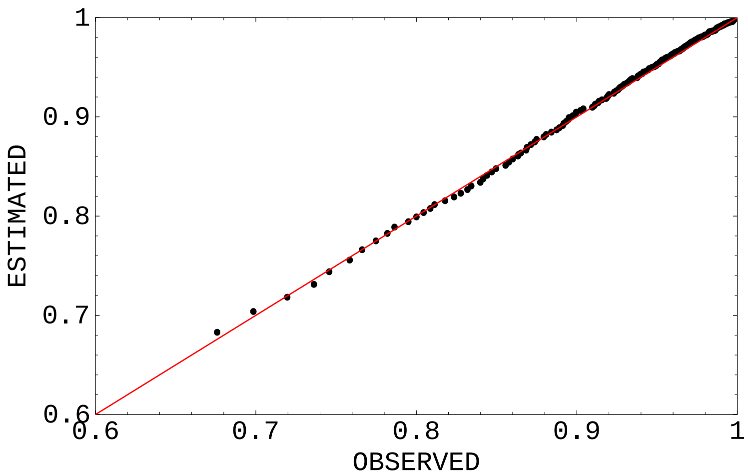
Some comments

- Boostrap can be used for determining confidence intervals
- Spatial correlation has to be considered
- !!! The estimated model does not provide good predictions !!! $\Rightarrow$
  - other atmospheric data
  - downscaling (Hughes *et al.*, 1999)
- Transformation of data to improve the fit

Transformation to improve de fit



Real data



Transformed data

Some comments

- Boostrap can be used for determining confidence intervals
- Spatial correlation has to be considered
- !!! The estimated model does not provide good predictions !!! $\Rightarrow$
  - other atmospheric data
  - downscaling (Hughes *et al.*, 1999)
- Transformation of data to improve the fit
- Bayesian Inference

References

Hughes J.P., Guttorp P., Charles S.P. (1999) *A nonhomogeneous hidden Markov model for precipitation occurrence.* J. Roy. Satist. Soc. C, 48, 15–30.
MacDonald I.L., Zucchini W. (1997) *Hidden Markov and Other Models for Discrete Time Series.* Chapmann & Hall, London.
Zucchini W., Guttorp P. (1991) *A hidden Markov model for space–time precipitation.* Water Resources Research, 27, 1917–1923.

Altman MCK. (2004) *Assessing the Goodness-of-Fit of Hidden Markov Models.* Biometrics, 60, 444–450.

Betrò B., Bodini A., Gullà G., Terranova O. (2006) *Analysis of daily rainfall occurrence over southern Calabria Ionica via a Hidden Markov Model.* Technical report 06-02, CNR-IMATI, Milan. *http://www.mi.imati.cnr.it/iami/abstracts/06-02.html*