Stochastic weather generators
with non-homogeneous hidden Markov switching
Application to temperature series

Valérie Monbet¹, Pierre Ailliot²

¹IRMAR, Rennes 1
²LMBA, UBO
Stochastic weather generators?

- **Stochastic** tools for generating sequences of meteorological variables

### Historical data

- 40 years of daily mean temperature at Brest

### Stochastic model

- SWGEN calibrated on historical data

### Synthetic data

- Large number of synthetic temperature sequences with statistical properties similar to original data

---

Monbet, Ailliot

Switching autoregressive models
Stochastic tools for generating sequences of meteorological variables

Used in impact studies when climate is involved and not enough data to estimate the quantities of interest

Synthetic weather conditions → System → Distribution of the response

Most usual applications: hydrology and agriculture
- Meteorological variables: rainfall, temperature, solar radiation, humidity, wind speed,...

In this talk, firstly focus on daily mean temperature at a single location
- Brest, 40 years, $\Delta t = 1$ day
- Preprocessing step to remove seasonal components (+ trends, daily components)?
  In this talk, data blocked by month, results for January
Some existing stochastic weather generators

- **Data oriented** (non parametric): random sampling with replacement from the original data, also called **bootstrapping**
  - Marginal distribution of the data reproduced by construction
  - Cannot create unobserved meteorological situations
  - It may be difficult to restore the dependence structure (especially for long events)

Ref: Rajagopolan and Lall (1999)

- **Model oriented** (parametric)
  - Difficult to build realistic multi-site and multi-parameter generators
  - Can create unrecorded situations
  - Interpretable

Ref: Wilks and Wilby (1999), Srikanthan et al. (2001)
Transformed Gaussian processes

- Classical approach for modeling time series
- Simulation of the stationary time series \( \{Y_t\} \) in three steps:
  - **First step:** find a deterministic monotonic function \( g \) such that
    \[
    X_t = g(Y_t)
    \]
  - "Normal score transformation": \( g = \Phi^{-1} \circ F_Y \)
    \[
    F_Y(y) = P[Y_t < y], \Phi \text{ cdf of } \mathcal{N}(0, 1)
    \]
  - "Box-cox transformation": \( g(y) = \sqrt[\lambda]{y} - 1 \)
  - **Second step:** \( \{X_t\} \) is a stationary process with (approximately) Gaussian margins. Further assume that \( \{X_t\} \) is a Gaussian process and simulate this process.
    - Exact simulation
    - ARMA models
  - **Third step:** apply the inverse transform \( g^{-1} \) to the simulated Gaussian sequence

- **Temperature data**
  - Transformation to Gaussian margins: \( Y_{t}^{(transf)} = g(Y_t) = \sqrt{13 - Y_t} \)
  - Time series model AR(2): \( Y_t^{transf} = 0.57 + 0.83 Y_{t-1}^{transf} - 0.07 Y_{t-2}^{transf} + 0.41 \epsilon_t \)
  - Inverse transformation \( \hat{Y}_t = g^{-1}(Y_t^{(transf)}) = 13 - (Y_t^{(transf)})^2 \)

Ref: Brockwell and Davies (2002), Monbet et al. (2007)
Transformed Gaussian processes

- Temperatures are too hot: transformation?
- Correlation function decreases too slow for short lags and is good for large lags: mixture of AR models?
- Transformed Gaussian process can not reproduce some non-linearities:
  - A stationary Gaussian process has a symmetric behavior below \( \alpha \) and above \( 1 - \alpha \) quantile
  - The same holds true for transformed Gaussian processes (since \( g \) is monotonic)
  - Observed hot regimes have shorter durations than cold regimes

Data, simulation

Monbet, Ailliot
Switching autoregressive models
Transformed Gaussian processes

+ Easy to implement
+ Can be used for multivariate data (several sites and/or variables)
  - May be hard to interpret
  - Can not reproduce some non-linearities

Alternatives?
  - More general transformation $X_t = g(Y_{h(t)})$ with $g$ and $h$ (stochastic?) transformations such that $\{X_t\}$ is a Gaussian process
  - Non-linear time series models
Events-based models
  - Summarize original data by a succession of "events"
    e.g. rain events, swell/wind sea systems, tropical cyclones
  - Model intensity/time of arrival/duration/shape/... of the events

Ref: Onof et al. (2000)

Weather type models
  - Weather type = typical weather situation
    - e.g. dry/wet days, convective/frontal rainfall, cyclonic/anticyclonic conditions,...
    - Generally between 2 and 10 weather types are used; enough to summarize climate
  - Weather type models characterized by
    - A model for the weather type sequence
      Markov chain or semi-Markov model.
    - A model for the meteorological variables conditionally to the weather type
      Independence conditionally to the state or autoregressive models.
  - Weather type models are widely used for meteorological variables
    - Various strategies available for defining the weather type
      - A priori classification (e.g. wet/dry)
      - Latent variable (e.g. HMM)
Weather type models

- An example: the chain dependent model (Richardson, 1981)
  - **Weather type** $S_t=$ {Dry, Wet}
    - First order Markov chain with 2 states
  - **Rainfall amount** $R_t$ conditionally independent in time given the weather type sequence
  - **Temperature, solar radiation, wind speed** $Y_t$
    - Residual correlation between successive observations modeled as an AR process with time varying coefficients
      $$ Y_t = \beta_0^{(S_t)} + \beta_1^{(S_t)} Y_{t-1} + \sigma^{(S_t)} \epsilon_t $$

  $(\beta_i^{(s)}), (\sigma^{(s)})$ unknown parameters and $\{\epsilon_t\}$ a Gaussian white noise

  $Rainfall \quad \cdots \quad R_{t-1} \quad R_t \quad R_{t+1} \quad \cdots$

  $\uparrow \quad \uparrow \quad \uparrow$

  $Weather \ type \ (wet/dry) \quad \cdots \rightarrow S_{t-1} \quad \rightarrow S_t \quad \rightarrow S_{t+1} \quad \rightarrow \cdots$

  $\downarrow \quad \downarrow \quad \downarrow$

  $Temperature, \ (wind, \ ...) \quad \cdots \rightarrow Y_{t-1} \quad \rightarrow Y_t \quad \rightarrow Y_{t+1} \quad \rightarrow \cdots$

- Temperature dynamics modeled as a mixture of two AR(2) processes
Richardson’s model for temperature time series

- The selected model is not easily interpretable (results for January at Brest): the regimes are close to each other

\[ Y_t = \begin{cases} 1.26 + 0.88 Y_{t-1} - 0.08 Y_{t-2} + 1.96 \epsilon_t & (S_{t-1} = \text{dry}), \quad \mu^{(1)} = 6.41 \\ 1.28 + 1.01 Y_{t-1} - 0.18 Y_{t-2} + 2.10 \epsilon_t & (S_{t-1} = \text{wet}), \quad \mu^{(2)} = 7.39 \end{cases} \]

Transition matrix: \[
\begin{bmatrix}
0.55 & 0.45 \\
0.33 & 0.77
\end{bmatrix}, \text{ stationary distribution } \begin{bmatrix}
0.24 \\
0.76
\end{bmatrix} \]
Richardson’s model for temperature time series

- Introduction of weather states gives more flexibility to the model → marginal distribution better reproduced than in Box and Jenkins model → idem for low level up crossings.
- But, the dry/wet classification seems not convenient for the temperature for Brest’s climate.
- The non linearities are still not reproduced by the model.

- Weather types can be obtained from other variables?
Clustering on meteorological variables (e.g. pressure,...)

*Pressure, Wind, Rainfall*  
\[ \cdots R_{t-1} \downarrow \cdots R_t \downarrow \cdots R_{t+1} \cdots \]

*Weather type*  
\[ \cdots \rightarrow S_{t-1} \downarrow \rightarrow S_t \downarrow \rightarrow S_{t+1} \rightarrow \cdots \]

*Temperature*  
\[ \cdots \rightarrow Y_{t-1} \downarrow \rightarrow Y_t \downarrow \rightarrow Y_{t+1} \rightarrow \cdots \]

<table>
<thead>
<tr>
<th>Pression</th>
<th>Wind</th>
<th>Precipitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regime 1 (anticyclonic)</td>
<td>1021</td>
<td>8.5 ms(^{-1})</td>
</tr>
<tr>
<td>Regime 2 (cyclonic)</td>
<td>1007</td>
<td>12.4 ms(^{-1})</td>
</tr>
</tbody>
</table>

**Model**

\[
Y_t = \begin{cases} 
1.42 + 0.73Y_{t-1} + 0.02Y_{t-2} + 1.90\varepsilon_t & (S_{t-1} = \text{anticyclonic}), \\
1.26 + 0.88Y_{t-1} - 0.08Y_{t-2} + 1.96\varepsilon_t & (S_{t-1} = \text{cyclonic}),
\end{cases}
\]

with non parametric semi-Markov transitions.

*Anticyclonic* vs *Cyclonic*
This classification is close to the wet/dry's one?

Clustering on meteorological variables (e.g. pressure,...)
This classification with semi Markov model for the chain improves the results compared to Richardson’s model:
→ Correlation is well restored,
→ mean number of up-crossings is improved but still to symmetric,
→ not enough high temperatures.

Latent weather type can help to better reproduce non linearities?
Introduce the weather type as a latent (hidden) variable: "Markov Switching AutoRegressive" model (MS-AR)

- Hidden weather type modeled as a first order Markov chain
- Linear Gaussian AR(p) model for the temperature evolution conditionally to the weather type

\[ Y_t = \beta_0^{(S_t)} + \beta_1^{(S_t)} Y_{t-1} + \ldots + \beta_p^{(S_t)} Y_{t-p} + \sigma^{(S_t)} \epsilon_t \]

- (\beta_i^{(s)}) and (\sigma^{(s)}) unknown parameters and \{\epsilon_t\} iid \mathcal{N}(0, 1) sequence
- Maximum likelihood estimates (MLE) can be computed with the EM algorithm
The selected model is interpretable (results for January at Brest)

\[ Y_t = \begin{cases} 
1.10 + 0.86Y_{t-1} - 0.09Y_{t-2} + 1.83\epsilon_t & (S_{t-1} = 1), \\
7.20 + 0.32Y_{t-1} - 0.03Y_{t-2} + 1.02\epsilon_t & (S_{t-1} = 2),
\end{cases} \]

\( \mu^{(1)} = 4.64 \quad \mu^{(2)} = 10.07 \)

Transition matrix: \[
\begin{bmatrix}
0.9 & 0.1 \\
0.4 & 0.6
\end{bmatrix}, \text{ stationary distribution } \begin{bmatrix}
0.8 \\
0.2
\end{bmatrix}
\]

- **Regime 1**: lower temperature with slow variations (anticyclonic)
- **Regime 2**: higher temperature with low and quick variation around the mean (cyclonic)
A MS-AR model for the temperature data, simulation

The model is able to reproduce some important statistical properties of the data.
- MS-AR models much better reproduce the observed asymmetry in up-crossing rate
  - No enough variability around middle levels?

- In MS-AR models the regime switchings are independent of past temperature conditions
  - The probability of staying in the cold regime at time $t$ is higher if the temperature is low at time $t - 1$?
  - Adding this in the model could create more asymmetry?
A non-homogeneous MS-AR models for the temperature

\[ \text{Hidden Regime} \quad \ldots \rightarrow S_{t-1} \rightarrow S_t \rightarrow S_{t+1} \rightarrow \ldots \]

\[ \downarrow \quad \uparrow \quad \downarrow \quad \uparrow \quad \downarrow \]

\[ \text{Temperature} \quad \ldots \rightarrow Y_{t-1} \rightarrow Y_t \rightarrow Y_{t+1} \rightarrow \ldots \]

- Logistic link function for the switching probabilities. For \( s \in \{1, 2\}, \)

\[
P(S_t = s | S_{t-1} = s, Y_{t-1} = y_{t-1}) = \pi(s) + \frac{1 - \pi(s) - \pi(s)}{1 + \exp\left(\lambda_0(s) + \lambda_1(s)y_{t-1}\right)}
\]

- Linear Gaussian AR(p) model for the temperature conditionally to the weather type

\[
Y_t = \beta_0^{(S_t)} + \beta_1^{(S_t)} Y_{t-1} + \ldots \beta_p^{(S_t)} Y_{t-p} + \sigma^{(S_t)} \epsilon_t
\]

- \((\beta_i^{(s)})\), \((\sigma^{(s)})\) unknown parameters and \((\epsilon_t)\) iid sequence of \(\mathcal{N}(0, 1)\) r.v.
Model fitted with the EM algorithm

- **Regime 1**: lower temperature with slow variations (anticyclonic)
- **Regime 2**: higher temperature with low and quick variation around the mean (cyclonic)

Higher probability of staying in regime 1 if low temperature
A non-homogeneous MS-AR model for the temperature data, simulation

- The model is able to reproduce asymmetries more accurately.
A non-homogeneous MS-AR models for the temperatures

- Durations in the regimes are no more geometric (especially in the 2nd regime)

- Create more asymmetry in the dynamics
Some extensions

- The model has been generalized to handle
  - Multivariate time series
    - Applied to bivariate wind data ($(u, v)$ components)
  - Non linear autoregressive models
    - e.g. wind direction with von-Mises distribution
  - Other link functions
Theoretical framework flexible enough to include models with exogenous covariates

For example non-homogeneous HMMs used for statistical downscaling

- **Covariates**
  - $\cdots \rightarrow Z_{k-1} \rightarrow Z_k \rightarrow Z_{k+1} \rightarrow \cdots$

- **Hidden Regime**
  - $\cdots \rightarrow X_{k-1} \rightarrow X_k \rightarrow X_{k+1} \rightarrow \cdots$

- **Output time series**
  - $\cdots \rightarrow R_{k-1} \rightarrow R_k \rightarrow R_{k+1} \rightarrow \cdots$

- **Covariates** = large scale information
  - Assumed to be an ergodic Markov process

- **Output time series** = local weather conditions
  - (rainfall, temperature, ...) at a meteorological station
Some scientist argued that occurrence of longer anticyclonic conditions may be linked with sunspots.
Non-homogeneous MS-AR model with sunspots

- **Sunspots**: $\ldots \rightarrow Z_{k-1} \rightarrow Z_k \rightarrow Z_{k+1} \rightarrow \ldots$
- **Weather type**: $\ldots \rightarrow X_{k-1} \rightarrow X_k \rightarrow X_{k+1} \rightarrow \ldots$
- **Temperatures**: $\ldots \rightarrow R_{k-1} \rightarrow R_k \rightarrow R_{k+1} \rightarrow \ldots$

Logistic link function for the switching probabilities. For $s \in \{1, 2\}$,

$$P(S_t = s | S_{t-1} = s, Y_{t-1} = y_{t-1}, Z_{t-1} = z_{t-1})$$

$$= \pi^{(s)} + \frac{1 - \pi^{(s)} - \pi^{(s)}}{1 + \exp \left( \lambda_0^{(s)} + \lambda_1^{(s)} y_{t-1} + \lambda_2^{(s)} z_{t-1} \right)}$$

Monbet, Ailliot

Switching autoregressive models
Transitions with sunspots

\[ \hat{\lambda}_2^{(1)} : 1.82 \ [0.69, 3.80], \hat{\lambda}_2^{(2)} : 1.35 \ [-1.13, 4.63] \]
MS-AR with sunspots

Sunspots, 1−max per month (duration of regime 1) NH MS-AR, NH MS-AR with sunspots
Pearson’s correlation test p-value 0.03, with sunspots $10^{-3}$
MS-AR with sunspots
Spatio-temporal model at the scale of France: MS-VAR(2) with homogeneous or non-homogeneous transition probabilities (covariable = temperature in Rennes).

- Same regime for all the sites.
- Does the model allow to "observe" motion of large scale events?
Cross-correlations

Homogeneous model

Non Homogeneous model
Cross-correlations

Cold Regime

Hot Regime
Mean number of up-crossings

Homogeneous model

Non Homogeneous model
Marginal distributions

Homogeneous model

Non Homogeneous model
• MS-VAR(2) spatio temporal models for temperature
• allows to quite well reproduce
  • Marginal distributions (high temperature have to be improved...)
  • Space-time correlation with delay
  • Some non linearities ($E(N_u)$)
• The non homogeneous transition probabilities improve the results.
Conclusions

- Weather type models provide a flexible and interpretable family of models for meteorological time series
- Models have been developed/validated for generating
  - (rainfall, temperature, solar radiation, humidity, wind speed) simultaneously at a single location with a priori clustering
  - Wind speed, wind direction, temperature independently at a single location with latent clustering
  - Rainfall and temperature independently at several locations simultaneously with latent clustering
- More works needed for
  - Developing multi-site and multi-parameters generators
  - Improving some aspects of existing models
    - Interannual variability underestimated ("overdispersion" phenomenon)
    - Probability of long "events" (dry spell, heat wave,...) generally underestimated but can be improved by introducing large scale covariate (e.g. sunspots)
- R package under development
- References