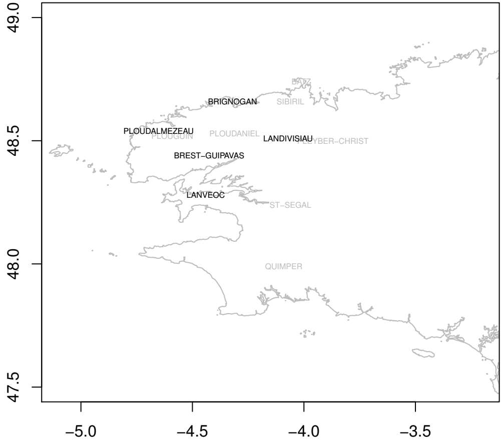
Time varying autoregressive models for multisite weather generators

V. Monbet [1], joint work with P. Ailliot[2], T. T. T. Chau [1], P. Tandéo[3]

[1] Université de Rennes 1/IRMAR & INRIA/Aspi,
[2] UBO/LMBA,
[2] Telecom

- Wind in Finistère



Multivariate
- (U_t, Φ_t): easy to interprete for meteorologists
- (u_t, v_t): easier to handle for statistical models

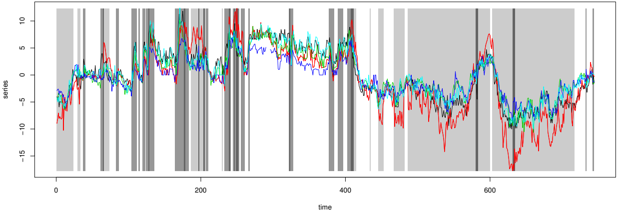
About 20 years of hourly data, missing data

Seasons
- Focus on january month

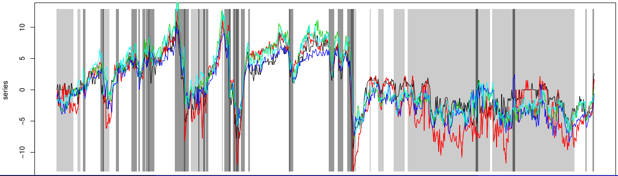
Example of multivariate time series

- The notion of regimes ou weather types is common in meteo. Different mean, volatility, spatial covariance...
- Can we model (u, v) with a finite number of regimes or should we consider a smoother model?

u component (West-East)



v component (North-South)



Local linear model

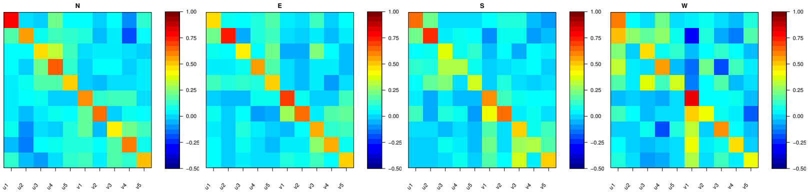
- Let us denote $X_t = (u_1(t), \cdots, u_d(t), v_1(t), \cdots, v_d(t))^T$
- General local linear model

$$X_t = A_0(X_{t-1}, S_t) + A_1(X_{t-1}, S_t)X_{t-1} + \cdots + A_p(X_{t-1}, S_t)X_{t-p} + \Sigma(X_{t-1}, S_t)^{1/2}\epsilon_t$$

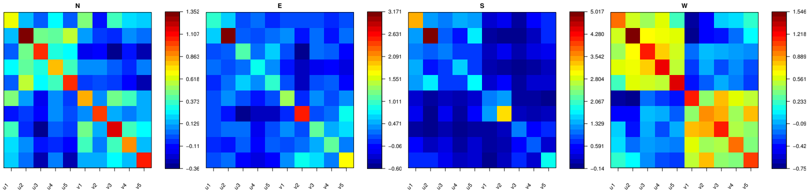
with $S_t \in \{1, \cdots, M\}$ a variable describing the regime.

- If A_0, A_1, $\cdots$, A_p and Σ are constant over time
  $\rightarrow$ Vector autoregressive model (VAR)
- If A_0, Σ only depend on X_{t-1} and $A_1 = \cdots = A_p = 0$
  $\rightarrow$ Locally constant model
- If A_0, A_1 and Σ only depend on X_{t-1}
  $\rightarrow$ Local Linear Regression model (LLR)
- If A_0, A_1 and Σ only depend on a Markov chain S_t
  $\rightarrow$ Markov Switching autoregressive model (MSAR)

- Consider X_{t-1} such that $U = 7m/s$ in each station and different directions.
- If we fit a VAR model conditionnaly to these conditions, one obtains
  - Autoregressive matrices $A_1(X_{t-1})$



  - Innovation covariance matrices $\Sigma(X_{t-1})$ (⚠ not same scale)



Models

We propose to compare the following models.

- Locally constant model
- Local Linear Regression model (LLR)
- Markov Switching autoregressive model (MSAR) (VAR is a particular case)

Locally constant model

- Algorithm: local bootstrap (LB) with a compact support kernel.
- Simulation algorithm "Analogs"
  Given a catalog of pairs of observations $\left\{(x^{(k-1)}, x^{(k)}), k = 1, \cdots, n\right\}$ such that $x^{(k)}$ is the successor of $x^{(k-1)}$ in time,

$$\hat{X}_t|\hat{X}_{t-1} = x \leftarrow \text{(weighted) bootstrap sampling in } \left\{x^{(k)}|x^{(k-1)} \in Knn(x)\right\}$$

  where $Knn(x)$ stands for the set of the K nearest neighbors of x for the euclidean distance. Weights may depend on the distance $d(x, x^{(k-1)})$

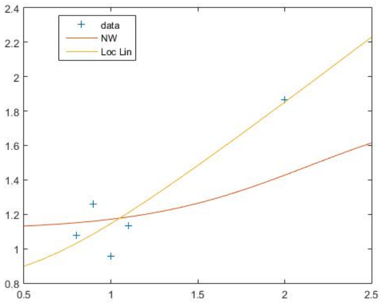
- It leads to the Nadaraya Watson estimate of the conditional mean

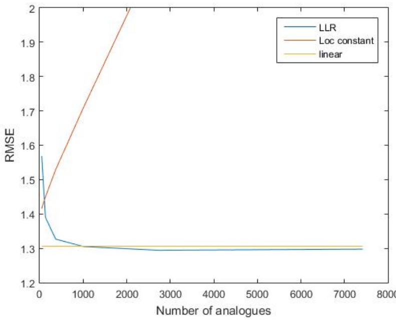
$$E(X_t|X_{t-1}) = A_0(X_{t-1}).$$

  The simulated process is Markovian.
- May be extended to higher order and covariables can be added.

Replace Local Bootstrap by Local Linear Regression?

- We expect that Local Linear Regression does better than LB
- Example : dimension 1, 4 observations





- LLR is better for forecasting (dynamics close to linear).
- Is it also better for simulating? How does it compare to MS-AR?

Local linear regression

- Simulation algorithm "LLR"
  Given a catalog of pairs of observations $\{(x^{(k-1)}, x^{(k)}), k = 1, \cdots, n\}$

  1. Estimation of the local parameters

  $$(\hat{A}_0(x), \hat{A}_1(x)) = arg \min_{A_0, A_1} \sum_{x^{(k-1)} \in Knn(x)} w_k \left\| x^{(k)} - \left(A_0 + A_1 x^{(k-1)} \right) \right\|^2$$

  with w_k weights depending on the distance $d(x^{(k-1)}, x)$.
  2. Bootstrap sampling of the innovation

  $$\hat{\epsilon} \leftarrow \text{bootstrap sampling in } \epsilon_k = x^{(k)} - \left(\hat{A}_0(x) + \hat{A}_1(x) x^{(k-1)} \right)$$

  3. Computation of the successor $\hat{X}_t$ of $\hat{X}_{t-1} = x$

  $$\hat{X}_t | \hat{X}_{t-1} = x \leftarrow \hat{A}_0(x) + \hat{A}_1(x) x + \hat{\epsilon}$$

  Alternative for residuals $\hat{\Sigma}_\epsilon \leftarrow \hat{B}_0(x) + \hat{B}_1(x) x$
- It leads to the LLR estimate of the conditional mean

  $$E(X_t | X_{t-1}) = A_0(X_{t-1}) + A_1(X_{t-1}) X_{t-1}$$

  The simulated process is Markovian.

$$\begin{array}{lccccccccccc} \textit{Weather type} & \cdots & \rightarrow & S_{t-1} & \rightarrow & S_t & \rightarrow & S_{t+1} & \rightarrow & \cdots \\ & & & \downarrow & \nearrow & \downarrow & \nearrow & \downarrow & & \\ \textit{Wind} & \cdots & \rightarrow & X_{t-1} & \rightarrow & X_t & \rightarrow & X_{t+1} & \rightarrow & \cdots \end{array}$$

- **Weather type** - Hidden weather type modeled as a first order Markov chain.
  Homogeneous transitions

  $$P(S_t = s_t | S_0 = s_0, \cdots, S_{t-1} = s_{t-1}, X_1 = x_1, \cdots, X_{t-1} = x_{t-1}) = P(S_t = s_t | S_{t-1} = s_{t-1})$$

  Non homogeneous transitions

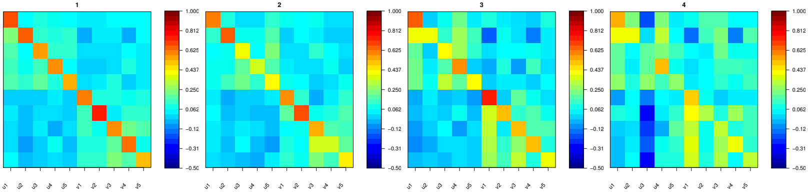
  $$\begin{aligned} & P(S_t = s_t | S_0 = s_0, \cdots, S_{t-1} = s_{t-1}, X_1 = x_1, \cdots, X_{t-1} = x_{t-1}) \\ & = P(S_t = s_t | S_{t-1} = s_{t-1}, X_{t-1} = x_{t-1}) \end{aligned}$$

- **Wind** - Linear Gaussian AR(p) model for the wind evolution conditionally to the weather type

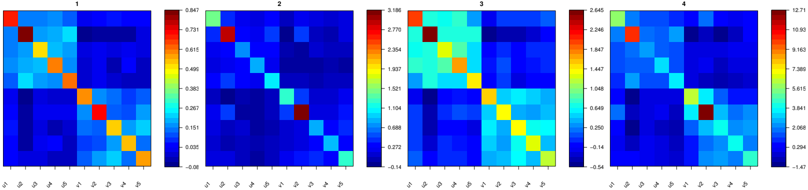
  $$Y_t = A_0(S_t) + A_1(S_t) X_{t-1} + \cdots + A_p(S_t) X_{t-p} + \Sigma(S_t)^{1/2} \epsilon_t$$

  $(A_i(s))$, $(\Sigma(s))$ for $s \in 1, \cdots, M$ unknown parameters and $\{\epsilon_t\}$ iid $\mathcal{N}(0,1)$ sequence
- **Inference**: Maximum Likelihood

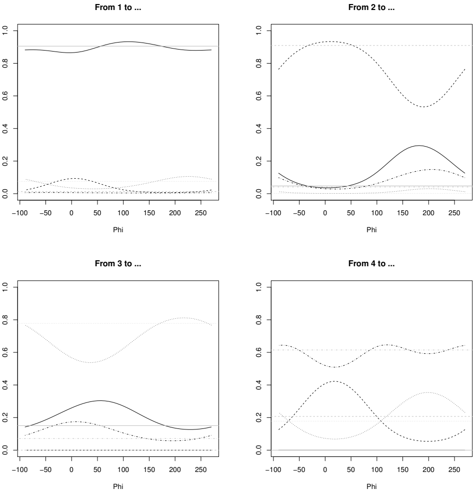
- Autoregressive matrices $A_1(s)$
  Regime 1 : wind blowing from the West; regime 2: wind blowing from the East; regime 3 and 4: larger volatility.



- Innovation covariance matrices $\Sigma(s)$ (⚠ not same scale)
  Regime 2 and 4 have high covariance of innovations. Regime 1 and 3: smaller covariance but larger spatial scale.



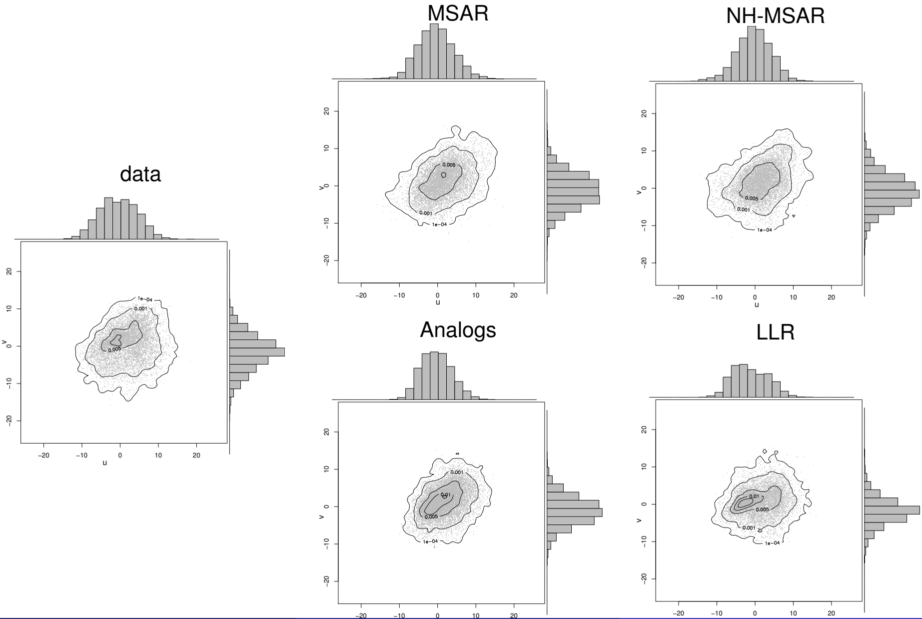
- Transition probabilities depends of the wind direction at Ploudamézeau (westliest station)



Stochastic weather generator

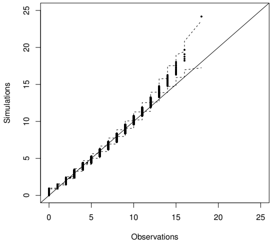
- One simulates sequences of (u, v) simultaneously at the 5 stations.
- Computational time
  - LLR $\sim$ 14 s for one month of January (hourly data). The computation of the distance and the sort are expensive. It can be improved.
  - MSAR (4 regimes, order 2) $\sim$ 40s to calibrate the model (done only one time), then 0.04 s for one month of January (hourly data).
- To validate the SWGEN one compares several statistics of the observations to the one of the data. The considered station is the westliest (Ploudalmézeau). The performances at other stations are similar.
  - the marginal distribution of (u, v) and U
  - spatial covariance
  - auto correlation of U and cross-correlation of U at Ploudalmézeau (westliest station) and U at Landivisiau (eastliest station)
  - cross-correlation of $\cos(\Phi$ and $\sin(\Phi)$
  - intensity of up-crossing for u and U

Marginal distributions (u,v), Ploudalmézeau

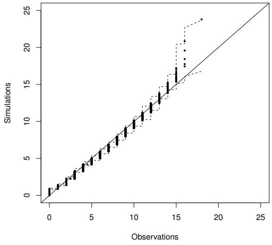


Marginal distribution U, Ploudalmézeau

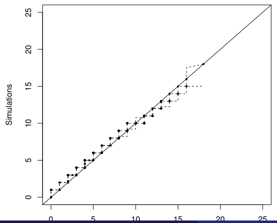
MSAR



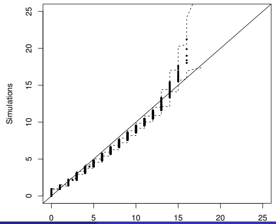
NH-MSAR



Analogs

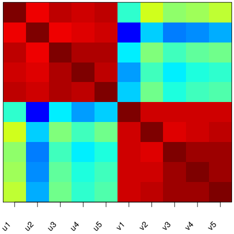


LLR



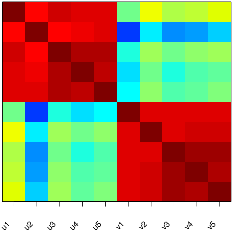
data

Data



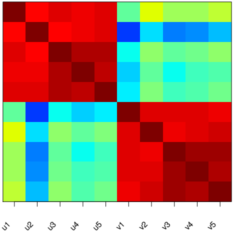
MSAR

Simulations



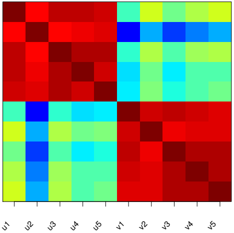
NH-MSAR

Simulations



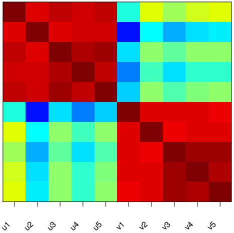
Analogs

Simulations



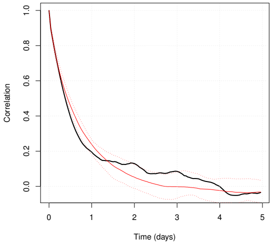
LLR

Simulations

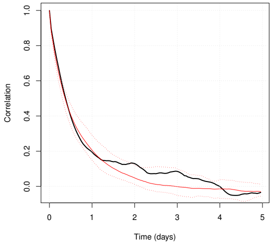


Correlation U, Ploudalmézeau

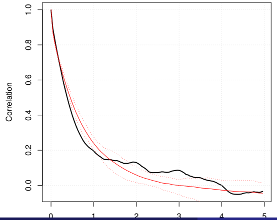
MSAR



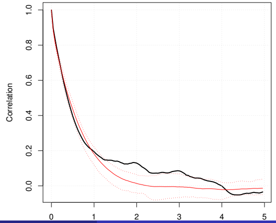
NH-MSAR



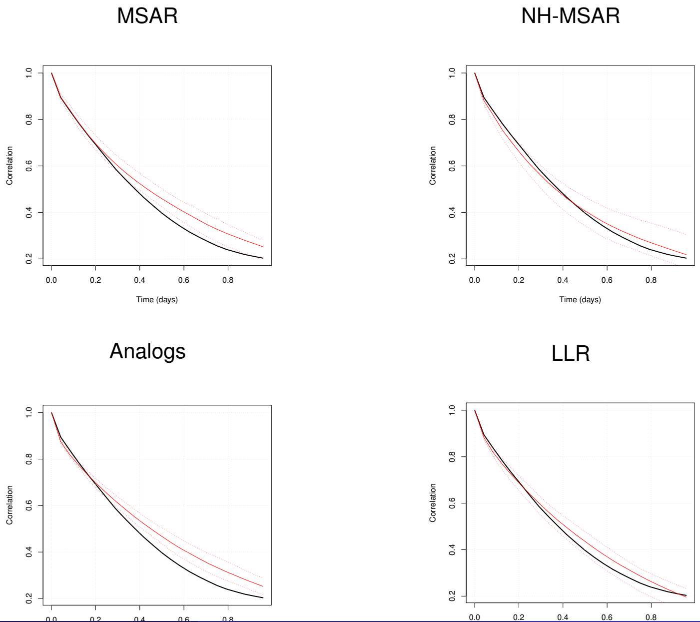
Analogs



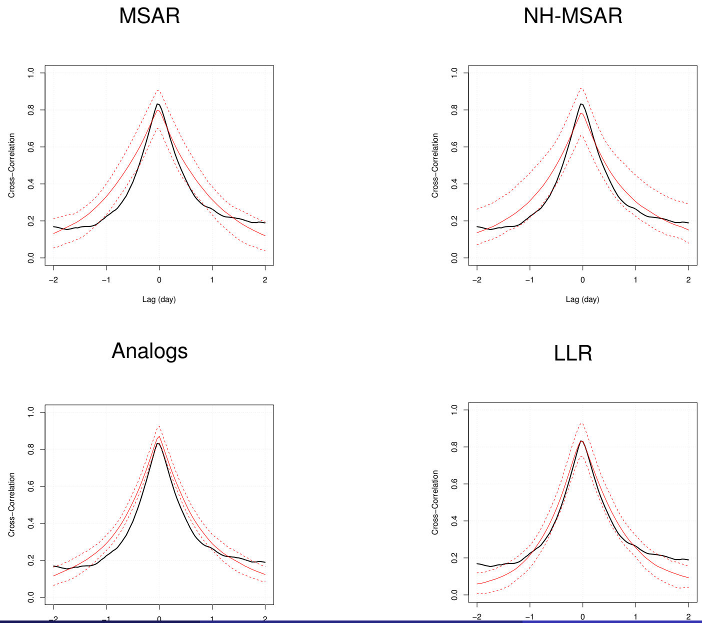
LLR



Correlation U, Ploudalmézeau - zoom

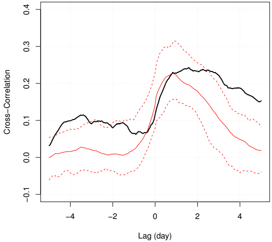


Cross-correlations, U, Ploudalmézeau-Landivisiau

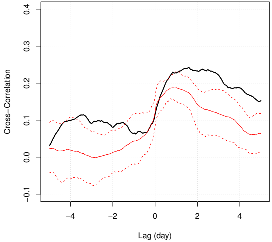


Cross-correlation (cos(Φ), sin(Φ)), Ploudalmézeau

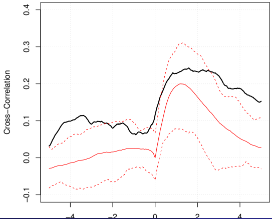
MSAR



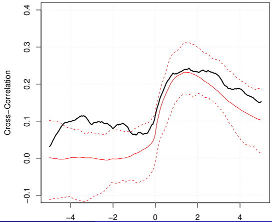
NH-MSAR



Analogs

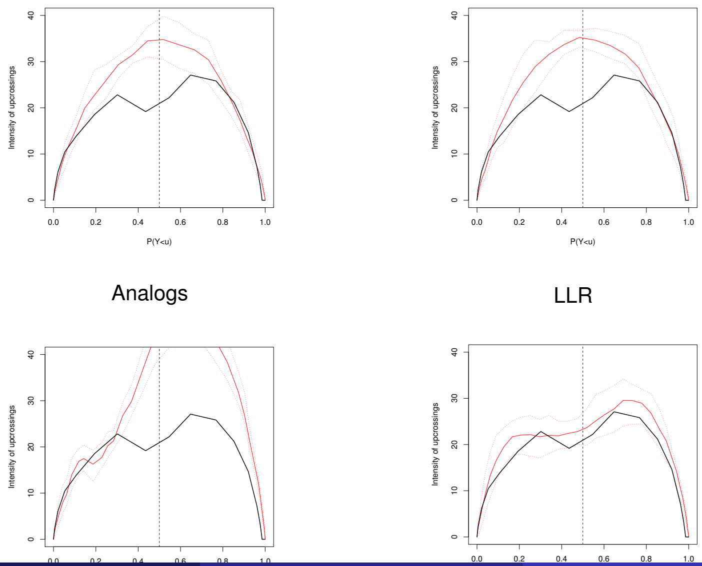


LLR



MSAR

NH-MSAR



Analogs

LLR

Models comparison

- **Inference**
  - Statistical properties of Analogs are well known: local bootstrap.
  - LLR only a few results for multivariate time series.
  - MSAR provides a proper model with well known statistical properties ans tools for inference.
- **Model/Interpretability**
  - Analogs and LLR are non parametric models difficult to interpret.
  - Regimes and parameters of MSAR models have physical interpretation.
- **Robustness**
  - LLR needs less calibration than MSAR.
  Easy to introduce covariables and to handle missing values.
  It may be difficult and time consuming to optimize the algorithm parameters. What is the good criteria for simulation?
  If the dimension increases, some work is needed to choose an appropriate distance.
  Ridge shrinkage easy to implement. Lasso?
  - MSAR has been fitted to many weather variables time series (ex: wind, temperature, rain).
  Possiblity to introduce covariables in the intercept and in the transitions.
  If the dimension increases, sparse versions can be used (Lasso).
- **Computational time**
  - MSAR is much faster.

Models comparison

- **Simulation performances**
  - Analogs algorithm reproduces the marginal distribution, but produces no new conditions.
  - LLR better reproduces the dynamics but not as good for the marginal distribution.
  - (NH) MSAR models are quite good in reproducing the second order structure but they would need a control of the marginal distribution (ABC algorithms).

	Analogs	LLR	MSAR	NH-MSAR
(u, v) pdf	+	++	+	++
U pdf	++			+
$C_{spat.}$	+	+	++	++
$C_{(U)}$		++		++
$C_{(UU)}$		++		+
$C_{(\cos(\Phi)\sin(\Phi))}$		++		
$E(N)$		++		

- Initial question : " Can we model (u, v) with a finite number of regimes or should we consider a smoother model? "
  It is difficult to give a clear-cut answer: NH-MSAR is somewhere betwenn MSAR and LLR and gives good results.