Geostatistics for space-time analysis

From kriging to ensemble Kalman filtering

Hans Wackernagel[1] & Laurent Bertino[2]

[1]MINES ParisTech, [2]Nansen Environmental and Research Center

Workshop on *Spatiotemporal modelling with Gaussian processes*
University of Sheffield, January 2014

http://hans.wackernagel.free.fr

Space-time Geostatistics

Central question:

Is a physical model describing the time evolution of the system available?

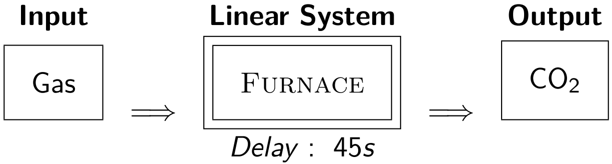
Topics:

- Covariance functions (space-time, multivariate)
- Kriging, filtering
- Geostatistical simulation
- Ensemble Kalman filtering

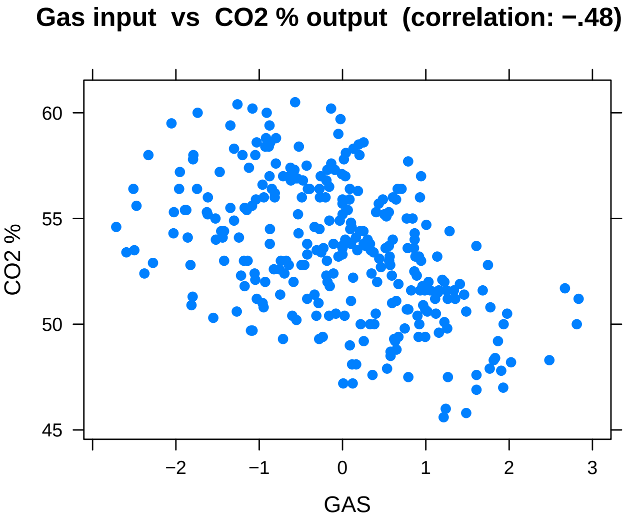
Now, a small warm-up example...

Gas furnace data (Box & Jenkins, 1970)

Gas input variation and percentage of CO2 in output have been measured every 9 seconds for a gas furnace.

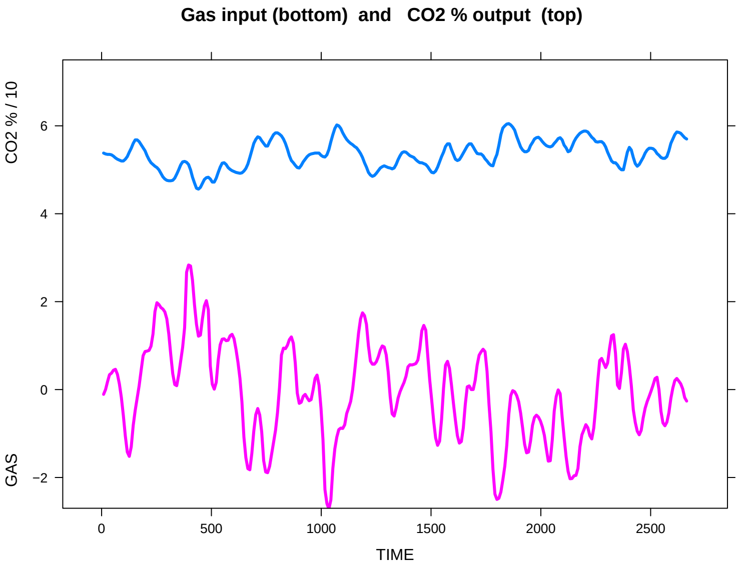


Diagram



Correlation is clearly negative, but does not seem very strong.

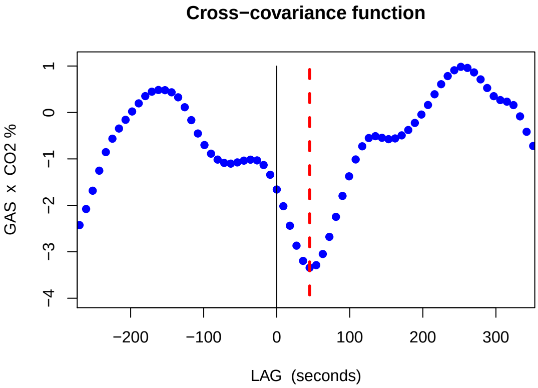
Time series of input and output



Cross-covariance function: shows the delay effect

Stationary random functions $Z_i(t)$ and $Z_j(t)$:

$$C_{ij}(\tau) = E[(Z_i(t) - m_i) \times (Z_j(t+\tau) - m_j)]$$



A 45s delay between GAS input variation and its effect on CO2 output !

Space-time covariance: simplifying assumptions

Let $Z(\mathbf{x}, t)$ with $(\mathbf{x}, t) \in \mathbb{R}^d \times \mathbb{R}$ be a space-time random function.
The following simplifying assumptions about the space-time covariance are useful in applications:

- Separability:

$$\mathrm{cov}(Z(\mathbf{x}_1, t_1), Z(\mathbf{x}_2, t_2)) = C_S(\mathbf{x}_1, \mathbf{x}_2) \cdot C_T(t_1, t_2)$$

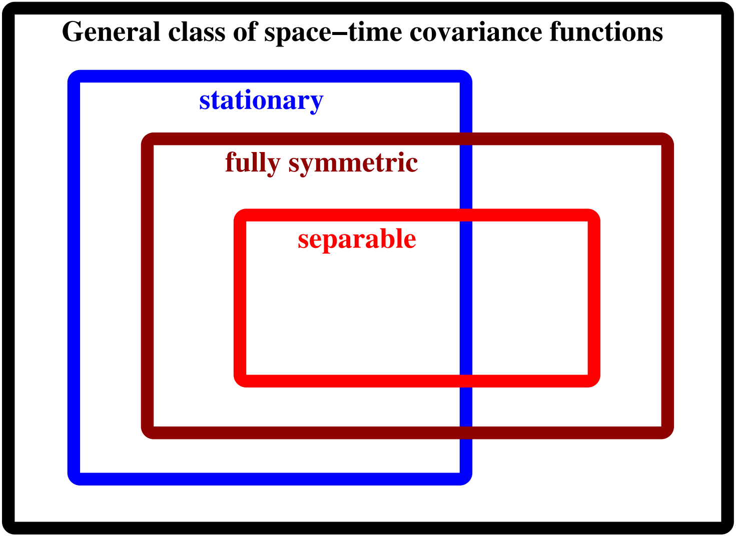
- Full symmetry:

$$\mathrm{cov}(Z(\mathbf{x}_1, t_1), Z(\mathbf{x}_2, t_2)) = \mathrm{cov}(Z(\mathbf{x}_1, t_2), Z(\mathbf{x}_2, t_1))$$

- Stationarity (translation invariance):

$$\mathrm{cov}(Z(\mathbf{x}_1, t_1), Z(\mathbf{x}_2, t_2)) = C(\mathbf{x}_1 - \mathbf{x}_2, t_1 - t_2)$$

Imbrication of the assumptions



From Gneiting, Genton & Guttorp (2007)

Irish wind case study - Gneiting, Genton & Guttorp (2007)

- Winds in Ireland are predominantly westerly, so that the velocity measures propagate from west to east.
- Temporal correlations lead or lag between W and E stations at a daily scale.
- Exploratory analysis shows a lack of full symmetry and thereby of separability in the correlation structure of the velocities.
- Fitting different parametric models: separable, fully symetric but not separable, stationary but not fully symmetric.
- Space-time simple kriging results show the best performance with the general stationary model in terms of four different preformance measures.

Separable mean field

In case of a non-stationary random function it is also possible to consider a separable mean field:

$$M(\mathbf{x}, t) = M(\mathbf{x}) + M(t)$$

To model the diurnal fluctuation of the magnetic field, Séguret & Huchon (1990) use a finite trigonometric expansion of the form:

$$M(t) = \sum_i A_i \cos(\omega, t) + \sum_i B_i \sin(\omega, t)$$

where

- ω_i are fixed angular frequencies (e.g. $2\pi/24$ for the daily cycle and t in hours),
- A_i, B_i are unknown (possibly random) coefficients.

Earth magnetism

Séguret & Huchon, JGR 1990

- Magnetic anomalies are essential to study earth history.
- Magnetism is influenced by several external factors like:
  - solar wind explaining daily fluctuations (period: 24 hours)
  - rotation of the moon around the earth (period: 28 days)
  - solar perturbartions (half-year cycle)

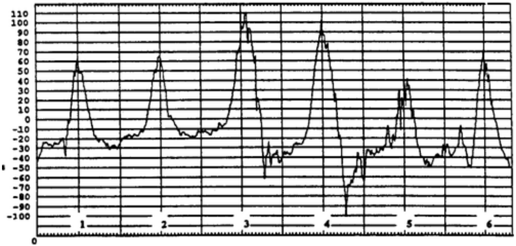
Available data:

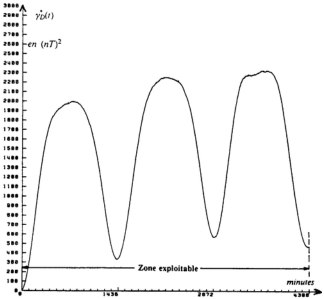
SEAPERC campaign (Ifremer, 1986) Data from a research vessel about magnetism over a fractured area of 111 km^2 off Peru.

Fluctuations of earth magnetism Measurements at a Peruvian observatory for the time period of the campaign.

Daily fluctuation of earth magnetism

Huancayo observatory (Peru): 22 to 28/08/1986



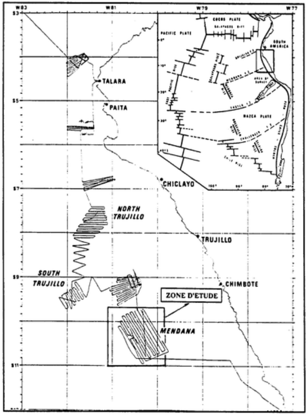


Time series (6 days)

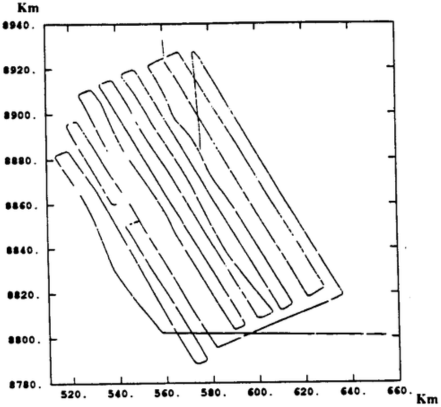
Variogram

SEAPERC campaign

Ship moves along a profile in 12 hours

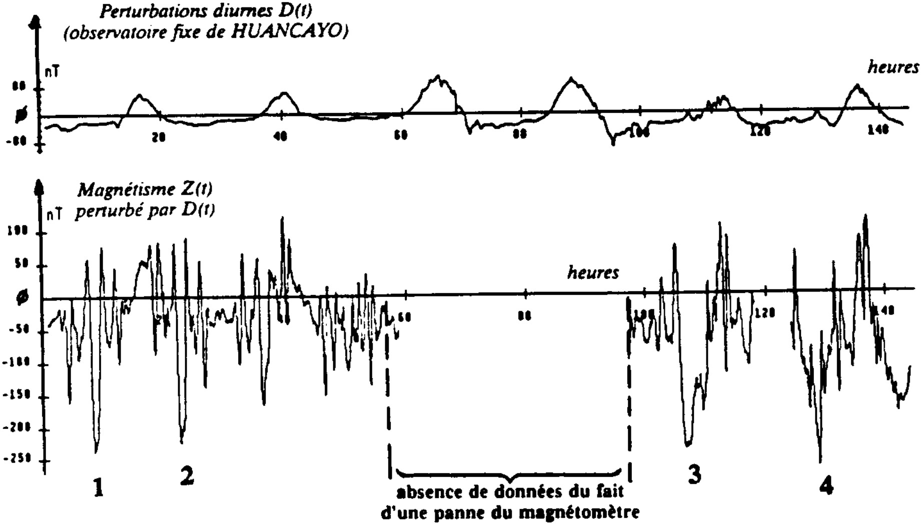


Map



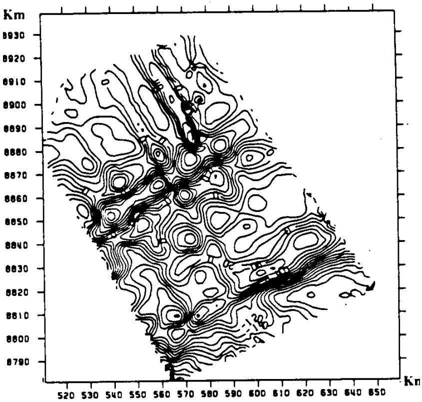
Study area

Measurements at observatory and along ship route

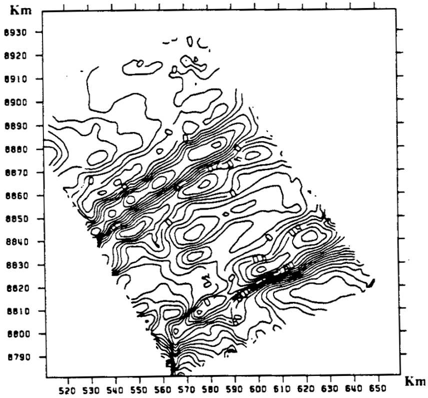


Filtering the daily fluctuations of magnetism

Space-time model: $Z(\mathbf{x}, t) = Y(\mathbf{x}) + m(t)$



With time perturbations



After geostatistical filtering

Simple kriging: filtering of measurement error

Simple kriging estimate:

$$z_x = m + (c_x + c_x^o)(C + C^o)^{-1}(z_\alpha - m)$$

where c_x^o, C^o are a vector and a matrix of observational error covariances (white or red noise).
Simple kriging is an exact interpolator.

Filtering of measurement error (by removing c_x^o):

$$\begin{aligned} y_x &= m + (c_x \qquad)(C + C^o)^{-1}(z_\alpha - m) \\ &= m + k\,(z_\alpha - m) \end{aligned}$$

Krige gain:

$$k = (c_x \qquad)(C + C^o)^{-1}$$

(also called the *kriging weights* vector in geostatistics)

Simple kriging: filtering of measurement error

Simple kriging estimate:

$$z_x = m + (c_x + c_x^o)(C + C^o)^{-1}(z_\alpha - m)$$

where c_x^o, C^o are a vector and a matrix of observational error covariances (white or red noise).
Simple kriging is an exact interpolator.

Filtering of measurement error (by removing c_x^o):

$$\begin{aligned} y_x &= m + (c_x \qquad)(C + C^o)^{-1}(z_\alpha - m) \\ &= m + k\,(z_\alpha - m) \end{aligned}$$

Krige gain:

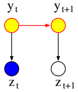
$$k = (c_x \qquad)(C + C^o)^{-1}$$

(also called the *kriging weights* vector in geostatistics)

Linear Kalman filter

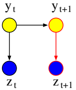
The recursions are initialized with $\hat{y}_{0|-1} = 0$ and $C_{0|-1} = C_0$.

The recursion equations:



Propagation

$$\begin{aligned} \hat{y}_{t+1|t} &= F\hat{y}_{t|t} \\ C_{t+1|t} &= FC_{t|t}F^\top + GQG^\top \end{aligned}$$



Update

$$\begin{aligned} \hat{y}_{t+1|t+1} &= \hat{y}_{t+1|t} + K_{t+1}(z_{t+1} - H\hat{y}_{t+1|t}) \\ C_{t+1|t+1} &= C_{t+1|t} - K_{t+1}HC_{t+1|t} \end{aligned}$$

Kalman gain:

$$K_{t+1} = (C_{t+1|t} \qquad)H^\top (HC_{t+1|t}H^\top + C^o)^{-1}$$

We note that observational error is filtered in the update (kriging) step. The observational error covariance matrix C^o does not appear in the numerator of the Kalman gain.

Ensemble Kalman filter

Evensen (1994)

In the Ensemble Kalman filter (EnKF) the non-linear dynamics are propagated by Monte-Carlo simulation.

This amounts to approximate the forecast distribution $F(\mathbf{y}_t|\mathbf{z}_{1:t-1})$ by an ensemble of N members $\mathbf{y}_t^{f,i}$.

Propagation

$$\{ \quad \mathbf{y}_t^{f,i} = \mathcal{M}(\mathbf{y}_{t-1}^{\star,i}, \mathbf{u}_t^i); \qquad i = 1, \ldots, N \quad \}$$

Update

$$\{ \quad \mathbf{y}_t^{\star,i} = \mathbf{y}_t^{f,i} + \mathbf{K}_t(\mathbf{z}_t - \mathcal{H}\,\mathbf{y}_t^{f,i} + \mathbf{u}_t^{o,i}); \qquad i = 1, \ldots, N \quad \}$$

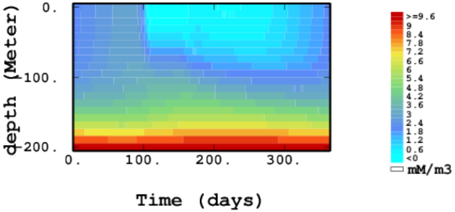
- The covariance matrices C_t^f of the forecasts $\mathbf{y}_t^{f,i}$ and C_t^o of the observation errors $\mathbf{u}_t^{o,i}$ are used to set up the Kalman gain $\mathbf{K}_t$.
- The covariance matrix $C_t^\star$ of the updated state vectors $\mathbf{y}_t^{\star,i}$ is computed directly on the ensemble.

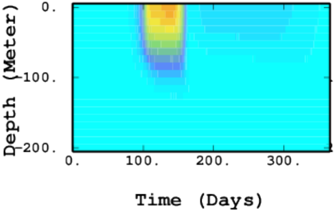
Properties of the Ensemble Kalman filter (EnKF)

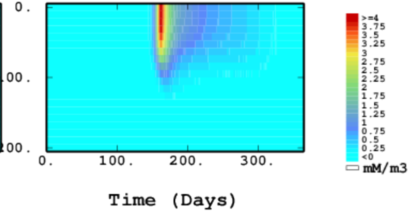
- The EnKF with an infinite ensemble will yield in the limit the same result as the linear KF.
- The EnKF is not a pure resampling of a Gaussian posterior: only the updates are linear and these are added to the prior non-Gaussian ensemble.
- The updated ensemble will thus inherit of many of the non-Gaussian properties from the forecast ensemble.
- In summary: the analysis in the EnKF is computationally efficient and avoids resampling of the posterior. The solution is midway between a linear update and a full Bayesian computation.

1D Ecological model

Nutrients







Phytoplankton Herbivores

EnKF: comparison

We test the performance of data assimilation with perturbed samples taken from the numerical output.

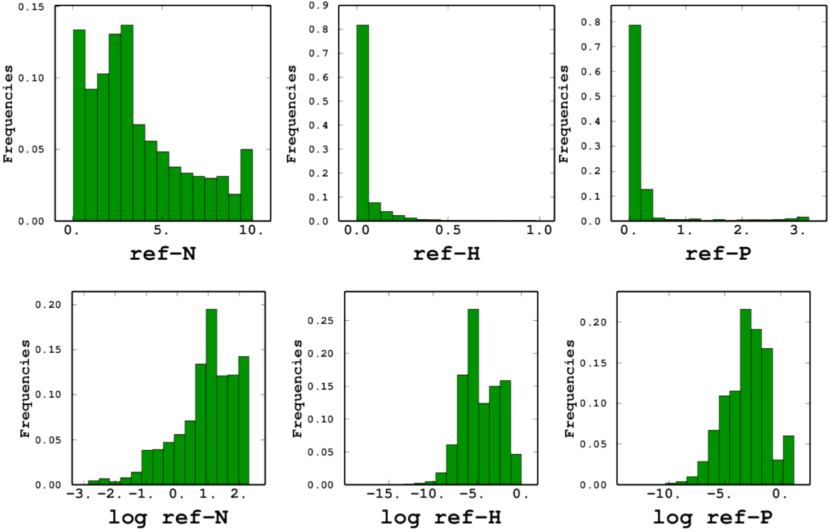
Data: samples taken every 10 days,
perturbed with a white noise.

Ensemble: 100 members.

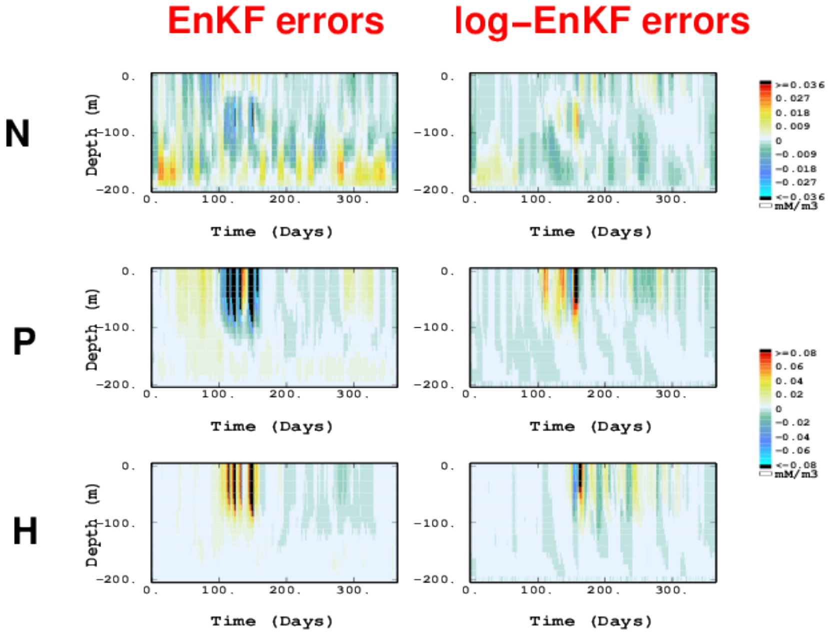
Comparison: EnKF with/without logarithmic transformation.

Logarithmic transformation

Nutrients, Herbivores, Phytoplankton

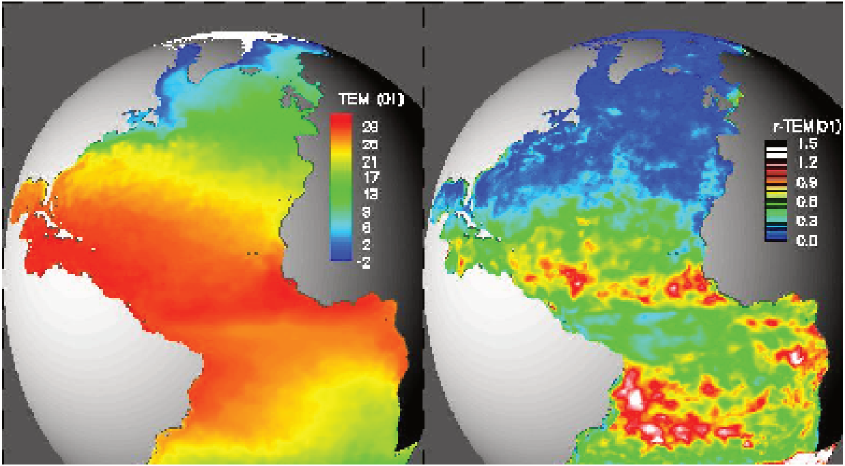


EnKF runs with and without anamorphosis



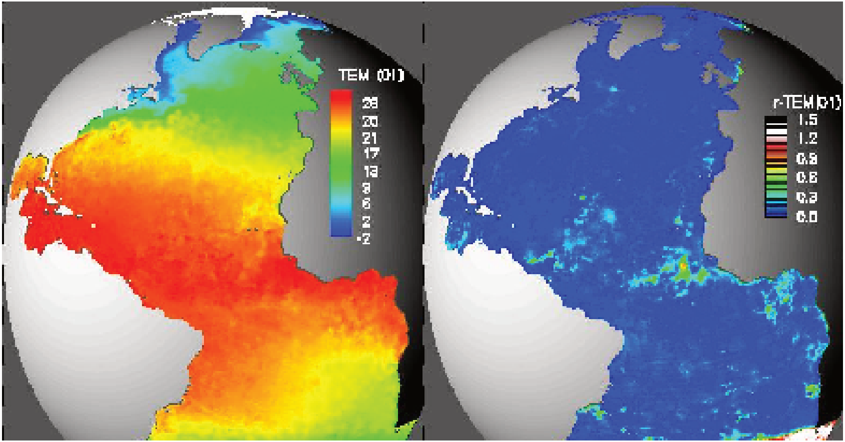
Atlantic: sea-surface temperature, C_t^f variance

http://topaz.nersc.no



Atlantic: sea-surface temperature, $C_t^\star$ variance

http://topaz.nersc.no



Summary

- The choice of a space-time method depends on whether or not a physical model describing the time evolution of the system is available.
- Exploratory analysis of the covariance structure may suggest simplifications, which will impact the numerical effort to produce forecasts, nowcasts or reanalysis.
- State-space models provide the framework for various flavours of Kalman filters for systems with nonlinear dynamics.

References

L Bertino, G Evensen, and H Wackernagel.
Sequential data assimilation techniques in oceanography.
International Statistical Review, 71:223–241, 2003.

J P Chilès and P Delfiner.
Geostatistics: Modeling Spatial Uncertainty.
Wiley, New York, 2nd edition, 2012.
(See in particular section 5.8, pp370–385, on Space-Time Models).

G Evensen.
Data Assimilation: the Ensemble Kalman Filter.
Springer, Berlin, 2007.

T Gneiting, M G Genton, and P Guttorp.
Geostatistical space-time models, stationarity, separability and full symmetry.
In B Finkenstaedt, L Held, and V Isham, editors, *Statistics of Spatio-Temporal Systems*, pages 151–175. CRC Press, 2007.

C Lantuéjoul.
Geostatistical Simulation: Models and Algorithms.
Springer-Verlag, Berlin, 2002.

P Sakov, F Counillon, L Bertino, K A Lisaeter, P R Oke, and A Korablev.
TOPAZ4: an ocean-sea ice data assimilation system for the North Atlantic and Arctic.
Ocean Sci., pages 633–656, 2012.

S Séguret and P Huchon.
Trigonometric kriging: a new method for removing the diurnal variation from geomagnetic data.
J. Geophysical Research, 32(B13):21.383–21.397, 1990.

H Wackernagel.
Multivariate Geostatistics: an Introduction with Applications.
Springer-Verlag, Berlin, 3rd edition, 2003.

Georges Matheron (1930-2000)

1955 First paper on geostatistics in *Annales des Mines.*

1962 *Traité de Géostatistique Appliquée*, BRGM (Technip).

1965 *Les variables régionalisées et leur estimation: une application de la théorie des fonctions aléatoires aux sciences de la nature*, Masson.

1967 *Eléments pour une théorie des milieux poreux*, Masson.

1968 Creation of Centre de Morphologie Mathématique at Ecole des Mines.

1970 The Theory of Regionalized Variables and its Applications, Cahiers du CMM (Oxford, 2014).

1975 *Random sets and integral geometry*, Wiley.

1978 *Estimating and choosing* (Springer, 1989; Presses des Mines, 2013).

1995 Retirement.

See also our electronic library:
`http://cg.ensmp.fr/bibliotheque`

Software

Geostatistics (free): `http://RGeoS.free.fr`

- package RGeoS, which runs in R (open source) available at: `http://www.r-project.org`

  By the way, R can be used in a Matlab-like graphical environement by installing additionnally: `http://www.rstudio.com/ide`

  RGeoS runs with N spatial coordinates.

Geostatistics (commercial): `http://www.geovariances.com`

- especially the standalone general purpose software Isatis.

Ensemble Kalman filter (free): `http://EnKF.nersc.no`

- code in Fortran 90 as well as a Matlab package.

Acknowledgements

- Thank you for inviting me to GPWS14 !
- Research on these topics will be pursued in a new franco-scandinavian project funded by NordForsk: *Ensemble-based Methods for Environmental Monitoring and Prediction* (2014-2018).