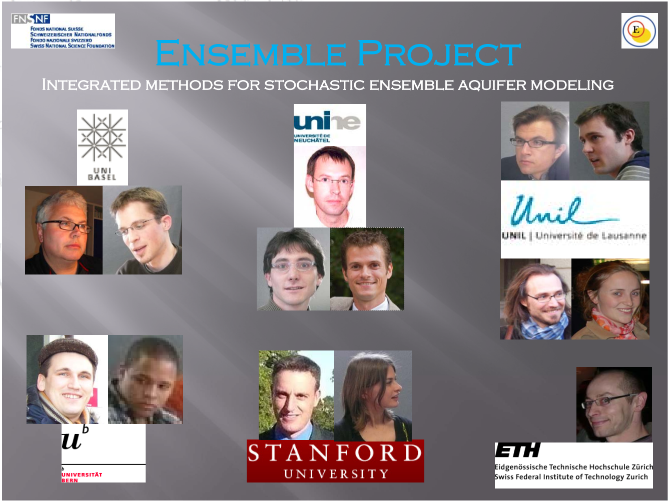
Méthodes de distances et métamodèles sur base de proxys pour la planification d'expériences

David Ginsbourger (Univ. Berne)

École de Physique des Houches

10 Avril 2013

Acknowledgements: "ENSEMBLE" project



Partly based on a paper with high-quality co-authors!

- Bastien Rosspopoff (was at Uni Bern)
- Guillaume Pirot (Uni Neuchâtel)
- Nicolas Durrande (Sheffield)
- Philippe Renard (Uni Neuchâtel)

Motivations of MDS and other distance methods

Given a sample of n high-dimensional and/or complicated "objects" $x_1, \ldots, x_n$ (say in a set E, e.g. $E \subset \mathbb{R}^p$ with $p >> 1$), and a "distance" (or *similarity measure*) on E, how to summarize this sample using low-dimensional, visualizable, representations?

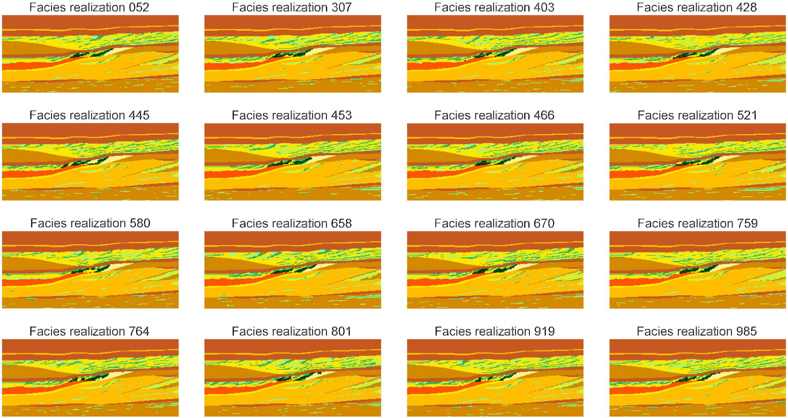
Motivations of MDS and other distance methods

Given a sample of n high-dimensional and/or complicated "objects" $x_1, \ldots, x_n$ (say in a set E, e.g. $E \subset \mathbb{R}^p$ with $p >> 1$), and a "distance" (or *similarity measure*) on E, how to summarize this sample using low-dimensional, visualizable, representations?

A few applications of distance methods (dixit Wikipedia!)

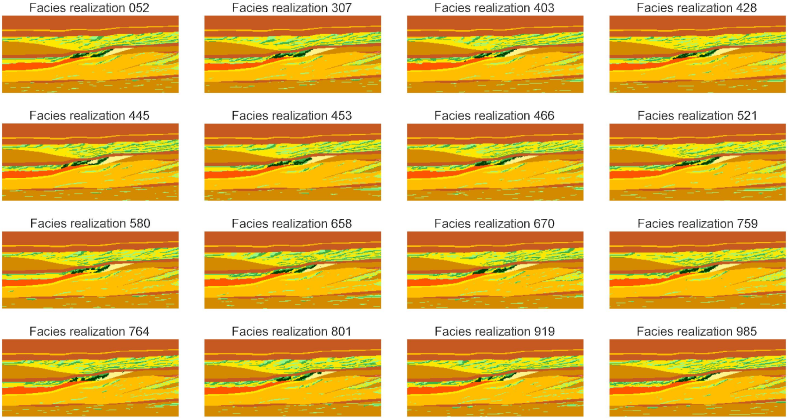
- Archeology: grouping items found in different search places into objects from the same period/place/dynasty
- Biology: constructing a phylogenetic tree based on sequences
- Marketing: representing preferences and perceptions of customers
- Geostatistics: diverse appl., e.g. modeling the variability of geological facies...

Motivating geostatistical application



First question: how to select a few ”representative ones”?

Motivating geostatistical application



First question: how to select a few ”representative ones”?

End motivation: which one(s) correspond(s) best to reality?

Outline

1. Classical multidimensional scaling: background
2. An application of MDS in stochastic hydrology
3. Proxy-based kriging and the ProKSI algorithm

Outline

1. Classical multidimensional scaling: background
2. An application of MDS in stochastic hydrology
3. Proxy-based kriging and the ProKSI algorithm

MDS consists in using pairwise distances (or *dissimilarities*) to set up an approximate representation of the x_i's in a low-dimensional Euclidean space.

Definition

An $(n \times n)$ matrix $\mathbf{D}$ is called a distance matrix if it is symmetric and

$$d_{i,i} = 0, \quad d_{i,j} \geq 0 \;\; i \neq j$$

MDS consists in using pairwise distances (or *dissimilarities*) to set up an approximate representation of the x_i's in a low-dimensional Euclidean space.

Definition

An $(n \times n)$ matrix $\mathbf{D}$ is called a distance matrix if it is symmetric and

$$d_{i,i} = 0, \quad d_{i,j} \geq 0 \quad i \neq j$$

Starting with a distance matrix $\mathbf{D}$, MDS aims at finding points $u_1, \ldots, u_n$ of the k-dimensional Euclidean space such that the distance matrix with entries

$$d_{\mathbb{R}^k}(u_i, u_j),$$

where $d_{\mathbb{R}^k}$ is the Euclidean distance over $\mathbb{R}^k$, is close (in some sense) to $\mathbf{D}$.

How does it work? Some theoretical results

Definition

A distance matrix **D** is called *Euclidean* if $\exists$ points $u_1, \ldots, u_n$ in a Euclidean space $\mathbb{R}^k$ (for some k) whose interpoint distances are given by **D**:

$$d_{i,j}^2 = (u_i - u_j)^T (u_i - u_j)$$

How does it work? Some theoretical results

Definition

A distance matrix **D** is called *Euclidean* if $\exists$ points $u_1, \ldots, u_n$ in a Euclidean space $\mathbb{R}^k$ (for some k) whose interpoint distances are given by **D**:

$$d_{i,j}^2 = (u_i - u_j)^T (u_i - u_j)$$

Let us set a few notations. A distance matrix **D** being fixed, let A be defined by

$$a_{i,j} = -\frac{1}{2} d_{i,j}^2$$

Furthermore, set

$$\mathbf{B} = \mathbf{HAH}$$

where $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the $(n \times n)$ *centring matrix*.

How does it work? Some theoretical results

Theorem

D is Euclidean if and only if **B** is positive semi-definite (p.s.d.). In particular:

How does it work? Some theoretical results

Theorem

D is Euclidean if and only if **B** is positive semi-definite (p.s.d.). In particular:

a) If **D** is a matrix of Euclidean interpoint distances for a configuration $\{u_1, \dots, u_n\} \in (\mathbb{R}^k)^n$, then

$$b_{i,j} = (u_i - \bar{u})^T (u_j - \bar{u})$$

whereof $\mathbf{B} = (\mathbf{HU})(\mathbf{HU})^T \geq 0$.

How does it work? Some theoretical results

Theorem

D is Euclidean if and only if **B** is positive semi-definite (p.s.d.). In particular:

a) If **D** is a matrix of Euclidean interpoint distances for a configuration $\{u_1, \dots, u_n\} \in (\mathbb{R}^k)^n$, then

$$b_{i,j} = (u_i - \bar{u})^T (u_j - \bar{u})$$

whereof $\mathbf{B} = (\mathbf{HU})(\mathbf{HU})^T \geq 0$.

b) If **B** is p.s.d. of rank k, denote $v_1, \dots, v_k$ its k first eigenvectors, normalized by their corresponding eigenvalues $\lambda_1, \dots, \lambda_k > 0$. Then the points $u_i = (v_{1,i}, \dots, v_{k,i}) \in \mathbb{R}^k$ $(1 \leq i \leq n)$ have interdistances given by **D**.

K.V. Mardia, J.T. Kent, and J.M. Bibby
Multivariate Analysis (1979)

A first example: US Flying Mileage

	Atl	Chi	Den	Hou	LA	Mia	NY	SF	Sea	DC
Atl	0	587	1212	701	1936	604	748	2139	2182	543
Chi	587	0	920	940	1745	1188	713	1858	1737	597
Den	1212	920	0	879	831	1726	1631	949	1021	1494
Hou	701	940	879	0	1374	968	1420	1645	1891	1220
LA	1936	1745	831	1374	0	2339	2451	347	959	2300
Mia	604	1188	1726	968	2339	0	1092	2594	2734	923
NY	748	713	1631	1420	2451	1092	0	2571	2408	205
SF	2139	1858	949	1645	347	2594	2571	0	678	2442
Sea	2182	1737	1021	1891	959	2734	2408	678	0	2329
DC	543	597	1494	1220	2300	923	205	2442	2329	0

A first example: US Flying Mileage

	Atl	Chi	Den	Hou	LA	Mia	NY	SF	Sea	DC
Atl	0	587	1212	701	1936	604	748	2139	2182	543
Chi	587	0	920	940	1745	1188	713	1858	1737	597
Den	1212	920	0	879	831	1726	1631	949	1021	1494
Hou	701	940	879	0	1374	968	1420	1645	1891	1220
LA	1936	1745	831	1374	0	2339	2451	347	959	2300
Mia	604	1188	1726	968	2339	0	1092	2594	2734	923
NY	748	713	1631	1420	2451	1092	0	2571	2408	205
SF	2139	1858	949	1645	347	2594	2571	0	678	2442
Sea	2182	1737	1021	1891	959	2734	2408	678	0	2329
DC	543	597	1494	1220	2300	923	205	2442	2329	0

Can we recover a map of the USA from that distance matrix?

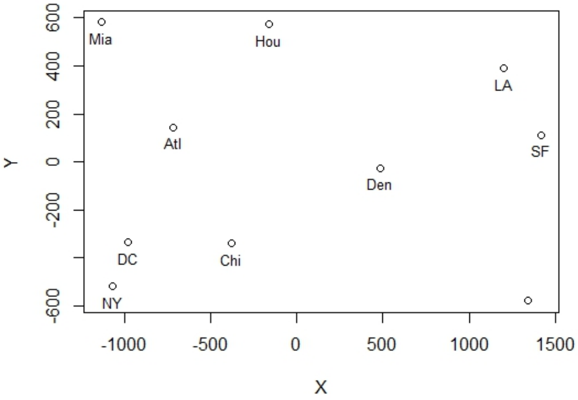
Source: Click here (website in French! :-)

```
D <- read.csv2("FlyingMileage.csv")
```

```
library(MASS)
res <- cmdscale(D[,2:11])

Y <- res[,2]
X <- res[,1]

plot(X,Y,type="p")
text(X,Y,D[,1],pos=1,cex=0.8)
```



How does it work? A practical algorithm

Given a distance matrix **D** (Euclidean or not), a classical solution to the MDS problem in p dimensions is summarized below:

a) Form **D** construct $\mathbf{A} = (-\frac{1}{2}d_{i,j}^2)$

b) Obtain **B** with elements $b_{i,j} = a_{i,j} - \bar{a}_{i,\cdot} - \bar{a}_{\cdot,j} + -\bar{a}_{\cdot,\cdot}$

c) Find the p largest eigenvalues of **B**, and the corresponding (normalized) eigenvectors $v_1, \ldots, v_p$.

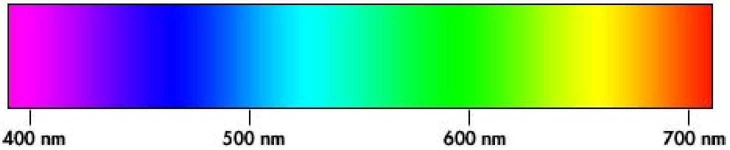
d) The required points are given by $u_i = (v_{1,i}, \ldots, v_{p,i}) \in \mathbb{R}^p$ $(1 \leq i \leq n)$

K.V. Mardia, J.T. Kent, and J.M. Bibby

Multivariate Analysis (1979)

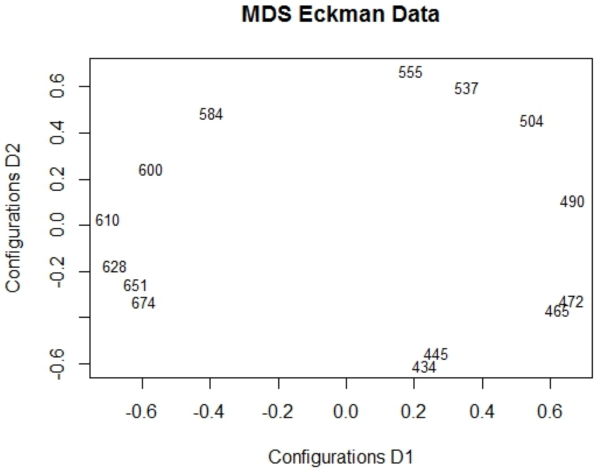
A second (historical) example: Ekman's color data

Ekman (1954) presents similarities for 14 colors (wavelengths from 434 to 674 nm).



Similarities are based on a rating by 31 subjects. Each pair of colors was rated on a 5-point scale (0 = no similarity up to 4 = identical).

J. de Leeuw, P. Mair
Multidimensional Scaling Using Majorization: SMACOF in R
Journal of Statistical Software (2009)



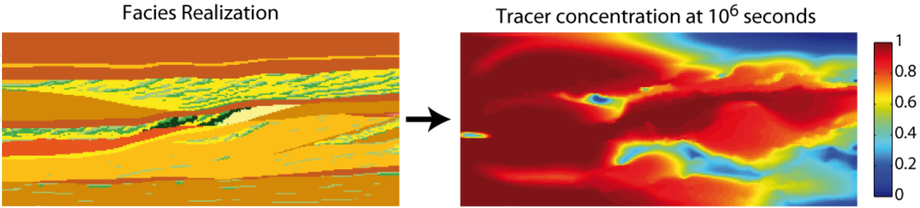

```
library("smacof") ; data("ekman")
D <- sim2diss(ekman, method = 1)
plot(smacofSym(D, metric = FALSE), main = "MDS Eckman Data")
```

Outline

Set-up of the forward flow simulation

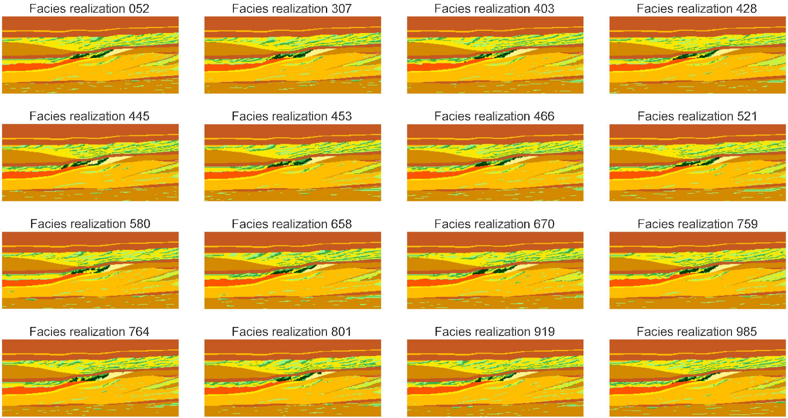
We now focus on numerical simulations taking a parameter field (or *map*, denoted by $x \in E$) as input and delivering a functional output:

$$f : x \in E \longrightarrow f_x(\cdot) \in F$$



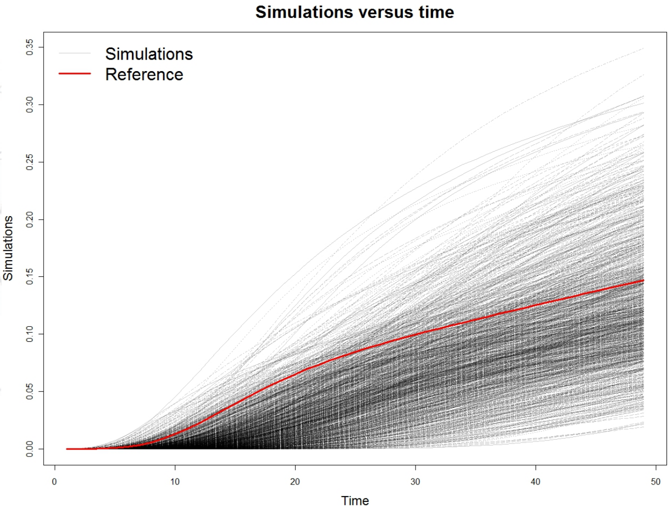
Multiple parameter fields may be candidate to model the subsurface . . .

16 among 1000 facies (multipoints simulations)



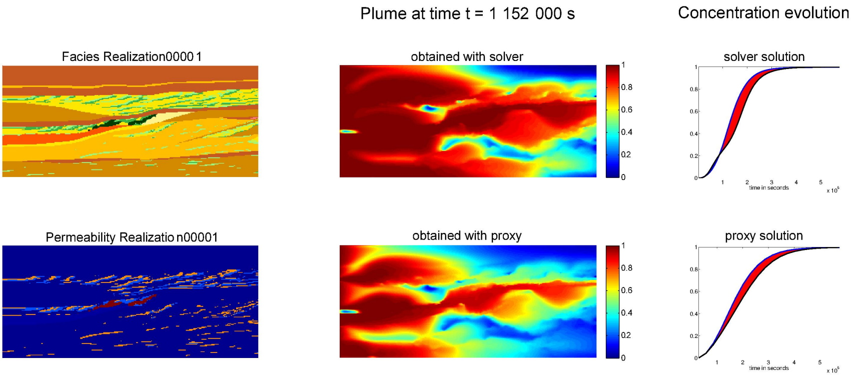
The candidate maps are noted x_i $(1 \leq i \leq n)$. Here $n = 1000$.

Corresponding distribution of outputs



How to capture the variability of the output relying on a few runs only?

Key idea of Scheidt et al.: using degraded simulations



If simulating the response precisely for the 1000 maps is *a priori* too long, doing rougher (proxy) simulations for all of them may be affordable.

Modeling

- The proxy simulator is denoted by p:

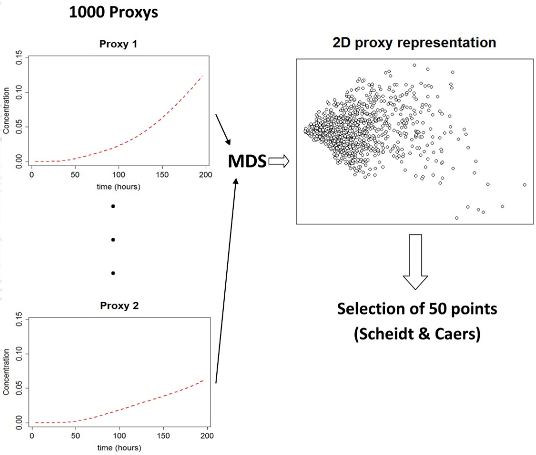
$$p : x \in E \longrightarrow p_x(\cdot) \in F$$

- E is equipped with a (pseudo-)distance:

$$d^2(x, y) := \int_{T_{\min}}^{T_{\max}} (p_x(t) - p_y(t))^2 \, dt$$

- We call $\mathbf{D}$ the $n \times n$ matrix of (pseudo-)distances^2 between the x_i's.

Proxy-based MDS



A clustering method allows defining a design of experiments reflecting the diversity of the n maps, according to the proxy pseudo-distance.

Some references on proxy-based distance methods

C. Scheidt, J. Caers
Representing spatial uncertainty using distances and kernels
Mathematical Geosciences 41 (4), 397-419

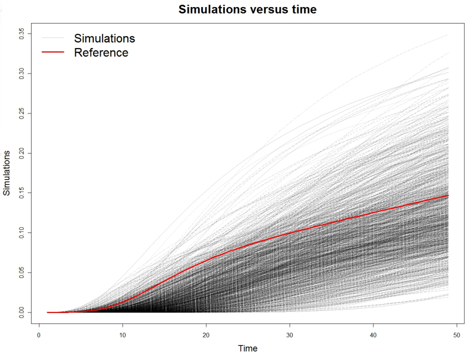
C. Scheidt, J. Caers
Uncertainty Quantification in Reservoir Performance Using Distances and Kernel Methods–Application to a West Africa Deepwater Turbidite Reservoir
SPE Journal 14 (4), 680-692

Outline

Inverse problem: identification of geological facies

One measures a response curve after a fluid injection at a boundary. A similar curve is then simulated for the candidate x's.

Comparing observed and simulated curves, one gets an idea of which parameter fields are "realistic" . . .



Modeling

- The reference curve is called f_{ref}
- The objective function to be minimized is called g:

$$g(x) := \int_{T_{\min}}^{T_{\max}} \left(f_x(t) - f_{\text{ref}}(t)\right)^2 dt$$

- Reminder: The candidate maps are noted x_i ($1 \leq i \leq n$). Here $n = 1000$.

Modeling

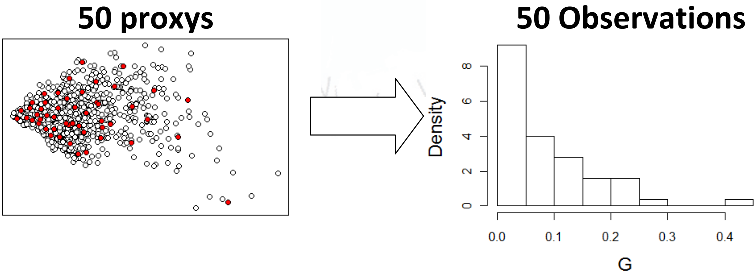
- The reference curve is called f_{ref}
- The objective function to be minimized is called g:

$$g(x) := \int_{T_{\min}}^{T_{\max}} \left(f_x(t) - f_{\text{ref}}(t)\right)^2 dt$$

- Reminder: The candidate maps are noted x_i ($1 \leq i \leq n$). Here $n = 1000$.

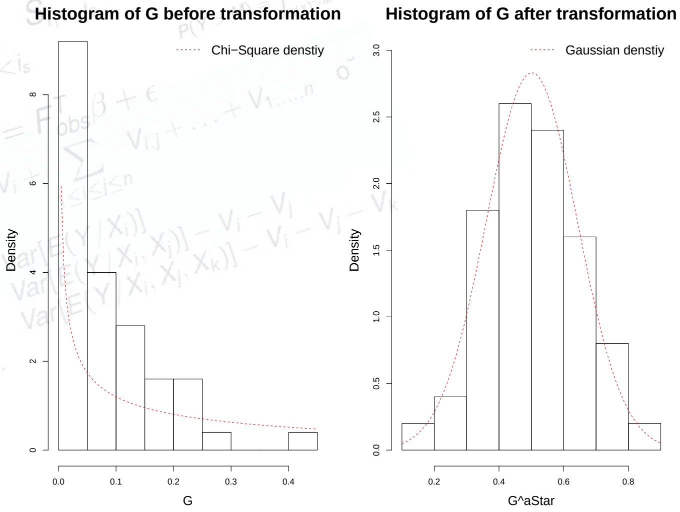
Problem: find, in a restricted number of evaluations (each simulation being very time consuming), as many x_i's as possible with small values of $g(x_i)$.

Initial design of experiments



An initial design is obtained by using Scheidt and Caers' approach.

Transformation of the objective function



Proxy-based kriging

Covariance kernel used

$$k(x, y) := \sigma^2 \exp\left(-\frac{1}{\theta^2}\int_0^T (p(x,t) - p(y,t))^2 dt\right) + \tau^2 \mathbf{1}_{x=y}$$

Proxy-based kriging

Covariance kernel used

$$k(x,y) := \sigma^2 \exp\left(-\frac{1}{\theta^2}\int_0^T (p(x,t)-p(y,t))^2 dt\right) + \tau^2 \mathbf{1}_{x=y}$$

Why is this kernel an admissible covariance over $E \times E$?

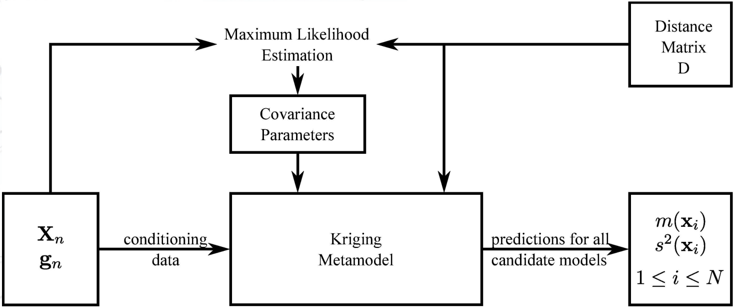
Proposition

Let E and F be two arbitrary spaces. Given a positive (semi-)definite kernel k_F on $F \times F$, the following kernel k_E:

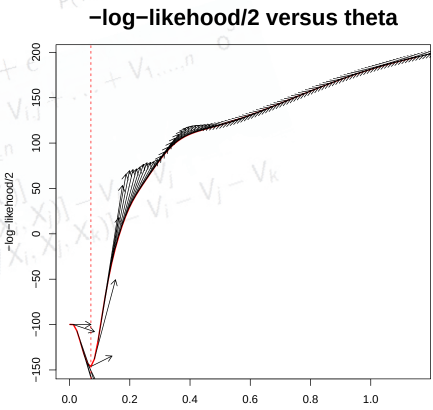
$$k_E(x,y) := k_F(p(x), p(y))$$

is positive (semi-)definite over $E \times E$, whatever the function $p : E \longrightarrow F$.

Implementation (transformation apart)

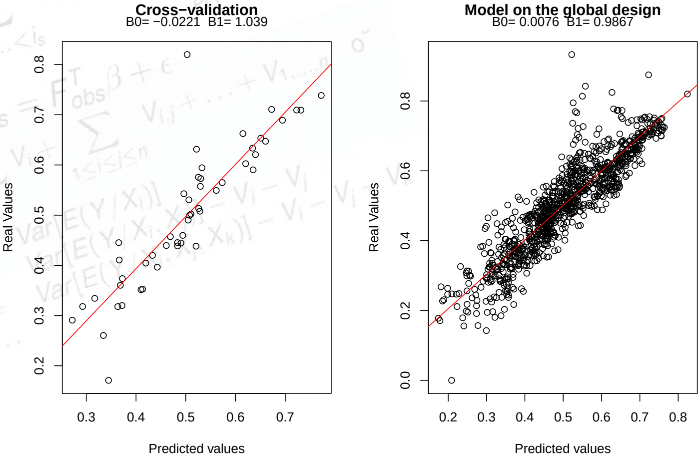


Estimation of kriging covariance parameters

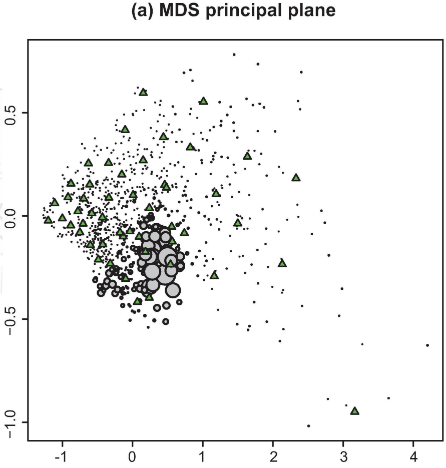


Theta*= 0.071 Sigma*= 0.007

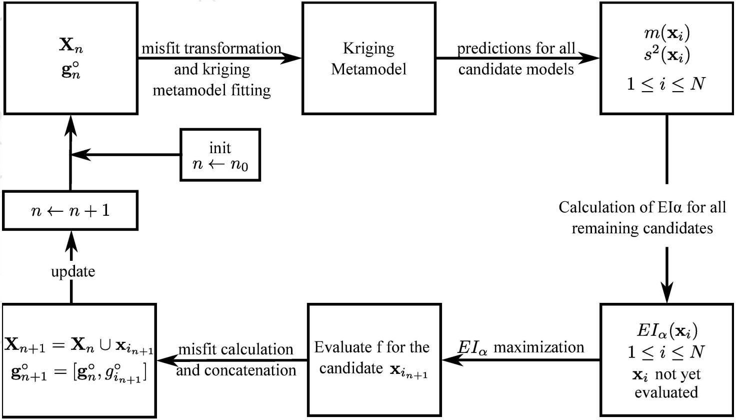
Validation of the Kriging model



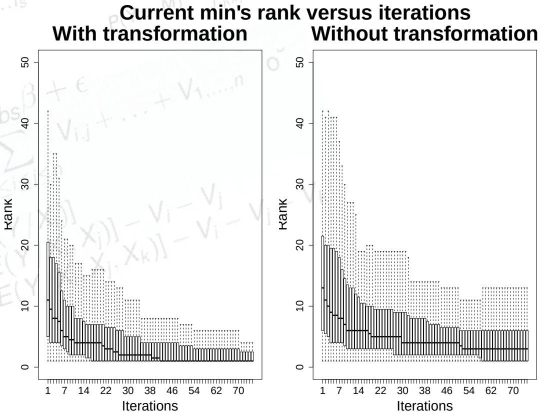
Expected Improvement in MDS space



Main loop of the ProKSI Algorithm

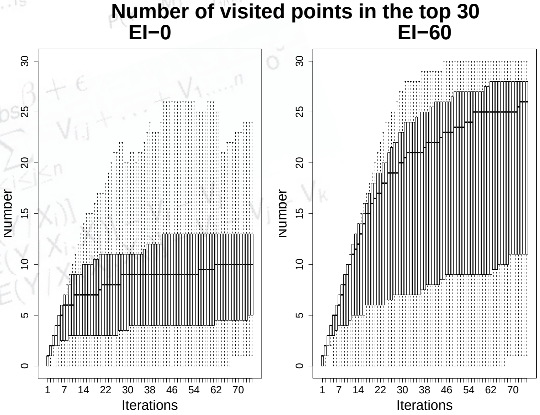


ProKSI Algorithm: Results based on 100 references



After 43 iterations, the global minimizer was visited in more than 50% of the cases. The performances are significantly better with transformation.

ProKSI Algorithm: Results based on 100 references



In 75 iterations of the (EI-60) strategy developed during B. Rosspopoff's internship, 25 of the 30 best maps are recovered (in median).

Conclusion and perspectives

For more detail, see also:

D. G., B. Rosspopoff, G. Pirot, N. Durrande, and P. Renard
Distance-based Kriging relying on proxy simulations for inverse conditioning (2013)
Advances in Water Resources (52), 275–291

Conclusion and perspectives

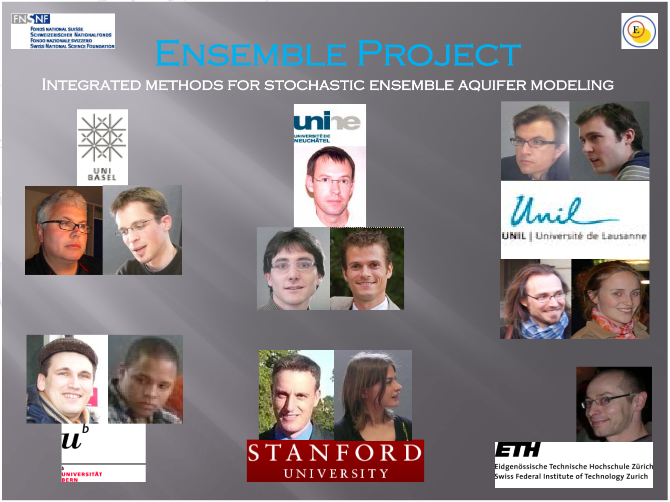
For more detail, see also:

D. G., B. Rosspopoff, G. Pirot, N. Durrande, and P. Renard
Distance-based Kriging relying on proxy simulations for inverse conditioning (2013)
Advances in Water Resources (52), 275–291

A few take home messages

- Distance methods deserve to be known; they are simple and useful!
- Distances can be adapted to the problem at hand; versatile methods . . . Further distance methods available (Clustering, non-metric MDS, etc.).
- Kriging and kriging-based optimization/inversion strategies are applicable in arbitrary dimensions provided that:
  - a) A suitable covariance kernel is available (or can be found)
  - b) The search is limited to a discrete subset of candidate inputs

Acknowledgements: ”ENSEMBLE” project



http://www.ensemble-modeling.org/