Metamodeling methods that incorporate qualitative variables for improved design of vegetative filter strips.

C. Lauvernet[1], C. Helbert[2], Zhu Xujia[3], B. Sudret[3]

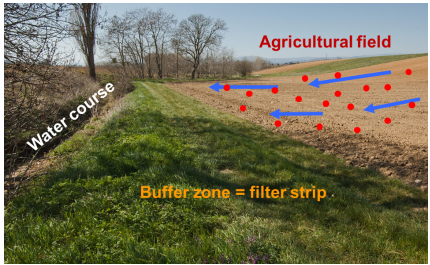
[1]RIVERLY, INRAE Lyon-Villeurbanne, FR
[2]Univ. Lyon, UMR CNRS 5208, Ecole Centrale de Lyon, FR
[3]ETH Zürich | Institute of Structural Engineering, Chair of Risk, Safety & Uncertainty Quantification, CH

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

Mitigation of non-point source inputs in France and in EU

- Significant amounts of pollutants are measured in surface water
- Vegetative filter strips (VFS) are identified as the BMP of Choice for Runoff mitigation
- VFSs are mandatory or advised depending on the country and conditions
- They need to be properly designed, considering the specific context



Mitigation of non-point source inputs in France and in EU

- Significant amounts of pollutants are measured in surface water
- Vegetative filter strips (VFS) are identified as the BMP of Choice for Runoff mitigation
- VFSs are mandatory or advised depending on the country and conditions
- They need to be properly designed, considering the specific context

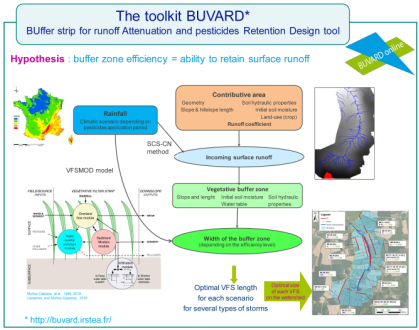
Development of a specific tool to design VFS, once a local diagnosis has been realized:
BUVARD [a]
BUffer strip for runoff Attenuation and pesticides Retention Design tool

[a] Carluer, N., Lauvernet, C., Noll, D, Muñoz-Carpena, R. Defining context-specific scenarios to design vegetated buffer zones that limit pesticide transfer via surface runoff Sc. of The Total Env., 2017, 575, 701-712

BUVARD issues for operational purposes

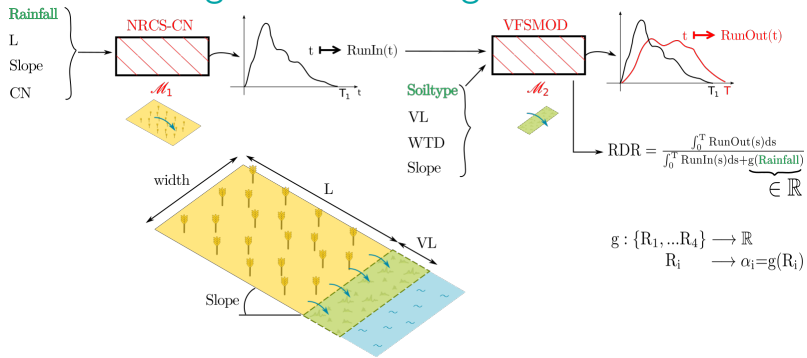
- Processes that drive the pesticide fate at the catchment scale are complex and interact : infiltration, surface runoff, sediment trapping, pesticide transfer, etc.
- BUVARD is in fact a chain of several models
- their description is based on non linear equations and/or conceptual and/or stochastic
- a large set of parameters that are difficult to measure/estimate
- inputs and outputs are dynamic (ex : rainfall)

⇒ **a high uncertainty in an operationnal context**



⇒ metamodeling BUVARD to bridge the gap between modeling and decision support

Challenges for the surrogate of BUVARD



- a chain of several models
- inputs are quantitative and qualitative (categorical)
- a huge number of zero values of Runin, Runout, and then RDR
- The output variable RDR has to range between 0 and 1

Problem description

Input

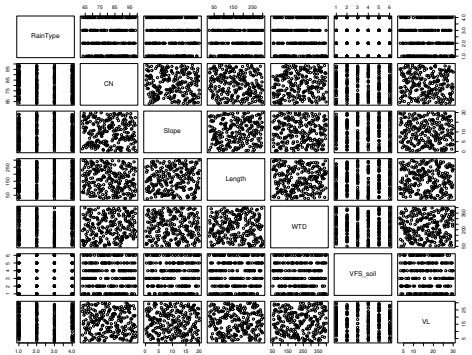
Variable	Name	Distribution	Parameters
X_1	Curve number (CN)	Uniform	$[63, 99]$
X_2	Slope	Uniform	$[0.1, 20]$
X_3	Length	Uniform	$[25, 300]$
X_4	Rainfall type	Categorical	4 levels with equal probability
X_5	Vegetative length (VL)	Uniform	$[3, 30]$
X_6	Water table depth (WTD)	Uniform	$[50, 400]$
X_7	Soil type	Categorical	6 levels with equal probability

Output

- R_{in} (depends on $X_1 - X_4$), R_{out} (depends on all the inputs), $Rain$ (depends on X_4, X_5) are recorded
- Model output: $\mathrm{RDR} = \frac{R_{out}}{R_{in}+Rain}$ which is between $[0, 1]$

Data

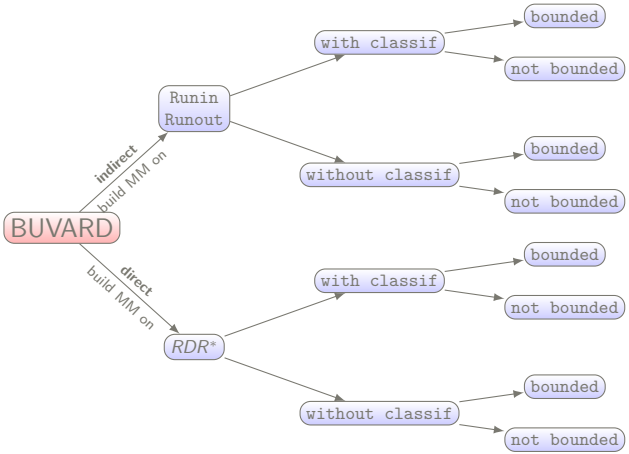
2400 training data and 960 test data



LHS Sampling approach is not too expensive, and adapted to irregular models.

Obj. = good projection properties on each axis : each 1D projection is Maximin-optimal

Metamodeling experiments



The metamodel is built using

- Gaussian Process regression / DeepGP
- Polynomial Chaos Expansion

$\rightarrow$ adapted to :

- mixed variables (quali/quanti)
- or by category

Metamodels

Approximation of a function $f : [0,1]^d \to \mathbb{R}$ from observations $\mathbf{y} = f(\mathbb{X})$ on a DoE $\mathbb{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$.

Gaussian Process regression (kriging)

- f is a realization of $(Y(\mathbf{x}))_{\mathbf{x}} \sim GP(m, k(.,.))$
- Prediction : $\hat{f}(x) = \mathbb{E}(Y(\mathbf{x})|Y(\mathbb{X}) = \mathbf{y})$
- Interpolation, non parametric approach, all is in the prior.

Polynomial Chaos Expansion (PCE)

- $\hat{f}(x) = \sum_{\alpha \in \mathbb{N}^d} c_\alpha \phi_\alpha(\mathbf{x})$ where ϕ_α are obtained by tensor product of polynomial chaos basis (Legendre, Hermite, ...).
- Estimation of $\boldsymbol{c}$ by least squares $\min \|\mathbf{y} - \boldsymbol{\Psi c}\|$, with a sparsity criterion (LASSO).
- Approximation approach.

Adaptation to categorical inputs

Assume that the categorical variable U having K levels $\{u_1, \ldots, u_K\}$

Kriging : adaptation of the covariance kernel[1]

$$k((\boldsymbol{x}, u_j), (\boldsymbol{x}', u_l)) = k_1(\boldsymbol{x}, \boldsymbol{x}')k_2(u_j, u_l)$$

k_2 is a specific covariance kernel for categorical variables, several choices are possible

Polynomial Chaos Expansion (PCE)[2]. The multivariate basis are given by

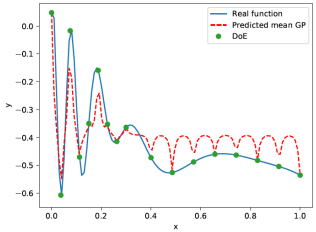
$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}, u) = \varphi_{\boldsymbol{\alpha}_x}(\boldsymbol{x}) \otimes \phi_{\boldsymbol{\alpha}_u}(u)$$

The estimation is done by group-LARS

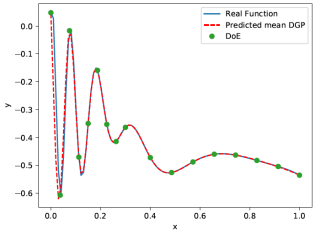
$$\hat{\boldsymbol{c}} = \arg\min_{\boldsymbol{c}} \|\boldsymbol{y} - \boldsymbol{\Psi}\boldsymbol{c}\| + \nu \sum_{\mathcal{G} \in \boldsymbol{\mathcal{G}}} \|\boldsymbol{c}_{\mathcal{G}}\|_{G_{\mathcal{G}}}$$

1. See Lauvernet, C., Helbert, C. Metamodeling methods that incorporate qualitative variables for improved design of vegetative filter strips Reliability Engineering System Safety, 2020, 204, 107083
2. See Xujia Zhu,Bruno Sudret presentation, just before me!

Presence of null observations - DeepGP for non stationarity



GP prediction of a non-stationary 1-D function.



DGP prediction of a non-stationary 1-D function.

Figure: Extracted from PhD defense of Ali Hebbal

Presence of null observations - DeepGP for non stationarity [Damianou and Lawrence, 2013]

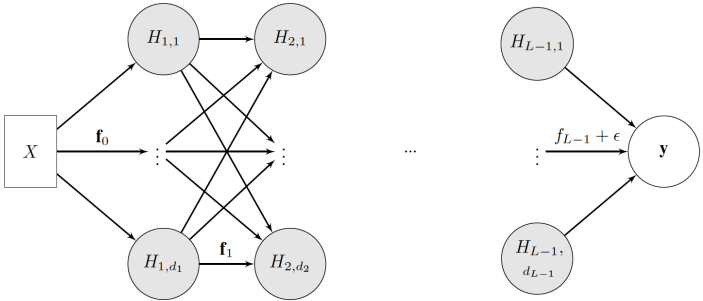
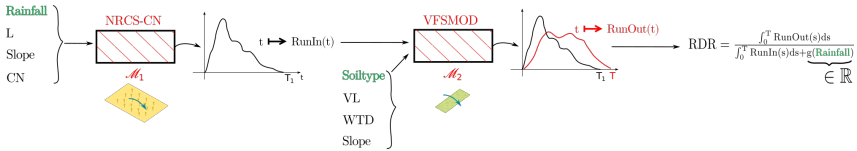
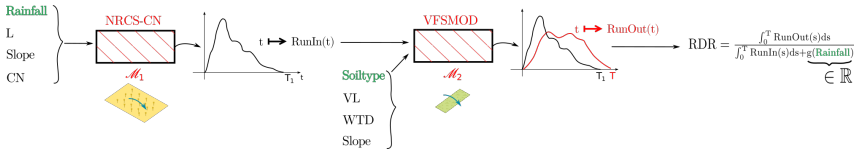


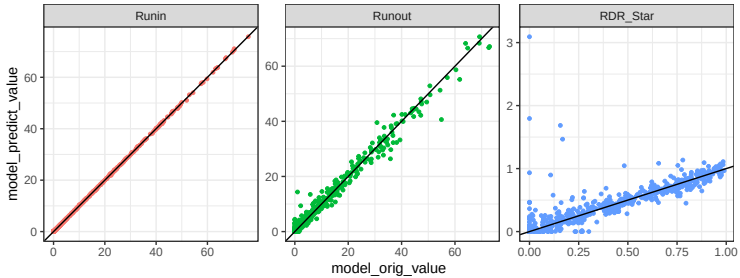
Figure: Extracted from PhD defense of Ali Hebbal

Direct MM vs indirect MM



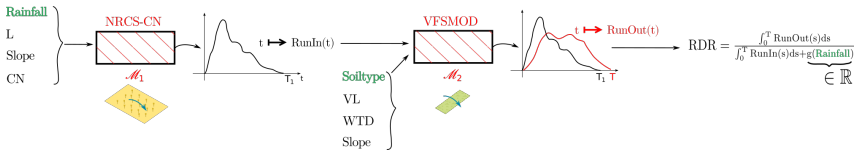
Direct MM vs indirect MM

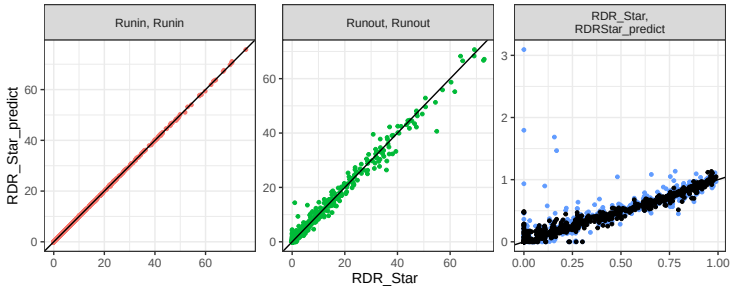




Kriging: $R^2 = 0.999$ $R^2 = 0.985$ $R^2 = 0.753$

Direct MM vs indirect MM

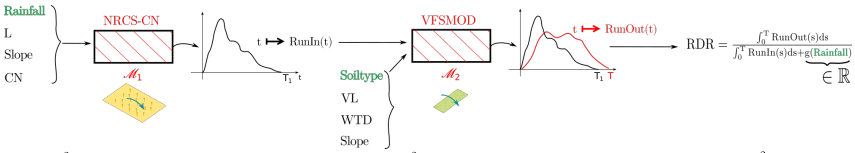


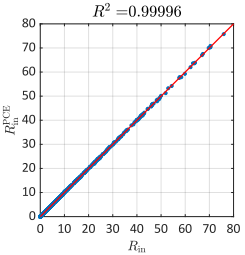


Kriging $R^2 = 0.999$ $R^2 = 0.985$ $R^2 = 0.753 \Rightarrow 0.96$

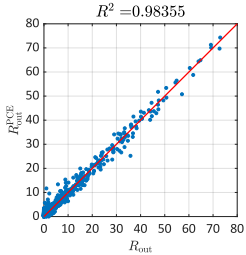
$\Rightarrow$ Surrogate of the ratio is much more reliable

Direct MM vs indirect MM

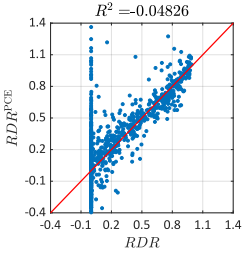




Direct PCE for Runin



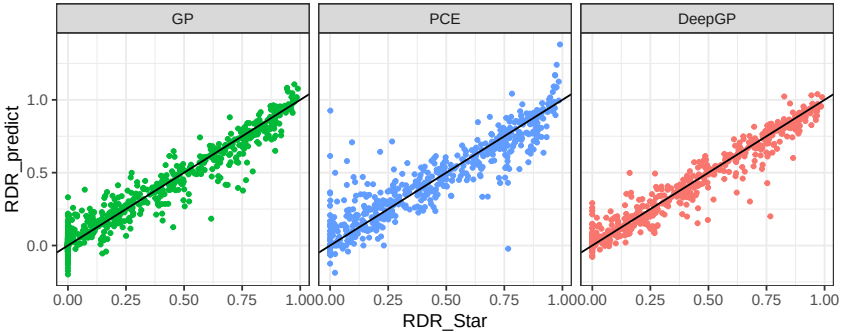
Direct PCE for Runout



Indirect surrogate for RDRStar

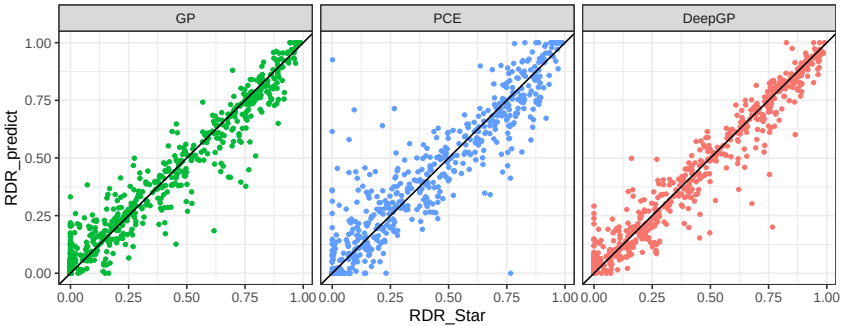
$\Rightarrow$ The same for PCE !

Results: MM with classif / boundaries ? comparison per category



$R^2 = 0.951$ $R^2 = 0.903$ $R^2 = 0.964$

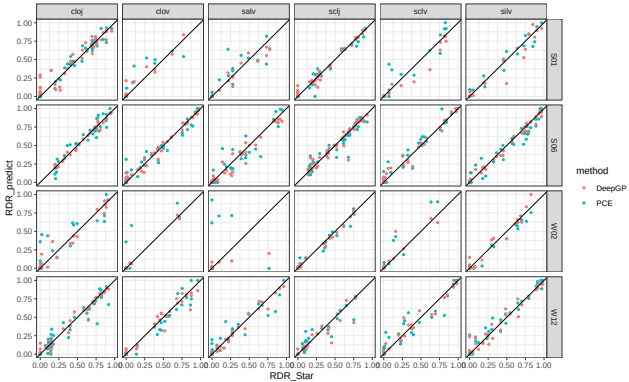
Results: MM with classif / boundaries ? comparison per category



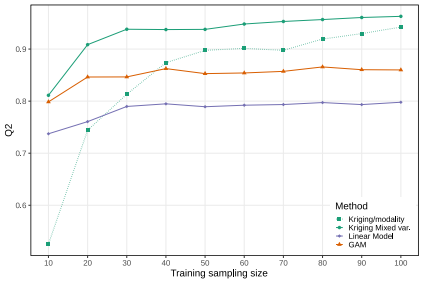
before	: $R^2 = 0.951$	$R^2 = 0.903$	$R^2 = 0.964$
bounded	: $R^2 = 0.955$	$R^2 = 0.911$	$R^2 = 0.964$

$\Rightarrow$ DeepGP does not need any classification or boundaries

Results : mixed variables vs by category?



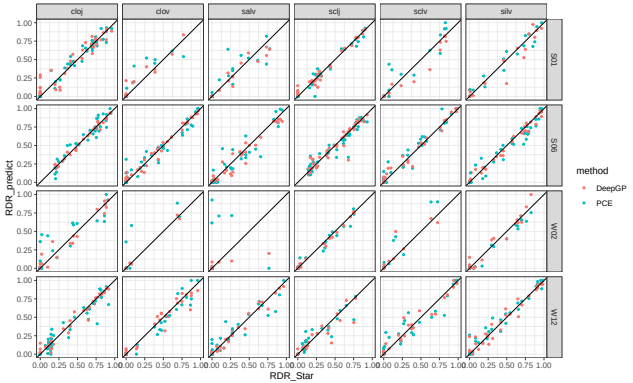
DeepGP and PCE MM by couple of category (Soil type x Rain type)



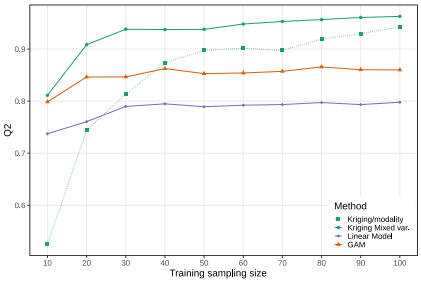
Effect of training sampling size on metamodel quality[a]

[a]extracted from Lauvernet and Helbert, 2020, RESS, 204

Results : mixed variables vs by category?



DeepGP and PCE MM by couple of category (Soil type x Rain type)



Effect of training sampling size on metamodel quality[a]

[a] extracted from Lauvernet and Helbert, 2020, RESS, 204

⇒ Both methods are in trouble with soils with a predominance of zeros
Mixed methods are more robust to the sampling size

Results : mixed variables vs by category?

Method	R^2 per category	R^2 for mixed var.
PCE	0.916	0.966
Kriging	0.955	0.964
DeepGP	0.964	-

⇒ Methods for mixed variables are more efficient and robust, and even more with smaller samplings

⇒ DeepGP performs well but needs repetitions for the worst soils, and is more costly numerically

Summary

- Categorical variables were proprely taken into account by the kriging and by the PCE adaptations
- Mixed variables methods outperform the MM by category
- Classification does not improve the surrogate
- Good quality of prediction (96 % of variance is explained)

$\Rightarrow$ Next step : DeepGP for categorical variables

Thank you!