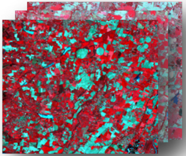
High Dimensional Kullback-Leibler divergence for grassland classification using satellite image time series with high spatial resolution

Presented by Mailys Lopes[1]
In collaboration with Mathieu Fauvel[1], Stéphane Girard[2] and David Sheeren[1]

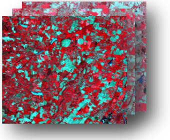
[1]Dynafor, INRA, University of Toulouse, France
[2]Team Mistis, INRIA Grenoble, France

ADVANCING THE UNDERSTANDING OF OUR LIVING PLANET
IGARSS JULY 10 - 15, 2016 ★ BEIJING, CHINA

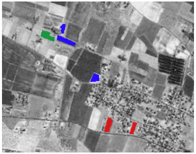


Study objectives

Agroecological application
Grassland management practices



Data
SITS* with high spatial ($\approx$ 10m) resolution
and temporal (2-3 images per month) resolution
*satellite image time series



Method
Supervised classification at the object scale

Context

High Dimensional Divergence for measuring grassland similarity

Experimental results

Conclusion

We propose to use **dense multispectral time series with high spatial resolution** to characterize grasslands.

Grasslands in Europe are:

- Relatively **small** ($\approx$ 100m $\times$ 100m) $\Rightarrow$ need **high spatial resolution images**
- **Heterogeneous** $\Rightarrow$ need **multispectral images**
- **Natural cycle** disturbed by **human activities** $\Rightarrow$ need **SITS**

Mowing

Grazing

We propose to use **dense multispectral time series with high spatial resolution** to characterize grasslands.

Grasslands in Europe are:

- Relatively **small** ($\approx$ 100m $\times$ 100m) $\Rightarrow$ need **high spatial resolution images**
- **Heterogeneous** $\Rightarrow$ need **multispectral images**
- **Natural cycle** disturbed by **human activities** $\Rightarrow$ need **SITS**

Mowing

Grazing

We propose to use **dense multispectral time series with high spatial resolution** to characterize grasslands.

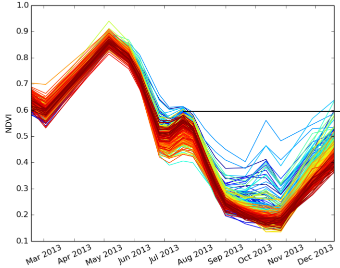
Grasslands in Europe are:

- Relatively **small** ($\approx$ 100m $\times$ 100m) $\Rightarrow$ need **high spatial resolution images**
- **Heterogeneous** $\Rightarrow$ need **multispectral images**
- **Natural cycle** disturbed by **human activities** $\Rightarrow$ need **SITS**

Mowing

Grazing

- Learn f such as $y_i = f(\mathbf{X}_i)$, where y_i is the predicted label
- $\mathbf{X}_i = \left[\mathbf{x}_{i1}|\ldots|\mathbf{x}_{in_i}\right]$ is a matrix of size $(n_i \times d)$ that contains all the pixels inside g_i



$$\mathbf{x}_{ik}(t) = \begin{pmatrix} x_{ik}(t_1) \\ \vdots \\ x_{ik}(t_d) \end{pmatrix} \in \mathbb{R}^d$$

with

- g_i grassland with index i,
- n_i number of pixels in grassland g_i
- k pixel index, $k \in \{1, \ldots, n_i\}$
- d length of time series
- $x_{ik}(t_l)$ NDVI value of pixel k at time l

Thematic contributions

- **Grassland management practices**
- **Sentinel-2** contribution

Methodological contributions

- **Model grassland distribution**
- Process grassland **supervised classification at the parcel scale**
- **Robust** to
  - the dimension of data (n_i pixels, d temporal variables with $n_i \approx d$),
  - the total number of grasslands pixels that might be large.

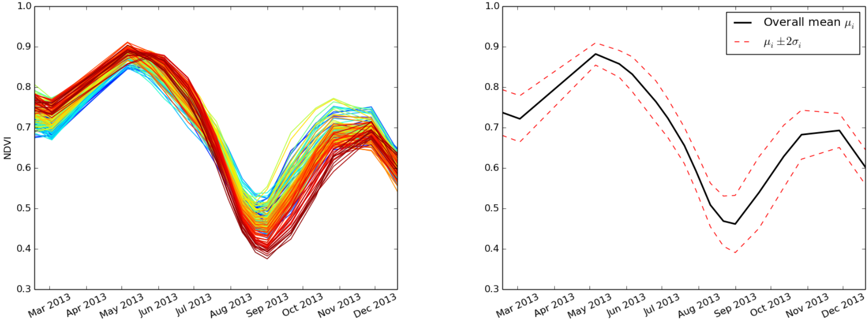
Context

High Dimensional Divergence for measuring grassland similarity

Experimental results

Conclusion

For each grassland g_i: each pixel $\mathbf{x}_{ik} \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$



Measuring similarity between g_i and $g_j \Rightarrow$ measuring similarity betwen $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$

Symmetrized Kullback-Leibler divergence:

$$KLD(g_i, g_j) = \frac{1}{2}\Big[\operatorname{Tr}\left[\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i\right] + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top\left(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1}\right)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\Big] - d$$

The number of pixels in a grassland is usually lower than the number of parameters to estimate!

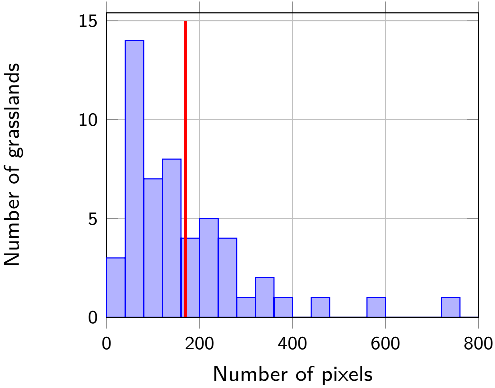


Figure : Histogram of grassland size in number of pixels n_i. Red line: number of parameters to estimate for each grassland for a multivariate Gaussian model. It is derived from the number of variables using the formula $d(d+3)/2 = 170$ for $d = 17$.

High Dimensional Discriminant Analysis[1] modelling is used: it **assumes that the last eigenvalues of the covariance matrix are equal**:

$$\mathbf{\Sigma}_i = \mathbf{P}_i \mathbf{D}_i \mathbf{P}_i^\top$$

$$\text{with } \mathbf{D}_i = \left(\begin{array}{cc} \boxed{\begin{array}{ccc}\lambda_{i1} & & 0\\ & \ddots & \\ 0 & & \lambda_{ip_i}\end{array}} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{array}{cccc}\lambda_i & & & 0\\ & \ddots & & \\ & & \ddots & \\ 0 & & & \lambda_i\end{array}}\end{array}\right) \begin{array}{l} \left.\vphantom{\begin{array}{c}a\\b\\c\end{array}}\right\} \; p_i \\ \left.\vphantom{\begin{array}{c}a\\b\\c\\d\end{array}}\right\} \; (d - p_i)\end{array}$$

- p_i is the number of non-equal eigenvalues,
- λ_i is the multiple eigenvalue corresponding to the noise term (last and equal eigenvalues),
- $\lambda_{ij} \geq \lambda_i$, for $j = 1, ..., p_i$

[1] C. Bouveyron, S. Girard and C. Schmid, "High-dimensional discriminant analysis", Communications in Statistics - Theory and Methods, vol. 36, no. 14, pp. 2607–2623, 2007

High Dimensional Discriminant Analysis[1] modelling is used: it **assumes that the last eigenvalues of the covariance matrix are equal**:

$$\boxed{\mathbf{\Sigma}_i = \mathbf{Q}_i \mathbf{\Lambda}_i \mathbf{Q}_i^\top + \lambda_i \mathbf{I}_d}$$

- $\mathbf{I}_d$ is the identity matrix of size d,
- $\mathbf{Q}_i = \left[\mathbf{q}_{i1}, \ldots, \mathbf{q}_{ip_i}\right]$,
- $\mathbf{\Lambda}_i = \mathrm{diag}\left[\lambda_{i1} - \lambda_i, \ldots, \lambda_{ip_i} - \lambda_i\right]$.

Folllowing this model, $\mathbf{\Sigma}_i^{-1}$ can be computed explicitly:

$$\boxed{\mathbf{\Sigma}_i^{-1} = -\mathbf{Q}_i \mathbf{V}_i \mathbf{Q}_i^\top + \lambda_i^{-1} \mathbf{I}_d}$$

with $\mathbf{V}_i = \mathrm{diag}\left[\frac{1}{\lambda_i} - \frac{1}{\lambda_{i1}}, \ldots, \frac{1}{\lambda_i} - \frac{1}{\lambda_{ip_i}}\right]$

[1] C. Bouveyron, S. Girard and C. Schmid, "High-dimensional discriminant analysis", Communications in Statistics - Theory and Methods, vol. 36, no. 14, pp. 2607–2623, 2007

To compute HDKLD, only the p_i first eigenvalues/eigenvectors are required:

- the number of parameters to estimate is reduced,
- the unstable estimation of the eigenvectors associated to small eigenvalues is avoided.

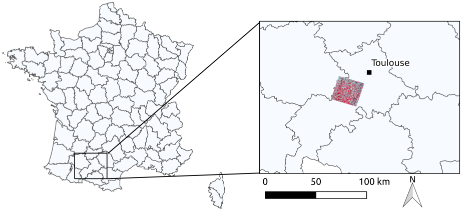
$$HDKLD(g_i, g_j) = \frac{1}{2}\Big[-\|\mathbf{\Lambda}_j^{\frac{1}{2}}\mathbf{Q}_j^\top\mathbf{Q}_i\mathbf{V}_i^{\frac{1}{2}}\|_F^2 - \|\mathbf{\Lambda}_i^{\frac{1}{2}}\mathbf{Q}_i^\top\mathbf{Q}_j\mathbf{V}_j^{\frac{1}{2}}\|_F^2 \\ + \lambda_i^{-1}\,\mathrm{Tr}\left[\mathbf{\Lambda}_j\right] - \lambda_j\,\mathrm{Tr}\left[\mathbf{V}_i\right] + \lambda_j^{-1}\,\mathrm{Tr}\left[\mathbf{\Lambda}_i\right] - \lambda_i\,\mathrm{Tr}\left[\mathbf{V}_j\right] \\ - \|\mathbf{V}_i^{\frac{1}{2}}\mathbf{Q}_i^\top(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\|^2 - \|\mathbf{V}_j^{\frac{1}{2}}\mathbf{Q}_j^\top(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\|^2 \\ + \frac{\lambda_i + \lambda_j}{\lambda_i\lambda_j}\|(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\|^2 + \frac{\lambda_i^2 + \lambda_j^2}{\lambda_i\lambda_j}d\Big] - d$$

where $\|L\|_F^2 = \mathrm{Tr}(L^\top L)$ is the Frobenius norm.

HDKLD is used to build a positive definite kernel function. This kernel function can be used in any kernel method, such as SVM.

- (HD)KLD is a semi-metric.
- $K(g_i, g_j) = \exp\left[-\frac{(HD)KLD(g_i, g_j)^2}{\sigma}\right]$ with $\sigma \in \mathbb{R}_*^+$

Study site

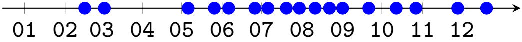


Field data 52 parcels with 3 management practices (field survey, 2015):

Class	Nb of grasslands
Mowing	34
Grazing	10
Mixed (mowing & grazing)	8

Satellite data **Formosat-2**

- Spatial resolution: 8m
- 4 spectral bands (B, G, R, NIR)
- $d = 17$ images from Feb. 16 to Dec. 20, 2013:



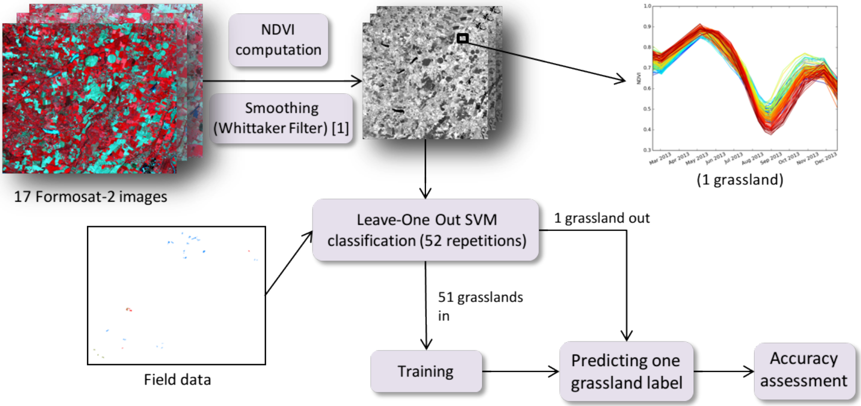
Comparison with other methods:

Method	**p-SVM**	μ-**SVM**	**KLD-SVM**	**HDKLD-SVM**
Scale	Pixel	Object	Object	Object
Expl. variable	$\mathbf{x}_{ik}$	$\boldsymbol{\mu}_i$	$\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$	$\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
Kernel	RBF	RBF	$K(g_i, g_j)$	$K(g_i, g_j)$
Nb of samples	8741	52	52	52

$+$ *majority voting rule*

Optimal hyperparameters have been optimized by cross-validation.

Leave-One-Out Cross-Validation for error estimation.



HDKLD outperforms KLD, but the results are not significantly different than the other methods.

Classification accuracy

	P-SVM			μ-SVM			KLD-SVM			HDKLD-SVM		
	REF			REF			REF			REF		
PRED	**32**	4	2	**31**	6	3	**32**	8	8	**33**	4	4
PRED	1	**4**	1	1	**0**	0	1	**0**	0	0	**3**	0
PRED	1	0	**7**	2	2	**7**	1	0	**2**	1	1	**6**
OA	0.83			0.73			0.66			0.81		
κ	0.64			0.41			0.09			0.57		

Test of significance of observed differences

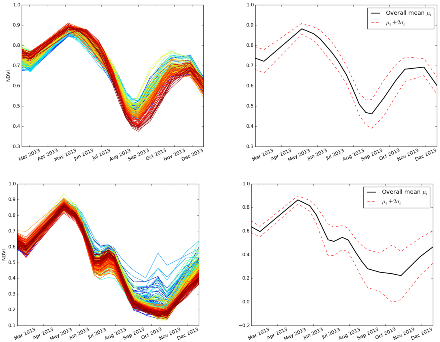
Z	p-SVM	μ-SVM	KLD-SVM	HDKLD-SVM
p-SVM	-	1.47	**3.67**	0.43
μ-SVM	1.47	-	**2.04**	1.01
KLD-SVM	**3.67**	**2.04**	-	**3.11**
HDKLD-SVM	0.43	1.01	**3.11**	-

$$Z = \frac{|\hat{\kappa}_m - \hat{\kappa}_n|}{\sqrt{\hat{\mathrm{var}}(\hat{\kappa}_m) + \hat{\mathrm{var}}(\hat{\kappa}_n)}}^{*}$$

(*from Congalton and Green, *Assessing the Accuracy of Remotely Sensed Data, Principles and Practices*, 2009)

The proposed model

- enables a **proper grassland modelling at the parcel scale**,



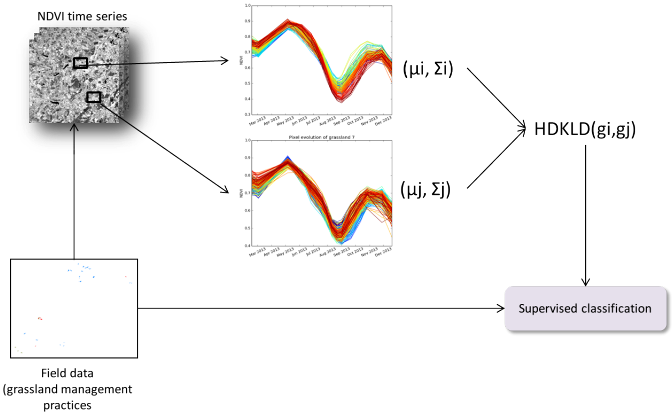
- **reduces the number of elements** to be processed by SVM.
  For HDKLD-SVM: $G = 52$ grasslands processed
  For P-SVM: $N = 8741$ pixels processed

Context

High Dimensional Divergence for measuring grassland similarity

Experimental results

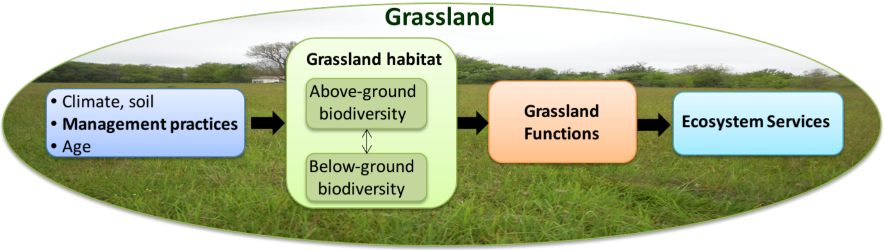
Conclusion

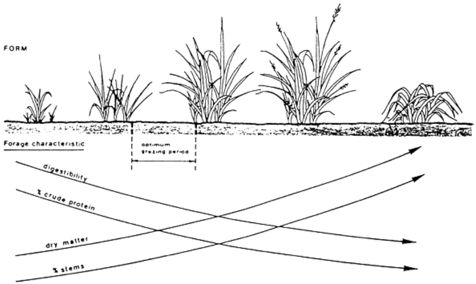


- **Gaussian modelling** seems accurate to **model grassland pixels distribution**.
- **HDKLD is efficient** for measuring **grassland proximity** and **robust to the dimension of data**.
- Best results are not **significantly different**.
- **HDKLD is better than KLD**.

- The method will be tested on a **larger dataset**.
- The method will be further extended to **multispectral data**.
- The method will be used for **unsupervised classification**.


Thank you for your attention.


Grassland
Grassland habitat
Above-ground biodiversity
Below-ground biodiversity
• Climate, soil
• Management practices
• Age
Grassland Functions
Ecosystem Services


FORM
Forage characteristic
optimum grazing period
digestibility
% crude protein
dry matter
% stems

Estimation of HDKLD parameters:

$$\begin{aligned}HDKLD(g_i, g_j) = \frac{1}{2}\Big[& - \|\mathbf{\Lambda}_j^{\frac{1}{2}} \mathbf{Q}_j^\top \mathbf{Q}_i \mathbf{V}_i^{\frac{1}{2}}\|_F^2 - \|\mathbf{\Lambda}_i^{\frac{1}{2}} \mathbf{Q}_i^\top \mathbf{Q}_j \mathbf{V}_j^{\frac{1}{2}}\|_F^2 \\ & + \lambda_i^{-1} \operatorname{Tr}\left[\mathbf{\Lambda}_j\right] - \lambda_j \operatorname{Tr}\left[\mathbf{V}_i\right] + \lambda_j^{-1} \operatorname{Tr}\left[\mathbf{\Lambda}_i\right] - \lambda_i \operatorname{Tr}\left[\mathbf{V}_j\right] \\ & - \|\mathbf{V}_i^{\frac{1}{2}} \mathbf{Q}_i^\top (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\|^2 - \|\mathbf{V}_j^{\frac{1}{2}} \mathbf{Q}_j^\top (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\|^2 \\ & + \frac{\lambda_i + \lambda_j}{\lambda_i \lambda_j} \|(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\|^2 + \frac{\lambda_i^2 + \lambda_j^2}{\lambda_i \lambda_j} d \Big] - d\end{aligned}$$

- $\hat{\lambda}_{ij}$ and $\hat{\boldsymbol{q}}_{ij}$ are the first eigenvalues/eigenvectors of $\hat{\mathbf{\Sigma}}_i$, $j \in \{1, \ldots, p_i\}$,
- $\hat{p}_i$ corresponds to the number of eigenvalues needed to reach a given percentage of variance t, $\frac{\sum_{j=1}^{\hat{p}_i} \hat{\lambda}_{ij}}{\operatorname{Tr}(\hat{\mathbf{\Sigma}}_i)} \geq t$, t being a user defined parameter,
- $\hat{\lambda}_i = \frac{\operatorname{Tr}(\hat{\mathbf{\Sigma}}_i) - \sum_{j \leq \hat{p}_i} \hat{\lambda}_{ij}}{d - \hat{p}_i}$.

Optimal parameters have been optimized during cross-validation given this search grid:

Parameter	p-SVM	μ-SVM	KLD-SVM	HDKLD-SVM
σ	$\{2^{-5}, 2^{-4}, \ldots, 2^{5}\}$		$\{2^{8}, 2^{9}, \ldots, 2^{12}\}$	
C	$\{1, 10, 100\}$			
t			$\{0.80, 0.85, 0.90, 0.95, 0.99\}$	

Bhattacharyya distance:

$$D_B(g_i, g_j) = \frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2} ln\left(\frac{det(\boldsymbol{\Sigma})}{\sqrt{det(\boldsymbol{\Sigma}_i)det(\boldsymbol{\Sigma}_j)}}\right)$$

with

$$\boldsymbol{\Sigma} = \frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2}$$