



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

Year : 2013

« Kernel-based methods for change detection
in remote sensing images »

Michele Volpi

Volpi, Michele (2013) ; Kernel-based methods for change detection in remote sensing images

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive.

<http://serval.unil.ch>

Droits d'auteur

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.





UNIL | Université de Lausanne

Faculté des Géosciences et de l'Environnement
Centre de Recherche en Environnement Terrestre

Kernel-based methods for change detection in remote sensing images

Thèse de doctorat

Présentée à la

Faculté des Géosciences et de l'Environnement
de l'Université de Lausanne

pour l'obtention du grade de

Docteur en Sciences de l'Environnement

par

Michele Volpi

B.Sc., M.Sc. Université de Lausanne, Suisse

Jury

Président du colloque	Prof. François BUSSY
Directeur de Thèse:	Prof. Mikhail KANEVSKI
Co-directeur de Thèse:	Dr. Devis TUIA
Expert externe:	Prof. Gustavo CAMPS-VALLS
Expert externe:	Prof. Jean-Philippe THIRAN

Lausanne, 2013



UNIL | Université de Lausanne
Décanat Géosciences et de l'Environnement
bâtiment Géopolis
CH-1015 Lausanne

IMPRIMATUR

Vu le rapport présenté par le jury d'examen, composé de

Président de la séance publique :	M. le Professeur François Bussy
Président du colloque :	M. le Professeur François Bussy
Co-Directeur de thèse :	M. le Professeur Mikhail Kanevski
Co-Directeur de thèse :	M. le Docteur Devis Tuia
Expert externe :	M. le Professeur Gustavo Camps-Valls
Expert externe :	M. le Professeur Jean-Philippe Thiran

Le Doyen de la Faculté des géosciences et de l'environnement autorise l'impression de la thèse de

Monsieur Michele VOLPI

Titulaire d'une

*Maîtrise universitaire ès Sciences en géosciences de l'environnement
Université de Lausanne*

intitulée

**KERNEL-BASED METHODS FOR CHANGE DETECTION IN
REMOTE SENSING IMAGES**

Lausanne, le 26 août 2013

Faculté des géosciences et de l'environnement

Professeur François Bussy, Doyen

Kernel-based methods for change detection in remote sensing images

Michele Volpi

Centre for Research on Terrestrial Environment

Summary

Nowadays, the joint exploitation of images acquired daily by remote sensing instruments and of images available from archives allows a detailed monitoring of the transitions occurring at the surface of the Earth. These modifications of the land cover generate spectral discrepancies that can be detected via the analysis of remote sensing images. Independently from the origin of the images and of type of surface change, a correct processing of such data implies the adoption of flexible, robust and possibly nonlinear method, to correctly account for the complex statistical relationships characterizing the pixels of the images.

This Thesis deals with the development and the application of advanced statistical methods for multi-temporal optical remote sensing image processing tasks. Three different families of machine learning models have been explored and fundamental solutions for change detection problems are provided.

In the first part, change detection with user supervision has been considered. In a first application, a nonlinear classifier has been applied with the intent of precisely delineating flooded regions from a pair of images. In a second case study, the spatial context of each pixel has been injected into another nonlinear classifier to obtain a precise mapping of new urban structures. In both cases, the user provides the classifier with examples of what he believes has changed or not.

In the second part, a completely automatic and unsupervised method for precise binary detection of changes has been proposed. The technique allows a very accurate mapping without any user intervention, resulting particularly useful when readiness and reaction times of the system are a crucial constraint.

In the third, the problem of statistical distributions shifting between acquisitions is studied. Two approaches to transform the couple of bi-temporal images and reduce their differences unrelated to changes in land cover are studied. The methods align the distributions of the images, so that the pixel-wise comparison could be carried out with higher accuracy. Furthermore, the second method can deal with images from different sensors, no matter the dimensionality of the data nor the spectral information content. This opens the doors to possible solutions for a crucial problem in the field: detecting changes when the images have been acquired by two different sensors.

Kernel-based methods for change detection in remote sensing images

Michele Volpi

Centre de Recherches en Environnement Terrestre

Résumé

L'exploitation conjointe des images de télédétection acquises sur une base journalière et de celles présentes dans les archives permettent un suivi détaillé des transformations survenant à la surface de la Terre. Les modifications des classes de couverture du sol engendrent des divergences dans l'information spectrale qui peuvent être détectées par l'analyse d'images de télédétection. Indépendamment de l'origine de l'image ou du type de changement au sol, le traitement de ce type de données implique l'utilisation de méthodes flexibles, robustes et potentiellement non-linéaires, ainsi qu'une bonne prise en compte des relations statistiques complexes qui caractérisent les pixels des images.

Cette Thèse aborde le développement et l'application de méthodes statistiques avancées pour le traitement d'images optiques multi-temporelles. Trois différentes familles de modèles d'apprentissage par ordinateur ont été explorées et solutions aux problèmes fondamentaux pour la détection de changements sont proposées.

Dans la première partie, la détection de changements est réalisée sous la supervision de l'utilisateur. La première application présentée exploite un classificateur non-linéaire pour la cartographie des zones inondées à partir d'un couple d'images. Dans le deuxième exemple, le contexte spatial de chaque pixel est injecté dans un autre classificateur non-linéaire pour obtenir une carte précise des nouvelles structures urbaines. Dans les deux cas, l'utilisateur fournit aux classificateurs des exemples de ce qu'il croit avoir pas changé ou non.

Dans la deuxième partie, une approche complètement automatique et non-dirigée est proposée pour la détection binaire. Cette méthode est particulièrement précise sans nécessiter l'intervention de l'utilisateur. Un tel algorithme se révèle utile quand le temps de réaction du système est réduit.

Dans la troisième partie, le problème des distributions statistiques qui changent d'une acquisition à l'autre pour des classes stables dans le temps est abordé. Les deux méthodes présentées alignent ces distributions de façon à améliorer la précision de la comparaison par pixels pour détecter les changements. De plus la deuxième méthode est capable de traiter des images avec des différentes dimensionalités et informations spectrales. Cela permet d'envisager des pistes de solutions au problème crucial de la détection de changements dans des images provenant de capteurs différents.

“ [...] car rien ne se crée, ni dans les opérations de l’art, ni dans celles de la nature, et l’on peut poser en principe que, dans toute opération, il y a une égale quantité de matière avant et après l’opération; que la qualité et la quantité des principes est la même, et qu’il n’y a que des changements, des modifications.”

Antoine Laurent de Lavoisier,
Traité élémentaire de chimie (1789), p. 101.

Acknowledgements

The accomplishment this Thesis would not have been possible without the priceless support and help received from others, whether scientific or not.

First of all, I would like to thank my supervisors Prof. Mikhail Kanevski and Dr. Devis Tuia. Mikhail believed in me since my Master project and gave me the opportunity of pursue a Ph.D. He gave me invaluable advices on how to do scientific research and how to understand the academic life. I met Devis in the same period, when he co-advised my Master project. He coached me with the research, writing papers, teaching assistance classes, how to manage efforts. I am really honoured that I had these advisors.

Some parts of this Thesis have been done during research visits at the University of Trento and at the University of València. I would like to express my deepest gratitude to the people who made possible these stays, and in particular those who advised me. Thanks to Dr. Francesca Bovolo and Prof. Lorenzo Bruzzone of the University of Trento, and to all the RSlab team. Then it came the great IPL lab in València: exponential thanks to Prof. Gustavo Camps-Valls (who additionally had to read all these pages), Prof. Jordi Muñoz-Marí and Dr. Luis Gómez-Chova. Thanks for all the scientific and non-scientific discussions, for letting me discover the true paella and for teaching how to prepare the tortilla. I'm very looking forward to see you all soon again.

I would like also to thank all the colleagues and friends from the University of Lausanne, who have been always so kind to never refuse going out for a beer.

Last but not least, there is a number of people which have nothing to do with my research, but have lots to do with my life. I would like to thanks my mother Cristina and my father Giuseppe for the long lasting support (both moral and financial!) during my studies. I would like to thank my cool bro Stefano, great brother and great friend. And thank you Sara ♡, you followed me during this Ph.D. adventure and you will probably follow me even farther. Brave girl!

Then, thanks to all my friends who gave me good times and the opportunities to chill, relax and having fun during these years. In particular, I would like to thank my great friends of the “Localino” team, Mosi, Nico, Steff, Crivi, Copa, Giò, Ricks, Massa, and the all the “Valgi” friends.

Michele Volpi, September 2013

Financial Support

The author of the Thesis would like to kindly acknowledge the SNSF for the financial support. This thesis has been supported by two Swiss National Science Foundation projects, (kernelcd and kernelcd II) “Change detection in remote sensing images using kernel-based machine learning algorithms” (see www.kernelcd.org), under the grants number 200021-126505 and 200020-144135. Co-advisor Dr. Devis Tuia was sponsored by the SNSF grant Ambizione PZ00P2-136827. I would also thank the support of Prof. Kanevski, Dr. Tuia and in particular Dr. Vittorio Ferrari at the CALVIN Lab (University of Edinburgh) for helping me in obtaining the SNSF Early PostDoc Mobility grant “Robust segmentation-based methods for remote sensing VHR image analysis”, number P2LAP2-148432.

Contents

List of Figures	xiii
List of Tables	xv
I Introduction	1
1 Introduction to the Thesis	3
1.1 Motivation	3
1.2 Objectives	4
1.3 Contributions	5
1.3.1 Chapter 6	6
1.3.2 Chapter 7	6
1.3.3 Chapter 8	7
1.3.4 Other unrelated yet related work	7
1.4 Outline	8
2 Introduction to remote sensing	9
2.1 An overview of the acquisition systems	9
2.1.1 The EM radiations in the optical regime	10
2.1.2 Spectral signatures: characterizing materials	14
2.2 Optical remote sensing systems	16
2.2.1 A characterization of optical sensors	17
2.2.2 The optical data as grayscale images	19
2.3 Change detection in remote sensing data	21
2.3.1 Standard approaches to change detection	22
2.3.2 Geometrical requirements for change detection	24
2.3.3 Spectral and radiometric requirements for change detection	24
2.4 Some considerations	26

Contents

II	Machine learning and kernel-based algorithms	29
3	Machine learning	31
3.1	Learning from data	31
3.1.1	A practical example	35
3.2	Connections with regularization theory	38
3.3	Hyperparameters optimization	40
3.4	Models of machine learning	41
3.4.1	Parametric and non-parametric inference	41
3.4.2	Supervised, unsupervised and semi-supervised models	42
3.4.3	Linear and nonlinear models for data analysis	45
4	Learning with kernels	47
4.1	From least-squares to kernel ridge regression	47
4.2	Kernel methods: theory and regularization	50
4.2.1	Reproducing kernel Hilbert space	51
4.2.2	Operations in the RKHS	52
4.2.3	The kernel functions	54
4.2.4	Ad hoc-kernel functions and closure properties	56
4.2.5	The Gram matrix	57
4.2.6	On the choice of the kernel function and its parameters	57
4.3	Some considerations	58
III	Kernel-based methods for change detection	61
5	State-of-the-art in change detection	63
5.1	Learning from pixels	63
5.2	Supervised change detection	65
5.2.1	Post classification comparison	65
5.2.2	Direct multi-date classification	66
5.2.3	Supervised difference image analysis	67
5.3	Automatic and unsupervised change detection	67
5.3.1	Clustering and unsupervised classification	68
5.3.2	Novelty detection	69
5.4	Feature extraction for multi-temporal applications	70
5.5	Some considerations	71
6	Supervised change detection	73
6.1	Supervised change detection for monitoring	73
6.2	Supervised flooded area extraction	74
6.2.1	The regularized kernel Fisher's discriminant classifier	74
6.2.2	Experimental setup	78

6.2.3	Results	79
6.2.4	Discussion	81
6.3	Supervised change detection for urban monitoring	85
6.3.1	The support vector machines for classification	85
6.3.2	Textural features	88
6.3.3	Mathematical morphology	90
6.3.4	Experimental setup	94
6.3.5	Results	96
6.3.6	Discussion	100
6.4	Conclusions	101
7	Unsupervised change detection	103
7.1	Clustering for automatic change detection	103
7.2	The proposed unsupervised kernel-based change detection scheme	104
7.2.1	A partitioning algorithm: the kernel k -means	104
7.2.2	The initialization	105
7.2.3	The unsupervised cost function	106
7.2.4	Feature maps	107
7.3	Experimental setup	109
7.4	Results and experimental validation	110
7.4.1	Case studies	110
7.4.2	The cost function	114
7.4.3	Cluster separability	116
7.5	Conclusions	117
8	Feature extraction for change detection	119
8.1	Adjusting radiometric differences	119
8.2	Relative radiometric normalization using kernels	120
8.2.1	The kernel principal component analysis	121
8.2.2	Multivariate alignment for change detection	123
8.2.3	Experimental setup	124
8.2.4	Results	126
8.2.5	Discussion	130
8.3	Multi-sensor alignment for change detection	131
8.3.1	Regularized kernel canonical correlation analysis	131
8.3.2	Semi-supervised relative alignment via manifold regularization	134
8.3.3	Heterogeneous alignment for change detection	136
8.3.4	Discussion	140
8.4	Conclusions	142

Contents

IV	Conclusions	145
9	Conclusions	147
9.1	A new generation of change detection systems	147
9.1.1	On supervised change detection	147
9.1.2	On unsupervised change detection	148
9.1.3	On multi-sensor change detection	149
9.2	Contributions of the Thesis	149
9.3	Future perspectives	150
	Appendices	153
A	The learning sets	155
B	Accuracy evaluation	157
C	Datasets	161
C.1	Brüttisellen	161
C.2	Steinacker	163
C.3	Missouri flooding	164
C.4	Gloucester flooding	165
C.5	Brüttisellen 2 dataset	166
C.6	Greek island forest fire	167
C.7	Greek fires dataset	168
	References	169

List of Figures

2.1	The electromagnetic spectrum	10
2.2	Atmospheric transmittance	12
2.3	Solar irradiance	13
2.4	Example of spectral signatures	15
2.5	RS image evolution	17
3.1	Bias-variance dilemma	34
3.2	Structural risk minimization	36
3.3	k NN classification errors	37
3.4	Bayes optimal classification	38
3.5	GMM-based clustering	44
3.6	Semi-supervised classification toy example	46
4.1	Polynomial kernel map	55
4.2	Sigma parameter of Gaussian RBF kernel	59
5.1	PCC, DMC and DIA schemes	64
6.1	FDA graphical interpretation	76
6.2	Supervised flood mapping results	80
6.3	Scatterplot matrix of the multi-temporal Landsat TM flooding data	83
6.4	Subsets of the Landsat TM flooding scene	84
6.5	SVM graphical interpretation	87
6.6	Multi-scale occurrence texture statistic	89
6.7	Multi-scale co-occurrence (GLCM) texture statistic	90
6.8	Multi-scale opening and closing morphological operators	91
6.9	Multi-scale opening and closing by reconstruction morphological operators	92
6.10	Multi-scale classwise multi-temporal signal	93
6.11	Increased separability of similar classes	95
6.12	Details of the Brüttisellen change detection maps.	96
6.13	Test accuracies for urban monitoring datasets	97
6.14	Details of the Steinacker change detection maps	99
6.15	Outcomes of the McNemar tests for the urban monitoring datasets	100

List of Figures

7.1	Magnitude-based initialization of the kernel k -means	106
7.2	The block diagram of the proposed change detection scheme	109
7.3	Change maps for the kernel-based automatic framework	113
7.4	ROC curves for the three datasets with the automatic approach	114
7.5	Unsupervised cost function example	115
7.6	Separability of clusters in the input space and in the RKHS	117
8.1	Statistically aligned images using the kPCA approach	124
8.2	kPCA-based statistical relative radiometric normalization	125
8.3	Scatterplot of the difference image without and with kPCA alignment . . .	126
8.4	CVA magnitude with histogram matching and with kPCA alignment	127
8.5	Change detection accuracies for the kPCA-based relative normalization . .	128
8.6	Change maps with histogram matching and kPCA alignment	129
8.7	Change detection maps for the tested SSkCCA alignment approaches	139
8.8	Dependence on the number of labelled and unlabelled pixels	140
8.9	Accuracy as a function of the dimensionality of the projections	141
8.10	Accuracy as a function of the dimensionality of the projections	142
C.1	The Brüttisellen dataset	162
C.2	The Steinacker dataset	163
C.3	Subsets of the Landsat TM James River flooding	164
C.4	The Gloucester subset	165
C.5	The Brüttisellen 2 dataset	166
C.6	The Greek Island dataset	167
C.7	Greece fires dataset	168

List of Tables

2.1	Categorization of sensors based on the number of spectral channels	18
2.2	Categorization of sensors based on the spatial resolution	19
6.1	Average error matrices for flood mapping problem	81
6.2	Contextual information feature blocks	94
7.1	Figures of merit for the Gloucester dataset	111
7.2	Figures of merit for the Brüttisellen 2 dataset	112
7.3	Figures of merit for the Greek Island dataset	114
8.1	Multi-source change detection results	137
B.1	Confusion matrix	157

Glossary

Glossary

AUC	Area under the ROC curve. See ROC curve. The AUC is the area under the ROC in $[0, 1]$, page 158
AVIRIS	Airborne visible and infrared imaging spectrometer. NASA operated airborne imaging spectrometer, page 16
BRDF	Bidirectional reflectance density function. Function describing the light reflection by an opaque surface, page 14
CCA	Canonical correlation analysis. A feature extraction method mapping two different feature sets of the same samples into a subspace in which they are maximally correlated, page 120
CD	Change detection. The process of identifying ground cover transitions in a couple (or series) of remote sensing images of the same geographical area acquired at different time instants, page 21
CVA	Change vector analysis. Change detection method relying on the threshold of the norm and angle distributions of the spectral change vectors derived from the difference image, page 23
DIA	Difference image analysis. Approach to bi-temporal change detection involving the analysis of the difference image, page 65
DMC	Direct multivariate classification. Direct supervised classification of the stack of multi-temporal images, page 65
DN	Digital Number. Values indicating average radiance for each pixel, corresponding to quantized electric signals from the sensor, page 14
EM	Electromagnetic energy. Qualifying one of the fundamental interactions occurring in nature between electrically charged particles, page 9
EO	Earth observation. Science studying Earth processes by remote and non-intrusive observations, page 10
ERM	Empirical risk minimization. Direct minimization of the loss function during the training of a model, from the statistical learning theory [Vapnik, 1998], page 32
ETM+	Enhanced thematic mapper plus. Sensor used in Landsat programs 6 (lost at launch) and 7, page 18
FOV	Field of View. Angle covering the image size in the cross-track direction, page 17
GFOV	Ground-projected field of view. Ground projection of cross-track section (angle) of a sensor, page 17
GIFOV	Ground-projected instantaneous field of view. Cross-track size of the IFOV projection to the ground, pixel size in [m], page 16
GIS	Geographical information system. Computer-based environment for the analysis and processing of geographically referenced and spatial data, page 65
GMM	Gaussian mixture models. Parametric and generative method to clustering, based on the estimation of a mixture of k Gaussians, page 43
GPS	Global positioning system. A satellite constellation for absolute geographical coordinate retrieval, page 9
GSD	Ground sample distance. See GIFOV, page 18
GSI	Ground sample interval. Spacing between the centres of two adjacent pixels, page 18

Glossary

IFOV	Instantaneous field of view. Smallest solid angle in which radiations are measurable by the sensor. It defines the pixel size, see GIFOV, page 12	NIR	Near infrared. Shortest wavelengths composing the infrared radiation, behaving as visible light (surface reflection), from 0.7 to 2.5 [μm], page 10
IR	Infrared. EM radiations from 0.7[μm] to 1[mm], page 18	NMI	Normalized mutual information. Information theoretic multi-class measure of agreement, page 158
kCCA	Kernel canonical correlation analysis. The nonlinear (kernel-based) extension of the canonical correlation analysis (see CCA), page 120	OA	Overall accuracy. Proportion of correctly classified samples over the total, page 157
KDE	Kernel density estimation. A non-parametric (data driven) model of density estimation, page 42	OAA	One-against-all. Technique to solve multi-class problems by dividing the classification into $ Y $ binary sub-problems, page 88
kFDA	Kernel Fisher's discriminant analysis. A supervised method for the extraction of class-wise discriminant features and their classification, page 73	OAo	One-against-one. Technique to solve multi-class problems by dividing the classification into $ Y (Y -1)/2$ binary sub-problems., page 88
KkM	Kernel k-means. A clustering algorithm based on the iterative partitioning of the dataset into k groups, page 104	PCC	Post classification comparison. A change detection approach involving the independent classification of the multi-temporal images and the consequent logical comparison of the obtained maps, page 65
KL	Kullback-Leibler distance. A distance between probability density distributions, page 43	RBF	Radial basis function. Gaussian-like bell-shaped multi-variate function, page 55
kPCA	Kernel principal component analysis. The nonlinear (kernel-based) extension of the principal component analysis (see PCA), page 120	RKHS	Reproducing kernel Hilbert space. A Hilbert space endowed with the reproducing property, page 51
LOO-CV	Leave-one-out cross-validation. A method to approximate the generalization error using the training set, used to set hyperparameters, page 40	ROC	Receiver operating characteristic curve. It evaluates the performance of a binary classifier by plotting the true positive / false positive rates for varying decision thresholds, page 158
ML	Machine learning. See Chapter 3, page 27	ROSiS	Reflective optics system imaging spectrometer. Airborne imaging spectrometer (hyperspectral) sensor, page 18
MLC	Maximum likelihood classifier. Parametric and generative supervised classifier based on the assumption of Gaussian-distributed classes, page 65	SAR	Synthetic aperture radar. Active, microwave sensor, page 9
NDVI	Normalized difference vegetation index. Vegetation index reacting in the presence of vegetation. It is correlated to the amount of vegetation in each pixel, page 64	SNR	Signal-to-noise ratio. A measure of the quality of the signal, estimating the proportion of the signal versus noise, page 20

SRM	Structural risk minimization. The process of minimizing a loss functional under some capacity constraint, from the statistical learning theory [Vapnik, 1998], page 34		learning theory, page 52
SVDD	Support vector domain description. One-class classification method fitting a minimum radius ball enclosing the target class or the background, page 69	TIR	Thermal infrared. Low energy infrared radiations emitted by materials with an intensity proportional to material temperature, page 10
SVM	Support vector machines. A classifier inspired from the Vapnik's statistical learning theory. It belongs to the family of large margin classifiers, page 52	TM	Thematic mapper. Sensor used in Landsat programs 4 and 5, page 19
SVR	Support vector machines for regression. A regression method inspired from the Vapnik's statistical	VHR	Very high spatial resolution. Imagery with meter- or sub-meter spatial resolution, see Tab. 2.2, page 18
		VIS	Visible. Part of spectrum seen by human eye, from 0.4 to 0.7 [μm], page 10
		VNIR	Visible and near-infrared. Part of the spectrum related to reflection of EM energy emitted by the sun, from 0.4 to 2.5 [μm], page 10

List of symbols

Scalars, vectors and matrices			
n	number of samples	d	number of dimensions
\mathbf{x}_i	i th pixel vector, $\in \mathbb{R}^{d \times 1}$	x_{ib}	i th pixel at b th channel
Y	set of labels	\mathbf{y}	vector of labels
y_i	i th label corresponding to \mathbf{x}_i	$ Y $	number of classes
X^{t_1, t_2}	image at time t_1 or t_2	$\mathbf{X}_{1,2}$	data matrix of X^{t_1, t_2} , $\in \mathbb{R}^{n \times d}$
\mathbf{w}	primal weight vector	$\boldsymbol{\alpha}$	dual weight vector
Φ	data matrix of mapped samples	$\mathbf{w}^{\mathcal{H}}$	primal weight vector in \mathcal{H}
\mathbf{D}	difference image	$\mathbf{0}$	vector of n zeros
\mathbf{K}	kernel matrix, Gram matrix	\mathbf{S}	Scatter matrix
$\mathbf{1}$	vector of n ones	\mathbf{I}	$n \times n$ identity matrix
$\mathbf{1}_c$	$n_c \times 1$ vector with entries $1/n_c$	\mathbf{I}_c	$n_c \times n_c$ matrix with entries $1/n_c$
$\boldsymbol{\mu}$	sample average	$\boldsymbol{\mu}^{\mathcal{H}}$	sample average in \mathcal{H}
$\boldsymbol{\Sigma}$	sample covariance	λ	Lagrange multipliers
s_c	scatter of class c	\mathcal{M}	the graph Laplacian
Θ	model parameters	Θ_h	model hyperparameters
σ	Gaussian kernel bandwidth	\mathbf{m}_c	projected mean of class c
b	bias of a classifier	$\boldsymbol{\xi}$	slack variables
γ	regularisation parameter		

Functions			
$f, f(\cdot)$	general learning function	$g, g(\cdot)$	general learning function
$k(\cdot, \cdot)$	kernel function	$L(\cdot)$	Lagrangian problem
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$	$\mathbb{E}(\cdot)$	expectation operator
$p(\cdot)$	probability density function	$\mathbb{P}(\cdot)$	probability distribution
$\langle \cdot, \cdot \rangle$	inner product	$\mathcal{L}(\cdot)$	loss function
$\Omega(\cdot)$	regularisation function	$\ \cdot\ _p$	ℓ_p -norm
$\phi(\cdot)$	mapping function	$\varphi(\cdot)$	mapping function

Other symbols			
\mathcal{X}	input space	\mathcal{Y}	output space
\mathbb{R}	field of real numbers	\mathbb{R}^d	product of d \mathbb{R} sets
\mathcal{F}	hypothesis space of functions f	\mathcal{H}	reproducing kernel Hilbert space

Part I

Introduction

Chapter 1

Introduction to the Thesis

1.1 Motivation

Since the advent of new generation satellites, the science of Earth observation has known an unprecedented progress. Images are acquired and stored in archives for future use and their joint exploitation allows a very precise temporal and geographical monitoring of the evolution of the surface of the Earth.

Remote sensing images are encountered in many different aspects of daily life: from the visualization to recover the path to a friend's house, to the extraction of physical parameters used in numerical models providing scenarios of climate change. Independently on the degree and complexity of the use, remote sensing image processing is central in each one of these tasks. From the acquisition of the image to the delivery of a product (e.g. a map) the analyst relies on those methods to transform, enhance and process the datasets.

In this Thesis, we tackle the topic of multi-temporal processing and change detection in optical remote sensing images. By multi-temporal processing we intend all the tools that are explicitly designed to account for the temporal component of the image data, during their processing. In this sense, change detection is a particular instance of this family of methods, and it aims at detecting and mapping the changes occurred in the ground cover between the considered acquisitions over the same geographical area. High societal value applications such as environmental and urban monitoring, post-catastrophe assessments, natural hazard quantification, crop monitoring and surveillance application are greatly dependent on the methods used for the multi-temporal image analysis and change detection.

The above observations underline the diversity of the application fields in which those methods have to be applied, but a common observation joins them: the manual screening of the images to map the differences is not a feasible option. To this end, automatic methods are truly needed. These approaches should be able to process newly acquired data, but also have to solve problems requiring the use of older data stored in archives. In particular, if one may want to study the evolution of the ground cover of a particular region, then archives are a primary source of information. They may also provide some additional

1. Introduction to the Thesis

information prior to the processing, useful to drive the analyses. The user is expecting to be able to apply methods providing very accurate solutions, so that his particular monitoring application will be correctly carried out. In other situations, the practitioner dealing with change detection may not dispose of information about the processes occurring at the surface of the Earth, but still require an automatic precise cartography of changes. Furthermore, in particular situations such as post catastrophe assessment or natural hazard quantification, the time available for the analyses may be a constrain for the processing. Consequently, the user may want to apply methods specifically developed to provide high readiness of the system while accurate enough to correctly map changes.

To obtain the most accurate detection even in the most challenging situation, we take advantage of the recent developments in a domain of computer sciences intermingled with mathematics and statistics, known as machine learning. Theoretical advances in data analysis as the ones provided by this field are of great interest in many disciplines, ranging from economy to biology [Blundell and Duncan, 1998; Camps-Valls et al., 2007b; Keshet and Bengio, 2008; Schölkopf et al., 2004], and they are also of great interest also for remote sensing image processing tasks [Camps-Valls and Bruzzone, 2009]. These methods are able to learn a model from the data and their interrelations, thus not requiring computationally heavy numerical simulations and physical models. Moreover, they provide tools able to solve many analysis situations (classification, function estimation, extraction of relevant information, etc.), such as those considered in this Thesis.

Specifically, we aim at relating more closely the field of change detection and a specific field of the machine learning research: the kernel methods. This family of algorithms allows elegant and robust solutions for most of the multi-temporal processing tasks we are interested to, and fit well the many open issues in remote sensing data analysis, such as change detection between multiple sensors, accurate and automatic partitioning of changes and precise monitoring. Kernel methods provide a common modelling solution to all these problems making a simple assumption: similarity between pixels is the only information needed.

1.2 Objectives

This Thesis aims at relating more closely the field of kernel methods to multi-temporal image processing tasks. We believe that this framework is robust and flexible enough to positively contribute with methods able to encode and exploit the versatile nature of remote sensing data. In particular, we aim at proposing solutions to both supervised and automatic (unsupervised) change detection. Although in the machine learning literature a variety of benchmark problems are efficiently and accurately solved by kernel methods, only few contributions are found in the field of multi-temporal processing and change detection. Thus, the general objectives of the Thesis are twofold: contribute in the theoretical development of kernel methods for change detection tasks and provide real solutions to two main families of problems encountered: supervised and unsupervised change detection problems.

The first family is intended for monitoring purposes, in which the accuracy of the products is of central importance and the computational time is not in general a limiting factor. The adopted system, to guarantee the maximal precision of the final thematic product, must be able to deal with heterogeneously distributed classes in possibly high dimensional spaces, a setting that usually lowers the performance of standard algorithms. In this case, one is able to retrieve sub-meter resolution maps exhaustively summarizing the observed processes, that may be composed of distinct class transitions and permanent (stable) ground cover classes. These products are usually employed in consequent analyses, ranging from the study of ecological systems to the mapping of re-/de-forestation. The scientists involved often extrapolate additional information from these maps, for instance to parametrize other models or describing particular phenomena, e.g. the spread rate of invasive weed species. It clearly appears that an accurate mapping is needed to fully support those extrapolations. Consequently, one objective of the Thesis is to develop kernel-based systems for accurate supervised change detection.

Conversely, other applications may require a rapid mapping of changes, in which one does not dispose of examples exploitable to learn models and the time available for the mapping is limited. In this case, one usually looks for a binary mask, indicating whether a pixel has changed or not, without having any prior information about the location and the type of transition that may have occurred between the acquisitions. Phenomena such as earthquakes, tsunamis, landslides, avalanches and many other processes generating abrupt changes may generate modifications of the landscape and damages to human infrastructures. Consequently, the user has no access to information (ground reference data) to initialize or validate the adopted change detection methods. In these cases, a change detection system should be able to provide within a short time interval highly reliable (thus accurate) change masks. These are to be used to either support rescue teams, assess and rapidly quantify the damages or plan the physical access to the involved regions without disposing of anything but the couple of pre- and post-event images. When facing such applications of change detection the maps have to be obtained by completely automatic methods, allowing also inexperienced users to use them. The second domain in which the Thesis aims at contributing is the automatic and unsupervised processing for change detection.

The above objectives resulted in three main contributions of the Thesis, as depicted in the next Section.

1.3 Contributions

In the Part III of the Thesis, the main contributions are presented. Here, we briefly recall the main points of each and list the publications and conference proceedings related to each topic.

1. Introduction to the Thesis

1.3.1 Chapter 6

This Chapter proposes, studies and evaluates supervised approaches for precise monitoring of natural and urban processes. In a first application, the supervised kernel Fisher's discriminant analysis classifier is studied with the aim of flood mapping. In a second study support vectors machines are used for supervised change detection and multi-temporal classification of urban scenes. In the latter, the use of very high spatial resolution data required the adoption of spatial regularization schemes. To this end, spatial context features of different nature are studied and evaluated in two change detection schemes.

The Chapter is based directly and indirectly on the following works:

[Volpi et al., 2009] **Volpi, M.**; Tuia, D.; Kanevski, M.; Bovolo, F. & Bruzzone, L.; Supervised Change Detection in VHR Images: a Comparative Analysis; In *IEEE International Workshop on Machine Learning for Signal Processing MLSP 2009, Grenoble (F)*, pp. 1-6, **2009**.

[Volpi et al., 2013c] **Volpi, M.**; Tuia, D.; Kanevski, M.; Bovolo, F. & Bruzzone, L.; Supervised Change Detection in VHR Images Using Contextual Information and Support Vector Machines, *International Journal of Applied Earth Observation and Geoinformation*, vol. 20, pp. 77-85, **2013a**.

[Volpi et al., 2013d] **Volpi, M.**; Petropoulos, G. P. & Kanevski, M.; Flooding Extent Cartography with Landsat TM Imagery and Regularized Kernel Fisher's Discriminant Analysis, *Computers and Geosciences*, vol. 57, pp. 24-31. **2013b**.

[Longbotham et al., 2012] Longbotham, N.; Pacifici, F.; Glenn, T.; Zare, A.; **Volpi, M.**; Tuia, D.; Christophe, E.; Michel, J.; Inglada, J.; Chanussot, J. & Du, Q.; Multi-modal Change Detection, Application to the Detection of Flooded Areas: Outcome of the 2009-2010 Data Fusion Contest, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 6, pp. 331-342, **2012**.

1.3.2 Chapter 7

This Chapter presents and validates an unsupervised approach to automatic change detection. The standard difference image is reformulated into higher dimensional feature spaces, the reproducing kernel Hilbert spaces, and a difference kernel defined to implicitly work in that space is exploited. Additionally, to tackle the issue of tuning the hyperparameters, a completely automatic cost function inspired from the geometry of the problem has been developed.

The Chapter is based directly and indirectly on the following works:

[Volpi et al., 2010a] **Volpi, M.**; Tuia, D.; Camps-Valls, G. & Kanevski, M.; Unsupervised change detection by kernel clustering, In *SPIE Image and Signal Processing for Remote Sensing XVI, Toulouse (F)*, 7830, **2010**.

[Volpi et al., 2011] **Volpi, M.**; Tuia, D.; Camps-Valls, G. & Kanevski, M.; Unsupervised Change Detection in the feature space using kernels, In *IEEE International Geosciences and Remote Sensing Symposium IGARSS, Vancouver (CAN)*, pp. 106-109, **2011**.

[Volpi et al., 2012b] **Volpi, M.**; Tuia, D.; Camps-Valls, G. & Kanevski, M.; Unsupervised change detection with kernels, *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 9, pp. 1026-1030, **2012a**.

1.3.3 Chapter 8

This Chapter studies a kernel-based feature extraction framework to improve the change detection process. Two different cases are presented: in the first, the use of a standard kernel-based feature extraction method allows a simple yet effective alignment of the statistical distribution of unchanged samples prior to the detection of changes. In the second, an extension of the above reasoning using a different kernel method yields to a system allowing the projection of heterogeneous images into a common subspace, thus permitting to perform change detection between two different sensors.

The Chapter is based directly and indirectly on the following works:

[Volpi et al., 2012a] **Volpi, M.**; Matasci, G.; Tuia, D. & Kanevski, M.; Enhanced change detection using nonlinear feature extraction, In *IEEE International Geosciences and Remote Sensing Symposium IGARSS, Munich (D)*; pp. 6757-6760, **2012b**.

[Volpi et al., 2013a] **Volpi, M.**; de Morsier, F.; Camps-Valls, G.; Kanevski, M. & Tuia, D.; Multi-sensor change detection based on nonlinear canonical correlations, In *IEEE International Geosciences and Remote Sensing Symposium IGARSS, Melbourne (AUS)*, **2013c**.

[Volpi et al., 2013b] **Volpi, M.**; Matasci, G.; Kanevski, M. & Tuia, D.; Multi-view feature extraction for hyperspectral image classification, In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning ESANN, Bruges (B)*, pp. 11-16, **2013d**.

[Matasci et al., 2011] Matasci, G.; **Volpi, M.**; Tuia, D. & Kanevski, M.; Transfer Component Analysis for Domain Adaptation in Image Classification, In *SPIE Image and Signal Processing for Remote Sensing XVII, Prague (CZ)*, 8180, **2011**.

[Trolliet et al., 2013] Trolliet, M.; Tuia, D. & **Volpi, M.**; Classification of urban multi-angular image sequences by aligning their manifolds, In *Joint Urban Remote Sensing Event, Sao Paolo (BRA)*, **2013**.

[Matasci et al., 2013] Matasci, G.; Bruzzone, L.; **Volpi, M.**; Tuia, D. & Kanevski, M.; Investigating feature extraction for domain adaptation in remote sensing image classification, In *International Conference on Pattern Recognition Application and Methods ICPRAM, Barcelona (SP)*, **2013**.

[Tuia et al., 2013a] Tuia, D.; Trolliet, M. & **Volpi, M.**; Multisensor alignment of image manifolds, In *IEEE International Geosciences and Remote Sensing Symposium, Melbourne (AUS)*, **2013**.

1.3.4 Other unrelated yet related work

Besides change detection and multi-temporal image classification, the Thesis project allowed also to contribute in other remote sensing processing studies. In particular, topics

1. Introduction to the Thesis

related to hyper- and multi-spectral image thematic classification, specifically active learning and feature learning, were also explored.

- [Volpi et al., 2010b] **Volpi, M.**; Tuia, D.; Kanevski, M.; Advanced Active Sampling for Remote Sensing Image Classification, In *IEEE International Geosciences and Remote Sensing Symposium IGARSS, Honolulu (USA)*, pp. 1414-1417, **2010**.
- [Volpi et al., 2012c] **Volpi, M.**; Tuia, D. & Kanevski, M.: Memory-Based Cluster Sampling for Remote Sensing Image Classification, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 8, 3096-3016, **2012**.
- [Copa et al., 2010] Copa, L.; Tuia, D.; **Volpi, M.** & Kanevski, M.; Unbiased query-by-bagging active learning for VHR image classification, In *SPIE Image and Signal Processing for Remote Sensing XVI, Toulouse (F)*, 7830, **2010**.
- [Tuia et al., 2011] Tuia, D.; **Volpi, M.**; Copa, L.; Kanevski, M. & Muñoz-Marí, J.; A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification, *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 606-617, **2011**.
- [Tuia et al., 2012] Tuia, D.; **Volpi, M.**; Dalla Mura, M.; Rakotomamonjy, A. & Flamary, R.; Discovering relevant spatial filterbanks for VHR image classification, In *International Conference on Pattern Recognition ICPR, Tsukuba (JAP)*, **2012**.
- [Tuia et al., 2013b] Tuia, D.; **Volpi, M.**; Dalla Mura, M.; Rakotomamonjy, A. & Flamary, R.; Create the relevant spatial filterbank in the hyperspectral jungle, In *IEEE International Geosciences and Remote Sensing Symposium IGARSS, Melbourne (AUS)*, **2013**.
- [Penna et al., 2013] Penna, I. M.; Derron, M.-H.; **Volpi, M.** & Jaboyedoff, M.; Analysis of past and future dam formation and failure in the Santa Cruz River (San Juan province, Argentina), *Geomorphology*, vol. 186, pp. 28-30, **2013**.

1.4 Outline

The Thesis is organized in four parts. In Part I, Chapter 2 provides a general introduction to the field of remote sensing and to the types of imagery derived from optical sensors. In Part II, Chapter 3 provides general concepts of machine learning. In Chapter 4, the family of kernel methods is presented. Finally, Part III reviews the main contributions of the Thesis. Chapter 5 presents principal elements of the state-of-the-art literature in change detection and multi-temporal processing. Chapter 6 illustrates the supervised methods. Chapter 7 presents the developed unsupervised approaches, while Chapter 8 explain the feature extraction-based methods for the alignment of unchanged spectral information. Finally, Part IV concludes the Thesis. Chapter 9 summarizes the main results and contributions, and it states possible future research directions in the field of remote sensing image analysis.

Chapter 2

Introduction to remote sensing imagery and to change detection

This Chapter introduces the basic notions of remote sensing imaging. Section 2 recalls the principles of electromagnetic radiation for optical remote sensing, Section 2.2 characterizes passive remote sensing systems and Section 2.3 illustrates change detection in optical data. In the latter, the reader is introduced to change detection by exploiting the concept of difference image, and the main preprocessing considerations are extrapolated from this basic but universal representation.

2.1 An overview of the acquisition systems

Remote sensing may be defined as the ensemble of the technologies, analogical or digital, allowing the distant acquisition of informations about an object or a process of interests. Therefore, the term remote sensing could refer to different systems, acquiring signals as diverse as from differential GPS¹ system for precise geographical coordinate retrieval, to microwave sounding of the atmosphere. Remote acquisitions may be consequently performed by ground networks, by aircraft (airborne) or by satellites (spaceborne). In this Thesis, we will refer to remote sensing as the ensemble of airborne or spaceborne technologies permitting the collection of imagery of the Earth surface. In this sense, two different families are distinguished: active and passive sensors [Lillesand et al., 2004; Schowengerdt, 2007; Woodhouse, 2006].

Active systems are imaging sensors that process electromagnetic energy (EM) emitted by an antenna, usually in the microwave region of the spectrum, as illustrated in Figure 2.1. The system interprets the sensed energy reflected back to the receiver, after interacting with the surface of the Earth, to form an interpretable signal: the radar image. The most advanced radar imaging systems are the synthetic aperture radar (SAR), that, in contrast to real aperture radars, exploits the movement of the sensor carrier to form much

¹For all the abbreviations, see the glossary on page xix

2. Introduction to remote sensing

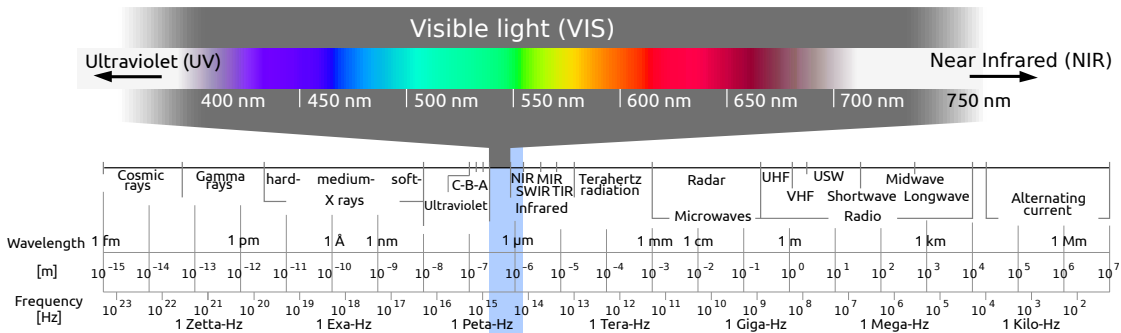


Figure 2.1: The electromagnetic spectrum - Electromagnetic spectrum and its principal characteristics, with emphasis on the visible region. The light blue vertical shade represents the optical regime.

higher spatial resolution images. It is worth mentioning that radar acquisitions are almost insensible to atmospheric conditions and to sun illumination. For additional details, see [Jensen, 2007; Woodhouse, 2006]. On the contrary, passive systems exploit the energy coming from the Earth, that is either composed by reflected radiations from the Sun or heat emitted by the surface, to create a spatial representation of it: the optical image.

In the next Sections, the origin of the spectral signatures is summarized by recalling the main physical properties of the EM spectrum, how it interacts with the atmosphere and finally by describing its interplay with the surface of our planet.

2.1.1 The EM radiations in the optical regime

The sun, thanks to complex thermo-nuclear processes, emits energy in the full EM spectrum. This radiation propagates throughout the space until an interaction with it occurs, e.g. by the atmosphere of a planet. These radiations are characterized by properties such as the wavelength (λ), frequency (ν) and amplitude. The velocity of propagation in the vacuum is the speed of light $c = 299'792.5 \cdot 10^3$ [m/s], and in the Earth atmosphere the attenuation is negligible. These quantities are related by the fundamental equation $c = \lambda\nu$.

Optical remote sensing-based Earth observation (EO) studies the spectral signatures of the materials contained in each pixel, the smallest spatial element composing an image. It is usually defined through the visible (VIS) and near infrared (NIR) to the thermal infrared (TIR). In this introduction we will mainly focus on the reflective portion of the spectrum, defined in the VIS and NIR wavelengths (VNIR). The most of the images processed and used in this thesis do not include thermal channels.

The radiations considered here cover only a small portion of the EM spectrum that comes towards Earth, as illustrated in Figure 2.1¹. This interval is composed by two distinct physical behaviours: the VNIR part of the spectrum is called solar-reflective region. The considered VNIR EM interval is reflected by most of the materials at the

¹Modified from:

http://upload.wikimedia.org/wikipedia/commons/0/00/Electromagnetic_spectrum_sRGB.svg

2.1 An overview of the acquisition systems

surface of the Earth, and in the visible wavelengths corresponds to the perceived colors. Although not recognized by the human eye, the NIR behaves similarly to visible light in terms of reflections. The second category refers to mid- to thermal infrared radiations, corresponding to the emission of heat by the surface of the Earth and the objects on it. This last category is only marginally influenced by the direct reflection of solar radiations, the only exception being found when objects behaves as specular reflectors, that is, the solar radiation is redirected directly to the remote sensor.

The reflected EM energy that reaches the sensor can be decomposed in different processes. They are caused by interactions that the light experiences in its path to and from the Earth surface, as well as the processes that take place when the radiations hit the surface.

The interaction between electromagnetic energy and the atmosphere. The very first obstacle that the light encounters in its path to the Earth surface are the atmospheric layers. The atmosphere, being composed by gases, molecules, micro- and macroscopic solid particles such as ashes, droplets and ice, has a large impact on the amount of energy that effectively hits the ground and is scattered back to the sensor. The quantity of EM radiation that effectively illuminate the sensor can be decomposed in three atmosphere-related components. The total energy received by the sensor is given by the sum of three elements: $E_{\lambda}^{A,tot} = E_{\lambda}^{A,sr} + E_{\lambda}^{A,ds} + E_{\lambda}^{A,us}$.

The first, defined as $E_{\lambda}^{A,sr}$, corresponds to the energy that is transmitted throughout the atmosphere, interacts with the surface, and travels back to the sensor. The fraction of the solar radiation that effectively reach the Earth surface is provided by the atmospheric transmittance, illustrated in Figure 2.2¹. This quantity varies as a function of the wavelength, and is given by the transparency of the atmosphere to specific wavelengths. Gases such as ozone (O₃), carbon dioxide (CO₂) and water vapour (H₂O) may completely absorb or strongly attenuate the incoming energy at some wavelengths, corresponding to the absorption losses depicted in Figure 2.2. As an example, the atmospheric layer mostly composed by O₃ prevent the dangerous ultraviolet radiations to reach the Earth surface. The regions that are poorly affected by these effects are called atmospheric windows. The EM energy corresponding to the VNIR spectrum is only poorly affected by gas absorption and passes through the atmosphere (the optical window). Another important absorption-free region is the so-called radio window, corresponding to almost a full atmospheric transparency for microwave radiations (λ from roughly 1[cm] to 10[m]). The absorption windows, in which the spectrum is almost completely absorbed, are often exploited for remote sensing of clouds and atmosphere.

The second phenomena describing $E_{\lambda}^{A,ds}$ is the portion of the energy that is first down-scattered in the atmosphere then it is reflected and up-scattered by the ground surface to the sensor. This radiation is also known as diffuse light, and it is caused by the Rayleigh scattering. This effect concerns all the wavelengths of the spectrum, but it is significant

¹modified from http://commons.wikimedia.org/wiki/File:Atmospheric_window_EN.svg

2. Introduction to remote sensing

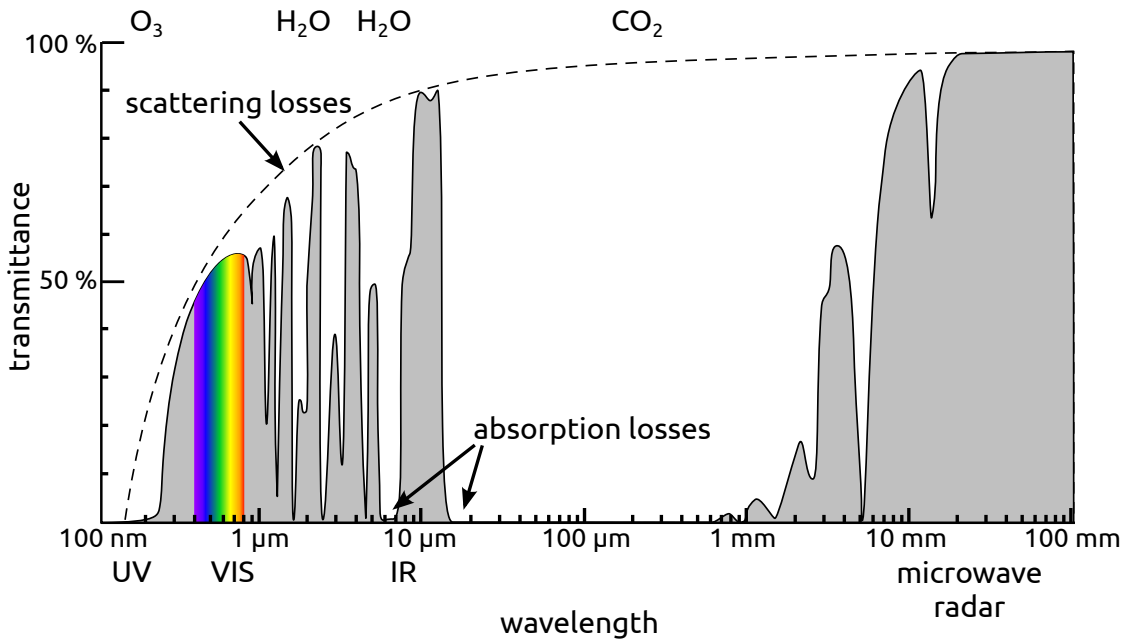


Figure 2.2: Atmospheric Transmittance - Atmospheric transmittance (in %) as a function of the wavelength λ . The dashed line approximates the losses due to scattering, while the principal molecule causing absorption is highlighted at the top of the figure. Atmospheric windows correspond to the gray regions of non-zero transmittance.

only in higher energy radiations, being proportional to the inverse of the fourth power of the wavelength. These losses of energy are due to the continuous scattering of the light by molecules and atoms that are much smaller than the wavelength, in particular by gases such as N_2 , O_2 or by very small dust particles.

In the presence of much larger particles such as smoke, dust, water droplets and pollen, approximately of the same size of the wavelength, another type of energy diffusion known as the Mie scattering occurs. The interactions are much more complex than the ones occurring in a Rayleigh situation, and are very localized spatially. They mostly depend on factors such as wind, anthropization, seasonality, humidity, etc. Both Rayleigh and Mie scattering occur at the same time, and cause the losses due to scattering illustrated in the dashed line of Figure 2.2. Additionally, large particles such as dust in sandstorms, snow, haze and clouds generate a wavelength-independent non-selective obstruction of the light, i.e. causing shadows.

Finally, the third component $E_{\lambda}^{A,us}$ corresponds to the part of the light that is completely up-scattered by the atmosphere, reaching the sensor without interacting with the ground. However, the size of a satellite field of view is often too small to observe spatial variations of this quantity and its contribution is assumed as constant.

Summing up, the energy that comes into the instantaneous field of view of the satellite (IFOV) is strongly influenced by the atmosphere. In particular most of these effects are proportional to wavelength, acquisition geometry and Sun angles, both defining the length

2.1 An overview of the acquisition systems

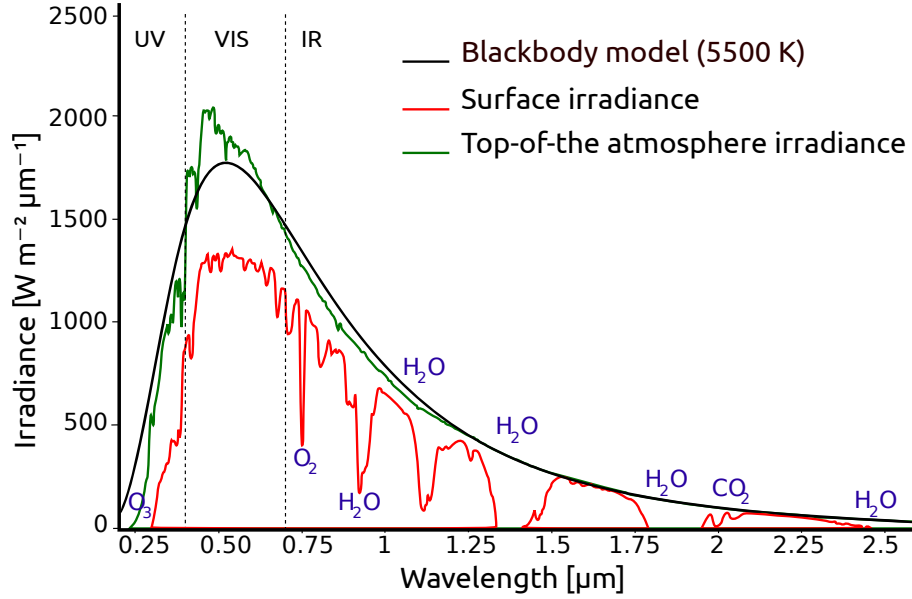


Figure 2.3: Solar irradiance - This plot depicts the incoming energy at the top-of-the-atmosphere (green line) along with the surface irradiance (red line). The black line depicts the model for solar irradiance at the top-of-the-atmosphere, a blackbody of 5500 [K]. The absorption windows of the atmosphere are clearly visible.

that the light has to travel in the atmosphere. Also, the season and the weather conditions strongly affect the atmosphere, thus directly influencing the acquisition of images.

An illustration summarizing the top-of-the-atmosphere irradiance (incoming energy) and the atmospheric effects, by plotting the surface irradiance, is shown in Figure 2.3. For the Sun EM emission, the maximal illumination occurs in the visible region, while for mid-IR and larger wavelength the magnitude of the radiations is much lower.

The interaction between electromagnetic energy and the Earth surface. A portion of the quantity measured by the sensor is influenced by the sensed surface by processes such as reflection, absorption and transmission of the incident radiance. Other interactions such as fluorescence are not reviewed here and one can find additional information in [Campbell and Wynne, 2011; Lillesand et al., 2004].

Depending on the type of material, these three physical processes vary as a function of the wavelength and allow us to distinguish the different objects composing a remotely sensed image. By recurring again to the principle of energy conservation, we can decompose the surface irradiance as $E_{\lambda}^{S,tot} = E_{\lambda}^{S,sr} + E_{\lambda}^{S,sa} + E_{\lambda}^{S,tr}$. In other words, the reflected energy equals the amount of incoming radiation (irradiance) minus the either absorbed or transmitted energy.

The amount of reflected energy $E_{\lambda}^{S,sr}$ directly depends on the surface roughness (at given wavelengths), and varies between an ideally specular reflector (a mirror-like situation) to the perfectly diffuse reflection (Lambertian surface). However, both cases are very

2. Introduction to remote sensing

rare in nature, being the most probable observable situation a combination of the two. A more precise approximation, accounting for the reflective behaviour of the irradiant energy as a function of the surface type, topography and geometry of acquisition, is mathematically described using the bidirectional reflectance distribution function (BRDF).

Since the most of these effects are measurable or at least an approximation can be obtained, and since the total energy going through the sensor $E_\lambda^{A,tot}$ is known, the atmospheric contributions and the BRDF may be estimated and the data compensated (or corrected) for their effects. To obtain these quantities an accurate knowledge of acquisition, atmospheric, topographic parameters and the BRDF itself are needed. From the simple (unitless) numbers composing a raw image (the DN) the at-sensor radiance in $[W m^{-2} sr^{-1} \mu m^{-1}]$ can be computed by knowing sensor coefficients gain and offset. After the compensations of the atmospheric effects, the empirical reflectance may be extrapolated as the proportion of the energy reflected by the surface, as described by the BRDF, by considering the atmospheric attenuation. For details, see [Martonchik et al., 2000; Schaepman-Strub et al., 2006].

Once pixel values are converted from the raw DN to reflectance, the image is expressed using an absolute and in principle invariant reference for each wavelength. However, as mentioned above, to obtain these values large efforts in collecting the adequate prior information and computationally intensive physical models have to be made, to estimate the compensation coefficients. This large amount of information is often neither accessible nor provided, and this largely justify statistical and data driven approaches for the processing of remote sensing images. At least in a relative manner, the pixel values may be used to extrapolate measures of interest or thematic classification maps, as we will see in the following.

2.1.2 Spectral signatures: characterizing materials

Independently on the type of information carried by each pixel the sensor measures sampled parts of the continuous spectrum reflected by the surface. For each pixel and for each wavelength interval (the band), this amount of energy gives us the spectral signature.

The series of sampled values are very characteristic of the ground cover composing the scene. If reflectance values are used, these measures are generalisable also to other observations from other satellites, but if DN or radiance data are considered, these are only discriminative in a relative manner, for the considered dataset. Practically, for a given pixel at given geographical coordinates, the spectral signature is a vector: being \mathbf{x}_i the i th pixel indexed by i indicating indirectly its geographical location, the spectral values are defined (here and throughout the thesis) as $\mathbf{x}_i = [x_{i1} \ x_{i2} \ \dots \ x_{id}]'$, i.e. a d -dimensional column vector defined in \mathbb{R}^d , d being the number of spectral channels. By considering all the pixels together for a given spectral band, they are organized in d two-dimensional arrays (d spectral bands) corresponding to d graylevel images¹.

¹For a list of the most used symbols, see Table on page xx

2.1 An overview of the acquisition systems

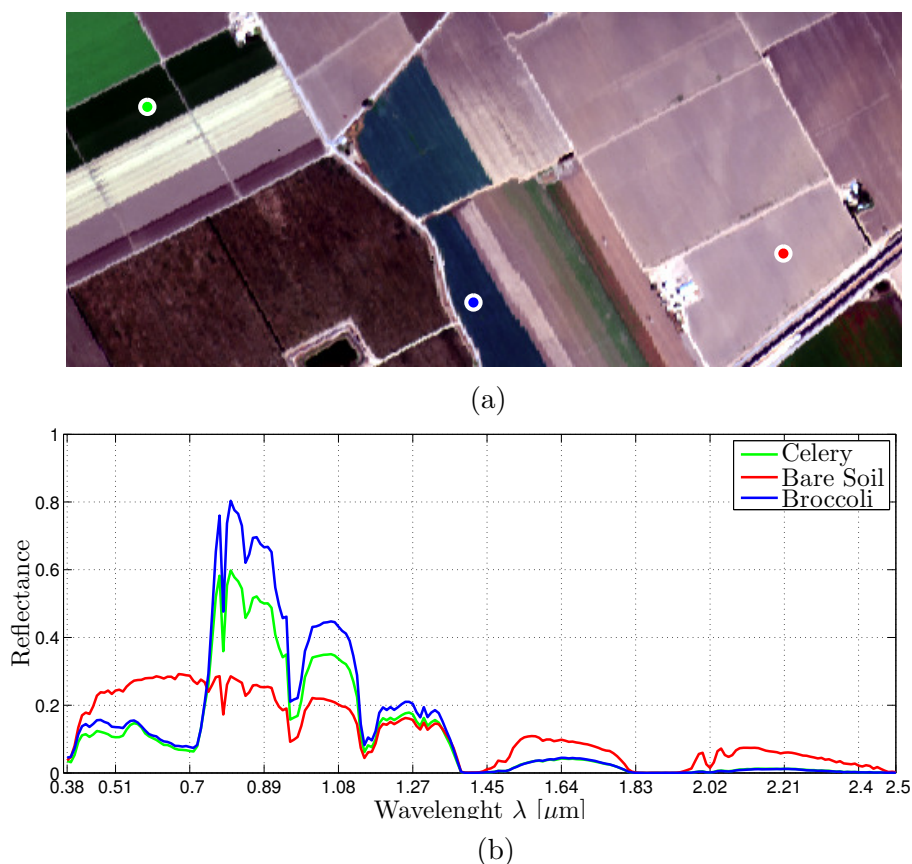


Figure 2.4: Example of spectral signatures - The image shows three different spectral signatures for the hyperspectral dataset Salinas acquired by the AVIRIS sensor over the Salinas Valley, California, USA, illustrated in (a). In (b), the three different lines correspond to the spectral signature of the pixels corresponding to three different ground covers, highlighted by the colored circle. Namely, from top to bottom: Celery, Bare Soil and Broccoli.

Using a widely cited sentence, Parker and Wolff [1965] define the bases for remote sensing by introducing the concept of spectral signature as:

“Everything in nature has its own distribution of reflected, emitted and absorbed radiation. These spectral characteristics can – if ingeniously exploited – be used to distinguish one thing from another or to obtain information about size, shape, and other physical and chemical properties” (citation from [Campbell and Wynne, 2011])

Traditional remote sensing bases most of the processing hypotheses and algorithms on this seminal statement, assuming in a wide sense the uniqueness of the spectral signature with respect to a given material. In Figure 2.4 a visual example is given¹: three different ground covers – two of them related to cultivated crops, one to bare soil – are plotted. The

¹dataset freely available online at:
http://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

2. Introduction to remote sensing

image was acquired by the AVIRIS¹ hyperspectral sensor (see the following for a formal definition). The x -axis illustrates the wavelength, in $[\mu\text{m}]$ for each measured sampled wavelength interval. The y -axis shows the corresponding reflectance in the IFOV (colored circles in Figure 2.4). This sensor samples a very detailed spectrum, by measuring the EM energy in $10[\text{nm}]$ ($0.01 [\mu\text{m}]$) wide bandpass, for a total of 224 different spectral channels from 380 to 2500 $[\text{nm}]$. By analysing the spectral signature of the two differently vegetated cultivations, it is relatively easy to discriminate them one from each other. In more detail, “celery” and “broccoli” classes show a similarly shaped signature, but a different EM response for wavelengths ranging from 400 to 1370 $[\text{nm}]$ and particularly in the NIR-MIR region (from 700 to 1400 $[\text{nm}]$). The spectral signature corresponding to the “bare soil” class shows a completely different behaviour, since related to a very dissimilar ground cover class. By exploiting these properties, we are able to discriminate these three ground cover all over the image, by comparing the spectral signatures of these three references to all the other pixels within the image. This is the principle of pixel-wise thematic classification.

Although for a same spectral class, in relative or in absolute terms, the spectral signature behave very similarly from pixel to pixel, one can note slight differences in the signature of pixels sampled over a same ground cover. This issue can be related to effects caused by topography, shadowing and by mixed pixels. The latter is caused by the fact that, in the IFOV of the satellite, it is very rare (at all the possible resolutions), to observe only a single and pure spectral signature in the sensed pixel. As an example, think to the Broccoli cultivation above: since the ground projected IFOV (GIFOV) of the sensor corresponds to a pixel size of $5[\text{m}]$, it results impossible to observe pixels containing only broccoli in an uniform manner. In practice, it is very likely to observe a mixing of the spectral signatures of broccoli, bare soil, small weeds and the attenuation introduced by their shadows. It results that, for a land use cover corresponding to “broccoli cultivation”, the spectral signature is a (mostly) linear mixing of the pure spectral signatures related to the different sub-classes contained in the pixel. The branch of the remote sensing science that is devoted to decode these signals, i.e. finding the percentage of pure signatures that compose the observed mixed pixel signature, is referred to as signal unmixing, and it is intimately related to source separation problems in signal processing. The interested reader is referred to [Bioucas-Dias et al., 2012] for more details.

2.2 Optical remote sensing systems

Evolutions and innovations in remote sensing technologies, from aerospace engineering to camera sensors, from data transmission protocols to optical elements such as lens and mirrors, allowed an incredible improvement in the optical data quality. In Figure 2.5(a) (one of) the first image of the Earth from the space is visualized alongside one of the most recent in Figure 2.5(b). In the first image clouds are barely distinguishable, while in the second the cars, building façades (thanks to the off-nadir acquisition angle) and road signs

¹Airborne visible and infrared imaging spectrometer

2.2 Optical remote sensing systems

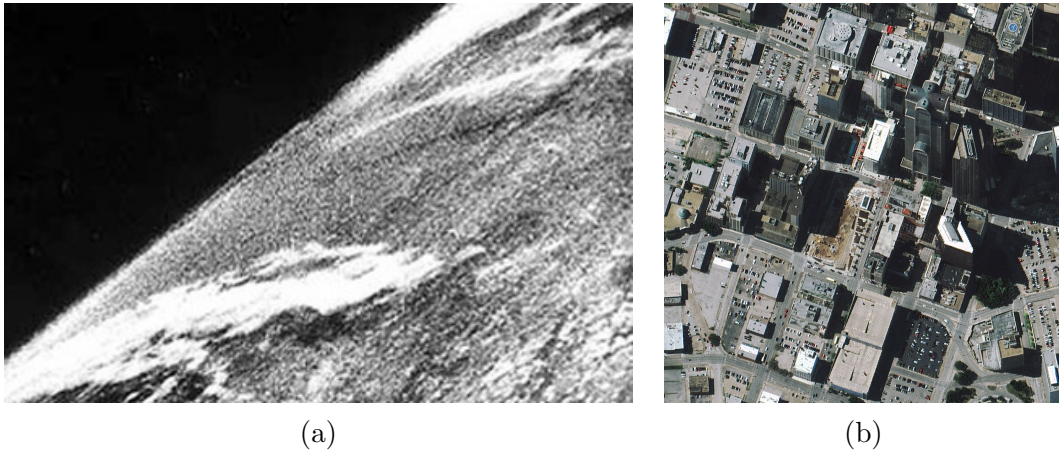


Figure 2.5: RS image evolution - In (a) the first image acquired from space by a 35mm camera, from a V-2 rocket launched the October 24, 1946, from a New Mexico (USA) missile range. Image from <http://www.airspacemag.com/space-exploration/FEATURE-FirstPhoto.html>. In (b) an image acquired by the WorldView-II sensor over Dallas, Texas (USA) on October 19, 2009. ©DigitalGlobe, from <http://www.geovar.com/wv-2.htm>.

are clearly visible. This sensor comes with a maximal spatial resolution of 0.46[m] for the high resolution panchromatic band. Furthermore, it comes with a total of 8 spectral channels in the VNIR region, with 1.8[m] of spatial resolution.

As deducible from these paragraphs, it is clear that the optical images coming from different sensors could differ largely one from the another. So far, we mentioned hyperspectral, very high resolution, multi-spectral and other characterizing terms without properly defining them. In the next Sections, these main characteristics are detailed, by presenting the four types of resolutions that characterize these kinds of data.

2.2.1 A characterization of optical sensors

The sensors populating our sky acquire the images in different ways, depending on the manner of sensing the Earth surface. There are mainly two different scanning schemes: the whiskbroom scanners acquire pixel values by mirroring in the cross-track direction (perpendicular to the sensor movement) the reflected light into several detectors. The pushbroom scanners analyze separate lines of pixels covering all the ground-projected field of view (GFOV). It corresponds to projecting the light into a linear array of detectors, covering the area of the field of view (FOV), coinciding with the angle covered by the sensor in the cross-track direction, i.e. the image width. The natural displacement of the sensor allows to scan spatially contiguous lines of pixels, until the predefined area (or strip) is sensed. This last scheme is usually adopted by airborne hyperspectral sensors.

On the basis of the number of spectral channels, different families of sensors can be distinguished. This first categorization is based on the absolute number of spectral bands acquired. In this sense, we usually make distinction between multi-, super-, hyper- and ultra-spectral sensors, as summarized in Table 2.1.

2. Introduction to remote sensing

Category	Number of bands
RGB (standard image)	3
Multi-spectral	4-15
Super-spectral	16-50
Hyper-spectral	51-500
Ultra-spectral	501- >1000

Table 2.1: Categorization of sensors based on the number of spectral channels.

One should be cautious in thinking that a large number of spectral channels correspond generally to better data. Usually, the bandpass width trades-off with the spatial resolution, to keep an optimal signal-to-noise ratio. Thus, many spectral bands usually corresponds to lower spatial resolution. To obtain (very) high spatial resolution hyperspectral data, one may recur to the use of airborne sensors with adjustable optics, such as the ROSIS¹ [Kunkel et al., 1988], with 115 bands in 5[nm] intervals on the VNIR, with a maximal spatial resolution of 1.2[m] (IFOV of 0.56[mrad]).

The different sensors are usually categorised on the basis of the number of spectral channels. Standard RGB imagery is usually acquired by airborne cameras and it is used to retrieve very high spatial resolution images (VHR) in the range of 0.1-0.5[m]. At the opposite situation we have spectrometers² such as the MetOP-IASI sensor, sampling 8461 spectral channels in the infrared (IR, in 3.62 - 15.5[μm]). It is used for meteorological applications and to retrieve atmospheric parameters (e.g. temperature, ozone, humidity) [Camps-Valls et al., 2012].

The spatial resolution. It is usually defined as the size of the smallest spatial element that can be distinguished by observing the image. However, as an objective simplification, it is often assumed that the spatial resolution corresponds to the GIFOV (also known as ground sample distance, GSD), the actual size of the ground projected pixels. It is common to design sensors such that the GIFOV corresponds to the distance between two pixel centres (defined as the ground sample interval, GSI), so that the image is composed by an array of adjacent pixels with a common boundary [Schowengerdt, 2007]. In Table 2.2 a distinction of the sensors based on their spatial resolution is given.

The spectral resolution. The spectral resolution is defined by the ability of the sensor to sample the EM radiation with the smallest possible bandpass. The smaller this interval, the more precise spectral details are. To define this type of resolution using an example, think to the spectrum sensed by the AVIRIS used in Fig. 2.4. In this example, bands are 0.01[μm] wide, allowing a very fine sensing. In contrast, a spectral band of a standard multispectral sensor such as the Landsat's ETM+, say the NIR band, is constituted by a

¹Reflective optics system imaging spectrometer

²Imaging spectrometers are often referred to as hyperspectral sensors

2.2 Optical remote sensing systems

Category	Image GIFOV [m]
Very high resolution	< 2.5
High resolution	2.5-10
Medium resolution	10-50
Low resolution	50-100
Very low resolution	> 100

Table 2.2: Categorization of sensors based on the spatial resolution.

passband of $0.15[\mu\text{m}]$, between 0.75 and $0.9[\mu\text{m}]$. The AVIRIS sensor covers this interval by 15 distinct channels: by considering a spectral property visible only in one of these intervals, solely AVIRIS can resolve it. On the contrary, the ETM+ would simply have averaged out such fine wavelength details. Summing up, hyperspectral images provide a high-resolution sensing of spectrum, defined by the fine sampling and the large number of channels, while multi-spectral sensor provide few bands with wider bandpass. For such sensors the signal-to-noise ratio is improved and smaller spatial details can be resolved.

The radiometric resolution. It is related to the ability of the detector to quantize, for each spectral band, the EM energy into distinct graylevel values. The more these intervals are, the better this resolution is, since the spatial variations of the quantity of EM received by the sensor are more detailed. The quantity of these intervals is given by the number of bits used to code the information, as for the most of digital data. The number of bits, say B , gives the number of distinct values of the sampled, as 2^B , in an interval $\text{DN}_{\text{range}} = [0, 2^B - 1]$. For a same continuous signal, for larger B the quantization is more detailed and higher is the radiometric resolution. Usually, remote sensing images are coded using 8, 11, 12 or 16 bits per channel.

The temporal resolution. It is defined by the shortest time that the acquisition system needs in order to acquire an image of a same geographical area sensed previously. It has been greatly improved by the introduction of adjustable systems allowing the acquisition of images at off-nadir angles. However, for large angles, geometrical detail degrades and parallax effects should be taken into account when processing the data. Indicatively, large distortions start to appear when acquiring off-nadir images with angles wider than $\pm 25^\circ$. For instance, fixed angle acquisition system such as the TM or the ETM+ can acquire an image of a given geographical area each 16 days ([d]). The commercial sensors QuickBird and WorldView-II, depending on the latitude and on the off-nadir angle, can provide data in 1 to 3.5[d].

2.2.2 The optical data as grayscale images

A remote sensing dataset can be seen as a collection of single grayscale images. Consequently, we should briefly recall some properties common to all types of image data. One

2. Introduction to remote sensing

of the most important effects introduced by the sensor is the imaging noise. That is, some random variations to the true pixel signal are introduced, resulting in histograms that instead of showing a single peak have some spread. For optical data, pixel values will most likely follow a Gaussian distribution. In this case, noise is composed by independent and identically distributed (iid) realizations from a Gaussian probability density function (PDF) with mean 0 and standard deviation σ . It is a special case of the white noise: the intensity of the noise does not change with the spatial frequency at which it is observed (constant power spectrum). Moreover, it follows an additive model, i.e. it is not dependent on the pixel signal. The process generating the pixel values can be described, for a grayscale pixel x_i , as $x_i = g_i + \epsilon_i$. The true signal g_i is uncorrelated from the noise ϵ_i , drawn from a zero mean Gaussian distribution. A measure of the noise intensity is readily obtained from the model above: the signal-to-noise ratio (SNR). It is defined as the ratio between the variation of the signal g_i and the variation of the noise ϵ_i [Schowengerdt, 2007]:

$$\text{SNR} = \frac{n^{-1} \sum_i (g_i - n^{-1} \sum_i g_i)}{n^{-1} \sum_i (\epsilon_i - n^{-1} \sum_i \epsilon_i)}. \quad (2.1)$$

However, its practical estimation requires the knowledge of the true signal and noise distributions. A priori, this is often unknown, but it can be estimated from the image directly. By introducing the concept of spatial autocorrelation, we know that neighbourhooding pixels will probably have a closer value to each other than to two pixels which are far away. Additionally, due to the noise properties, we may observe that the empirical average of pixels covering a homogeneous area is close to the true underlying signal. In these terms, for a given homogeneous area within a neighbourhood W , the signal noise can be estimated as the deviation of the samples from the mean of pixels in W . Note that the noise variance is assumed to be constant on the whole image, and the average is assumed to be an unbiased estimate of the true signal corresponding to the pixels to which a sufficiently large W corresponds:

$$\mathbb{E}[x_i] = \mathbb{E}[g_i + \epsilon_i] = \mathbb{E}[g_i] + \mathbb{E}[\epsilon_i] \overset{0}{=} \mathbb{E}[g_i]. \quad (2.2)$$

The SNR can then be approximated as:

$$\text{SNR} = \frac{n^{-1} \sum_i (g_i - n^{-1} \sum_i g_i)}{n^{-1} \sum_i (g_i - n^{-1} \sum_{j \in W} g_j)}. \quad (2.3)$$

The SNR is commonly evaluated in [dB] after a nonlinear transform [Sonka et al., 1999]:

$$\text{SNR}_{\text{dB}} = 10 \log_{10} \text{SNR}. \quad (2.4)$$

It is worth mentioning that there exist different estimations of the SNR [Atkinson et al., 2005]. Since noise is assumed to be constant in the images and caused by the sensor, degradation of channel quality is a consequence of the low signal present in it, for two main reasons. The first is that, atmospheric scattering may strongly reduce the amount of signal for channels corresponding to shorter wavelengths, and secondly because the light

2.3 Change detection in remote sensing data

measured by a sensor could be reduced to zero depending on the molecular absorption windows of the atmosphere. In remote sensing image processing, a direct application of the SNR measure is for detecting uninformative spectral channels, such as the ones affected by water vapour absorption. In both cases, the imaging noise is not directly caused by the wavelength, but the signal is weakened consequently decreasing the SNR. One can take advantage of the above observations to improve SNR of the image. Some advanced methods, such as the SNR-based rotations, oriented PCA, or minimum noise fraction rotations and their nonlinear extensions [Canty, 2007; Gómez-Chova et al., 2011; Green et al., 1998; Nielsen, 2011] take SNR estimators as objective functions to estimate a projection of the original data minimizing the noise.

2.3 Change detection in remote sensing data

Now that the data in which changes are to be detected have been introduced, we detail the general problem of change detection.

A human interpreter may detect changes in the shape and state of objects with a very high accuracy, by simply comparing of two images roughly covering the same geographical area. This holds for images with arbitrary size, different spectral channels (the shape of objects is invariant to colors), spatial resolution, independently from seasons and Sun elevation levels. This is due to the ability of our brain to interpret the image and to extract the relevant information from both images, such as the relative localization, shape and color information (if required), and to discard the uninteresting effects, such as shadows, differences in illumination, high level of details. Finally, our brain makes a decision on whether a change is relevant and if it should be qualified as such [Rensink, 2002].

Still, our brain is able to interpret only a portion of the information carried in a remote sensing image, that is the spatial aspect of the objects and their relative position and their color, if relevant. However, this last characteristic is defined by the visualization system, that usually rely on a RGB composition of the available spectral channels. Consequently, many bands may not be considered in the process of change detection by the human brain, simply because not visualized. In this sense, other changes related to a more intimate state of the object (e.g. thermal variations) are only visible by manually analysing and comparing the spectral bands related to the wavelengths at which the phenomena is observable, or by properly considering all the bands at once. Additionally, images usually cover many square kilometres, and it is almost impossible to manually screen bi-temporal couples of images with the aim of change detection.

In change detection (CD) terms, there are many phenomena that can generate changes between acquisition. Raw differences are simply given by deviations in the DN numbers for the pixels at the same spatial coordinates. However, most of them are not related to actual changes or transitions in the ground cover. Therefore, we should first define which transitions are of interest. Change detection in remote sensing data has been considered for many applications, principally for urban monitoring and mapping [Nemmour and Chibani, 2006; Schneider, 2012], crop and environmental monitoring [Kennedy et al., 2009;

2. Introduction to remote sensing

Koppe et al., 2013; Zhang and Jia, 2013], natural hazard detection and quantification [Gianinetto and Villa, 2007; Metternicht et al., 2005; Tralli et al., 2005] and post-catastrophe assessment [Gillespie et al., 2007; Guchi et al., 2003; Suppasri et al., 2012]. However, even if these application fields are very different, all of them rely on the detection of specific changes and transitions, that are the ones the user is interested in.

Without entering now in methodological details, we can readily classify changes depending on the phenomena generating them. The changes in which we are interested in are related to structural differences, generated by processes such as addition and removal of materials or object motion [Rensink, 2002]. The latter deserves a precise definition, since motion does not corresponds in general to actual changes in land use and ground cover. Specifically, two images of a car moving on a road generates a structural change, while ash plumes or river streams, even if obviously moving, are not considered as a changes. Dynamic processes generating transitions in ground cover may not be considered as changes depending on the acquisition time instants. Using again the river stream as example, two remote sensing images acquired over it within a few days will not show any change, if no abrupt process such as a flooding occurred, while the same area imaged ten years apart will likely show structural differences due to river bed erosion.

The intervention of external effects such as different illuminations, Sun elevation, parallax effects, registration errors and noise in general, may generate detectable changes from the radiometric point of view, since spectrally distinguishable, but without belonging to a semantic or to a specific thematic information class transition. The most evident example resides in the shadow: a very easily detectable change from the spectral point of view, but semantically inconsistent (shadow is not a ground cover). Regarding changes due to uniform transformations of the image values, due for instance to homogeneous atmospheric effects, may generate strong differences in object and notably to colors. Without a proper preprocessing of the data and the application of advanced methods this may result in false detections. As introduced in Section 2.1.1 the obvious solution is to work with reflectance values, but this kind of preprocessing is very costly and not always applicable. In this Thesis, we consider as changes of interest all the transitions generated by a structural modification of the objects, in ground cover state and in land use, all of them detectable by image analysis and image comparison. Illumination differences and shadows, if not stated differently, are not considered as changes.

2.3.1 Standard approaches to change detection

To detect changes occurred in a pair of images (bi-temporal change detection), different approaches may be exploited. Ideally, two main paradigms for CD exist: feature-based and pixel-based. The former, extract a series of features independently from the original images, such as structure descriptors, edges and object identifiers. Then, these features are compared one to each other and changes are detected if modifications in their shape and values are observed. The latter family includes the approaches developed in this Thesis and builds on the assumption that changes are directly detectable by comparing each pixel

2.3 Change detection in remote sensing data

at the same spatial coordinates from both images. To make a decision on whether the land cover to which the pixel belongs has changed or not a similarity measure, a change metric or a decision function, is computed. These values are finally thresholded in order to separate changed from unchanged pixels. Note that this definition places the change detection process very close to standard machine learning tasks such as classification, clustering, or density estimation task [Camps-Valls et al., 2011].

To provide an introduction of many crucial concepts for CD, we analyze a simple yet effective family of methods, those relying on the difference image [Bovolo and Bruzzone, 2007; Bruzzone and Fernández-Prieto, 2000; Coppin et al., 2004; Malila, 1980; Mas, 1999; Radke et al., 2005; Singh, 1989]. The image comparison is based on a differencing operator, i.e., for general d -dimensional images X^{t_1} and X^{t_2} , acquired at times instants t_1 and t_2 , the difference image \mathbf{D} is computed as:

$$\mathbf{D} = \mathbf{X}_2 - \mathbf{X}_1, \quad (2.5)$$

where \mathbf{X}_1 and \mathbf{X}_2 are the data matrices of the image, i.e. pixels are rearranged from a stack of d two-dimensional arrays into a $n \times d$ matrix where each one of the n lines is a pixel and its d columns are the values of the d spectral channels.

Ideally, a multi-variate difference close to $\mathbf{0}$ indicates that at the spatial coordinate to which the pixel belongs a change has not occurred, while for values significantly different than $\mathbf{0}$, say larger than an optimal discriminating threshold, will probably indicate a ground cover change. To compress the change information from a d -dimensional space into a 1-dimensional measure easily interpretable, spectral change vectors contained in \mathbf{D} are decomposed into a magnitude (the difference pixel vector norm) as $\Delta_i = \|\mathbf{D}_i\|_2$ and orientation $\Xi_i = \arccos(\frac{\sum_j \mathbf{D}_{ij}}{\Delta_i})$. Transitions can be discriminated by inferring a binary decision, as:

$$\hat{y} = \begin{cases} 1 & \text{if } \Delta_i \geq \theta; \\ 0 & \text{otherwise.} \end{cases} \quad (2.6)$$

where θ is a user defined threshold discriminating high magnitudes related to pixel change. In parallel, the same can be performed on the angles Ξ to discriminate groups of pixels probably belonging to the same transitions. In this case multiple thresholds quantize the angle values into distinct spectral classes of change. By considering the information class to which they belong, pseudo-classes of changes may be detected [Bovolo et al., 2012], as well as artificial differences [Marchesi et al., 2010]. This approach is also known as the change vector analysis or CVA. For a description of main approaches of change detection and for a state-of-the-art literature review, see Chapter 5.

For different methods for CD, the task is often very similar, and some general statements can be made. Key points of paramount importance that the algorithm and CD systems must face are the robustness with respect to noise and to different illumination conditions, to enhance the ability of detecting true structural and semantically coherent changes. Noise is the main reason why $\mathbf{D} \approx \mathbf{0}$ and not $\mathbf{D} = \mathbf{0}$. In this case, one must pay attention to the fact that noise must not be considered as change, of course, even if large

2. Introduction to remote sensing

deviations from the no change score are observed. Concerning differences in illumination, the method should possess some invariance properties to global transformations of the image, such as those due to Sun elevation level and seasonality. In the next Sections, spatial and spectral requirements for change detection are reviewed.

2.3.2 Geometrical requirements for change detection

Since many change detection approaches are based on pixelwise processing of the multi-temporal images or a their transformation, either by stacking images, computing the difference or comparing independently obtained classification maps (see Chapter 5.1), the precision of the spatial correspondence of pixels is of crucial importance. A perfect superposition ensures that the comparison of each pixel is related to an absolute geographical location, and artificial changes due to misregistrations are not introduced in the process. This geometrical preprocessing step is known as co-registration if the images are referenced in a relative manner and (absolute) registration or georeferencing when geographical coordinates in a given geographical projection system are assigned to each pixel. Detailed studies on the effects of the misregistration errors on change detection can be found in [Bovolo et al., 2009; Dai and Khorram, 1998]. To this end, different manual or automatic techniques exist [Campbell and Wynne, 2011; Lillesand et al., 2004; Schowengerdt, 2007]. The choice between manual or automatic (co-)registration is usually made on the basis of the amount of deformation that has to be corrected. If images were acquired with significantly different acquisition angles, a preprocessing step exploiting digital surface elevation models known as orthorectification may be needed before co-registering the images, so that different perspectives does not affect image geometry and change detection [Schowengerdt, 2007].

As a general observation, the higher is the resolution of the image, the challenging is the spatial matching processing. Moreover, VHR systems have the ability of tilt the sensor to large angles, and the parallax effects are hardly compensable in particular for a highly variable surface, such as in urban areas with tall skyscrapers or in mountainous areas. On the contrary, mid to low resolution fixed-angle sensors such as the Landsat TM are usually easier to match, and only require global linear shifts after orthorectification to compensate the misregistrations.

2.3.3 Spectral and radiometric requirements for change detection

Similarly to geometrical properties, radiometric values of the spectral channels should be matched so that the relative comparison of images can be carried out meaningfully. Since the conditions might vary from one acquisition to another, it is important to compensate shifts that make the values of same classes differ from one acquisition to the other. As introduced, these shift are caused by atmospheric conditions and differences in illumination. These adjustments are of particular importance for unsupervised methods, since basing the CD process on a direct detection of the deviation between pixel values (see Chapter 5 and Chapter 7). For supervised classification-based change detection methods this step

2.3 Change detection in remote sensing data

is less crucial, since the feature vector is modelled directly without any assumption on the value it should take (see Chapter 5 and Chapter 6). However, it has been demonstrated that for multi-temporal analyses, a preprocessing aiming at matching the image values is always beneficial, with a worst case scenario of no difference between analysing preprocessed or raw images [Song et al., 2001]. Three main approaches are usually applied: atmospheric compensation to obtain reflectance, data normalization and histogram matching. The first transforms values in an absolute, global, sensor independent reference, unique for a given wavelength and ground cover class. In this case, one does not need any further transformation and all the spectral band values are in $[0, 1]$. Regarding the second and the third approach, they aim at matching the radiometry of the image in a relative manner, so that for a same spectral channel and a same ground cover class the spectral values are as closest as possible and thus comparable. Data normalization aim at applying some data-dependent or fixed transformation function to pixels such that the feature vector values are adjusted to a common scale, while histogram matching infers the distribution of each channel to the corresponding one of the other image. If changed areas affect large regions of the images, or if changes are generated by new and previously unseen classes, parameters for the scaling of the data may be extrapolated manually from unchanged regions, so that differences may be detected more effectively. In detail, principal normalizations used in the field of data analysis are:

Centering The pixel values are translated or centered such that their mean is $\boldsymbol{\mu} = \mathbf{0}$, i.e. $\mathbf{x}_i - \boldsymbol{\mu}$, $\forall i$. This provides a homogenization of the variable means. In change detection this is particularly useful since for a large scene, if changed regions are only a fraction of the total and belong to the same spectral classes in both images, the average of the bands for each image should correspond. Centring is a useful transformation when external effects are considered homogeneous and the average of the channel values are stationary over the time (again, if changed pixels are few in a large image, and not due to novel classes). This might correspond to a very basic relative atmospheric compensation and radiometric correction.

Standard scores The pixel values are transformed such that each channel is rescaled to mean $\boldsymbol{\mu} = \mathbf{0}$ and standard deviation $\boldsymbol{\Sigma} = \mathbf{I}$. The data are centered (as described above) and further normalized by the standard deviation of each channel so that the data range is also matched. It is well suited to deal with data following a normal distribution, since after the transformation the samples will follow $\mathcal{N}(\mathbf{0}, \mathbf{1})$, although it can be applied without assuming any prior probability density. Standardization is often required depending on the adopted method, in particular when variables should possess the same scale.

Unit norm This type of normalization maps independently each feature vector (pixel) on the unit hypersphere, so that $\|\mathbf{x}_i\|_2 = 1$. This is useful when the data sample magnitude is not important and only the angle between vectors should be discriminative of their properties (e.g. spectral angle mapper classification). The relative importance of the variables is preserved. The normalization might provide some helpful illumination invariant properties.

2. Introduction to remote sensing

Data rescaling This widely used data stretching does not affect relative importance or statistical behaviour of data samples and variables. Usually, a stretching function is applied so that the set of images are bounded by, for instance, $\tilde{\mathbf{x}}_i \in [-1, 1]$ or $\tilde{\mathbf{x}}_i \in [0, 1]$, where $\tilde{\mathbf{x}}_i, \forall i$ denotes the stretched data sample.

In some situations the normalization of data is not only useful to enhance the separability of samples (e.g. improve classification accuracy), but it can reduce the computational time of some algorithms [Graf and Borer, 2001; Villa et al., 2008].

Finally, one of the most widely used radiometric preprocessing of remote sensing data is to transform the image histograms by either histogram equalization or histogram matching. Again, these approaches are valid under the assumption that small changed areas, not due to novel classes, affect the image. The former applies a function on the image histogram aiming at linearizing it, such that the distribution of values is approximated by a uniform distribution. The cumulative density function (CDF) is simply applied to the image whose distribution has to be equalized. The specified uniform histogram values \tilde{x} with $p_1(x)$ being the density function of the image, are given by:

$$\tilde{x}_s = \text{CDF}_1(x_s) = \int_0^s p_1(w)dw. \quad (2.7)$$

The discrete approximation of the analytical specification of a uniform histogram is:

$$\tilde{x}_s = \text{CDF}_1(x_s) = \sum_{i=0}^s p_1(x_s) = \sum_{i=0}^s \frac{n_s}{n}, \quad (2.8)$$

with $s = 0, 1, \dots, 2^B - 1$ and B the number of bits of the image and n_s the number of pixels having a value of s . A way to match the values for unchanged areas of the image is to apply independently the histogram equalization.

However, a better approach preserving maximally the original distributions, avoiding artefacts and color distortion, is the histogram matching procedure. This technique is used to specify the PDF of an image to a second one, by inferring the inverse CDF of the image from which the histogram is to be transferred to the equalized histogram of the image to be transformed:

$$\tilde{x} = \text{CDF}_2^{-1}(\text{CDF}_1(x)), \quad (2.9)$$

where CDF_1 and CDF_2 are obtained as in Equation (2.7) or Equation (2.8). In practice, the inverse $\text{CDF}_2^{-1}(\cdot)$ is not needed since the support, being bounded in $[0, 2^B - 1]$, allows the explicit computation and storage of all the possible values of $\text{CDF}_2(\cdot)$. Alternatively, when the image radiometric resolution is very high, one may recur to the use of very small quantiles (binning) to estimate the inverse of the CDF.

2.4 Some considerations

As discussed in this Chapter, the processing of remote sensing data and in particular the detection of changes involves a series of important observations. From the semantics

2.4 Some considerations

behind observed changes and their detection, going through data normalization and the computation of a robust change indicator used to map changes, the practitioner may end up in situations in which standard change detection methods are not effective, in particular for new generation data and for high level requirements to obtain accurate and detailed outputs for a very specific application.

Nowadays, many researchers are involved in the application of machine learning approaches for processing remote sensing data, mostly for thematic classification problems. Fundamental limitations in such systems were underlined by trying to obtain high level products as required by modern geoscientists and planners by exploiting standard image classification tools, incapable of overcoming the complexity of most tasks. The pattern recognition and statistical machine learning fields are offering solutions to these issues, and proposing mathematical tools able to solve many of such problems. For change detection purposes, the same observation could be made: fundamental limitations in the efficiency and accuracy of standard methods are appearing clearly, since users are requiring products of increasing quality standards.

With the increase of the computational power, many researchers adopt the machine learning (ML) paradigms to process remote sensing data and images. In particular, early developed methods based on standard statistical and signal processing models are rapidly being replaced by more powerful and versatile algorithms from the advanced ML theory. The application of ML tools to remote sensing data is still an open research field and methods aiming at solving specific processing tasks are still needed to be studied, developed and verified for operational use [Richards, 2005]. Nowadays, a standard laptop provides enough computational power to solve in a fast and reliable manner processing problems, by exploiting such power to solve complex optimizations and learning problems. Furthermore, a fast moving research area in remote sensing data analysis deals with parallel implementation of processing algorithms, and with the integration of the high performance computing through the distribution of computations to graphical processing units [Plaza et al., 2010]. For these reasons, machine learning and pattern recognition based processing methods are promising tools to solve also the new challenges in multi-temporal remote sensing image processing tasks.

2. Introduction to remote sensing

Part II

Machine learning and kernel-based algorithms

Chapter 3

Machine learning methods for data analysis

This Chapter introduces the main concepts of statistical machine learning and of regularization theory used throughout the Thesis. Section 3.1 summarizes the fundamental concepts of machine learning, and Section 3.2 couples them with regularization theory. In Section 3.3 the model selection issues are discussed. Finally, Section 3.4 briefly reviews the main families of machine learning data analysis methods.

3.1 Learning from data

Given a learning task, we look for the functional f of $\mathbf{x} \in X \subset \mathcal{X}$ that best fits inputs to outputs. The task may be to classify a pixel, to estimate a quantity of interest or to find a the subspace on which the data live. In other words [Bishop, 2006]:

Definition 1 (Learning function) *A learning task is solved by a function f that takes input vectors \mathbf{x} and maps data samples from an input space \mathcal{X} with realizations \mathbf{X} to an output space \mathcal{Y} with realizations Y , with minimal error. It is instantiated by corresponding input-outputs pairs (\mathbf{x}, y) :*

$$\begin{aligned} f : \mathcal{X} &\rightarrow \mathcal{Y} \\ \mathbf{x} &\mapsto \hat{y} = f(\mathbf{x}). \end{aligned} \tag{3.1}$$

Here, $\mathcal{X} \in \mathbb{R}^d$ is the d -dimensional input space, while \mathcal{Y} may vary from task to task, for instance $\mathcal{Y} \in \mathbb{N}$ for thematic classification (a given information class is recoded through discrete labels), $\mathcal{Y} \in \mathbb{R}$ for regression and $\mathcal{Y} \in \mathbb{R}^q$, with $q \ll d$ for feature extraction / selection or for multi-output regression predicting q variables.

To find the optimal form of f , a function estimating the disagreement between estimated and true outputs, $f(\mathbf{x})$ and y respectively, has to be optimized. This step is known as the training step, and it allows to optimize the internal parameters Θ of the model

3. Machine learning

f so that the best possible approximation (or fit) of the data is obtained. This is performed through the optimization of a loss function $\mathcal{L}(f(\Theta, \mathbf{x}), y)$ on the training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_s}$ ¹. Keep in mind that many models also have some free hyperparameters to be manually selected. For now, consider that those to be set by the user are fixed.

Definition 2 (Loss function, [Schölkopf and Smola, 2002]) *A loss function denoted as $\mathcal{L}(f(\Theta, \mathbf{x}), y)$ is an integrable, nonnegative function quantifying the error or fit of the model $f(\Theta, \mathbf{x})$ over known example pairs (\mathbf{x}, y) . It evaluates the disagreement in the form of a function $\mathcal{L} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^{(0,+)}$ (or $[0, \infty)$). The value of $\mathcal{L}(\cdot)$ ranges from 0 (no error, perfect fit) to any larger value corresponding to larger errors (bad fit).*

A classical example of loss function, widely used in classification models with true and predicted labels $y \in [-1, 1]$, is the 0-1 loss $\mathcal{L}(f(\Theta, \mathbf{x}), y) = \frac{1}{2}|f(\Theta, \mathbf{x}) - y|$. It returns a value of 0 if the true label y is correctly estimated by $f(\Theta, \mathbf{x})$, 1 otherwise. Another well-known function is the quadratic loss: $\mathcal{L}(f(\Theta, \mathbf{x}), y) = \frac{1}{2}(f(\Theta, \mathbf{x}) - y)^2$ used in least-squares regression.

Starting from an appropriate loss function quantifying the errors occurring between $f(\Theta, \mathbf{x}_i)$ and y_i , it should be further extended to approximate the error on all the samples coming from the observed distribution $p_{\text{emp}}(\mathbf{x}, y)$. It is assumed that the samples used to learn the model are from an underlying process generating the data $P(\mathbf{x}, y)$. For a fixed amount of observed examples n_s , the empirical risk (training error) of a model $f(\Theta, \mathbf{x})$, is:

$$R_{\text{emp}}(f) = \int_{\mathcal{X} \times \mathcal{Y}} \mathcal{L}(f(\Theta, \mathbf{x}), y) p_{\text{emp}}(\mathbf{x}, y) d\mathbf{x}dy = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathcal{L}(f(\Theta, \mathbf{x}_i), y_i). \quad (3.2)$$

However, a direct minimization of the empirical risk (the ERM principle) will lead to solutions not representative of the true underlying distribution, since we usually dispose only of a small training set from $P(\mathbf{x}, y)$ [Schölkopf and Smola, 2002; Vapnik, 1998]. Consequently, the empirical risk $R_{\text{emp}}(f)$ on a finite set alone is not a good approximation of the true risk of the model with respect to $P(\mathbf{x}, y)$. To verify if this situation occurred, we may want to compute the generalization error, also known as risk. It corresponds to the empirical risk $R_{\text{emp}}(f)$ evaluated over all the possible outcomes of the underlying distribution function $P(\mathbf{x}, y) = \int_{\mathcal{X} \times \mathcal{Y}} p(\mathbf{x}, y) d\mathbf{x}dy$:

$$R(f) = \int_{\mathcal{X} \times \mathcal{Y}} \mathcal{L}(f(\Theta, \mathbf{x}), y) p(\mathbf{x}, y) d\mathbf{x}dy = \mathbb{E}(R_{\text{test}}). \quad (3.3)$$

Recall that examples modelled by f are also generated from the governing distribution $P(\mathbf{x}, y)$. The rightmost part in Equation (3.3) provides another look at this integral of the loss over the density $p(\mathbf{x}, y)$. It can be interpreted as the expectation of the test error estimated using (infinite) iid realizations of the governing process $P(\mathbf{x}, y)$. However, as we will see, one usually dispose only of finite sets, making impossible a direct estimation of Equation 3.3 since $p(\mathbf{x}, y)$ is not accessible.

¹For a formal definition of the sets used for learning, see Appendix A.

3.1 Learning from data

To estimate the generalization ability of the model, we usually employ a finite test set of n_t elements. This set allows to estimate the model performance on previously unseen data by mimicking another independent draw from $P(\mathbf{x}, y)$, thus providing an approximation of $\mathbb{E}(R_{\text{test}})$. Note that this set has never been used to train the model. In this case, we replace the expectation with the sample average, to obtain the test, or generalization, error.

Definition 3 (generalization error) *The generalization error of a trained model f , denoted as $R_{\text{test}}(f)$, is defined as the rate of errors over the total of predictions, as:*

$$R_{\text{test}}(f) = \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(f(\Theta, \mathbf{x}), y) \quad (3.4)$$

As one remarks, there is no apparent difference with the empirical error of Equation (3.2). However, note that the parameters defining the function $f(\Theta, \mathbf{x})$ in Equation 3.4 are fixed, indicating a trained model.

We may now fix two important concepts in machine learning: under- and over-fitting. Figure 3.1 illustrates the behaviour of the train and test errors for models of increasing complexity. If a model is not complex enough (or if its class of hypothesis is not rich enough), it underfits the examples and fails to model the true underlying data structure, resulting in high generalization error. This situation is related to a high bias of the model, meaning that we would observe a large generalization error even if we train the model on a very large set. In this case the variance of the model is low, since for other realizations of the training data the estimations would not differ largely. The opposite situation is met at the rightmost part of Figure 3.1. A model complex enough fits the training set always perfectly, but fails in capturing the true structure of the data, which results in a poor generalization ability. For different realizations of the training data such model undergoes to large variations, while showing a low bias since it easily adapts to very complex distributions.

Summing up, these extreme situations are very atypical in nature, and models should always avoid them. In most cases, the optimal solution providing the lowest generalization error is somewhere in between, meaning that a good model has to be sufficiently complex to capture the data structure but simple enough to guarantee generalization ability on new samples. Thus, a trade-off between bias and variance is needed. This is often referred to as bias-variance dilemma [Hastie et al., 2009]. These intuitions were already studied in the 13th century by William of Ockham, an English Franciscan friar. The principle is known as the Ockham's razor: the simplest model providing acceptable accuracies should be preferred, since is likely to possess the larger explanatory power [Duda et al., 2001].

Based on these observations, we may finally state that the generalization ability of a model trades-off with its complexity. By observing the error rate that one commits on the training and on the test sets, we may extrapolate some important informations. These are the bases of the probabilistic induction principle of the statistical learning theory [Evgeniou et al., 1999; Vapnik, 1998].

3. Machine learning

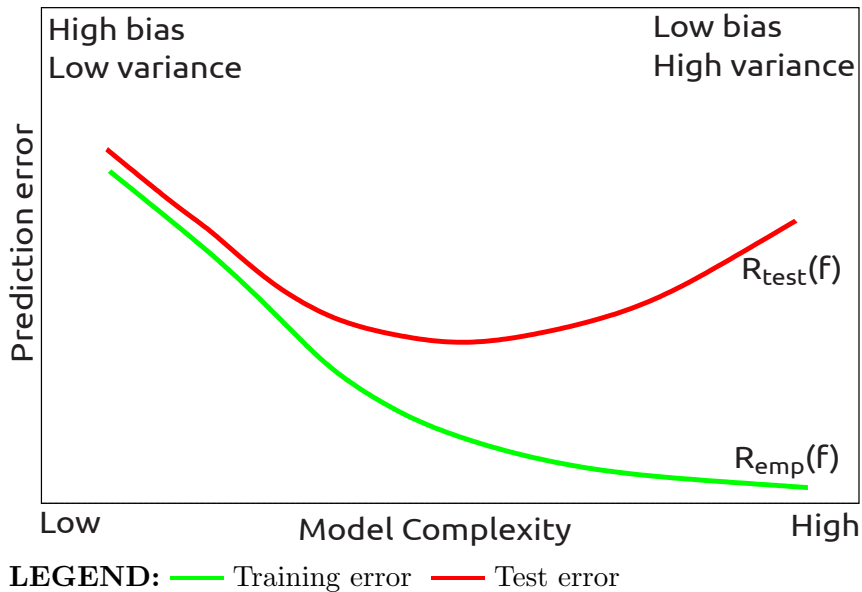


Figure 3.1: Bias-variance dilemma - Behaviour of training and test errors in a finite set situation for a general model f (modified from [Hastie et al., 2009]).

To avoid underfitting situation, it suffices to ensure that the family of models selected is rich enough to capture complex data relationships. To guarantee a certain degree of generalization, the final solution must choose a model $f \in \mathcal{F}$ that best approximates the expected error on unseen samples. In this setting, choosing the functional by a direct minimization of the ERM [Schölkopf and Smola, 2002; Vapnik, 1998]:

$$f^* = \min_{f \in \mathcal{F}} R_{\text{emp}}(f), \quad (3.5)$$

is not a good choice, since it provides an overfit of the data.

A solution to this issue is given by the structural risk minimization principle in the statistical learning framework (SRM) [Schölkopf and Smola, 2002; Vapnik, 1998]. One of the core concepts of the statistical learning theory builds on the consistency of the ERM principle. Vapnik [1998] stated formally what deducted above: as $n \rightarrow \infty$, minimizing Equation (3.5) converges towards the optimal (lowest achievable) $R(f^*)$, being f^* the optimal function. That is, ERM leads to the same solution as if minimizing directly $R(f)$. Formally, this may be presented as the following convergence in probability:

$$\lim_{n \rightarrow \infty} \mathbb{P}(\sup_{f \in \mathcal{F}} (R(f) - R_{\text{emp}}(f)) > \epsilon), \quad (3.6)$$

with $\epsilon > 0$. In this case, the consistency depends on the class of function \mathcal{F} . One needs to estimate feasible and admissible hypotheses spaces \mathcal{F} guaranteeing the convergence of the errors under the limit of $n \rightarrow \infty$. The problem is now to select the correct function among the family of functions guaranteeing this convergence.

The solution to this problem is to constraint the family of possible hypotheses \mathcal{F}^* to functions that minimize the observed error for the minimal complexity. Also, the choice

3.1 Learning from data

should be made in order to accommodate the consistency presented in Equation (3.6). The framework of the SRM translates these observations into the addition of a capacity term to the empirical risk $R_{\text{emp}}(f)$, which penalizes complex models that do not guarantee the uniform convergence above. In the SRM framework, Equation (3.5) is modified to take into account the model complexity with an additional term, a confidence interval depending on the model complexity defining an upper bound on the true risk [Schölkopf and Smola, 2002; Vapnik, 1998]:

$$R(f) \leq R_{\text{emp}}(f) + \psi\left(\sqrt{\frac{h}{n}}, (1 - \eta)\right). \quad (3.7)$$

Here $\psi\left(\sqrt{\frac{h}{n}}, (1 - \eta)\right)$ is an increasing function of $\frac{h}{n}$ and η . Specifically, being n the number examples, h is a capacity term (growing for increasing complexity) and $1 - \eta$ defines the probability of observing the above mentioned bound. As depicted from the uniform convergence of Equation (3.6), the capacity term goes to zero for $n \rightarrow \infty$.

A quantification of the capacity of the class of functions in \mathcal{F} may be given by the Vapnik-Chervonenkis (VC) dimension [Vapnik, 1998]. It is defined as the largest number of samples with binary labels in any configuration that can be shattered (solved in a classification sense) for a given \mathcal{F} . For instance, a linear separation in \mathbb{R}^2 has a VC dimension of 3 since it can separate any 3 points with binary labels in any configuration in \mathbb{R}^2 . The same holds for a VC dimension of 4 for planes (\mathbb{R}^3), and so on. For linear separating functions the VC dimension is $d + 1$.

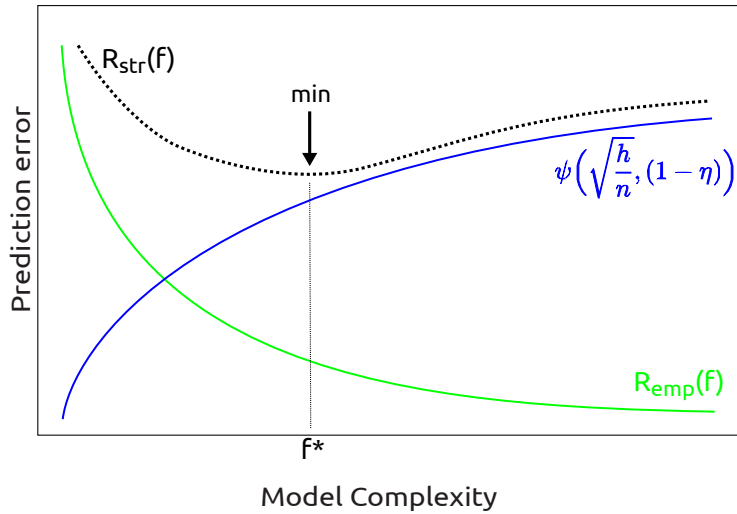
The SRM principle illustrated in Figure 3.2 can be seen as selecting a single model from the optimal set of functions \mathcal{F}^* satisfying the SRM, selected from a series of hypotheses families with increasing VC dimension (or capacity) $\mathcal{F}^1 \subset \mathcal{F}^2 \subset \dots \subset \mathcal{F}^\infty$, with $\mathcal{F} = \bigcup_{i=1}^{\infty} \mathcal{F}^i$. The choice of the optimal model will account for the variance-bias issue implicitly, as depicted in Figure 3.2. For a detailed derivation of the bounds, see [Alon et al., 1997; Evgeniou et al., 1999; Schölkopf and Smola, 2002].

Summing up, the statistical learning theory exploits the SRM principle to select models and their parameters, and it is implemented as a penalization term added to the empirical risk. However, even if this is an intuitive concept, it is hard to apply it to complex classes of functions, in particular since estimating the effective VC dimension is often infeasible. In Chapter 7 of [Hastie et al., 2009] an example of approximating the VC dimension by other complexity measures is given. In general, when applicable, the complexity of a model may be related to the internal parameters vector of a model. This will be the subject of Section 3.2.

3.1.1 A practical example

This Section is deemed to provide a short example of the general concepts introduced before. In this case, the k -Nearest Neighbour (k NN) classifier is employed. This method assigns to a previously unseen sample the label that is the most frequent among its k neighbours. The neighbourhood of \mathbf{x}_i corresponds to the k closest samples in terms of Euclidean distance, as $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$. Note that in this case, the model does not

3. Machine learning



LEGEND: — Training error — regularization term ··· Structural risk

Figure 3.2: Structural risk minimization - The regularized risk functional shows a minimum that trade-offs complexity and empirical risk minimization.

have any internal parameter to optimize and it does not require training (lazy learner). It only has a user defined hyperparameter k . Even if the k NN does not have internal parameters that can be exploited to estimate its complexity, the ratio n_s/k can be used instead, corresponding to the effective degrees of freedom [Hastie et al., 2009].

In Figure 3.3 three situations are illustrated. We observe that the model obtained with 100 NN separates the samples with a very smooth function, with training and test errors (in terms of the 0-1 loss) of 0.295 (29.5%), Figure 3.3(a), and a test one of 0.3, Figure 3.3(d). The model shows also a very low complexity, corresponding to $200/100 = 2$. In this case, the model is not flexible enough to learn an appropriate separation and it disregards the region of green samples in the lower left corner. In this case, we incur in underfitting.

The opposite situation is illustrated in Figure 3.3(c)-(d) for train and test errors respectively. It is observed when using a 1 NN classifier and the training samples are obviously always perfectly separated. In this case, the $R_{\text{emp}}(f) = 0$ and $R_{\text{test}}(f) = 0.310$. As further indicated by the large model complexity ($200/1 = 200$) this corresponds to overfitting.

Finally, the intermediate solution is given by a 15 NN, illustrated in Figure 3.3(b). The model is neither too complex nor too simple, with a complexity of $200/15 = 13.35$. The balanced errors $R_{\text{emp}}(f) = 0.215$ and $R_{\text{test}}(f) = 0.205$ also depicts that no extreme situations in the sense of Figure 3.1 occurred. Practically, the value of $k = 15$ has been obtained by minimizing the cross-validation error varying the value of k and retaining the one generating the best model in terms of accuracy. This issue will be discussed later in Section 3.3.

As mentioned, if the training set is infinitely large, the difference between empirical and expected risk will be reduced to zero. In this case we may compute the true risk since we know the distribution from which the data have been generated. Such error is called

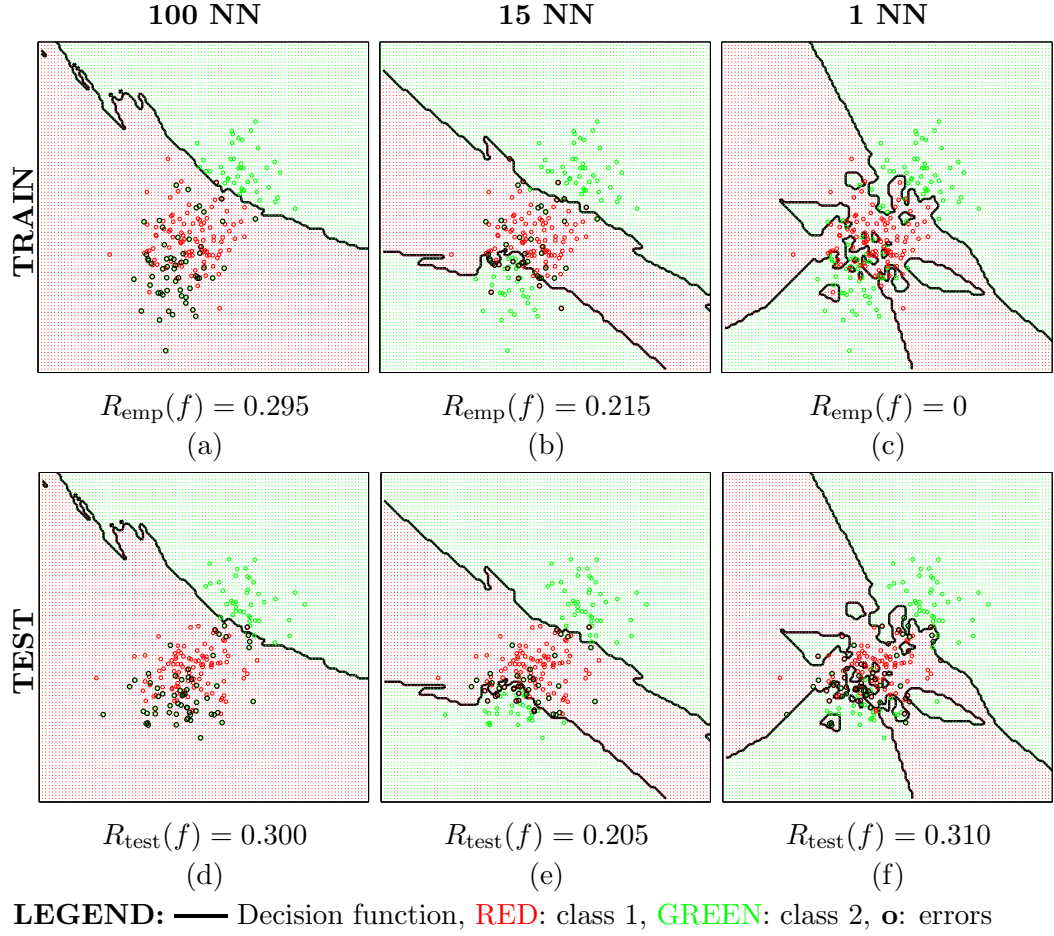


Figure 3.3: k NN classification errors - In (a), the model voting over 100 NN with training errors circled in black and corresponding $R_{\text{emp}}(f)$. In (b) and (c) the same but using 15 and 1 NN. In (d)-(f), the same as above but showing test errors and the corresponding $R_{\text{test}}(f)$. Train and test samples are two separate realizations from the same underlying $P(\mathbf{x}, y)$.

the Bayes rate, and model generating it is the optimal Bayes classifier.

The **RED** and **GREEN** classes have been generated accordingly to two bivariate normal distributions with $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and are illustrated in Figure 3.4.

- **RED CLASS.** Uni-modal distribution with 100 samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = (-0.5, -1), \boldsymbol{\Sigma} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$$

- **GREEN CLASS.** Bi-modal distribution with 50 samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (1, 1)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$ and 50 samples with $\boldsymbol{\mu} = (-1, -2)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$.

The optimal Bayes classifier is simply obtained by assigning a sample to the class maximizing the posterior probability, as:

$$f(\mathbf{x}) = \max_{y \in \{\text{RED}, \text{GREEN}\}} P(y|\mathbf{x}). \quad (3.8)$$

3. Machine learning

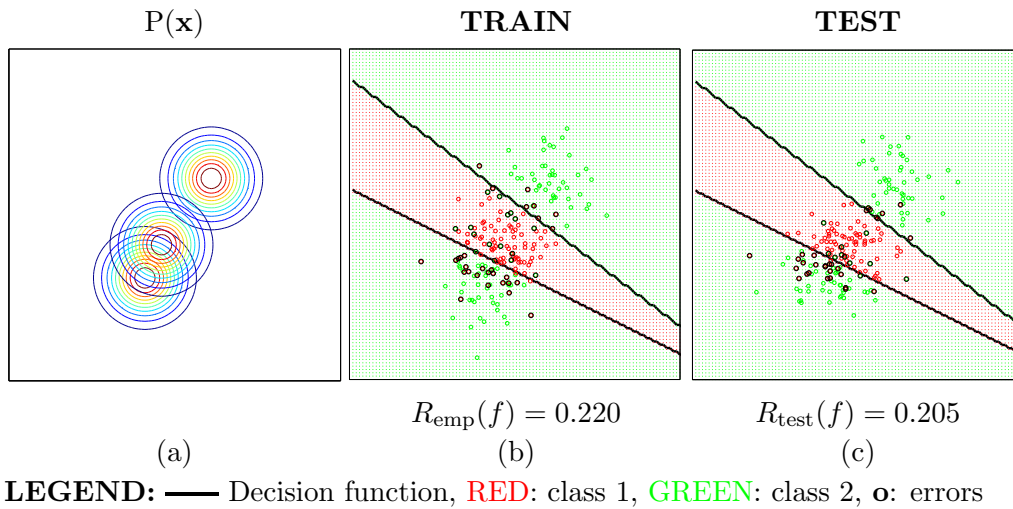


Figure 3.4: Bayes optimal classification - In (a), the underlying probabilities generating the data. In (b) the optimal Bayes model fitting the data with training errors black circles, and in (c) the same but outlining the test error (Bayes rate).

Since class covariances are equal, the decision boundaries are linear. Returning again to the k NN example, we may observe that the best model we obtained, is the one that better approximates the optimal Bayes classifier, both in sense of the error and of the decision function shape.

3.2 Connections with regularization theory

As mentioned, the estimation of the exact bound of Equation (3.7) may be a difficult task. To approximate the model complexity, we may want to replace the capacity term with a regularization penalizing complex models. The regularization theory has been introduced by Tikhonov and Arsenin [1977], and it was intended with the aim of solving ill-posed inversion problems issuing from the discretization of integral equations. The ill-posedness of a problem is given by Jacques Hadamard's definition of well-posedness, as follows [Hable and Christmann, 2011; Steinwart and Christmann, 2008]. An optimization problem is well-defined (or well-posed) if (i) a solution exists, (ii) it is unique and (iii) slight changes in the data generate very small changes in the model. By contradiction, an ill-posed problem is encountered when one of these conditions is not met.

Additionally, these concepts allow to introduce a problem often encountered in data analysis and machine learning: the curse of dimensionality [Hughes, 1968; Lee and Verleysen, 2007; Trunk, 1979]. As we will see in the next Chapter, it is easier to solve classification problems in high dimensional spaces. Since for a given and finite number of samples the space exponentially empties as the dimensions increase, linear separations are more likely to shatter samples. As the dimensionality increases, so does the number of valid separation of training sample, making the solution not unique (ill-posed problem

3.2 Connections with regularization theory

from condition (ii)). Additionally, this phenomenon makes the norm of vectors (of iid samples) grow proportionally to \sqrt{d} , while the variance remains constant. This heavily affects the meaning of Euclidean distances in high dimensional spaces [Lee and Verleysen, 2007]. Consequently, learning a model with a small training set and in high dimensions may be a very difficult or even an infeasible task, since the space is mostly empty and strong regularization to solve the problem is required. Finally, the penalization allows to avoid choosing models that would show small bias and large variance, ensuring that the problem is not ill-posed from condition (iii).

Regularization theory ensures the well-posedness of the problem by adding a term $\Omega(f)$ during the ERM to avoid training too complex models resulting in poor generalization ability, as indicated by the SRM theory. We may define a penalized or regularized risk as:

$$R_{\text{reg}}(f) = R_{\text{emp}}(f) + \gamma\Omega(f). \quad (3.9)$$

This is straightforward in the interpretation: the regularized risk minimization aims at finding a solution fitting well the training data ($R_{\text{emp}}(f)$ data dependent fitting) but at the same time finding a solution that do not incur in complexity penalizations ($\gamma\Omega(f)$ restricts the function class). The term γ is a user defined trade-off parameter controlling the amount of penalization of complex models. Provided that $R_{\text{emp}}(f)$ is convex, one chooses $\Omega(f)$ to be also convex, so that a unique minimum of $R_{\text{reg}}(f)$ exists.

There are different families of $\Omega(f)$, and a small review of main families of parametric and non-parametric penalties may be found in [Cherkassky and Mulier, 2007]. In general, we may encounter regularizers that enforce some prior guess on the distribution of model weights, or penalizers in the form of differential operators discarding models with high frequency in the input domain, such as Fourier-based [De Canditiis and De Feisb, 2006]. However, for many learning algorithms, the most of these capacity constraints are in the form of ℓ_p norms of the weight vector of the model as:

$$\ell_p(\mathbf{w}) = \|\mathbf{w}\|_p := \left(\sum_{i=1}^d |w_i|^p \right)^{\frac{1}{p}}. \quad (3.10)$$

Regularizers of this form with $p \neq 0$ often lead, provided the convexity of the loss function, to convex optimization problems. Equation (3.10) is illustrated for a vector of parameters $\mathbf{w} \in \mathbb{R}^d$ defining f , and $p \geq 1$. The family of ℓ_p norms possesses many interesting features, for instance a minimization using the ℓ_1 -norm (approximation of the usually infeasible ℓ_0 regularization) favours coefficients of \mathbf{w} to become 0, since for growing values of $\|\mathbf{w}\|_1$ the solution may only be encountered when some coefficients of \mathbf{w} are zero. For ℓ_2 norms, the solution is not sparse since when $\|\mathbf{w}\|_2$ grows the solution is met also with a dense \mathbf{w} . Moreover, these solutions from the regularization theory show connections with the SRM principle for a variety of important classes of functions [Evgeniou et al., 1999, 2002]. Parameter γ may be chosen so that nested subspaces of hypotheses spaces with growing complexity are created.

It is worth mentioning that other techniques aiming at controlling the complexity of the models exists, besides SRM and regularization theory. For example, we may mention

3. Machine learning

techniques such as the early stopping criterion, widely used in neural networks [Haykin, 1999]. In the Bayesian inference context one exploits prior information on the form of the admissible solution to shrink \mathcal{F} [Bishop, 2006], or noise injection, which is a complexity control simulating new noisy samples from $P(\mathbf{x}, y)$, building directly from the last Hadamard’s well-posedness [Bishop, 1995].

However, as a matter of fact, we simply added an additional hyperparameter to be estimated to the minimization problem. In the next Section, a brief summary of the model selection procedure is provided.

3.3 Hyperparameters optimization

As mentioned above, there are many methods which require the fitting of some extra hyperparameters before estimating their internal parameters. These hyperparameters strongly depend on the data at hand, and, when possible, should be set so that the final model generalizes optimally, i.e. provides a generalization error the closest possible to the expected risk. Practically, one looks for the hyperparameters Θ_h^* among all the possible Θ_h that satisfy:

$$\Theta_h^* = \arg \min_{\Theta_h} \mathcal{L}(f, \Theta_h, y). \quad (3.11)$$

where f corresponds, as before, to $f(\Theta, \mathbf{x})$.

In this case we evaluate a model learned on the training set using all the precautions given by Equation (3.9), but we evaluate it on the validation one. In this case, we hope that:

$$R(f) \approx \mathcal{L}(f(\Theta, \mathbf{x}), y). \quad (3.12)$$

where the \mathbf{x} and y are from the validation set, and the model parameters Θ are learned on the training set by the minimization of Equation (3.9).

In this case we assume that by employing the validation set as the testing one, it would provide an approximation of the true generalization error. In the most optimistic situation, and by disposing of very large sets, $R_{\text{validation}}(f) \approx R_{\text{test}}(f)$. However, disposing of an independent held-out validation set is rare, since labelled samples are costly. It is consequently hard to split the dataset into training, validation and testing set. However, samples required by a validation set may be used to complete the population of the training or the test set.

An important observation, directly relating to the ERM principle, is that one can not select the hyperparameters on the basis of the minimization of $R_{\text{reg}}(f)$. In this case, the selected model would bring to a severe overfitting of the training samples and consequently leading to a suboptimal model. We mention two main strategies. The first is the leave-one-out cross-validation (LOO-CV): one of the examples of the training set is kept apart and the model is trained using the $n_s - 1$ remaining samples. Then, the error is evaluated on the single held-out sample and its output stored. This is repeated for all the n_s combinations, and the corresponding estimated outputs are used to compute the validation error. Since

3.4 Models of machine learning

this could be computationally expensive even for moderately large training sets (e.g. think as $n_s = 100$ with $|\Theta_h| = 100$ parameters to test, it would result in training and predicting 10000 models), more efficient schemes exist. The second approach, probably the most used, is the generalization of the LOO-CV, the k -fold cross-validation. In this case, the subset splitting is performed for k random folds. The model is trained with the $k - 1$ groups and tested on the k th remaining block. The final validation error is the average of the errors obtained from each held-out fold. Finally, the hyperparameters defining a model minimizing the error are then retained to train the final model.

Other approximations of the expected generalization error may be obtained using bootstrap based estimations [Hastie et al., 2009], techniques close to the aforementioned ones. For instance, the leave-out bootstrap approach draws randomly a given number of samples from \mathbf{X}_s but with replacement, then it evaluates the model trained on the remaining part.

3.4 Models of machine learning

So far, we only considered the broad family of supervised algorithms in a general manner, i.e. an arbitrary model is learned from the training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$ exploiting the knowledge of the labels y from \mathcal{Y} . In the literature there exist many distinctions between methods, and this Section is deemed to introduce the main families of learning models. The first differentiation is between parametric and nonparametric models.

3.4.1 Parametric and non-parametric inference

Parametric models. Algorithms of this family are characterized by the availability of some prior knowledge in the form of the distribution generating the data. The modelling process aims estimating a finite set of parameters from the data, assumed to be a realization of the true and a-priori known Θ -parametrized distribution. Thus, the fitting of the data is based on an estimation of the parameters of the joint distribution $P(\Theta|\mathbf{x}, y)$ from its density $p(\Theta|\mathbf{x}, y)$, that best approximates the data.

A classical example for the classification of optical imagery is to fit the class-conditional probability by a multivariate normal distribution of the data with $\Theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. In this case, $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\Sigma}$ the covariance matrix of the d -variate distribution. It is assumed that $\{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})|y\}$. The modelling task is thus to find a f_Θ that belongs to the restricted class of functions:

$$\mathcal{F} = \left\{ f_\Theta = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \middle| \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \succ 0 \right\}, \quad (3.13)$$

with $\boldsymbol{\Sigma} \succ 0$ defining the positive definiteness of the covariance matrix¹. There are many models that belong explicitly or implicitly to this family, from both the supervised and the unsupervised family. The above example corresponds to the maximum likelihood classification (MLC) and may be solved by fitting the parameters by the expectation-maximization algorithm [Dempster et al., 1977].

¹A matrix \mathbf{M} is said to be positive definite if, for any nonzero column vector \mathbf{z} , $\mathbf{z}'\mathbf{M}\mathbf{z} > 0$.

3. Machine learning

Nonparametric models. While models belonging to the parametric family fit a pre-defined distribution to the data, nonparametric models fit the data by optimizing a functional directly on the observations, with a distribution-free approach. The adoption of this paradigm is often motivated by the too restrictive and rigid hypotheses given by parametric models. In this sense, nonparametric models try to directly fit a descriptive model to the data, without imposing any prior restriction on the form of the generating process. The number of total parameters may vary depending on the data characteristics, such as number of dimensions for linear models. Nonparametric models may be more difficult to interpret due to the strong data-dependent setting and additional information may be rarely retrieved from a trained model. For this reason, they are often qualified as black box models. However, non-parametric data-driven density estimation such as the kernel density estimator (KDE) may estimate the probability density function using only observed samples (thus in principle with infinite parameters, as $n \rightarrow \infty$). In addition, methods to retrieve posterior probabilities from fitted nonparametric decision functions are available. This category of methods is very large and comprehends the family of the kernel methods [Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004] and Gaussian processes [Rasmussen and Williams, 2006], neural networks [Haykin, 1999] and many others.

3.4.2 Supervised, unsupervised and semi-supervised models

Supervised learning. Models discussed so far require the presence of labels to exploit in the learning, i.e. pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are used for training, validation and testing. Many real world examples are characterized as supervised problems: for instance, let :
 $\mathcal{X} = \{\text{coordinates; soil type; porosity; water content; depth; Cesium content}\}$, with
 $\mathcal{Y} = \{\text{Lead content}\}$. The modelling of such relationships may be motivated by the fact that simply sampling the data in \mathbf{x} is much less costly than also measuring the corresponding amount of lead, or field / laboratory estimations of \mathbf{x} do not allow to retrieve directly lead content. In this case we want to predict the dependent variable y_i by learning the relationship with \mathbf{x}_i and generalize the model for the other available \mathbf{x} to obtain their estimated y . In addition, and most importantly, classification and regression (or spatial interpolation in this case) provide also generalization in the spatial domain of punctual measurements, thus generating maps and cartographic products [Kanevski et al., 2007, 2009]. Supervised learning is not limited to regression and classification problems, but it comprehends also novelty detection [Tax and Duin, 2004], density estimation [Schölkopf et al., 2001], feature extraction [Mika et al., 1999] and feature selection [Camps-Valls et al., 2010].

Unsupervised learning. At the opposite situation, we may encounter learning problems characterized by the availability of feature vectors \mathbf{x} alone, without the corresponding y . In this case, one wants to discover some hidden structures and partitioning in the data, without any knowledge of the y . This is known as unsupervised learning. The most of these approaches to data analysis are related to clustering (e.g. k -means [MacQueen,

3.4 Models of machine learning

1967], Gaussian mixture models (GMM) [Dempster et al., 1977], spectral clustering [Ng et al., 2002]), i.e. the task of automatically grouping similar samples into disjoint sets (see example below). However, many efforts have been also devoted to feature extraction, dimensionality reduction and manifold learning (e.g. PCA [Hotelling, 1933], Laplacian eigenmaps [Belkin and Niyogi, 2003]), and to density estimation (e.g. GMM, KDE). In Figure 3.5, an example of parametric clustering (based on GMM) is given. In this case, a mixture of three Gaussians has been fitted to the data, as:

$$p(\mathbf{x}) = \sum_{i=1}^k \pi_i \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad \sum_{i=1}^k \pi_i = 1, \quad 0 \leq \pi_i \leq 1, \quad (3.14)$$

with $k = 3$ for the example illustrated below. Here, π_i are the mixing coefficient of each distribution. This may be seen as a density estimation step and then an inference process assigning regions of the space maximizing the posteriors probability to belong to a given component of the mixture to the identifier of such component, the cluster index. In this example, its performance is compared to the supervised Bayes classifier (providing very similar outcomes for unimodal classes), since the true distribution is known.

To measure the difference between the fitted distribution $p_{\text{GMM}}(\mathbf{x})$ and the original $p_{\text{true}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the Kullback-Leibler divergence (also known as marginal entropy) $\text{KL}(\mathcal{P}_{\text{GMM}}||\mathcal{P}_{\text{true}})$ has been adopted [Bishop, 2006]. For two multivariate (d -dimensional) normal distributions $\mathcal{N}_1(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}_2(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ with unequal covariances, the KL distance is:

$$\begin{aligned} \text{KL}(\mathcal{N}_1(\mathbf{x})||\mathcal{N}_2(\mathbf{x})) &= \int \mathcal{N}_1(\mathbf{x}) \log \frac{\mathcal{N}_1(\mathbf{x})}{\mathcal{N}_2(\mathbf{x})} d\mathbf{x} \\ &= \frac{1}{2} \left(\text{trace}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \ln \left(\frac{\det(\boldsymbol{\Sigma}_1)}{\det(\boldsymbol{\Sigma}_2)} \right) - d \right). \end{aligned} \quad (3.15)$$

In this case, KL distance has been computed for each pair of normal distributions, between the one generating the cluster and the fitted one, and then averaged. It gives a value $\text{KL}(\mathcal{N}_1||\mathcal{N}_2) = 0.055$, indicating a very good fit.

Semi-supervised learning. A third family of learning algorithms issues from an intermediate situation, in which a very large amount of samples are available, but only a small fraction of it is labelled, as shown in Figure 3.6(a). This is the case in many real world situations, and in particular for remote sensing image processing: all the pixels composing the image are available, but usually only a small portion of them is labelled, since assigning ground truth to pixels is a very costly or time consuming process. One can improve a failing supervised model, Figure 3.6(b), by inferring some knowledge extracted from the unlabelled data. Additional information corresponds to the geometrical distribution of the unlabelled samples providing insights on the form of the space in which data lie (the manifold). Many approaches implement this intuition by adding some marginal information: in the example illustrated in Figure 3.6 a graph is estimated using 5 NN, i.e. putting a link between samples \mathbf{x}_i and \mathbf{x}_j if they are among their 5 NN. One can in

3. Machine learning

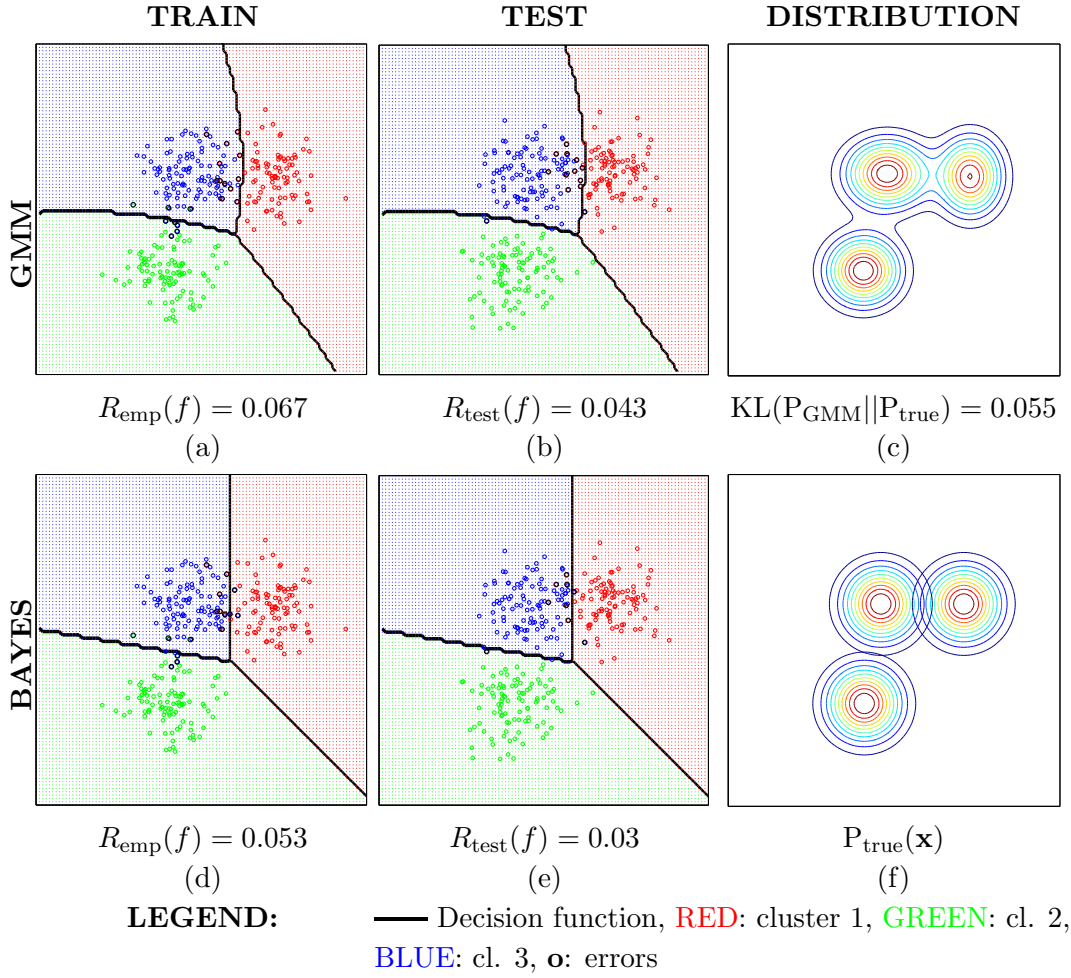


Figure 3.5: GMM-based clustering - In (a) and (b) the partitioning of the space using a GMM with training (not used to estimate the distributions) and testing samples respectively, along with their errors. Note that the training error has been computed only for illustrative purposes. In (c) the density is estimated by the GMM. In (d)-(f), the same as above but showing the Bayes classification, based on the known class distributions.

- **RED** 100 samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (2, 1)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$
- **GREEN** 100 samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (-1, -2)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$
- **BLUE** 100 samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (-0.5, 1)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$

the following propagate the known labels via the paths linking the samples in the graph. The recently developed semi-supervised paradigm is generally based on three principal assumptions [Chapelle et al., 2006]:

Cluster assumption and low density separation This assumption states that two samples lying in the same cluster are likely to belong to the same class. This statement can also be seen from the opposite point of view: separating boundaries should be favoured to lie in low density areas. These observations enable the use of transductive methods, i.e. learning by incrementally updating a model on the basis of the gradual labelling of the unlabelled set and to shift separating boundaries accommodating this assumption.

Smoothness assumption If two points lie in the same high density region, so will their output. This enables a series of generative models for semi-supervised learning.

Manifold assumption Disregarding the statistical distribution of the samples, one can note that high dimensional data will likely lie on a lower dimensional representation, i.e. lying close on a manifold. These samples close on the manifold will likely share some properties, such as the label. This motivates the use of graph-based learning methods in supervised settings (see example in Figure 3.6).

Recall that supervised, unsupervised and semi-supervised models may be either parametric or nonparametric.

3.4.3 Linear and nonlinear models for data analysis

A last distinction that has to be made is between linear and nonlinear models. These categories are usually discriminated by the form of the decision function they draw in the input space. Consequently, nonlinear models are able to model and deal with nonlinear relationships between data samples. Models producing a linear separation (e.g. linear discriminant) or a linear function approximation (e.g. linear regression) are assumed to be the simplest ones with low complexity, since usually depending on few parameters. For instance, the linear regression $\mathbf{y} = \mathbf{X}\mathbf{w}$ needs only the estimation of d weight parameters \mathbf{w} to interpolate the data in \mathbf{X} . On the other hand, while being much more powerful, nonlinear models are usually complex and need computationally expensive training algorithms. For instance, when using neural networks, the models need to be specified by an architecture (e.g. number of hidden layers, number of neurons). Then, the learning step estimates weights linking the neurons, by iteratively fitting an error function. During training, one should pay attention to stop the learning process before overfitting the data, since the learning step has to fit a number of model weights growing exponentially with the data dimensionality and the number of neurons per layer. More details can be found in [Haykin, 1999]. Although the cost may seem high, these models are very powerful and can describe data in any form, following any distribution and providing accurate solutions to supervised classification and regression problems. Additionally, once trained, the prediction step is usually fast, being a weighted sum of the inputs. For these reasons, they have been successfully applied in a variety of problems.

3. Machine learning

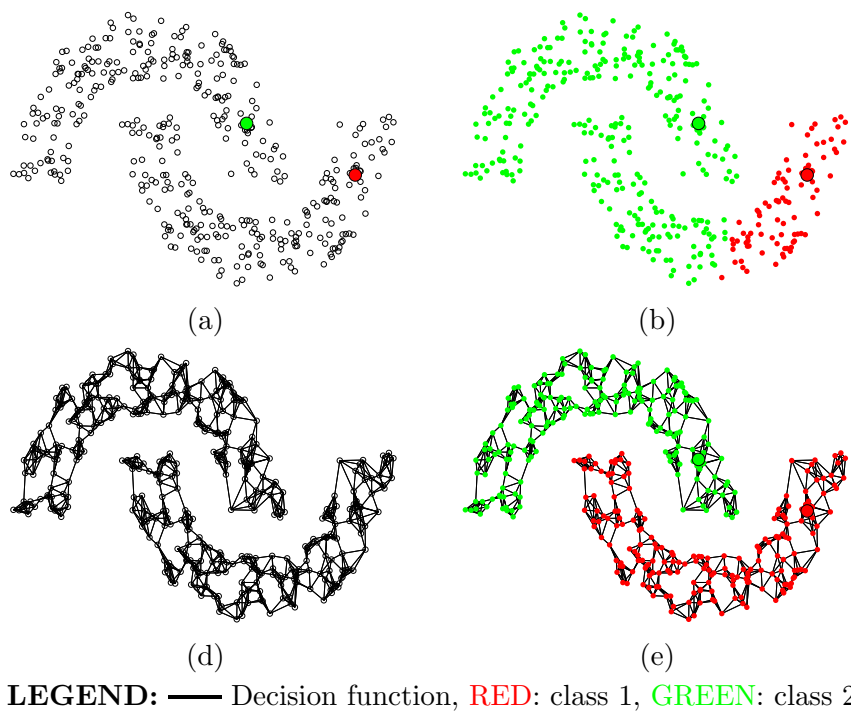


Figure 3.6: Semi-supervised classification toy example - (a) available information: distribution $p(\mathbf{x})$ and 1 labeled sample per class. (b) a suboptimal supervised model (minimum Euclidean distance), (c) the 5 NN graph and (d) output of a semi-supervised model accounting for the local structure (label propagation on connected regions, e.g. [Camps-Valls et al., 2007a]).

Another important family of nonlinear algorithms are the kernel methods [Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004], that will be further described in the next Chapter. The underlying idea of this family is that a nonlinear analysis in the input space can be obtained by running standard linear algorithms in some higher dimensional feature space [Aizerman et al., 1964].

Chapter 4

Learning with kernels

This Chapter introduces to kernel methods by illustrating main features and properties of this family of learning algorithms. In Section 4.1 the kernel ridge regression is exploited to introduce the reasoning behind kernel methods. Then, Section 4.2 reports main characteristics and properties of these nonlinear algorithms, in relation to Chapter 3. Section 4.3 draws some considerations.

4.1 A motivating example: from least-squares linear regression to the kernel ridge regression

As introduced in the previous Chapter, the theory for machine learning, pattern recognition and data mining is well defined and robust. Within this framework, the family of kernel-methods is a unifying theory for linear and nonlinear analysis for general data (e.g. vectors, strings, sets, structured data, graphs, etc.) offering advanced tools for many learning problems such as classification, regression, clustering, density estimation, etc. using a common and well-founded theoretical framework.

Kernel methods may be summarized in five different points [Campbell, 2002; Camps-Valls and Bruzzone, 2009; Hoffmann et al., 2008; Schölkopf and Smola, 2002; Schölkopf et al., 1999; Shawe-Taylor and Cristianini, 2004]:

1. They map samples into an embedding higher dimensional vector space \mathcal{H} .
2. In \mathcal{H} , the relationships among data samples are likely to be linear. A linear algorithm in \mathcal{H} suffices to solve the learning task.
3. The theory reformulate the learning task such that the explicit coordinates in \mathcal{H} are not needed.
4. Kernel methods depend only on inner products between samples in \mathcal{H} , and they can be computed efficiently using kernel functions taking as argument only the data in their original input space.
5. Depending on the kernel function, learning in \mathcal{H} returns a nonlinear solution in the input space.

4. Learning with kernels

Let us introduce kernel methods using a practical example, allowing to draw a direct link with the concepts presented in Chapter 3. This is done by formulating the kernel ridge regression from its standard least-squares counterpart.

A function estimation (regression) problem may be approached by estimating a functional

$$f : \mathbb{R}^d \rightarrow \mathbb{R}$$

$$\mathbf{x} \mapsto f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}'\mathbf{x} = \sum_{i=1}^{n_s} w_i x_i. \quad (4.1)$$

The operator $\langle \cdot, \cdot \rangle$ indicates the inner product between weights \mathbf{w} and observations \mathbf{x} .

Definition 4 (Inner product, [Axler, 1997], Chapter 6) *The inner product operator $\langle \cdot, \cdot \rangle$ generalizes the dot product to abstract vector spaces \mathcal{H} (in this thesis limited to the field of \mathbb{R}). This operation is defined as $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$, and satisfies the following properties. Let \mathbf{a}, \mathbf{b} and \mathbf{c} be three vectors in \mathcal{H} and a scalar $v \in \mathbb{R}$:*

1. *Positive definiteness:* $\langle \mathbf{a}, \mathbf{a} \rangle_{\mathcal{H}} \geq 0$, with $\langle \mathbf{a}, \mathbf{a} \rangle_{\mathcal{H}} = 0$ only if $\mathbf{a} = \mathbf{0}$
2. *Symmetry:* $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathcal{H}} = \langle \mathbf{b}, \mathbf{a} \rangle_{\mathcal{H}}$
3. *Linearity (additivity + homogeneity):* $\langle v\mathbf{a} + \mathbf{b}, \mathbf{c} \rangle_{\mathcal{H}} = v\langle \mathbf{a}, \mathbf{c} \rangle_{\mathcal{H}} + \langle \mathbf{b}, \mathbf{c} \rangle_{\mathcal{H}}$

By generalization of the dot product (recall that we limit to real spaces), the following properties also hold:

4. $\|\mathbf{a}\|_{\mathcal{H}} = \sqrt{\langle \mathbf{a}, \mathbf{a} \rangle_{\mathcal{H}}}$
5. $\cos \omega = \frac{\langle \mathbf{a}, \mathbf{b} \rangle_{\mathcal{H}}}{\|\mathbf{a}\|_{\mathcal{H}} \|\mathbf{b}\|_{\mathcal{H}}}$
6. $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathcal{H}} = 0$ if $\mathbf{a} \perp \mathbf{b}$

Going back to our regression problem, the issue is to find the weights \mathbf{w} that produces the best fit of the training pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_s}$. In vector notation, one wants to minimize the misfits between the true outputs \mathbf{y} and the predicted ones $\mathbf{X}\mathbf{w}$. Note that we have now the output vector $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{w} \in \mathbb{R}^d$. By considering the sum-of-squares loss function $\mathcal{L}(\mathbf{X}\mathbf{w}, \mathbf{y}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$, the optimal \mathbf{w} can be obtained by equating the derivative of the loss with respect to the parameters, as $\frac{\partial \mathcal{L}(\mathbf{X}\mathbf{w}, \mathbf{y})}{\partial \mathbf{w}} = 0$:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 &= \frac{\partial \mathcal{L}}{\partial \mathbf{w}} (\mathbf{y}'\mathbf{y} + (\mathbf{X}\mathbf{w})'(\mathbf{X}\mathbf{w}) + 2(\mathbf{X}\mathbf{w})'\mathbf{y}) \\ &= \frac{\partial \mathcal{L}}{\partial \mathbf{w}} (\mathbf{y}'\mathbf{y} + \mathbf{w}'\mathbf{X}'\mathbf{X}\mathbf{w} + 2\mathbf{w}'\mathbf{X}'\mathbf{y}) = 2\mathbf{X}'\mathbf{X}\mathbf{w} - 2\mathbf{X}'\mathbf{y}. \end{aligned} \quad (4.2)$$

Letting Equation (4.2) equal to 0 gives:

$$2\mathbf{X}'\mathbf{X}\mathbf{w} - 2\mathbf{X}'\mathbf{y} = 0 \Leftrightarrow \mathbf{X}'\mathbf{X}\mathbf{w} = \mathbf{X}'\mathbf{y}.$$

$$\boxed{\mathbf{w} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}. \quad (4.3)$$

As it was introduced in the previous Chapter, this may lead to either to overfitting or to ill-posed situations, in particular if \mathbf{X} is high-dimensional with $d > n$. In the first case, it may be solved with zero error. In the second case, depending on the row rank of

4.1 From least-squares to kernel ridge regression

\mathbf{X} , $\mathbf{X}'\mathbf{X}$ may not be invertible. One possible approach is to reduce the capacity of the learner, by imposing the regularization term $\gamma\Omega(f) = \gamma\|\mathbf{w}\|^2$. In this case, we obtain the regularized squared loss $\mathcal{L}(\mathbf{X}\mathbf{w}, \mathbf{y}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \gamma\|\mathbf{w}\|^2$, and it favours solutions with a small weight vector in norm, that is, no features participate with very large weights (or dominates) to the solution.

Thus, the regularized least squares regression problem, also known as a ridge regression, is the following:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} (\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \gamma\|\mathbf{w}\|^2) &= \frac{\partial \mathcal{L}}{\partial \mathbf{w}} (\mathbf{y}'\mathbf{y} + \mathbf{w}'\mathbf{X}'\mathbf{X}\mathbf{w} + 2\mathbf{w}'\mathbf{X}'\mathbf{y} + \gamma\mathbf{w}'\mathbf{w}) \\ &= 2\mathbf{X}'\mathbf{X}\mathbf{w} - 2\mathbf{X}'\mathbf{y} + 2\gamma\mathbf{w}. \end{aligned} \quad (4.4)$$

With Equation (4.4) equal to 0:

$$\begin{aligned} 2\mathbf{X}'\mathbf{X}\mathbf{w} - 2\mathbf{X}'\mathbf{y} + 2\gamma\mathbf{w} &= 0 \Leftrightarrow \mathbf{X}'\mathbf{X}\mathbf{w} + \gamma\mathbf{w} = \mathbf{X}'\mathbf{y}. \\ (\mathbf{X}'\mathbf{X} + \gamma\mathbf{I})\mathbf{w} = \mathbf{X}'\mathbf{y} &\Leftrightarrow \boxed{\mathbf{w} = (\mathbf{X}'\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}}. \end{aligned} \quad (4.5)$$

The output of a previously unseen test sample \mathbf{x} can now be estimated by using $y = \mathbf{w}'\mathbf{x}$. By taking advantage of the dual properties of the weight vectors, we know that the minimum norm weight will always lie in the span of \mathbf{X} [Shawe-Taylor and Cristianini, 2004]. We can rewrite $\mathbf{w} = \mathbf{X}'\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \in \mathbb{R}^n$ are the dual weights [De Bie et al., 2004; Guyon et al., 1992]. In Shawe-Taylor and Cristianini [2004] this is outlined by verifying that \mathbf{w} is a linear combination of \mathbf{X} , from Equation (4.3).

By plugging the dual weights in Equation (4.4) we obtain:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}} (\|\mathbf{y} - \mathbf{X}\mathbf{X}'\boldsymbol{\alpha}\|^2 + \gamma\|\mathbf{X}'\boldsymbol{\alpha}\|^2) &= \frac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}} (\mathbf{y}'\mathbf{y} + \boldsymbol{\alpha}'(\mathbf{X}\mathbf{X}'\mathbf{X}\mathbf{X}')\boldsymbol{\alpha} - 2\boldsymbol{\alpha}'\mathbf{X}\mathbf{X}'\mathbf{y} + \gamma\boldsymbol{\alpha}'\mathbf{X}\mathbf{X}'\boldsymbol{\alpha}) \\ &= 2(\mathbf{X}\mathbf{X}'\mathbf{X}\mathbf{X}')\boldsymbol{\alpha} - 2\mathbf{X}\mathbf{X}'\mathbf{y} + 2\gamma\mathbf{X}\mathbf{X}'\boldsymbol{\alpha}. \end{aligned} \quad (4.6)$$

Putting Equation (4.6) equal to 0 gives now

$$\begin{aligned} 2(\mathbf{X}\mathbf{X}'\mathbf{X}\mathbf{X}')\boldsymbol{\alpha} - 2\mathbf{X}\mathbf{X}'\mathbf{y} + 2\gamma\mathbf{X}\mathbf{X}'\boldsymbol{\alpha} &= 0 \Leftrightarrow (\mathbf{X}\mathbf{X}'\mathbf{X}\mathbf{X}')\boldsymbol{\alpha} + \gamma\mathbf{X}\mathbf{X}'\boldsymbol{\alpha} = \mathbf{X}\mathbf{X}'\mathbf{y}. \\ \mathbf{X}\mathbf{X}'(\mathbf{X}\mathbf{X}' + \gamma\mathbf{I})\boldsymbol{\alpha} = \mathbf{X}\mathbf{X}'\mathbf{y} &\Leftrightarrow \boxed{\boldsymbol{\alpha} = (\mathbf{X}\mathbf{X}' + \gamma\mathbf{I})^{-1}\mathbf{y}}. \end{aligned} \quad (4.7)$$

Now we can rewrite the functional of the original regression problem:

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle = \boldsymbol{\alpha}'\mathbf{X}\mathbf{x} = \langle \mathbf{X}'(\mathbf{X}\mathbf{X}' + \gamma\mathbf{I})^{-1}\mathbf{y}, \mathbf{x} \rangle = \mathbf{x}'\mathbf{X}'(\mathbf{X}\mathbf{X}' + \gamma\mathbf{I})^{-1}\mathbf{y}. \quad (4.8)$$

Interestingly, the solution of any test point depends on its dot product with training samples $y = \boldsymbol{\alpha}'\mathbf{X}\mathbf{x}$, and $\boldsymbol{\alpha}$ is in the span of the rows of the data matrix \mathbf{X} .

If we define the kernel function $k(\mathbf{x}, \mathbf{x}_i) = \langle \mathbf{x}, \mathbf{x}_i \rangle_{\mathcal{H}}$ further generalizing the inner product operator to some unknown space \mathcal{H} , we can see that Equation (4.8) yields:

$$f(\mathbf{x}) = \langle \mathbf{X}'\boldsymbol{\alpha}, \mathbf{x} \rangle = \boldsymbol{\alpha}'\mathbf{X}\mathbf{x} = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i). \quad (4.9)$$

4. Learning with kernels

That is the form of the kernel ridge regression. By replacing all the inner product matrices $\mathbf{X}\mathbf{X}'$, we obtain $\boldsymbol{\alpha} = (\mathbf{K} + \gamma\mathbf{I})^{-1}\mathbf{y}$ where \mathbf{K} is the matrix containing all the evaluations of the kernel function $k(\cdot, \cdot)$ between the training samples in \mathbf{X} as $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. As a side note, from Equation (4.7) one can see that ridge regression can naturally predict d_{out} multiple outputs $\mathbf{Y} = [\mathbf{y}_1 \dots \mathbf{y}_{d_{\text{out}}}]$ by replacing \mathbf{y} with \mathbf{Y} .

In this example, we discussed the principles of the kernelization of an algorithm, in this case the ridge regression. As mentioned in the introduction, by adopting a specific nonlinear $k(\cdot, \cdot)$, the ridge regression becomes a nonlinear function estimation method. The ingredients of the kernel methods used throughout this Thesis, will be detailed in the next Section.

4.2 Kernel methods: theory and regularization

In the previous Section, we illustrated how we can transform an algorithm from its standard form, the primal, to a more flexible and nonlinear formulation that can be expressed in terms of kernel functions, the dual [Suykens and Alzate, 2010]. Kernel functions provide a measure of similarity between the samples, exactly as the inner product does. Furthermore, such use of the inner products, extends the development of kernel learning algorithms by using simple analytic geometry and linear algebra [Schölkopf et al., 1999]. In fact, any algorithm that can be reformulated so that data matrices appear only in the form of inner products, can be rewritten using kernel functions. This is the kernel trick [Aizerman et al., 1964].

Definition 5 (Kernel function) *Let ϕ be a feature map to the vector space \mathcal{H} endowed with an inner product such that:*

$$\begin{aligned} \phi : \mathcal{X} &\rightarrow \mathcal{H} \\ \mathbf{x} &\mapsto \phi(\mathbf{x}) \end{aligned} \tag{4.10}$$

The similarity can be evaluated by the inner product into this space, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. For two data samples $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$, we define the kernel function as

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\rightarrow \mathbb{R} \\ (\mathbf{x}_i, \mathbf{x}_j) &\mapsto k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}}. \end{aligned} \tag{4.11}$$

Note that for notational convenience, the subscript \mathcal{H} is dropped and the space in which the inner product is evaluated will be clear from the context.

Additionally, linking Definition 4 to Definition 5, we may observe that the kernel function is positive definite, thus generating a kernel matrix of inner products (a Gram matrix) that is in turn positive definite. This implies positivity on the diagonal and symmetry of the kernel operator, $k(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_j, \mathbf{x}_i)$ and thus $\mathbf{K} = \mathbf{K}'$.

It appears from the Definition 5 that by replacing dot products of any algorithm by kernel evaluations, we may reformulate the linear method to work in some feature

4.2 Kernel methods: theory and regularization

space, possibly of much higher (even infinite) dimensionality, defined by the mapping $\phi(\cdot)$. However, the optimal function $\phi(\cdot)$ projecting data samples in \mathcal{H} is not known a-priori. The only information about the desirable function is that it must enforce linear relationships in the projected space. However, as mentioned in the introduction of this Chapter, explicit coordinates are not needed, as we will see in the following.

To obtain the same formulation of the kernel ridge regression but starting from a different point of view, we can rewrite the introductory example in Section 4.1 by using the samples mapped to \mathcal{H} as defined in Definition 5, $\mathbf{x} \mapsto \phi(\mathbf{x})$, and by building our data matrix of mapped samples Φ , which contains the coordinates of the samples in \mathcal{H} . By solving and exploiting Definition 5, the solution of the kernel ridge regression is obtained.

4.2.1 Reproducing kernel Hilbert space

To each (valid) kernel corresponds a unique hypothesis space \mathcal{H} , as implicitly assumed in the Definition 5, for a specific feature map $\phi(\cdot)$. This space of functions \mathcal{H} is said to be a Hilbert space if the inner product is a valid operation, explicitly defining an inner product space as illustrated in Definition 4. It is a generalization of the Euclidean space to an abstract vector space, composed by a finite or infinite number of dimensions [Aronszajn, 1950; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004].

By definition, \mathcal{H} is the space of functions (or hypotheses) to be used in the learning process. These are linear combinations of the weight vectors and the training samples as

$$\mathcal{H} = \left\{ f(\cdot) = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i) = \sum_{i=1}^n \alpha_i k(\cdot, \mathbf{x}_i) \mid \mathbf{x}_i \in \mathcal{X}, \alpha_i \in \mathbb{R}, i = 1, \dots, n \right\}. \quad (4.12)$$

In this space, an addition of functions f and g is expressed as $(f + g)(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$, and the multiplication of a function by a scalar $(\lambda f)(\mathbf{x}) = \lambda f(\mathbf{x})$ underlining that \mathcal{H} is a vector space. Finally, as observed above, a Hilbert space is qualified as reproducing kernel Hilbert space (RKHS), if there exist a kernel function k which satisfy the reproducing property $\langle f(\cdot), k(\cdot, \mathbf{x}) \rangle = f(\mathbf{x})$ and specifically $\langle k(\cdot, \mathbf{x}_i), k(\cdot, \mathbf{x}_j) \rangle = k(\mathbf{x}_i, \mathbf{x}_j)$.

If we now define two functions that are elements of \mathcal{H} as $f(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{x})$ and $g(\mathbf{x}) = \sum_{j=1}^l \beta_j k(\mathbf{z}_j, \mathbf{x})$, their inner product is:

$$\begin{aligned} \langle f, g \rangle &= \sum_{i=1}^n \sum_{j=1}^l \alpha_i \beta_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{z}_j) \rangle \\ &= \sum_{i=1}^n \sum_{j=1}^l \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{z}_j) = \sum_{j=1}^l \beta_j f(\mathbf{z}_j) = \sum_{j=1}^l \alpha_j g(\mathbf{x}_j) \geq 0. \end{aligned} \quad (4.13)$$

Since f, g are positive definite, so is \mathbf{K} , and $\langle f, f \rangle = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) = \alpha' \mathbf{K} \alpha \geq 0$.

Following Shawe-Taylor and Cristianini [2004], the reproducing property of the kernel may be observed again from the Equation (4.13), with $k(\cdot, \mathbf{x}) = \langle \cdot, \phi(\mathbf{x}) \rangle$

$$\mathcal{H}_f = \langle f, k(\cdot, \mathbf{x}) \rangle = \left\langle \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i), \phi(\mathbf{x}) \right\rangle = \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{x}) = f(\mathbf{x}). \quad (4.14)$$

4. Learning with kernels

This also defines the reproducing kernel of \mathcal{H}_f , i.e. it is possible to work in a unique \mathcal{H}_f by adopting any valid kernel. As outlined in [Hastie et al., 2009] the $f(\mathbf{x})$ is a solution of any convex problem in the RKHS of the form

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\mathbf{x}_i, f(\mathbf{x}_i)) + \gamma \mathbf{\Omega}(f), \quad (4.15)$$

providing a property called the representer theorem [Schölkopf and Smola, 2002]. Because of $\mathbf{\Omega}(f) = \|f\|^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)$ this may be rewritten as:

$$\min_{\boldsymbol{\alpha}} \mathcal{L}(\mathbf{K}, \boldsymbol{\alpha}, \mathbf{y}) + \boldsymbol{\alpha}' \mathbf{K} \boldsymbol{\alpha}. \quad (4.16)$$

Interestingly, the minimizer of Equation 4.16 $f(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{x})$ gives a representation of the function f as a weighted sum of kernels centred on each training data point.

Note that the regularization is now implicit in the formulation itself by considering the term $\mathbf{\Omega}(f) = \|f\|^2 = \boldsymbol{\alpha}' \mathbf{K} \boldsymbol{\alpha}$, and the coefficients of the solution optimize the regularized risk. In the setting above, the squared norm of the functional is penalized, corresponding to a solution with small norm of the primal weight vector in the RKHS $\mathbf{w}^{\mathcal{H}}$ due to the primal-dual relationship [Suykens and Alzate, 2010]. This norm decreases as the smoothness of the function increases, i.e. it varies slowly between two close (similar) points, by avoiding variables taking large coefficients of the weight vector.

This result is very useful, because a direct minimization of the weight norm is infeasible since $\mathbf{w}^{\mathcal{H}}$ may live in an infinite dimensional space. However, minimizing in the dual formulation the solutions with small $\boldsymbol{\alpha}' \mathbf{K} \boldsymbol{\alpha}$ corresponds to $\boldsymbol{\alpha}' \boldsymbol{\Phi} \boldsymbol{\Phi}' \boldsymbol{\alpha}$, which, in turn is related to $\mathbf{w}'^{\mathcal{H}} \mathbf{w}^{\mathcal{H}}$. This further illustrates the reproducing property illustrated in Equation (4.14). Also, by the connection between Equation (4.16) and Equation (4.14), is clear that by adopting a kernel function one implicitly adopts a form on the hypothesis space.

By plugging the sum-of-squares error in Equation (4.16) and solving it, it boils down to the solution of the kernel ridge regression directly (in its dual form is also known as regularization networks [Evgeniou et al., 2002]). Moreover, by plugging specific cost functions different models can be retrieved, for instance, support vector machines (SVM) by plugging the hinge loss (see Chapter 6) or SVM for regression (SVR) by adopting an ϵ -insensitive loss. For detailed proofs and high level explanations, see [Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Suykens and Alzate, 2010].

4.2.2 Operations in the RKHS

It was shown that in the RKHS we may solve infinite dimensional problems with a linear combination of kernel functions only requiring samples in their finite dimensional input space. In other words, a possibly infinite dimensional minimization problem reduces to minimizing over \mathbb{R}^n .

In the RKHS a series of basic operations stems from the fact that the RKHS \mathcal{H} is a vector space endowed with an inner product (and consequently with a norm). Following

4.2 Kernel methods: theory and regularization

[Camps-Valls and Bruzzone, 2009] and [Gómez-Chova et al., 2010] the operations that are important for the rest of the Thesis are detailed here. For two samples $\mathbf{x}, \mathbf{z} \in \mathcal{X}$, we have:

Translation A sample in \mathcal{H} can be translated by

$$\tilde{\phi}(\mathbf{x}) = \phi(\mathbf{x}) + \mathbf{\Gamma} \text{ with } \mathbf{\Gamma} \in \mathcal{H} \quad (4.17)$$

where $\mathbf{\Gamma} = \{\gamma_1, \dots, \gamma_n\}$ is a vector of translations restricted to lie in the span of $\{\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_n)\}$. Then, the inner product between translated maps is:

$$\begin{aligned} \langle \tilde{\phi}(\mathbf{x}), \tilde{\phi}(\mathbf{z}) \rangle &= \langle \phi(\mathbf{x}) + \mathbf{\Gamma}, \phi(\mathbf{z}) + \mathbf{\Gamma} \rangle \\ &= \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle + \langle \phi(\mathbf{x}), \mathbf{\Gamma} \rangle + \langle \mathbf{\Gamma}, \phi(\mathbf{z}) \rangle + \langle \mathbf{\Gamma}, \mathbf{\Gamma} \rangle \\ &= k(\mathbf{x}, \mathbf{z}) + \sum_{i=1}^n \gamma_i k(\mathbf{x}, \mathbf{x}_i) + \sum_{i=1}^n \gamma_i k(\mathbf{x}_i, \mathbf{z}) + \sum_{i,j=1}^n \gamma_i \gamma_j k(\mathbf{x}_i, \mathbf{x}_j) \\ &= \tilde{k}(\mathbf{x}, \mathbf{z}). \end{aligned} \quad (4.18)$$

Centering By exploiting the above property we can centre the data (zero mean) in the RKHS by letting $\mathbf{\Gamma} = -\boldsymbol{\mu}^{\mathcal{H}}$, with:

$$\boldsymbol{\mu}^{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^n \phi(\mathbf{x}_i) \quad (4.19)$$

since $\boldsymbol{\mu}^{\mathcal{H}}$ is a linear combination of the samples (the sample average)

$$\tilde{k}(\mathbf{x}, \mathbf{z}) = k(\mathbf{x}, \mathbf{z}) - \frac{1}{n} \sum_{i=1}^n k(\mathbf{x}, \mathbf{x}_i) - \frac{1}{n} \sum_{i=1}^n k(\mathbf{z}, \mathbf{x}_i) + \frac{1}{n^2} \sum_{i,j=1}^n k(\mathbf{x}_i, \mathbf{x}_j)$$

Distances The distance between the RKHS maps of two samples can be naturally expressed as:

$$\begin{aligned} d(\phi(\mathbf{x}), \phi(\mathbf{z})) &= \|\phi(\mathbf{x}) - \phi(\mathbf{z})\| = \sqrt{\langle \phi(\mathbf{x}) - \phi(\mathbf{z}), \phi(\mathbf{x}) - \phi(\mathbf{z}) \rangle} \\ &= \sqrt{\langle \phi(\mathbf{x}), \phi(\mathbf{x}) \rangle + \langle \phi(\mathbf{z}), \phi(\mathbf{z}) \rangle - 2\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle} \\ &= \sqrt{k(\mathbf{x}, \mathbf{x}) + k(\mathbf{z}, \mathbf{z}) - 2k(\mathbf{x}, \mathbf{z})}. \end{aligned} \quad (4.20)$$

Therefore, an algorithm relying on distances, can be run in \mathcal{H} by adopting such distance (e.g. the kernel k NN).

Normalization An additional operation that can be carried out in the RKHS is the evaluation of the similarity between two normalized samples \mathbf{x}, \mathbf{z} in \mathcal{H} , by scaling their feature vectors to the unit norm (mapping them to the unit sphere in the RKHS):

$$\tilde{k}(\mathbf{x}, \mathbf{z}) = \left\langle \frac{\phi(\mathbf{x})}{\|\phi(\mathbf{x})\|}, \frac{\phi(\mathbf{z})}{\|\phi(\mathbf{z})\|} \right\rangle = \frac{k(\mathbf{x}, \mathbf{z})}{\sqrt{k(\mathbf{x}, \mathbf{x})k(\mathbf{z}, \mathbf{z})}}. \quad (4.21)$$

Note that $\tilde{k}(\mathbf{x}, \mathbf{z}) = \cos^{\mathcal{H}}(\theta)$, where $\theta = (\angle(\phi(\mathbf{x}), \phi(\mathbf{z})))$, from the definition of the inner product.

4. Learning with kernels

4.2.3 The kernel functions

As it was discussed, the kernel function allows to use a linear algorithm in some higher dimensional space to perform nonlinear analyses, relying on the assumptions stated by the Cover's theorem [Cover, 1965]. This mapping is implicitly performed by any kernel function adopted, provided its validity by the Mercer's theorem [Mercer, 1909], ensuring that the RKHS of a kernel is unique. It states that the symmetric kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ can be expressed as the $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$ only if the kernel function is positive definite, i.e. $\int_{\mathcal{X} \times \mathcal{X}} k(\mathbf{x}_i, \mathbf{x}_j) f(\mathbf{x}_i) f(\mathbf{x}_j) d\mathbf{x}_i d\mathbf{x}_j \geq 0$.

If one adopts the standard dot product as the kernel function, no higher dimensional mapping is performed and the original linear algorithm is recovered in its dual form, providing up to a scaling factor the same solution as for the algorithm in the primal. The use of the standard inner product results in the linear kernel:

$$k(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle = \mathbf{x}'\mathbf{z}. \quad (4.22)$$

It measures the similarity as the collinearity of the two vectors, as illustrated in Definition 4, point 5. It ranges from 0 for orthogonal vectors (maximally dissimilar samples) to a value equal to the product of the two vector norms, since $\langle \mathbf{a}, \mathbf{b} \rangle = \cos \theta \|\mathbf{a}\|_2 \|\mathbf{b}\|_2$ with $\theta = 0$. This is kernel function perform a mapping defined as $\phi(\mathbf{x}) = \mathbf{x}$.

Another well known kernel is the polynomial one:

$$k(\mathbf{x}, \mathbf{z}) = (\langle \mathbf{x}, \mathbf{z} \rangle + b)^p, \quad (4.23)$$

with the hyperparameters $p > 1$ and b to be tuned. When $b = 0$ it is known as homogeneous polynomial, otherwise as a non-homogeneous. Shawe-Taylor and Cristianini [2004] illustrates that these kernels return the value of a dot product in a space with dimension $\binom{d+p-1}{p}$ for the homogeneous polynomial kernel and $\binom{d+p}{p}$ for a non-homogeneous one. Their explicit feature maps can be obtained with a similar reasoning.

A classical example is as follows. Consider the dataset composed of three classes in \mathbb{R}^2 , with a class-conditional distribution as concentric circles shown in Figure 4.1(a). In this case, the expansion of the form $\phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$ maps the bi-variate samples into a three-dimensional space. It follows that:

$$\begin{aligned} \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle &= \langle (x_1^2 \ x_2^2 \ \sqrt{2}x_1x_2), (z_1^2 \ z_2^2 \ \sqrt{2}z_1z_2) \rangle \\ &= x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 \\ &= (\mathbf{x}'\mathbf{z})^2 = k(\mathbf{x}, \mathbf{z}). \end{aligned} \quad (4.24)$$

For this homogeneous kernel of degree 2 its feature map is of dimensionality 3, as it is verified by $\binom{2+2-1}{2} = \binom{3}{2} = 3$. Computing the mapping explicitly for all the samples illustrates that the problem in in Figure 4.1(a) can be solved linearly, as showed in Figure 4.1(b).

4.2 Kernel methods: theory and regularization

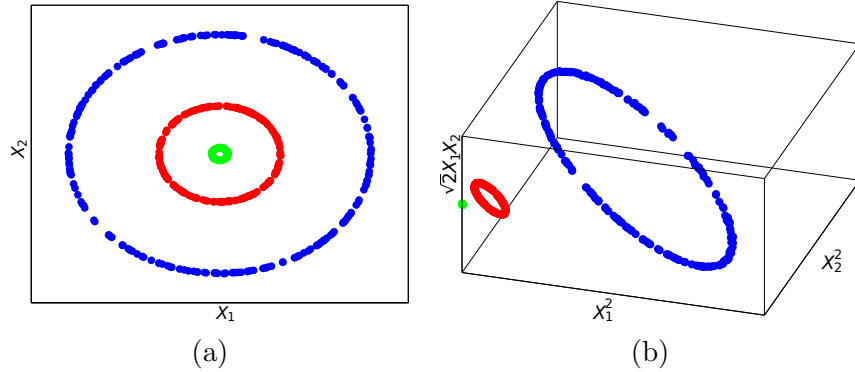


Figure 4.1: Polynomial kernel map - Feature map $\phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$ of the 2D coordinates of the 3 concentric circles (a) not separable using linear functions, to a 3D space (b), in which classes are linearly solvable by a simple planar (linear) decision.

Still, the most known and used kernel function in machine learning is the Gaussian radial basis function (RBF), and it is expressed as:

$$k(\mathbf{x}, \mathbf{z}) = \exp\left(\frac{-\|\mathbf{x} - \mathbf{z}\|^2}{2\sigma^2}\right) \quad (4.25)$$

where $\sigma > 0$ is a hyperparameter to be tuned. Its values range from 0 to 1, since two very dissimilar samples having large $\frac{-\|\mathbf{x} - \mathbf{z}\|^2}{2\sigma^2}$ will result in a value of $k(\mathbf{x}, \mathbf{z})$ close to 0, while at the other limit case where $\mathbf{x} = \mathbf{z}$, the $\frac{-\|\mathbf{x} - \mathbf{z}\|^2}{2\sigma^2}$ makes $\exp(0) = 1$. The in-between situations provide a measure of pairwise similarity. The scaling factor σ is very important and it may be interpreted as a control of the scaling of the Euclidean distance of the numerator [Bavaud, 2011], controlling the degree of nonlinearity of the kernel map. The use of this kernel function is well documented and motivated in the literature and it is supported by good performances and successful applications [Kanevski et al., 2009]. Additionally, Bach and Jordan [2002b] reported that when using this kernel, the projected data in the RKHS are likely to be normally distributed. In light of this observation, this kernel is often preferred by the consistency of the hypotheses made for the primal algorithm, even if, by definition, kernel-methods do not assume any prior distribution. A multi-modal distribution in the input space is consequently represented by a uni-modal distribution in the RKHS, provided that an adequate σ , as illustrated in Appendix A of [Cremersa et al., 2003]. This means that when the input space distribution is not Gaussian and possibly nonlinearly shaped, the Gaussianity holds for mapped data. Obviously, Gaussian data in the input space is still mapped into Gaussian distributions. This may ease the interpretation of many kernel methods when using the Gaussian RBF function. In this Thesis, the kernel Fisher's discriminant (Chapter 6) and the kernel k -means (Chapter 7) will both benefit from these observations.

Regarding the kernel hyperparameter, we may observe that small values of σ may lead to overfitting situation. In the extreme case, the sample is only similar to itself and accommodates any solution during the training step of a model. In this case, the kernel

4. Learning with kernels

matrix \mathbf{K} tends to the identity matrix. On the contrary, very large values of σ may provide underfitted situation since the models using such parameter tend towards a linear solution, by making the kernel close to a constant function. In this case the geometry of the input space is preserved and the map into a Gaussian RKHS is not verified. Note that the distance in the numerator of the kernel in Equation (4.25) can be itself expressed as a distance in the RKHS as in Equation (4.20), thus allowing extensions to the use of non-Euclidean distances and on non-vectorial data. As illustrated in [Francois et al., 2005], the original Gaussian RBF may be suboptimal in a very high dimensional input space, since the range of similarities (the histogram of the kernel values), as d grows, may tend to saturate around the mean. This is due to inflation of the Euclidean distance at the denominator, as the dimension of the sample vectors grow. As they show, this corresponds to hardly distinguishable small and large distances in the RKHS, that is, the first and last quantiles of the distribution of the similarity values tend to become empty. The authors propose to use a p RBF kernel function: it corresponds to a standard Gaussian RBF, but with the distance at the numerator and the σ at the denominator both elevated to the p th power. This ensure the locality of the kernel also for large d .

For the Gaussian kernel function, it has been proven that the induced dimensionality of such kernel is infinite. A way to see this important observation comes from the analysis in terms of Taylor's series expansion. It can be demonstrated that the Gaussian RBF kernel is a homogeneous polynomial kernel of infinite degree, since $\exp(\langle \mathbf{x}, \mathbf{z} \rangle) = \sum_{i=0}^{\infty} \frac{1}{i!} \langle \mathbf{x}, \mathbf{z} \rangle^i$. Details can be found in [Belanche, 2013; Steinwart et al., 2006].

Many other valid kernel functions exist. One can find a comprehensive review and explanation in [Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004].

4.2.4 Ad hoc-kernel functions and closure properties

An interesting property of kernel methods is that, by respecting some base principles making the Mercer's conditions hold, one can create his own kernel function adapted to specific problems. Some prior information on the data structure and their similarity may be available, and it can be introduced in the computation of the kernel to obtain a better representation. One can develop its own kernel functions by observing the following rules, also known as the closure properties [Camps-Valls and Bruzzone, 2009; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004].

Let $k_1(\mathbf{x}, \mathbf{z})$ and $k_2(\mathbf{x}, \mathbf{z})$ be two valid kernel functions, defined over $\mathcal{X} \times \mathcal{X} \subseteq \mathbb{R}^d$, scalars $v, p > 0$, a real-valued function $g(\cdot)$, a symmetric positive definite matrix \mathbf{A} and a valid distance metric $d(\mathbf{x}, \mathbf{z})$. Then, the following properties lead to valid kernels $k(\mathbf{x}, \mathbf{z})$:

$$\begin{array}{ll}
 k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z}) + k_2(\mathbf{x}, \mathbf{z}) & k(\mathbf{x}, \mathbf{z}) = \mathbf{x}' \mathbf{A} \mathbf{z} \\
 k(\mathbf{x}, \mathbf{z}) = vk_1(\mathbf{x}, \mathbf{z}) & k(\mathbf{x}, \mathbf{z}) = (k_1(\mathbf{x}, \mathbf{z}) + v)^p \\
 k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z})k_2(\mathbf{x}, \mathbf{z}) & k(\mathbf{x}, \mathbf{z}) = \exp(-d(\mathbf{x}, \mathbf{z})) \\
 k(\mathbf{x}, \mathbf{z}) = g(\mathbf{x})g(\mathbf{z}) & k(\mathbf{x}, \mathbf{z}) = \exp(k_1(\mathbf{x}, \mathbf{z})).
 \end{array}$$

4.2 Kernel methods: theory and regularization

Other important kernel properties follow [Gómez-Chova et al., 2010]:

$$k(\mathbf{x}, \mathbf{z}) = \sum_{r=1}^R d_r k_r(\mathbf{x}, \mathbf{z}), \quad \sum_{r=1}^R d_r = 1, \quad d_r \geq 0 \quad (4.26)$$

which defines a convex combination of R base kernels. Approaches aiming at optimize this convex combination are known as multiple kernel learning [Bach et al., 2004; Raketomamonjy et al., 2008; Tuia et al., 2010a].

Also, the deformation of a kernel with another positive matrix is a valid kernel. It is particularly useful in semi-supervised learning, in which marginal information coming from unlabelled samples may be integrated into the kernel. A classical approach is to deform the original kernel by the geometrical information carried by the empirical graph Laplacian \mathcal{M} as $\tilde{\mathbf{K}} = \mathbf{K} + \delta \mathbf{K} \mathcal{M} \mathbf{K}$. In Chapter 8 this property is exploited for regularization.

Finally, by exploiting again the above properties, one can construct kernels between distributions: for $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ as $k(p_1(\mathbf{x}), p_2(\mathbf{x})) = \langle p_1(\mathbf{x}), p_2(\mathbf{x}) \rangle = \int_{\mathcal{X}} p_1(\mathbf{x}) p_2(\mathbf{x}) d\mathbf{x}$.

It is possible to take advantage of these properties to define a very specific representation of the problem at hand, by combining and weighting different forms and sources of information [Camps-Valls et al., 2006, 2008]. This can be done for a variety of data types, e.g. using string kernels for gene prediction [Leslie and Kuang, 2004], kernels built on trees [Shin et al., 2008], kernels on graphs [Vishwanathan et al., 2010], kernel over sets for structured predictions [Ricci et al., 2008]. For a review, see [Belanche, 2013].

4.2.5 The Gram matrix

Accordingly to the kernel methods literature, the symmetric matrix $\mathbf{K}_{ij} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ is often referred to as a Gram matrix. Any Gram matrix is defined as the positive definite matrix in which entries ij correspond to the inner products of vectors \mathbf{x}_i and \mathbf{x}_j . Obviously, this holds also for the projection of $\mathbf{x}_i \mapsto \phi(\mathbf{x}_i)$ and thus \mathbf{K}_{ij} is a Gram matrix.

The Mercer's theorem allows us to interpret the kernel Gram matrix as a covariance operator (see Chapter 3 in Shawe-Taylor and Cristianini [2004]). By applying the spectral decomposition of the sample covariance matrix $\mathbf{\Sigma} = (n-1)^{-1} \mathbf{X}' \mathbf{X}$ and of the centred Gram matrix $\mathbf{G} = \mathbf{X} \mathbf{X}'$, one can verify, by requiring eigenvectors normalized to unit norm, that the eigenvalues are the same, and the eigenvectors are proportional for both representations. This can be easily verified by observing the relationships between primal and dual PCA modes (see Chapter 8).

4.2.6 On the choice of the kernel function and its parameters

By either adopting a pre-existent kernel function or by developing a custom similarity measure, one imposes a form on the data representation. Even before selecting the correct hyperparameters of the kernel, one should select an appropriate function. In particular, as pointed out in Section 4.2.1, by selecting a kernel function one defines a hypothesis space in which the learning task will be accomplished. This is a crucial task, since the form

4. Learning with kernels

of the kernel function effectively corresponds to the prior information available about the problem [Schölkopf and Smola, 2002]. By using a Gaussian kernel, its implicit regularization penalizes derivatives of all the orders, favouring low energy of high frequencies of the Fourier spectrum of the function, by selecting a polynomial kernel of degree p one looks for the p th order relationship of the data or when using a linear kernel, the flatness of the final function is favoured [Evgeniou et al., 1999; Schölkopf and Smola, 2002].

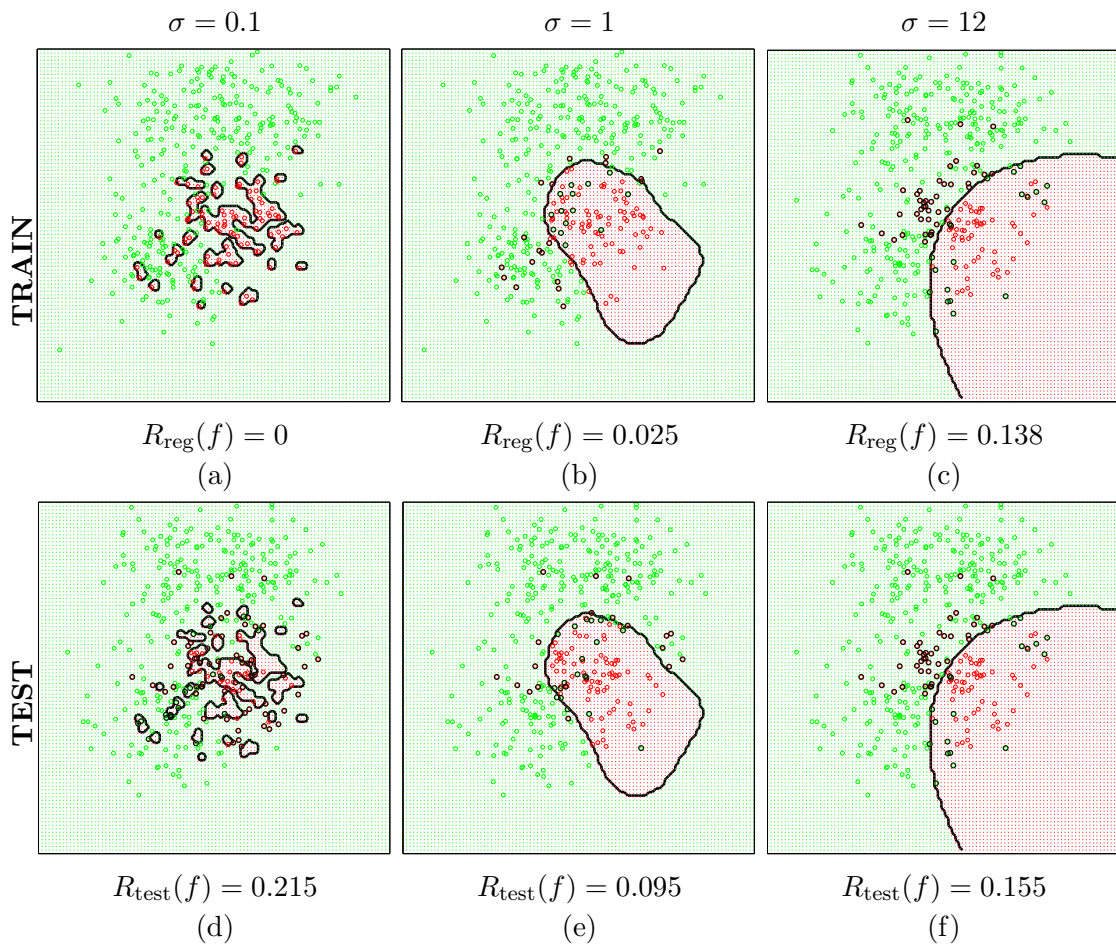
Once the appropriate kernel function is chosen, its hyperparameters have to be fitted during the model selection step, as illustrated in Section 3.3. This restricts the hypothesis space, while the model optimization provides the solution. Finally, note that many hyperparameters may have to be optimized when adopting combinations of kernels. In this case, fast model selection procedures are proposed, as in [Chapelle and Rakotomamonjy, 2008; Chapelle et al., 2002; Rakotomamonjy et al., 2008].

An example of the influence of the Gaussian RBF σ parameter is illustrated in Figure 4.2, by using a SVM classifier (see Section 6) with a C parameter equal to 10. Similarly to what is observed in the example of Figure 3.3 regarding under- and over-fitting, a correct hyperparameter tuning gives the minimum generalization error.

4.3 Some considerations

As we illustrated, kernel methods are a particular family of machine learning algorithms with a robust and well founded theory. The use of kernels provide nonlinear solutions to many learning problems, from classification to clustering, from feature extraction to regression, with formulations varying smoothly one from the other sharing the same underlying form, the representer theorem (Section 4.2.1). Furthermore, they allow the inclusion of precise and ad-hoc information about the problem at hand, by either designing kernel function encoding the structure of the particular data, by satisfying Mercer's conditions (Section 4.2.3), or by blindly combining different kernels and letting the algorithm to find the optimal combination as in a multiple kernel learning approach, making use of the closure properties (Section 4.2.4).

In the remote sensing image processing literature kernels are gaining growing interest in many application fields, and they became an active research area for both theoretical and fundamental developments and scientific applications [Camps-Valls and Bruzzone, 2009]. As the title of this Thesis suggests, kernel methods are very useful also for change detection analyses. In the rest of the Thesis, after a brief literature review to put the reader aware of the main strategies for change detection, the solutions proposed within the context of the Thesis are presented.



LEGEND: — Decision function, RED: class 1, GREEN: class 2, o: errors

Figure 4.2: Sigma parameter of Gaussian RBF kernel - Influence of the σ parameter using a SVM classifier. In (a)-(c) the training error for an increasing σ parameters, while (d)-(f) illustrates the test error. Note the decreasing model complexity for larger σ parameters.

4. Learning with kernels

Part III

Kernel-based methods for change detection

Chapter 5

A review of the approaches for remote sensing multi-temporal image analysis

This Chapter introduces and discusses the state-of-the-art for remote sensing multi-temporal data analysis, with particular focus on statistical approaches. Section 5.1 gives a general introduction to the task. Then, the literature concerning supervised, unsupervised and feature extraction-based approaches is reviewed in Section 5.2, Section 5.3 and Section 5.4, respectively. Section 5.5 provides some concluding remarks.

5.1 Learning from pixels

As introduced in Chapter 2, each pixel comes with a series of measurements directly related to the ground cover type, which constitute its spectral signature. Depending on the nature of the processing task, we may want to estimate the probability of appearance of given pixel values, to approximate a function discriminating classes of samples, or to build an inversion model predicting biophysical parameters. In any way, these values are given to a processing algorithm which models the relationships of interest. In multi-temporal image analysis, the input of a learning system vary from single-time imagery to bi- and multi-temporal (time series) data compositions.

It is reasonable to qualify the data preparation task, defining the inputs of the learning system, as a preprocessing step. For standard image classification tasks, the inputs of the learning algorithm (training, testing and, if any, validation) are the pixels coming from an image $X^{t_0} \in \mathbb{R}^d$, where t_0 stands for a general time 0, being this problem independent from the temporal component. In the case of bi-temporal analyses, that is the setting of standard change detection, the two images and their pixels may be combined in different ways. For many automatic algorithms the input is the difference image \mathbf{D} . As mentioned in Section 2.3.1, this strategy does not allow a multi-class categorization of

5. State-of-the-art in change detection

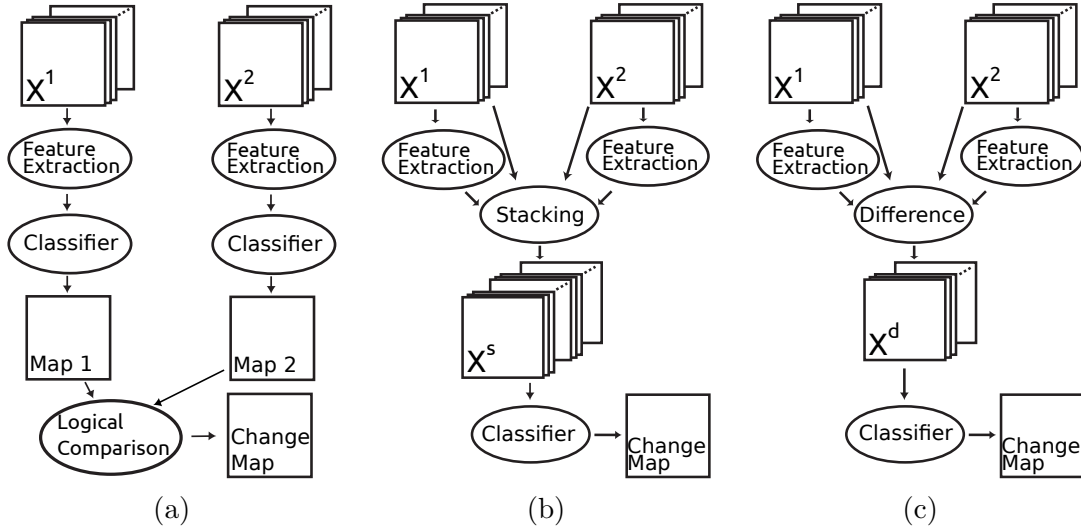


Figure 5.1: PCC, DMC and DIA schemes - General flowchart for: (a) Post-classification comparison, (b) direct multi-date classification and (c) difference image analysis. Note that the feature extraction step is not mandatory.

unchanged samples, since $\mathbf{D} \approx \mathbf{0}$ for the stable transitions indistinctly. If an exhaustive map summarizing both changed and stable classes from a bi-temporal set is required, one might want to adopt their stack as input, as $X^s = \bigcup_{i=1}^T X^{t_i} = X^{t_1} \cup X^{t_2}$. In this case, all the radiometric information is preserved and all the classes can be modelled, provided that an exhaustive training set exists. For general multi-temporal images and time series analysis involving the modelling of T distinct images, the stack approach is still valid since the concatenation operation $X^s = \bigcup_{i=1}^T X^{t_i}$ is not affected by the number or dimensionality of the single images, contrarily to image differencing. However, if the time series is large, e.g. when analysing temporal trajectories, one may fall in issues related to the high dimensionality. For these reasons, one may prefer to model independently the pixels as multivariate time series and adopting some generalization schemes on the rest of the dataset [Lhermitte et al., 2011]. Another approach is to extract some lower dimensional representation such as the the NDVI from the images corresponding to each time point and to study the derived temporal sequence [Verbesselt et al., 2010]. Finally, additional approaches are to study some time series-derived indicators in given temporal intervals, such as trends or cycles [Mello et al., 2013; Wessels et al., 2012], or to embed the pixel time series into a lower dimensional space [Small, 2012].

Note that the radiometric normalization techniques defined in Section 2.3.3 are still required, in particular when dealing with the difference image. Similarly, when analysing image time series, global normalization schemes should be applied carefully since the relative change from one date to another might smaller than the acquisition dependent deformations. In the next Section, a review of supervised approaches to change detection is presented.

5.2 Supervised change detection

Supervised change detection methods can be grouped in three categories: post-classification comparison (PCC), direct multirate classification (DMC) and supervised difference image analysis (DIA) [Coppin et al., 2004; Singh, 1989].

5.2.1 Post classification comparison

Methods from this family are the most simple and used approaches to operational change detection for monitoring purposes. To this end, PCC compares classification maps obtained from each image separately. The resulting output is a summary of a series of maps, encoded by an appropriate legend. For instance, Lee [2008] coded the classification maps obtained from a series of Landsat images, all acquired in November to minimize the phenological differences, into a single map describing the forest evolution over 30 years. The uni-temporal maps of trees categories were obtained by NDVI minimum-error thresholding. Serra et al. [2003] studied a maximum likelihood classification (MLC) based approach to PCC with particular attention to the error generation and propagation phenomena in the PCC process. Since images are classified independently using independent training sets, the worst possible error may result from the multiplication of the per-class error rates (errors are independent). Additionally, this is studied under a multi-sensor perspective, which is natural for PCC schemes. Similarly, Alphan et al. [2009] and Fichera et al. [2012] applied the MLC for a PCC analysis for environmental monitoring. In [Anhed et al., 2008] the PCC is used into a geographical information system (GIS). First, images are segmented and labelled using GIS layers, then the PCC summarizes land cover evolutions. Chen et al. [2011] proposed a probabilistic approach to PCC by analysing the differences in posterior probabilities of the pixels given the class, obtained again by the MLC.

As one can see, PCC-based approaches offer a straightforward solution to change detection, also in multi-modal (multi-sensor) and time series scenarios. Furthermore, the ground truth is collected independently for each image without the need of labelling all the observed transitions. Another interesting property of PCC is the possibility to avoid atmospheric compensations or relative radiometric normalizations. However, depending on the classifier adopted, these specific preprocessing methods may still be needed: statistical models do not need them, while classifications based on the comparison with reflectance databases obviously need calibrated values [Song et al., 2001]. In both cases one has to pay attention to the ability of each classifier to solve accurately the single classification tasks, because of the error propagation thorough the process of map comparison. For this reason, a final manual selection of plausible and realistic changes has to be carried out carefully, resulting in a strong limitation of this family of methods. For instance, a PCC involving the analysis of two maps classified into k and q classes respectively may generate in a worst case situation $k \times q$ transitions, but only a small part of them are observable or even realistic.

5. State-of-the-art in change detection

5.2.2 Direct multi-date classification

In DMC change detection, one performs a direct categorization exploiting a single model, classifying the stack of original or transformed images. The change map is obtained directly and the classes modelled are only those present in the training set.

This approach to change detection has been considered since early developments of satellite-based precision monitoring, such as in [Salem et al., 1995], in which the authors monitored citrus agriculture fields by the use of the MLC on stacked multi-temporal vegetation indexes. In the following years, neural networks classifiers were also considered [Long Dal and Khorram, 1999; Sucharita and Woodcock, 1996]. In the former, DMC aimed at analysing the development of the city of Wilmington (USA). The comparison of the neural approach versus the MLC demonstrated a sharp increase of almost 10% in accuracy, thus clearly pointing out the advantages of nonlinear classification also in multi-temporal applications. In the latter, a neural network of similar architecture was employed to model forest changes due to conifer mortality caused by a severe drought in the Lake Tahoe (USA) basin. Another approach based on the direct multi-date analysis of image time series is presented in [Elmore et al., 2000]. Authors used spectral unmixing to analyze the abundance of vegetation in the time series images. The relative maps of vegetation abundance at each time point composed another time series on which changes have been modelled. In [Yuan et al., 2005], an hybrid DMC and PCC scheme was adopted to monitor the cities of Minneapolis and Saint Paul, USA. Four couples of images were independently classified using DMC schemes, and their maps were then compared with PCC. In Nemmour and Chibani [2006] supervised change detection is implemented using a cascade of binary SVM to solve multi-class problems. The comparison with a neural network classifier proved that SVM were less prone to overfit the data and training issues related to non-convex error functions were avoided, providing a better generalization. The same authors extended their analyses in [Nemmour and Chibani, 2010] by testing different multi-class SVM architectures.

DMC has been adopted less frequently than PCC, in particular since data dimensionality doubles for bi-temporal applications, or grows proportionally to the number of dimensions of the considered dates, increasing the requirements in the modelling process. The most of the aforementioned literature adopts approaches reducing original data into low dimensional variables, such as spectral indexes. This greatly facilitates the supervised analysis of time series and image stacks, in particular with respect to the training set size requirements thanks to the lower dimensional space. However, a large loss of information may harm the process since only few channels are usually considered to compute such variables. Thus, a strong problem depended component is always present in the choice of the representation and modelling of the data.

The benefits of kernel methods and in particular classifiers such as SVM were only poorly analyzed. In the perspective of the most recent high resolution imagery, DMC approaches may be difficult to implement, since VHR data is prone to possess low between-class separability due to the large amount of spatial detail, resulting in large within-class

5.3 Automatic and unsupervised change detection

variances. In Chapter 6, an approach to supervised change detection in VHR images exploiting spatial context features to cope with this problem will be presented.

5.2.3 Supervised difference image analysis

In the DIA setting, one aims at modelling only the changes enhanced by the difference image. This approach is limited to the analysis of two images at a time, but it has the advantage of dealing with datasets of dimensionality equal to the one of the original single-time images. In the literature of supervised change detection, supervised DIA approaches are only marginally studied. They can be seen as a particular case of the DMC, in which the particular temporal composition of the images emphasizes only changes (provided an adequate preprocessing), whilst minimizing the unchanged areas signals.

Coppin and Bauer [1994] presented an approach based on the supervised thresholding of the normalized difference of vegetation indexes, for forest canopy monitoring purposes. Prior to change detection, authors performed feature selection among a variety of spectral indexes by quantifying their correlation with ground observations. Dale et al. [1996] applied a similar approach for the monitoring of wetlands in Australia. Guerra et al. [1998] studied the changes in a vegetated environment by proposing a multi-temporal normalized vegetation index, in which changes were successfully discriminated. Finally, Camps-Valls et al. [2008] proposed a complete framework for multi-temporal classification and change detection based on composite kernels [Camps-Valls et al., 2006]. Although the approach can be adapted to a variety of learning problems, experiments in the study exploit supervised classifiers. Authors formulated the difference, the ratio and the stack operators for change detection directly into the RKHS, thus generalizing to abstract vector spaces the notion of multi-temporal image compositions. By the adoption of nonlinear kernels, the change information can be modelled nonlinearly through ad-hoc kernel functions representing the nature of the problem. Multi-modal change detection can also be performed explicitly, by combining in a weighted fashion information from different sensors. Finally, Du et al. [2012] presented an interesting approach relying on the fusion of multiple difference images to perform multi-modal change detection. The difference images are computed via various spectral distance indexes (e.g. absolute distance, ratio, Chi-squared distance, etc.), and the subsequent fusion improved robustness to noise and to outliers.

5.3 Automatic and unsupervised change detection

The approaches of this family exploit unsupervised and automatic algorithms to detect changes. Due to the appealing setting of complete automation, and since obtaining an appropriate labelling of changes is often infeasible, unsupervised approaches are probably the most prominent part of the change detection literature. They may be divided into two principal categories: the first relies on clustering and unsupervised classification methods, while the second reformulates the problem as a novelty detection process. Note that both categories may comprehend the application of a series of supervised algorithms initialized

5. State-of-the-art in change detection

by exploiting a standard change detection approach to provide a suboptimal training set, from which start the learning step. The obtained pseudo-training samples are then employed by a more robust supervised algorithm to refine the partitioning of the multi-temporal data.

In general, the most of these approaches rely on the difference image, and consequently the final output is a binary change map indicating the spatial occurrence changed pixels.

5.3.1 Clustering and unsupervised classification

In his seminal work, Fung [1990] proposed different approaches to change detection. One of them relied on the unsupervised thresholding of the difference image using a distribution-based fitting. Then, in [Bruzzone and Fernandèz-Prieto, 2000] and [Melgani et al., 2002] different approaches to the automatic thresholding of the magnitude of the difference image were reviewed. Hazel [2001] studied an object-based approach to detect changes in two coregistered images, further extended in [Bovolo, 2009] by proposing a spatially aware CVA system. These schemes first segment independently the bi-temporal images and then performs change detection. Similarly, but with a segmentation step occurring after the computation of the difference image magnitude, was presented in Desclée et al. [2006]. After segmentation, the mean values of the regions were clustered using a variant of the k -means algorithm. Dalla Mura et al. [2008] proposed a similar approach based on the filtering of the difference image with morphological reconstruction operators prior to a CVA. Melgani and Bazi [2006] presented an approach to change detection via the fusion of change maps obtained via independent CVA, relying on different thresholding methods. The fusion was based on Markov random fields, so that the spatial structure of the phenomena was considered while being robust to outliers. In [Im et al., 2008] an algorithm for change detection based on logical reasoning was proposed. In their model, the magnitude of the difference image and the local spatial autocorrelation were considered simultaneously to optimize a threshold for the CVA. Other context-driven approaches were presented in [Celik, 2009a,b]. In the former, the spatial component of the difference image magnitude was assimilated through a wavelet transform, while in the latter a parcel-based principal component transformation summarized simultaneously spatial and pixel intensity information. Once the improved difference images were obtained, k -means was adopted in both cases to compute the change maps. For both methods, a trade-off between spatial smoothing and detail preservation has to be manually tuned. An approach based on the automatic level set segmentation of the difference image was presented in [Bazi et al., 2010]. Finally, in [de Morsier et al., 2012] the unsupervised hierarchical support vector clustering was performed on the difference image. A criterion based on the between-cluster distance was proposed to group changed areas.

In [Ghosh et al., 2007], the change detection task was performed by initializing a Hopfield neural network via CVA, as presented in [Bruzzone and Fernandèz-Prieto, 2000]. The system modelled the spatial autocorrelation of the difference image, so that the context of each pixel improved the coherence of the final map. Bovolo et al. [2008] proposed a

5.3 Automatic and unsupervised change detection

nonlinear and automatic system relying on the semi-supervised transductive SVM. The method was again initialized on the outcome of a CVA, so that the first transduction model could exploit some pseudo-training samples for learning. Advantages over standard SVM with the same initialization were detailed. In [Huo et al., 2010], the method described above was applied to objects issued from multi-temporal segmentation.

In this Thesis, the Chapter 7 is deemed to present an automatic change detection method based on the clustering of the difference image computed in the RKHS. To enforce the stability of the algorithm, an initialization similar to the one adopted by the methods described above is employed.

5.3.2 Novelty detection

Novelty detection approaches to the analysis of changes have received the most of the attention for anomalous change detection in hyperspectral images. However, this problem setting is now becoming of broad interest also for researchers involved in the development of multi-spectral change detection system. These methods usually rely on the analysis of the unchanged information, while changes are considered as outliers, i.e. samples deviating significantly from the background. To this end, one-class classification methods are often employed [Muñoz-Mari et al., 2010], while statistical measures of deviation are employed in hyperspectral target and anomalous change detection. In this latter setting, the most of the approaches aim at detecting the apparition of new classes. Note that these methods may also be classified as feature extraction based, since they often rely on subspace projections to detect anomalies.

Authors in [Bovolo et al., 2010] adopted a supervised novelty detection method, the support vector domain description (SVDD), initialized using the CVA technique. Their approach modelled changes as targets, while unchanged pixels were detected as those lying outside the hypersphere containing changed samples. To tune hyperparameters, both classes issuing from the CVA initialization were employed. Pacifici and Del Frate [2010] proposed an approach relying on the unsupervised pulse coupled neural network. This algorithm flags spatial patches of the images in which a change occurred. In [de Morsier et al., 2013] a semi-supervised extension of the cost-sensitive version of the SVM, i.e. considering class-wise cost, is proposed for change detection purposes. The SVM needs only samples belonging to the background composed by unchanged samples, while the semi-supervised criteria automatically detects changed regions.

Regarding methods of anomalous change detection in bi-temporal hyperspectral data, one can find a review in [Theiler, 2008; Theiler and Perkins, 2007]. These approaches find changed samples by comparing a transformation of the spectral channels at both times. The type of transformation of the original images defines the sensitivity of the anomaly detection scheme to detect outliers. Depending on the measure from which estimate the transformation the original data, different methods are obtained: standard difference, chronochrome [Theiler, 2008], RX detector, of which a kernel extension exist [Kwon and Nasrabadi, 2005] and many others. However, as the name of this family of

5. State-of-the-art in change detection

approaches suggest, changes are detected only if they are anomalous, i.e. they are a set of rare occurrences corresponding to areas that changed into a new spectral class. In this sense, many methods of anomaly detection may be reformulated to perform change detection. For instance, Wu et al. [2013] adopted the orthogonal subspace projection to perform change detection, a method that has also been generalised to nonlinear and semi-supervised problems [Capobianco and Camps-Valls, 2009].

5.4 Feature extraction for multi-temporal applications

Linear and nonlinear feature extraction methods have been widely utilized in the context of dimensionality reduction for hyperspectral image classification. In this case, the aim is to reduce the many spectral bands to few variables maximizing some statistical measure. In change detection and multi-temporal image analysis literature, these approaches have been only partly explored. The goal of this family of models is twofold: (i) apply a transformation explicitly designed to enhance changed areas or (ii) apply the transformation to obtain a relative radiometric normalization to maximally align unchanged samples. After the transformation, the change map can be obtained by thresholding a single projected variable or by running a standard classification or clustering algorithm on the set of projected variables. Also, since the change information appears clearly in the features extracted, a RGB composition of the new multi-temporal variables may be sufficient for a visual discrimination. More insights of these algorithm will be provided in Chapter 8.

First feature extraction based approaches to multi-temporal image analysis were proposed by using PCA rotations in the work of Fung [1990]. By diagonalizing the covariance matrix of the stacked multi-temporal set, the author observed that the component of the transformation related to largest variance summarized the information about unchanged areas. Since changes are located in a different region of the spectral space and are usually uncorrelated from unchanged areas, these are found in the second largest component, orthogonal to the first one. In this sense, under low noise conditions and a linear relationship between unchanged samples, a simple threshold is able to provide a change map. A nonlinear extension of these assumptions was proposed in [Nielsen and Canty, 2008] by adopting the kernel version of the PCA. In this case, the extracted components have to be analyzed manually since the first and second directions of maximal variance may not correspond to unchanged and changed areas respectively. Nielsen [2002] proposed the analysis of an image time series by applying a relative radiometric normalization via multi-set canonical correlation analysis. This method is able to discover a liner map of the different datasets to a space in which their projections are maximally and mutually correlated. A RGB composition of the variables corresponding to the components of largest correlation provide a visual (but not discrete) indicator of changed areas. Nielsen [2007] presented an extension of this approach developed for bi-temporal change detection, the iteratively reweighted multi-variate alteration detection. The approach aims at iteratively enhance the separability between changed and unchanged pixels by matching the unchanged samples distribution. Then, a threshold based on the fit of the canonical variables distribution,

provides the change map. A review of the feature extraction based methods proposed by these authors are summarized in [Canty and Nielsen, 2012].

Zhong and Wang [2006] classified with a MLC the components derived from the independent component analysis of the stacked images. The more compact and improved representativeness of the new features allowed a better classification than by the original input stack. Deng et al. [2008] presented an approach relying on the supervised classification of the first components of the PCA transformation of the multi-temporal stack. In [Marchesi and Bruzzone, 2009] an approach to the analysis of different multi-temporal compositions (stacked and difference images) based on the independent component analysis and its kernel extension are reviewed. To obtain change maps, manually selected features corresponding to change directions are thresholded as in the CVA. Finally, in [Gómez-Chova et al., 2012, 2013] a nonlinear kernel-based feature extraction approach was explicitly designed to extract the change information from a pair of difference images, in which changes of interest are contained in the second image. Experiments on cloud detection demonstrate the high discriminative power of the change components.

Chapter 8 presents two approaches relying on feature extraction for change detection applications. Rather than exploiting labels of unchanged and changed regions, only few samples corresponding to unchanged areas are exploited. Then, the feature extraction method is aimed to maximally align the information carried by those pixels in order to obtain a more discriminative representation of changes.

5.5 Some considerations

The statistical change detection literature is rapidly evolving. It progresses proportionally to the advances in the pattern recognition and machine learning literature, trying to overcome fundamental limitations of current techniques to face the processing challenges generated by the most recent acquisition systems and the new applications issuing from the increasing needs of the users. As it clearly appears, it is hard to be exhaustive in reviewing the state-of-the-art in such an evolving research field. However, clear trends are remarked: the consideration of the spatial context, in particular for VHR images, and the adoption of nonlinear methods to process the data. In the late 1990 and the early 2000, the literature exploiting nonlinear models was mainly dedicated to neural networks, while in the last years kernel methods started to be considered also in this processing problem.

Currently, attention is being paid to the scarcity of labelled information, by either using unsupervised or semi-supervised models, or by enhancing the changed data. Moreover, great efforts are paid to the development of models allowing change detection between different sensors. Emerging trends in this field are certainly to be researched in merging the aforementioned considerations. In the next Chapters, the methods developed in this Thesis to tackle these issues are presented.

5. State-of-the-art in change detection

Chapter 6

Supervised change detection. Representativeness of the input space¹

This Chapter presents two change detection methods based on two supervised classifiers: kernel Fisher’s discriminant and support vector machines. After briefly introducing the tasks to be solved in Section 6.1, the adopted classifiers and the obtained results are presented in Section 6.2 and Section 6.3 respectively. General conclusions are drawn in Section 6.4.

6.1 Supervised change detection for monitoring

In this Chapter, two case studies involving supervised change detection are presented. In the former, an approach to flooded area extraction using regularized kernel Fisher’s discriminants (kFDA) is proposed, with emphasis on the comparison between uni- and multi-temporal approaches. In the latter, we study the contribution of contextual features into a SVM-based multi-class supervised change detection for VHR urban monitoring purposes.

In the flood mapping scenario we aim at delineating the zones that have been inundated by a river flood. To this end, two approaches are studied: uni- and multi-temporal image classification. Depending on the use of the derived map, both approaches are valid: in the former, the classification of the post-event image provides a cartography delineating the water extent at the time of the acquisition. The extracted map (regardless of the permanent standing water) is useful in particular ecological applications, e.g. if only a water mask is needed [Chormanski et al., 2011; Dey et al., 2009; Khan et al., 2011]. In general, uni-temporal approaches are usually preferred when time constrains the process

¹This Chapter is based on the following publications: [Volpi et al., 2013c] and [Volpi et al., 2013d]. See Section 1.3.1 for the details.

6. Supervised change detection

or the producer only dispose of limited data, memory and computational power [Ip et al., 2006]. The second mapping scenario is implemented as a change detection problem, in which only the non-permanent standing water is targeted as “flooded” (changed) areas. This solution makes sense when a precise cartography of the exceeding amount of water is needed [Hudson and Colditz, 2003; Sanders et al., 2005; Zwenzner and Voigt, 2009].

In the second case study, the aim is to extend the supervised DMC studied in the first monitoring task to account for the spatial context of each pixel in order to process VHR data. To this end, we studied local indicators of texture [Haralick et al., 1973]; and regional smoothing based on local extremes, the mathematical morphology [Soille, 2004; Soille and Pesaresi, 2002]. To assess the improved informativeness of the proposed input spaces, accounting for different multi-scale representations of the images, two schemes for the combination of the multi-temporal information are adopted: DMC and supervised DIA. Their role in the process of precise cartography of the changes in a urban scenario are deeply discussed, and parallels with standard spatio-spectral image classification are drawn [Benediktsson et al., 2005; Pacifici et al., 2009; Tuia et al., 2009, 2010b]. The role of these variables have been only poorly studied in multi-temporal applications and in particular for change detection analyses.

6.2 Supervised approaches for flooded area extraction

In this case study, the kFDA will be exploited to assess the flood mapping task by providing theoretical analysis of the classification setting and by examining the role of the permanent standing water. The discrimination problem can be efficiently solved by the kFDA, offering a regularized and nonlinear solution. This also provides low sensitivity against high dimensional datasets and robustness to over-fitting issues by controlling the complexity of the model [Bandos et al., 2009]. The use of the kFDA is further motivated by its simplicity, while keeping the advantages of kernel methods. However, the black box application of the kFDA does not allow a clear understanding of the flood cartography process. For this reason, the temporal component of input space, the role of the permanent standing water and the linear / nonlinear classification models are discussed for each setting implemented (uni- and multi-temporal classification). To this end, the “permanent standing water” class as been recoded to “flooded” in the uni-temporal case, while it has been assigned to “not flooded” in the DMC.

6.2.1 The regularized kernel Fisher’s discriminant classifier

The Fisher’s discriminant analysis (FDA) [Fisher, 1936] can be used either as a linear supervised dimensionality reduction or a linear classification technique. In its standard binary classification formulation, it aims at finding a unidimensional projection of the training pixels $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $y_i \in \{-1, +1\}$, that maximally separates the two class means. Once this direction is found, a threshold suffices to classify the projected data. The standard linear decision function can be expressed in the form $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$, \mathbf{w} being

6.2 Supervised flooded area extraction

the projection vector to the discriminant subspace (a weight vector of an hyperplane, as in SVM) and b a bias term. It is possible to define the mean of class c , composed by n_c samples in the subset of the training set $X_c = \{(\mathbf{x}, y) \in (X, Y) | y = c\}$, as $\boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{\mathbf{x}_i \in X_c} \mathbf{x}_i$. The projection onto the discriminant direction is consequently $\mathbf{m}_c = \frac{1}{n_c} \sum_{\mathbf{x}_i \in X_c} \mathbf{w}'\mathbf{x}_i = \mathbf{w}'\boldsymbol{\mu}_c$. Un-normalized variance (scatter) of the projected data can be consequently defined as $s_c^2 = \sum_{\mathbf{x}_i \in X_c} (\mathbf{w}'\mathbf{x}_i - \mathbf{m}_c)^2$. After the definition of these class-wise measures, the objective function of the Fisher's discriminant can be formulated as the maximization of:

$$\arg \max_{\mathbf{w}} \frac{(\mathbf{m}_1 - \mathbf{m}_2)^2}{(s_1^2 + s_2^2)} = \frac{\mathbf{w}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\mathbf{w}}{\mathbf{w}'(\sum_c \sum_{\mathbf{x}_i \in X_c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)')\mathbf{w}}. \quad (6.1)$$

The optimal separation is given by the direction that maximizes the distance between the means but also minimizes the scatter around them, that is, an optimization of the separation / overlap ratio, as depicted in Figure 6.1. The solution \mathbf{w} corresponds to the direction in which the between-class variance \mathbf{S}_b is maximized and the total within-class variance \mathbf{S}_w is minimized. It corresponds to the following Rayleigh quotient [Mika et al., 1999; Shawe-Taylor and Cristianini, 2004]:

$$\arg \max_{\mathbf{w}} \frac{\mathbf{w}'\mathbf{S}_b\mathbf{w}}{\mathbf{w}'\mathbf{S}_w\mathbf{w}}, \quad (6.2)$$

where the between - and within-scatter matrices are defined respectively as:

$$\mathbf{S}_b = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)' \quad (6.3)$$

$$\mathbf{S}_w = \sum_c \sum_{\mathbf{x}_i \in X_c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)'. \quad (6.4)$$

One can observe that the norm of \mathbf{w} at the denominator $\mathbf{w}'\mathbf{S}_w\mathbf{w}$ is not important to find the direction of the discriminant subspace, since it always points in the direction of $(\mathbf{m}_1 - \mathbf{m}_2)$ [Mika, 2002]. Therefore, one can set $\mathbf{w}'\mathbf{S}_w\mathbf{w} = 1$ without loss of generality. The problem may now be reformulated as a constrained optimization:

$$\arg \max_{\mathbf{w}} \mathbf{w}'\mathbf{S}_b\mathbf{w} \quad (6.5)$$

$$\text{s.t. } \mathbf{w}'\mathbf{S}_w\mathbf{w} = 1 \quad (6.6)$$

The solution of the above optimization may be found by its Lagrangian:

$$L(\mathbf{w}, \lambda) = \mathbf{w}'\mathbf{S}_b\mathbf{w} - \lambda(\mathbf{w}'\mathbf{S}_w\mathbf{w} - 1). \quad (6.7)$$

By equating to 0 the partial derivative of the function with respect to the parameters, the optimality conditions give:

$$\frac{\partial L}{\partial \mathbf{w}} = 2\mathbf{S}_b\mathbf{w} - 2\lambda\mathbf{S}_w\mathbf{w} = 0, \quad (6.8)$$

and it can be solved, for instance, by the generalised eigenvalue problem (note that both \mathbf{S}_b and \mathbf{S}_w are symmetric and positive definite):

$$\mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w}. \quad (6.9)$$

6. Supervised change detection

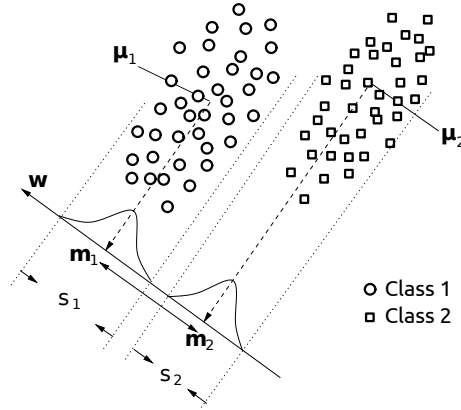


Figure 6.1: FDA graphical interpretation - An example of classification using FDA. The different statistical measures are used to find the weights \mathbf{w} that maximally separate the two groups in the projected space.

Here, λ are the eigenvalues and \mathbf{w} the eigenvectors of the system. The solution optimizing the projections defined in Equations (6.5) and (6.7) are given by the projections corresponding to the largest eigenvalues in λ . As illustrated above, Fisher's discriminant are interesting since they provide a solution in a closed form, with a global minimum. In addition, it has been demonstrated that for normally distributed classes with equal covariances, Fisher's discriminant corresponds to the Bayes classifier by setting a threshold corresponding to $b = \frac{1}{2}(\mathbf{m}_1 - \mathbf{m}_2)'(\mathbf{m}_1 - \mathbf{m}_2)$, corresponding to $p(+1|\mathbf{x}_i) > 0.5$.

In the linear case, the final class assignment is given by the sign of $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$, that is, indicates to which projected mean the sample is closest. However, in this form, it still limited by the linearity of the projection. Moreover, as clearly indicated by the use of the scatter matrices, multi-modal or strongly skewed and asymmetric distributions will affect the discriminant ability of \mathbf{w} . In order to overcome these problems and to take advantage of the flexibility and nonlinearity offered by kernels within the Fisher's discriminant, one may recur to the solution proposed by [Mika et al., 1999, 2000]. Original scatter matrices in Equation (6.3) and Equation (6.4) are replaced by their counterparts computed in the RKHS. To derive the dual formulation enabling the use of kernels, we switch from the primal weights to the dual ones (representer theorem) with $\mathbf{w}^{\mathcal{H}} = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i)$, by already considering samples projected to the RKHS by means of the map $\phi(\cdot)$. It is possible to compute the mean value for class c in \mathcal{H} by $\boldsymbol{\mu}_c^{\mathcal{H}} = \frac{1}{n_c} \sum_{\mathbf{x}_i \in X_c} \phi(\mathbf{x}_i)$. Therefore, the value of the projected class average onto the discriminant subspace in the RKHS is:

$$\mathbf{m}_c^{\mathcal{H}} = \mathbf{w}^{\mathcal{H}} \boldsymbol{\mu}_c^{\mathcal{H}} = \frac{1}{n_c} \sum_{i=1}^n \sum_{\mathbf{x}_j \in X_c} \alpha_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = \frac{1}{n_c} \sum_{i=1}^n \sum_{\mathbf{x}_j \in X_c} \alpha_i k(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\alpha}' \bar{\mathbf{k}}_c \quad (6.10)$$

where $\bar{\mathbf{k}}_c$ is the column vector corresponding to $\bar{\mathbf{k}}_c = \frac{1}{n_c} \sum_{i=1}^n \sum_{\mathbf{x}_j \in X_c} \alpha_i k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{1}_c \mathbf{K}_c$, that is, the average value of kernel evaluations between the samples belonging to class c and all the training samples, in short \mathbf{K}_c . Here, $\mathbf{1}_c$ corresponds to a vector of length n_c with entries $1/n_c$. It is now possible to rewrite the numerator of Equation (6.2) by

6.2 Supervised flooded area extraction

considering $\mathbf{S}_b^{\mathcal{H}}$ in the RKHS as:

$$\begin{aligned}\mathbf{w}'^{\mathcal{H}} \mathbf{S}_b^{\mathcal{H}} \mathbf{w}^{\mathcal{H}} &= \mathbf{w}'^{\mathcal{H}} (\boldsymbol{\mu}_1^{\mathcal{H}} - \boldsymbol{\mu}_2^{\mathcal{H}}) (\boldsymbol{\mu}_1^{\mathcal{H}} - \boldsymbol{\mu}_2^{\mathcal{H}})' \mathbf{w}^{\mathcal{H}} \\ &= (\mathbf{w}'^{\mathcal{H}} \boldsymbol{\mu}_1^{\mathcal{H}} - \mathbf{w}'^{\mathcal{H}} \boldsymbol{\mu}_2^{\mathcal{H}}) (\mathbf{w}'^{\mathcal{H}} \boldsymbol{\mu}_1^{\mathcal{H}} - \mathbf{w}'^{\mathcal{H}} \boldsymbol{\mu}_2^{\mathcal{H}})' \\ &= \boldsymbol{\alpha}' (\bar{\mathbf{k}}_1 - \bar{\mathbf{k}}_2) (\bar{\mathbf{k}}_1 - \bar{\mathbf{k}}_2)' \boldsymbol{\alpha} \\ &= \boldsymbol{\alpha}' \mathbf{M} \boldsymbol{\alpha}\end{aligned}\tag{6.11}$$

$$\text{with } \mathbf{M} = (\bar{\mathbf{k}}_1 - \bar{\mathbf{k}}_2) (\bar{\mathbf{k}}_1 - \bar{\mathbf{k}}_2)'\tag{6.12}$$

Similarly, we can rewrite the denominator as:

$$\begin{aligned}\mathbf{w}'^{\mathcal{H}} \mathbf{S}_w^{\mathcal{H}} \mathbf{w}^{\mathcal{H}} &= \mathbf{w}'^{\mathcal{H}} \left(\sum_{c=1,2} \sum_{\mathbf{x}_j \in X_c} (\phi(x_j) - \boldsymbol{\mu}_c^{\mathcal{H}}) (\phi(x_j) - \boldsymbol{\mu}_c^{\mathcal{H}})' \right) \mathbf{w}^{\mathcal{H}} \\ &= \left(\sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i) \right)' \left(\sum_{c=1,2} \sum_{j \in X_c} (\phi(x_j) - \boldsymbol{\mu}_c^{\mathcal{H}}) (\phi(x_j) - \boldsymbol{\mu}_c^{\mathcal{H}})' \right) \left(\sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i) \right) \\ &= \sum_{c=1,2} \sum_{i=1}^n \sum_{j \in X_c} \left((\alpha_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle - \frac{1}{n_c} \alpha_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle) \right. \\ &\quad \left. (\alpha_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle - \frac{1}{n_c} \alpha_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle)' \right) \\ &= \left(\boldsymbol{\alpha}' \mathbf{K}_c \mathbf{K}'_c \boldsymbol{\alpha} + \frac{1}{n_c^2} \boldsymbol{\alpha}' \mathbf{K}_c \mathbf{K}'_c \boldsymbol{\alpha} - \frac{2}{n_c^2} \boldsymbol{\alpha}' \mathbf{K}_c \mathbf{K}'_c \boldsymbol{\alpha} \right) \\ &= \left(\boldsymbol{\alpha}' \mathbf{K}_c - \frac{1}{n_c} \boldsymbol{\alpha}' \mathbf{K}_c \right) \left(\boldsymbol{\alpha}' \mathbf{K}_c - \frac{1}{n_c} \boldsymbol{\alpha}' \mathbf{K}_c \right)' \\ &= \boldsymbol{\alpha}' \left(\sum_{c=1,2} \mathbf{K}_c \mathbf{I} \mathbf{K}'_c - \mathbf{K}_c \mathbf{I}_c \mathbf{K}'_c \right) \boldsymbol{\alpha} \\ &= \boldsymbol{\alpha}' \left(\mathbf{K} (\mathbf{I} - \mathbf{I}_1 - \mathbf{I}_2) \mathbf{K}' \right) \boldsymbol{\alpha} \\ &= \boldsymbol{\alpha}' \mathbf{N} \boldsymbol{\alpha}\end{aligned}\tag{6.13}$$

with $\mathbf{N} = \mathbf{K} (\mathbf{I} - \mathbf{I}_1 - \mathbf{I}_2) \mathbf{K}'$.

where \mathbf{I}_c is a $n_c \times n_c$ matrix with entries equal to $1/n_c$ on the diagonal.

The Rayleigh ratio in Equation (6.2) in the RKHS can be rewritten as:

$$\max_{\boldsymbol{\alpha}} \frac{\boldsymbol{\alpha}' \mathbf{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}' \mathbf{N} \boldsymbol{\alpha}},\tag{6.14}$$

solved again by the generalised eigendecomposition $\mathbf{M} \boldsymbol{\alpha} = \lambda \mathbf{N} \boldsymbol{\alpha}$ [Mika, 2002; Shawe-Taylor and Cristianini, 2004]. Finally, the projection of a new sample \mathbf{x} onto the kernel discriminant component and its classification is given by the sign of $f(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b$. In this case, b is set as half the distance between the RKHS mean projections.

Since the problem of estimating covariances in a possibly infinite dimensional space using n samples is ill-posed, and since the computation of the dual weights $\boldsymbol{\alpha}$ might be infeasible, \mathbf{N} must be regularized to ensure its non-singularity [Bandos et al., 2009; Friedman, 1989; Mika et al., 2000]. The introduction of a regularization parameter γ additionally controls the capacity when working in RKHS, alleviating over-fitting caused by

6. Supervised change detection

the curse of dimensionality. Here, we adopted a ridge penalization in Equation (6.14), as $\mathbf{N}_\gamma = \mathbf{N} + \gamma\mathbf{I}$, where \mathbf{I} is the identity matrix of size $n \times n$ and γ the penalty parameter to be tuned by the user. In this case γ penalizes large norms of the vector of dual coefficients $\boldsymbol{\alpha}$, since $\boldsymbol{\alpha}'(\mathbf{N} + \gamma\mathbf{I})\boldsymbol{\alpha} = \boldsymbol{\alpha}'\mathbf{N}\boldsymbol{\alpha} + \gamma\|\boldsymbol{\alpha}\|^2$. However, note that different penalization schemes exist [Mika et al., 1999, 2000]. In particular, it is worth mentioning the Tikhonov regularization: $\mathbf{w}^{\mathcal{J}c}(\mathbf{S}_w^{\mathcal{J}c} + \gamma\mathbf{I})\mathbf{w}^{\mathcal{J}c} = \mathbf{w}^{\mathcal{J}c}\mathbf{S}_w^{\mathcal{J}c}\mathbf{w}^{\mathcal{J}c} + \gamma\|\mathbf{w}^{\mathcal{J}c}\|^2$ that, when switching to the dual expression, becomes $\boldsymbol{\alpha}'\mathbf{N}\boldsymbol{\alpha} + \gamma\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} = \boldsymbol{\alpha}'(\mathbf{N} + \gamma\mathbf{K})\boldsymbol{\alpha}$. This solution penalizes the norm (of the possibly infinite dimensional) weight vector $\mathbf{w}^{\mathcal{J}c}$. The use of the latter regularization shows a link between the regularized kFDA and the least squares SVM [Gua et al., 2010; Mika et al., 2000; Van Gestel et al., 2002]. In parallel, as illustrated in [Bandos et al., 2009; Friedman, 1989], one might want to add the regularization to decrease the bias between the eigenvalues of the empirical covariance matrix and of the ones of the true covariance, since the largest eigenvalues of both matrices does not converge to the same value as $n \rightarrow \infty$.

As mentioned, it may happen that non-linear, multi-modal and heavily asymmetric and skewed distributions reduce the effectiveness of the linear FDA. An effective approach to alleviate these issues is to deform the empirical scatter matrices by local information issuing from the manifold distribution by adopting a graph Laplacian deformation. This is known as locality preserving Fisher’s discriminant [Sugiyama, 2007]. However, by adopting the kernel-based extension, these limitations are strongly relaxed. With the use of Gaussian RBF, it has been verified that the data in the RKHS are normally distributed [Bach and Jordan, 2002b; Cremers et al., 2003; Kwon and Nasrabadi, 2005]. Since non-Gaussianity in input space corresponds to Gaussianity in the RKHS, the kFDA results optimal, provided the correct hyperparameters. Note that the aforementioned graph regularization may still be employed to include information regarding the manifold into the kernel to enforce smoothness and locality preservation properties of the projections.

6.2.2 Experimental setup

The Landsat TM dataset used in the analyses is presented in Appendix B. It corresponds to a recent flooding occurred in a tributary of the Missouri River in South Dakota (USA).

Training and testing labels were carefully selected by visual inspection from two spatially disjoint regions of the image in order to avoid spatial autocorrelation when estimating figures of merit (see Appendix B for the details). They are composed by three classes: “flooded”, “not flooded” and “permanent standing water”. Recall that in the experiments, the latter class is recoded either to “flooded” or “not flooded” depending on the temporal composition to be classified. The training set is composed of 48’379 examples while the testing of 88’501. Labelled areas were chosen so that the class variability is well represented in both train and test regions, in particular for the heterogeneous “not flooded” class. Specifically for the “flooded” class, different water colours have been included in the sets. Pixels corresponding to regions that are only partially flooded or covered by shallow water presented a spectral contamination by the ground cover before

6.2 Supervised flooded area extraction

the event, attenuated by the water absorption proportionally to its depth. This further increases variance and class overlap with the permanent standing water, for which the same phenomenon is observed. For the illustration of the original details used to visually validate the mapping settings, see Figure 6.4 (Ex. 2,3,4).

The regularized kFDA is trained with random subsets composed of 10 to 1000 examples per class (in a balanced classification setting), allowing us to evaluate its robustness and sensitivity under different training sets sizes. Final numerical accuracies are averaged over 10 independent realizations of such sets, to have a robust estimate and a confidence interval over the values. The upper limit of 1000 training samples is given by a plateau effect observed on the numerical accuracies for larger sets. When exceeding this size, only a decrease in the standard deviation of the accuracy measure has been observed. For both the uni- and multi-temporal setting, digital numbers for each spectral band were mean-centred and scaled to unit variance prior to experiments.

Linear and nonlinear models, applied to the uni- and multi-temporal problems, resulted in four independent mapping tasks. Since the Gaussian RBF kernel has been adopted, two hyperparameters have to be set: the σ and the regularization parameter γ . Model selection by grid search within a 3-fold cross validation scheme has been applied. The σ was optimized in $\{0.5\sigma_e, 0.6\sigma_e, \dots, 1.5\sigma_e\}$, where σ_e is the median Euclidean distance between 5000 randomly selected pixel from the initial dataset. This choice avoids falling in over-/under-fitting situations caused by a bad choice of the kernel bandwidth, in particular for small training sets. The γ parameter was optimized in the range $\{10^{-3}, 10^{-2}, \dots, 10^2\}$. The outcomes are then numerically evaluated and compared by considering the estimated Cohen’s Kappa statistic (κ), the error matrices and the McNemar test (see Appendix A).

6.2.3 Results

Figure 6.2 illustrates the estimated κ coefficient as a function of the training set size. Since for each of the tested settings the training sets are composed by the same pixels co-ordinates, curves are comparable. Still, the label switch for the permanent standing water pixels should be always kept in mind.

Results suggest that the most accurate mapping method resides in the multi-temporal nonlinear kFDA classification, with a peak of 0.937 average κ points for a training set composed by 500 pixels. For larger sizes the model shows a plateau effect around $\kappa = 0.93$. Its linear counterpart performs poorly. Even though the accuracy for the smallest training sets (10, 50 pixels) is comparable, the nonlinear kFDA rapidly outperforms the linear algorithm for larger sets. The Gaussian kFDA false alarm rate is also strongly reduced, a fact also depicted by the “flooded” class user’s accuracy, increasing from 67.52% to 89.42%. This ability is also underlined by observing the label assignment for only the “permanent standing water” pixels, which are classified correctly on the average the 10% of the times for the linear model (into “not flooded” class) and 57.27% by the nonlinear one. Even if the linear model seems to be generally less accurate, it provides a higher detection rate for the flooding class, corresponding to few missed detections of the flooded

6. Supervised change detection

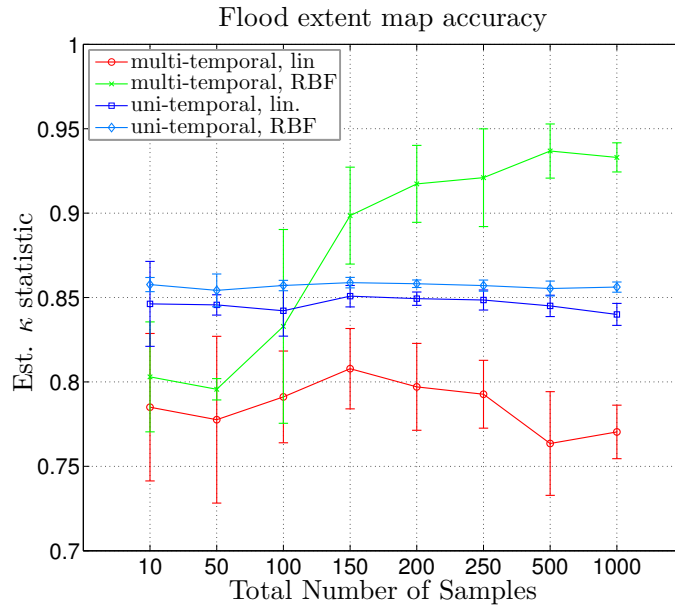


Figure 6.2: Supervised flood mapping results - Estimated κ statistic and standard deviation (error bars) for the flood mapping tests.

area, which are comparable to the nonlinear kFDA outcomes. A McNemar test at level of $p = 0.001$ indicates a significantly higher average accuracy of the nonlinear approach for models built with 100 or more training samples per class.

In the uni-temporal perspective, the problem reduces to a binary classification of water against the rest. This problem can be solved easily, since the spectral signature of the water is usually well distinguishable from other land covers. This results in accuracy curves for the linear and nonlinear models that behave similarly, with a slight improvement for the nonlinear one. The performance still significantly lower than the aforementioned multi-temporal Gaussian kFDA (in the range of 0.05 to 0.1 κ points). Nevertheless, the lower performance of these models is counterbalanced by the stable accuracy with respect to the class sizes, i.e. models trained on 10 and 50 samples performs as the ones using 1000 pixels. This leads to two observations: on the one hand the class water is (mostly) linearly separable and easily discriminable, thanks to the strongly clustered distribution. On the other hand, some issues related to class overlap seem to limit the performance of the uni-temporal models. The close performances of both the linear and nonlinear models can also be observed from the error matrices (Table 6.1). The nonlinear model performs significantly better at $p = 0.01$ (but not at $p = 0.001$) for all the different training set sizes. The linear model appears again to be more conservative in the prediction of flooded pixels (detection rates are higher than in the nonlinear scheme). This is reflected in a higher user’s accuracy. For the nonlinear kFDA the producer’s accuracy for the “flooded” class is slightly inferior, but the overall accuracy results larger.

6.2 Supervised flooded area extraction

		uni-temporal			multi-temporal			
		<i>True</i>			<i>True</i>			
		F	NF	Us.	F	NF	Us.	
Lin. <i>E_{st.}</i>	F	10842.9	4890.4	68.92	F	10841.6	5225.1	67.52
	NF	5.1	72762.6	99.99	NF	6.4	72427.9	99.99
	Pr.	99.95	93.70	94.47	Pr.	99.94	93.27	94.09
RBF <i>E_{st.}</i>	F	10806.9	4584.5	70.22	F	10739	1284.7	89.42
	NF	41.1	73068.5	99.94	NF	109	76368.2	99.86
	Pr.	99.62	94.10	94.77	Pr.	98.99	98.35	98.43

Legend	F	Flooding	NF	Not flooding
	Est.	Predicted label	True	True label
	lin.	Linear	RBF	Radial basis function
	Pr.	Producer’s accuracy	Us.	User’s accuracy

Table 6.1: Average error matrices - Obtained using models trained on 500 pixels. Accuracy values are expressed in [%], in bold the overall accuracy.

6.2.4 Discussion

The classification of the permanent standing water as “not flooded” requires nonlinear strategies in a multi-temporal setting. Otherwise, standard single image classification methods can easily detect pure water pixels. However, the latter only provides a cartography of the water bodies present in the post-event scene, after the flood occurred, and cannot be considered as a proper flood extent map by itself without performing further adjustments. However, even with very small training sets, single time image classification using kFDA appears to be robust to the size of the training sets.

In the multi-temporal scenario the ability of the Gaussian kFDA to exploit all the spectral information to solve the mixed / ambiguous samples in the input stack makes the approach very accurate, but only when considering 150 or more samples to train the models. The linear model cannot separate the permanent standing water from the examples belonging to the flood class, making the maps less pertinent. By observing Figure 6.3, to separate the permanent standing water from the “flooded” class in the different spectral variables, a nonlinear boundary is often required, due mainly by the high class mixing, when merging permanent standing water samples and unchanged pixels. The classification task can also cope with the high spectral variance due to the large heterogeneity of the multi-modal classes. The separation is more complex than simple water discrimination in single image classification. It results clearly that the suboptimal model tends to assign “not flooding” pixels to the “flooded” class, instead of wrongly predicting mixed pixels and shallow water to dry regions. In particular, the class distributions represented in Figure 6.3 show different behaviours: If the flood and permanent standing water can be

6. Supervised change detection

roughly approximated by two overlapping normal distributions with a larger variance in the second time direction, the same can not be stated for the “not flooding” class, strongly multi-modal and scattered density. This explains why the linear model fails in discriminating only the flooded areas from the permanent standing water. For the nonlinear kFDA, without any assumption on the underlying distribution, examples are mapped into a Gaussian RKHS, resulting in a correct maximization of the separability of the classes [Huang and Hwang, 2006].

The nonlinear multi-temporal setting provides a reliable flood extent map correctly delineating only the exceeding water. In more detail, Figure 6.4 Ex. 1, shows that the small water basins appearing far from the river bed are correctly delineated by both multi- and uni-temporal approaches. The small river in the upper-left part of the image does not present a general augmentation of the surface due to flooding, but only a small area in the upper part is related to exceeding water. In this case, the permanent standing water has been again correctly discriminated by the nonlinear classifier using stacked input data.

In Figure 6.4 Ex. 2, a larger portion of flooded river is considered. The river in 2005 is clearly visible meandering on the alluvial valley. Again, the multi-temporal kFDA does not detect it as flooded area, correctly delineating the exceeding water. On the right hand side of the image, a variety of small water bodies appear, and in this case all the approaches detect well the most of the puddles.

The last two examples in Figure 6.4 Ex. 3 and Figure 6.4 Ex. 4 show a combination of the two aforementioned examples, with the addition of permanent lakes. Clearly, the uni-temporal approaches cannot discriminate them, since no temporal component is exploitable. The multi-temporal Gaussian kFDA correctly assigns the permanent standing water to the class “not flooded”, generating pertinent maps of the event.

Future developments may be conducted mainly by improving the representativeness of the training samples by optimizing their input space. This can be achieved by injecting into the classification problem relevant information as features related to physical properties of the phenomenon, such as normalized difference vegetation index (NDVI), normalized water difference index (NDWI), or surface temperature and elevation models. Also, the complementarity with SAR data could be exploited, that has proven to be useful but suboptimal in flooded area extraction tasks. Additionally, to reduce the heterogeneity of classes and in particular for the “not flooded” regions, spatial filters smoothing the spectral information as a function of the spatial context of each pixel may be considered. To this end, data fusion and multi-source methods are a worthwhile research direction.

In the next case study, involving supervised change detection for monitoring the urban area of Zurich, these possibilities are investigated for VHR images. In particular, we examine the improved informativeness of the input space when filling the lacks of spectral information by injecting the spatial context of pixels.

6.2 Supervised flooded area extraction

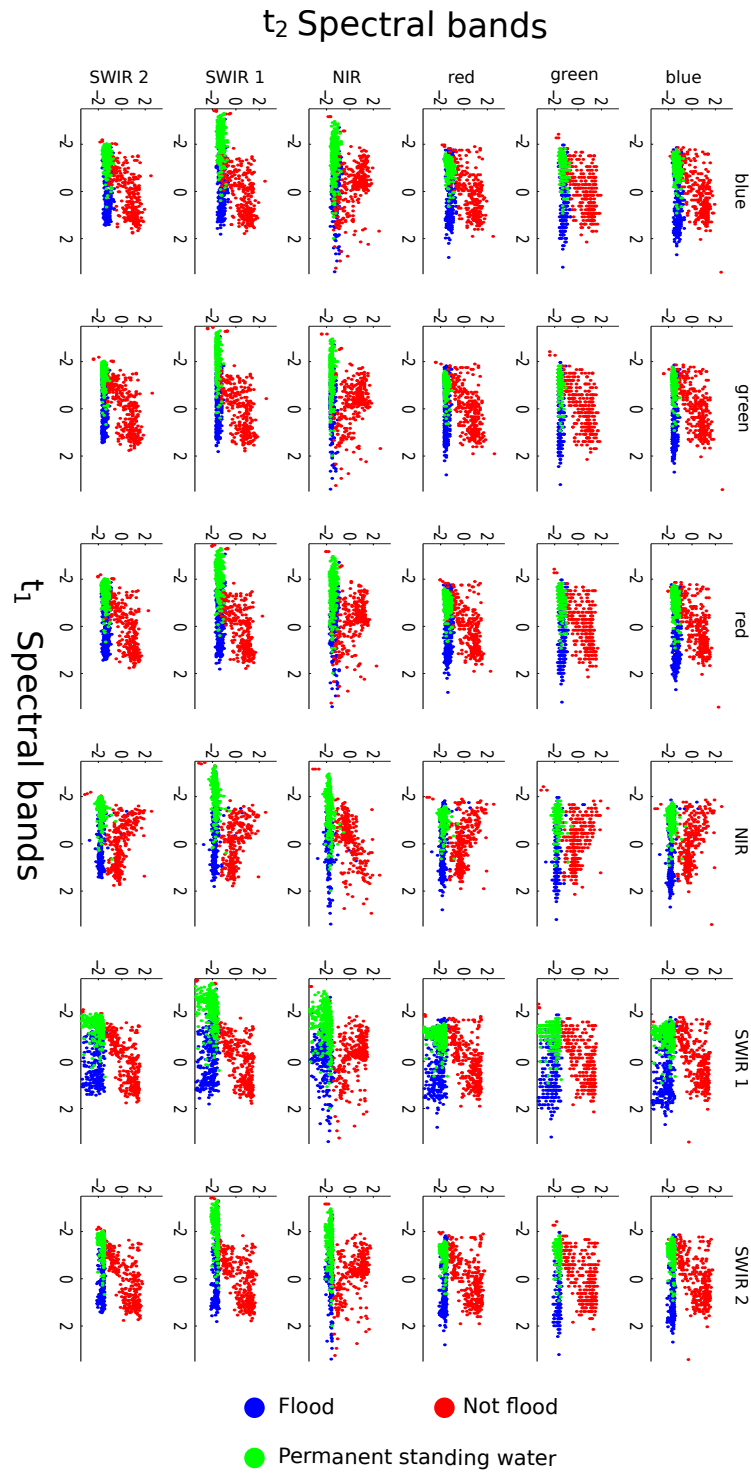


Figure 6.3: Scatterplot matrix of the multi-temporal Landsat TM flooding data - Obtained by subsampling the test set at 300 pixels per class after normalization (mean centered and unit variance).

6. Supervised change detection

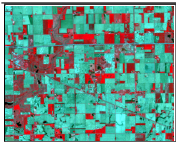
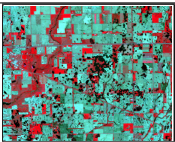
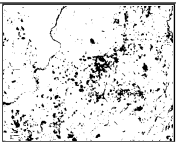
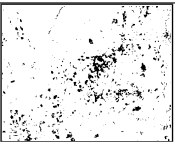
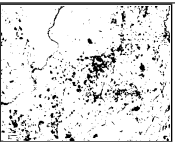
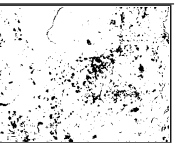

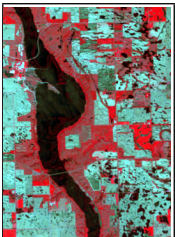




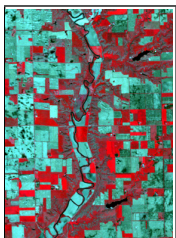
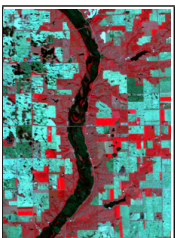




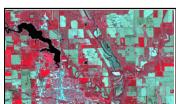


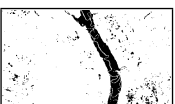

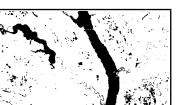
pre-event	post-event	multi-temporal		uni-temporal	
		linear $\kappa = 0.782$	Gaussian $\kappa = 0.952$	linear $\kappa = 0.850$	Gaussian $\kappa = 0.859$
					
Ex. 1a	Ex. 1b	Ex. 1c	Ex. 1d	Ex. 1e	Ex. 1f
					
Ex. 2a	Ex. 2b	Ex. 2c	Ex. 2d	Ex. 2e	Ex. 2f
					
Ex. 3a	Ex. 3b	Ex. 3c	Ex. 3d	Ex. 3e	Ex. 3f
					
Ex. 4a	Ex. 4b	Ex. 4c	Ex. 4d	Ex. 4e	Ex. 4f

Figure 6.4: Subsets of the Landsat TM flooding scene - (a)-(b) Subset of the original images; (c)-(f) detail of the flood extent map.

6.3 Exploiting the spatial context in VHR supervised change detection for urban monitoring

As mentioned in the concluding remarks of the Section before, the spatial context of each pixel may alleviate issues related to low class separability and to the heterogeneity of the class-conditional distributions. The exploitation of this information in multi-temporal and change detection applications is poorly documented in the literature, as pointed out in Section 5.2, although the benefits of considering such variables are clearly demonstrated in classification tasks, particularly for VHR images [Benediktsson et al., 2005; Pacifici et al., 2009; Tuia et al., 2009, 2010b].

In this Section, two change detection architectures are considered: direct multi-date classification (DMC) and supervised difference image analysis (DIA) (see Section 5.1). The rationale of the approach is to exploit the benefits of the improved representativeness of the input space, while exploiting the properties of the SVM classifier, proved to be a suitable tool in many remote sensing applications [Camps-Valls and Bruzzone, 2009].

6.3.1 The support vector machines for classification

SVM are a non-parametric binary classifier relying on Vapnik’s statistical learning theory [Vapnik, 1998] (see Chapter 3 of this Thesis). This method aims at building a linear separation rule of the form $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ between examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_s}$. The final decision whether a sample belongs to the class $y_i \in \{+1; -1\}$ is given by the sign of the decision function. The issue resides in finding the weight vector \mathbf{w} and bias term b , as in the kFDA, defining the separating hyperplane. Following Chapter 3, the solution guaranteeing the optimal generalization ability is the ones that finds a trade-off between the minimization of the training error and a control of the complexity. The SVM problem may be formulated as a regularized problem of the form:

$$\min_{\mathbf{w}, b, \xi} \frac{1}{n_s} \sum_{i=1}^{n_s} \mathcal{L}(f(\mathbf{x}_i), y_i) + \gamma \|f\|^2 \quad (6.15)$$

In the SVM formulation, the hinge loss is adopted: $\mathcal{L}(f(\mathbf{x}_i), y_i) = \max(1 - y_i f(\mathbf{x}_i), 0)$ [Boser et al., 1992; Cortes and Vapnik, 1995]. This function penalises samples that lie inside the 1-margin of the model f and increases as the decision function grows with wrong sign (recall the sign of $f(\mathbf{x})$ corresponds to the predicted class). By doing so, SVM fits a separating boundary with the largest margin between the examples of the two classes. To make the classifier robust to outliers by allowing some training errors, it is possible to relax the $\mathcal{L}(f(\mathbf{x}_i), y_i)$ with slack variables ξ_i . The SVM problem becomes:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{n_s} \sum_{i=1}^{n_s} \xi_i + \gamma \|f\|^2 \\ \text{s.t.} \quad & \mathcal{L}(f(\mathbf{x}_i), y_i) \geq \xi_i \end{aligned} \quad (6.16)$$

6. Supervised change detection

By recasting the problem into \mathcal{H} , the nonlinear formulation in the original space is obtained. The solution is given by the hyperplane defined by $\mathbf{w}^{\mathcal{H}}$, and $f(\mathbf{x}) = \langle \mathbf{w}^{\mathcal{H}}, \phi(\mathbf{x}) \rangle + b$. We can further modify the objective function as:

$$\begin{aligned} \min_{\mathbf{w}^{\mathcal{H}}, b, \xi} \quad & C \sum_{i=1}^{n_s} \xi_i + \frac{1}{2} \|\mathbf{w}^{\mathcal{H}}\|^2 \\ \text{s.t.} \quad & 1 - y_i \left(\langle \mathbf{w}^{\mathcal{H}}, \phi(\mathbf{x}) \rangle + b \right) \geq \xi_i \\ & \xi_i \geq 0. \end{aligned} \quad (6.17)$$

Note that in Equation (6.17) the two terms have been multiplied by $\frac{1}{2\gamma}$ and replaced by a cost term $C = \frac{1}{2\gamma n_s}$. This manipulation simplifies the derivation of the expression for the subsequent optimization. Now, similarly to γ , C has to be tuned by the user and it controls the trade-off between the maximization of the hyperplane margin and the number of allowed training errors. This further strengthens the generalization ability on previously unseen data from $P(\mathbf{x}, y)$. This constrained quadratic optimization problem may be solved by introducing the Lagrange multipliers α for the first constraint and β for the second:

$$\begin{aligned} \min_{\mathbf{w}^{\mathcal{H}}, b, \xi} \max_{\alpha, \beta} \quad & L(\mathbf{w}^{\mathcal{H}}, b, \xi, \alpha, \beta) = C \sum_{i=1}^{n_s} \xi_i + \frac{1}{2} \|\mathbf{w}^{\mathcal{H}}\|^2 \\ & - \sum_{i=1}^{n_s} \alpha_i \left(\xi_i - 1 + y_i \left(\langle \mathbf{w}^{\mathcal{H}}, \phi(\mathbf{x}) \rangle + b \right) \right) - \sum_{i=1}^{n_s} \beta_i \xi_i. \end{aligned} \quad (6.18)$$

The optimum is given by the saddle point of $L(\mathbf{w}^{\mathcal{H}}, b, \xi, \alpha, \beta)$. By fixing α and β , the partial derivatives of L with respect to $\mathbf{w}^{\mathcal{H}}, b$ and ξ are then equated to 0:

$$\frac{\partial L}{\partial \mathbf{w}^{\mathcal{H}}} = \mathbf{w}^{\mathcal{H}} - \sum_{i=1}^{n_s} \alpha_i y_i \phi(\mathbf{x}_i) = 0 \quad (6.19)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n_s} \alpha_i y_i = 0 \quad (6.20)$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \quad (6.21)$$

Using the optimality condition in Equation (6.19), $\mathbf{w}^{\mathcal{H}} = \sum_{i=1}^{n_s} \alpha_i y_i \phi(\mathbf{x}_i)$. Finally, the problem in the dual space is obtained by replacing the above derivatives into Equation 6.18 and solving. The expression is optimized by finding the α maximizing (note that by replacing $\beta_i \xi_i = (C - \alpha_i) \xi_i$, the β disappeared) [Boser et al., 1992; Schölkopf and Smola, 2002; Suykens and Alzate, 2010]:

$$\max_{\alpha} \quad \sum_{i=1}^{n_s} \alpha_i - \frac{1}{2} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} \alpha_i \alpha_j y_i y_j \langle \phi(\mathbf{x}_i) \phi(\mathbf{x}_j) \rangle \quad (6.22)$$

where α_i are the coefficients determining the solution of the optimization problem. As illustrated in the Chapter 4, we may now apply the kernel trick to obtain the final kernel

6.3 Supervised change detection for urban monitoring

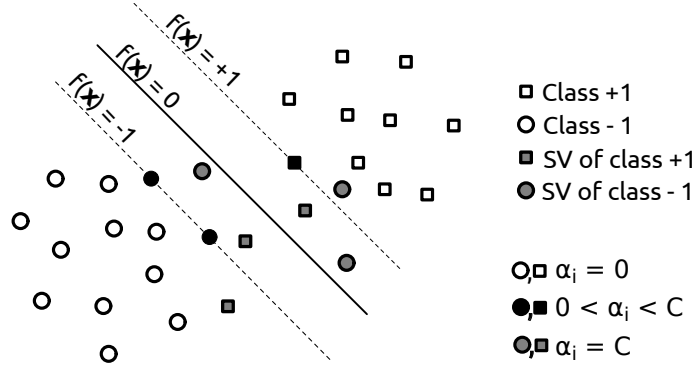


Figure 6.5: SVM graphical interpretation - An example of classification using SVM. The different situation involving the α corresponding to specific training samples.

formulation as:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n_s} \alpha_i - \frac{1}{2} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_{i=1}^{n_s} \alpha_i y_i = 0. \end{aligned} \quad (6.23)$$

When the solution to Equation (6.23) is found, the label of an unknown sample \mathbf{x} is given by the position with respect to the separating hyperplane:

$$\hat{y} = \text{sign}(f(\mathbf{x})) = \text{sign} \left(\sum_{i=1}^{n_s} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b \right). \quad (6.24)$$

Thanks to the primal-dual relationship, Equation 6.24 corresponds to the solution of $\hat{y} = \text{sign}(f(\mathbf{x})) = \text{sign}(\langle \mathbf{w}^{\mathcal{J}}, \phi(\mathbf{x}) \rangle + b)$. We obtain that for any training sample \mathbf{x}_i with $0 < \alpha_i < C$ (an unbounded support vector), ξ_i and $1 - y_i (\langle \mathbf{w}^{\mathcal{J}}, \phi(\mathbf{x}) \rangle + b)$ are both equal to 0. The offset b may be obtained as $b = y_i - \langle \mathbf{w}^{\mathcal{J}}, \phi(\mathbf{x}_i) \rangle = y_i - \sum_{j=1}^{n_s} y_j \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)$. By observing the solutions α_i corresponding to samples \mathbf{x}_i , we have three different situations:

- $\alpha_i = 0$ The training sample is correctly classified, i.e. it lies on the correct side of the separating hyperplane with $y_i f(\mathbf{x}_i) > 1$. It does not contribute to the decision function.

- $0 < \alpha_i < C$ The sample \mathbf{x}_i is an unbounded support vectors, implying that it lies exactly on the class margin, and $y_i f(\mathbf{x}_i) = 1$.

- $\alpha_i = C$ The example \mathbf{x}_i is a bounded support vector, lying inside or outside (but on the wrong side) the separating boundary. They correspond to training errors, and $y_i f(\mathbf{x}_i) < 1$.

A graphical interpretation of SVM classifiers, according to the definitions given above, is given in Figure 6.5.

Depending on the implementations, multi-class SVM may be obtained by reformulating the problem involving $|Y|$ classes into different binary sub-problems. The most used

6. Supervised change detection

approaches are the one-against-all (OAA), solving $|Y|$ binary sub-problems (shattering each class from all the others at a time) and assigning the label as the class with the largest decision function, and the one-against-one (OAO), that builds $|Y|(|Y|-1)/2$ binary separations discriminating one class from the other and assign the label as the class with the most frequent outcome. Note that direct multi-class optimization exists, and it is based on a multi-class hinge loss function [Suykens and Alzate, 2010]. In the next Sections, the contextual filters considered to account for the spatial information are presented.

6.3.2 Textural features

Occurrence and co-occurrence textural statistics (TXT) [Baraldi and Parmiggiani, 1995; Haralick et al., 1973] are local indexes computed on the basis of overlapping moving windows of size $P \times Q$ (usually $P = Q$). The resulting variables emphasize the local texture structures of the graylevel image. The image from which the statistics are retrieved can be of different forms: in the case of multi-spectral VHR scenes it is common to use the panchromatic band, the first principal component or a task-dependent discriminative band or combinations of them (e.g. NDVI).

Occurrence statistics These measures are computed on the intensity values contained in the moving window centred on the pixel x_{ij} . They return a local texture value defined by the statistic T at x_{ji} , as x_{ij}^T . In this Section, two occurrence indicators are considered, mean and variance:

$$x_{ij}^M = \frac{1}{|\mathcal{V}|} \sum_{p,q \in \mathcal{V}} x_{pq} \quad (6.25)$$

$$x_{ij}^{\text{VAR}} = \frac{1}{|\mathcal{V}|} \sum_{p,q \in \mathcal{V}} (x_{pq} - x_{ij}^M)^2 \quad (6.26)$$

where \mathcal{V} denotes the neighbourhood of the pixel x_{ij} (note that, unless stated otherwise, ij are the spatial coordinates of the pixel), and $|\mathcal{V}|$ their number ($|\mathcal{V}| = P \cdot Q$). The local average (M) reduces effects of noise and outliers such as saturated pixels, by smoothing their large values. The local variance (VAR) indicator summarizes differences in the graylevel values contained in the considered patch, emphasizing edges between objects at different scales. Other indicators such as skewness or kurtosis can be considered for additional information on the graylevel distribution [Haralick et al., 1973].

Co-occurrence statistics These indicators are based on the graylevel co-occurrence matrix (GLCM), that represents the relative occurrence frequency $p(m, n)$ of two graylevel values m and n in the $P \times Q$ window at a given angular neighbourhood (note that the radiometric scale of the image values may be changed to avoid null occurrences). The lag is given by a connecting vector (δ_x, δ_y) in x and y spatial coordinates. Many statistical texture descriptors can be extracted on the basis of the GLCM [Haralick et al., 1973;

6.3 Supervised change detection for urban monitoring



Figure 6.6: Multi-scale occurrence texture statistic - Four examples of moving window-based occurrence statistic, the mean (M). For each image, the outcome for the 3 considered window sizes (squares of size 3, 7, 15) are illustrated in growing order (M3-M15) along the original image (IM).

Petrou and Sevilla, 2006]. In this paper three descriptors are adopted: entropy (ENT), angular second moment (ASM) and homogeneity (HOM).

$$x_{ij}^{\text{ENT}} = - \sum_{m,n} p(m,n) \log p(m,n) \quad (6.27)$$

$$x_{ij}^{\text{ASM}} = \sum_{m,n} p(m,n)^2 \quad (6.28)$$

$$x_{ij}^{\text{HOM}} = \sum_{m,n} \frac{p(m,n)}{1 + |m - n|}. \quad (6.29)$$

ENT is a measure of information content and can be interpreted as a the randomness of the graylevel values. Regions with high variance of the graylevels will result in high entropy, while smooth patches correspond to low entropy. ENT is a good indicator of the intensity of the texture in the considered patch. ASM indicates the local contrast. It provides an accurate estimate on the degree of uniformity of the values of the GLCM. A low ASM value indicates that no spatial coherence characterizes the patch. HOM measures the variance around the diagonal of the GLCM. In homogeneous patches, the values are clustered around the diagonal resulting in high HOM statistic value. Other GLCM-based indicators can be used, such as correlation or contrast [Haralick et al., 1973], but have been disregarded since highly correlated to the ones listed above.

6. Supervised change detection



Figure 6.7: Multi-scale co-occurrence (GLCM) texture statistic - Four examples of moving window-based GLCM statistics, the HOM feature. For each image, the outcome for the 3 considered window sizes (squares of size 3, 7, 15) are illustrated in growing order (HOM3-HOM15) along the original image (IM).

6.3.3 Mathematical morphology

Many textural indices may present similar statistics for different classes. Consequently, they are insufficient to describe properly the spectral classes. To solve this issue, the joint use of texture indicators with multi-band morphological profiles [Benediktsson et al., 2005; Fauvel et al., 2008] is proposed. The mathematical morphology (see [Soille, 2004; Soille and Pesaresi, 2002] for details) defines a family of operators that aim at emphasizing homogeneous spatial structures in a graylevel image. The resulting variables present higher autocorrelation for neighbouring pixels in the same object, reducing noise and within-class variance. Since a multi-band approach is adopted, the between-class variance may ensue increased, improving separability. These filters are based on a moving window of given shape and size called the structuring element S .

Basic operations are erosion and dilation, respectively denoted as $\epsilon_S(x_{ij})$ and $\delta_S(x_{ij})$. They are defined as follows:

$$\epsilon_S(x_{ij}) = \min(x_{ij}, x_s) \quad \forall x_s \in S_{ij} \quad (6.30)$$

$$\delta_S(x_{ij}) = \max(x_{ij}, x_s) \quad \forall x_s \in S_{ij}, \quad (6.31)$$

they return respectively the minimum and the maximum value between pixel x_{ij} and the ones contained in the structuring element S_{ij} centred on x_{ij} .

6.3 Supervised change detection for urban monitoring

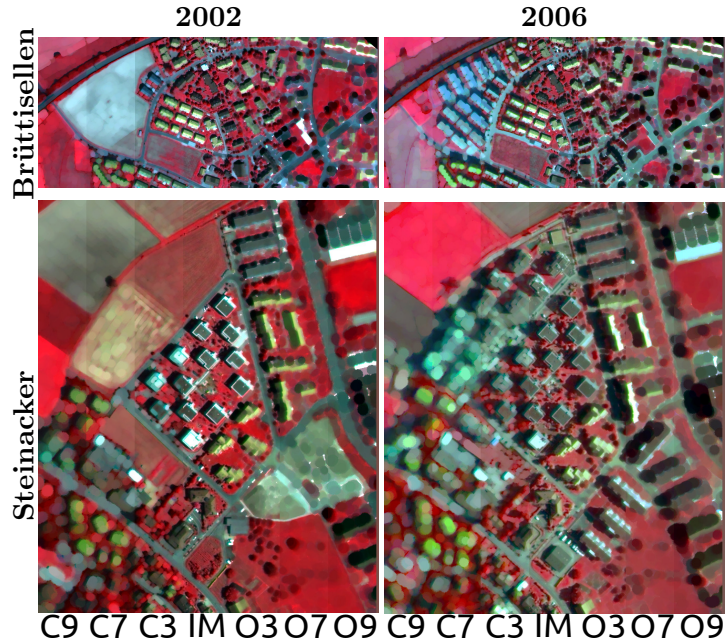


Figure 6.8: Multi-scale opening and closing morphological operators - Four examples of opening and closing morphological operators. For each image, the outcome for the 3 considered structuring element sizes (disks of radius 3, 7, 9) are illustrated in decreasing order for the closing operator (C9-C3). The original image is in the centre (IM), while on the right the opening operator for the same structuring element illustrated in increasing order (C3-C9).

Opening and closing (OC) These two filters are the concatenation of erosion and dilation:

$$\gamma_S(x_{ij}) = \delta_S(\epsilon_S(x_{ij})) \quad (6.32)$$

$$\omega_S(x_{ij}) = \epsilon_S(\delta_S(x_{ij})). \quad (6.33)$$

The opening $\gamma_S(x_{ij})$ of the graylevel image filters out elements that are brighter than the ones contained in the neighbourhood defined by the structuring element S . Closing $\omega_S(x_{ij})$ filters out darker elements in the same range.

Opening and closing by reconstruction (OCR) Although emphasizing meaningful contextual information, opening and closing do not preserve the shape of objects represented in the image. To provide the same level of smoothing but preserving the geometrical information at precise object level, the use of reconstruction filters is proposed [Fauvel et al., 2008; Soille, 2004].

Opening and closing by reconstruction are noted as $\rho_{\delta_S}(I_M)$ and $\rho_{\epsilon_S}(I_M)$ respectively. These operations reconstruct the original image by iterative cycles of erosions or dilations on a marker image I_M . If I_M is an erosion of the original image ($I_M = \epsilon_S(x_{ij})$), the latter

6. Supervised change detection

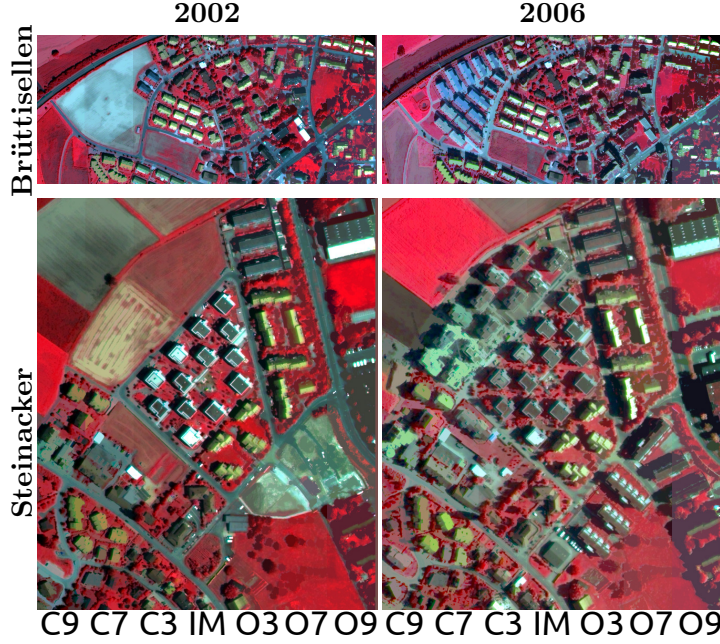


Figure 6.9: Multi-scale opening and closing by reconstruction morphological operators - Four examples of opening and closing by reconstruction operators. For each image, the outcome for the 3 considered structuring element sizes (disks of radius 3, 7, 9) are illustrated in decreasing order for the closing operator (C9-C3). The original image is in the centre (IM), while on the right the opening operator for the same structuring element illustrated in increasing order (C3-C9).

is reconstructed by iterative series of dilations of I_M as $I_M^k = \delta^1 \delta^2 \delta^3 \dots \delta^k(I_M)$ resulting in the opening by reconstruction:

$$\rho_{\delta_S}^k(\epsilon_S(x_{ij})) = \min(I_M^k, x_{ij}) \quad (6.34)$$

and the process is iterated until $\rho^k = \rho^{k-1}$. Similarly, closing by reconstruction reconstructs the graylevel image starting from its dilated version $I_M = \delta_S(x_{ij})$ iteratively performing erosions of the marker image I_M as $I_M^k = \epsilon^1 \epsilon^2 \epsilon^3 \dots \epsilon^k(I_M)$:

$$\rho_{\epsilon_S}^k(\delta_S(x_{ij})) = \max(I_M^k, x_{ij}), \quad (6.35)$$

converging to the final filtering when $\rho^k = \rho^{k-1}$. As for the OC operators, opening and closing by reconstruction filter out brighter and darker elements smaller than S_{ij} , but preserving the original spatial structures larger than S , since the reconstruction is constrained by values of the original images.

In all the cases, the signal of the spatial-context-augmented input vectors allows a better discrimination among changed and unchanged classes, as depicted in Figure 6.10. For the DMC approach, the per-pixel input signal is considered as a whole, while for the DIA only the punctual differences are used to discriminate the different changed areas.

6.3 Supervised change detection for urban monitoring

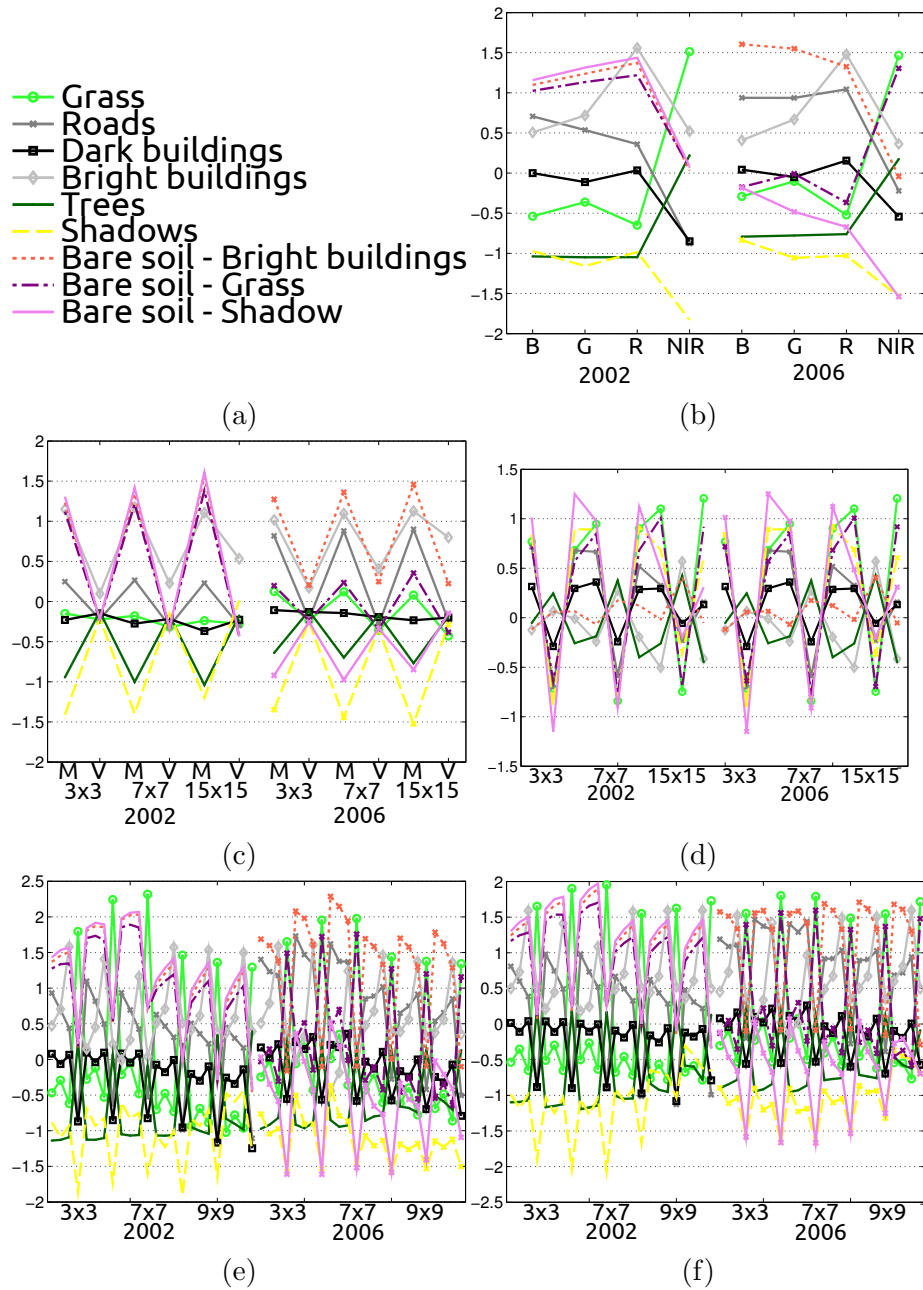


Figure 6.10: Multi-scale class-wise multi-temporal signal - Example of the newly created input space, illustrated by the average value of the features for each class, in 2002 and 2006. In (a) the legend of the classes, in (b) the average spectral signature per class, in (c) the average occurrence features (OCC) Mean (M) and Variance (V), (d) GLCM co-occurrence (for each window size, ordered as ENT ASM HOM), (e) opening-closing (for each structuring element size, O C for each band) and (f) opening and closing by reconstruction (same as for OC), again illustrated on a per-class average basis. Note that values are standardized.

6. Supervised change detection

Set Name	Dimensions	Description
IMM	4	Pansharpened bands
TXT	15 (+4)	6 occurrence and 9 co-occurrence
OC	24 (+4)	Opening and closing
OCR	24 (+4)	Opening and closing by reconstruction
OCOOCR	48 (+4)	OC and OCR stacked
OCTXT	39 (+4)	OC and TXT stacked
OCRTXT	39 (+4)	OCR and TXT stacked
OCOVRTXT	63 (+4)	OC, OCR and TXT stacked

Table 6.2: Contextual information feature blocks - The number of the features refers to a single date. For both dates, same features with same parameters are extracted. For each set of features, the pansharpened image is included (+4, the IMM set).

Furthermore, as depicted by the scatterplots in Figure 6.11, the inclusion of the spatial context eases the process of class separation. Specifically, multi-channel reconstruction operators allow a better classification by both reducing within-class variability and by including discriminant information, thus increasing the distance between classes.

6.3.4 Experimental setup

Textural features are computed on the corresponding panchromatic bands (2002 and 2006). For each occurrence statistic, three window sizes are considered (3×3 , 7×7 and 15×15), resulting in 6 variables per date as illustrated in Figure 6.6 with the corresponding signal in Figure 6.10(c). Regarding co-occurrence indicators, the average of the statistics computed in four directions (0° , 45° , 90° and 135°) has been considered, with a shift in horizontal and vertical directions proportional to the moving window size. The reason of considering the average on four directions is that, since the GLCM-based indicators are symmetric, e.g. $\hat{x}_{ij}(0^\circ) = \hat{x}_{ij}(180^\circ)$, their average is invariant to rotation. Three window sizes have been utilized for computing the GLCM (3×3 with a shift of 1 pixel, 7×7 with a shift of 2 pixels and 15×15 with a shift of 5 pixels) resulting in 9 co-occurrence variables as depicted in Figure 6.7 and Figure 6.10(d). The choice of the window size is related to the resolution of the objects represented in the scene. To preserve the level of details, 3×3 pixels windows have been computed (roughly corresponding to squares of 2[m] of side) to provide information about small patches as trees and small buildings, along with abrupt variations in object borders. The 7×7 window accounts for local structures in a range of 5[m], including information at building and road level, as well as smooth changes among different texture classes. Finally, the 15×15 window provides textural information for larger regions (approximately 10[m]) accounting for trends in fields and grasslands as well as commercial buildings. Larger windows have not been considered since the scenes are mainly characterized by small and medium sized objects. Finally, morphological filters have been implemented with three different disk-shaped structuring elements, with radius 3, 7 and 9 pixels, independently for all the spectral channels of the images. The size of the

6.3 Supervised change detection for urban monitoring

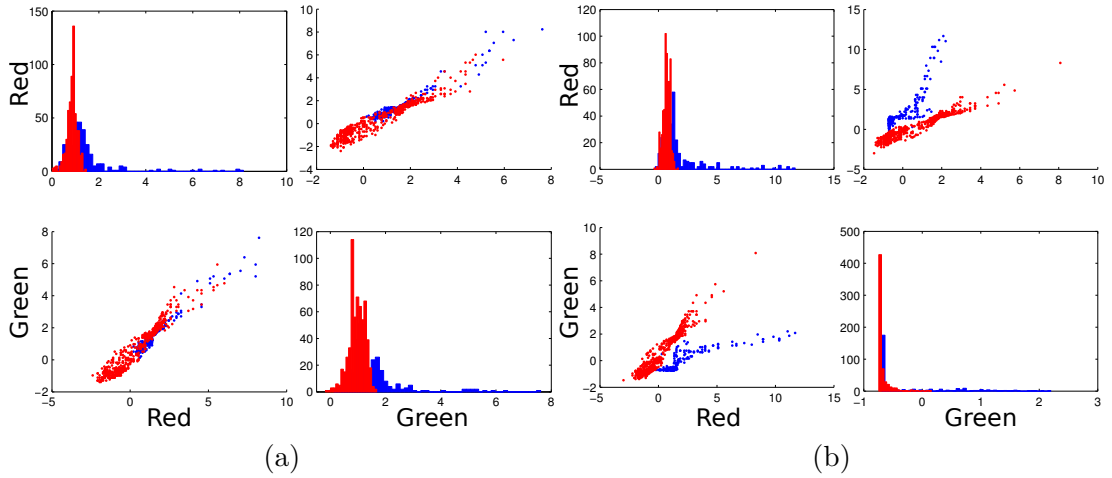


Figure 6.11: Increased separability of similar classes by the inclusion of morphological features - The separability of “Bright building” and “Roads” classes, for the image acquired in 2006, results increased. Specifically, for the green and red spectral channels, a lower within class variance is accompanied by a better separation of the two data clouds.

structuring element is again proportional to the size of the object of interest. The OC and OCR features, both composed by 24 variables per date, are visualized in Figure 6.8 and Figure 6.9 respectively, while their signal is shown in Figure 6.10(e) and Figure 6.10(f).

To allow fair comparisons between DMC and DIA, where unchanged pixels are treated as single class, a third approach referred to as reduced DMC is considered: all the samples representing unchanged classes are assigned to the class unchanged, and the change detection is performed as for the standard DMC scheme.

To better understand the role of the spatial-contextual information within the process of supervised change detection, blocks of features and their combinations are tested independently and in growing order. Furthermore, for each feature block, eight experimental conditions are tested, accounting for different sizes of the training sets: 5, 10, 20, 50, 100 and 200 labelled examples per class randomly extracted from the available training ground truth. The size of the sets varies from very small to large, and for the smaller ones the number of dimensions can be larger than the one of training samples (e.g. the Brüttsellen OC set accounts for 56 multi-temporal features and just 45 training samples for 9 classes in the smallest complete DMC setting). Classification results are consequently very sensitive to the representativeness of training set. To provide robust statistical estimates, results are averaged on 10 independent experiments.

SVM hyperparameters are selected by a 3-fold cross-validation. The C parameter is selected by exhaustive search in the range $\{1, 10, 20, \dots, 1000\}$. To mitigate overfitting, in particular for small training sets, an initial guess on the Gaussian kernel bandwidth σ_p has been obtained by computing the median distance on 3000 randomly chosen coordinates for the considered dataset. A refined search around this initial guess, in $\{0.5\sigma_p, \sigma_p, 1.5\sigma_p\}$, has been performed and the parameters producing minimal error were retained. The free

6. Supervised change detection

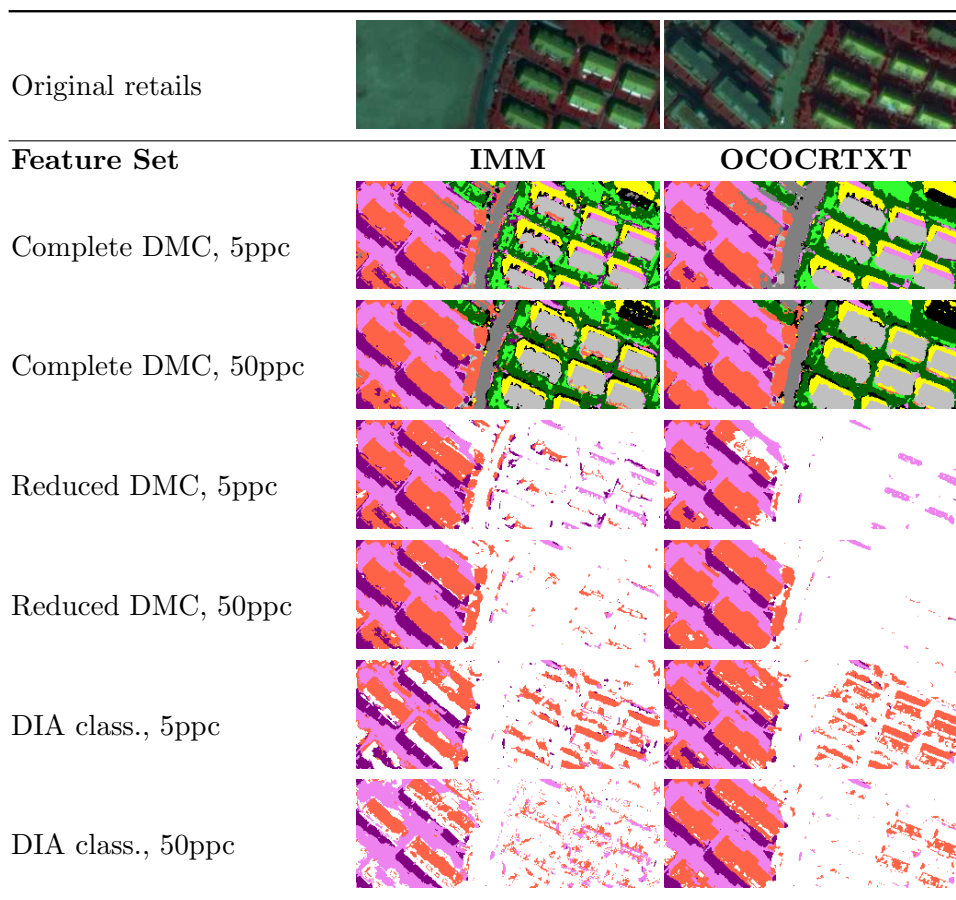


Figure 6.12: Details of the Brüttisellen change detection maps - Note that on the left column of the maps, the IMM set has been used, while on the right one the OCOCRTXT provided the maps. For the legend please refer to Appendix B.

Torch 3 library has been used to solve the SVM optimization [Collobert et al., 2002].

The generalization accuracy is evaluated in terms of estimated Cohen’s Kappa statistic (κ) [Foody, 2004]. To assess the significance of differences in accuracy, the McNemar test is reported in Table 6.15 (see Appendix A). This table shows if the average accuracy is significantly higher (+), lower (-) or statistically similar (o) to the one obtained using the pure spectral baseline set (IMM).

6.3.5 Results

Brüttisellen dataset results The accuracies for the Brüttisellen experiments are reported in Figure 6.13(a),(c),(e) as a function of the per class training set size.

The complete DMC shows very good classification performances saturating around a $\kappa = 0.9$, in particular for the composite textural and morphological spatio-spectral sets. The sets showing the lowest accuracy are the pure spectral and spectral-textural,

6.3 Supervised change detection for urban monitoring

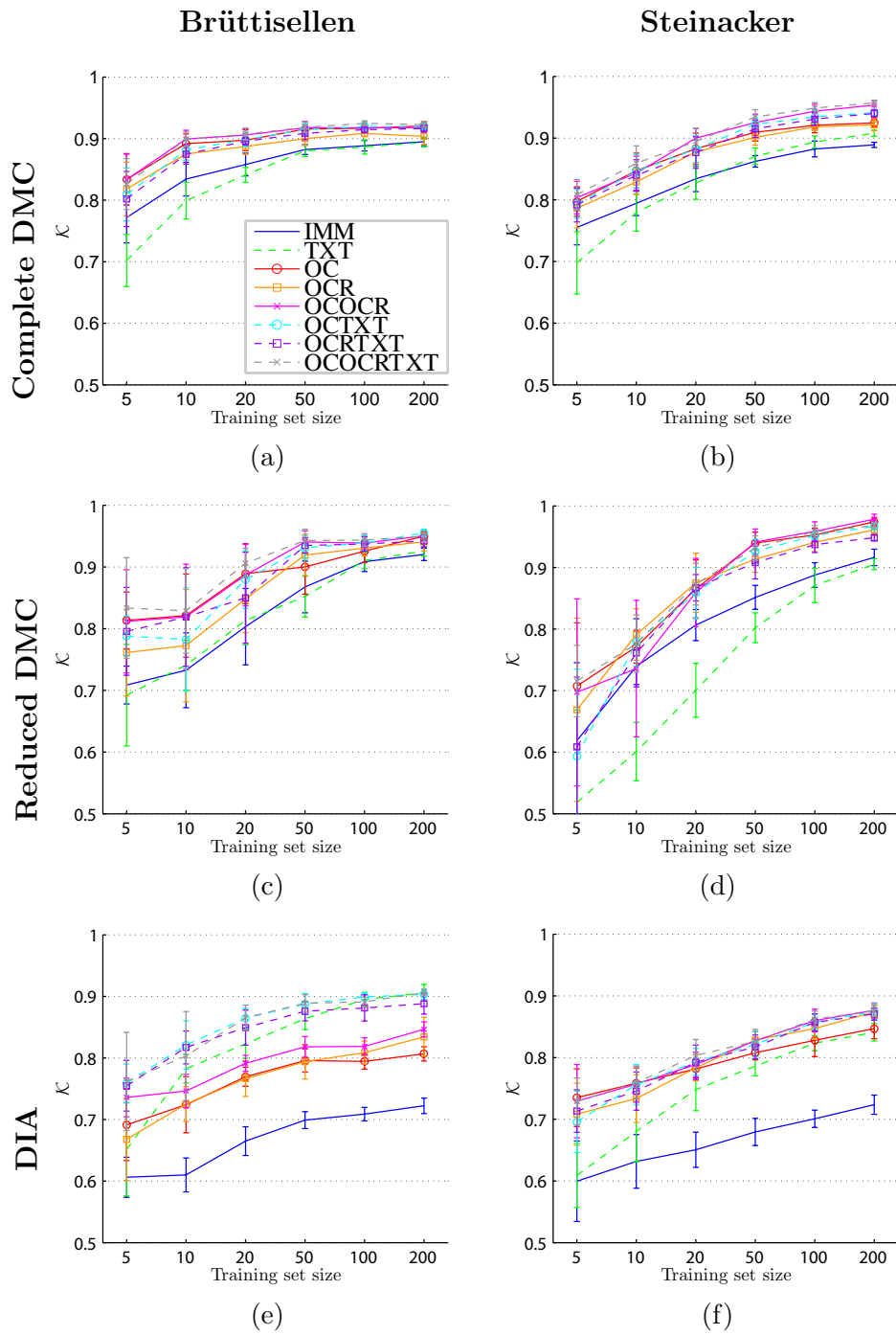


Figure 6.13: Test accuracies for urban monitoring datasets - Test accuracies as a function of the per class training set size: Brüttisellen (a),(c),(e) and Steinacker (b),(d),(f), illustrating the complete DMC, reduced DMC and DIA accuracies for the tested input spaces, respectively.

6. Supervised change detection

indicating that texture alone does not help in discriminating all the classes. The McNemar test reported in Table 6.15 indicates that, except for the TXT, all the contextual features improve significantly the DMC results without contextual information.

The reduced DMC shows similar trends. It is worth mentioning that, since the number of classes is different (4 instead of 9), no direct comparisons on the absolute accuracies observed above can be made. In particular, note that classification errors within the unchanged classes are removed. Again, the baseline IMM set performs worse than the others, with a very close performance of the TXT set. The contextual information improves significantly the accuracies, as depicted in Table 6.15.

Regarding DIA, different observations can be made. As in the previous experiments, the pure spectral IMM feature set performs significantly worse than the others. The three morphological sets (OC, OCR and OCOCR) show similar κ scores and standard deviations, and provide accuracies from 0.7 to 0.8 κ as the training set size increases. The best approaches are again the composite textural-morphological, improving significantly the classification provided by morphological sets. The texture seems an important information to mitigate the ambiguity of the spectral change vector representations and, if combined to other measures, reduces greatly the false alarm rate. In this case, the TXT set accuracy grows rapidly to the performance of the most accurate feature sets. By comparing the reduced DMC and the DIA schemes it appears clearly that the difference in accuracy of 0.03-0.07 κ points is related to richness of the multi-temporal signal, preserved in the former. On the other hand, even if the accuracy provided by the latter is lower the dimensionality of the dataset is the half, mitigating issues related to low sample conditions.

Figure 6.12 compares the classification maps obtained by the pure spectral input set (IMM) and morphological composite OCOCRTXT sets for 5 and 50 samples per class. This last size has been chosen since a plateau effect on the accuracy is observed (see Figure 6.13). The change detection maps show an improved spatial coherence when adding contextual information, and, when the training sets better represent the variance of the class, higher accuracies are obtained.

Steinacker dataset results As observed for the previous dataset, the complete DMC performances of IMM and TXT sets are significantly lower than the other tested feature. The sets providing the most accurate results are those composed by the mixed textural-morphological and spectral information. For training sets larger than 20 samples per class, standard deviations are very low, indicating stable classification models. As for the previous dataset, Table 6.15 confirms the significance of the improvements in change detection accuracy when adding spatial information.

In the reduced DMC setting, the TXT feature set provides the worse results (significantly worse than the IMM features) for each training set size considered. The baseline IMM block performs in the range of the other sets when considering 5 and 10 examples per class, then worsen from 20 samples per class on. The best accuracies are again obtained by models that include contextual information.

Regarding DIA setting, trends are similar to those observed for the previous dataset.

6.3 Supervised change detection for urban monitoring

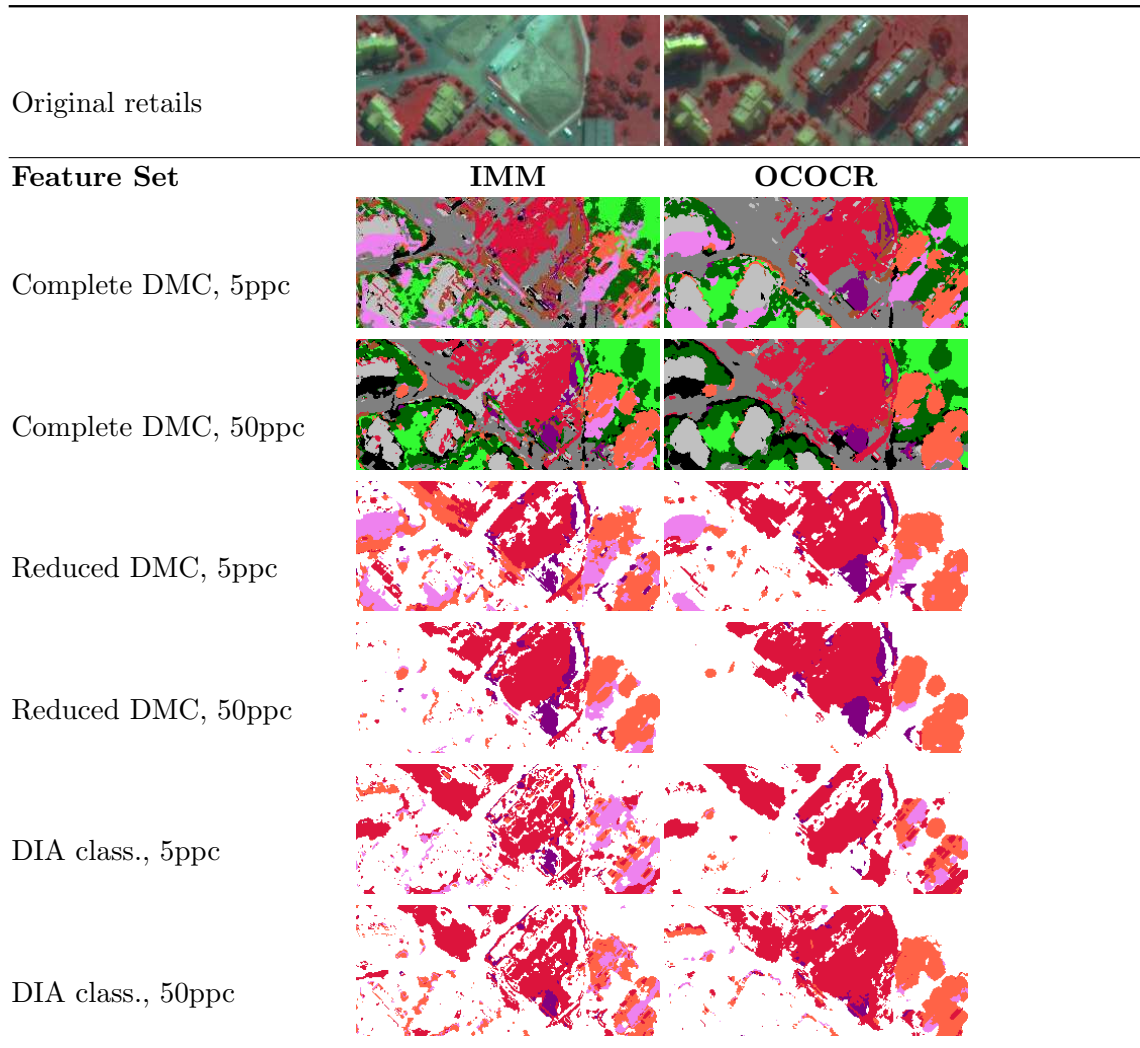


Figure 6.14: Details of the Steinacker change detection maps - Note that on the left column of the maps, the IMM set has been used, while on the right one the OCOCR provided the maps. For the legend please refer to the Appendix B.

The IMM set performs constantly worse than the rest and the TXT set increases to the best accuracies when adding training samples. All tested variables, except TXT with 5 training samples per class, are significantly better than the pure IMM information. Morphological-textural composite sets behave very similarly, indicating again the appropriateness of this information for the DIA setting. As for the previous experiments, the differences between reduced DMC and DIA are related to the loss in information when adopting the difference image, in contrast to all preserved information for DMC schemes. Figure 6.12 reports the change detection maps produced with training sets of 5 and 50 samples per class. The spatial coherence of the basic spectral change detection map is again greatly improved by the inclusion of morphological contextual information (OCOCR).

6. Supervised change detection

Method	Complete DMC						Reduced DMC						DIA							
	per class size	5	10	20	50	100	200	5	10	20	50	100	200	5	10	20	50	100	200	
Brüttsellen	TXT	-	-	-	o	o	o	o	+	+	-	o	+	+	+	+	+	+	+	
	OC	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
	OCR	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
	OCOCR	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
	OCTXT	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
	OCRTXT	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	OCOCRTXT	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Steinacker	per class size	5	10	20	50	100	200	5	10	20	50	100	200	5	10	20	50	100	200	
	TXT	-	-	-	+	+	+	-	-	-	-	-	o	o	+	+	+	+	+	
	OC	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
	OCR	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	
	OCOCR	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
	OCTXT	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	
	OCRTXT	+	+	+	+	+	+	o	+	+	+	+	+	+	+	+	+	+	+	
OCOCRTXT	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		

Figure 6.15: Outcomes of the McNemar tests for the urban monitoring datasets
- The + indicates that the tested set of features is significantly better than the baseline IMM set with $z > 1.96$ at $\alpha = 0.05$ level, while - indicates that IMM is better than the compared approach $z < -1.96$. The o indicates no significant difference.

6.3.6 Discussion

The experiments on the VHR multi-temporal datasets provided interesting insights about the inclusion of spatial context information in the process of supervised change detection. From Table 6.15, it is clear that considering such information significantly improves the accuracy of the process in the most of the tested settings.

The complete DMC setting has the advantage of predicting a full map by considering each stable and transitional class. If the ground truth has been created carefully, the different classes are well-defined and separability is increased by including spatial information. The usefulness of the pixel context is also beneficial for obtaining smooth change detection maps, eliminating spurious changes and salt-and-pepper noise, while reducing the false alarm rate, as shown in the change map details in Figure 6.12 and Figure 6.14.

Regarding the reduced DMC setting, performance is even higher thanks to the easier classification problem given by the lower number of class, due to the aggregation of all the permanent land covers into a single class. However, problems may arise when the training sets are small, as underlined by the corresponding high variances of the κ score. This is mainly due to the multi-modal distribution of the unchanged class, becoming sparse and clustered in the high dimensional input space. As a consequence, even a robust method such as the SVM needs many training samples to discover correct separating hyperplanes.

For the DIA approach it can be noticed that the inclusion of composite contextual information is always beneficial, reducing the effects of ambiguity and increased class overlapping. The comparisons with the reduced DMC scheme suggest that DIA can pro-

vide high accuracies by utilizing only textural information, allowing the use of simpler classifiers due to the lower dimensionality of the dataset. The increased separability when considering pixel context is evident.

When only few samples composed the training set, the dimensionality was often higher than the number of samples. Even if SVM are robust to the curse of dimensionality [Hughes, 1968; Trunk, 1979], one has to control the n_s/d ratio (number of samples / dimensions) by providing enough examples to model correctly the class boundaries. In our experiments a n_s/d ratio lower than 0.6 - 0.7 provided the less stable solutions. This fact is underlined by the decrease of the standard deviation for larger training sets, indicating stable models. However, note that the most of the considered set sizes were too small for many classifiers (e.g. Fisher's discriminant). Hence, SVM classifiers are strongly recommended due to their robustness against the curse of dimensionality.

6.4 Conclusions

As discussed in this Chapter, kernel-based supervised change detection is a stable and effective way to obtain exhaustive and very accurate maps. Considering both persistent and transitional classes as for the Zurich case study or by looking for a semantically coherent map of changes, such as depicted in the James River case study, both the nonlinear kFDA and SVM provided very good results. They showed robustness to high dimensional input spaces, in particular for the SVM case study, thanks to the implicit kernel mapping and the possibility to control the capacity of the classifier.

As it has been observed, the training sample selection issue must be addressed carefully. While for pansharpened VHR images (GSD of roughly 0.6[m]) the user might be able to correctly label pixels by photointerpretation, when using medium resolution images even the most trained and experienced user might fail in correctly assigning labels to pixels. In this case, terrain campaigns are needed. For the case involving the urban monitoring task, the addition of discriminant features acted as an additional regularization, penalizing spatial variability and noise, greatly improving the class discrimination process. However, adequate training sets must be provided: they should be large enough to be representative of the class distribution and to provide to the SVM an appropriate number of candidate support vectors. In this sense, many pixels coming from a large homogeneous region (further smoothed by spatial filters) would likely provide only a small fraction of support vectors, while samples coming from more ambiguous and spatially varying areas are likely to possess more information about the geometrical limits of the correct class separation, in particular if those possess high variance as in VHR images [Foody and Mathur, 2004]. In this sense, active learning may be an interesting solution. It is a family of iterative sampling schemes that, on the basis of the classifier confidence, return some unlabelled samples to the user asking for the label. Then, the model is retrained with the largest set, until some stopping criterion is met [Tuia et al., 2011; Volpi et al., 2012b]. These schemes are promising for the data classification, and there is no apparent constraint for their application in supervised change detection and multi-temporal classification scenarios.

6. Supervised change detection

For the flood mapping scenario, it resulted clearly that the precision of supervised classifier is often counterbalanced by the difficulty of obtaining exhaustive training sets. However, in many change detection studies one may not be interested in exhaustive maps, but only in a binary discrimination of the type “change”-“no change”. In the flood mapping example, labels only referred to easily discriminable classes: water at the two dates and general unchanged pixels. This latter class suggested the adoption of the kernel extension of the FDA, to be able to work with normally distributed classes in the RKHS, and therefore covariance based operations are effective.

In both cases, the difficulty in correctly labelling pixels appeared evident. For many change detection applications assuming available labelled information prior to the analysis is often an unrealistic assumption. In the next Chapter, a completely automatic and unsupervised approach to change detection is presented, with the aim of contributing to applications where the readiness of the system is of paramount importance.

Chapter 7

Unsupervised change detection. Automatic clustering the difference image in the RKHS¹

This Chapter presents an unsupervised and automatic approach to non-linear change detection. First, the kernel-based clustering method is briefly introduced in Section 7.2. Then, the elements composing the change detection algorithm, i.e. initialization and hyperparameters optimization, are illustrated in Section 7.2.2 and Section 7.2.3 respectively. The feature maps studied to perform nonlinear clustering are presented in Section 7.2.4. Section 7.3 presents the experimental setup, while Section 7.4 reviews three case studies involving SPOT, QuickBird and Landsat TM exploited to validate the proposed change detection scheme. Finally, Section 7.5 draws the conclusions.

7.1 Clustering for automatic change detection

As summarized in the state-of-the-art review in Chapter 5, many efforts in change detection research are put into unsupervised methods, which require no or minimal user intervention in the process. As illustrated in the previous Chapter 6 for supervised multi-temporal classification, the collection of ground truth samples allowing to correctly train a model is often very difficult to be properly carried out. In particular, for many change detection applications related to catastrophes and natural hazards, it is unrealistic to assume the availability of ground truth samples. Moreover, to precisely define transitions in a supervised context, ground truth samples need to be spatially registered. While it may be possible to perform terrain campaigns for the most recent acquisition, information about the landcover prior to the event under study is usually not available.

¹This Chapter is based on the following publication: [Volpi et al., 2012b]. See Section 1.3.2 for the details.

7. Unsupervised change detection

Therefore, many recently developed automatic change detection systems rely on clustering algorithms to partition the multi-temporal data into changed and unchanged regions. In this Chapter, we introduce an unsupervised approach to change detection relying on kernels. Kernel k -means clustering is used to partition a selected subset of pixels of the image, representing changed and unchanged areas with high probability. Once the optimal clustering is obtained, the estimated representatives (or centroids) of each group are used to assign all the others pixels composing the multi-temporal scenes to their class. We review different ways to encode the multi-temporal information and, in particular, we show the superiority of computing the difference image directly in the RKHS by adopting a difference kernel approach [Camps-Valls et al., 2008]. Moreover, we propose an effective way to cope with the estimation of the hyperparameters of the kernel function (e.g. Gaussian RBF bandwidth) in a completely unsupervised way. Experiments on three datasets (a very high, a high and a medium resolution image) validate the proposed system.

7.2 The proposed unsupervised kernel-based change detection scheme

The proposed scheme relies on three different steps: (i) initialization, (ii) estimation of the kernel parameters and clustering, and (iii) final assignment of the pixels to their classes.

7.2.1 A partitioning algorithm: the kernel k -means

The kernel k -means partitioning (KkM) extends the standard linear k -means [MacQueen, 1967] to higher dimensional RKHS denoted as \mathcal{H} . Based on the criteria discussed in Chapter 4, the k -means formulation can be expressed solely in terms of inner products between samples, and consequently kernel functions can replace them as $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = k(\mathbf{x}_i, \mathbf{x}_j)$. In clustering problems, we do not dispose of labels during training: the learning set is composed by $\{\mathbf{x}_i\}_{i=1}^{n_s} \in X$ only.

Let k denote the total number of desired clusters. The kernel k -means algorithm can be formulated as the minimization of the loss function corresponding to the sum of squares of the distance between mapped samples $\phi(\mathbf{x}_i)$ in cluster c , denoted as X_c , to their mean in \mathcal{H} , $\boldsymbol{\mu}_c^{\mathcal{H}} = \frac{1}{n_c} \sum_{\mathbf{x}_i \in X_c} \phi(\mathbf{x}_i)$ [Girolami, 2002]:

$$\mathcal{L}(\phi(\mathbf{x}_i), k) = d^2(\phi(\mathbf{x}_i), \boldsymbol{\mu}_c^{\mathcal{H}}) = \sum_{c=1}^k \sum_{\mathbf{x}_i \in X_c} \|\phi(\mathbf{x}_i) - \boldsymbol{\mu}_c^{\mathcal{H}}\|^2. \quad (7.1)$$

As illustrated in Chapter 4, a distance in the RKHS can be expressed using kernels. In this case, the distance to the mean of group c is:

$$\begin{aligned} d^2(\phi(\mathbf{x}_i), \boldsymbol{\mu}_c^{\mathcal{H}}) &= \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_i) \rangle - \frac{2}{n_c} \sum_{\mathbf{x}_j \in X_c} \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle + \frac{1}{n_c^2} \sum_{\mathbf{x}_j, \mathbf{x}_l \in X_c} \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}_l) \rangle \\ &= k(\mathbf{x}_i, \mathbf{x}_i) - \frac{2}{n_c} \sum_{\mathbf{x}_j \in X_c} k(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n_c^2} \sum_{\mathbf{x}_j, \mathbf{x}_l \in X_c} k(\mathbf{x}_j, \mathbf{x}_l), \end{aligned} \quad (7.2)$$

7.2 The proposed unsupervised kernel-based change detection scheme

This problem may be solved iteratively, by alternating the computation of $d^2(\phi(\mathbf{x}_i), \boldsymbol{\mu}_c^{\mathcal{H}})$ and then reassigning the samples to the closest center, until convergence [MacQueen, 1967]. Since the coordinates of $\phi(\mathbf{x}_i)$ are used implicitly and are not explicitly known, we can not obtain the true cluster center. We approximate the centroid as the sample of cluster c closest to its true center in the feature space $\boldsymbol{\mu}_c^{\mathcal{H}} \in \mathcal{H}$, i.e. $\mathbf{x}_c = \arg \min_{\mathbf{x}_i \in X_c} d^2(\phi(\mathbf{x}_i), \boldsymbol{\mu}_c^{\mathcal{H}})$.

After the final centroids \mathbf{x}_c are obtained, a new unseen sample \mathbf{x} is assigned to the cluster c whose centroid is the closest, as $\arg \min_c d^2(\phi(\mathbf{x}), \phi(\mathbf{x}_c))$. This last expression can be seen as a kernel minimum distance classification, with the class representatives estimated by the kernel k -means. As for many iterative partitioning clustering algorithms, the initialization of the cluster centroids strongly affects the convergence to the global minimum of the cost function. If the algorithm is initialized in a suboptimal manner, it may happen that the convergence is reached at a local minima of the cost function. In the next Section, we present a strategy able to initialize in a robust manner the partitioning.

7.2.2 The initialization

In an optimal situation, cluster centers should be initialized close to the true representatives of the group structure of samples. In a supervised context, empirical estimates of such representatives can be obtained by computing the class average or mode. Since we do not dispose of label information, a suboptimal partitioning of the multitemporal image allowing a correct unsupervised selection of training samples – a pseudo-training set – is adopted. This way, information on the change detection problem can be included to alleviate the issue of bad initializations. This procedure returns a set of pixels expected to belong to the two classes (change and no-change, respectively y_m and y_l) with high confidence. This is performed by selecting two subsets of the original bi-temporal data on the basis of the statistical distribution of the difference image magnitude. Using this pseudo-training set, the centroids of the clusters are estimated using the kernel k -means, used in the following to partition the rest of the bi-temporal images.

Let X^{t_1} and X^{t_2} be the $n \times d$ (n pixels and d spectral channels) coregistered and radiometrically matched images at times t_1 and t_2 . The magnitude of the i th pixel is computed on the basis of the ℓ_2 -norm of the d -dimensional difference image \mathbf{D} as introduced in Section 2.3.1. In this unidimensional representation, low values (ideally near 0) correspond to unchanged pixels, while large values (usually larger than a given threshold) correspond to pixels whose radiometric differences indicate a change between the two acquisitions. This distribution can be approximated by a mixture of two univariate Gaussian distributions [Bruzzone and Fernández-Prieto, 2000], as $p(\delta) = p(y_m)p(\delta|y_m) + p(y_l)p(\delta|y_l)$, whose parameters can be estimated using the Expectation-Maximization algorithm [Dempster et al., 1977]. Since image noise, differences in illumination and in particular outliers can affect the tails of the magnitude distribution (e.g. the tail of $p(\delta|y_m)$ may represent false changes related to saturated pixels), we choose the pseudo-training samples in the most dense regions of the distribution, by selecting a threshold proportional to the standard deviation around the means of the components of the bimodal distribution. The sketch of

7. Unsupervised change detection

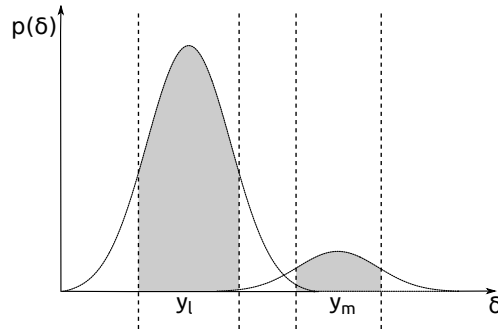


Figure 7.1: Magnitude-based initialization of the kernel k -means - Threshold of the magnitude of the difference image to obtain a raw initialization of samples belonging to changed areas y_m and unchanged zones y_l .

the proposed initialization is illustrated in Figure 7.1 .

To lighten the computational load, a random subset sufficiently large of the selected regions of $p(\delta)$ can be chosen without losing the representativeness. The KkM applied to this subset returns the centroids of the samples closest to the cluster mean in the RKHS. As mentioned in the previous Chapter, and as illustrated in [Bach and Jordan, 2002b; Cremers et al., 2003], the use of a Gaussian RBF kernel makes the assumption of Normality in the RKHS consistent. In this case, samples closest to the mode of the cluster in RKHS will show a kernel value close to one, while for tails of the cluster, such value decreases. Ideally, the similarity between modes has to be zero. Therefore, errors and noise included in the pseudo-training set, if they are only a fraction of the total number of samples, should be placed in the tails of the distribution. The KkM should return centroid consistent with the densities of problem at hand, provided a good hyperparametrization. Since the final class assignment is performed in the same RKHS the retrieved centroids are still adequate for the successive classification.

7.2.3 The unsupervised cost function

As for all the kernel-based algorithms, the choice of the kernel hyperparameters plays a central role for the success of the method. When dealing with labelled data, i.e. in a supervised framework, the parameters can be estimated by minimizing an error function over a given subset for example by adopting leave-one-out or cross-validation estimations, as introduced in Chapter 3. In unsupervised problems, as the one considered here, the issue of fitting hyperparameters is usually addressed by expert knowledge or by trial and error. To avoid such a heuristic strategy and to obtain an objective and data driven solution, we propose to fit the kernel hyperparameter(s) Θ_h by optimizing a geometrical criterion. Such a function favours mappings enhancing geometrical configurations adapted to the partitioning task, which in the case of KkM corresponds to the definition of far clusters (high between-cluster distance) showing low within-cluster variance. We propose to minimize the difference between the average within-cluster distances from each center and the between-cluster distance in the RKHS.

7.2 The proposed unsupervised kernel-based change detection scheme

As illustrated in the previous Chapter 6 for the kernel Fisher’s discriminant classifier, this situation is optimal to discover the class structure in the RKHS. In this case, no labelled information is available to extrapolate such measure from the data at hand. By exploiting the partitioning provided by the KkM, the cluster centroids are used to evaluate this geometrical loss. The set of kernel hyperparameters Θ_h satisfying the following relationship are retained:

$$\arg \min_{\Theta_h} \frac{1}{n} \sum_{c=1}^k \sum_{\mathbf{x}_i \in X_c} d^2(\phi(\mathbf{x}_i), \boldsymbol{\mu}_c^{\mathcal{H}}) - \sum_{c \neq q} d^2(\boldsymbol{\mu}_c^{\mathcal{H}}, \boldsymbol{\mu}_q^{\mathcal{H}}). \quad (7.3)$$

Equation (7.3) is evaluated by subsequent runs of the KkM using the same initial pseudo-training samples and varying the set of kernel free parameters Θ_h . The hyperparameters minimizing the above expression are retained, and the corresponding centroids are used to partition the multi-temporal dataset into changed and unchanged classes. Note that the KkM maximizes by definition the distance between the cluster centers, and, since the partitioning is performed using the isotropic distance function, an hypersphere defines the labellings of the samples closest to their center of mass. The optimal situation for clustering is obtained when finding the kernel hyperparameter that offer an optimal trade-off between the clusters compactness, i.e. the average of the hypersphere radius, and the maximization of the distance between the centroids. By analyzing the above formulation, underfitting may be defined as the situation in which the implicit map performed by the kernel function projects samples into a space overestimating the similarities and resulting distances are null, e.g. all the samples are equally similar in \mathcal{H} (e.g. using very large σ for a Gaussian RBF). On the contrary, a situation in which samples are similar only to themselves is likely to provide a feature space in which clusters are not separable. In this case, all the samples are scattered around their mean which are mutually superimposed. This situation is likely to be provided by a too small σ parameter.

Note that, even if exact coordinates of $\boldsymbol{\mu}_c^{\mathcal{H}} \in \mathcal{H}$ are not retrievable, exact distances between two mapped samples and between the two centers can be obtained, as illustrated by Equation 7.2 and as discussed in Section 4.2.2. It is worth observing that the proposed cost function is independent on the form of the kernel function adopted, and the geometrical tenet holds for different kernels, their combinations and multiple parameters. However, the use of Gaussian RBF kernels should enforce the properties of the KkM in the RKHS, that is, the cluster normality. In its linear version, the k -means is globally optimal only when samples are generated according to Gaussian distributions.

7.2.4 Feature maps

In addition to the kernel parameters, the temporal information must be correctly encoded to detect changes accurately. In this Section we present two kinds of feature maps used for automatic change detection: the first correspond to mapping into the RKHS the standard difference image, while the second defines the difference image directly into RKHS.

7. Unsupervised change detection

Difference image in the input space. To perform this mapping, images are first subtracted pixelwise to obtain $\mathbf{D} \in \mathcal{X}$ (see Section 7.2.2). The difference pixels \mathbf{x}^d are mapped to \mathcal{H} as $\phi(\mathbf{x}) = \phi(\mathbf{x}^d)$. This approach aims at defining a multivariate threshold, similarly to CVA, that discriminates changes (linearly or not linearly depending on the type of mapping function used) on a linear combination of the images in their input spaces known to emphasize changes occurred between the two acquisitions (see Section 2.3.1). Although this approach is widely used, nonlinear relationships hidden in single images and in the bi-temporal dataset cannot be discovered and correctly modelled. Problems related to the ambiguity of the difference image can affect the process, since the same difference values may be either related to actual processes occurred on the ground or to radiometric differences not related to land cover transitions.

Difference image in the feature spaces. This feature map is built explicitly to account for linear and nonlinear dependencies between the single and the bi-temporal pixels. This mapping function computes the difference image in the higher dimensional feature space, known to enforce linear relationships among the different structures in the data. The RKHS feature vector $\phi(\cdot)$ corresponding to the difference pixel induced by (possibly different) mappings of uni-temporal pixels $\varphi(\cdot)$ can be defined, for a given sample \mathbf{x}_i , as:

$$\phi(\mathbf{x}_i) = \mathbf{H}^{(t_2)}\varphi(\mathbf{x}_i^{(t_2)}) - \mathbf{H}^{(t_1)}\varphi(\mathbf{x}_i^{(t_1)}), \quad (7.4)$$

where $\mathbf{H}^{\{t_1, t_2\}}$ are positive and symmetric projection matrices to match the feature mappings. Then, the similarity of two difference vectors in feature spaces $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$ is evaluated by $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. By solving the inner product with Equation (7.4), and exploiting the closure properties introduced in Section 4.2.4, we obtain the corresponding kernel evaluating the similarity among pixels composing the difference image in the RKHS:

$$\begin{aligned} k(\mathbf{x}_i, \mathbf{x}_j) &= \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle \\ &= \left\langle \left(\mathbf{H}^{(t_2)}\varphi(\mathbf{x}_i^{(t_2)}) - \mathbf{H}^{(t_1)}\varphi(\mathbf{x}_i^{(t_1)}) \right), \left(\mathbf{H}^{(t_2)}\varphi(\mathbf{x}_j^{(t_2)}) - \mathbf{H}^{(t_1)}\varphi(\mathbf{x}_j^{(t_1)}) \right) \right\rangle \\ &= \varphi(\mathbf{x}_i^{(t_2)})'\mathbf{H}'^{(t_2)}\mathbf{H}^{(t_2)}\varphi(\mathbf{x}_j^{(t_2)}) + \varphi(\mathbf{x}_i^{(t_1)})'\mathbf{H}'^{(t_1)}\mathbf{H}^{(t_1)}\varphi(\mathbf{x}_j^{(t_1)}) \\ &\quad - \varphi(\mathbf{x}_i^{(t_2)})'\mathbf{H}'^{(t_2)}\mathbf{H}^{(t_1)}\varphi(\mathbf{x}_j^{(t_1)}) - \varphi(\mathbf{x}_i^{(t_1)})'\mathbf{H}'^{(t_1)}\mathbf{H}^{(t_2)}\varphi(\mathbf{x}_j^{(t_2)}) \\ &= \varphi(\mathbf{x}_i^{(t_2)})'\mathbf{H}'^{(t_2)}\varphi(\mathbf{x}_j^{(t_2)}) + \varphi(\mathbf{x}_i^{(t_1)})'\mathbf{H}'^{(t_1)}\varphi(\mathbf{x}_j^{(t_1)}) \\ &\quad - \varphi(\mathbf{x}_i^{(t_2)})'\mathbf{H}^{(t_2, t_1)}\varphi(\mathbf{x}_j^{(t_1)}) - \varphi(\mathbf{x}_i^{(t_1)})'\mathbf{H}^{(t_1, t_2)}\varphi(\mathbf{x}_j^{(t_2)}) \\ &= k(\mathbf{x}_i^{(t_2)}, \mathbf{x}_j^{(t_2)}) + k(\mathbf{x}_i^{(t_1)}, \mathbf{x}_j^{(t_1)}) - k(\mathbf{x}_i^{(t_2)}, \mathbf{x}_j^{(t_1)}) - k(\mathbf{x}_i^{(t_1)}, \mathbf{x}_j^{(t_2)}) \end{aligned} \quad (7.5)$$

Kernel functions composing the above expression can be of different nature and form, since no restriction has been put on $\varphi(\cdot)$. The difference kernel needs the estimation of the corresponding parameters (e.g. 4 bandwidths when using 4 RBF kernels). The cost function proposed in Section 7.2.3 depends only on the cluster assignments and can be used directly to estimate multiple parameters of different kernels. However, the kernels composing Equation (7.5) can be grouped in two categories, uni-temporal kernels –

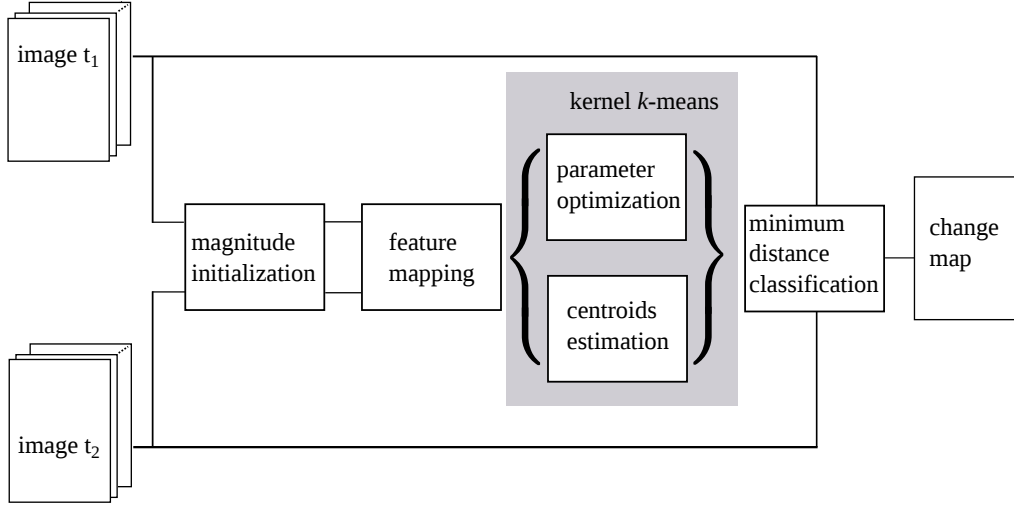


Figure 7.2: The block diagram of the proposed change detection scheme.

$k(\mathbf{x}_i^{(t_1)}, \mathbf{x}_j^{(t_1)})$ and $k(\mathbf{x}_i^{(t_2)}, \mathbf{x}_j^{(t_2)})$ – and cross-time kernels – $k(\mathbf{x}_i^{(t_1)}, \mathbf{x}_j^{(t_2)})$ and $k(\mathbf{x}_i^{(t_2)}, \mathbf{x}_j^{(t_1)})$ – their parameters are assumed to be shared and in the experiments reported only a search among 2 kernel hyperparameters $\Theta_h = \{\theta_h^{\text{single}}, \theta_h^{\text{cross}}\}$ is performed.

7.3 Experimental setup

Three multi-temporal images are considered for testing the unsupervised scheme: The Gloucester flooding, the Brüttisellen 2 and the Greek islands subsets, detailed in Appendix B. In Figure 7.2 the main steps of the proposed system are summarized.

In order to test the sensitivity of the proposed approach to initial conditions, different initial pseudo-training sets have been considered by sampling different quantities of pixels to define y_m and y_l . After experimental evaluation, we report results only on a single set size, since many pseudo-training samples can be obtained at zero cost once the thresholds are estimated, and a plateau effect on the accuracy was observed for sets larger than the ones considered, for each multi-temporal pair of images. It is recommended to sample a balanced number of pixels to cover data variability but also to allow fast computations, regulated by the computation and storage of the kernel matrix. For all the nonlinear cases, Gaussian RBF kernels were adopted. RBF bandwidths are optimized in the interval in $\{0.1, 0.2, \dots, 10\}$ using the cost function defined in Equation (7.3). In order to have robust statistical estimates of the accuracy, 10 runs of each experiment have been performed (each one considering a different realization of the pseudo-training set over the modes $p(\delta)$). The average of the skill scores and its standard deviation are reported. For each scene, a ground truth has been visually extracted in order to validate the outcomes of the change detection schemes. Overall Accuracy (OA), estimated Cohen’s κ statistic [Foody, 2004], ROC Curve, Area Under the ROC Curve (AUC) [Fawcett, 2006] and adjusted Rand index (AR) [Rand, 1971] are used as figures of merit (see Appendix A for a list). Change maps

7. Unsupervised change detection

are produced as the sum of the clustering outcomes, with values for each pixel ranging from 1 to 10. The colour ranges from black (the pixel was never classified as changed), to white (the pixel was always classified as changed). Then, from darker to brighter colours, it indicates how many times the pixels have been detected as changed (e.g. purple means 1 out of 10, orange 5 out of 10).

The proposed approaches are tested versus the linear counterpart of the considered mappings (both resulting in standard k -means on the difference image) providing a baseline accuracy, and against two automatic change detection methods: the standard CVA [Bovolo and Bruzzone, 2007] and the approach presented in [Celik, 2009a]. The former puts a threshold in the magnitude distribution as in [Bruzzone and Fernández-Prieto, 2000]. The latter relies on a patch-based PCA transformation of the difference image of the intensities followed by standard binary k -means. Additionally, tests using the fully supervised SVM classifier introduced in Section 6 are also provided, defining a best possible scenario, with models trained with the same number of samples used for testing the proposed KkM but coming from pre-defined ground truth regions. Since the approach of [Celik, 2009a] is designed for univariate intensity images, an investigation to select the best unidimensional representation of changes has been carried out: among single band differences and the magnitude, the latter resulted in higher accuracies and has been used in the experiments.

7.4 Results and experimental validation

7.4.1 Case studies

The Gloucester flooding (DFC dataset). The pseudo-training sets are composed by 500 randomly selected pixels, 250 per mode of the magnitude distribution. By observing figures of merit reported in Table 7.1, the *Diff. Lin.* approach shows a relatively high κ value coupled to the lowest standard deviation, indicating that it is the most stable approach. It also exhibits a high AUC value, suggesting a low missed detections rate, confirmed by the ROC curves in Figure 7.4(a). The nonlinear *Diff. RBF* accuracy suggests that nonlinearly cluster the difference image did not improve significantly the change detection process, if compared to the *Diff. Lin.* The *Ker. Diff. RBF* approach is the most accurate and it illustrates clearly the improvements when considering the difference image representation in the feature spaces. The κ score increased by 0.065 with respect to the *Diff. RBF*. Standard CVA and the approach from [Celik, 2009a] provided the lowest accuracies, caused respectively by the high false alarm rate and by the weak detection rate. Supervised SVM outperformed all considered change detection approaches, except for *Ker. Diff. RBF*, which shows equal accuracy but higher standard deviations. It underlines the good performances of the change detection computed in the difference image in RKHS, without exploiting any label.

Change maps are reported in Figure 7.3, Gloucester (a)-(f), corresponding to the sum of the 10 independent binary maps. By observing the change maps, the *Ker. Diff. RBF* is characterized by the lowest false alarm rates, making it the most accurate approach, as il-

7.4 Results and experimental validation

	Diff. Lin.	Diff. RBF	Ker. Diff. RBF	CVA	[Celik, 2009a]	SVM
OA	87.49 (1.63)	87.72 (1.71)	90.93 (2.04)	48.52 (-)	80.31 (-)	<i>90.83</i> (0.36)
κ	0.749 (0.03)	0.754 (0.04)	0.819 (0.04)	0.527 (-)	0.609 (-)	<i>0.817</i> (0.01)
AUC	0.952 (0)	0.955 (0.01)	0.975 (0.01)	0.864 (-)	0.896 (-)	<i>0.967</i> (0.01)
AR	0.563 (0.05)	0.570 (0.05)	0.672 (0.07)	0.293 (-)	0.367 (-)	<i>0.667</i> (0.01)

Table 7.1: Figures of merit for the automatic change detection methods for the Gloucester dataset - The most accurate and the second most accurate are outlined in bold and italic, respectively.

illustrated in Table 7.1. Maps of the supervised SVM are very similar to the aforementioned ones, but providing a much lower standard deviation of the outcomes, visually defined by the predominance of black and white colours in the sum-of-changes map. By comparing the obtained maps to the CVA, it is visible that the improvements in accuracy are given by the lower commission errors, that greatly penalised the CVA. Finally, the approach of [Celik, 2009a] indicated clearly where changes occurred, but at the price of strong spatial smoothing, highlighting the difficulty of finding a trade-off between the filtering of noise and consideration of spatial context to remove false detections and preservation of the geometrical resolution of the original images.

Brüttisellen subset. For this case study involving again a Zurich neighbourhood, since the images are smaller, 100 random pixels per mode of the histogram are selected. The same number of samples has been used to train the SVM. Globally, on the average, all the tested approaches provided very good results, except for the CVA and the approach presented in [Celik, 2009a], as illustrated in Table 7.2. By observing the difference between the *Diff. Lin.* and *Diff. RBF*, it appears that nonlinear clustering provides a smoother solution, in the range of 0.11 κ better. As for the previous case study, the *Ker. Diff. RBF* provided the highest accuracy among the unsupervised approaches, this time with a very low standard deviation. Its accuracy further improves the *Diff. RBF* by 0.1 κ and results only 0.02 points inferior to SVM. As for the previous case study, although valid approaches to highlight main changes, CVA and the approach of [Celik, 2009a] result in the poorer performances. CVA suffers from the ambiguity of the representation, and the method from [Celik, 2009a] again suffers from a very hardly tunable balance of the spatial smoothing at the price of the preservation of the geometrical accuracy. In this case, note that the ground truth used for testing the method, in particular for the class “change”, respects well the geometry of the objects and consequently penalizes spatial over-smoothing.

By comparing the change maps in Figures 7.3, Brüttisellen 2 (a)-(f), the same observations made by analyzing the figures merit are remarked. In particular, the low standard deviation and very high accuracy of the *Ker. Diff. RBF* and of the SVM provided the best maps. SVM further reduced false alarms and differences between outcomes. In general, the most accurate approaches improve the maps by providing lower rates of false alarms. This observation is underlined also by looking at the ROC curves in Figure 7.4(b). How-

7. Unsupervised change detection

	Diff. Lin.	Diff. RBF	Ker. Diff. RBF	CVA	[Celik, 2009a]	SVM
OA	91.13 (16.1)	95.13 (12.24)	<i>98.60</i> (0.3)	95.67 (-)	86.35 (-)	99.48 (0.02)
κ	0.752 (0.46)	0.867 (0.35)	<i>0.968</i> (0.01)	0.903 (-)	0.651 (-)	0.988 (0)
AUC	0.998 (0)	0.986 (0.04)	<i>0.999</i> (0)	0.978 (-)	0.893 (-)	0.999 (0)
AR	0.750 (0.42)	0.859 (0.32)	<i>0.944</i> (0.01)	0.832 (-)	0.505 (-)	0.979 (0)

Table 7.2: Figures of merit for the automatic change detection methods for the Brüttisellen 2 dataset - The most accurate and the second most accurate are outlined in bold and italic, respectively.

ever, for the couples *Ker. Diff. RBF* - SVM and *Diff. Lin.* - *Diff. RBF* the differences in the selected run (an average performance of *Ker. Diff. RBF*) are too close to make general conclusions.

Greek island. For these experiments, 100 samples (50 per mode) compose the pseudo-training set. As for the previous case, the smallest set reaching the plateau in accuracy is reported. The numerical performances illustrated in Table 7.3, indicate that the *Diff. Lin.* approach performed better than its nonlinear counterpart, thanks to an improved detection rate. However, its standard deviation is higher, indicating that in one or more runs the algorithm converged unevenly to different solutions. The *Diff. RBF* approach provided the most stable solution among the proposed methods. Again, nonlinear partitioning of the difference image in the input space, as for the first case study, did not significantly improve the change detection process. The *Ker. Diff. RBF* approach is again the most accurate among unsupervised methods, confirming the better representation for the change detection problem. For this method, the κ score increased of a sharp 0.22 κ points with respect to the *Diff. RBF* method. The methods used for the comparison showed again lower accuracy, confirming the complexity of the scene composed by a large region of water, strongly clustered in the spectral domain, and by changes related to a small patch of burned forest. The approach of [Celik, 2009a], even with a smaller accuracy, provided less missed detections and more false alarms than the CVA. This is also visible in the ROC curves reported in Figure 7.4(c). SVM are again the best approach, suggesting that the use of the labels to fit a separating boundary obviously improves the detection rate. This contrast when exploiting supervision may be caused by a slightly multi-modal distribution for the class of changes, not harming the SVM.

The change maps shown in Figures 7.3, Greek Island (a)-(f), clearly illustrates the strength of the proposed approach: by using nonlinear clustering and in particular by adopting the better representation provided by the *Ker. Diff. RBF*, small deviations to the unchanged class can be clustered as such, making the detection of large deviations, corresponding to changes, more accurate. In this case, both the CVA and the method of [Celik, 2009a] suffered from the not normalizable differences that appeared on the island, that make the standard difference image and the magnitude suboptimal due to ambiguity in the representation. Therefore, methods relying solely on the magnitude are supposed

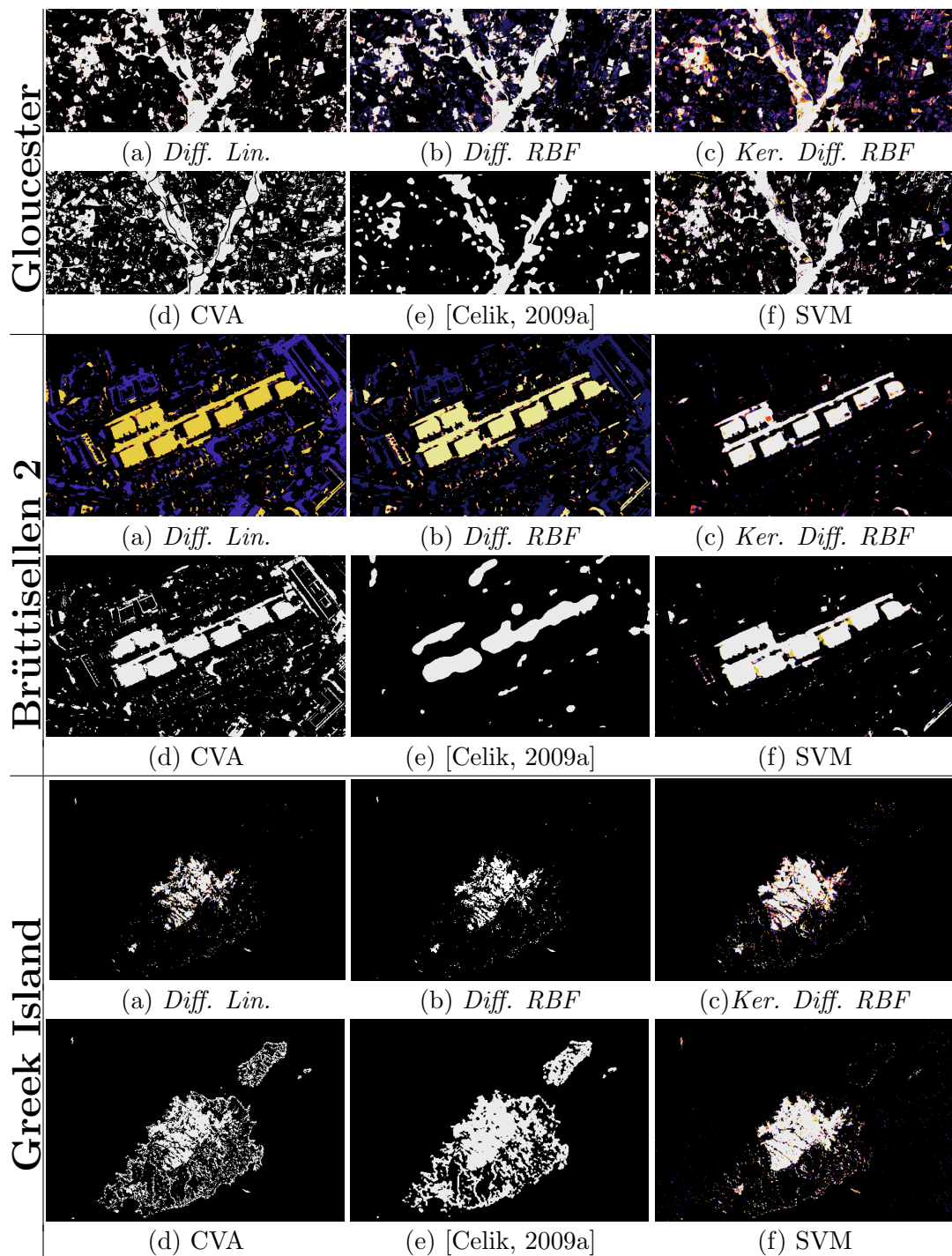


Figure 7.3: Change maps for the three different datasets using the proposed automatic kernel-based framework - For each dataset, the maps correspond to: (a) the *Diff. Lin.* approach, (b) the *Diff. RBF*, (c) the *Ker. Diff. RBF*, (d) the CVA, (e) the approach proposed by Celik [2009a] and finally (f) refers to the map obtained by SVM.

7. Unsupervised change detection

	Diff. Lin.	Diff. RBF	Ker. Diff. RBF	CVA	[Celik, 2009a]	SVM
OA	86.54 (1.24)	85.57 (0.37)	<i>88.77</i> (1.3)	77.84 (-)	76.90 (-)	99.44 (0.47)
κ	0.607 (0.04)	0.573 (0.01)	<i>0.793</i> (0.02)	0.516 (-)	0.503 (-)	0.818 (0.02)
AUC	0.964 (0)	0.958 (0)	<i>0.968</i> (0)	0.872 (-)	0.894 (-)	0.979 (0)
AR	0.484 (0.04)	0.448 (0.01)	<i>0.582</i> (0.03)	0.310 (-)	0.286 (-)	0.818 (0.02)

Table 7.3: Figures of merit for the automatic change detection methods for the Greek Island dataset - The most accurate and the second most accurate are outlined in bold and italic, respectively.

to provide worse results than other approaches, as the latter fully exploit the information content of the data. The map issuing from the SVM is again the most accurate, with the lowest deviation from one map to the other.

7.4.2 The cost function

As introduced previously, the final outcome of many kernel methods strongly depends on the hyperparameters of the kernel function, here optimized by the geometrical loss described in Section 7.2.3. Recall that the proposed cost function can be adopted for any kind and number of kernel functions, since it relies only on distances.

In this part, we study the properties of the cost function proposed, by analyzing it when applied to the Gloucester dataset. As illustrated in the Figure 7.5, the minimization of the proposed cost function corresponds to the correct kernel parameters in terms of accuracy, as illustrated by the figures of merit for the three different case studies. For the optimization in the *Diff. RBF* case (illustrated in Figure 7.5(a)), the fitted values are around the average Euclidean distance of the pixels of the difference image in standard scores, corresponding to 2.43. By observing the plot of the distances related to the cost function, the role of the RBF bandwidth is understandable, since it relates directly to distances. For small σ values the clusters are not separable, since mapped in a space

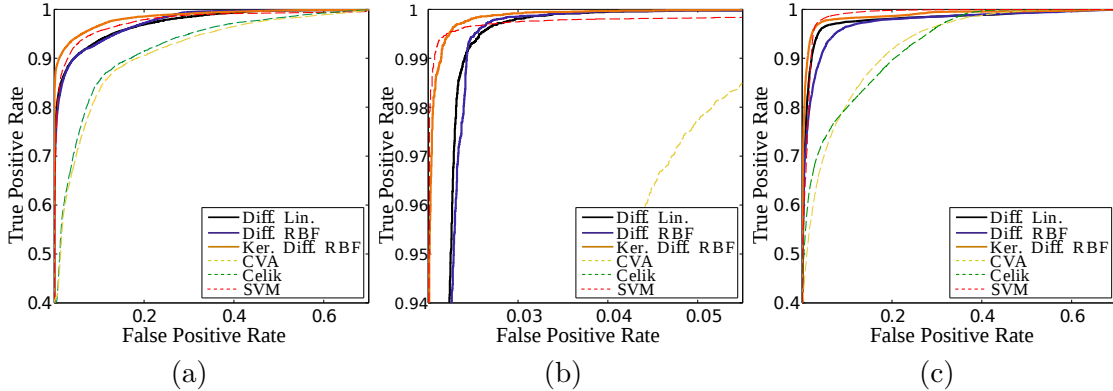


Figure 7.4: ROC curves for the three datasets used in the automatic change detection experiments - (a) Gloucester, (b) Brüttisellen 2 and (c) Greek Island.

7.4 Results and experimental validation

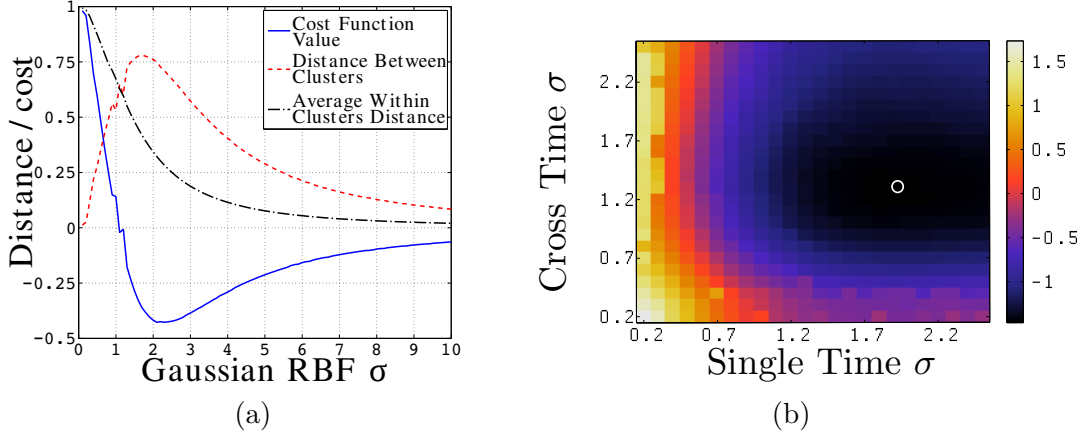


Figure 7.5: Unsupervised cost function example - It depicts a single random run on the Gloucester dataset: (a) line search of kernel bandwidth for the Gloucester case study, input space difference image setting, and (b) grid search of the difference kernel parameters. In (a) the minimum corresponds to $\sigma = 2.2$. In (b) the white circle indicates the minimum of the cost function, corresponding to $\theta^{\text{single}} = 1.9$, $\theta^{\text{cross}} = 1.3$.

where their shape is arbitrary and the average distance to the center is maximal (equal to 1 for the Gaussian RBF function), they lead to a situation of overfitting. Moreover, the distance between the two centers is 0, making low sigma values become bad candidates for the clustering step. In this case, the similarity is underestimated and each pixel is similar only to itself. For larger hyperparameters, the similarity is overestimated and the clusters are mapped again very close one to each other (distance between clusters near 0) in small punctual clusters (average distance near 0). The optimal separability between clusters indicated by the minimum of the cost function is reached when the parameter is in the range of the average Euclidean distance, correctly encoding the local similarities of the samples. The hyperparameter minimizing the function is close to the one producing the best trade-off between distance of the centers and compactness of the clusters.

For the difference kernel approach, the training set pixels at t_1 are distant on the average 1.81 between themselves, while at t_2 1.99. The optimal RBF bandwidth of the single time kernels is of 1.9, respecting the average distance among samples. For the cross-kernels, different values are automatically chosen depending on the dataset and the covariance between times. For the Gloucester case study, smaller parameters (with respect to the single-kernel ones) are often selected. The surface of the cost function value in Figure 7.5(b) indicates that, for the difference kernel approach, an optimal separability is reached (the cost function is lower than 0 for the chosen combination) and the empirical analysis carried out in the next paragraphs indicates that the situation is optimal also in terms of the accuracy provided by the cluster representatives retained.

7. Unsupervised change detection

7.4.3 Cluster separability

To better understand the influence of the mapping function within the partitioning scheme, Figure 7.6 illustrates the distance of each pixel to the two estimated cluster centroids. Recall that in KkM algorithm the exact coordinates of the centers are not retrievable, but exact distances from them can be obtained easily, as depicted by Equation (7.2). To cluster data, the centroids are used as an approximation of the coordinates of the true center, since they are the samples closest to the cluster mean in the RKHS. In this case, the distance from the representative is encoded in two different ways: the first is to compute the distance between the samples belonging to the difference image in the original input space after the projection into the RKHS, while the second evaluates the distance of the centroids to the pixels of the difference image computed directly into the RKHS. As introduced in Chapter 4, this should provide a better representation of the data, in particular since it is assumed that the relationships between the multi-temporal data are linearized in the RKHS. The actual average distances of the samples to each centroid and the distance between centroids can be used as a measure of the cluster separability, as for the cost function presented above.

The distances to the means of the estimated components of the mixture of univariate Gaussian distributions (the CVA case) are illustrated in Figure 7.6(a) as well as the map issued from the thresholding and the corresponding estimated κ statistic. In Figure 7.6(b) and Figure 7.6(c) the distances to the centroids by using respectively the *Diff. RBF* and the *Ker. Diff. RBF* are shown. Note that, since the algorithms work distances computed in different spaces, the estimated final centroids may also differ. The advantages of using the difference image computed in the RKHS appears clearly by observing the improved contrast in the values of the flooded region. For the cluster corresponding to changed regions the distances of the pixels to their representatives are low. In parallel, the separation with respect to the other centroid is larger and consequently it corresponds to a generalised improvement in the clustering solution. This uniformity in estimating correctly the cluster related to unchanged pixels is the reason of the strong decrease of false alarms when compared to the *Diff. RBF* of Figure 7.6(b). For this method the two clusters are less separable, generating a large number of field patches incorrectly clustered. As an example, note the differences for the unchanged areas visible in the upper right and lower right parts of Figure 7.6(b)-(c) respectively. In these areas, the *Diff. RBF* provides noisy outcomes and, as in the region between the two arms of the river, gives worse results than the CVA. However, globally, it yields less false alarms. These observations further support the intuition issuing from the experiments on the three datasets, i.e. nonlinear models strongly improve the accuracy of the change detection thanks to a better delineation of unchanged pixels.

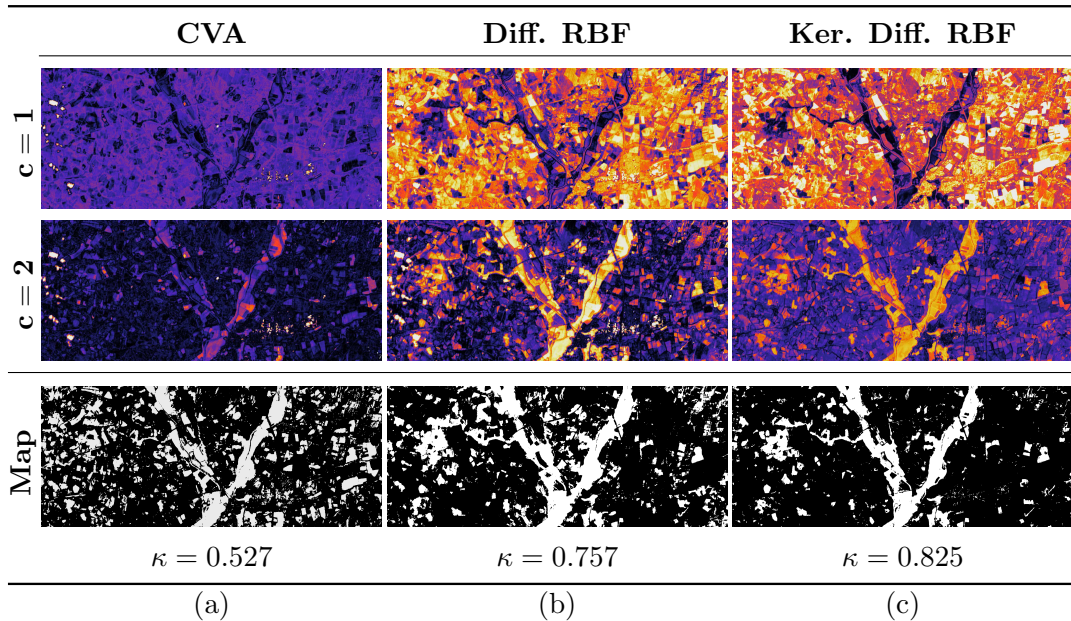


Figure 7.6: Separability of clusters in the input space and in the RKHS - In (a) the distances and the map of the CVA, running in the original input space, while in (b) the *Diff. RBF* approach, relying on the difference image mapped to the RKHS. Finally, (c) depicts the process of computing and clustering the difference image directly into RKHS, the *Ker. Diff. RBF*. In the first row the distance to the “change” cluster representative is illustrated, while in the second the distance to “no change” centroid is mapped. The corresponding map as well as its estimated κ statistics are shown in the third row. For each method, colors are scaled so that relative distances within a dataset are comparable.

7.5 Conclusions

This Chapter presented an automatic kernel-based approach to unsupervised change detection. By exploiting a proper initialization, the kernel k -means partitioning algorithm is used to estimate the centroids representing the clusters of interest, namely, changed and unchanged regions. The main issue related to the estimation of the kernel hyperparameters has been tackled by encoding a geometrical criterion, favouring dense and far clusters, into a function showing the minimum when this convenient geometrical representation is achieved. Kernel hyperparameters enforcing this profitable situation are then utilized to partition the whole bi-temporal image and to consequently generate the change map.

When estimating the similarity between pixels composing the difference image in the feature spaces (the *Ker. Diff. RBF* approach) performances are much better than simply clustering the original difference image using either linear or nonlinear models. This indicates that a better representation can be obtained by considering simultaneously single- and cross-time relationships among the pixels composing the multi-temporal scenes. As a consequence, the decrease in false alarms rate is stronger, as the use of separate kernels better depicts the nature of the change detection problem: single time kernels observe the

7. Unsupervised change detection

similarity of the pixels at the single times separately. For a same couple of samples, a different value of these kernel functions indicates that a change probably occurred. Cross-time kernels quantify the similarity of the same pixels but across the two acquisitions, indicating if both samples have changed (thus regularizing possibly large values of both single time kernels) or if only one pixels changed. Moreover, the cross-kernels account also for global differences between the acquisitions, such as illumination conditions and slightly different atmospheric situations. Even though the approach may seem complex, no user intervention is required, and the partitioning of large images (the Gloucester flood dataset is composed by 1,234,608 pixels) can be achieved in a couple of minutes, by using no particularly efficient implementation.

Further research might be spent to investigate spatial contextual relationships and their influence in kernel-based change detection. By exploiting the composite kernel framework exploited in this Chapter, contextual and multi-scale approaches [Bovolo, 2009] can be included in the process by combining the specific kernel functions, as proposed in [Camps-Valls et al., 2006, 2008] or in [Tuia et al., 2010a]. Furthermore, kernel functions encoding different aspects of the problem (e.g. single-time information, cross-information, cross-spatial, spectral-spatial) can be built depending on the user requirements, the type of changes and their direction (class type), allowing the application of this change detection system to large VHR images, in which the exploitation of the spatial context is crucial to solve these complex scenes.

Chapter 8

Statistical alignment for change detection using nonlinear feature extraction methods¹

This Chapter presents two approaches for relative radiometric normalization. Section 8.1 introduces to the general task of relative alignment. In Section 8.2 we consider an approach matching unchanged pixels from the multi-temporal images through the use of the kernel PCA. In Section 8.3, an extension of the former approach for heterogeneous domains is presented. This method allows the computation of the difference image without renouncing to any available information, even when using images from different sensors. Finally, Section 8.4 draws some concluding considerations.

8.1 Adjusting radiometric differences

In this Chapter we propose two methods for relative radiometric normalization for change detection in remote sensing images. These methods may be utilised as a preprocessing step, that has to be applied before change detection algorithms. As introduced in Section 2.3.3, there exist different methods to transform the data prior to the analysis, so that unchanged samples are the most similar among themselves. We recall the use of physical models to retrieve pixels absolute reflectance values, not needing additional relative compensations. However, as introduced in Chapter 2, the use of such models is costly and require a large amount of prior information. Due to their simplicity and good performance, statistical methods matching the pixel distributions from the two images are gaining interest. The underlying assumption is that for unchanged pixels the statistical distribution generating the data is the same, and the occurred shifts are only due to external factors that can be

¹This Chapter is based on the following publications: [Volpi et al., 2012a], [Volpi et al., 2013a] and [Volpi et al., 2013b]. See Section 1.3.3 for the details.

8. Feature extraction for change detection

compensated and corrected by an alignment approach. In other words, we can state that the pixel density shifted, while conditional distributions of unchanged areas remain the same on both images, i.e. $p(\mathbf{x}^{t_1}|y_l) = p(\mathbf{x}^{t_2}|y_l)$ [Quiñonero-Candela et al., 2009].

Among the different methods aiming at aligning the image distributions, we recall the use of the widely applied histogram matching, introduced in Section 2.3.3. Despite of its simplicity, histogram matching completely disregards the multi-variate nature of remote sensing images, the band covariance and higher order statistics when matching the distributions. To this end, Inamdar et al. [2008] proposed an approach matching the multi-variate image histograms. Moreover, depending on the type and size of changes, one may want to manually select unchanged regions to perform the statistical alignment, in order to not contaminate the relative matching with changed pixels. Another widely used family of methods, to which the approaches presented in this section may be related, are the (multi-variate) linear regressions, aiming at predicting the values of the second image pixels starting from those of the first, on the basis of some examples of unchanged areas [Heo and Fitzhugh, 2000; Singh, 1989]. A last approach aiming at matching the multi-variate values of the remote sensing images rely on the graph matching strategy [Conte et al., 2004; Tuia et al., 2013a]. In this case the goal is to match the pixels living on the manifold of each image by local shifts of the data cloud.

However, in both cases, the involved relative normalization may hardly accommodate all the differences that do not correspond to changes, in particular if those are large. In particular, seasonality effects, shadows and different illumination conditions may introduce not normalizable differences or enforce nonlinear relationships, for instance by occlusions, between the multi-temporal images [Theiler and Perkins, 2007]. In this sense, methods relying on image differencing, that despite the enticing simplicity is a delicate operation, are prone to fail or to provide suboptimal change maps if the aforementioned issues are not specifically addressed. Even if image histograms are matched, such not normalizable radiometric differences introduce disturbances in such representation. As a consequence, the ambiguity problem of the difference image may be enforced increasing the overlap of the “change” and “no change” class distributions, making the detection of absolute transitions more difficult.

In the following, we propose two methods: the former aligns unchanged pixels values on the direction of maximal variance using the PCA. The issues related to the nonlinearity of the data samples in the temporal component of the images are solved by adopting the kernel extension, i.e. the kernel PCA (kPCA). The second considers the problem of change detection with different (optical) sensors. To this end, the kernel canonical correlation analysis (kernel CCA, kCCA in short) is adopted.

8.2 Relative radiometric normalization using kernels

As introduced in Section 5.4, feature extraction methods in change detection systems may be used in two different ways: (i) by enhancing explicitly the signal of the changed classes, possibly using some available examples and then thresholding the component explaining

8.2 Relative radiometric normalization using kernels

the most of the changes [Gianinetto and Villa, 2007; Gómez-Chova et al., 2012; Marchesi and Bruzzone, 2009; Nielsen and Canty, 2008], or (ii) by applying a feature extraction / regression methods to perform a relative radiometric normalization and detect changes in a separate step [Heo and Fitzhugh, 2000; Nielsen, 2002, 2007; Singh, 1989]. In this Section, we propose an approach to perform a (nonlinear) statistical alignment of the unchanged areas, by adopting a framework issuing from the domain adaptation literature [Pan and Yang, 2010; Pan et al., 2011; Quiñonero-Candela et al., 2009]. These last approaches may be categorised in the feature-representation-transfer family, whose aim is to learn a new latent representation for the datasets in which tasks may be transferred from one domain to the other without losing performance.

By following the above literature, we exploit the kPCA [Schölkopf et al., 1998; Shawe-Taylor and Cristianini, 2004] to find a common mapping of the images, where the divergence between the probability distributions of unchanged pixels is reduced.

8.2.1 The kernel principal component analysis

The kPCA is the dual version of the PCA, that aims at finding a rotation of the data that maximizes the variance of the projections $\mathbf{X}\mathbf{w}$, under an orthogonality constraint $\mathbf{w}'\mathbf{w} = 1$. For a centred data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ (i.e. with a mean equal to 0), the PCA can be defined in its primal form as:

$$\arg \max_{\mathbf{w}} \frac{(\mathbf{X}\mathbf{w})'(\mathbf{X}\mathbf{w})}{\mathbf{w}'\mathbf{w}} = \frac{\mathbf{w}'\mathbf{X}'\mathbf{X}\mathbf{w}}{\mathbf{w}'\mathbf{w}} = \frac{\mathbf{w}'\mathbf{S}\mathbf{w}}{\mathbf{w}'\mathbf{w}}, \quad (8.1)$$

where \mathbf{S} is the scatter matrix (unnormalized variance) of the data in \mathbf{X} . As for the kFDA in Chapter 6, the ratio is optimized by the direction and not by the norm of \mathbf{w} . It is possible to reformulate the Rayleigh ratio in Equation (8.1) by exploiting that $\mathbf{w}'\mathbf{w} = 1$ simply as:

$$\begin{aligned} \arg \max_{\mathbf{w}} \quad & \mathbf{w}'\mathbf{S}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{w} = 1. \end{aligned} \quad (8.2)$$

In this case, note that the only difference to kFDA is the constraint $\mathbf{w}'\mathbf{N}\mathbf{w} = 1$, that imposes the orthogonality in the anisotropic metric defined by \mathbf{N} [De Bie et al., 2004]. Since \mathbf{N} corresponds to the within class scatter, this reduces to constrain the solution to lie in the direction of the minimal within class variance. For the PCA, $\mathbf{N} = \mathbf{I}$, which makes the \mathbf{w} follow the direction of the maximal variance. Practically, the PCA can be easily oriented by replacing the identity matrix in the orthogonality constraint with some positive definite matrix. For instance, by adding the noise covariance $\mathbf{w}'\mathbf{S}_{\text{noise}}\mathbf{w} = 1$, we obtain the minimum-noise-fraction transformation (or noise oriented PCA) for uncorrelated noise (diagonal $\mathbf{S}_{\text{noise}}$) or correlated (full $\mathbf{S}_{\text{noise}}$) [Green et al., 1998; Mika, 2002].

By introducing the Lagrange multipliers, Equation 8.2 can be rewritten as:

$$L(\mathbf{w}, \boldsymbol{\lambda}) = \mathbf{w}'\mathbf{S}\mathbf{w} - \lambda(\mathbf{w}'\mathbf{w} - 1). \quad (8.3)$$

8. Feature extraction for change detection

At the optimality the derivative with respect to the parameters vanishes:

$$\frac{\partial L}{\partial \mathbf{w}} = 2\mathbf{S}\mathbf{w} - 2\lambda\mathbf{w} = 0, \quad (8.4)$$

which can be solved easily by the following system of linear equations, the symmetric eigendecomposition:

$$\mathbf{S}\mathbf{w} = \lambda\mathbf{w} \quad (8.5)$$

where \mathbf{w} are the weight vectors of the projection of the data samples into the directions of maximal variance (eigenvectors) and λ is a diagonal matrix of eigenvalues corresponding to the value of the objective function (the scaled variances).

To obtain the kernel formulation, Equation 8.5 is transformed to the dual problem, by replacing \mathbf{w} with the expansion $\mathbf{X}'\boldsymbol{\alpha}$, and by left-multiplying with \mathbf{X} . This gives the following:

$$\begin{aligned} \mathbf{S}\mathbf{w} &= \lambda\mathbf{w} \\ \mathbf{S}\mathbf{X}'\boldsymbol{\alpha} &= \lambda\mathbf{X}'\boldsymbol{\alpha} \\ \mathbf{X}\mathbf{S}\mathbf{X}'\boldsymbol{\alpha} &= \lambda\mathbf{X}\mathbf{X}'\boldsymbol{\alpha} \\ \mathbf{X}\mathbf{X}'\mathbf{X}\mathbf{X}'\boldsymbol{\alpha} &= \lambda\mathbf{X}\mathbf{X}'\boldsymbol{\alpha} \end{aligned} \quad (8.6)$$

The Gram matrix $\mathbf{X}\mathbf{X}'$ contains all the dot products of samples \mathbf{x}_i and \mathbf{x}_j , as illustrated in Chapter 4. The \mathbf{X} can be replaced by their counterparts containing all the mapped samples to the RKHS, as $\mathbf{X} \rightarrow \boldsymbol{\Phi}$. The kernel trick can be directly applied, to obtain the nonlinear PCA as a symmetric eigenvalue decomposition [Schölkopf et al., 1998]:

$$\begin{aligned} \boldsymbol{\Phi}\boldsymbol{\Phi}'\boldsymbol{\Phi}\boldsymbol{\Phi}'\boldsymbol{\alpha} &= \lambda\boldsymbol{\Phi}\boldsymbol{\Phi}'\boldsymbol{\alpha} \\ \mathbf{K}^2\boldsymbol{\alpha} &= \lambda\mathbf{K}\boldsymbol{\alpha}. \end{aligned} \quad (8.7)$$

If \mathbf{K} is full rank, we can left multiply by \mathbf{K}^{-1} to obtain:

$$\mathbf{K}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}. \quad (8.8)$$

Note that here we assume a centred kernel matrix, as described in Section 4.2.2. If \mathbf{K} is not full rank, we can still solve Equation (8.8) and ignoring the null space of \mathbf{K} . The eigenvectors projecting the data to this null space do not contribute to the final directions of variance, since the null space is orthogonal to the subspace spanned by projected samples [De Bie et al., 2004]. Note that, up to a normalization factor, $\mathbf{X}\mathbf{w}^{\mathcal{Jc}} = \mathbf{X}\mathbf{X}'\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}$, relating again primal and dual forms in kernel methods. If vectors composing the rotation matrix $\boldsymbol{\alpha}$ are scaled to unit length, then $\|\mathbf{w}^{\mathcal{Jc}}\|^2 = (\mathbf{X}'\boldsymbol{\alpha})'(\mathbf{X}'\boldsymbol{\alpha}) = \boldsymbol{\alpha}\mathbf{X}\mathbf{X}'\boldsymbol{\alpha} = \boldsymbol{\alpha}\mathbf{K}\boldsymbol{\alpha} = \boldsymbol{\alpha}\lambda\boldsymbol{\alpha} = \lambda$. Conversely, to obtain a unit length primal vector $\mathbf{w}^{\mathcal{Jc}}$, as stated in Equation (8.2), the vectors in $\boldsymbol{\alpha}$ are scaled using the corresponding eigenvalue, as $1/\sqrt{\lambda_j}$.

Finally, the projection of a test sample \mathbf{x} into the kernel principal component space $\mathbf{w}'^{\mathcal{Jc}}\boldsymbol{\phi}(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{x}) = \mathbf{K}_t\boldsymbol{\alpha}$, where \mathbf{K}_t is the kernel matrix evaluating the similarity between training and testing samples. Data can also be represented by a lower

8.2 Relative radiometric normalization using kernels

dimensional set of features, more suitable to a direct pixelwise comparison. As in standard dimensionality reduction techniques, by computing the full kPCA rotation matrix $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1 \ \boldsymbol{\alpha}_2 \ \dots \ \boldsymbol{\alpha}_n]$, we obtain a number of eigenvectors equal to the size of the kernel Gram matrix. However, by already dropping the eigenvectors corresponding to the null space of \mathbf{K} , or to the eigenvalues equal to 0, the data may be rotated to a space with $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1 \ \dots \ \boldsymbol{\alpha}_q]$ corresponding to the q largest eigenvalues.

8.2.2 Multivariate alignment for change detection

Instead of analysing directly a multi-temporal composition, as in DMC schemes, the feature extraction method is exploited to find a common projection for pixels coming from both images. Specifically, the common subspace is obtained by applying the kPCA on a subset of samples from the two images simultaneously (the learning set). These examples are sampled at the same geographical coordinates of both images and represent unchanged areas. Pixels are stacked element-wise (pooled, in contrast to variable-stacking) as $\mathbf{X}' = [\mathbf{x}_i^1 \ \mathbf{x}_i^2]_{i=1}^n$, to obtain a $2n \times d$ matrix composed of $2n$ samples of d spectral channels (n pixels from each image).

The directions representing the axes of maximal variance (of unchanged samples) for both sets are then used as new uncorrelated bases for rotating the two images. Once these bases have been computed, the images are mapped independently using the common projection matrix, to obtain two datasets showing unchanged samples with maximally similar sample values. Note that the physical meaning of the images is no longer maintained.

The choice of the nonlinear PCA with respect to its linear counterpart is motivated by the fact that the kPCA is much more flexible in extracting (nonlinear) structures from the data. PCA simply finds a rotation around the mean of the data matching the axes of maximal variance in the input space, and does not guarantee an increase in superposition of the unchanged samples distribution. In the proposed setting, the alignment is performed by using pixels coming from areas that have not changed between the acquisitions, which mutually belong to the same spectral class. This choice ensures that after the projection these samples have a closer value in the transformed space. In change detection terms, this means that the representation obtained by subtracting the transformed images becomes more reliable, since pixels belonging to unchanged areas are very likely to be grouped around low values observing a better deviation of changed samples. As a consequence, separability increases. However, note that if the classes present in the bi-temporal images are the same, meaning that changes are due to differences in the geographical locations of classes (i.e. no novel spectral class appears), the proposed approach may work also by sampling randomly pixels on the image. In this case, no matter where classes appear, their coordinates would occupy approximately the same regions of the spectral space (if compared to sampling couples of pixels), and the kPCA step does not change significantly. One should only ensure that all the classes present in the image are sufficiently represented in the data matrix \mathbf{X} .

8. Feature extraction for change detection

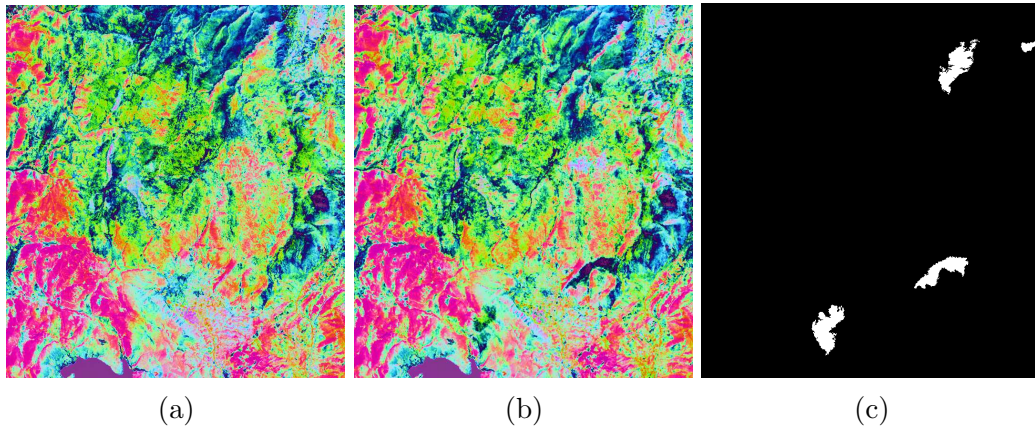


Figure 8.1: Statistically aligned images using the kPCA approach - The reprojected data in a false RGB color composite, with principal components illustrated in decreasing eigenvalue order (R: first, G: second and B: third kernel principal component). (a) 1987 and (b) 1991 transformed images, (c) ground truth of changes.

8.2.3 Experimental setup

In these experiments, the Greece fires multi-temporal images have been used (see Appendix C.7). Data are projected into the first 3 principal components for illustration purposes, and are depicted in Figure 8.1(a)-(b). As term of comparison, changes we are looking for are illustrated in Figure 8.1(c). To ensure fair comparisons, the standard difference image has been computed after histogram matching.

To assess the suitability of the proposed alignment approach, the difference of the first principal components and the standard difference image are used as inputs for different change detection methods: the CVA and the supervised one-class support vector domain description (SVDD) [Tax and Duin, 2004] (both linear and nonlinear). This last approach consists in finding, during the training, a hypersphere with a minimum radius length containing all the unchanged samples. During the test step, SVDD attributes the class “changed” to pixels lying outside the hypersphere and “unchanged” to those lying inside it [Tax and Duin, 2004]. The SVDD models the “unchanged” class boundaries by exploiting only some labels from this class. Changed pixels are detected by thresholding the decision function allowing a given fraction of outliers.

To perform kPCA, a Gaussian RBF kernel has been used. The bandwidth σ has been set as the median Euclidean distance among pixels randomly chosen in the entire image (20% of the available pixels). To estimate the projection matrix, 200 samples are chosen in a supervised way from unchanged areas (100 samples per date at the same coordinates). The same pixels are then used to train the SVDD methods using either the transformed or the original difference image, respectively.

Note that the selection of samples for computing the kPCA can be easily extended to be unsupervised by using a pseudo-training sampling criteria (see Chapter 7). To ensure the best possible performance of the change detectors, both the threshold on the CVA

8.2 Relative radiometric normalization using kernels

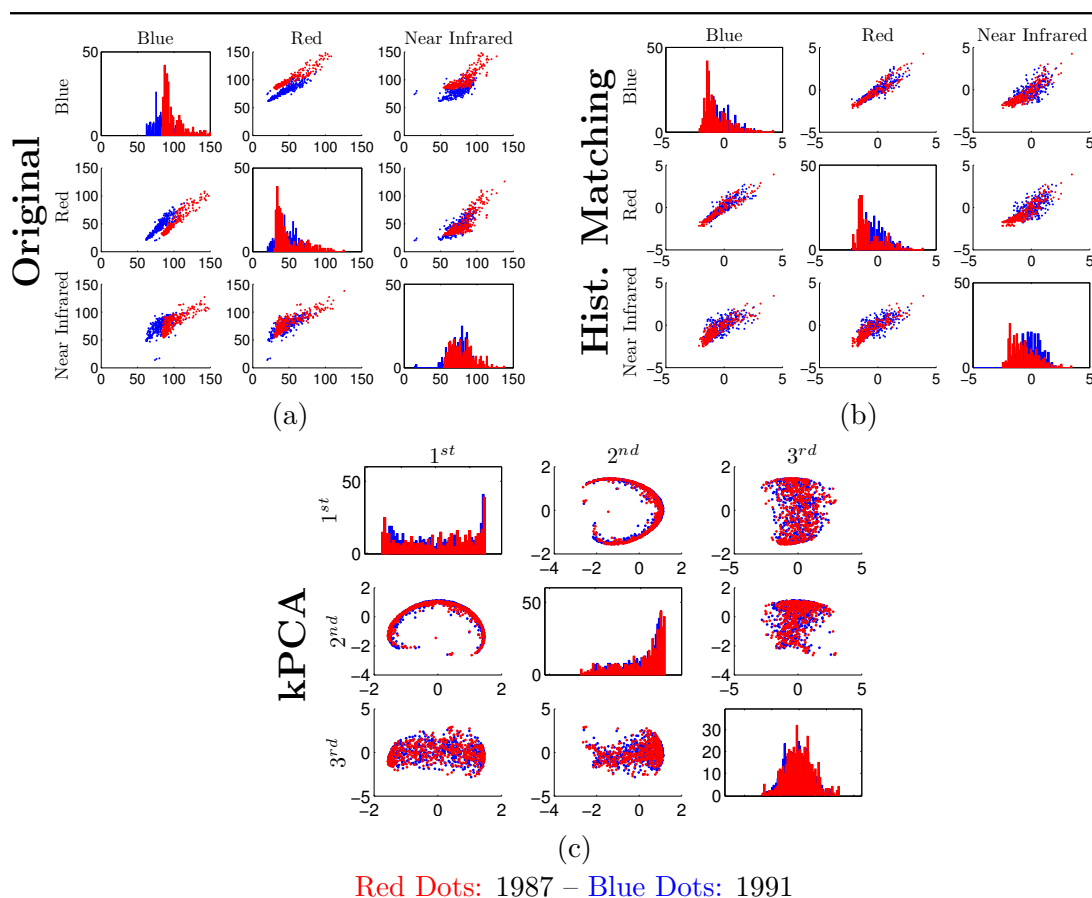


Figure 8.2: kPCA-based statistical relative radiometric normalization - Unchanged pixels represented with (a) Original DN values, (b) after histogram matching and (c) after kPCA-based normalization.

magnitude and SVDD hyperparameters (the rejection rate and the kernel width for the nonlinear RBF SVDD) are tuned in a supervised way by using 100 independent validation coordinates, equally belonging to changed and unchanged classes. An exhaustive line/grid search by cross-validation has been adopted. Note that since the kernel principal projections are scaled differently than the original pixel/bands values, the kernel width in the RBF SVDD has been re-estimated. However, automatic and efficient methods to estimate the kernel width exist (e.g. [Khazai et al., 2012]). The retained dimensionality of the projections is varied from 2 to 5 and the best change maps are presented.

Estimated Kappa statistic (κ) is used on an independent and common test set ($\sim 60'000$ pixels from Figure C.7(c)) to compare performances. Average scores obtained after 10 independent realizations of the training set are presented in the following.

8. Feature extraction for change detection

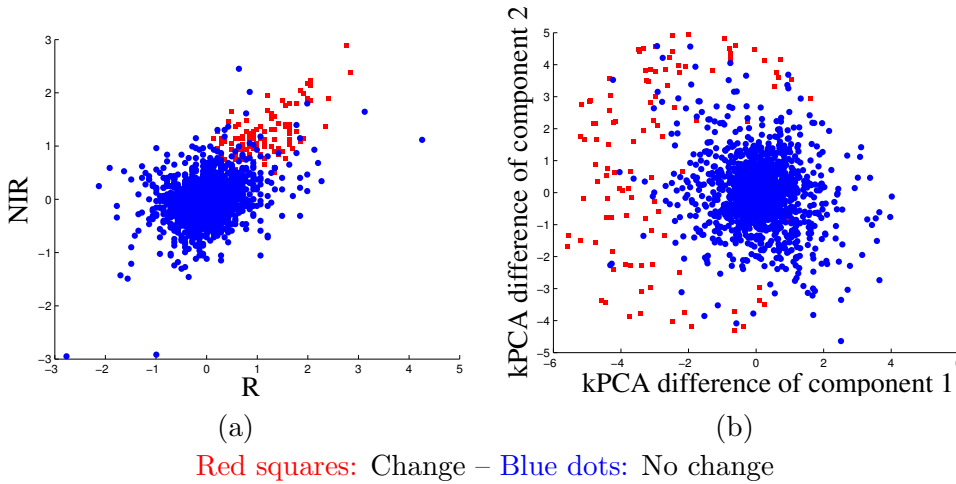


Figure 8.3: Scatterplot of the difference image without and with kPCA alignment - (a) Scatterplot of the difference of the NIR-R bands and (b) scatterplot of first 2 kernel principal components difference.

8.2.4 Results

The data transformation Figure 8.2(a)-(c) illustrates the main properties of the proposed transformation. Figure 8.2(a) represents a cross-scatterplot of the NIR-R-B bands (selected from the 6 original channels) in their original space, for the class “no change” (couples of pixels at same spatial coordinates are selected). Figure 8.2(b) illustrates the same samples after histogram matching. It appears clearly that the means are better aligned. However, small differences still persist and in particular in the data covariance, since as introduced in Section 8.1 the cross-relationships of the data are not explicitly taken into account. Also, by looking at the uni-temporal histograms (R-R and NIR-NIR), small differences in the mode are visible. Finally, in Figure 8.2(c), the proposed kPCA-based alignment is illustrated. Even if following a more complex distribution the scatterplots of the no change samples show a better alignment. This is due to the kPCA, that in this case consider higher order relationships based on covariance structures in possibly infinite dimensional RKHS.

By disregarding the raw data (we assume that is always possible to perform histogram matching), the benefits of the transformation are illustrated in Figure 8.3. Although empirically the distribution of unchanged samples seems the same, roughly $\mathcal{N}(0, 1)$ since the datasets have been transformed to standard scores, the changed samples tend to be scattered farther. In particular, the separation from unchanged samples by the SVDD hypersphere or simply by thresholding the unidimensional magnitude (the distance of the difference samples from the origin, CVA) becomes easier.

In Figure 8.4, the above scatterplots are translated into the spatial domain. The magnitude of the difference image in Figure 8.4(a) corresponds to the standard difference image (after histogram matching). Changes are clearly visible, but a large amount of

8.2 Relative radiometric normalization using kernels

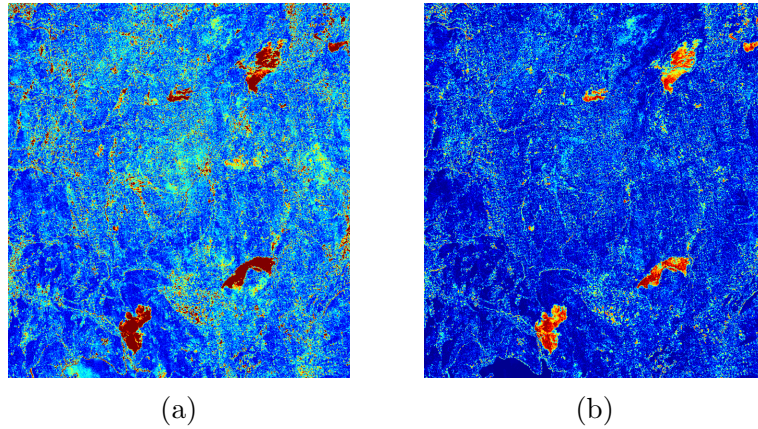


Figure 8.4: CVA magnitude for the histogram matched data and using the kPCA-based alignment - (a) CVA magnitude after histogram matching and (b) kPCA-based difference image magnitude of the first 3 kernel principal components. Note that the colours are rescaled so that are comparable between the two magnitudes.

spurious noise affects the rest of the image, in particular for the samples belonging to unchanged areas. Figure 8.4(b) depicts the magnitude of the difference image after kPCA projection: changes are still clearly visible, even if they possess a larger range of values, but still more easily discriminable thanks to the large reduction of the background noise. Some artefacts such as the image striping are still visible, but note that the approach has not developed to reduce image noise (if striping can be considered as such). The new representation seems more appropriate for change detection for two main reasons: firstly, the kPCA alignment considers the relationships between the samples in a multi-variate manner, while the standard histogram matching does not. Secondly, in computing the magnitude in Figure 8.4(b), only 2 principal components are used, in contrast to the 6 spectral channels used for the magnitude in Figure 8.4(a). Even if the dimensionality of the original data is not too large, the ℓ_2 -norm used to compute the magnitude of the transformed data is less affected by the noise in each channel, that inflates the magnitude of the vector even for unchanged samples. The kPCA allows to work in a lower dimensional space, ending up with a less noisy magnitude image. To be fair in the visualization, note that few aberrant values (outliers) have been removed from the CVA by rescaling the colours. The same outliers were not present in the kPCA-based transformation.

Numerical accuracies Change map accuracy plots are illustrated in Figure 8.5. The CVA approach on the transformed image produced homogeneous κ scores along different dimensions of the projection. The average score is $\kappa = 0.362$, which is 0.072 higher than the average CVA accuracy applied to the original difference image. The change maps in Figure 8.6(a) illustrate the CVA applied on both types of difference images: in the transformed space, CVA provided less stable but often more accurate results, in particular thanks to a lower false alarm rate. It is worth mentioning that the hit rate of the CVA

8. Feature extraction for change detection

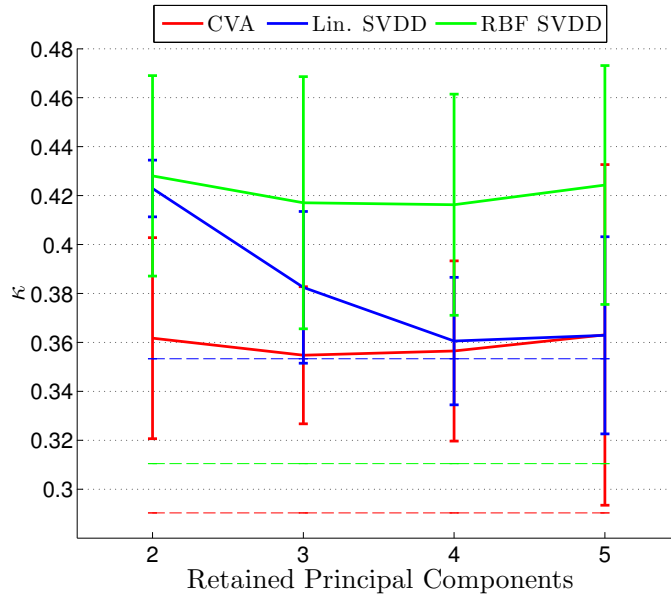


Figure 8.5: Change detection accuracies for the kPCA-based relative normalization - Dashed lines correspond to the methods in the original input space (standard difference image), while the solid lines correspond to the methods applied on the difference image after kPCA relative radiometric normalization.

performed on both types of difference image is high. Regarding the SVDD approaches, the linear SVDD (*Lin. SVDD*) performed clearly better when considering, the transformed difference image, but only the two first aligned components. It results in an average increase of performance of 0.07κ with respect to the original difference image model. It is worth mentioning that the *Lin. SVDD* model acts very similar to the CVA: it fits a spherical separating boundary around the ‘no change’ class. However, note that, similarly to SVM, the SVDD formulation allows for slack variables accounting for training errors. For this reason, the SVDD may be more robust in terms of generalization accuracy, since the separating sphere is not influenced by outliers and noise. This could explain the much higher accuracy of the *Lin. SVDD* with respect to the CVA. When considering the nonlinear *RBF SVDD*, the improvements are again very clear: the *RBF SVDD* applied to the transformed image (by retaining 2 principal components) performed 0.12κ better than when applied on the standard difference image, with a $\kappa = 0.428$ for the aligned *RBF SVDD* and $\kappa = 0.310$ for the standard *RBF SVDD*, respectively.

The maps in Figure 8.6(a)-(c) illustrate the average reduction of false detections by adopting the proposed transformed difference image. They are illustrated in the same scale as the ones in Figure 7.3, Section 7.4.1, that is, white corresponds to pixels always detected as “change”, while black characterize the ones always classified as “unchanged”. The colours in between, from purple to yellow, indicates the number of times the corresponding pixel has been classified as “changed”.

For the CVA, illustrated in Figure 8.6(a), it is clearly visible that the detection of false

8.2 Relative radiometric normalization using kernels

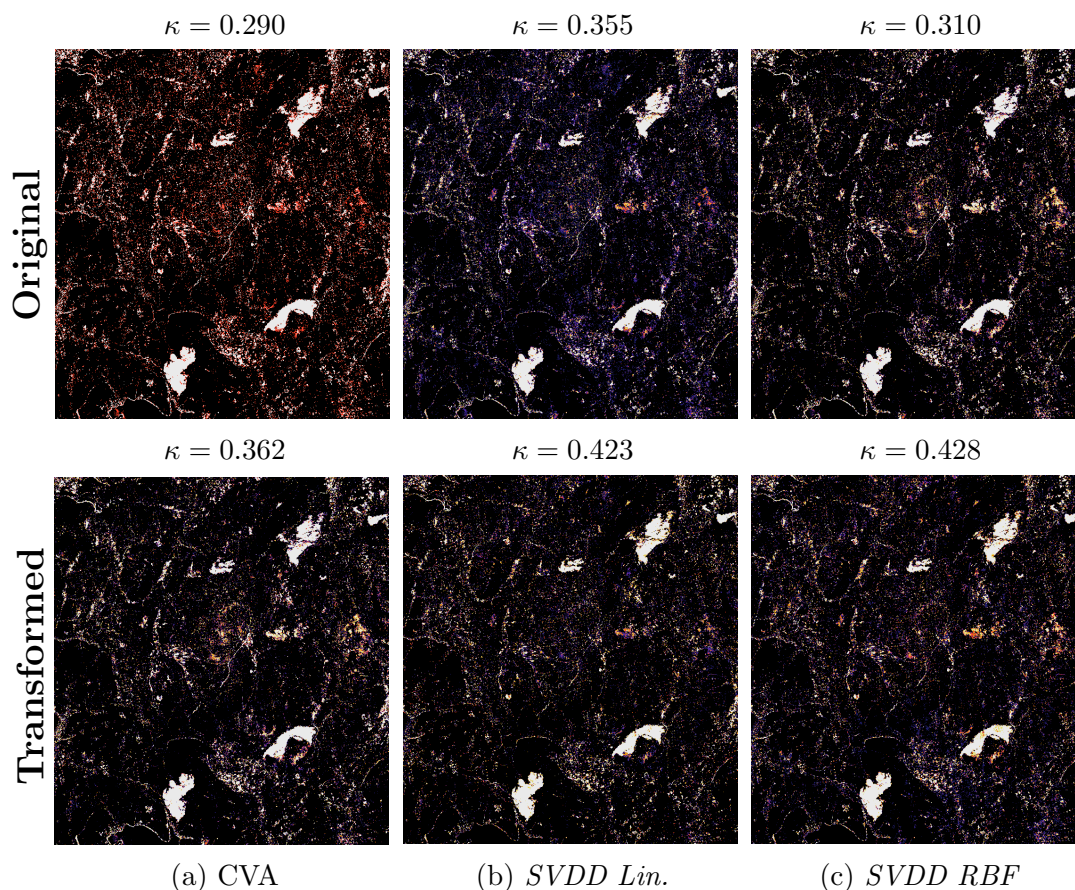


Figure 8.6: Sum of 10 change maps for the original histogram matching and kPCA relative normalization - Sum-of-change maps for the (a) CVA, (b) linear SVDD and (c) RBF SVDD, with the first row corresponding to results on the histogram matched data while on the second row the images adjusted with the proposed kPCA-based radiometric alignment, by retaining 2 dimensions. The brighter the pixel is, the more often it has been detected as changed.

8. Feature extraction for change detection

changes is generally reduced by considering the aligned images. Note that the CVA on the original dataset produced different outcomes, i.e. it has a standard deviation, since the validation sets used to fit the threshold have been varied along the iterations. Recall that this small validation set is the same adopted also for the SVDD-based approaches. The same observations may be made for the *Lin. SVDD* and for the *RBF SVDD*, for which using the transformed space produced less false alarms. On the average, the latter approach improves the most and in a more stable manner with respect to the growing dimensionality of the transformed space.

8.2.5 Discussion

In this case study, we have presented a strategy to align the common information carried by unchanged pixels. By aligning what has not changed, changes tend to be more distinguishable. In all the experiments, the accuracy of the detection in the transformed space is superior to the one obtained with the direct difference image analysis, independently from the dimensionality retained or the method used. Experiments indicated that for the tested dataset, the maximal accuracy occurs when retaining two dimensions for computing the transformed difference image. It has been also observed that, when considering more than 10 dimensions, the change detection accuracy decreases under the baseline, that is the method applied in the original space. The noise present in high frequency components contaminates the transformed difference image.

On the other hand, the tested system showed high variance of the final accuracy. This is probably due to the difficulty of sampling, at each run, pixels providing the same information. In particular, when selecting randomly the unchanged pixels used to learn the kPCA transformation and to train the SVDD, the ground cover classes represented may vary from one draw to the other, influencing the projections and the change detection outcome. To solve this issue, regularization penalizing the spatial variability of the multi-temporal signal may be considered. Another solution could be sampling using some additional information, such as unsupervised initializations, to obtain informative samples but keeping working with small matrices, or to sample much larger regions if the computational power is not an issue.

Moreover, the way of combining the temporal component of the images can also be criticized. For large shifts in the distribution of each image, the obtained aligned features may be suboptimal, since the projections to the directions of the pooled maximal variance may largely differ to the ones of the single time images. For these reasons, in the next Section we present an approach developed specifically to solve these issues. It aims at (i) maximizing the correlation of the projections between the unchanged samples independently, instead of projecting data onto the common maximal variance direction; (ii) it accounts for a regularization term favouring smooth projections following the geometrical nature of the data (the manifold) and (iii) it is able to work with data with different input spaces (e.g. images from multiple sensors).

8.3 Relative alignment for change detection in heterogeneous sources

As the Earth observation technologies evolve, a new processing trend is observed in the recent years. To fully exploit all the remote sensing data that has already been collected and will continue to be gathered by future missions, studies involving multi-source imagery are starting to receive attention in the community. Multi-source and multi-modal acquisitions are nowadays standard sources of information, but their systematic assimilation in real world systems is still limited by the complexity and ad-hoc nature of the most of data fusion methods [Gao et al., 2006].

In this Section, we propose a method able to align heterogeneous data sources, i.e. for images with different spectral channels, and to perform change detection exploiting the derived images in a successive step. Specifically, we look for joint mappings of the original data sources that maximize the correlation between unchanged pixels at both dates. To this end, we use the regularized non-linear kernel CCA [Bach and Jordan, 2002a; Hotelling, 1936]. Manifold regularization using the graph Laplacian has been considered to find projections that respect well the manifold structure of the data [Belkin et al., 2006; Blaschko et al., 2011]. Also, it allows to relax problems related to small sample conditions, since it allows to select pixels randomly from all the image, and reduce overfitting issues, by penalizing complex projections. The performance of the proposed semi-supervised kernel CCA (SSkCCA) is illustrated through a challenging example using Landsat images.

8.3.1 Paired multi-view learning and regularized canonical correlation analysis

Canonical correlation analysis. The CCA is a multi-view learning method developed to study the relationships between two paired datasets, composed by two different sets of features (views) describing the same examples [Hotelling, 1936]. The aim of the CCA is to find joint projections \mathbf{w}_k , for each group of features k , by minimizing the angle among the mapped vectors $\mathbf{X}_k \mathbf{w}_k$, with $k \in \{1, 2\}$. This corresponds to the maximization of the cosine between the mapped vectors, or, equivalently, of the correlation between the projected vectors as:

$$\begin{aligned} \arg \max_{\mathbf{w}_1, \mathbf{w}_2} \cos(\angle(\mathbf{X}_1 \mathbf{w}_1, \mathbf{X}_2 \mathbf{w}_2)) &= \frac{(\mathbf{X}_1 \mathbf{w}_1)'(\mathbf{X}_2 \mathbf{w}_2)}{\sqrt{(\mathbf{X}_1 \mathbf{w}_1)'(\mathbf{X}_1 \mathbf{w}_1)} \sqrt{(\mathbf{X}_2 \mathbf{w}_2)'(\mathbf{X}_2 \mathbf{w}_2)}} & (8.9) \\ &= \frac{\mathbf{w}_1' \mathbf{X}_1' \mathbf{X}_2 \mathbf{w}_2}{\sqrt{(\mathbf{w}_1' \mathbf{X}_1' \mathbf{X}_1 \mathbf{w}_1)} \sqrt{(\mathbf{w}_2' \mathbf{X}_2' \mathbf{X}_2 \mathbf{w}_2)}} \\ &= \frac{\mathbf{w}_1' \mathbf{S}_{12} \mathbf{w}_2}{\sqrt{(\mathbf{w}_1' \mathbf{S}_{11} \mathbf{w}_1)} \sqrt{(\mathbf{w}_2' \mathbf{S}_{22} \mathbf{w}_2)}} \end{aligned}$$

where $\mathbf{X}_1 \in \mathbb{R}^{n \times d_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{n \times d_2}$ are the two data matrices of the bi-temporal sets (time t_1 and t_2 , $d_1 \neq d_2$) of the mean-centred multi-variate coregistered pixels. The matrix \mathbf{S}_{kq} is the empirical scatter between views k and q .

8. Feature extraction for change detection

By setting the norm of the projected features to be unit in \mathbf{S}_{kk} ($\mathbf{w}'_1 \mathbf{S}_{11} \mathbf{w}_1 = 1$ and $\mathbf{w}'_2 \mathbf{S}_{22} \mathbf{w}_2 = 1$), this optimization can also be seen as the minimization of the (Mahalanobis) distance among projections [Kuss and Graepel, 2003].

Similarly to the FDA and the PCA, the Lagrangian formulation of this constrained optimization problem is, with the two Lagrangian multipliers λ_1 and λ_2 , as:

$$L(\mathbf{w}_1, \mathbf{w}_2, \lambda_1, \lambda_2) = \mathbf{w}'_1 \mathbf{S}_{12} \mathbf{w}_2 - \frac{1}{2} \lambda_1 (\mathbf{w}'_1 \mathbf{S}_{11} \mathbf{w}_1 - 1) - \frac{1}{2} \lambda_2 (\mathbf{w}'_2 \mathbf{S}_{22} \mathbf{w}_2 - 1), \quad (8.10)$$

by remarking that $\frac{1}{2} \mathbf{w}'_2 \mathbf{S}_{11} \mathbf{w}_1 + \frac{1}{2} \mathbf{w}'_1 \mathbf{S}_{22} \mathbf{w}_2 = 1$. At the optimality, we have $\partial L / \partial \mathbf{w}_1 = 0$ and $\partial L / \partial \mathbf{w}_2 = 0$:

$$\begin{cases} \frac{\partial L}{\partial \mathbf{w}_1} = \mathbf{S}_{12} \mathbf{w}_2 - \lambda_1 \mathbf{S}_{11} \mathbf{w}_1 = 0 \Rightarrow \mathbf{S}_{12} \mathbf{w}_2 = \lambda_1 \mathbf{S}_{11} \mathbf{w}_1 \\ \frac{\partial L}{\partial \mathbf{w}_2} = \mathbf{S}_{21} \mathbf{w}_1 - \lambda_2 \mathbf{S}_{22} \mathbf{w}_2 = 0 \Rightarrow \mathbf{S}_{21} \mathbf{w}_1 = \lambda_2 \mathbf{S}_{22} \mathbf{w}_2. \end{cases} \quad (8.11)$$

Since $\mathbf{w}'_1 \mathbf{S}_{11} \mathbf{w}_1 = \mathbf{w}'_2 \mathbf{S}_{22} \mathbf{w}_2 = 1$ and using what observed in Equation (8.11), $\lambda_1 = \lambda_2$ and $\mathbf{S}_{12} \mathbf{w}_1 = \lambda \mathbf{S}_{11} \mathbf{w}_1 = \lambda \mathbf{S}_{22} \mathbf{w}_2 = \mathbf{S}_{21} \mathbf{w}_1$. We can then reformulate the problem as:

$$\mathbf{S}_{12} \mathbf{w}_1 + \mathbf{S}_{21} \mathbf{w}_1 = \lambda (\mathbf{S}_{11} \mathbf{w}_1 + \mathbf{S}_{22} \mathbf{w}_2), \quad (8.12)$$

or, in matrix form, as:

$$\begin{pmatrix} \mathbf{0} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{S}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix}. \quad (8.13)$$

This system of equations can be solved as a generalized eigenvalue decomposition [De Bie et al., 2004; Shawe-Taylor and Cristianini, 2004]. The projections of the variables \mathbf{X}_k into the space in which the correlation is mutually maximized are called canonical variates [Hotelling, 1936], and the projection in this space is performed simply as in the definition of the problem in Equation (8.9), $\mathbf{X}_k \mathbf{w}_k$. Note that $\text{corr}(\mathbf{X}_1 \mathbf{w}_1, \mathbf{X}_2 \mathbf{w}_2) = \mathbf{w}'_1 \mathbf{S}_{12} \mathbf{w}_2 = \lambda_1 \mathbf{w}'_1 \mathbf{S}_{11} \mathbf{w}_1 = \lambda_2 \mathbf{w}'_2 \mathbf{S}_{22} \mathbf{w}_2 = \lambda$, indicating the correlation of the projections is equal to λ . Thus, the larger the eigenvalue, the largest the correlation between the considered projections.

Kernel canonical correlation analysis (KCCA). To obtain the standard two-set kCCA algorithm, the primal in Equation (8.13) is replaced with its dual by plugging

8.3 Multi-sensor alignment for change detection

$\mathbf{w}_k = \mathbf{X}'_k \boldsymbol{\alpha}_k$, and by left multiplying by $\begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix}$:

$$\begin{aligned}
& \begin{pmatrix} \mathbf{0} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{S}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} \\
\Rightarrow & \begin{pmatrix} \mathbf{0} & \mathbf{X}'_1 \mathbf{X}_2 \\ \mathbf{X}'_2 \mathbf{X}_1 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{X}'_1 \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}'_2 \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} \\
\Rightarrow & \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{X}'_1 \mathbf{X}_2 \\ \mathbf{X}'_2 \mathbf{X}_1 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{X}'_1 \boldsymbol{\alpha}_1 \\ \mathbf{X}'_2 \boldsymbol{\alpha}_2 \end{pmatrix} \\
& = \lambda \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \mathbf{X}'_1 \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}'_2 \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \mathbf{X}'_1 \boldsymbol{\alpha}_1 \\ \mathbf{X}'_2 \boldsymbol{\alpha}_2 \end{pmatrix} \quad (8.14) \\
\Rightarrow & \begin{pmatrix} \mathbf{0} & \mathbf{X}_1 \mathbf{X}'_1 \mathbf{X}_2 \\ \mathbf{X}_2 \mathbf{X}'_2 \mathbf{X}_1 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{X}'_1 \boldsymbol{\alpha}_1 \\ \mathbf{X}'_2 \boldsymbol{\alpha}_2 \end{pmatrix} \\
& = \lambda \begin{pmatrix} \mathbf{X}_1 \mathbf{X}'_1 \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \mathbf{X}'_2 \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \mathbf{X}'_1 \boldsymbol{\alpha}_1 \\ \mathbf{X}'_2 \boldsymbol{\alpha}_2 \end{pmatrix} \\
\Rightarrow & \begin{pmatrix} \mathbf{0} & \mathbf{X}_1 \mathbf{X}'_1 \mathbf{X}_2 \mathbf{X}'_2 \\ \mathbf{X}_2 \mathbf{X}'_2 \mathbf{X}_1 \mathbf{X}'_1 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix} \\
& = \lambda \begin{pmatrix} \mathbf{X}_1 \mathbf{X}'_1 \mathbf{X}_1 \mathbf{X}'_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \mathbf{X}'_2 \mathbf{X}_2 \mathbf{X}'_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}
\end{aligned}$$

The kernel trick can be applied, replacing the $\mathbf{X}_k \mathbf{X}'_k$ terms with a centred kernel matrix \mathbf{K}_{kk} of inner products between the mapped data matrices $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$, obtaining:

$$\begin{pmatrix} \mathbf{0} & \mathbf{K}_{11} \mathbf{K}_{22} \\ \mathbf{K}_{22} \mathbf{K}_{11} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{K}_{11} \mathbf{K}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{22} \mathbf{K}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}. \quad (8.15)$$

Note that this problem is not regularized. When performing kCCA it should be preferred to work with a regularized solution, in order to avoid trivial or degenerate solutions on the training samples, consequently leading to poor projections for test data. To see this, we can rewrite the kernel CCA problem as:

$$\arg \max_{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2} \frac{\boldsymbol{\alpha}'_1 \mathbf{K}_{11} \mathbf{K}_{22} \boldsymbol{\alpha}_2}{\sqrt{(\boldsymbol{\alpha}'_1 \mathbf{K}_{11} \mathbf{K}_{11} \boldsymbol{\alpha}_1)} \sqrt{(\boldsymbol{\alpha}'_2 \mathbf{K}_{22} \mathbf{K}_{22} \boldsymbol{\alpha}_2)}}. \quad (8.16)$$

As for Equation (8.9), the denominator can be scaled so that $(\boldsymbol{\alpha}'_1 \mathbf{K}_{11} \mathbf{K}_{11} \boldsymbol{\alpha}_1) = 1$ and $(\boldsymbol{\alpha}'_2 \mathbf{K}_2 \mathbf{K}_2 \boldsymbol{\alpha}_2) = 1$. In this case, if \mathbf{K}_{kk} it is full rank (e.g. by using a Gaussian kernel), we derive from the first part of the system in Equation (8.15) that $\boldsymbol{\alpha}_1 = \frac{1}{\lambda} \mathbf{K}_{11}^{-1} \mathbf{K}_{22} \boldsymbol{\alpha}_2$ and thus, replacing for the second view, $\mathbf{K}_{22}^2 \boldsymbol{\alpha}_2 = \lambda^2 \mathbf{K}_{22}^2 \boldsymbol{\alpha}_2$. This holds for all the solutions $\boldsymbol{\alpha}_2$, with $\lambda = 1$. Consequently, regularization is really needed to avoid such perfect correlation among the projections that would result in an overfit of the data.

8. Feature extraction for change detection

Regularization of Equation (8.16) can be performed by adding a term $\Omega(f) = \|\mathbf{w}_k^{\text{Jc}}\|^2$ (Tikhonov regularizer) at the denominators, penalizing large norms of \mathbf{w}_k^{Jc} . By transforming again the weight vectors \mathbf{w}_k^{Jc} in their dual form $\mathbf{X}'_k \boldsymbol{\alpha}_k$, this results in:

$$\begin{aligned} \arg \max_{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2} & \frac{\boldsymbol{\alpha}'_1 \mathbf{K}_{11} \mathbf{K}_{22} \boldsymbol{\alpha}_2}{\sqrt{(\boldsymbol{\alpha}'_1 \mathbf{K}_{11} \mathbf{K}_{11} \boldsymbol{\alpha}_1) + \gamma \boldsymbol{\alpha}'_1 \mathbf{K}_{11} \boldsymbol{\alpha}_1} \sqrt{(\boldsymbol{\alpha}'_2 \mathbf{K}_{22} \mathbf{K}_{22} \boldsymbol{\alpha}_2) + \gamma \boldsymbol{\alpha}'_2 \mathbf{K}_{22} \boldsymbol{\alpha}_2}} \\ & = \frac{\boldsymbol{\alpha}'_1 \mathbf{K}_{11} \mathbf{K}_{22} \boldsymbol{\alpha}_2}{\sqrt{\boldsymbol{\alpha}'_1 (\mathbf{K}_{11} \mathbf{K}_{11} + \gamma \mathbf{K}_{11}) \boldsymbol{\alpha}_1} \sqrt{\boldsymbol{\alpha}'_2 (\mathbf{K}_{22} \mathbf{K}_{22} + \gamma \mathbf{K}_{22}) \boldsymbol{\alpha}_2}} \end{aligned} \quad (8.17)$$

The canonical variate for a test sample \mathbf{x} in view k is $f(\mathbf{x}) = \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x})$. In matrix form, this is expressed as $\mathbf{K}_k \boldsymbol{\alpha}_k$, where \mathbf{K}_k represents the kernel matrix evaluating the similarity between training and test samples in view k and $\boldsymbol{\alpha}_k$ is the corresponding collection of the leading q eigenvectors $[\boldsymbol{\alpha}_k^1, \dots, \boldsymbol{\alpha}_k^q]$.

8.3.2 Semi-supervised relative alignment via manifold regularization

To obtain a fully regularized version of the kCCA, the expression of the generalized canonical correlation problem in Equation (8.13) is considered [Bach and Jordan, 2002a]. Instead of maximizing the correlation of the projection of only the two disjoint feature sets, the mutual correlation of k blocks can be maximized simultaneously, thus generalizing the CCA to multiple sets [Kettenring, 1971].

The problem is formulated starting from:

$$\begin{aligned} \arg \max_{\mathbf{w}_k} \cos(\angle(\sum_{kq} \mathbf{X}_k \mathbf{w}_k, \mathbf{X}_q \mathbf{w}_q)) & = \frac{\sum_{kq} (\mathbf{X}_k \mathbf{w}_k)(\mathbf{X}_q \mathbf{w}_q)}{\sum_k \sqrt{(\mathbf{X}_k \mathbf{w}_k)(\mathbf{X}_k \mathbf{w}_k)} \sqrt{\sum_q (\mathbf{X}_q \mathbf{w}_q)(\mathbf{X}_q \mathbf{w}_q)}} \quad (8.18) \\ & = \frac{\sum_{kq} \mathbf{w}'_k \mathbf{S}_{kq} \mathbf{w}_q}{\sum_k \sqrt{(\mathbf{X}_k \mathbf{w}_k)(\mathbf{X}_k \mathbf{w}_k)} \sqrt{\sum_q (\mathbf{X}_q \mathbf{w}_q)(\mathbf{X}_q \mathbf{w}_q)}} \end{aligned}$$

If we limit ourselves to the two set case of the above formulation, this results in optimizing:

$$\begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} = (1 + \lambda) \begin{pmatrix} \mathbf{S}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix}. \quad (8.19)$$

Equation (8.19) allows a more flexible formulation of the CCA enforcing the desired regularization. The above expression is readily kernelized from Equation (8.18):

$$\begin{pmatrix} \mathbf{K}_{11} \mathbf{K}_{11} & \mathbf{K}_{11} \mathbf{K}_{22} \\ \mathbf{K}_{22} \mathbf{K}_{11} & \mathbf{K}_{22} \mathbf{K}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{K}_{11} \mathbf{K}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{22} \mathbf{K}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}. \quad (8.20)$$

Finally, by exploiting all the relationships illustrated above, we can solve the problem in Equation (8.17), ending up in:

$$\begin{pmatrix} \mathbf{K}_{11} \mathbf{K}_{11} + \mathbf{R}_{11} & \mathbf{K}_{11} \mathbf{K}_{22} \\ \mathbf{K}_{22} \mathbf{K}_{11} & \mathbf{K}_{22} \mathbf{K}_{22} + \mathbf{R}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{K}_{11} \mathbf{K}_{11} + \mathbf{R}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{22} \mathbf{K}_{22} + \mathbf{R}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}, \quad (8.21)$$

8.3 Multi-sensor alignment for change detection

where \mathbf{R} is the regularization, defined before as $\Omega(f) = \|\mathbf{w}\|^2$.

In this work, we consider a semi-supervised extension of the kCCA (SSkCCA), allowing the projection vectors to account for the geometrical distribution of the data, thanks to a manifold regularization [Belkin et al., 2006]. By changing the regularization that brings to Equation 8.17, we can see that, as for other methods illustrated in this Thesis, we can penalize differently the projection vectors. Again, for instance, we might adopt directly $\|\boldsymbol{\alpha}\|^2$ to obtain small dual weights bringing to $\mathbf{R}_{kk} = \gamma \mathbf{I}_{kk}$.

Belkin et al. [2006] proposed a complete framework to achieve solution that vary smoothly when moving between close samples on the manifold. Their proposition is that the function that maps to the manifold, say f , should be smooth and vary only a little for samples being close on the data manifold. As stated in [Belkin et al., 2006], the regularizer should enforce small $\|\nabla f(\mathbf{x}_i, \mathbf{x}_j)\|^2 = \|\nabla f_{ij}\|^2$, thus penalizing projections that maps samples lying close on the manifold far one to each other. Equivalently, we want to penalize solutions evaluated on close samples that vary rapidly, to enforce the manifold (or smoothness) assumption of semi-supervised learning (see Chapter 3.4.2). By letting the weights q_{ij} indicate if samples \mathbf{x}_i and \mathbf{x}_j are neighbours (i.e. 1 if they lie among the k NN neighbours, 0 otherwise), the following penalization functional may be defined by approximating the Laplace-Beltrami operator:

$$\begin{aligned}
 & \sum_{ij}^n q_{ij} (f(\mathbf{x}_i) - f(\mathbf{x}_j))^2 = \sum_{ij}^n q_{ij} (f(\mathbf{x}_i)^2 + f(\mathbf{x}_j)^2 - 2f(\mathbf{x}_i)f(\mathbf{x}_j)) \\
 & = \sum_i^n q_{ij} f(\mathbf{x}_i)^2 - \sum_{ij}^n q_{ij} f(\mathbf{x}_i)f(\mathbf{x}_j) \\
 & = f_i' \left(\sum_i^n q_{ij} - \sum_{ij}^n q_{ij} \right) f_j = \mathbf{f}'(\mathbf{G} - \mathbf{Q})\mathbf{f} = \mathbf{f}'\mathcal{M}\mathbf{f}.
 \end{aligned} \tag{8.22}$$

where \mathcal{M} is the empirical graph Laplacian, computed as $\mathcal{M} = \mathbf{G} - \mathbf{Q}$. Here, \mathbf{G} is the degree matrix, the sum of the rows of \mathbf{Q} in the diagonal, that in turn is the adjacency matrix between samples \mathbf{x}_i and \mathbf{x}_j indicating if they are neighbours. Finally, by adopting the projection function of the kCCA, i.e. $f(\mathbf{x}) = \mathbf{K}\boldsymbol{\alpha}$, we can rewrite the Equation 8.22 as $\|\nabla \mathbf{f}\| = \|\nabla \mathbf{K}\boldsymbol{\alpha}\| = \boldsymbol{\alpha}'\mathbf{K}'\mathcal{M}\mathbf{K}\boldsymbol{\alpha}$. Summing up, we can include this additional manifold regularizer in Equation (8.21), with $\mathbf{R}_{\bar{k}\bar{k}} = \gamma \mathbf{K}_{\bar{k}\bar{k}} + \delta \mathbf{K}_{\bar{k}\bar{k}} \mathcal{M}_{\bar{k}\bar{k}} \mathbf{K}_{\bar{k}\bar{k}}$, with hyperparameters γ and δ to be tuned, controlling the penalization of large norms of $\mathbf{w}_k^{\mathcal{J}}$ and the deformation by the graph Laplacian respectively. Note that the subscript \bar{k} indicates the expanded training set \mathbf{X} using samples chosen randomly from the k th view, resulting in a set $\mathbf{X}_{\bar{k}} \in \mathbb{R}^{(n_s+u) \times d}$. The kernel $\mathbf{K}_{\bar{k}\bar{k}}$ contains the evaluations between \mathbf{X}_k and $\mathbf{X}_{\bar{k}}$. The graph Laplacian has been estimated using standard k NN links [Belkin et al., 2006]. A similar formulation of the regularizer leading to the Laplacian SVM [Belkin et al., 2006] has been adopted for remote sensing image classification purposes in [Gómez-Chova et al., 2008], verifying the intuitions that by penalizing highly varying solutions $\mathbf{K}\boldsymbol{\alpha}$ an improved generalization may be obtained.

8. Feature extraction for change detection

The final formulation of the semi-supervised kernel CCA is:

$$\begin{pmatrix} \mathbf{K}_{\bar{1}\bar{1}}\mathbf{K}_{\bar{1}\bar{1}} + \mathbf{R}_{\bar{1}\bar{1}} & \mathbf{K}_{\bar{1}\bar{1}}\mathbf{K}_{\bar{2}\bar{2}} \\ \mathbf{K}_{\bar{2}\bar{2}}\mathbf{K}_{\bar{1}\bar{1}} & \mathbf{K}_{\bar{2}\bar{2}}\mathbf{K}_{\bar{2}\bar{2}} + \mathbf{R}_{\bar{2}\bar{2}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{K}_{\bar{1}\bar{1}}\mathbf{K}_{\bar{1}\bar{1}} + \mathbf{R}_{\bar{1}\bar{1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{\bar{2}\bar{2}}\mathbf{K}_{\bar{2}\bar{2}} + \mathbf{R}_{\bar{2}\bar{2}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}, \quad (8.23)$$

Thanks to this double regularization, the solution favors small norms of $\mathbf{w}^{\mathcal{H}}$ and at the same time forces samples lying close on the manifold structure to be projected nearby. This property results particularly useful in multi-source change detection. Assuming that pixels lie in a lower dimensional subspace, and that the manifolds coming from heterogeneous sources behave similarly (e.g. class distributions), the SSkCCA solutions optimizing Equation (8.21) promote a solution lying on the geometrical structures and less affected, thanks to the CCA itself and by the regularization, to overfitting induced by noise and high data variances.

8.3.3 Heterogeneous alignment for change detection

In contrast to multivariate alteration detection approaches [Nielsen, 2007; Nielsen et al., 1998], a measure of change as the variance in the projected space is not directly optimized. A statistical alignment of unchanged samples is optimized instead. Since the eigenvectors corresponding to leading eigenvalues are retained, it results into a more stable solution, and the change information can be obtained as the difference of the first q canonical variates. However, this comes at the cost of obtaining a set of some labelled unchanged pixels.

Experimental setup To test the ability of the proposed system to perform multi-sensor change detection, we considered four settings using the Greece Island dataset (see Appendix C.6). First, change detection aligning all the available bands of the bi-temporal images (without the thermal band) has been carried out, as the baseline indicating performances when disposing of a maximal amount of input information. It corresponds to bands (1-5,7) of the TM sensor (6 vs 6 bands setting). In the second setting, the Landsat TM image at t_1 is complete, while for the image at t_2 only channels 1-3 are retained (6 vs 3 bands). In the third case, the same problem is considered but t_2 is composed now by bands 2-4 of the TM sensor (6 vs 3 bands). A last experiment involving an extreme alignment, is performed by transforming the original t_1 against bands 5 and 7 of the TM sensor (6 vs 2 bands).

In all the cases, SSkCCA results are compared to those obtained by aligning the datasets with standard (primal) linear CCA and to those obtained with the original image after histogram matching (HM). However, this last setup for the spectrally downsampled images requires the same number of bands. Therefore, in the multi-sensor experiments, the HM models are obtained by spectrally downsampling of the t_1 acquisition to match the t_2 data. It is referred to as REDU hereafter.

8.3 Multi-sensor alignment for change detection

ID	Method	Dims.	κ (std)	NMI (std)
Single-sensor 6/6 (1-5,7)	LDA + SSkCCA	7	0.86 (0.05)	0.64 (0.08)
	LDA + CCA	6	0.86 (0.03)	0.65 (0.04)
	LDA + HM	6	0.72 (0.01)	0.44 (0.01)
	CVA + SSkCCA	3	0.55 (0.07)	0.26 (0.06)
	CVA + CCA	3	0.47 (0.10)	0.23 (0.10)
	CVA + HM	6	0.32 (0.05)	0.80 (0.02)
Multi-sensor 6/3 (1,2,3)	LDA + SSkCCA	3	0.82 (0.02)	0.58 (0.02)
	LDA + CCA	6	0.71 (0.05)	0.46 (0.05)
	LDA + REDU	3	0.66 (0.01)	0.36 (0.01)
	CVA + SSkCCA	2	0.50 (0.17)	0.23 (0.11)
	CVA + CCA	3	0.34 (0.04)	0.10 (0.03)
	CVA + REDU	3	0.28 (0.02)	0.07 (0.01)
Multi-sensor 6/3 (2,3,4)	LDA + SSkCCA	6	0.90 (0.01)	0.71 (0.04)
	LDA + CCA	3	0.78 (0.03)	0.50 (0.04)
	LDA + REDU	3	0.70 (0.02)	0.41 (0.02)
	CVA + SSkCCA	3	0.60 (0.07)	0.31 (0.08)
	CVA + CCA	3	0.44 (0.04)	0.19 (0.04)
	CVA + REDU	3	0.29 (0.05)	0.07 (0.02)
Multi-sensor 6/2 (5,7)	LDA + SSkCCA	10	0.77 (0.07)	0.51 (0.04)
	LDA + CCA	2	0.63 (0.04)	0.33 (0.04)
	LDA + REDU	2	0.57 (0.01)	0.27 (0.01)
	CVA + SSkCCA	2	0.55 (0.07)	0.26 (0.07)
	CVA + CCA	2	0.43 (0.09)	0.26 (0.07)
	CVA + REDU	2	0.16 (0.03)	0.03 (0.01)

Table 8.1: Relative radiometric normalization in heterogeneous sources, change detection results - Change detection results using original images and using three settings simulating heterogeneous images. Here, $n = 50$ and $u = 200$.

All kernels are Gaussian RBF with a scale parameter equal to the median distance among 3000 pixels randomly selected from the corresponding image. To test the sensitivity to the size of both sets, composed by the samples from the unchanged regions and the ones added to estimate the regularizer, their number has been varied. For this study, the regularization parameters have been tuned by cross-validation on the experiment involving $n_s = 50$ labelled and $u = 200$ unlabelled samples, resulting in $\gamma = 0.1$ and $\delta = 0.001$. The number of neighbours used to compute the graph Laplacian is 10. Finally, to detect changes, the standard change vector analysis (CVA) [Bovolo and Bruzzone, 2007] and the supervised linear discriminant classification (LDA) [Shawe-Taylor and Cristianini, 2004] are used on the difference image. For the former, 100 randomly selected validation pixels (50 per class) have been used to tune the threshold of the CVA norm. The same 100

8. Feature extraction for change detection

pixels have been used to train the LDA. The test set is common to all the experiments, while the training sets are varied five times to account for stability with respect to random initializations. Numerical results are assessed by the estimated Cohen’s Kappa coefficient κ and by the Normalized Mutual Information (NMI) between the predictions and the corresponding ground truth samples (see Appendix B).

Results and discussions Table 8.1 reports the performances of change detection methods applied in the settings tested. The number of dimensions used to compute the difference image are illustrated in the column “Dims.”. They correspond to the dimension providing the best average accuracy when using 1 to 20 dimensions for the projection of the aligned images. In Figure 8.9(a)-(b) and Figure 8.10(c)-(d), the complete accuracy curves are illustrated.

First setting: When using all the available information, SSkCCA + LDA and CCA + LDA perform very similarly, as depicted by the accuracy scores. The LDA on the original data after histogram matching performs worse, with a loss of approximately 0.14 κ points. By observing the CVA performance, which is indicative of the degree of separation of the classes in the projected space, the proposed SSkCCA + CVA performs around 0.23 κ better than its CVA + HM counterpart. The tested baselines, except for the LDA + CCA approach, are significantly less accurate than the adopted method.

Second setting: The setting involving the alignment of the original TM data to the first three channels (roughly corresponding to RGB components), the LDA + SSkCCA and CVA + SSkCCA performed again very well, with 0.82 and 0.50 κ , respectively. These accuracies are only slightly inferior to the ones obtained with the full sets. Linear CCA applied with LDA and CVA performed again significantly better than HM on the reduced dataset, but with accuracies significantly worse to the ones obtained with a comparison involving 6 bands on both dates.

Third setting: This experiment provided accuracies surprisingly higher than the ones obtained on the full images. However, this is only verified for the change detection methods applied on the transformed images after the SSkCCA alignment. The LDA + SSkCCA performed 0.04 κ scores better than the full-band alignment counterpart, while the CVA + SSkCCA improved by 0.05 κ the accuracy.

Fourth setting: In the last experiment, an extreme situation involving the alignment of two sets of 6 and 2 variables respectively is illustrated. In this case, the LDA + SSkCCA accuracy is the worse of all the similar experiments, but recall that here only 2 bands at t_2 are available for the alignment. The CVA + SSkCCA, on the contrary, performed as in the original 6 vs 6 channel matching, clearly demonstrating that the adopted system is able to improve the separability of the classes by leveraging all the available information, from both the spectral and the geometrical distributions.

By looking at the change maps obtained after the SSkCCA-based alignment, the spatial coherence relates well with the accuracies of Table 8.1. The most accurate maps among the 5 runs are illustrated in Figure 8.7. From a spatial homogeneity perspective, the best maps are the one obtained by the SSkCCA on the NIR-R-G set (the third setting).

8.3 Multi-sensor alignment for change detection

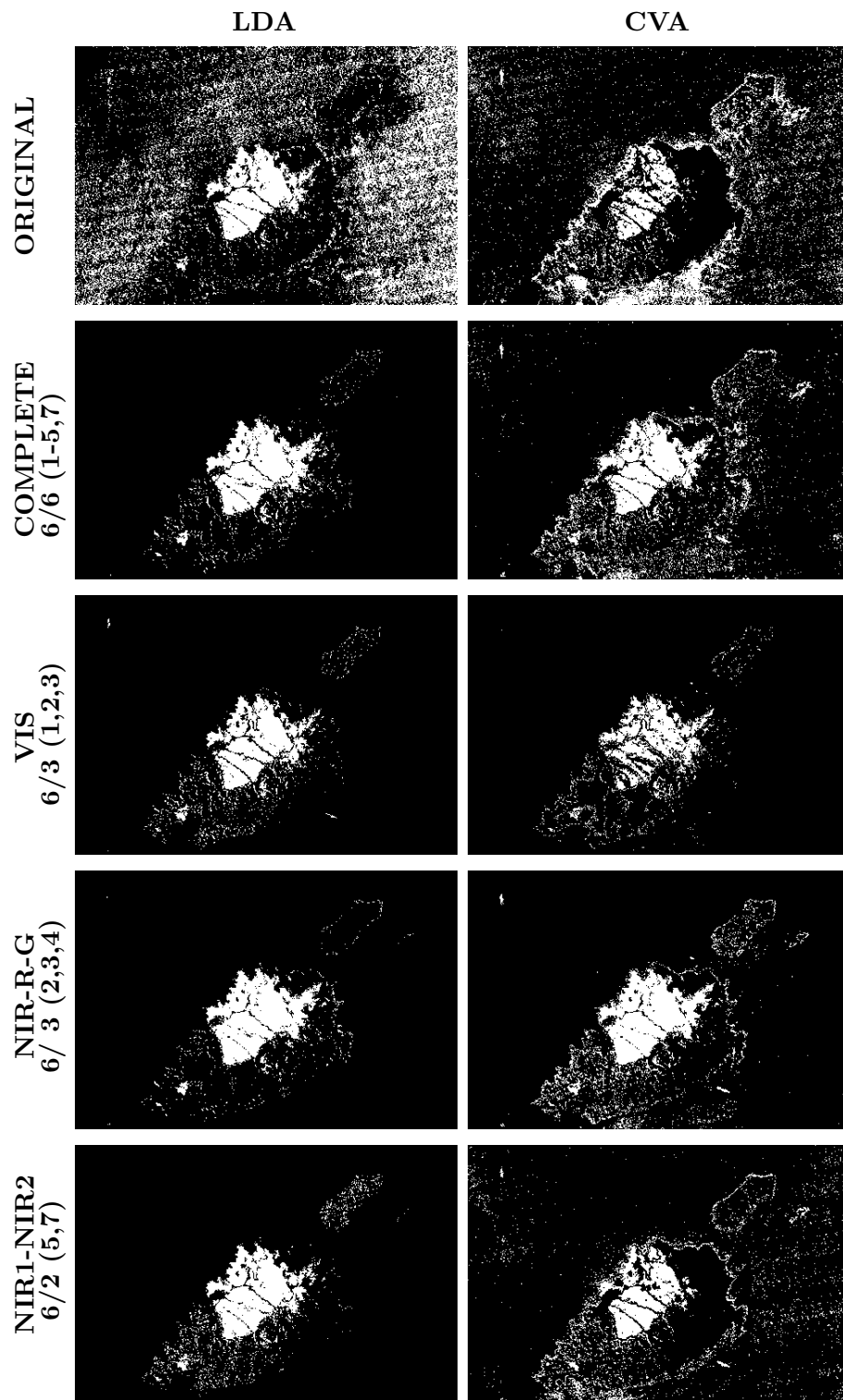


Figure 8.7: Change detection maps for the tested SSkCCA alignment approaches.

8. Feature extraction for change detection

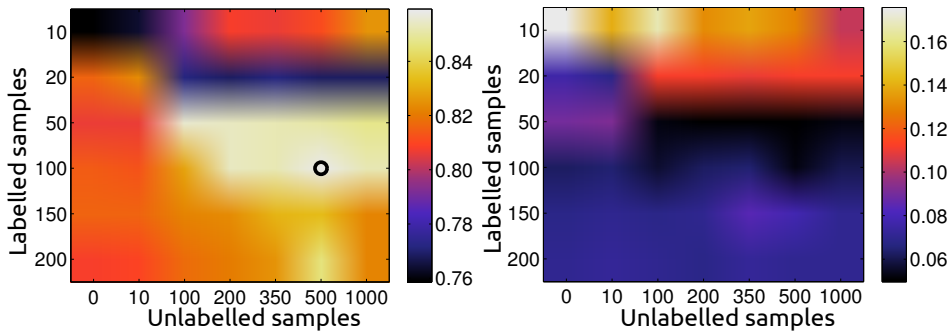


Figure 8.8: Dependence on the number of labelled and unlabelled pixels - Performance of the LDA + SSkCCA in the first experiment involving 6 vs 6 bands. Left panel represents the κ coefficient and right panel its standard deviation for the 5 runs. The maximal accuracy is represented by the black circle ($\kappa = 0.86$).

Figure 8.9(a)-(b) and Figure 8.10(c)-(d) illustrate the sensitivity of the change detection average accuracy to the number of retained dimensions for the SSkCCA-based alignment methods. While LDA is relatively robust to the number of dimensions, the CVA suffers large dimensionality. As discussed for the kPCA approach, this is related to the inflation of the difference image magnitude, worsening the discriminative information. The higher the discriminative information contained in the sets to be aligned, the more stable and higher are the accuracy curves.

Figure 8.8 studies the role of the number of labelled (unchanged) and unlabelled pixels used in the SSkCCA + LDA in the first setting (6 vs 6 bands). For the other experiments, a similar structure but with higher variance was observed. The method needs a minimal number of labelled unchanged pixels (typically 20) to find a proper normalization of the heterogeneous dataset. The contribution of the unlabelled samples is underlined by observing, for a given n_s , the increase of the κ score with respect to the size of u (left panel). Also the standard deviation greatly decrease as the size of the sets increases (right panel). For training sets larger than 50 samples from the unchanged regions plus 100 unlabelled pixels, the accuracy is stable around 0.82-0.84 κ reaching its maximum for $n_s = 100$ and $u = 500$ with a $\kappa = 0.86$ (and stabilizing for $u \geq 500$).

8.3.4 Discussion

The proposed multi-temporal transformation improved the performance of the change detection process. The benefits of the nonlinear relative normalization appears clearly, thanks to the regularization on the manifold penalizing noise and outliers, while favouring smooth solutions. Moreover, depending on the data to which the images are matched to (e.g. VIS, NIR-R-G, NIR-NIR, etc.) the enhancement of the discrimination by the proposed alignment technique is further boosted. For instance, by considering a problem of change detection in vegetated areas, the IR channels may provide a very discriminant view useful to align properly the other data to information correlated with the event

8.3 Multi-sensor alignment for change detection

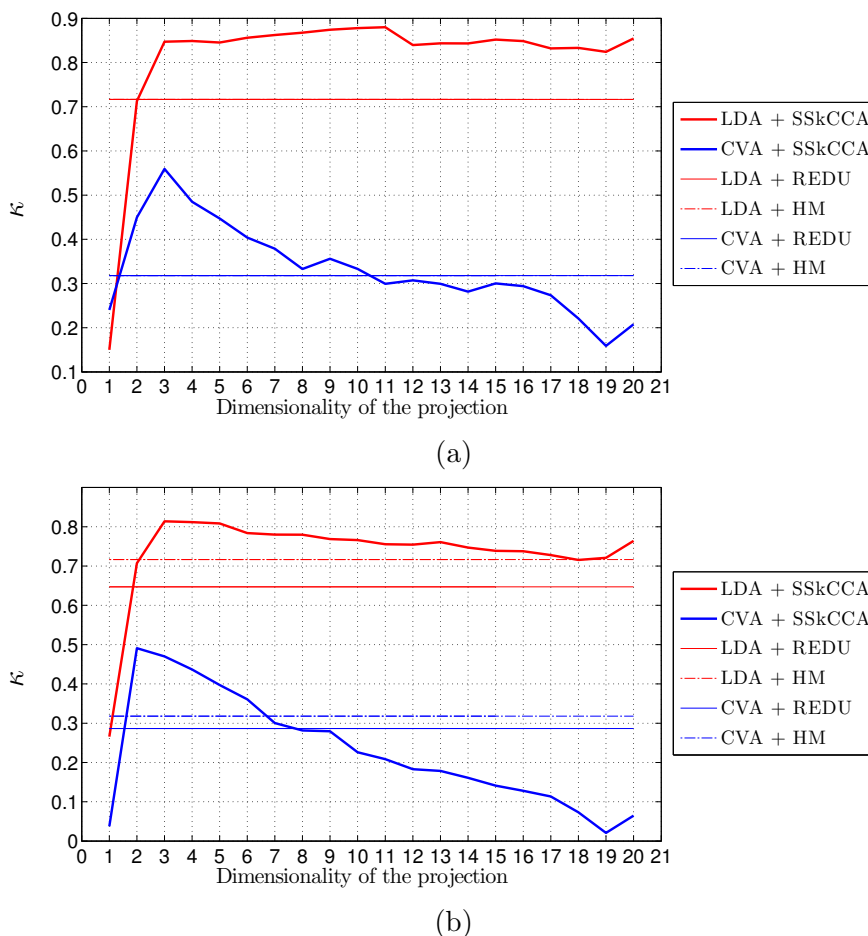
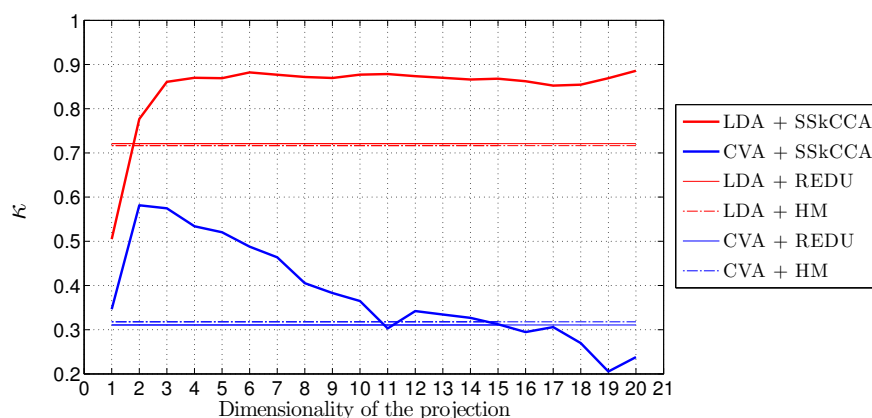


Figure 8.9: Accuracy as a function of the dimensionality of the projections - In (a) experiments involving the alignment of the two complete images. In (b) alignment of the original data to the VIS set, corresponding to TM bands 1-3.

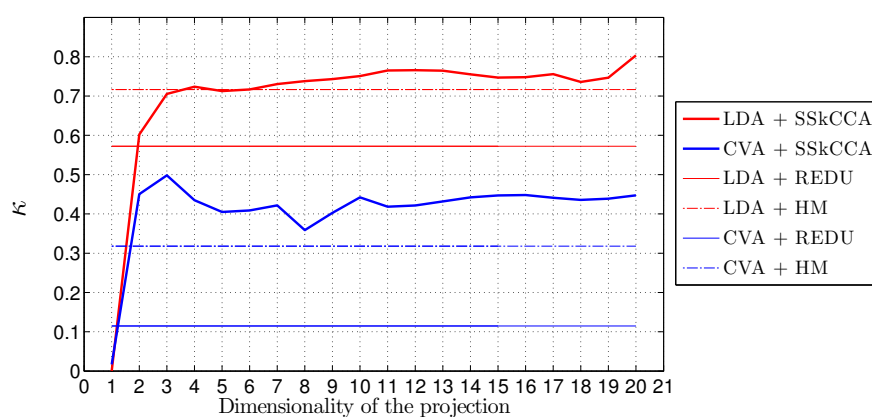
generating the changes. This is the case of the difference in accuracy of the second and the third experiment (VIS and NIR-R-G). In this case, the final change detection accuracy is very high, since it correlates unchanged samples more easily. This suggests that by adding additional – but relevant – information, the creation of a more discriminant difference image is possible.

However, to benefit from all the good properties of the presented approach, the free hyperparameters of the classifiers / detectors should be correctly tuned. In these experiments some labels were available: some unchanged pixels to train the feature extraction and another small set containing examples from both the classes to tune the hyperparameters. The fitting was possible thanks to cross-validation on the final change detection error, for both the CVA and the LDA. Optimal regularization parameters and dimensionality were easy to obtain in such setting. Note that the Gaussian kernel bandwidth was set as the median Euclidean distance among a randomly chosen set of samples.

8. Feature extraction for change detection



(c)



(d)

Figure 8.10: Accuracy as a function of the dimensionality of the projections - In (c) experiments involving the alignment of the original image to the NIR-R-G (TM bands 2-4). In (d) alignment of the original data to the NIR1-NIR2 (TM bands 5,7) set.

It is worth concluding by pointing out that the change detection step was applied independently to the transformation, so that the user may adopt its preferred change detection methods. The latter may rely on the difference of reprojected images or to the stack of transformed images. This approach allows to implement multi-sensor change detection approaches by exploiting known and standard change detection systems, by simply enhancing and matching the input images after a nonlinear transformation.

8.4 Conclusions

In this Chapter we discussed two approaches for the relative radiometric normalization through the use of nonlinear feature extraction techniques. The former, relying on the kernel PCA, finds a nonlinear projection on the basis of the variance of a joint set of unchanged pixels from the bi-temporal images. The images are then mapped into a new

space, in which data are rotated towards the direction aligning the most.

In the second case, the above mentioned approach is extended to account for better and more stable projections, by coupling it to a regularization accounting for the geometrical distribution of the data through its manifold. This way, we can use additional pixels without the need of knowing their label by adding them to the data matrix collecting unchanged samples. Additionally, the adopted technique explicitly considers the two images as disjoint sets of variables, i.e. as different views of the same examples (pixels). For this reason, the approach is able to handle data with different dimensionality by construction and allowed the computation of an enhanced difference image even if the spectral channels and spectral information of the original images are not the same. The change detection step is an independent procedure which enable the user to apply its own preferred change detection technique, either supervised or unsupervised.

In this Chapter we demonstrated that the change detection algorithm itself is not the only important issue to consider to obtain an accurate change map. Here, we paid attention to the creation of a space in which the images were the most comparable, and allowing an enhancement of standard methods. In all the experiments presented in this Chapter, the projection of the data into a common subspace improved the detection of changes by the use of the standard difference image.

8. Feature extraction for change detection

Part IV

Conclusions

Chapter 9

Conclusions

9.1 A new generation of change detection systems

As discussed along the Chapters of this Thesis, kernel methods have greatly contributed in remote sensing image processing tasks, by providing flexible and nonlinear solutions to complex data analysis problems. The range of applications in which these methods may positively contribute is spreading each year, but up to now only few systematic studies have been addressed to multi-temporal image processing and in particular to change detection with kernel-based algorithms. This Thesis contributed in a better understanding of the issues involving multi-temporal analysis and kernel methods, and it constitutes a step further towards real world implementations of kernel-based change detection systems. It is also emphasized that many kernel algorithms are obtained by reformulating standard methods known to work well on real world problems (e.g. PCA, k -means, etc.). However, a series of issues have first to be carefully addressed, and these may range from the choice or creation of an adapted kernel function, the optimisation of the corresponding hyperparameters or the representation of the input data, and they all vary strongly depending on the considered task.

In this Thesis, we considered one of these tasks: change detection. In particular, the topic has been studied by attacking the problem from three different perspectives: by adopting supervised classification models, by reformulating unsupervised clustering and applying feature extraction algorithms. In most of the experiments aimed at validate the proposed algorithms, kernel-based methods improved and outperformed the baseline models found in the literature.

9.1.1 On supervised change detection

Supervised classifiers provide very accurate and exhaustive thematic classification of multi-temporal datasets. In Chapter 6 we adopted and improved such system by enriching the input space of the classifier with appropriate spatial and contextual information extracted from the images. By doing so, we observed that the multi-temporal classification may be carried out accurately also on VHR images. To handle the higher dimensional data

9. Conclusions

spaces, the use of a robust classifier was mandatory. The price of the computational time, for both the filtering and classification, was consequently high. However, we can fairly say that for many monitoring tasks the increase in accuracy is worth the computational price. In addition, for classifiers such as SVM, large scale and fast implementations exist, and it is likely that the trend continues in the next future, by the constantly increasing computational power offered by personal computers.

By solving the problems of supervised change detection in VHR data, the addition of spatial features has been demonstrated to be very beneficial. However, to further generalise these approaches, the subjective bias caused by the intervention of the user when selecting the appropriate spatial filters and their parameters should be removed. Recently, automatic schemes have been developed and provided promising results. In particular, we mention the multiple and composite kernel frameworks [Camps-Valls et al., 2006; Tuia et al., 2010a] or automatic feature learning schemes [Tuia et al., 2012]. In both cases, the classifier optimizes its input space on their contribution to the overall classification, by either weighting the kernel corresponding to a particular group of features or information sources, or by selecting the filters and operators improving a large margin separation between classes.

These systems could be successfully applied for supervised change detection. Promising future directions rely on the exploitation and inclusion of complementary information sources into the multi-temporal analysis process, such as radar images, digital elevation models or spectral indexes. In general, kernel methods offer the tools to perform such integration.

9.1.2 On unsupervised change detection

In contrast to applications that require powerful methods to correctly exploit complex and high dimensional input spaces, there is a series of important real world scenarios that rely on a fast and reliable mapping of the events. Automatic and unsupervised change detection methods are of paramount importance for post-catastrophe and natural hazard related applications. In these cases, the methods must be able to provide accurate solutions within a short time instant, with possibly limited or no user intervention. Moreover, the changes to be detected may be spectrally ambiguous. This is the case of earthquakes, where destroyed building are hardly discriminable from the intact ones.

Again, kernel methods provide a robust and simple formulation allowing for automatic, rapid and accurate change detection. As illustrated in Chapter 7, simple algorithms may be (re-)formulated using kernels, ensuring the correct modelling of nonlinear relationships and only requiring a little more computational efforts. A classical domain specific expression such as the difference image has been reformulated through the use of kernel functions, corresponding to a difference image computed in RKHS providing an improved representation. In this high dimensional space, standard and fast algorithms provide much more accurate results. Nevertheless, one drawback observed was the increase of the amount of free hyperparameters to fit. In Chapter 7 this problem has been tackled by proposing a

heuristic cost function relying on geometrical criteria showing a minimum when the optimal clustering is encountered, thus implicitly defining the set of optimal hyperparameters to retain.

In future studies, unsupervised and automatic methods should be extended to the inclusion of ancillary data such as SAR images and digital surface model, in order to exploit the data complementarity and to improve the change detection with easily available data. To this end, besides standard approaches to data fusion, the methods proposed in Chapter 8 may result very useful.

9.1.3 On multi-sensor change detection

The last research topic studied in the Thesis aimed at the statistical alignment of the distribution of unchanged samples, to improve change detection models based on direct pixelwise image comparison. In particular, two feature extraction methods have been studied: kernel principal component analysis and kernel canonical correlation analysis. The use of the former verified the assumptions that, by mapping the multi-temporal image into a common subspace, algorithms may benefit of an improved data separability.

The second approach established such alignment on the maximization of the correlation of two mutual projections of the original input images. It allowed, by construction, the alignment of datasets of different dimensionality, yielding natural multi-source change detection schemes using standard methods. The tested experimental setting illustrated promising results even when the number and type of spectral channels employed was drastically different. As for the first method, the images are projected into a common subspace where unchanged samples are maximally close, and discrepancies in multi-temporal information are enhanced for changed areas only.

The accuracy of supervised and unsupervised methods for change detection has been drastically improved by preprocessing the images exploiting these findings. In this Thesis, the experiments have been limited to change detection tasks, but the last approach could be applied for correlating the images to different sources of information, thus resulting in a general purpose data fusion method. The fused information may be used for subsequent classification, regression or density estimation tasks by letting the user choosing its preferred algorithms. Images can be aligned to other corresponding spatial data, such as radar images, digital elevation models, precipitation and temperature maps, to enhance the detection of specific ground cover classes or processes. Further research has to be deployed in the analysis of the statistical behaviour of real world image fusion with heterogeneous sources.

9.2 Contributions of the Thesis

The main contributions of the Thesis can be summarized as follows:

- The development of novel insights on the application of kernel-based algorithm to a variety of problems encountered in multi-temporal image analysis.

9. Conclusions

- The adaptation and verification of the use of spatial context into multi-temporal classification and supervised change detection systems. Such an inclusion resulted in a scheme well suited for monitoring purposes relying on VHR images.
- The analysis of two kernel-based classifiers for multi-temporal monitoring purposes in relation to the input space provided.
- The development of an automatic and unsupervised change detection method relying on kernels. It provides a fast and stable result also in challenging situations with a small computational effort.
- The development of a cost function for selecting kernel hyperparameters in an unsupervised partitioning framework, avoiding the user intervention.
- The study of kernel-based feature extraction methods for the improvement of the statistical alignment between unchanged regions.
- The development of a kernel-based system allowing the projection of multi-source images into a common subspace, in which standard methods are effective.

9.3 Future perspectives

In change detection As illustrated along the Thesis, kernel methods offer a complete and robust framework for the analysis of multi-temporal remote sensing images. Although the methods proposed seem promising, fundamental research is still required to overpass some issues clearly limiting current change detection techniques. In particular, we may emphasize the following key points:

Cross the limit of the perfect coregistration. For tasks involving the detection of novel classes, i.e. in an anomaly detection setting, the strict coregistration is not required and the methods adopted may work on the pooled datasets. However, when changes are due only to differences in the spatial location of the same classes present in both images, the detection has to be performed pixelwise, consequently suffering from errors due to spatial misregistrations. Usually, a perfect co-registration is always assumed, and the preprocessing required to obtain errors at pixel level is costly and time consuming. Recent works have been devoted to study these issues, e.g. [Bovolo et al., 2009; Theiler and Wohlberg, 2012] and future extension of change detection algorithms should be able to improve the robustness with respect to registration errors.

Exploit all the information. As mentioned many times, Earth observation applications require products of increasing temporal coverage, quality and accuracy. To this end, the very frequent acquisitions, more and more similar to continuous streams of remote sensing images, have to be efficiently processed. Since atmospheric conditions may strongly limit the use of single sensor imagery (e.g. clouds), a logical solution to overcome these gaps is to use images from multiple sensors. In terms of change detection algorithms, new methods should be able to efficiently compare images with different spectral and spatial resolutions and provide maps indicating where changes occurred. Possibly, to be able to exploit and assimilate the frequent acquisitions,

algorithms should be able to adapt themselves to the data at hand with minimal user supervision. Observations issued from the Chapter 8 are very promising, and should be further considered for an efficient integration of multi-modal data in the change detection process. In this sense, recently developed fields of multi-view learning and domain adaptation could provide insights for the implementation of multi-sensor change detection.

In remote sensing image processing In a more general context, the processing of new generation remote sensing VHR images is still limited by strong assumptions that are no longer valid for these images. When we look at VHR data, we can easily extract infinite amounts of information, by simply looking at the objects, their colours, their spatial locations, their shape, mutual similarities and so on. On these bases, we may underline the following key points to improve the methods of remote sensing image analysis:

Generalise from pixels to objects. VHR data are complex, due to their intrinsic multi-scale nature. To be able to extract relevant information, objects and semantically coherent regions composing the image should be considered and processed as entities. Future methods involving the processing of VHR data should switch from a pixel-based representation to an object-based one. By doing so, a more compact, meaningful and realistic representation of the image is obtained, while keeping different degrees of scale information and respecting the precise spatial arrangement of the pixels. For these developments, computer vision approaches could greatly help for spatially and semantically coherent analyses and understanding of images.

Analyze relationships between objects. By assuming a perfect segmentation, the extraction of the relevant information from the semantically coherent regions needs the use of adapted analysis methods. To be able to obtain a high level processing, features characterizing the objects, such as texture, size, shape, colour, location, edges, orientation, and, more importantly, the spectral information, should be correctly considered. To deal with such a newly created feature space, kernel methods offer many state-of-the-art solutions. In particular, we may mention the use of structured classifiers respecting class hierarchies, multi-task learning, manifold learning of objects, data fusion, domain adaptation and multi-view learning. In other words, we want to transform unstructured and uninformative objects (the pixels) to a structured and valuable representation, from which extract the information needed to process the data. By understanding the relationships between objects and correctly encoding the discriminant characteristics, the need of labelled information may be strongly reduced and the transfer of information from one task to another one can be accomplished more easily, thus leading to more general methods of remote sensing image interpretation.

Only a more strict synergy between the (already strongly related) domains of image processing, vision and machine learning to the field of remote sensing image processing and interpretation could bring to a significant improvement in the tools available for the processing of new generation data.

9. Conclusions

Appendices

Appendix A

The learning sets

When approaching a statistical data modelling problem, one usually takes advantage of the data samples that dispose at the moment of learning. Here, we give a formal definition of the different subsets required during the training and assessment of the generalization ability of a model. In particular, we distinguish between training, testing and validation sets. Formal explanation of the process in which they are implied, are deemed to Section 3.

Dataset The dataset is a collection of all the available samples, labelled or not. In this Thesis, we distinguish between the image X and the data matrix \mathbf{X} . The two sets of samples only differ in their organization: while an image $X \in \mathbb{R}^{(N \times M \times d)}$ is arranged into d spectral channels of size $M \times N$ pixels each, the data matrix is reshaped so that $\mathbf{X} \in \mathbb{R}^{(N \cdot M \times d)}$. In each row of \mathbf{X} one finds the pixels as \mathbf{x}' , while in the columns the d spectral values for each sample.

Training set The training set \mathbf{X}_s is a labelled subset drawn from the original data \mathbf{X} , and it is used to train a model, i.e. estimating model parameters to fit the data. As it will be detailed in the dedicated sections, in supervised learning we dispose of n_s training pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_s}$, thus $(\mathbf{X}_s \times Y_s) \subset (\mathbf{X} \times Y) \in \mathcal{X} \times \mathcal{Y}$, corresponding to input-output couples from their respective spaces (Chapter 6). Otherwise, when disposing only of samples $\{\mathbf{x}_i\}_{i=1}^{n_s} \in \mathcal{X}$ one recurs to the use of unsupervised techniques (Chapter 7). If a situation between the two occurs, the use of semi-supervised models may be foreseen (Chapter 8).

Test set This set is another disjoint labelled subset of the original dataset $(\mathbf{X}_t \times Y_t) \subset (\mathbf{X} \times Y) \in \mathcal{X} \times \mathcal{Y}$ used to evaluate the generalization accuracy of the final model (an approximation of the goodness of the model with respect to new samples). A test set is composed by n_t pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_t}$, so that a generalization error can be estimated and used to rank final models. It is important to point out that the samples of this set have never been used neither in the training nor in the choice of the hyperparameters. In this case the test set is independent from the training process of a model, and it can assess its accuracy with respect to the process generating the data $P(\mathbf{x}, y)$.

In remote sensing applications it is often required that this set comes from regions that are spatially disjoint from the ones producing the training samples. By doing

A. The learning sets

so, biases in the estimation of the generalization error due to training-testing spatial autocorrelation can be avoided. However, since the labelling process is costly, this requirement is not always fulfilled.

Validation (or development) set The validation set $(\mathbf{X}_v \times Y_v) \subset (\mathbf{X} \times Y) \in \mathcal{X} \times \mathcal{Y}$ is an additional labelled subset, independent from the training and test sets, used to estimate the performance of the model $f(\mathbf{x})$ trained on \mathbf{X}_s under different hyperparametrizations. The lowest validation error defines the hyperparameters that will be retained for training the final model. However, since labelled data are usually scarce in real world scenarios, one may not dispose of such set (see Section 3.3).

Appendix B

Accuracy evaluation

At the end of the change detection process, one must evaluate the goodness of a map by a measure of performance. On the basis of this score, one is able to decide whether the map is accurate enough to be used, or simply to rank and compare different models. It is a good practice to use different scores when evaluating a classification map, since they provide different insights on the models, in particular for unbalanced problems or correlated errors. The figures of merit used in this Thesis are based on the comparison between the predicted and true labels (the ground truth) for the test set, composed by n_t samples. This comparison is made through the use of an error matrix (Confusion matrix, Table B.1), counting the number of times a true sample has been assigned into the different predicted classes.

For binary problems, it can be reduced to a 2×2 table summarizing the correct predictions (true positives and true negatives) and the wrong ones (false positives and false negatives). By exploiting the frequencies of these categories, the following metrics can be computed:

Overall Accuracy (OA) is expressed as the ratio of correctly classified samples over the grand total of test pixels, ranging in $[0, 1]$ or expressed in percentages. It has a straightforward interpretation, but note that OA is biased for unbalanced class

		Observed				
Class		1	2	\dots	c	User acc.
Predicted	1	n_{11}	n_{12}	\dots	n_{1c}	$n_{11}/\sum_i n_{1i}$
	2	n_{21}	n_{22}			$n_{22}/\sum_i n_{2i}$
	\vdots	\vdots		\ddots		\vdots
	c	n_{c1}	n_{c2}		n_{cc}	$n_{cc}/\sum_i n_{ci}$
Producer acc.		$n_{11}/\sum_i n_{i1}$	$n_{22}/\sum_i n_{i2}$	\dots	$n_{cc}/\sum_i n_{ic}$	$\sum_i n_{ii}/n$

Table B.1: Confusion matrix - Observed and predicted classes are compared in order to establish a prediction accuracy.

B. Accuracy evaluation

problems (change occurs usually in a fraction of the total image). It is computed as, for n_{ii} correctly detected samples for class i and n_t total pixels:

$$\text{OA} = \frac{\sum_{i=1} n_{ii}}{n_t}. \quad (\text{B.1})$$

The overall error is defined by $1 - \text{OA}$ (or $100 - \text{OA}[\%]$). The marginal classwise accuracies are denoted as user's (commission rate) and producer's accuracies (omission rates).

Cohen's Kappa statistic (κ) [Foody, 2004] estimates the agreement between two maps by compensating the overall accuracy by the chance of random agreement. By this correction, the effect of large classes is partially compensated. It ranges in $[-1, 1]$, with values -1 if the models specularly disagree, 0 if the agreements are due to chance, and 1 if models perfectly match. It is evaluated as:

$$\kappa = \frac{p(c) - p(r)}{1 - p(r)}, \quad (\text{B.2})$$

where $p(c)$ is the agreement rate (the overall accuracy, expressed in $[0, 1]$) and $p(r)$ is the agreement due to chance, computed as the product of the class-wise fractions of correctly detected classes (over n_t), plus the product of the fractions of the predicted classes (over n_t).

Rand's Index was introduced by W. M. Rand in 1971 [Rand, 1971]. It is designed to penalise correct outcomes due to chance. It ranges in $[0, 1]$, with value 0 for completely random outcomes and 1 for perfect matches. It is also well suited to evaluate unbalanced classification problems. It is calculated as:

$$\text{Rand I} = \frac{\sum_i n_{ii}}{\sum_i n_{ii} + \sum_{i \neq p} n_{ip}}, \quad (\text{B.3})$$

where n_{ii} is the number of agreements between the model and the ground truth. The second term at the denominator counts the number of disagreements.

Normalized mutual information (NMI) is a multi-class measure of agreement relying on information theory [Cover and Thomas, 1991]. The agreement score is given by normalizing the mutual information $\text{MI}(\hat{Y}, Y)$ between the predicted \hat{Y} and true Y class assignments, with the average entropy of the independent labellings (true labels and predicted ones) $\text{H}(\hat{Y})$ and $\text{H}(Y)$ respectively. It ranges in $[0, 1]$ and is very appropriate for unbalanced problems. It is estimated as:

$$\text{NMI} = \frac{\text{MI}(\hat{Y}, Y)}{\text{H}(\hat{Y}) + \text{H}(Y)}, \quad (\text{B.4})$$

ROC curves and AUC. The receiver operating characteristic curve (ROC) and the corresponding area under the ROC curve (AUC) are measures used to assess the performance of binary classifiers. The true positive and false positive rates are analysed by varying the decision threshold and after plotting them on a true positive - false

positive plane, they result in a curve [Fawcett, 2006]. The area under the curve is a measure of classification accuracy, and it ranges in $[0.5, 1]$. It indicates the performance from random (0.5) to perfect (1).

To further support the ranking of models, statistical significance tests may be needed to assess whether a difference in accuracy is significant or not. In this Thesis, the statistical significance is assessed through the use of the one-tailed McNemar test for related samples, with a squared test statistic z^2 following a χ^2 distribution with 1 degree of freedom [Foody, 2004]:

$$z^2 = \frac{(M_{12} - M_{21})^2}{M_{12} + M_{21}}. \quad (\text{B.5})$$

where M_{12} is the number of samples wrongly classified by model 2 but correctly allocated by model 1, and M_{21} refers to the inverse situation. The hypothesis of a better accuracy of model 1 is then compared to tabulated values for different confidence levels. When the number of test samples is small, one may want to adopt the continuity corrected statistic:

$$z^2 = \frac{(|M_{12} - M_{21}| - 1)^2}{M_{12} + M_{21}}. \quad (\text{B.6})$$

B. Accuracy evaluation

Appendix C

Datasets

C.1 Brüttisellen

The Brüttisellen multi-temporal images are a subset of two QuickBird scenes, acquired in August 2002 and October 2006 respectively. They have been both pansharpened using the Gram-Schmidt transformation, resulting in approximately 0.7[m] of pixel size. The subsets have size of 521×1188 pixels, accounting for NIR-R-G-B channels. By visual inspection, a total of 9 classes have been detected, of which 3 are changed and 6 unchanged areas (see Figure C.1). The set available for training is composed of 57'587 pixels, while the spatially independent test set accounts for 58'293 samples.

The changed regions delineate a group of newly constructed houses in a bare soil region. The scene is challenging since bare soil can partially dissimulate radiometric changes related to the buildings. Other changes are related to grassland and bare soils, while a different shadowing causes radiometric differences in unchanged zones. Unchanged areas represent a typical low density residential surface. The different acquisition times do not raise issues related to phenological differences (since not modelled). Figure C.1 illustrates the datasets and the training/testing regions.

C. Datasets

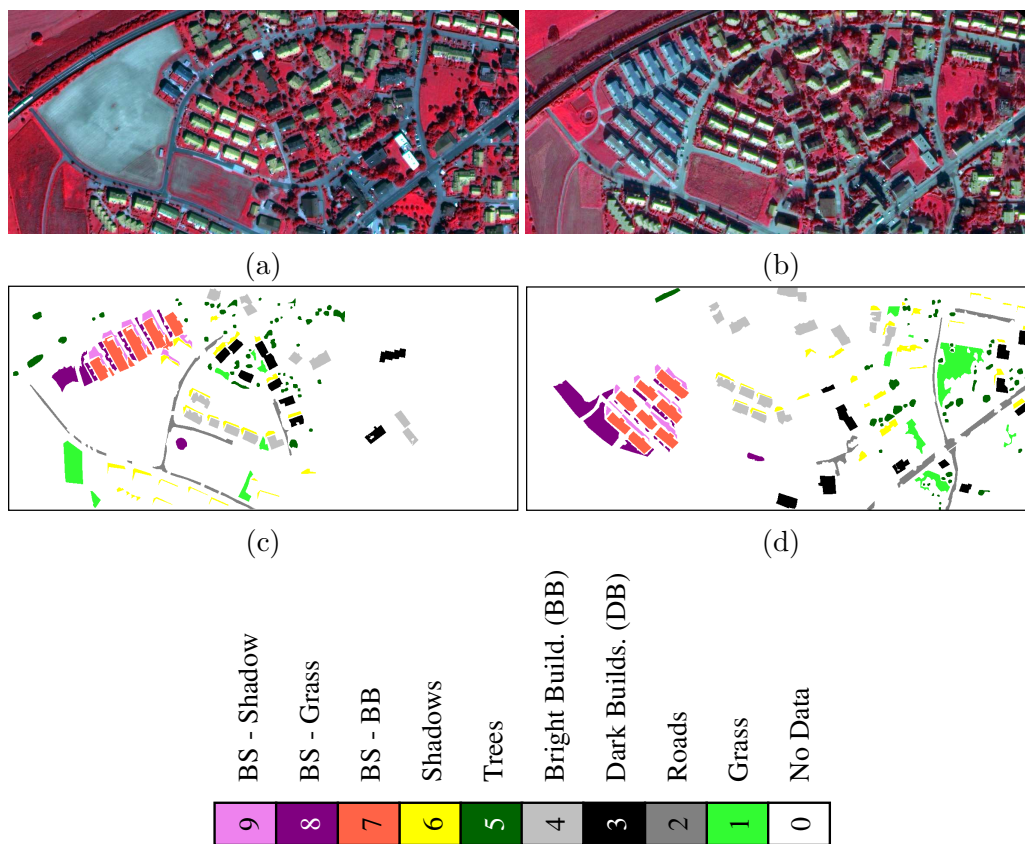


Figure C.1: The Brüttisellen dataset - In (a) 2002 and (b) 2006 datasets. In (c) and (d) the regions used for training and testing, respectively. In the legend, BS refers to bare soil.

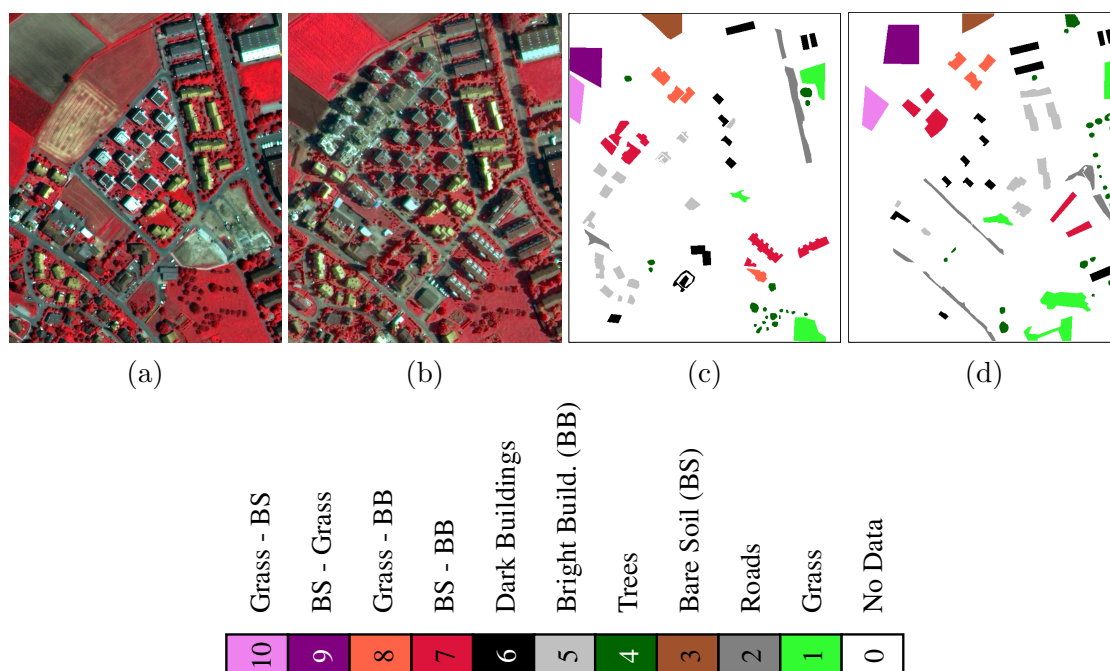


Figure C.2: The Steinacker dataset - In (a) 2002 and (b) 2006 datasets. In (c) and (d) the regions used for training and testing, respectively.

C.2 Steinacker

The Steinacker dataset is extracted from the same pansharpened QuickBird from which the Brüttisellen dataset has been selected. The scenes account for 4 classes related to ground cover changes and 6 to unchanged areas, both discovered by visual inspection of the two 784×649 scenes. The set available for training is composed of 52'564 pixels, while the spatially independent test set accounts for 58'293 samples.

The observed transitions are related to cultivated crops (vegetated and not) and to the construction of new buildings over a cultivated crop showing both grass and bare soil covers. Also, the construction site in the lower right corner has been completed. The rest of the image presents differences due to the sun elevation level and small changes due to urban dynamics. Figure C.2 illustrates the datasets and the training/testing regions.

C. Datasets

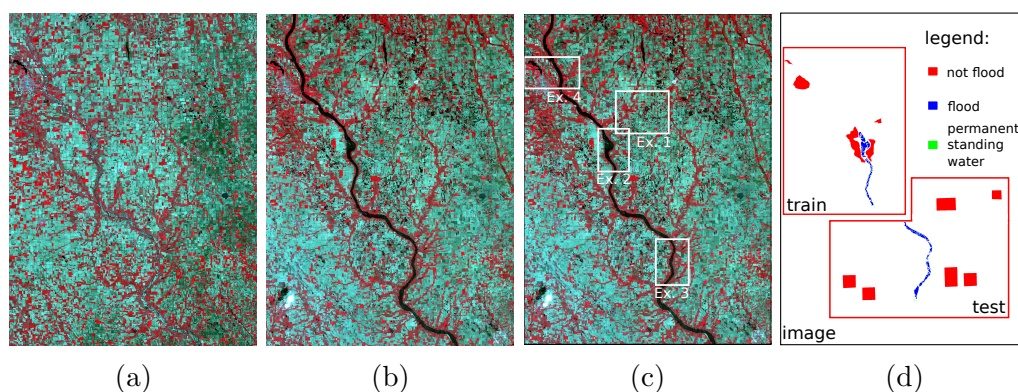


Figure C.3: Subsets of the Landsat TM James River flooding - (a) 2005 and (b) 2011 acquisitions. In (c) location of the frames used for visual validation of the results (see Figure 6.4 for the 4 details); and (d) the ground truth locations used to numerically validate the outcomes.

C.3 Missouri flooding

The James River is a tributary of the Missouri River, South Dakota, USA. In summer 2011 significant rainfalls affected the region and the river rose above the flood stages, inundating the alluvial valley. Damages to cultivated crops were reported due to the rise of the water table level and to the heavy precipitations. The National Aeronautics and Space Administration (NASA) reported that in some points the flooded river reached the kilometer wide¹. Both images used in this study have been acquired by the Landsat TM sensor, providing images with a spatial resolution of 30[m]. A subset of size 2800×2100 pixels covering the James River has been retained from the original data. The pre-event image has been acquired in May 19th, 2005 and the post-event image depicts the situation on June 5th, 2011, shortly after the flooding. The set available for training is composed of 47'162 pixels, while the spatially disjoint test set is composed of 80'282 pixels. This dataset is challenging since small differences in phenology and crop rotation introduce land cover changes that are of no interest for the flood mapping task. The 'not flooded' class includes consequently all the uninteresting changes. By observing the river path, valley morphology and structure, no significant changes occurred between the two acquisitions.

¹<http://earthobservatory.nasa.gov/NaturalHazards/view.php?id=50901>

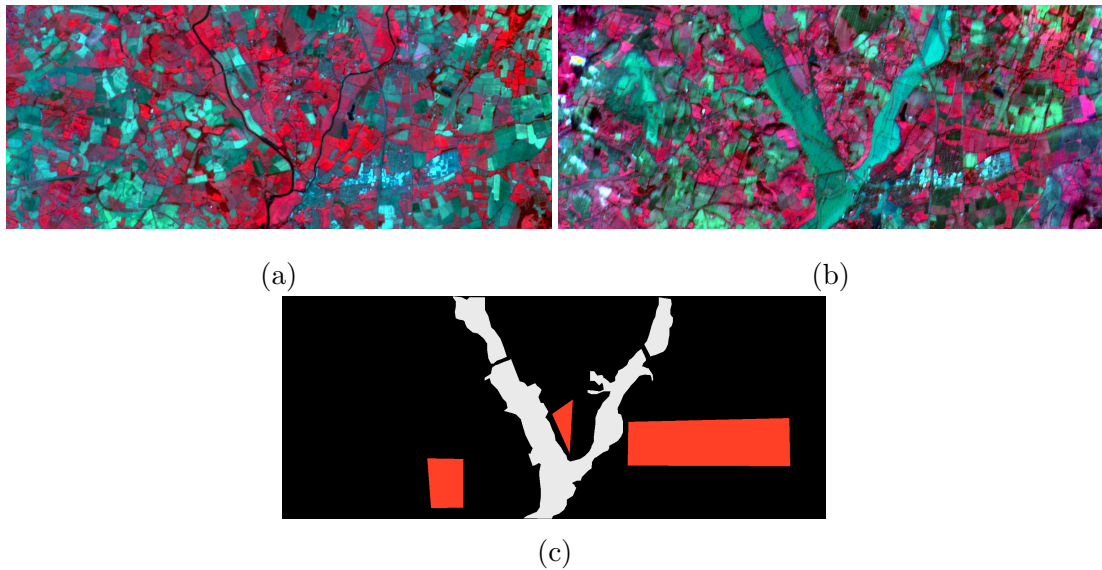


Figure C.4: The Gloucester subset - (a) 1999 and (b) 2000 acquisitions along with the ground truth, in (c). Changes are labelled in white while red refers to unchanged areas.

C.4 Gloucester flooding

This dataset consists in a subset of the image provided for the 2009 IEEE GRS-S data fusion contest [Longbotham et al., 2012]. The bi-temporal dataset is composed of two 3-bands SPOT XS images of size 712×1734 , in the range NIR-R-G. They come with a spatial resolution of 20[m]. The scenes were acquired before and after a flooding event occurred in Gloucester (UK), in 2007. The ground truth for the changed class is composed of 103'702 pixels, while 97'769 are available for the unchanged class. The change detection problem consists in correctly mapping the flood extent. The task is challenging since the scene presents many changes due to crop rotation and the spectral information available for modelling is low.

C. Datasets

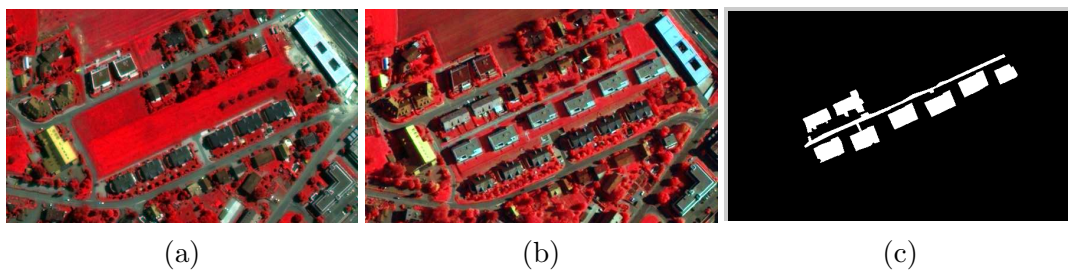


Figure C.5: The Brüttisellen 2 dataset - (a) 2002 and (b) 2006 acquisitions along with the ground truth, in (c). Changes are labelled in white while black refers to unchanged areas.

C.5 Brüttisellen 2 dataset

This bi-temporal dataset is a subset of a couple of pansharpened QuickBird images (NIR-R-G-B) of a neighborhood of Zurich (Switzerland), acquired respectively in August 2002 and October 2006 (as for the other Brüttisellen data, Section C.1). Their size is 362×598 , and represent general changes related to urban dynamics and in particular to the construction of a group of familiar houses. The spatial resolution of the images is of about 0.7[m], making the problem of change detection hard due to high variances of the class of pixels, as well as general differences caused by illumination differences. The ground truth for the changed class includes 12'309 pixels, while 204'167 are available for the unchanged one.

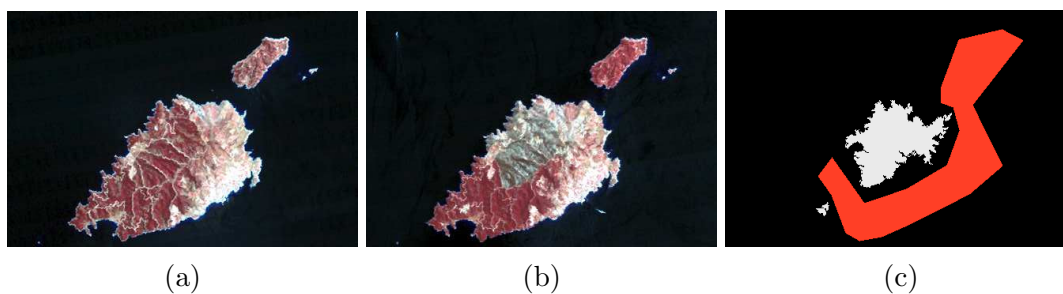


Figure C.6: The Greek Island dataset - (a) 1987 and (b) 1991 acquisitions along with the ground truth, in (c). Changes are labelled in white while red refers unchanged areas.

C.6 Greek island forest fire

This dataset is a portion of two Landsat TM images of a small island in Peloponnese, Greece, acquired respectively in 1987 and 1991 respectively. The scenes are 444×300 pixels, with a spatial resolution of 30[m]. Prior to analysis the low resolution thermal band has been removed, resulting in 6 bands. The change detection problem consists in delineating a post-fire region on the north-west flank of the island. The ground truth for the changed class is composed of 7'274 pixels, while 19'256 are available for the unchanged class. Train and test samples are selected randomly among the two class labels.

C. Datasets

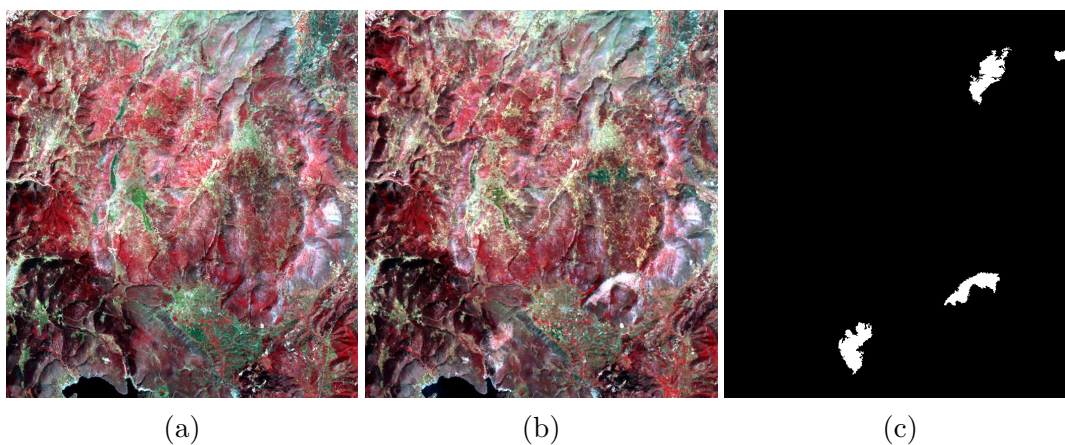


Figure C.7: Greece fires dataset - (a) 1987 and (b) 1991 acquisitions along with the ground truth, in (c). Changes are labelled in white while the remaining black pixels refer to unchanged areas.

C.7 Greek fires dataset

This dataset is a portion of two Landsat TM images acquired over the Peloponnese, Greece, in 1987 and 1991 respectively. The region has been struck by fires in 1989, 1990 and 1991. The scenes are 783×711 pixels, with a spatial resolution of 30[m]. Prior to analysis the low resolution thermal band has been removed, resulting in 6 bands. The change detection problem consists in delineating a 4 different post-fire regions, showing diverse re-vegetation situation. The ground truth for the changed class is composed of 10'457 pixels, while 564'256 are available for the unchanged class.

References

- M. A. Aizerman, E. A. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. In *Automation and Remote Control*, number 25, pages 821–837, 1964. 46, 50
- N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the Association for Computing Machinery*, 44(4):615–631, 1997. 35
- H. Alphan, H. Doygun, and Y. I. Unlukaplan. Post-classification comparison of land cover using multi-temporal Landsat and ASTER imagery: the case of Kahramanmaraş, Turkey. *Environmental Monitoring and Assessment*, 151:327–336, 2009. 65
- A. Anhed, P. Sabatier, F. Sèdes, and J. Inglada. Post-classification and spatial reasoning: new approach to change detection for updating GIS databases. In *International conference on Information and Communication Technologies: From Theory to Applications ICTTA, Damascus, (Syria)*, 2008. 65
- N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. 51
- P. M. Atkinson, I. M. Sargent, G. M. Foody, and J. Williams. Interpreting image-based methods for estimating the signal-to-noise ratio. *International Journal of Remote Sensing*, 26(22):5099–5115, 2005. 20
- S. Axler. *Linear algebra done right*. Springer, second edition, 1997. 48
- F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002a. 131, 134
- F. R. Bach and M. I. Jordan. Learning graphical models with mercer kernels. In *Advances in Neural Information Processing Systems NIPS, Vancouver (CAN)*, volume 15, 2002b. 55, 78, 106
- F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *International Conference on Machine Learning ICML, Banff (CAN)*, 2004. 57
- T. V. Bandos, L. Bruzzone, and G. Camps-Valls. Classification of hyperspectral images with regularized linear discriminant analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 47(3):862–873, 2009. 74, 77, 78
- A. Baraldi and F. Parmiggiani. An investigation of the textural characteristics associated with gray level co-occurrence matrix statistical parameters. *IEEE Transactions on Geoscience and Remote Sensing*, 33(2):293–304, March 1995. 88
- F. Bavaud. On the Schoenberg transformations in data analysis: Theory and illustrations. *Journal of Classification*, 28(3):297–314, 2011. 55
- Y. Bazi, F. Melgani, and H. D. Al-Sharari. Unsupervised change detection in multispectral remotely sensed imagery with level set methods. *IEEE Transactions on Geoscience and Remote Sensing*, 48(8):3178–3187, 2010. 68
- L. A. Belanche. Developments in kernel design. In *European symposium on artificial neural networks, computational intelligence and machine learning ESANN, Bruges (B)*, pages 369–378, 2013. 56, 57
- M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. 43
- M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006. 131, 135
- J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson. Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 43:480–490, 2005. 74, 85, 90
- J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):354–379, 2012. 16
- C. M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995. 40
- C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer Berlin, Information Science and Statistics, 2006. 31, 40, 43
- M. B. Blaschko, J. A. Shelton, A. Bartels, C. H. Lampert, and A. Gretton. Semi-supervised kernel canonical correlation analysis with application to human fMRI. *Pattern Recognition Letters*, 32(11):1572 – 1583, 2011. 131

References

- R. Blundell and A. Duncan. Kernel regression in empirical microeconomics. *The Journal of Human Resources*, 33(1):62–87, 1998. 4
- B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *5th ACM Workshop on Computational Learning Theory COLT, Pittsburg (USA)*, pages 144–152, 1992. 85, 86
- F. Bovolo. A multilevel parcel-based approach to change detection in very high resolution multitemporal images. *IEEE Geoscience and Remote Sensing Letters*, 6(1):33–38, 2009. 68, 118
- F. Bovolo and L. Bruzzone. A theoretical framework for unsupervised change detection based on change vector analysis in polar domain. *IEEE Transactions on Geoscience and Remote Sensing*, 45(1):218–236, 2007. 23, 110, 137
- F. Bovolo, L. Bruzzone, and M. Marconcini. A novel approach to unsupervised change detection based on a semi-supervised SVM and a similarity measure. *IEEE Transactions on Geoscience and Remote Sensing*, 46(7):2070–2082, 2008. 68
- F. Bovolo, L. Bruzzone, and M. Marchesi. Analysis and adaptive estimation of the registration noise distribution in multitemporal VHR images. *IEEE Transactions on Geoscience and Remote Sensing*, 47(8):2658–2671, 2009. 24, 150
- F. Bovolo, G. Camps-Valls, and L. Bruzzone. A support vector domain method for change detection in multitemporal images. *Pattern Recognition Letters*, 31(10):1148–1154, 2010. 69
- F. Bovolo, M. Marchesi, and L. Bruzzone. A framework for automatic and unsupervised detection of multiple changes in multitemporal images. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2196–2212, 2012. 23
- L. Bruzzone and D. F. Fernández-Prieto. Automatic analysis of the difference image for unsupervised change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 38(3):1171–1181, 2000. 23, 68, 105, 110
- C. Campbell. Kernel methods, a survey of current techniques. *Neurocomputing*, 48:63–84, 2002. 47
- J. B. Campbell and R. H. Wynne. *Introduction to Remote Sensing*. The Guilford Press, 2011. 13, 15, 24
- G. Camps-Valls and L. Bruzzone, editors. *Kernel Methods for Remote Sensing Data Analysis*. J. Wiley & Sons, 2009. 4, 47, 53, 56, 58, 85
- G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla. Composite kernels for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 3(1):93–97, 2006. 57, 67, 118, 148
- G. Camps-Valls, T. V. Bandos, and D. Zhou. Semi-supervised graph-based hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3044–3054, 2007a. 46
- G. Camps-Valls, J. Rojo-Álvarez, and M. Martínez-Ramón, editors. *Kernel methods in bioengineering, signal and image processing*. Idea Group Inc., Hershey, PA, USA, 2007b. 4
- G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, L. Rojo-Álvarez, and M. Martínez-Ramón. Kernel-based framework for multi-temporal and multi-source remote sensing data classification and change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6):1822–1835, 2008. 57, 67, 104, 118
- G. Camps-Valls, J. Mooij, and B. Schölkopf. Remote sensing feature selection by kernel dependence measures. *IEEE Geoscience and Remote Sensing Letters*, 7(3):587–591, 2010. 42
- G. Camps-Valls, D. Tuia, L. Gómez-Chova, S. Jimenez, and J. Malo. *Remote sensing image processing*, volume 5 of *Synthesis Lectures on Image, Video, and Multimedia Processing*. Morgan & Claypool Publishers, 2011. 23
- G. Camps-Valls, J. Muñoz-Marí, L. Gómez-Chova, L. Guanter, and X. Calbet. Nonlinear statistical retrieval of atmospheric profiles from MetOp-IASI and MTG-IRS infrared sounding dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 50(5):1759–1769, 2012. 18
- M. J. Canty. *Image analysis, classification and change detection in remote sensing*. Taylor and Francis, 2007. 21
- M. J. Canty and A. A. Nielsen. Linear and kernel methods for multivariate change detection. *Computers and Geosciences*, 38(1):107–114, 2012. 71
- L. Capobianco and G. Camps-Valls. Target detection with semisupervised kernel orthogonal subspace projection. *IEEE Transactions on Geoscience and Remote Sensing*, 47(11):3822–3833, 2009. 70
- T. Celik. Unsupervised change detection in satellite images using principal component analysis and k -means clustering. *IEEE Geoscience and Remote Sensing Letters*, 6(4):772–776, 2009a. 68, 110, 111, 112, 113, 114

- T. Celik. Multiscale change detection in multitemporal satellite images. *IEEE Geoscience and Remote Sensing Letters*, 6(4):820–824, 2009b. 68
- O. Chapelle and A. Rakotomamonjy. Second order optimization of kernel parameters. In *Advances in Neural Processing Information Systems NIPS, Workshop on Kernel Learning, Vancouver (CAN)*, 2008. 58
- O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46:131–159, 2002. 58
- O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006. 45
- J. Chen, X. Chen, X. Cui, and J. Chen. Change vector analysis in posterior probability space: A new method for land cover change detection. *IEEE Geoscience and Remote Sensing Letters*, 8(2):317–321, 2011. 65
- V. Cherkassky and F. Mulier. *Learning from data: concepts, theory and methods*. John Wiley & sons, inc., 2007. 39
- J. Chormanski, T. Okruszko, S. Ignar, O. Batelaan, K. T. Rebel, and Wassen M. J. Flood mapping with remote sensing and hydrochemistry: A new method to distinguish the origin of flood water during floods. *Ecological Engineering*, 37(9):1334–1349, 2011. 73
- R. Collobert, S. Bengio, and J. Mariéthoz. Torch: a modular machine learning software library. Technical Report IDIAP-RR 02-46, IDIAP, 2002. URL <http://www.torch.ch>. 96
- D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(3):265–298, 2004. 120
- L. Copa, D. Tuia, M. Volpi, and M. Kanevski. Unbiased query-by-bagging active learning for VHR image classification. In L. Bruzzone, editor, *SPIE Image and Signal Processing for Remote Sensing XVI, Toulouse (F)*, volume 7830(1), 2010. 8
- P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin. Digital change detection methods in ecosystem monitoring: a review. *International Journal of Remote Sensing*, 25:1565–1596, 2004. 23, 65
- P. R. Coppin and M. E. Bauer. Processing of multitemporal Landsat TM imagery to optimize extraction of forest cover change features. *IEEE Transactions on Geoscience and Remote Sensing*, 32(4):918–927, 1994. 67
- C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. 85
- T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991. 158
- T.M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14(3):326–334, 1965. 54
- D. Cremers, T. Kohlberger, and C. Schnorrb. Shape statistics in kernel space for variational image segmentation. *Pattern Recognition*, 36:1929–1943, 2003. 55, 78, 106
- X. Dai and S. Khorram. The effects of image misregistration on the accuracy of remotely sensed change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 36:1566–1577, 1998. 24
- P. E. R. Dale, A. L. Chandica, and M. Evans. Using image subtraction and classification to evaluate change in sub-tropical intertidal wetlands. *International Journal of Remote Sensing*, 17(4):703–719, 1996. 67
- M. Dalla Mura, J. A. Benediktsson, F. Bovolo, and L. Bruzzone. An unsupervised technique based on morphological filters for change detection in very high resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 5(3):433–437, 2008. 68
- T. De Bie, N. Cristianini, and R. Rosipal. *Handbook of computational geometry for pattern recognition, computer vision, neurocomputing and robotics*, chapter Eigenproblems in pattern recognition, pages 129–171. Springer Verlag, 2004. 49, 121, 122, 132
- D. De Canditiis and I. De Feisb. Pointwise convergence of Fourier regularization for smoothing data. *Journal of Computational and Applied Mathematics*, 196(2):540–552, 2006. 39
- F. de Morsier, D. Tuia, V. Gass, J.-P. Thiran, and M. Borgeaud. Unsupervised change detection via hierarchical support vector clustering. In *IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS), Tsukuba (JAP)*, 2012. 68
- F. de Morsier, D. Tuia, M. Borgeaud, V. Gass, and J.-P. Thiran. Semi-supervised novelty detection using SVM entire solution path. *IEEE Transactions on Geoscience and Remote Sensing*, 51(4):1939–1950, 2013. 69
- A. P. Dempster, M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977. 41, 43, 105
- J. S. Deng, K. Wang, Y. H. Deng, and G. J. Qi. PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data. *International Journal of Remote Sensing*, 29(16):4823–4838, 2008. 71

References

- B. Desclée, P. Bogaert, and P. Defourny. Forest change detection by statistical object-based method. *Remote Sensing of Environment*, 102:1–11, 2006. 68
- C. Dey, X. Jia, D. Fraser, and L. Wang. Mixed pixels analysis for flood mapping using extended support vector machines. In *Digital image computing: techniques and applications DICTA, Melbourne (AUS)*, pages 291–295, 2009. 73
- P. Du, S. Liu, P. Gamba, K. Tan, and J. Xia. Fusion of difference images for change detection over urban areas. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(4):1076–1086, 2012. 67
- R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, second edition, 2001. 33
- A. J. Elmore, S. J. Mustard, J. F. Manning, and D. B. Lobell. Quantifying vegetation change in semiarid environments: Precision and accuracy of spectral mixture analysis and the normalized difference vegetation index. *Remote Sensing of Environment*, 73: 87–102, 2000. 66
- T. Evgeniou, M. Pontil, and T. Poggio. A unified framework for regularization networks and support vector machines. Technical Report A.I. Memo No. 1654, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1999. 33, 35, 39, 58
- T. Evgeniou, T. Poggio, M. Pontil, and A. Verri. Regularization and statistical learning theory for data analysis. *Computational statistics and data analysis*, 38:421–432, 2002. 39, 52
- M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson. Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 46(11):3804 – 3814, 2008. 90, 91
- T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–674, 2006. 109, 159
- C. R. Fichera, G. Modica, and M. Pollino. Land cover classification and change-detection analysis using multi-temporal remote sensed imagery and landscape metrics. *European Journal of Remote sensing*, 45:1–18, 2012. 65
- R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936. 74
- G. M. Foody. Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering and Remote Sensing*, 70(5):627 – 633, 2004. 96, 109, 158, 159
- G. M. Foody and A. Mathur. Toward intelligent training of supervised image classifications: directing training data acquisition for svm classification. *Remote Sensing of Environment*, 93:107–117, 2004. 101
- D. Francois, V. Wertz, and M. Verleysen. About the locality of kernels in high-dimensional spaces. In *International Symposium on Applied Stochastic Models and Data Analysis ASDMA, Brest (F)*, pages 238–245, 2005. 56
- J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84 (405):165–175, 1989. 77, 78
- T. Fung. An assessment of TM imagery for land-cover change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 28(4):681–684, 1990. 68, 70
- F. Gao, J. Masek, M. Schwaller, and F. Hall. On the blending of the Landsat and MODIS surface reflectance: predicting daily Landsat surface reflectance. *IEEE Transactions on Geoscience and Remote Sensing*, 44(8):2207–2218, 2006. 131
- S. Ghosh, L. Bruzzone, S. Patra, F. Bovolo, and A. Ghosh. A context-sensitive technique for unsupervised change-detection based on Hopfield-type neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 45(3):778–789, 2007. 68
- M. Gianinetto and P. Villa. Rapid response flood assessment using minimum noise fraction and composed spline interpolation. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3204–3211, 2007. 22, 121
- T. W. Gillespie, J. C. Elizabeth, and F. D. Thomas. Assessment and prediction of natural hazards from satellite imagery. *Progress in Physical Geography*, 31(5):459–470, 2007. 22
- M. Girolami. Mercer kernel-based clustering in feature space. *IEEE Transactions on Neural Networks*, 13 (3):780–784, 2002. 104
- L. Gómez-Chova, G. Camps-Valls, J. Muñoz-Marí, and J. Calpe-Maravilla. Semisupervised image classification with laplacian support vector machines. *IEEE Geoscience and Remote Sensing Letters*, 5(3):336–340, 2008. 135
- L. Gómez-Chova, J. Muñoz, V. Laparra, J. Malo, and G. Camps-Valls. *Optical Remote Sensing Advances in Signal Processing and Exploitation Techniques*, chapter A Review of Kernel Methods in Remote Sensing Data Analysis, pages 171–206. Springer-Verlag, 2010. 53, 57

References

- L. Gómez-Chova, A. A. Nielsen, and G. Camps-Valls. Explicit signal-to-noise ratio in reproducing kernel Hilbert spaces. In *IEEE International Geoscience and Remote Sensing Symposium IGARSS, Vancouver (CAN)*, pages 3570–3573, 2011. 21
- L. Gómez-Chova, J. Amorós-López, E. Izquierdo-Verdiguier, J. C. Jiménez-Muñoz, and G. Camps-Valls. Cloud screening from multispectral image time series. In *EARSeL SIG Workshop on Temporal analysis of satellite images, Mykonos (GR)*, 2012. 71, 121
- L. Gómez-Chova, E. Izquierdo-Verdiguier, J. Amorós-López, and G. Camps-Valls. Kernel change discriminant analysis for multitemporal cloud masking. In *IEEE International Geoscience and Remote Sensing Symposium IGARSS, Melbourne (AUS)*, 2013. 71
- A. B. A. Graf and S. Borer. Normalization in support vector machines. In *Symposium of the German Association for Pattern Recognition DAGM, Munich (D)*, 2001. 26
- A. A. Green, M. Berman, P. Switzer, and M. D. Craig. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, 26(1):65–74, 1998. 21, 121
- S. Gua, Y. Tana, and X. Hea. Discriminant analysis via support vectors. *Neurocomputing*, 73:1669–1675, 2010. 78
- R. T. Guchi, C. K. Huyck, B. J. Adams, B. Mansouri, B. Houshmand, and G. M. Shinozuka. *MCEER Research Progress and Accomplishments, 2001-2003*, chapter Resilient Disaster Response: Using Remote Sensing Technologies for Post-Earthquake Damage Detection, pages 125–137. State University of New York at Buffalo, 2003. 22
- F. Guerra, H. Puig, and R. Chaume. The forest-savanna dynamics from multi-date Landsat TM data in Sierra Parima, Venezuela. *International Journal of Remote Sensing*, 19(11):2061–2075, 1998. 67
- I. Guyon, B. E. Boser, and V. Vapnik. Automatic capacity tuning of very large VC-dimension classifiers. In *Advances in Neural Information Processing Systems NIPS, Denver (USA)*, pages 147–155, 1992. 49
- R. Hable and A. Christmann. On qualitative robustness of support vector machines. *Journal of Multivariate Analysis*, 102(6):993–1007, 2011. 38
- R.H. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on System, Man and Cybernetics*, 3(6):610–621, November 1973. 74, 88, 89
- T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning. Data mining, inference, and prediction*. Springer Berlin, Information Science and Statistics, 2 edition, 2009. 33, 34, 35, 36, 41, 52
- S. Haykin. *Neural Network: A Comprehensive Foundation*. Prentice Hall, 1999. 40, 42, 45
- G. G. Hazel. Object-level change detection in spectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 39(3):553–561, 2001. 68
- J. Heo and T. W. Fitzhugh. A standardized radiometric normalization method for change detection using remotely sensed imagery. *Photogrammetric Engineering and Remote Sensing*, 66(2):173–181, 2000. 120, 121
- T. Hoffmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The annals of Statistics*, 36(3):1171–1220, 2008. 47
- H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 25:417–44, 1933. 43
- H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:312–377, 1936. 131, 132
- S.-Y. Huang and C.-R. Hwang. Kernel Fisher’s discriminant analysis in Gaussian reproducing kernel Hilbert spaces - Theory. Technical report, Institute of statistical sciences, Academia sinica, Taipei, China., 2006. 82
- P. F. Hudson and R. R. Colditz. Flood delineation in a large and complex alluvial valley, lower Pánuco basin, Mexico. *Journal of Hydrology*, 280:229–245, 2003. 74
- G. F. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, 1968. 38, 101
- C. Huo, Z. Zhou, H. Lu, C. Pan, and K. Chen. Fast object-level change detection for VHR images. *IEEE Geoscience and Remote Sensing Letters*, 7(1):118–122, 2010. 69
- J. Im, J. R. Jensen, and M. E. Hodgson. Optimizing the binary discriminant function in change detection applications. *Remote Sensing of Environment*, 112(6):27612776, 2008. 68
- S. Inamdar, F. Bovolo, L. Bruzzone, and S. Chaudhuri. Multidimensional probability density function matching for pre-processing of multitemporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(4):1243–1252, 2008. 120

References

- F. Ip, J. M. Dohm, V.R. Baker, T. Doggett, A. G. Davies, R. Castaño, S. Chien, B. Cichy, R. Greeley, R. H. Sherwood, D. Tran, and G. Rabideau. Flood detection and monitoring with the autonomous sciencecraft experiment onboard EO-1. *Remote Sensing of Environment*, 101:463–481, 2006. 74
- J. R. Jensen. *Remote Sensing of the Environment: An Earth Resource Perspective*. Prentice Hall, seconds edition, 2007. 10
- M. Kanevski, A. Pozdnoukhov, V. Timonin, and M. Maignan. Mapping of environmental data using kernel-based methods. *International Journal of Geomatics and Spatial Analysis*, 17(3-4):309–331, 2007. 42
- M. Kanevski, A. Pozdnoukhov, and V. Timonin. *Machine learning algorithms for spatial environmental data. Theory, applications and software*. EPFL Press, Lausanne, 2009. 42, 55
- R. E. Kennedy, P. A. Townsend, J. E. Gross, W. B. Cohen, P. Bolstad, Y. Q. Wang, and P. Adams. Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects. *Remote Sensing of Environment*, 113(7):1382–1396, 2009. 21
- J. Keshet and S. Bengio, editors. *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods*. John Wiley & Sons, 2008. 4
- J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58:433–451, 1971. 134
- S. I. Khan, Y. Hong, J. Wang, K. K. Yilmaz, J. J. Gourley, R. F. Adler, G. R. Brakenridge, F. Policelli, S. Habib, and D. Irwin. Satellite remote sensing and hydrologic modeling for flood inundation mapping in Lake Victoria basin: Implications for hydrologic prediction in ungauged basins. *IEEE Transactions on Geoscience and Remote Sensing*, 49(1):85–95, 2011. 73
- S. Khazai, A. Safari, B. Mojaradi, and S. Homayouni. Improving the svdd approach to hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 9(4):594–598, 2012. 125
- W. Koppe, M. L. Gnyp, C. Hütt, Y. Yao, Y. Miao, X. Chen, and G. Bareth. Rice monitoring with multi-temporal and dual-polarimetric TerraSAR-X data. *International Journal of Applied Earth Observation and Geoinformation*, 21:568–576, 2013. 22
- B. Kunkel, F. Blechinger, R. Lutz, R. Doerffer, and H. van der Piepen. ROSIS (reflective optics system imaging spectrometer) - a candidate instrument for polar platform missions. In J. S. Seeley and S. C. Bowyer, editors, *Society of Photo-Optical Instrumentation Engineers SPIE*, pages 134–141, 1988. 18
- M. Kuss and T. Graepel. The geometry of kernel canonical correlation analysis. Technical Report 103, Max Planck Institute for Biological Cybernetics, 2003. 132
- H. Kwon and N. M. Nasrabadi. Kernel RX-algorithm: A nonlinear anomaly detector for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 43(2):388–397, 2005. 69, 78
- H. Lee. Mapping deforestation and age of evergreen trees by applying a binary coding method to time-series Landsat November images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(11):3926–3936, 2008. 65
- J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer, 2007. 38, 39
- C. Leslie and R. Kuang. Fast string kernels using inexact matching for protein sequences. *Journal of Machine Learning Research*, 5:1435–1455, 2004. 57
- S. Lhermitte, J. Verbesselt, W. W. Verstraeten, and P. Coppin. A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. *Remote Sensing of Environment*, 115(12):3129–3152, 2011. 64
- T. M. Lillesand, R. W. Kiefer, and J. W. Chipman. *Remote sensing and image interpretation*. John Wiley & Sons, 2004. 9, 13, 24
- X. Long Dal and S. Khorram. Remotely sensed change detection based on artificial neural networks. *Photogrammetric Engineering and Remote Sensing*, 65(10):1187–1194, 1999. 66
- N. Longbotham, F. Pacifici, T. Glenn, A. Zare, M. Volpi, D. Tuia, E. Christophe, J. Michel, J. Inglada, J. Chanussot, and Q. Du. Multi-modal change detection, application to the detection of flooded areas: Outcome of the 2009-2010 data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(6):331–342, 2012. 6, 165
- J. MacQueen. Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967. 42, 104, 105
- W. A. Malila. Change vector analysis: An approach for detecting forest change with Landsat. In *IEEE Proceedings of Annual Symposium on Machine Processing of Remotely Sensing Data*, pages 326–336, 1980. 23
- S. Marchesi and L. Bruzzone. ICA and kernel ICA for change detection in multispectral remote sensing images. In *IEEE International Geosciences and Remote Sensing Symposium IGARSS*, volume 2, pages II–980–II–983, 2009. 71, 121

- S. Marchesi, F. Bovolo, and L. Bruzzone. A context-sensitive technique robust to registration noise for change detection in VHR multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 19(7):1877–1889, 2010. 23
- J. V. Martonchik, C. J. Bruegge, and A. H. Strahler. A review of reflectance nomenclature used in remote sensing, remote sensing reviews. *Remote Sensing Reviews*, 19(1):9–20, 2000. 14
- J.-F. Mas. Monitoring land-cover changes: A comparison of change detection techniques. *International Journal of Remote Sensing*, 20(1):139–152, 1999. 23
- G. Matasci, M. Volpi, D. Tuia, and M. Kanevski. Transfer component analysis for domain adaptation in image classification. In L. Bruzzone, editor, *SPIE Image and Signal Processing for Remote Sensing XVII, Prague (CZ)*, volume 8180, 2011. 7
- G. Matasci, L. Bruzzone, M. Volpi, D. Tuia, and M. Kanevski. Investigating feature extraction for domain adaptation in remote sensing image classification. In *International Conference on Pattern Recognition Application and Methods ICPRAM 2013, Barcelona (SP)*, 2013. 7
- F. Melgani and Y. Bazi. Markovian fusion approach to robust unsupervised change detection in remotely sensed imagery. *IEEE Geoscience and Remote Sensing Letters*, 3(4):457–461, 2006. 68
- F. Melgani, G. Moser, and S. B. Serpico. Unsupervised change-detection methods for remote-sensing images. *Optical Engineering*, 41(12):3288–3297, 2002. 68
- M. P. Mello, C. A. O. Bernardo Vieira, M. T. Rudorff, P. Aplin, R. D. C. Santos, and D. A. Aguiar. STARS: A new method for multitemporal remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2013. 64
- J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209:415–458, 1909. 54
- G. Metternicht, L. Hurni, and R. Gogu. Remote sensing of landslides: An analysis of the potential contribution to geo-spatial systems for hazard assessment in mountainous environments. *Remote Sensing of Environment*, 98(2-3):284–303, 2005. 22
- S. Mika. *Kernel Fisher Discriminants*. PhD thesis, Elektrotechnik und Informatik der Technischen Universität Berlin, 2002. 75, 77, 121
- S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In *IEEE Neural Networks for Signal Processing, Madison (USA)*, pages 41–48, 1999. 42, 75, 76, 78
- S. Mika, G. Rätsch, and K.-R. Müller. A mathematical programming approach to the kernel Fisher algorithm. In *Advances in Neural Information Processing Systems NIPS, Denver (USA)*, pages 591–597, 2000. 76, 77, 78
- J. Muñoz-Marí, F. Bovolo, L. Gómez-Chova, L. Bruzzone, and G. Camps-Valls. Semisupervised one-class support vector machines for classification of remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 48(8):3188–3197, 2010. 69
- H. Nemmour and Y. Chibani. Multiple support vector machines for land cover change detection: an application for mapping urban extensions. *ISPRS Journal of photogrammetry and remote sensing*, 61(2):125–133, 2006. 21, 66
- H. Nemmour and Y. Chibani. Support vector machines for automatic multi-class change detection in Algerian capital using Landsat TM imagery. *Journal of the Indian Society of Remote Sensing*, 38(4):585–591, 2010. 66
- A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems NIPS, Vancouver (CAN)*, volume 14, 2002. 43
- A. A. Nielsen. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE Transactions on Image Processing*, 11(3):293–305, 2002. 70, 121
- A. A. Nielsen. The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data. *IEEE Transactions on Image Processing*, 16(2):463–478, 2007. 70, 121, 136
- A. A. Nielsen. Kernel maximum autocorrelation factor and minimum noise fraction transformations. *IEEE Transactions on Image Processing*, 20(3):612–624, 2011. 21
- A. A. Nielsen and M. J. Canty. Kernel principal component analysis for change detection. In L. Bruzzone, editor, *SPIE Image and Signal Processing for Remote Sensing XIV, Cardiff (UK)*, volume 7109, 2008. 70, 121
- A. A. Nielsen, K. Conradsen, and J. J. Simpson. Multivariate alteration detection (MAD) and MAF post-processing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sensing of Environment*, 64(1):1–19, 1998. 136
- F. Pacifici and F. Del Frate. Automatic change detection in very high resolution images with pulse coupled neural networks. *IEEE Geoscience and Remote Sensing Letters*, 7(1):58–62, 2010. 69

References

- F. Pacifici, M. Chini, and W.J. Emery. A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. *Remote Sensing of Environment*, 113:1276–1292, 2009. 74, 85
- S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. 121
- S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011. 121
- D. C. Parker and M. F. Wolff. Remote sensing. *International science and technology*, 43:20–31, 1965. 15
- I. M. Penna, M.-H. Derron, M. Volpi, and M. Jaboyed-off. Analysis of past and future dam formation and failure in the Santa Cruz River (San Juan province, Argentina). *Geomorphology*, 186:28–30, 2013. 8
- M. Petrou and P.G. Sevilla. *Dealing With Texture*. Image Processing Series. John Wiley and Sons, 2006. 89
- A. Plaza, J. Plaza, and H. Vegas. Improving the performance of hyperspectral image and signal processing algorithms using parallel, distributed and specialized hardware-based systems. *Journal of Signal Processing Systems*, 50:293–315, 2010. 27
- J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, editors. *Dataset Shift in Machine Learning*, volume 229. MIT Press, 2009. 120, 121
- R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: A systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, 2005. 23
- A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008. 57, 58
- W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846850, 1971. 109, 158
- C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. 42
- R. A. Rensink. Change detection. *Annual Review of Psychology*, 53:245–277, 2002. 21, 22
- E. Ricci, T. De Bie, and N. Cristianini. Magic moments for structured output prediction. *Journal of Machine Learning Research*, 9:2803–2846, 2008. 57
- J. A. Richards. Analysis of remotely sensed data: the formative decades and the future. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):422–432, 2005. 27
- B. B. Salem, A. El-Cibahy, and M. El-Raey. Detection of land cover classes in agro-ecosystems of northern Egypt by remote sensing. *International Journal of Remote Sensing*, 16(14):2581–2594, 1995. 66
- R. Sanders, F. Shaw, H. MacKay, H. Galy, and M. Foote. National flood modelling for insurance purposes: using IFSAR for flood risk estimation in Europe. *Hydrology and Earth System Sciences*, 9(4):449–456, 2005. 74
- G. Schaepman-Strub, M. E. Schaepman, T. H. Painter, S. Dangel, and J. V. Martonchik. Reflectance quantities in optical remote sensing – definitions and case studies. *Remote Sensing of Environment*, 103:27–42, 2006. 14
- A. Schneider. Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach. *Remote Sensing of Environment*, 124:689–704, 2012. 21
- B. Schölkopf and A. Smola. *Learning with Kernels. Support vector machines, regularization, optimization, and beyond*. MIT Press, 2002. 32, 34, 35, 42, 46, 47, 51, 52, 56, 58, 86
- B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998. 121, 122
- B. Schölkopf, S. Mika, C. J. C. Burges, P. Knirsch, K. R. Müller, G. Rätsch, and A. J. Smola. Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, 10(5):1000–1017, 1999. 47, 50
- B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001. 42
- B. Schölkopf, K. Tsuda, and L.-P. Vert, editors. *Kernel methods in computational biology*. MIT Press, Cambridge, London, 2004. 4
- R. A. Schowengerdt. *Remote sensing: models and methods for image processing*. Academic Press, Elsevier Inc., 2007. 9, 18, 20, 24
- P. Serra, X. Pons, and D. Saurí. Post-classification change detection with data from different sensors: some accuracy considerations. *International Journal of Remote Sensing*, 24(16):3311–3340, 2003. 65

- J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. 42, 46, 47, 49, 51, 52, 54, 56, 57, 75, 77, 121, 132, 137
- K. Shin, M. Cuturi, and T. Kuboyama. Mapping kernels for trees. In *International Conference on Machine Learning ICML, Helsinki (FIN)*, pages 961–968, 2008. 57
- A. Singh. Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10(6):989 – 1003, 1989. 23, 65, 120, 121
- C. Small. Spatiotemporal dimensionality and time-space characterization of multitemporal imagery. *Remote Sensing of Environment*, 124:793–809, 2012. 64
- P. Soille. *Morphological image analysis: Principles and Applications*, volume 391. Springer-Verlag, Berlin-Heidelberg, 2nd edition, 2004. 74, 90, 91
- P. Soille and M. Pesaresi. Advances in mathematical morphology applied to geoscience and remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 40(9):2042–2055, September 2002. 74, 90
- C. Song, C. E. Woodcock, K. C. Seto, M. P. Lenney, and S. A. Macomber. Classification and change detection using Landsat TM data - when and how to correct atmospheric effects? *Remote Sensing of Environment*, 75(2):230–244, 2001. 25, 65
- M. Sonka, V. Hlavac, and R. Boyle. *Image processing, analysis and machine vision*. PWS Publishing, 1999. 20
- I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008. 38
- I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643, 2006. 56
- G. Sucharita and C. Woodcock. Remote sensing of forest change using artificial neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 34(2):398–404, 1996. 66
- M. Sugiyama. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *Journal of Machine Learning Research*, 8:1027–1061, 2007. 78
- A. Suppasri, S. Koshimura, M. Matsuoka, H. Gokon, and D. Kamthonkiat. *Remote Sensing of Planet Earth*, chapter Application of Remote Sensing for Tsunami Disaster, pages 143–170. IntechOpen, 2012. 22
- J. A. K. Suykens and C. Alzate. Primal and dual model representation in kernel-based learning. *Statistics Surveys*, 4:148–183, 2010. 50, 52, 86, 88
- D. M. J. Tax and R. P. W. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, 2004. 42, 124
- J. Theiler. Quantitative comparison of quadratic covariance-based anomalous change detectors. *Applied Optics*, 47(28):F12–F26, 2008. 69
- J. Theiler and S. Perkins. Resampling approach for anomalous change detection. In S. S. Shen and P. E. Lewis, editors, *Proceedings of SPIE, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIII, Orlando (USA)*, volume 6565, 2007. 69, 120
- J. Theiler and B. Wohlberg. Local co-registration adjustment for anomalous change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 50(8):3107–3116, 2012. 150
- A. N. Tikhonov and V. Y. Arsenin. *Solution of ill-posed problems*. Winston, 1977. 38
- D. M. Tralli, R. G. Blom, V. Zlotnicki, A. Donnellan, and D. L. Evans. Satellite remote sensing of earthquake, volcano, flood, landslide and coastal inundation hazards. *ISPRS Journal of photogrammetry and remote sensing*, 59(4):185–198, 2005. 22
- M. Trolliet, D. Tuia, and M. Volpi. Classification of urban multi-angular image sequences by aligning their manifolds. In *Joint Urban Remote Sensing Event JURSE 2013, Sao Paolo (BRA)*, 2013. 7
- G. V. Trunk. A problem of dimensionality: a simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(3):306–307, 1979. 38, 101
- D. Tuia, F. Pacifici, M. Kanevski, and W. J. Emery. Classification of very high spatial resolution imagery using mathematical morphology and support vector machines. *IEEE Geoscience and Remote Sensing Letters*, 47(11):3866–3879, 2009. 74, 85
- D. Tuia, G. Camps-Valls, G. Matasci, and M. Kanevski. Learning relevant image features with multiple-kernel classification. *IEEE Transactions on Geoscience and Remote Sensing*, 48(10):3780–3791, 2010a. 57, 118, 148
- D. Tuia, F. Ratle, A. Pouzdroukhov, and G. Camps-Valls. Multisource composite kernels for urban image classification. *IEEE Geoscience and Remote Sensing Letters*, 7(1):88–92, 2010b. 74, 85

References

- D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Muñoz Marí. A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):606–617, 2011. 8, 101
- D. Tuia, M. Volpi, M. Dalla Mura, A. Rakotomamonjy, and R. Flamary. Discovering relevant spatial filterbanks for VHR image classification. In *International Conference on Pattern Recognition ICPR 2012, Tsukuba (JAP)*, 2012. 8, 148
- D. Tuia, J. Muñoz-Marí, L. Gómez-Chova, and J. Jesus Malo. Graph matching for adaptation in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):329–341, 2013a. 7, 120
- D. Tuia, M. Volpi, M. Dalla Mura, A. Rakotomamonjy, and R. Flamary. Create the relevant spatial filterbank in the hyperspectral jungle. In *IEEE International Geosciences and Remote Sensing Symposium IGARSS 2013, Melbourne (AUS)*, 2013b. 8
- T. Van Gestel, G. Suykens, J. A. K. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle. Bayesian framework for least-squares support vector machine classifiers, gaussian processes, and kernel fisher discriminant analysis. *Neural Computation*, 14:1115–1147, 2002. 78
- V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998. xvii, xix, 32, 33, 34, 35, 85
- J. Verbesselt, R. Hyndman, G. Newnham, and D. Culvenor. Detecting trend and seasonal changes in satellite image time series. *Remote Sensing of Environment*, 114(1):106–115, 2010. 64
- A. Villa, M. Fauvel, J. Chanussot, P. Gamba, and J. A. Benediktsson. Gradient optimization for multiple kernels parameters in support vector machines classification. In *IEEE International Geoscience and Remote Sensing Symposium IGARSS, Boston (USA)*, 2008. 26
- S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010. 57
- M. Volpi, D. Tuia, M. Kanevski, F. Bovolo, and L. Bruzzone. Supervised change detection in VHR images: a comparative analysis. In *IEEE International Workshop on Machine Learning for Signal Processing MLSP, Grenoble (F)*, pages 1–6, 2009. 6
- M. Volpi, D. Tuia, G. Camps-Valls, and M. Kanevski. Unsupervised change detection by kernel clustering. In L. Bruzzone, editor, *SPIE Image and Signal Processing for Remote Sensing XVI, Toulouse (F)*, volume 7830(1), 2010a. 6
- M. Volpi, D. Tuia, and M. Kanevski. Advanced active sampling for remote sensing image classification. In *IEEE International Geosciences and Remote Sensing Symposium IGARSS, Honolulu (USA)*, pages 1414–1417, 2010b. 8
- M. Volpi, D. Tuia, G. Camps-Valls, and M. Kanevski. Unsupervised change detection in the feature space using kernels. In *IEEE International Geosciences and Remote Sensing Symposium IGARSS, Vancouver (CAN)*, pages 106–109, 2011. 6
- M. Volpi, G. Matasci, D. Tuia, and M. Kanevski. Enhanced change detection using nonlinear feature extraction. In *IEEE International Geosciences and Remote Sensing Symposium IGARSS 2012, Munich (D)*, pages 6757 – 6760, 2012a. 7, 119
- M. Volpi, D. Tuia, G. Camps-Valls, and M. Kanevski. Unsupervised change detection with kernels. *IEEE Geoscience and Remote Sensing Letters*, 9(6):1026–1030, 2012b. 7, 101, 103
- M. Volpi, D. Tuia, and M. Kanevski. Memory-based cluster sampling for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(8):3096–3016, 2012c. 8
- M. Volpi, F. de Morsier, G. Camps-Valls, M. Kanevski, and D. Tuia. Multi-sensor change detection based on nonlinear canonical correlations. In *IEEE International Geosciences and Remote Sensing Symposium IGARSS, Melbourne (AUS)*, 2013a. 7, 119
- M. Volpi, G. Matasci, M. Kanevski, and D. Tuia. Multi-view feature extraction for hyperspectral image classification. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning ESANN, Bruges (B)*, pages 14–16, 2013b. 7, 119
- M. Volpi, G. P. Petropoulos, and M. Kanevski. Flooding extent cartography with Landsat TM imagery and regularized kernel Fisher’s discriminant analysis. *Computers and Geosciences*, 57:24–31, 2013c. 6, 73
- M. Volpi, D. Tuia, M. Kanevski, F. Bovolo, and L. Bruzzone. Supervised change detection in VHR images using contextual information and support vector machines. *International Journal of Applied Earth Observation and Geoinformation*, 20:77–85, 2013d. 6, 73
- K. J. Wessels, F. van den Bergh, and R. J. Scholes. Limits to detectability of land degradation by trend analysis of vegetation index data. *Remote Sensing of Environment*, 125:10–22, 2012. 64
- I. H. Woodhouse. *Introduction to microwave remote sensing*. CRC Press, Taylor & Francis, 2006. 9, 10

References

- C. Wu, B. Du, and L. Zhang. A subspace-based change detection method for hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, in press:-, 2013. 70
- F. Yuan, K. E. Sawaya, B. C. Loeffelholz, and M. E. Bauer. Land cover classification and change analysis of the Twin Cities (Minnesota) metropolitan area by multitemporal Landsat remote sensing. *Remote Sensing of Environment*, 98:317–328, 2005. 66
- A. Zhang and G. Jia. Monitoring meteorological drought in semiarid regions using multi-sensor microwave remote sensing data. *Remote Sensing of Environment*, 134:12–23, 2013. 22
- J. Zhong and R. Wang. Multi-temporal remote sensing change detection based on independent component analysis. *International Journal of Remote Sensing*, 27(10):2055–2061, 2006. 71
- H. Zwenzner and S. Voigt. Improved estimation of flood parameters by combining space based SAR data with very high resolution digital elevation data. *Hydrology and Earth System Sciences*, 13:567–576, 2009. 74