Universidade Estadual de Campinas
Instituto de Computação

Juan Felipe Hernández Albarracín

Genetic-Programming-Based Spectral Indices for Remote Sensing Image Classification

Índices Espectrais Baseados em Programação Genética para Classificação de Imagens de Sensoriamento Remoto

CAMPINAS
2017

Juan Felipe Hernández Albarracín

Genetic-Programming-Based Spectral Indices for Remote Sensing Image Classification

Índices Espectrais Baseados em Programação Genética para Classificação de Imagens de Sensoriamento Remoto

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Thesis presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientador: Prof. Dr. Ricardo da Silva Torres

Este exemplar corresponde à versão final da Dissertação defendida por Juan Felipe Hernández Albarracín e orientada pelo Prof. Dr. Ricardo da Silva Torres.

CAMPINAS
2017

Agência(s) de fomento e nº(s) de processo(s): CNPq, 134089/2015-4
ORCID: http://orcid.org/0000-0002-3997-4422

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

H43g

Hernández Albarracín, Juan Felipe, 1990-
Genetic-programming-based spectral indices for remote sensing image classification / Juan Felipe Hernández Albarracín. – Campinas, SP : [s.n.], 2017.

Orientador: Ricardo da Silva Torres.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Sensoriamento remoto. 2. Programação genética (Computação). 3. Redução de dimensionalidade (Estatística). 4. Imagens digitais. I. Torres, Ricardo da Silva,1977-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Índices espectrais baseados em programação genética para classificação de imagens de sensoriamento remoto
Palavras-chave em inglês:
Remote sensing
Genetic programming (Computer science)
Dimension reduction (Statistics)
Digital images
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Ricardo da Silva Torres [Orientador]
Thiago Sanna Freire Silva
Sandra Eliza Fontes de Avila
Data de defesa: 17-03-2017
Programa de Pós-Graduação: Ciência da Computação

Universidade Estadual de Campinas
Instituto de Computação

Juan Felipe Hernández Albarracín

Genetic-Programming-Based Spectral Indices for Remote Sensing Image Classification

Índices Espectrais Baseados em Programação Genética para Classificação de Imagens de Sensoriamento Remoto

Banca Examinadora:

- Prof. Dr. Ricardo da Silva Torres
  Instituto de Computação - UNICAMP
- Prof. Dr. Thiago Sanna Freire Silva
  Instituto de Geociências e Ciências Exatas - UNESP
- Profa. Dra. Sandra Eliza Fontes de Avila
  Instituto de Computação - UNICAMP

A ata da defesa com as respectivas assinaturas dos membros da banca encontra-se no processo de vida acadêmica do aluno.

Campinas, 17 de março de 2017

Acknowledgements

I would like to start by thanking UNICAMP and the Institute of Computing for giving me the opportunity of doing my Master in such a prestigious institution. I thank the professors I had contact with, who somehow contributed in my formation as a scientist; and the administrative and the support staff, as they were determinant in specific moments of the course.

Most of my gratitude goes, definitely, to Prof. Dr. Ricardo Torres, my supervisor and principal partner in this rewarding experience, for his instruction regarding the scientific endeavour, and for giving me his support and autonomy I needed in the research. Every lesson learned from him is something I sincerely appreciate.

I thank Prof. Dr. Jefersson dos Santos and M.Sc. Edemir Ferreira (DCC/UFMG), key contributors of our work, for all their collaboration. People from E-Tribes group, with whom I learned the importance of multi-disciplinary teams: Prof. Dr. Marina Hirota (FSC/UFSC), Prof. Dr. Rafael Oliveira (IB/UNICAMP), Prof. Dr. Vinicius Dantas (IG/UFU), and my colleague Alexandre Almeida, with whom in more than one occasion I joined forces to achieve various objectives. I thank Dr. João Vila e João Santos too, from Embrapa, for all the disposition and interest about this research.

The valuable contributions from the examining committee members, Prof. Dr. Thiago Silva and Prof. Dr. Sandra Avila to this research were fundamental in assuring its quality and pertinence. My gratitude goes to them too.

I am enormously grateful to my colleagues from RECOD Laboratory, for all the experiences we lived together that allowed me to learn many things about my profession. I thank my close friends and family too: anyone who knew about what I was doing in my Master contributed with something, academically or emotionally.

Finally, I would like to thank CAPES, CNPq, and the Microsoft-FAPESP virtual institute (grants #2013/50155-0 and #2013/50169-1) for the financial support.

Resumo

Sensoriamento remoto é o conjunto de técnicas que permitem, por meio de sensores, analisar objetos a longas distâncias sem estabelecer contato físico com eles. Atualmente, sua contribuição em ciências naturais é enorme, dado que é possível adquirir imagens de alvos em mais regiões do espectro eletromagnético além do canal visível. Trabalhar com imagens compostas por múltiplas bandas espectrais requer tratar grandes quantidades de informação associada a uma única entidade, coisa que afeta negativamente o desempenho de algoritmos de predição, fazendo nacessário o uso de técnicas de redução da dimensionalidade. Este trabalho apresenta uma abordagem de extração de características baseada em índices espectrais aprendidos por Programação Genética (GP), que projetam os dados associados aos pixels em novos espaços de características, com o objetivo de aprimorar a acurácia de algoritmos de classificação. Índices espectrais são funções que relacionam a refletância, em canais específicos do espectro, com valores reais que podem ser interpretados como a abundância de características de interesse de objetos captados à distância. Com GP é possível aprender índices que maximizam a separabilidade de amostras de duas classes diferentes. Assim que os índices especializados para cada par possível de classes são obtidos, empregam-se duas abordagens diferentes para combiná-los e construir um sistema de classificação de pixels. Os resultados obtidos para os cenários binário e multi-classe mostram que o método proposto é competitivo com respeito a técnicas tradicionais de redução da dimensionalidade. Experimentos adicionais aplicando o método para análise sazonal de biomas tropicais mostram claramente a superioridade de índices aprendidos por GP para propósitos de discriminação, quando comparados a índices desenvolvidos por especialistas, independentemente da especificidade do problema.

Abstract

Remote sensing is the set of techniques that allow, by means of sensor technologies, to analyze objects at long distances without making physical contact with them. Currently, its contribution for natural sciences is enormous, since it is possible to acquire images of target objects in more regions of the electromagnetic spectrum than the visible region only. Working with images composed of various spectral bands demands dealing with huge amounts of data associated with single entities, which affects negatively the performance in prediction tasks, and makes necessary the use of dimensionality reduction techniques. This work introduces a feature extraction approach, based on spectral indices learned by Genetic Programming (GP), to project data from pixel values into new feature spaces aiming to improve classification accuracy. Spectral indices are functions that map the reflectance of remotely sensed objects in specific wavelength intervals, into real scalars that can be interpreted as the abundance of features of interest. Through GP, it is possible to learn indices that maximize the separability of samples from two different classes. Once the indices specialized for all the pairs of classes are obtained, they are used in two different approaches to fuse them into a pixel classification system. Results for the binary and multi-class scenarios show that the proposed method is competitive with respect to traditional dimensionality reduction techniques. Additional experiments in tropical biomes seasonal analysis show clearly how superior GP-based spectral indices are for discrimination purposes, when compared to indices developed by experts, regardless the specificity of the problem.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Remote sensing is the set of techniques that allow to analyze objects at long distances. Making use of sensors (e.g., imaging, altimeters, spectroscopes), the captured radiation, most of the time from an upper perspective, is used to infer properties of the Earth surface [7]. Currently, its contribution for natural sciences is enormous, principally in those areas in which it is fundamental to take advantage of large amounts of data of the real world, acquired remotely. Some of those sciences are agriculture, geology, ecology, climatology, and medicine.

Imaging sensors can acquire images not only in the visible regions of the electromagnetic spectrum (red, green, and blue), but also in a much wider range. In addition, it is possible to acquire these data sequentially in different timestamps, so a temporal dimension may also be incorporated into the analysis.

Working with images composed of various spectral bands and in different timestamps demands the management of a big number of features of data associated with single entities. This affects negatively the performance in classification tasks, in terms of efficiency and accuracy. Normally, in multi-band image classification, the number of dimensions in which the entities (pixels or regions of interest) are represented, is unreasonably larger than the number of training entities, prompting the *Hughes phenomenon* (or *Curse of dimensionality*) [30], which causes a low generalization capacity due to very sparse data. Reducing dimensionality, with a minimum loss of discriminatory information, becomes necessary, since augmenting the training data normally is a costly option.

Due to the emission/reflectance of the target objects, for a specific classification task, some bands might be noisy or redundant and can be ignored [36]. Many physical or chemical properties of targets become evident in relations that may exist among two or more bands, provided by the spectral signature of objects in the real world [63]; however, deducing this kind of relations demands specialized knowledge, which normally is not available.

Existing approaches for dimensionality reduction are based principally on the idea that, most of the time, adjacent spectral bands contain similar information, so the state-of-the-art techniques look for intelligent ways to select spectral bands for classification purposes (see Section 3.1). A less popular approach consists in mapping the band space into a (normally) smaller space, through diverse transformations (see Section 3.2).

The present work introduces a spectral-index-based approach for classification of multi-

band images, using an evolutionary technique called Genetic Programming, which builds formulas (i.e., spectral indices) that arithmetically combine the bands and map entities into a one-dimensional space, where examples of different classes can be linearly separated.

Genetic programming is a bio-inspired search technique that comes up with solutions to complex problems by exploring a solution space, which generally encodes computer programs or formulas [44]. Inspired by the principles of the Darwinian theory of evolution, the algorithm starts with a randomly-generated initial population of candidate solutions to the problem (also known as *individuals*) and moves throughout the solution space relying on heuristics that tell where better solutions might be, by giving the chance to "fitter" solutions in the population, of being selected to "procreate" similar solutions that are hopefully better, and will be part of a next generation of individuals.

In the case of dimensionality reduction, a fit solution will be one that represents the entities in such a space where classification accuracy is improved. In the proposed technique, this desired representation will be achieved by a spectral index that combines the bands in such a way that samples from different classes are separated. Spectral indices can explore complex interactions between bands in order to find latent properties of real-world entities, yielding a more promising scenario than just excluding some bands.

It is important to notice that spectral indices are traditionally used to measure a feature of interest, so the process of learning an index is associated with the resolution of a regression problem. This makes the proposed framework general enough to address both problems: classification, because of the further application of the indices to this purpose, and regression, due to their functional nature.

The proposed approach transforms data and projects it into a smaller feature space, which potentially makes clear the key (but complex) discriminative properties that were latent in the original representation. This research aims to prove this as possible, as long as there exist the available data about the regions of the electromagnetic spectrum in which those variations occur. Two research questions are, therefore, formulated:

1. **Can Genetic Programming effectively select bands?**

   Genetic programming operates with all the bands of a scene, without applying prior knowledge about their relevance, and does not explicitly aim at ignoring those that are not discriminative. The algorithm can choose anytime a band that may be noisy. However, if the optimization objective is settled correctly, its dynamics may cause the irrelevant bands to be naturally ignored, after some iterations.

2. **Can a one-dimensional non-linear projection learned by Genetic Programming overcome one-dimensional linear ones?**

   The simplest way to project data into a one-dimensional space is by a linear combination of the spectral bands, aiming at maximizing inter-class distance, while minimizing intra-class distance. The optimal solution for those objectives, under normality assumptions, is already achievable [38]. Intuition tells that non-linear solutions are more powerful, but Genetic Programming is theoretically far away of reaching an optimal solution and may face serious difficulties in hard search spaces.

The structure of this document is described below. Chapter 2 introduces and explains the fundamental concepts involved in this research. Chapter 3 presents related work. The applications and different techniques developed are distributed in three chapters (5, 6, and 7). Each one of them presents its own pipeline and methodology, but they have the same basis (the Genetic-Programming-based index learning approach) as initial phase. The index learning phase is presented in Chapter 4, and its outcomes are explored in different applications, which are discussed in the next three chapters. Chapter 5 presents the application of the indices for binary classification. Chapter 6, in turn, presents their application in a multi-class classification scenario. Chapter 7 presents the potential uses of the indices in tropical biomes seasonal analysis. Chapter 8 presents the conclusions and draws possible future work.

Chapter 2

Background

This chapter introduces the fundamental concepts involved in the research. The broad area of interest is *Remote Sensing* (Section 2.1), the principal problems to be addressed are *index saturation* and the *Hughes phenomenon* (Section 2.2), and the proposed technique is *Genetic Programming* (Section 2.3).

2.1 Remote sensing

Remote sensing is the set of techniques used for the acquisition and analysis of information to deduce knowledge about objects or phenomena, without making physical contact with them [7]. Currently, the term *remote sensing* is mostly used to refer to the analysis of information collected by upper-perspective sensor technologies (e.g., satellites, aircraft platforms) about targets on the terrestrial surface. The data is mostly collected in the form of images in one or more regions of the electromagnetic spectrum, which encode the quality of the radiation emitted or reflected by them. Images of the same geographic region in different channels of the spectrum (bands) are known as *multi-* or *hyper-spectral scenes*, depending on the quantity and continuity of the bands that compose them. Temporal analysis is possible as well, thanks to the sequential acquisition of data from sensors.

The general process of remote sensing is depicted in Figure 2.1. An energy source (normally, the Sun) sends radiation that goes through the Earth's atmosphere before reaching an object of interest. The radiation resulting of the interaction of the source radiation and the target object (i.e., radiance) is captured by sensors (after going through the atmosphere again), which send data to processing units that prepare it to be interpreted and analysed, and thus applied for a particular task [7].

For the purposes of this research, there are two key concepts that will be explained in detail: *spectral imagery* and *spectral indices*.

2.1.1 Spectral imagery

Scenes collected by sensors can be either multi- or hyper-spectral, depending on the quantity of bands and continuity of the spectral ranges contained in them. Each band of the image encodes the aggregation of the emittance/reflectance in a specific range of the spectrum. The difference between multi- and hyper-spectral images is shown in Figure 2.2.

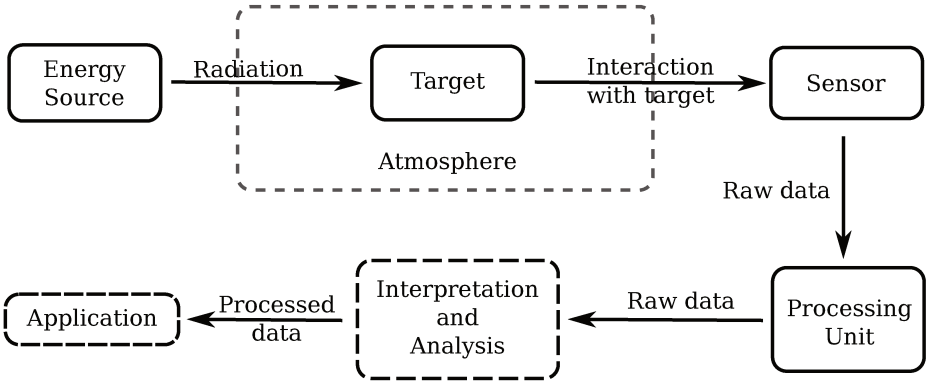
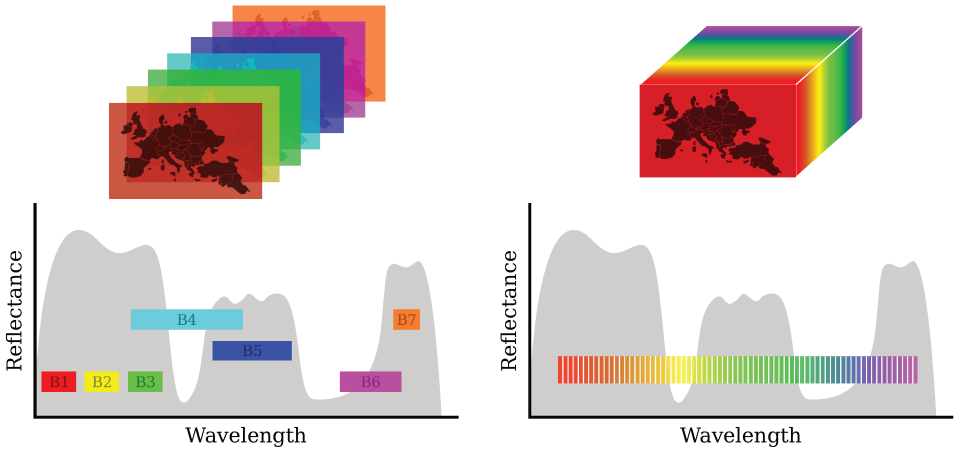


Figure 2.1: General process of remote sensing.



(a) Multi-spectral (b) Hyper-spectral

Figure 2.2: Spectral imagery. Bands of multi-spectral images are not necessarily homogeneous (a) while, for hyper-spectral images (b), they are.

The bands in multi-spectral images (Figure 2.2a) represent different regions of interest that normally are not homogeneous, since they are not equally distributed throughout the spectrum so they usually have different widths and some overlapping. This configuration is settled intentionally in each sensor, after careful planning, responding to requirements of specific applications. Multi-spectral images have few bands with respect to hyper-spectral ones (Figure 2.2b), since the latter quantize, homogeneously, and with high resolution, a large region of the spectrum.

2.1.2 Spectral indices

The acquisition of images in different spectral bands has a huge relevance, since materials and chemical substances can be detected by means of its *spectral signature*: the reflected/emitted electromagnetic radiation of an object in different wavelengths. The spectral signature is a function that helps identifying objects on the surface [7]. For this reason, a great part of the efforts in spectral image processing focuses on deducing associations between bands (or transformations of them) and physical/chemical features of

objects for their recognition.

Spectral indices are one of the most used tools in remote sensing to represent features of the objects in terms of their spectral signature. They are functions that map the reflectance/emittance at one or more wavelengths to real scalars that can be interpreted as the abundance of features of interest from target objects in the environment.

A relevant example is NDVI (acronym for *Normalized Difference Vegetation Index*), a vegetation spectral index for biomass measurement [58], based on the difference between the reflectance of an object in the visible (red) channel and its reflectance in the near-infrared channel:

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{2.1}$$

where NIR is the emission rate in the near-infrared channel and VIS is the emission rate in the visible (red) channel. Healthy vegetation in most cases presents $NDVI$ values close to 1, since it reflects big quantities of infrared radiance while absorbing big quantities in visible radiance [63]. Vegetation indices are the most common type of spectral indices and NDVI is only one of the vast amount of indices that exist for vegetation monitoring. Besides vegetation, spectral indices are widely used in other areas such as geology and urban planning and, although they are normally formulated by specialists, they can also be deduced automatically, relying on ground-truth data and pattern recognition techniques.

The application of indices to spectral imagery demands the calibration of the raw pixels, proceeding from the data collected by sensors, into physically meaningful units [56]. These issues are beyond the scope of this work, so the availability of calibrated data for the experiments is taken for granted.

2.2 Issues on remote sensing hyper- and multi-spectral image analysis

The present research aims to address two issues: *index saturation* and the *curse of dimensionality*. This section describes in detail both phenomena.

2.2.1 Index saturation

Human-developed spectral indices are normally purpose-specific and, although they are used in a wide range of applications, they are few in contrast to the huge number of specific needs present in natural sciences. One single index is used to describe different (but related) phenomena to the one it was originally developed, giving rise to a very common issue called *saturation*, in which the index is not descriptive enough of the phenomenon it is intended to point out.

Figure 2.3 illustrates the problem of saturation. The spectral index (y-axis), is supposed to describe a feature of interest whose real values were obtained through some in situ measure (x-axis). It can be seen that from one point on, the index fails at describing effectively the feature of interest, since its variation is significantly lower than the varia-

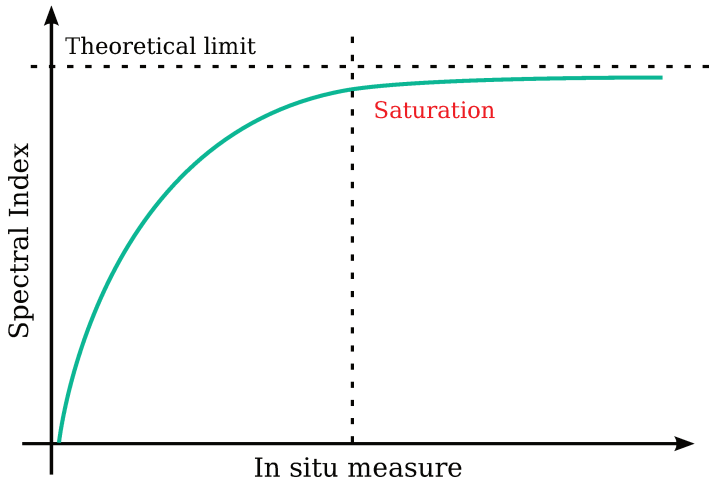


Figure 2.3: Index saturation phenomenon. From a point on, the index is not descriptive enough of the phenomenon it is intended to point out.

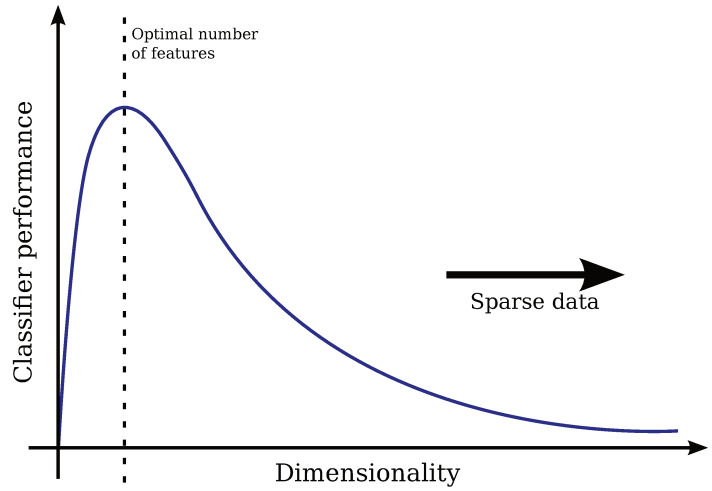


Figure 2.4: Curse of dimensionality. More sparse data yield poor prediction performance.

tion of the feature. It is expected that more specialized indices, which do not necessarily have to be human-developed, can be used to address this phenomenon.

2.2.2 Curse of dimensionality

The *Hughes phenomenon* (or *curse of dimensionality*) [30] is a widely known problem in machine learning that causes a decrement on the predicting power of the algorithms, due to highly sparse data. It occurs principally when the number of dimensions is extremely large with respect to the number of samples.

Figure 2.4 illustrates the behavior of the prediction performance in function of the number of dimensions, keeping the number of samples constant. When it is not possible to increase the volume of data, so the feature space stops being sparse, the one solution to address this problem is reducing dimensionality.

Remote sensing scenery is prone to experience this problem because, for any single entity (pixel or region of interest), a lot of information (coming from all the spectral

bands) is associated. This information, besides being redundant, is regularly misleading (noisy) for classification purposes. In this scenario, the use of dimensionality reduction techniques is mandatory.

Dimensionality reduction

Dimensionality reduction is a very common process performed in various machine learning tasks [11]. It consists in mapping entities from a high dimensional feature space into a space with less dimensions, aiming at maintaining relevant discriminative information. Some of the advantages of applying dimensionality reduction to data are: reducing processing time, saving storage space, reducing noise in data, and improving its visualization. In the context of spectral imagery, dimensionality reduction is widely used, since many of the bands are usually noisy or redundant for a specific purpose.

There are two main families of techniques used in this field [36]: *feature selection* techniques focus on selecting the best subset of features, excluding the less informative ones. *Feature extraction* techniques focus on generating new features by applying transformations to the original set of features. Both families contain supervised and unsupervised techniques; the former evaluates the performance of the reduction in terms of the distribution of the data while the latter, in terms of statistical measures such as correlation, mutual information or separability.

2.3 Genetic Programming (GP)

Genetic Programming (GP) [44] is a machine learning technique that belongs to the family of Evolutionary Algorithms (EA). EAs comprise a set of heuristic-based stochastic search techniques that are population-based and, inspired by the principles of Darwinian theory, simulate the dynamics of evolution of life forms, in order to guide searches in complex solution spaces.

In EAs, candidate solutions to a complex problem are represented as individuals within a population. A fitness function is defined as a criterion to select and maintain/reproduce the solutions of the current population that best solve the problem, according to the principle of survival of the fittest.

A typical EA starts with a population of randomly generated individuals (see Figure 2.5). Each iteration of the algorithm is considered a generation, in which individuals reproduce (Figure 2.5-b), creating new similar ones to form a new population (Figure 2.5-d). The fitness function is used to guide the selection of individuals to reproduce (i.e., create new solutions similar to the original), by attributing a higher chance of selection to those individuals with higher fitness (Figure 2.5-a). Different selection approaches can be used [44]. The algorithm keeps iterating until a stop condition is reached, e.g., a maximum number of generations, a specific fitness value, or the convergence of the fitness function.

Particularly in GP, the individuals are computer programs or formulas, whose fitness scores depend on the result of their execution. The next sections will describe in detail how the algorithm operates with the solutions.

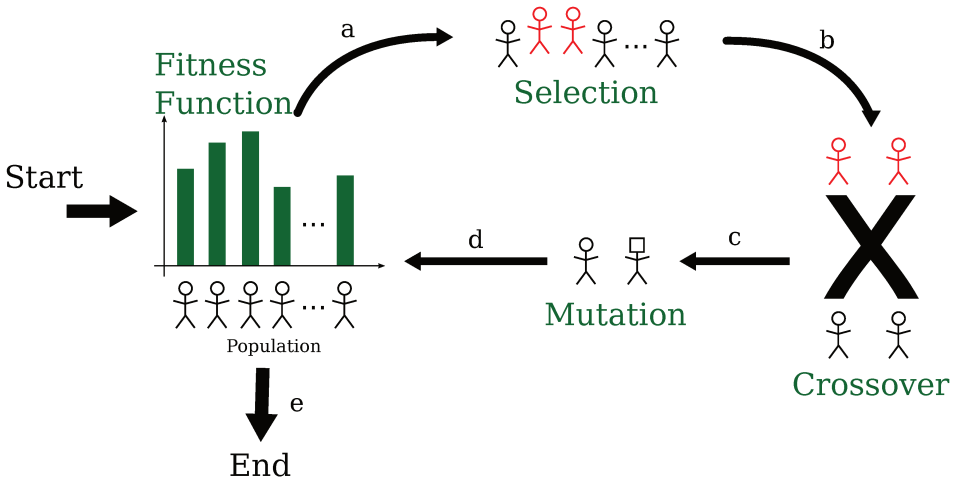


Figure 2.5: The general evolutionary algorithm process.

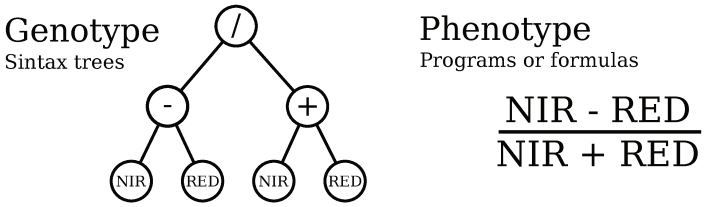


Figure 2.6: Genotype (computational representation) and phenotype (interpretation) of NDVI.

2.3.1 Computational representation of solutions

Candidate solutions need to be encoded into a manageable computational representation in order to operate with them. In EA terminology, the computational representation of the solution is known as *genotype* and the human-understandable representation is known as *phenotype* (e.g., the computer programs and the formulas, in the case of GP).

The most popular representation of GP solutions are syntax trees, which can encode efficiently instructions, functions, variables, and parameters present in computer programs or formulas. The leaves of the trees encode the parameters and variables that may be used in the program or formula, while the internal nodes represent the functions that will operate those parameters and variables.

A program or formula can also be represented linearly as a sequence of instructions in a specified format. Figure 2.6 shows an example of a syntax tree (genotype) and its corresponding formula (phenotype) of an individual, which will be the representation used in this work.

2.3.2 Genetic operators

Heuristic-based stochastic search relies on two fundamental activities: *exploration*, with which the algorithm moves dynamically throughout the solution space and *exploitation*, with which the algorithm focuses on exploring deeper a promising region [10]. In EA, these two activities are performed, respectively, by the *mutation* and *crossover* genetic

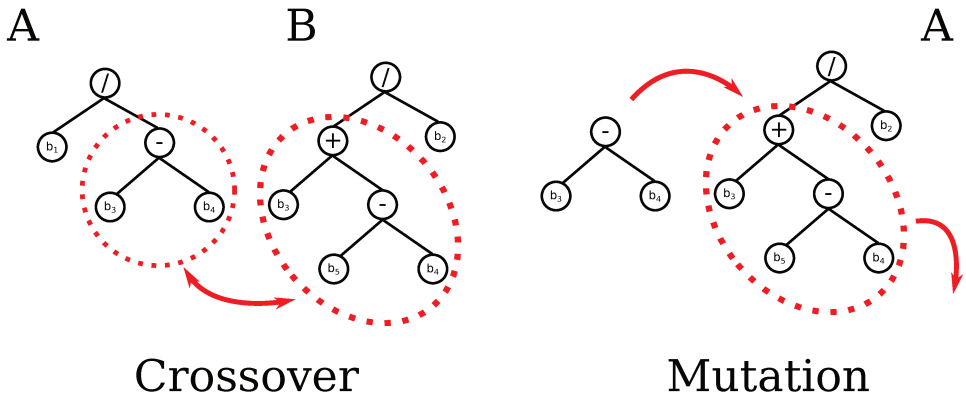


Figure 2.7: Crossover and mutation operators for syntax trees.

operators. The former helps to escape from local optima while the latter endeavors to advance to a better solution [51]. Both must be proceeded by a *selection* process.

Since it is desirable that the best individuals are suitable to reproduce maintaining the stochastic dynamics of the algorithm, an appropriate selection phase must be performed first. On the basis of the fitness values of all individuals in the population, a cumulative distribution function is built, so an individual can be sampled with a probability proportional to its fitness score with respect to the rest. If this sampling is repeated until there is enough selected individuals, this process is called *Roulette selection*, which is the most intuitive when selecting for the fittest candidates. However, this method can be inefficient for large populations, so the most popular alternative is *Tournament selection*, in which a random subset of individuals is sampled with equal probability, and the best of the set is selected.

Crossover is the operation with which two individuals reproduce, exchanging genetic information, and then creating new individuals, similar to the originals. In the case of syntax trees, the operation between two individuals is performed by exchanging sub-trees. The structure of the trees guarantees that this operation always creates a new syntactically correct individual. For the linear representation, small portions of the instructions are exchanged, and some verification tasks must be executed in order to guarantee syntactical consistency. The probability with which two selected individuals can go through crossover is a parameter of the algorithm, as well as the maximum size new formed individuals can have. Normally, the crossover probability is above 80% [44].

New formed individuals have a chance to go through *mutation*, by arbitrarily changing one of their sub-trees for a randomly generated (and syntactically correct) new one. In linear representation, a new portion of instructions is generated and replaces a portion of the original program; similar to the crossover operator, verification tasks must be performed. The probability of mutation is normally low, since mutations usually destroy good candidate solutions. Figure 2.7 illustrates the crossover and mutation operators for syntax trees.

Chapter 3

Related work

A common approach used to construct effective hyper- and multi-spectral remote sensing image classification systems relies on the selection of appropriate bands and their combination with the objective of providing a better discrimination among the pixel values observed for different classes. Many approaches have been proposed regarding this matter [46], and this chapter summarizes the most representative ones.

3.1 Band selection techniques

Band selection-based methods typically use a metric to select the most informative bands and provide a subset that could be the most appropriate. This strategy is based on the fact that some bands may contain noisy or redundant information, being bad candidates to be selected. It is very useful in many applications because it can not only reduce dimensionality, but also preserves relevant original information of the spectral bands. If prior knowledge is available, supervised methods [28, 31, 52, 54, 70] can be used to achieve good classification results (above 85% of accuracy, tested on several datasets).

Concerning the unsupervised band selection methods, some approaches exploit clustering techniques, in which the first step is to compute a distance measure for each pair of bands. With these metrics, the bands are grouped in disjoint clusters such that bands in a given cluster tend to be similar among them, according to these metrics, and bands in different clusters tend to be dissimilar. After the grouping step, a representative band from each cluster is chosen [16, 26, 33, 34, 72]. The results associated, with several datasets yield accuracies above 90%. Search-based methods, on the other hand, aim at finding a good set of bands by evaluating subset of features. Using exhaustive search strategies to find the best subset is normally unmanageable for this kind of data, however, several sub-optimal search strategies like sequential backward selection [43] or evolutionary techniques are used in this kind of problem [21, 71]. Rank-based methods use metrics, such as entropy [35], mutual information [4, 24, 25], and correlation [17], to sort the bands by their importance in the selection. The results associated with rank-based methods, with several datasets yield accuracies above 85%.

Many of the works presented in this section run experiments in famous hyperspectral

scenes,[1] that can be considered as benchmarks. In general, the techniques present in the literature have obtained competitive results on these datasets (above 80%).

3.2 Band combination techniques

Band combination techniques are included in the family of feature extraction techniques (see Section 2.2.2). The most conventional methods within this field are Principal Components Analysis (PCA) [38], Linear Discriminant Analysis (LDA) [50], Nonparametric Weighted Feature Extraction (NWFE) [45], and the Minimum Noise Fraction transformation (MNF) [23], which have been widely studied and currently are considered as baselines of more sophisticated methods.

In 2014, an unsupervised cluster-based feature extraction method (CBFE) was introduced [32]. This method uses K-means to cluster the feature vectors formed by the spectral bands and training data, and calculates the mean vector of each cluster, to be considered as part of the new representation. This technique obtained accuracies above 80%, outperforming various of the baselines described above.

Another work published in the same year exploited the fact that adjacent bands contain redundant information, and that there is a strong correlation between neighbor pixels [39]. This method consists in grouping adjacent bands and calculating their mean. The new representation is then processed by a Transform Domain Recursive Filter and the result is used as input to a SVM for region classification. This technique obtained accuracies above 98%.

A work introduced in 2015 [42] proposed a variation of the PCA-based extraction with pre-selection with a joint group sparse PCA (JGSPCA) algorithm, forcing the basic coefficients of groups of features to be jointly sparse and ensuring that the integrity of the features is preserved at most. The method outperforms other sparse-PCA-based techniques, obtaining accuracies above 80%.

Experiments run in these works are mainly on the same group of datasets introduced at the end of Section 3.1. In general, the techniques present in the literature have obtained competitive results on these datasets (above 80%).

3.3 Genetic programming and remote sensing

Various researches have been focusing on the use of GP for remote sensing image band combination. Usually, they aim to find formulas that measure the concentration of chemicals or the presence of an object.

Particularly, the method proposed by Fonlupt et al. uses GP to measure the concentration of phytoplankton, sediment, and yellow substance in oceans and coastal waters [19]. Using a dynamic fitness function to be optimized for each class at a time, it was possible to outperform traditional polynomial approximations, by reducing the root-mean-square error (RMSE) of the approximation almost to the half (0.012).

[1] Available at `http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes`. Last accessed on February, 2017.

Chion et al. [8] proposed the genetic programming-spectral vegetation index (GP-SVI), a method that evolves a regression model to describe the nitrogen level in vegetation. The fitness depends on the correlation with ground-truth data and the size of the formula. GP-SVI outperformed a group of regression methods, such as genetic algorithms for partial least squares regression (GA-PLS), multiple regression (MR), tree-based models (TBM), and some classical spectral vegetation indices such as NDVI, obtaining a slightly lower RMSE with respect to the baselines (0.15).

A very similar work was presented few years later by Puente et al. [58]. That work introduced a genetic programming vegetation index (GPSVI) to estimate the vegetation cover factor in soils to assess erosion. The proposed fitness function depended on the covariance with the cover-management factor obtained by in situ observations. The method outperformed vegetation indices such as NDVI and RVI.

A slightly different approach is presented by Ross et al. for mineral classification (3 classes) [62]. Binary classifiers are trained with GP and the fitness function depends on the rate of correctly classified examples.

Rauss et al. proposed to evolve an index in [61] that returns values greater than 0 when there is grass in the image, and values smaller than 0 in the other case. The fitness function depends on the number of correctly classified examples, relying on ground-truth data. No explicit quantitative results were reported.

Perkins et al., in turn, introduced GENIE [55], a hybrid algorithm based on linear genetic programming that finds a combination of successive image processing operations that can reproduce suitable feature planes for conventional classifiers.

The majority of the above initiatives uses GP to construct indices for regression tasks. The few approaches for classification [55,61,62] are applied in very close scenarios like few classes or isolated binary cases. Different from those methods, in this work, GP is used for classification purposes in general multi-class scenarios, relying on the quality of the feature space constructed by the band combination that was provided by the algorithm.

3.4 Spectral indices for classification tasks

Although spectral indices are widely used for continuous analysis purposes (e.g., regression), they have been used for classification tasks in various areas, mainly for plant coverage discrimination [15], but also in others such as astronomy [49] or flood monitoring [13].

Gerstmann et al. [22] introduces a method to find a general-form normalized difference index to separate cereal crops. It performs an exhaustive search in a set of possible permutations of bands and constants used in a general formula and measures the η^2 effect size to determine class separability. Classification was based on clustering of the pixels of the index image. The baselines used to assess the discrimination capacity of the new indices were NDVI, EVI (Enhanced Vegetation Index) and SAVI (Soil-Adjusted Vegetation Index), and the optimized index outperformed all of them.

Bhatt et al. [5] proposed a method to classify regions of interest integrating the scenes of vegetation, water, and built-up land spectral indices to classify soil coverage. The use of spectral indices reduces data correlation and redundancy between the bands, boosting

the discriminative capacity of their cluster-based classification algorithm.

Spectral indices have demonstrated ability to separate different classes of woody vegetation [47]. However, efforts that apply spectral indices for classification purposes normally rely on time series [9, 48, 57]. A single value in time of the indices about an object of interest is not enough to discriminate it. This is particularly true in vegetation analysis since an object of interest can change its state seasonally so, in order to classify it, it is necessary to look at its behavior though time. The possibility of novel indices that encode discriminative information invariant through time is open and could be achieved with the proposed technique (see Chapter 7).

Chapter 4

Spectral index learning via Genetic Programming

GP (see section 2.3) can build any kind of formula with few or no constraints, as long as the appropriate functions and parameters have been initially declared. It is expected that the learned indices (see section 2.1.2) through this method have a superior discriminative capacity than human-developed ones, since GP can explore the space of possible functions more freely. GP is a method that builds one-dimensional projections with many degrees of freedom, able to reproduce complex transformations that could be comparable, in terms of performance, with more sophisticated algorithms that consider more dimensions, since every classification task is at the end simple decision after complex transformations, regardless the dimensionality.

Therefore, this work introduces a GP-based technique for spectral imagery pixel-wise classification, aiming to find a spectral index that maximizes classification accuracy. The way to achieve it relies on finding an index that maximizes the separation of the distributions of different classes, so measures of separability of distributions will be used as fitness functions. This chapter describes in detail the first phase of the proposed technique: an early training step, which consists in the learning process of a spectral index.

4.1 Pipeline

Figure 4.1 shows the general pipeline of this phase. First, the labeled pixels of the scene are separated into training, validation, and test sets. For this phase, only training and validation sets are used.

Each individual of the population in every generation of the algorithm represents a candidate index (i.e., an arithmetic combination of the values of one pixel of the image in different spectral bands). In terms of its computational representation, they are syntax trees whose intern nodes stand for arithmetic operators and the leaves represent both variables b_i (corresponding to each spectral band) and real constants. For the first generation, randomly generated indices are used, and for the next ones, the population of indices will depend on the fitness function and the immediately previous population, as described in Section 2.3.

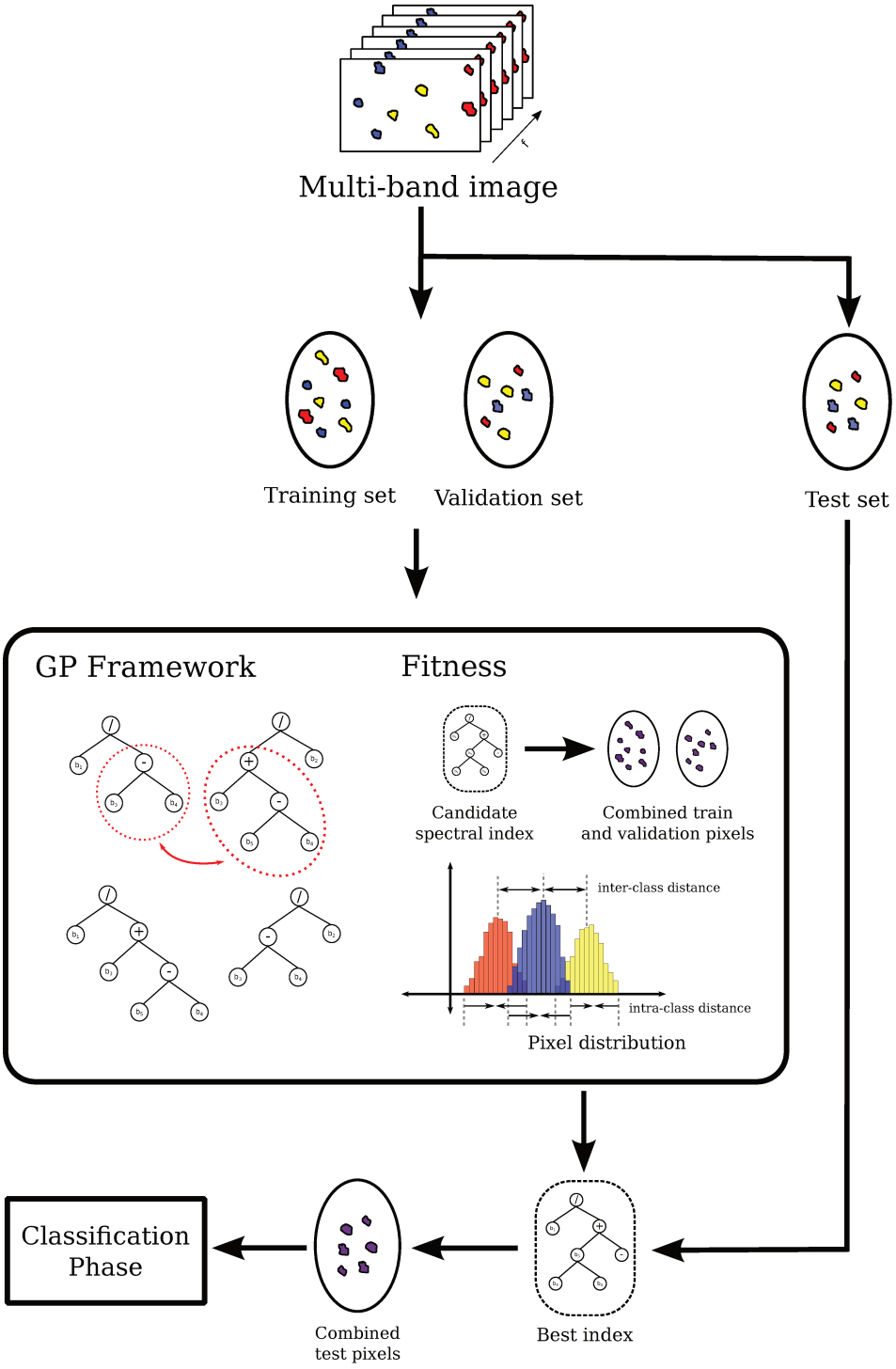
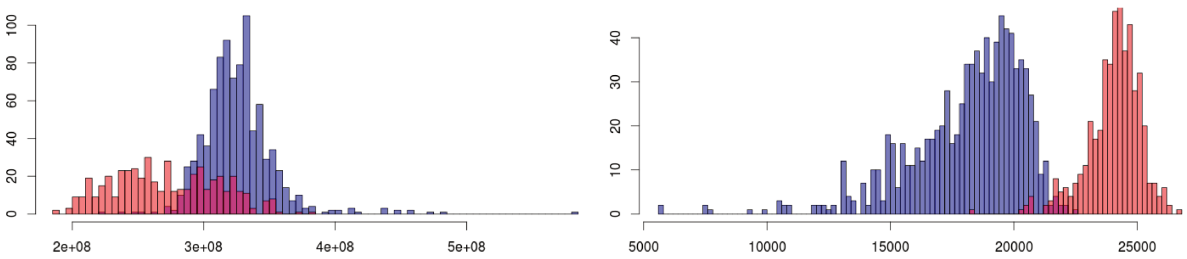


Figure 4.1: General pipeline of the index-learning phase.



(a) score = 0.64 (b) score = 0.84

Figure 4.2: Silhouette scores of class distributions in different moments of the evolution.

Each one of the candidate indices creates a single-band image, representing the combination of the bands for each pixel. Labeled data is necessary, since the main criterion to evaluate each index (by the fitness function) is the separability of the involved classes, which normally depends on inter and intra-class distance of the distribution of pixels from the generated gray-scale image: those of the same class should be close, and far away from pixels of different classes.

Training and validation sets are used in the GP framework to find the individual that best separates the distributions of the classes involved. The stop condition of the GP phase is the maximum number of generations reached, and once it is over, the best individual found is used to combine the bands of the test set pixel by pixel and the transformed image is used as input to a classification phase.

To avoid overfitting, a further selection of individuals is performed, using a validation set, as done by Torres et al. [12]. The validation consists in keeping the best k individuals through all generations and using each one of them to combine the pixels of the validation set. Let μ^i be the mean of the fitness values for the training and validation sets of individual i, and σ^i its standard deviation. The final score of the individual, S^i, is given by:

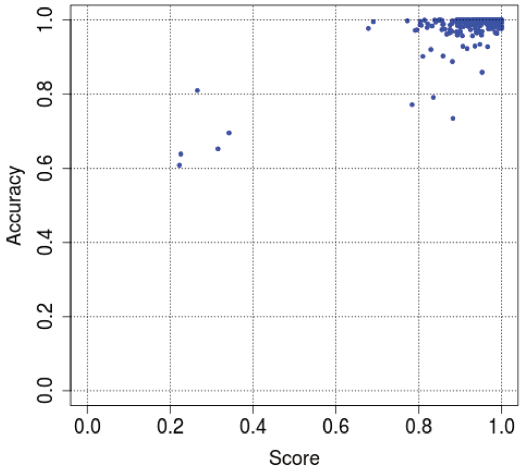
$$S^i = \mu^i - \sigma^i \tag{4.1}$$

The larger the difference between the fitness score for the training and the validation set, the larger σ^i and, probably, the smaller the final score. This score will impact negatively those individuals whose performance in the validation set is much smaller than the training set i.e., when an individual presents a bad generalization.

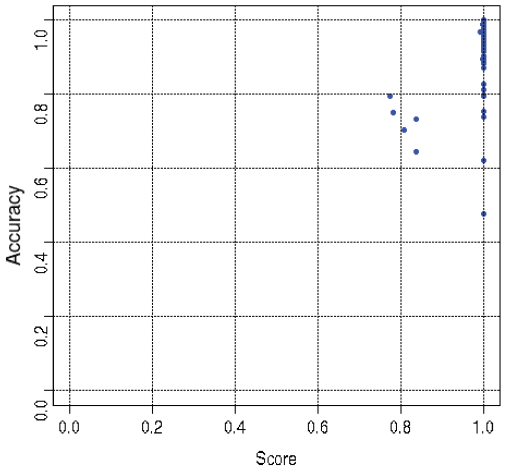
4.2 Fitness functions

Different separability measures were used as fitness function for a set of experiments. All of them associate a notion of distance between distributions with a real value, so it can be maximized in the GP phase. Figure 4.2 shows the distribution of the pixel values for two classes and the associated *silhouette score* (introduced later in this section) in two moments of the evolution of an index, intended to separate those distributions. It can be seen that the greater the value of the score, the more separated the classes are.

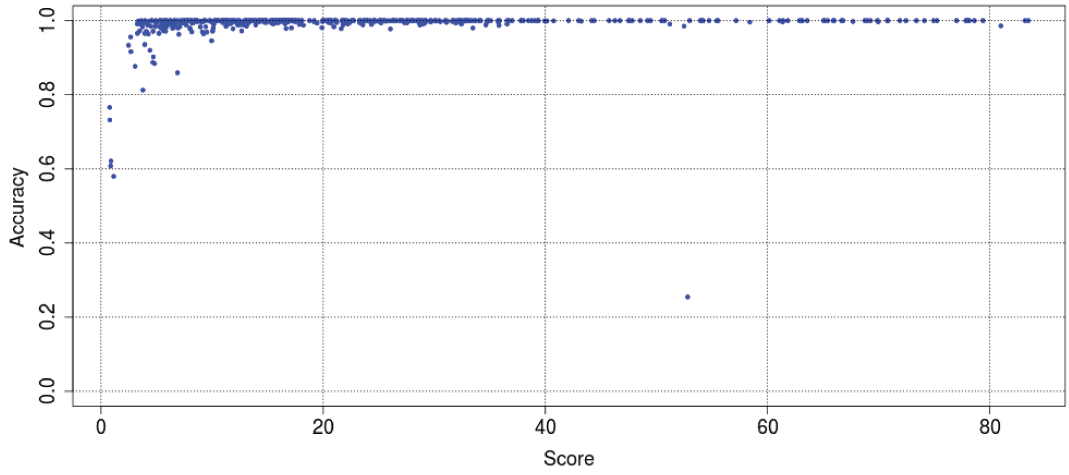
This section introduces briefly three of the best functions tested for a specific set of



(a) Silhouette score



(b) Classification with WARD method



(c) Normalized distance of means

Figure 4.3: Correlation of the silhouette score with the obtained accuracy for three different fitness functions.

experiments in order to be compared. The rest of fitness scores, which where considered as less effective, are presented in Appendix A. The functions were evaluated according to the mean accuracy obtained through the experiments, and the visualization of the accuracy obtained with the corresponding fitness score for each experiment. It is expected that high fitness scores are related to high accuracies. Figures 4.3 and A.1 show the scatter plots of the fitness scores versus the obtained accuracy in the same experiment.

4.2.1 Silhouette score

By considering each class as a distribution of values of pixels, the measure used to evaluate their separation in a space created by one spectral index is the *silhouette score* introduced by Rousseuw [64]. The silhouette, used principally for clustering, indicates how well-

assigned an object is to its cluster, with respect to the others, i.e., this score assesses how compact their distributions are and how far from the others.

Let the *average dissimilarity* of an object x to a cluster C_i be the mean of the distances of x to all the objects that belong to C_i. Let $a(x)$ be the average dissimilarity of x to its own cluster and $b(x)$ the smallest average dissimilarity of x to the other clusters, so $b(x)$ indicates how dissimilar x is with respect to the most similar cluster that is not its own. The silhouette of a single object x is defined as.

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \tag{4.2}$$

yielding values between -1.0 and 1.0. Negative values mean that there is at least one cluster that is more similar to x than the cluster to which it is assigned currently. The larger the silhouette, the better the assignment of x to its cluster. The overall score of the clustering method is the mean of the silhouettes of all the objects. The objective of the GP framework is to maximize this score.

This score can be applied to any number of classes and, since it has to compare all the pairwise distances, it runs in quadratic time. Figure 4.3a shows that there is a positive correlation between the silhouette score and the classification accuracy obtained.

4.2.2 Ward's method clustering-based classification

The use of a hierarchical clustering method was also tested as fitness function. This procedure consists in applying the Ward's Method [67], one of the most popular agglomerative clustering approaches, to the pixel values.

The final score depends on the *adherence* of the clustering to the real classes, i.e. how similar are the formed clusters to the real ones, provided by the ground truth. For a two-class scenario, this adherence is calculated in terms of classification accuracy: the clustering can be arranged in a confusion matrix and two accuracy scores (along both diagonals of the matrix) are calculated, choosing the greatest as the final score. In a multi-class scenario, another score, such as *Adjusted Rand Index* [60] may be used.

In detail, an agglomerative clustering method starts by assuming one cluster per sample and, relying on some determined criteria, merges the clusters in each iteration, until the target number of clusters is reached. *Ward's minimum variance criterion* consists in finding at each step the pair of clusters that leads to minimum increase in total within-cluster variance after merging and then proceed to join them. This algorithm has a complexity of $O(n^2 log(n))$, where n is the total number of examples. Figure 4.3b shows that the majority of experiments that obtained an accuracy of 1.0 yielded a silhouette of 0.8 or more.

4.2.3 Normalized distance of means

Let μ_a and μ_b be the mean values of all the pixel values of, respectively, classes a and b, and let σ_a and σ_b be their corresponding standard deviations. The score of separability of the two classes is given by:

$$S = \frac{|\mu_a - \mu_b|}{\max\{\sigma_a, \sigma_b\}} \tag{4.3}$$

The max function for the standard deviations was used instead of the sum or product, because a small standard deviation of one class can compensate the large standard deviation of the other, and they could obtain high separability values, even if they overlap. This measure returns real values greater or equal to 0. The larger the score, the better the separability.

This notion of distance cannot be extended to more than two classes, so the measure only works in a binary classification scenario. The calculation of this score runs in linear time. Figure 4.3c shows a positive correlation between the score and the obtained classification accuracy.

Chapter 5

Binary classification scenario

Once the spectral index is learned in the framework introduced in Chapter 4, it is used to project the data into a new feature space, where a classification algorithm is executed.

The classification stage for a binary scenario is introduced in this chapter. Spectral indices are projections into a one-dimensional space, and it is considered unnatural to distribute more than two classes in a one-dimensional space, unless there is a linear order relation between classes that indicates some notion of sequence (e.g., cold, warm, hot). Indexes separating more that two arbitrary classes can be learned and perform effectively, but may lack of sense in its interpretation. Far from being a disadvantage, the specialized indices learned for a binary scenario can be used as input to fusion approaches to yield a multi-class scenario, which is more realistic (see Chapter 6).

Experiments were executed on two widely used datasets with more than two classes so, for every possible pair of classes of the dataset, one index is learned.

This chapter describes the framework in detail, presents the experimental set up for discrimination between two classes, and presents and discusses the obtained results.

5.1 Pipeline

Counting on a good definition of the classes, provided by the best individual found in the GP phase, the classification task can be done with a Nearest Centroid Classifier (NCC), one of the simplest classification algorithms, in this new feature space.

Figure 5.1 illustrates the process. The best individual found is used to take the multi-band scene and combine each one of its pixels according to the formula of the index, forming a single-band image. The centroids of the two classes is obtained by calculating the mean of the training pixel values that belong to the same class. Once the centroids of each class are calculated, an unlabeled sample can be classified into the class of the closest centroid. Another central tendency measure, besides the mean, can be used to calculate the centroids of the classes.

The training data for this phase are the training and validation set of the GP phase.

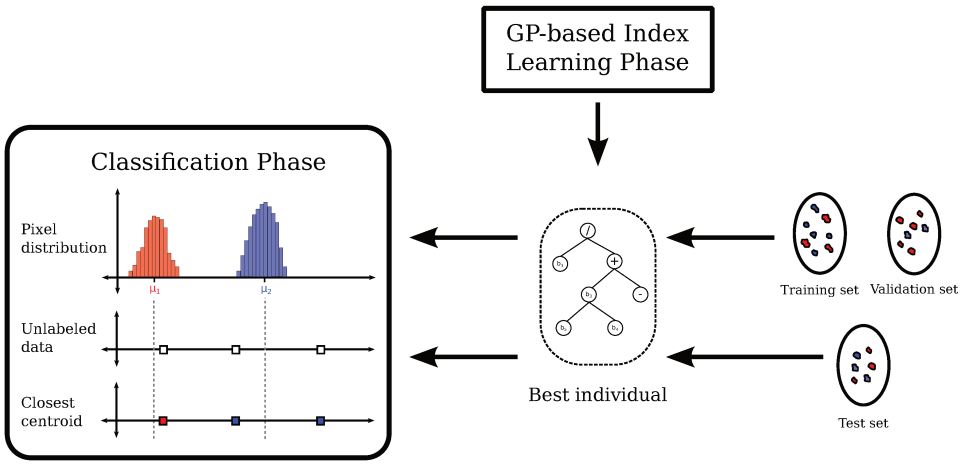


Figure 5.1: Pipeline for classification in a one-dimensional-space NCC algorithm.

(a) Salinas (b) Thetford Mines

Figure 5.2: Thematic maps of Salinas and Thetford Mines datasets.

5.2 Experimental protocol

This section describes how the experiments were set up, including description upon datasets, execution parameters and the evaluation protocol used.

5.2.1 Datasets

The method was tested on two datasets. The first one is a scene collected over Salinas Valley, California and the second one is a scene collected over an urban area near Thetford Mines in Quebec, Canada. For simplicity, the datasets will be called *Salinas* and *Thetford Mines*. Figure 5.2 shows the ground truth of both datasets. Details of each one will be presented below.

Table 5.1: Classes in the Salinas dataset and the number of examples per class.

	Class	**Num.**
1	Brocoli_green_weeds_1	2009
2	Brocoli_green_weeds_2	3726
3	Fallow	1976
4	Fallow_rough_plow	1394
5	Fallow_smooth	2678
6	Stubble	3959
7	Celery	3579
8	Grapes_untrained	11271
9	Soil_vinyard_develop	6203
10	Corn_senesced_green_weeds	3278
11	Lettuce_romaine_4wk	1068
12	Lettuce_romaine_5wk	1927
13	Lettuce_romaine_6wk	916
14	Lettuce_romaine_7wk	1070
15	Vinyard_untrained	7268
16	Vinyard_vertical_trellis	1807

Salinas

This scene was collected by the Airborne visible/infrared imaging spectrometer (AVIRIS) in 224 spectral bands at 10 nanometer intervals across the spectrum from 400 to 2500 nanometers. Its spatial resolution is 3.7 meters/pixel, forming a 512 x 217 image.[1] Its ground truth contains 16 classes, as shown in Table 5.1 and Figure 5.2a.

Thetford Mines

This scene was collected by a 84-channel sensor, with a spacial resolution of about 1 meter/pixel forming a 795×564 image. The wavelengths covered are between 7.8 and 11.5 micrometers. This is one of the data sets of the 2014 IEEE GRSS Data Fusion Contest.[2] Its ground truth contains 7 classes, as shown in Table 5.2 and Figure 5.2b.

5.2.2 Parameters

Table 5.3 summarizes the configuration of the GP algorithm. The parameters of the formulas (i.e., leaves of the trees) are random real numbers between 0 and 10^6, and the variables that indicate the spectral bands. Six possible arithmetical operators are considered for the formulas (i.e., internal nodes); three of them are the regular addition (+),

[1]Data retrieved from `http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes`. Last accessed on February, 2017.

[2]`http://www.grss-ieee.org/community/technical-committees/data-fusion/2014-ieee-grss-data-fusion-contest/`. Last accessed on February, 2017.

Table 5.2: Classes in the Thetford Mines dataset and the number of examples per class.

	Class	**Num.**
1	road	4293
2	trees	1027
3	red roof	1739
4	grey roof	1973
5	concrete roof	3797
6	vegetation	7167
7	bare soil	1711

Table 5.3: GP parameters setup.

Parameter	**Value**
Population size	200
Generations	300
Operators (intern nodes)	{+, -, *, %, $srt()$, $rlog()$}
Parameters (leaves)	$\{b_i : 0 \leq i < n\} \cup \{c_j \in [0, 10^6]\}$
Maximum initial tree depth	6
Maximum crossover depth	15
Selection method	Tournament $\times$ 3
Crossover rate	0.9
Mutation rate	0.1
Individuals kept for validation	10
Elitism	10

subtraction ($-$) and multiplication (*). The other three are *protected* versions of functions that are not defined for all the possible values of the operands. These new operators are recommended by Koza [44], since using them is more practical than controlling the values that enter to those functions. The *protected division* operation (%) avoids division by zero, by returning 1 every time the denominator is 0. The *protected square root* ($srt()$) operation avoids returning imaginary roots by calculating before the absolute value of the input value. The *protected natural logarithm* ($rlog()$) operation avoids the calculation of the natural logarithm for values equal or less than zero, by returning zero if the input value is zero, and returning the natural logarithm of the absolute value of the input if it is less than zero.

Experiments were performed with 200 individuals and 300 generations. New randomly generated trees will not have a depth greater than six, and crossover and mutation operations will not yield trees with a depth greater than 15.

The selection method is *tournament with three individuals*. It consists in selecting at random three individuals from the population and allowing the best one to go to crossover, as many times as a new generation is completed. Once two individuals are chosen, they have a probability of 0.9 to cross their genetic information and create new individuals.

Every new individual in the population has a probability of 0.1 to mutate.

For validation, the best ten individuals of all the generations are re-ranked according to the score described in Section 4.1. The best individuals found with and without validation are kept for classification, and their performances are compared. Taking advantage of this, elitism is included, by selecting the best 10 individuals of the population directly to the next generation without any modification.

The population size, the number of generations and the crossover and mutation rates are parameters that normally require tuning, regardless the domain of the problem [51]. The choice of the crossover and mutation rate constitutes a trade-of between the capacity of the algorithm to explore the search space and its capacity to conserve good solutions. The larger the population size and the number of generations, the better to achieve good solutions. However, a good combination of those rates is the key to make the algorithm converge in a satisfactory solution without the need of a high number of generations or big populations. The process of parameter tuning consisted in different runs of the evolution with a subset of the data in order to compare their convergence. Values presented in this section correspond to the best configuration found.

Experiments for each one of the fitness functions introduced in Section 4.2 where executed, obtaining highly equivalent results, in terms of accuracy, showing that none of them was significantly superior with respect to the others. For the sake of simplicity, only the experiment corresponding to the normalized distance of means fitness function (Section 4.2.3) is reported, because it is the less computationally expensive measure.

No stop condition different to the maximum number of iterations was considered, since it was not clear in the beginning what was the minimum fitness score needed to yield a good classification accuracy.

5.2.3 Evaluation protocol

Given a dataset with n classes, a total of $\frac{n(n-1)}{2}$ different experiments, considering every possible pair of classes were executed. The 5-fold cross validation approach was used to evaluate the performance of the classification of each pair of classes. The data is divided into three sets (folds), whose content is fixed for all the experiments performed on them. Three of the folds were used for training, one for validation, and one for testing. Five runs were executed in such a way that folds were shifted to be used for training, validation, and testing. The same evaluation protocol stands for the baselines, which are described below.

Two variations of the proposed method are compared: index learning with and without considering the validation set. The impact of the validation set is studied.

Baselines

Traditional band selection and/or combination methods are considered in order to measure the effectiveness of the GP-based band combination. They are:

1. **No selection (NS):** The raw values of the bands are used to form a feature vector of as many dimensions as number of bands.

Table 5.4: Results for binary classification scenario (% normalized accuracy). A ▲ symbol indicates that the proposed approaches were statistically superior to the corresponding baseline, **-** means that it was statistically inferior and • means that both methods were statistically tied.

	Salinas				T. Mines			
Method			μ	σ			μ	σ
NS	▲	▲	97.41	5.39	•	•	80.34	18.25
UFS	▲	▲	98.63	3.77	•	•	81.46	17.62
PCA	▲	▲	97.41	5.41	•	•	80.34	18.25
LDA	-	-	99.62	2.45	•	•	87.19	13.63
RFS	▲	▲	98.25	4.22	•	•	81.20	17.96
GP		•	99.19	3.01		•	84.16	15.68
GPVAL	•		99.17	3.09	•		83.95	15.79

2. **Uni-variate feature selection (UFS):** Unsupervised selection of the best ranked bands based on uni-variate statistical tests, in this case, the F-test.

3. **Principal component analysis (PCA):** Unsupervised orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables.

4. **Linear discriminant analysis (LDA):** Supervised linear projection of observations aiming to maximize class separability.

5. **Random forest selection (RFS):** Supervised band selection with the Random Forest classification algorithm, aiming at maximizing information gain.

The resulting feature vectors of each one of this methods are used as input to a NCC algorithm, as well as the GP-based approaches.

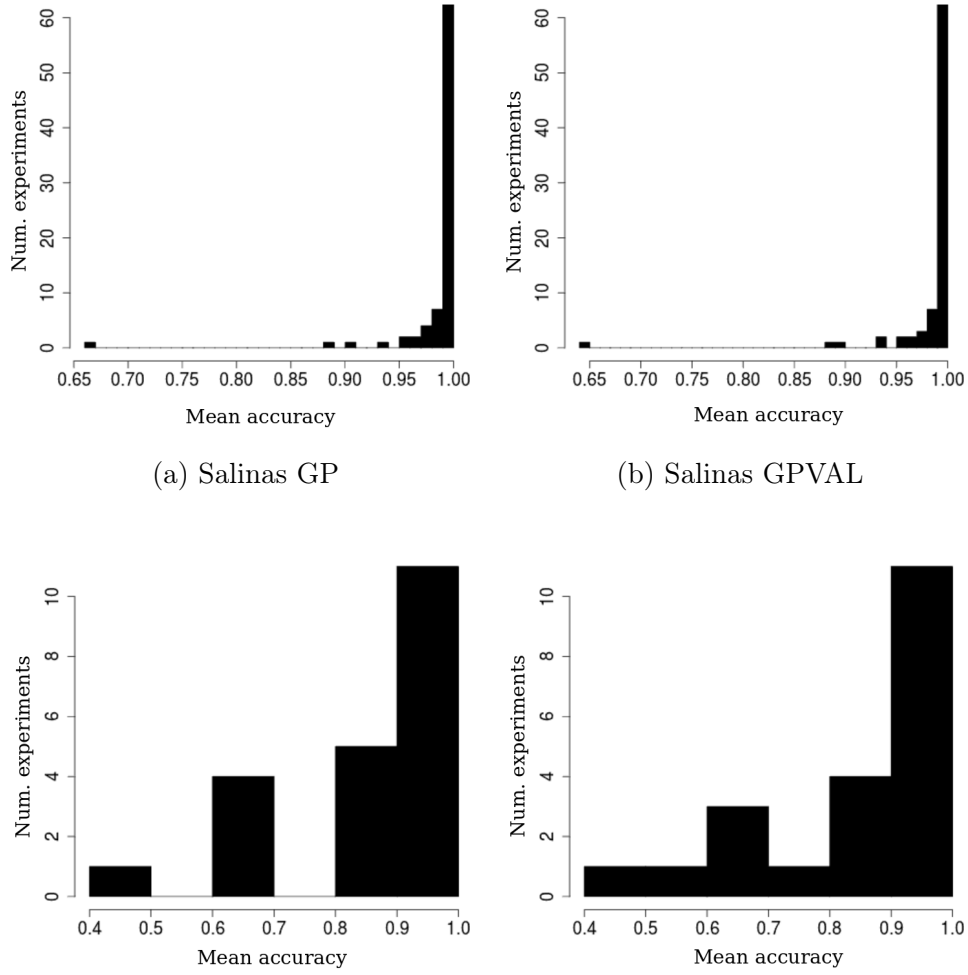
5.3 Results

In order to evaluate the impact of the validation set (see Section 4.1), this section reports the performance of two methods: NCC-based classification without validation set (GP) and NCC-based classification with validation set (GPVAL).

Additionally, the frequency of the selected bands in the GP-phase to build the spectral indices is analysed.

5.3.1 Classification accuracy

Due to the low inequality of the number of samples for each class in both datasets, the results consider the normalized accuracy, which is the mean of the rate of correctly classified samples for each class. For unbalanced datasets, this value is more reliable than the absolute accuracy, since it reduces the dominance of classes with more samples.



(a) Salinas GP

(b) Salinas GPVAL

(c) Thetford Mines GP

(d) Thetford Mines GPVAL

Figure 5.3: Histograms of the classification mean accuracies for all the possible pairs of classes in Salinas and Thetford Mines datasets.

A total of 120 independent runs of the cross validation were performed, corresponding to the different pairs that can be picked from the 16 classes of Salinas dataset. Similarly, a total of 21 independent runs of the cross validation were performed for the 7 classes of Thetford Mines dataset.

Figure 5.3 shows the distribution of the mean normalized accuracies of each cross validation. For Salinas dataset, it can be seen that at least half of the pairs obtained a mean classification accuracy of 100%, and only one pair obtained an accuracy smaller than 85% (*Grapes_untrained* x *Vinyard_untrained*). For Thetford Mines dataset, approximately half of the pairs obtained a mean accuracy of 100% and just a few obtained mean accuracies between 80% and 100%. There were pairs with a very poor performance, like *reed_roof* x *grey_roof* that obtained accuracies below 50%.

Table 5.4 shows the mean of the normalized accuracies for all the experiments associated with each method for both datasets. To the left of each of the mean accuracy of the baselines, there are two symbols, representing the statistical significance of the difference of each baseline with respect to the two methods: **GP** (first symbol) and **GPVAL** (sec-

ond symbol). A ▲ symbol indicates that the GP approach was statistically superior to the alternative, - means that it was statistically inferior and • means that both methods were statistically tied.

The statistical significance was estimated with the Friedman test for all the methods, considering each binary classification as a different problem. For Salinas dataset, the p-value associated is 1.5798×10^{-71} and for Thetford Mines dataset, it is 0.0017. Pairwise Wilcoxon Signed Rank tests (significance level of 0.05) with Bonferroni post hoc adjustment was performed in order to compare the proposed methods with the baselines and within themselves.

The GP approach was superior with respect to the rest of methods, except for LDA. For Thetford Mines dataset, the variability was high for all the methods, causing statistical tests to fail at coming up with conclusions. The comparison between the GP approaches with and without validation showed that they are not statistically different, even for Salinas dataset, which reports low variability, so it can be concluded that the use of the validation set had no effect in these experiments.

5.3.2 Band frequency within the indices

One of the advantages of spectral indices relies on the fact that they can be interpreted, through the arithmetic operations between the bands. This section presents the number of times a spectral band was used for all the indices learned in this set of experiments, as the simplest approach for analysing its relevance. For Salinas dataset, a total of 600 (120 pairs $\times$ 5 folds) indices were learned and for Thetford Mines dataset, a total of 105.

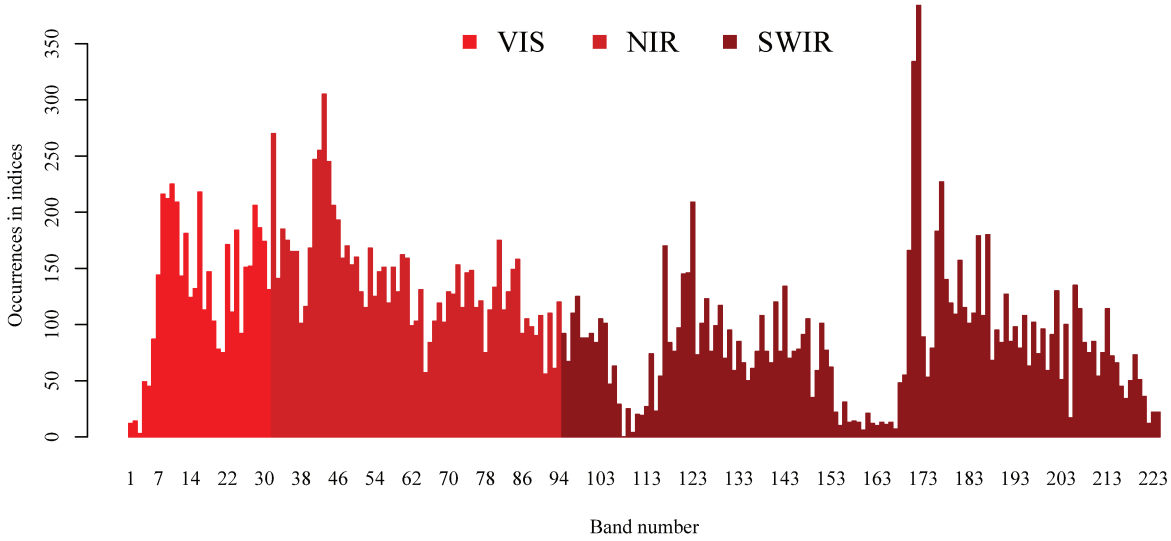
Figure 5.4 shows the histograms for both datasets, indicating the cumulative sum of the times a band is present in an index. The sensors of both datasets cover different regions of the infra-red: for Salinas, the information of the visible (VIS), near-infrared (NIR), and short-wavelength infrared regions is available. For Thetford Mines, only information of the thermal or Long-wavelength infrared (LWIR) is available.

The ten most frequent bands for Salinas dataset were, from the most to the less frequent, 172, 171, 43, 32, 42, 41, 44, 177, 10, and 16, from which two intervals can be highlighted: from $1.9661\mu m$ to $1.9758\mu m$ and from $0.7646\mu m$ to $0.7928\mu m$. For Thetford mines dataset, they were, from the most to the less frequent, 12, 11, 9, 69, 64, 45, 62, 10, 8, and 26, from which the interval from $8.1083\mu m$ to $8.2845\mu m$ can be highlighted. Additionally, two band intervals of notoriously low frequency were present for Salinas: 105-116 and 154-169. Both intervals match lightly the water absorption intervals 108-112 and 154-167 (according to the dataset specification[3]), which are generally ignored.

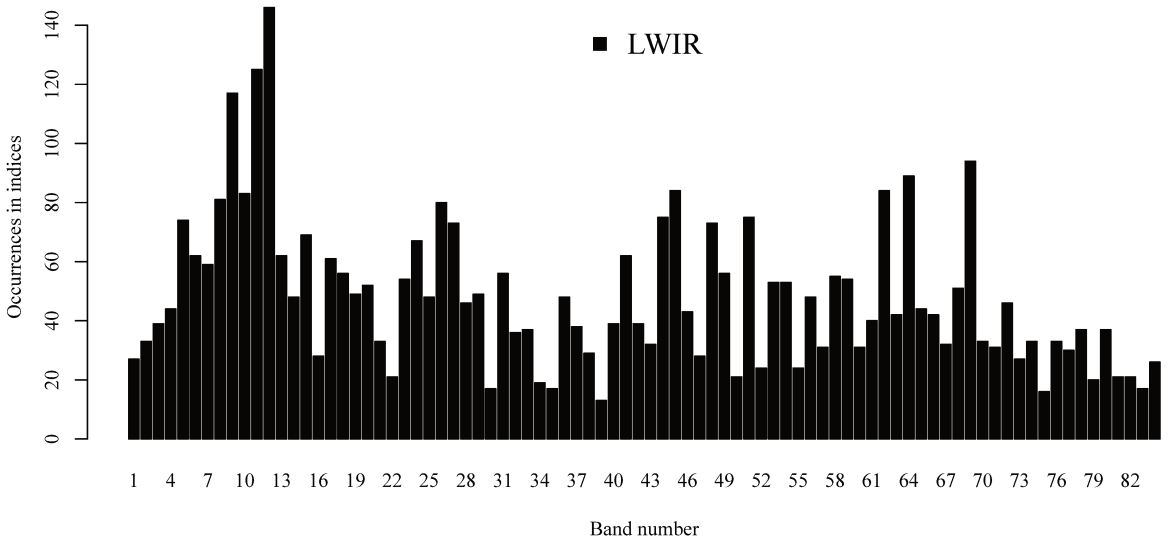
5.3.3 Examples of GP-based spectral indices

In this section, some of the spectral indices learned through the method introduced in Chapter 4, intended to separate samples of two classes, are visualized. Each one of the figures below has the the following structure: three images which, from left to right,

[3] `http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes`. Last accessed on February, 2017.



(a) Salinas



(b) Thetford Mines

Figure 5.4: Distribution of the frequencies of the spectral bands in the learned indices.

```
(b116 - b136) % (((b122 + b18) % ((b116 - b136) - ((((b116 - b40) - (b136 - b41)) - b189) * (((b189 +
b213) - ((((b116 - b40) - (b136 - b41)) - b189) * (b122 + ((b189 + b213) - b122)))) + (((b122 + b18) -
b41) - b122))))) * (b18 + (((b122 + b37) % (((b116 - b40) - (b136 - b41)) - (b189 + (b189 + ((((b189 +
b213) - ((b41 - b189) * (b122 + ((b189 + b213) - b122)))) - b122) % ((b116 - b40) + b37)))))) * (b122
                                        + (b122 + b37)))))
```

Figure 5.5: Visualization of the index obtained to separate *Brocoli_green_weeds_1* from *Brocoli_green_weeds_2* in the Salinas dataset. Mean accuracy: 0.9862.

represent the equalized image of the spectral index, the ground-truth of the classes to be separated and the equalization of the index in order to see the separation. The associated formula is presented below the images. In order to visualize the indices, the generated images of the corresponding scenes (using min-max normalization) had to be equalized, due to pixels with extremely low or high values with respect to the rest. This was necessary only for those indices that were complex.

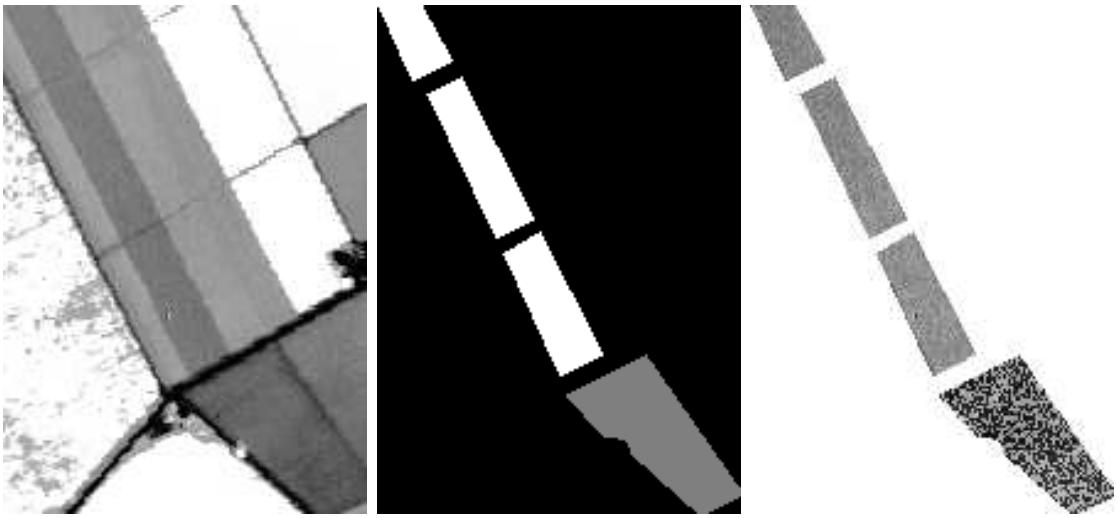
Some pairs of classes were easy to separate, so the learned associated indices were considered *trivial*, because they involved one spectral band only. For Salinas dataset, Figures 5.5 and 5.6 show two of the non-trivial indices learned that yielded classification accuracies above 0.9. Figure 5.7 shows one of the indices learned that yielded poor results. It can be seen that not even by binarizing the image, it was possible to evidence a good separation of the involved regions.

For Thetford Mines dataset, Figures 5.8 and 5.9 show two of the indices learned that yielded classification accuracies above 0.95. Resulting images were equalized and binarized as well. Figure 5.10 shows one of the indices learned that yielded poor results. As it can be observed, likewise the one shown for Salinas dataset in Figure 5.7, no good separation can be evidenced.

5.4 Discussion

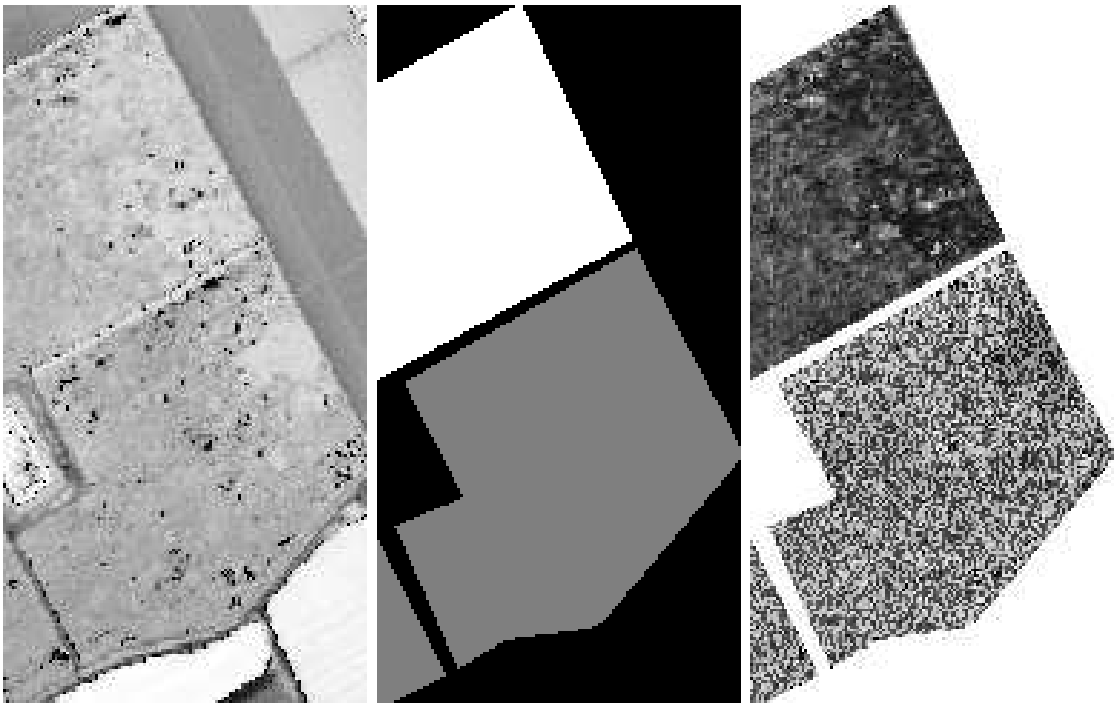
For the Salinas dataset, all the methods obtained mean accuracies higher than 0.98 (see Table 5.4), and LDA and the two proposed approaches presented the lowest standard deviations. The same three methods obtained the lowest deviations in the Thetford Mines dataset however, the deviations of all methods were so high that it was not possible to determine a statistical superiority of any of them, although the obtained mean accuracies were higher that 0.81.

For both datasets, the LDA algorithm for dimensionality reduction was the only one that the proposed method could not outperform significantly. LDA, when applied to binary classification, provides the optimal linear projection to a one-dimensional space,



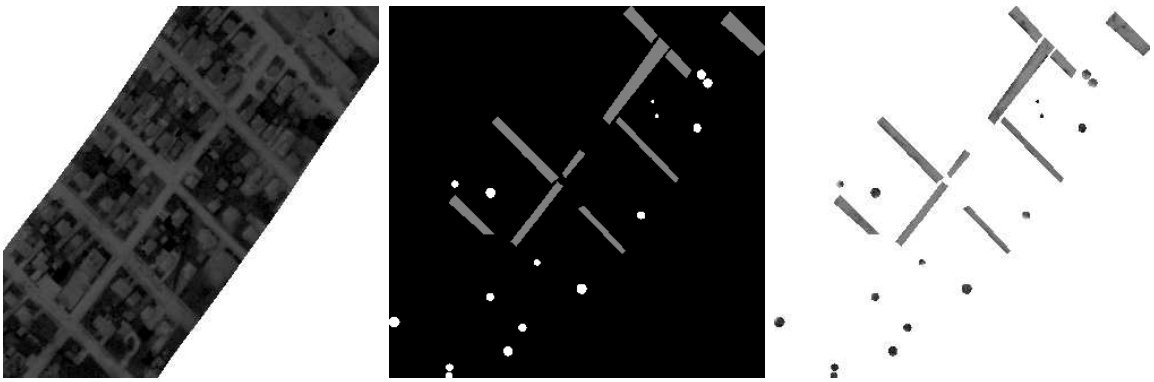
```
b133 - ((((((((((b200 % b133) % b133) * (b152 + b33)) % b133) % b133) * ((((((b200 % b133) % b133) *
((((((b200 % b133) % b133) * (b152 + b33)) % b133) * ((b81 * (b30 + b197)) + (((((b200 % b133) % b133)
    * ((((b200 % b133) % b133) * (b200 % b133)) * (b200 % b133))) % b133) % ((b33 % b133) * (b30 +
b197))))) * (b200 % b133))) % b133) % b133) * b33)) * b210) * (b152 + b33)) * b210) * (b212 * b191))
```

Figure 5.6: Visualization of the index obtained to separate *Fallow* from *Fallow_ smooth* in the Salinas dataset. Mean accuracy: 0.9321.



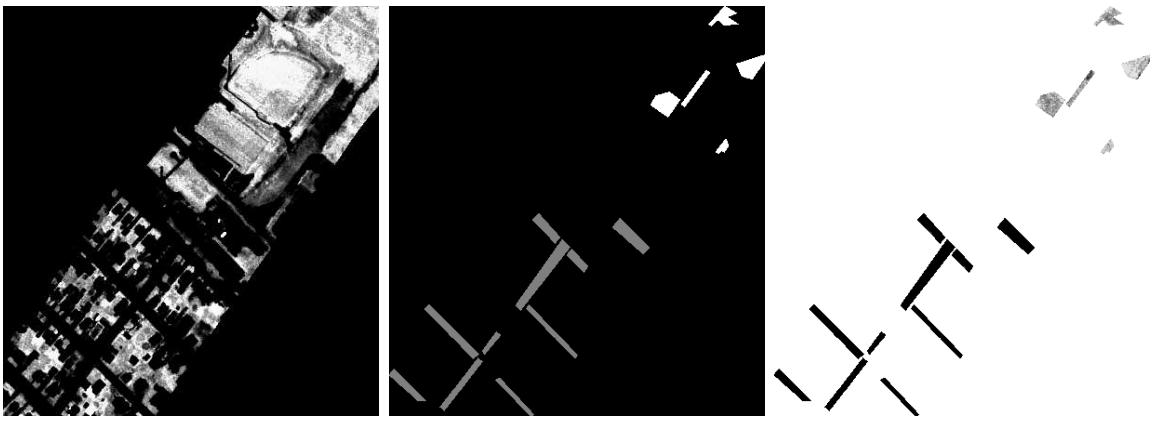
```
(((b60 - b101) % (b207 * b40)) % (b201 * ((b204 + b31) % b31))) % (b201 * (((b207 * b40) % ((b60 -
b101) % ((b61 % b70) % b70))) % (b201 + ((((((b61 % b70) % ((b61 % b70) % b70)) * (b61 % ((b61 % b70)
                    * b132))) * b31) - b31) % ((b140 % b32) % (b76 + b132))))))
```

Figure 5.7: Visualization of the index obtained to separate *Grapes_ untrained* from *Vineyard_ untrained* in the Salinas dataset. Mean accuracy: 0.6870.



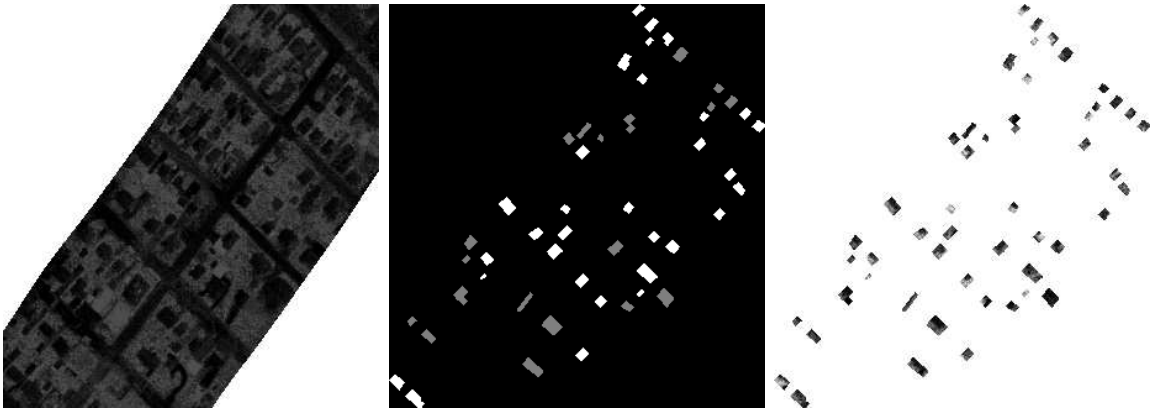
```
(((b0 - b41) - (b22 % b60)) % (((b14 + b10) * (b10 * (b27 * (b3 * b3)))) * (b10 * (b3 * b3)))) % ((b10
* (b3 * b3)) % (b3 + ((((b8 + b29) - b71) % ((b8 + b29) * ((b14 + b75) * (b8 + b29)))) % (((b14 + b10)
   * (b3 * (b44 + b55))) % ((((b71 % (b14 + b10)) % ((b0 - b41) * ((b8 + b29) * (b10 * (b27 * (b3 *
                              b3)))))) % (b10 * (b3 * b3))) + b55)))))
```

Figure 5.8: Visualization of the spectral index obtained to separate *road* from *trees* in Thetford Mines dataset. Mean accuracy: 0.9793.



```
((b61 + b17) - ((((b74 - (b18 + b6)) % (((b11 * b43) * (((b11 * b43) * ((((b11 * b43) * ((b11 * b43) *
((b53 + b5) * ((b53 + b5) * b61)))) * b61) * ((b6 * ((b6 * ((b11 * b43) * (b53 + b5))) * (b18 + b6)))
* b61))) * b43)) * (b48 + b15))) % b30) - (b48 + b15))) % ((((b11 * b43) * (b53 + b5)) * ((b6 * ((b53
                + b5) * ((b53 + b5) * b61))) * b61)) * ((b11 * b43) * (b61 + b17)))
```

Figure 5.9: Visualization of the spectral index obtained to separate *road* from *bare soil* in Thetford Mines dataset. Mean accuracy: 0.9930.



```
((b41 * (b68 * (b41 * b68))) * (b41 * b41)) % (((b28 - b51) % b81) % (b41 % (((((b28 - b51) *
(((((((((b28 - ((b28 - b51) % b41)) % (b41 * b41)) * ((b41 * b41) % (b41 * b41))) + b49) + b49) %
b68) % b41) % b41) % (b41 * b41)) % ((b41 * (b41 % b41)) * b41))) + b49) % (b41 * b41)) % ((b41 * b41)
% b41))))
```

Figure 5.10: Visualization of the spectral index obtained to separate *red roof* from *grey roof* in Thetford Mines dataset. Mean accuracy: 0.4663.

that minimizes intra-class distance and maximizes inter-class distance. GP can, in theory, come up with the LDA solution, while it is not possible the other way around, so the key aspect in this case is the search algorithm. For some hard-to-separate pairs of classes, the variability of the indices learned by GP was high (approximately 5% in classification accuracy), and the evolution curves showed that the fitness score of the executions did not improve from some generation on, which is a possible sign that the algorithm could have reach local optima.

The results presented in Table 5.4 also showed no statistically significant difference between the proposed method and its variation considering an adjustment by means of a validation set to avoid overfitting. These results, along with the fact that the proposed method was superior with respect to almost all the baselines, suggest that the method had a good generalization capacity by itself.

An interesting pattern with the AVIRIS bands was evidenced in Figure 5.4a for the Salinas dataset. The specification of the AVIRIS sensor[4] warns about three water absorption regions of the spectrum, which are normally ignored, due to its lack of information. The results obtained with the experiments of binary classification shows that GP attributed a low relevance to the bands that belong to those regions, by not selecting them in most of the indices.

[4] `https://aviris.jpl.nasa.gov/html/aviris.spectrum.html`. Last accessed on February, 2017.

Chapter 6

Multi-class classification scenario

Recalling that the classification of two or more classes is considered unnatural, the idea of indices specialized for two classes is preserved for the multi-class scenario. For this, the indices learned in the GP phase are combined to build a multi-class classification system. This chapter describes in detail two fusion methods that were proposed, as well as the comparison of their performances with respect to a set of baselines, equivalent to those used in the binary scenario.

6.1 Pipelines

The two proposed fusion methods are very straightforward: the first one simply builds an *One vs. One* (OVO) ensemble strategy, and the second concatenates the indices to build multi-dimensional feature vectors. This section describes both methods and their respective pipelines.

6.1.1 One-vs-One Classifier Fusion (GP-OVO)

Figure 6.1 illustrates the pipeline of the fusion method based in the OVO strategy [3]. This consists in, for every possible pair of classes, training a binary classifier and, for unlabeled examples, predicting the class with every classifier, and choosing the one with more votes.

The $\frac{n(n-1)}{2}$ learned indices (Step A in Figure 6.1) are used as separate feature spaces where the same number of Nearest Centroid Classifier algorithms (NCC) will be trained (Step B). Next, these classifiers will compose a one-vs-one ensemble strategy for classification (Step C).

6.1.2 Vector-based Fusion (GP-VBF)

For each pair of (n) classes, a specialized spectral index is learned, aiming at maximizing the separation between the training samples of both classes (Step A in Figure 6.2). A total of $\frac{n(n-1)}{2}$ indices learned are used to map each multi-band pixel into a feature space of the same number of dimensions (Step B). The resulting feature vectors are then used in a classification system (Step C).

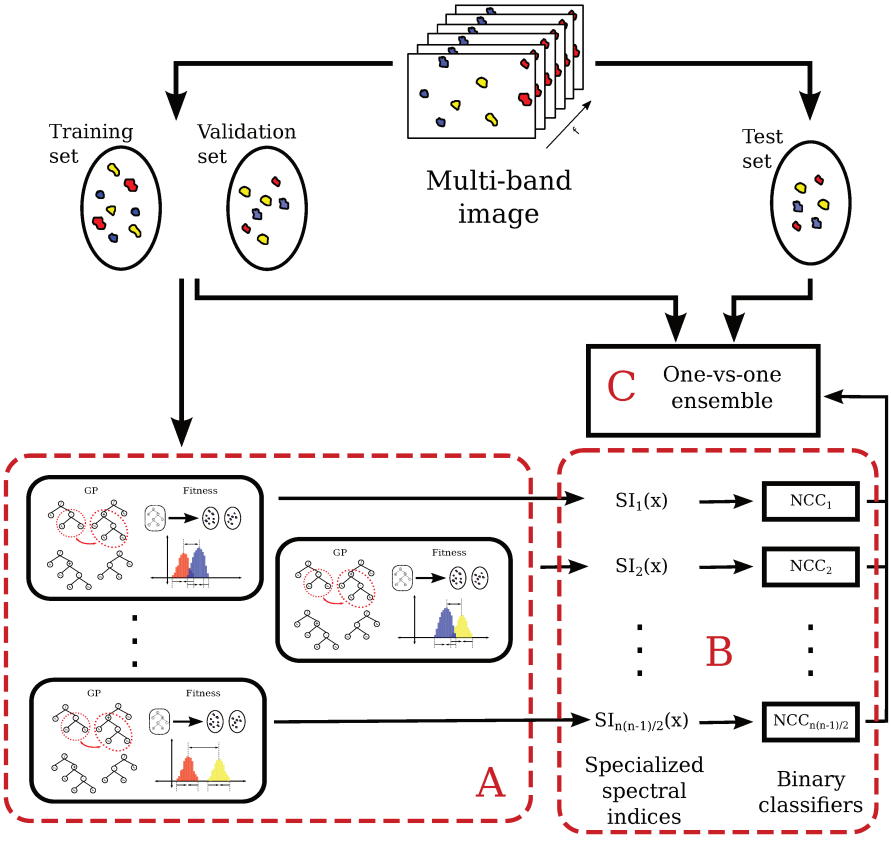


Figure 6.1: Pipeline for the GP-OVO classification system.

6.2 Experimental protocol

The experimental protocol for the multi-class scenario has various similarities with respect to the protocol used in the binary scenario (see Section 5.2) since, basically, exactly the same learned indices were combined through the proposed methods. The datasets used are the ones introduced in Section 5.2.1 and the 5-fold cross validation protocol is maintained.

6.2.1 Parameters

The GP parameters (including the fitness function) are exactly the same to the ones described in Section 5.2.2. For the classification phase, besides the NCC algorithm, the Random Forest (RF) classification algorithm was used. RF relies on the discriminative information provided by the features, so for a one-dimensional projection it made not so much sense to be used in the binary-scenario.

6.2.2 Evaluation protocol

The measure that will be compared is the normalized accuracy (see Section 5.3.1) and the same distribution of the folds in each run of the 5-fold cross validation, described in Section 5.2.3 is maintained.

The same five dimensionality reduction techniques used in the binary scenario were considered as baselines (see Section 5.2.3). Since now the RF classification algorithm is

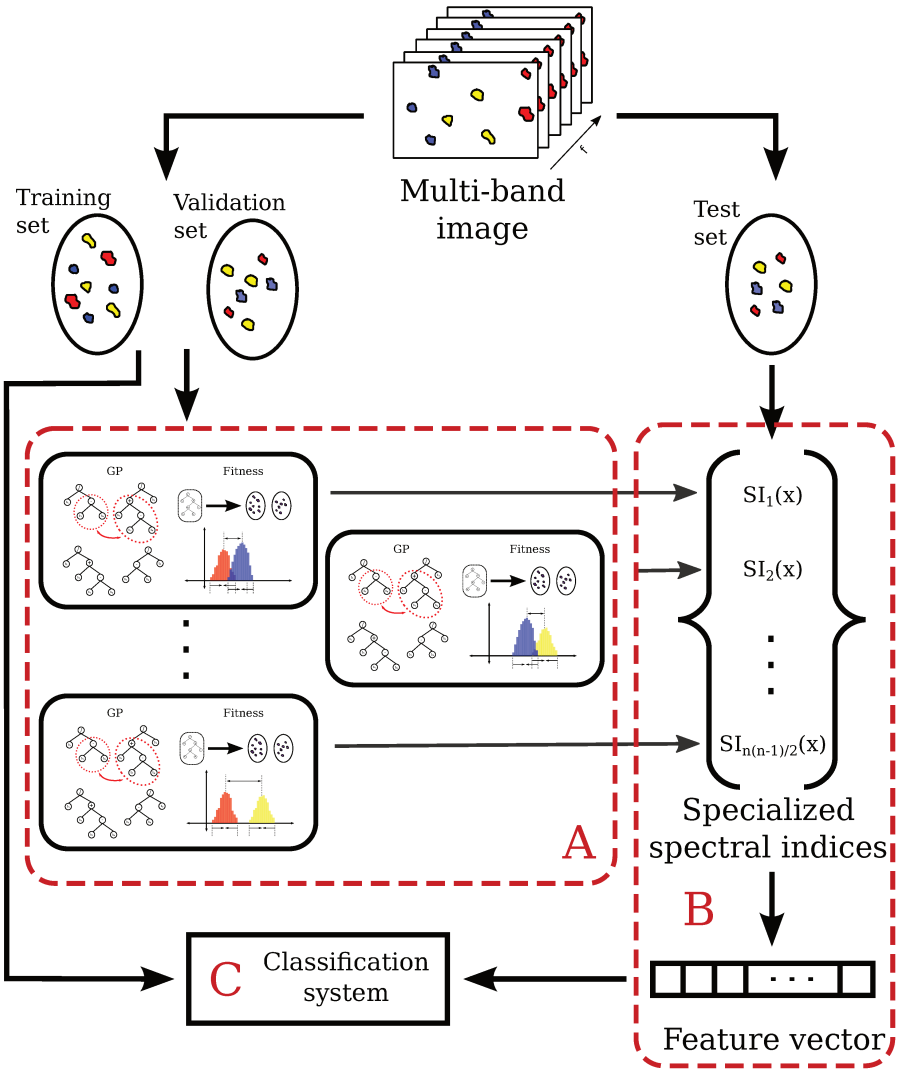


Figure 6.2: Pipeline for the GP-VBF classification system.

considered, besides the NCC, each reduction technique is associated with two classification algorithms, resulting in ten baselines.

6.3 Results

Table 6.1 shows the mean of the normalized accuracies obtained for each fold in the cross validation approach comparing the baselines and the proposed methods. The proposed fusion approaches are very effective in both collections, yielding better or comparable results with respect to the best baselines. A ▲ symbol indicates that the fusion approach was statistically superior to the alternative and • means that both methods are statistically tied. The first and the second symbol refer to the comparison with the GP-OVO+NCC and GP-VBF+RF methods, respectively. No significant difference was evidenced for the two proposed fusion methods.

The statistical significance was estimated with the Friedman test for all the methods, considering each fold as a different problem. For Salinas dataset, the p-value associated is

Table 6.1: Results for multi-class classification (% normalized accuracy). A ▲ symbol indicates that the proposed approaches were statistically superior to the corresponding baseline, **-** means that it was statistically inferior and • means that both methods were statistically tied.

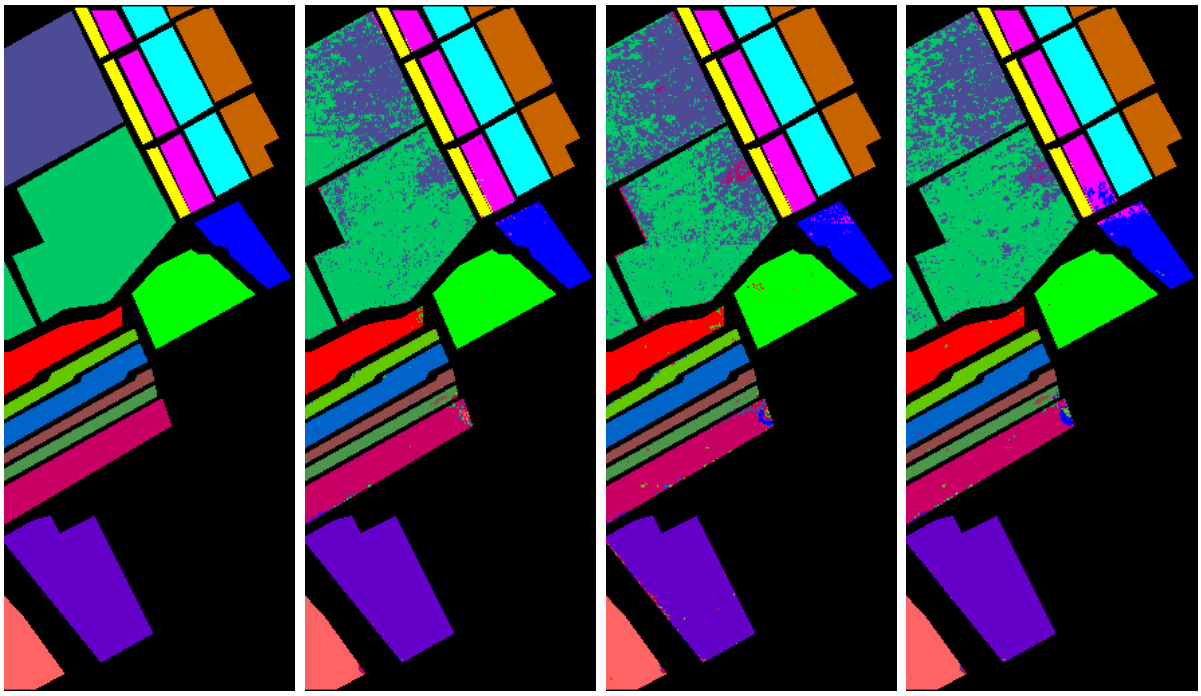
Method	Salinas				T. Mines			
			μ	σ			μ	σ
NS + NCC	▲	▲	78.86	3.63	▲	•	39.17	6.68
NS + RF	•	•	93.70	3.77	•	•	49.95	9.76
UFS + NCC	▲	▲	72.35	6.35	▲	▲	33.39	2.74
UFS + RF	▲	•	89.11	6.89	▲	▲	36.49	4.40
PCA + NCC	▲	▲	78.06	3.78	▲	•	39.11	6.66
PCA + RF	▲	•	90.55	4.78	▲	•	53.22	10.24
LDA + NCC	▲	•	93.77	3.65	•	•	54.14	10.74
LDA + RF	•	•	95.56	1.86	•	•	53.93	8.83
RFS + NCC	▲	▲	79.38	4.08	▲	•	36.50	5.29
RFS + RF	•	•	93.59	3.99	•	•	49.21	9.35
GP-OVO + NCC		•	94.46	2.23		•	54.08	10.40
GP-VBF + RF	•		95.09	3.07	•		54.10	9.63

1.94167×10^{-6} and for Thetford Mines, it is 5.647328×10^{-6}. Likewise the binary scenario, pairwise Wilcoxon Signed Rank tests (significance level of 0.05) with Bonferroni post hoc adjustment were performed in order to compare the proposed methods with the baselines and within themselves.

Figures 6.3 and 6.4 show the thematic maps of the ground truth, and the classification results of the best baselines and the proposed fusion schemes for the Salinas and Thetford Mines datasets, respectively. It can be observed that the results related to the proposed methods are quite similar to the ones of the best baselines.

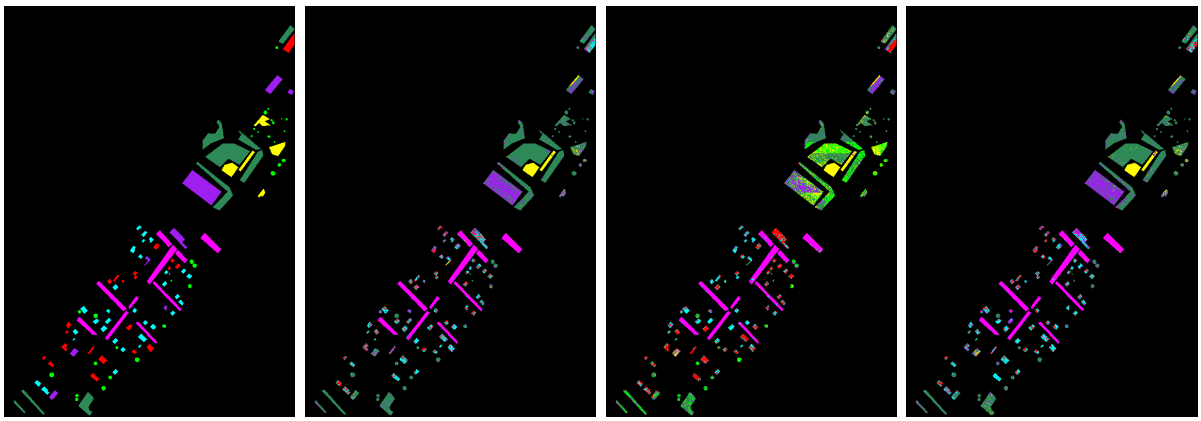
Tables 6.2, 6.3 and 6.4 show the confusion matrices of the first fold, for the two proposed methods and the best baseline in Salinas dataset. There is a clear confusion between the classes *Grapes_ untrained* and *Vinyard_ untrained*, regardless the method.

Tables 6.5, 6.6 and 6.7 show the confusion matrices of the first fold, for the two proposed methods and the best baseline in Thetford Mines dataset. For this dataset, the only class easily separable from the rest is *road* while the others are highly conflicting, particularly, *trees* and *vegetation*. Considering that, with Salinas (whose classes are species of crop), the method achieved good separation of most of the classes, supporting the known fact that NIR and SWIR are better at discriminating vegetation then LWIR.



(a) Ground truth (b) LDA + RF (c) GP-OVO (d) GP-VBF

Figure 6.3: Thematic map of the ground truth and classification of the two fusion methods and the best baseline for the Salinas dataset.



(a) Ground truth (b) LDA + NCC (c) GP-OVO (d) GP-VBF

Figure 6.4: Thematic map of the ground truth and classification of the two fusion methods and the best baseline for the Thetford Mines dataset.

Table 6.2: Confusion matrix: GP-OVO + NCC (Salinas).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	401	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	745	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	392	0	0	0	0	0	0	2	1	0	0	0	0	0
4	0	0	0	276	2	0	0	0	0	0	0	0	0	0	0	0
5	0	0	17	18	498	0	0	0	0	2	0	0	0	0	0	0
6	0	0	0	0	0	791	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	700	0	0	0	0	0	0	1	0	14
8	0	0	0	0	0	0	0	1924	0	57	0	0	0	0	273	0
9	0	0	0	0	0	0	0	0	1196	38	0	3	1	2	0	0
10	0	0	0	2	0	0	0	0	23	619	0	1	0	0	10	0
11	0	0	0	0	0	0	0	0	0	0	208	5	0	0	0	0
12	0	0	0	0	0	0	0	0	0	2	0	383	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	183	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	214	0	0
15	0	0	0	0	0	0	0	663	0	2	0	0	0	0	788	0
16	0	0	0	0	0	0	3	0	0	0	0	0	0	0	1	357

Table 6.3: Confusion matrix: GP-VBF + RF (Salinas).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	397	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	746	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	268	0	128	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	279	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	2	534	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	2	790	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	715	1	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	1494	0	34	0	0	0	0	727	0
9	0	0	0	0	0	0	0	0	1241	0	0	0	0	0	0	0
10	0	0	67	0	1	0	0	16	14	534	20	0	0	3	1	0
11	0	0	0	0	0	0	0	0	0	4	208	0	0	0	0	2
12	0	0	0	0	0	0	0	0	0	0	0	386	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	181	3	0	0
14	0	0	0	0	0	0	0	8	9	0	0	0	40	157	0	0
15	0	0	13	0	0	0	0	401	0	10	0	0	0	0	1021	9
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	362

Table 6.4: Confusion matrix: LDA + RF (Salinas).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	401	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	745	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	395	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	274	4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	1	18	512	3	0	0	0	0	0	0	0	0	1	0
6	0	0	0	0	0	791	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	715	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	2174	0	35	0	0	0	0	45	0
9	0	0	0	0	0	0	0	0	1229	9	2	0	0	0	0	0
10	0	0	0	0	0	0	0	17	43	593	0	2	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	211	2	0	0	0	0
12	0	0	0	0	0	0	0	0	0	1	0	384	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	2	177	4	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	214	0	0
15	0	0	0	0	0	0	0	1118	0	0	0	0	0	0	335	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	361

Table 6.5: Confusion matrix: GP-OVO + NCC (Thetford Mines).

	1	2	3	4	5	6	7
1	822	0	13	21	2	0	0
2	0	158	0	0	10	22	15
3	0	14	152	170	8	3	0
4	1	5	200	168	13	6	1
5	3	75	215	243	107	37	79
6	1	927	4	8	175	106	212
7	0	36	0	0	28	1	277

Table 6.6: Confusion matrix: GP-VBF + RF (Thetford Mines).

	1	2	3	4	5	6	7
1	844	0	1	2	2	4	6
2	0	0	0	0	12	190	4
3	9	0	118	90	16	115	0
4	8	0	138	117	73	59	0
5	5	1	1	2	272	431	48
6	12	1	8	14	102	1277	20
7	0	0	0	0	35	301	7

Table 6.7: Confusion matrix: LDA + NCC (Thetford Mines).

	1	2	3	4	5	6	7
1	849	0	5	3	0	0	1
2	0	74	0	0	18	112	1
3	0	8	152	172	3	12	0
4	0	12	225	136	9	11	1
5	5	71	224	203	71	95	90
6	1	523	0	1	184	660	64
7	0	70	0	0	23	5	244

6.4 Discussion

The two proposed fusion methods for the multi-class scenario obtained statistically equivalent results between them and, unlike the binary scenario (see Section 5.3), in which LDA (the best baseline) outperformed the proposed method. In this case they were statistically equivalent to LDA. This shows that the proposed methods are competitive with respect to traditional dimensionality reduction techniques.

Unlike Salinas dataset, which presented very similar classification results for each class between the binary and the multi-class scenario, the Thetford Mines dataset presented a significant drop (from more than 0.8 to less than 0.55) in this transition, for all the reported methods with no exception. The confusion matrices (Tables 6.5, 6.6 and 6.7) for this dataset show a clear interference between almost all the classes, that only was manifested when the discrimination among classes was not isolated.

One of the reasons for the low performance in the Thetford Mines dataset might be the range of bands that its sensor covers, since the most frequent seem to be concentrated in the region most contiguous to the SWIR, NIR, and VIS regions, which have shown to be discriminative enough in the Salinas dataset, while they are not present in the Thetford Mines dataset. In particular, for this dataset, the results obtained here are not comparable with the results obtained in the GRSS contest,[1] since it was about fusion techniques where, besides the hyper-spectral scene provided, the contestants used a high-spatial-resolution RGB scene, which did not match entirely to the hyper-spectral data.

[1] `https://www.grss-ieee.org/community/technical-committees/data-fusion/2014-ieee-grss-data-fusion-classification-contest-results/`. Last accessed on February, 2017.

Chapter 7

Tropical biomes discrimination and analysis

This chapter discusses the application of the technique introduced in Chapter 4 to characterize sampled quadrants of tropical vegetation labeled either as forests or savannas.

Three classification scenarios are considered. The first one corresponds to the aforementioned Forest/Savanna discrimination, in which the baselines are widely used. The other two scenarios go further with a more specific classification, in which the baselines normally get saturated, so they are not even considered for analysis. They consist in the discrimination of subcategories of savannas and forests with respect to its physiognomy: Evergreen forests vs. semi-deciduous forests, and savannas vs. woodland savannas.

The dataset used comprises a total of 294 regions (120 forests and 174 savannas) distributed in tropical South America (see Figure 7.1). The geographical coordinates used as ground truth are a modified subset of the inventory data used by Dantas et al. [14].

The proposed method will be used to learn vegetation spectral indices (GPVIs, which stands for Genetic-programming-based vegetation index) and compare its discriminative properties with respect to the most popular vegetation indices developed by specialists (NDVI, EVI, and EVI2), which will be the baselines.

Since it is a real-world problem and phenomena do not always belong completely to one class or another, the classification accuracy is considered but will not be determinant. The analysis presented in this chapter is qualitative, and its objective is to introduce the potential applications of the automatically-learned spectral indices.

7.1 Experimental protocol

Since these are binary classification scenarios, the experimental protocol for this analysis has various similarities with respect to the protocol presented in Section 5.2. Fundamental differences are described below.

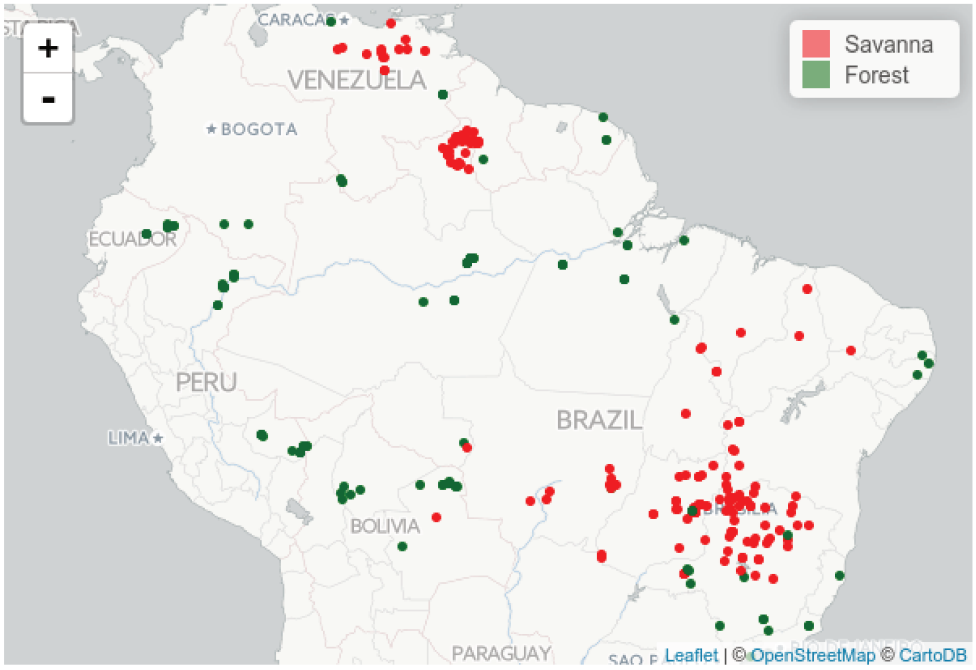


Figure 7.1: Geographical distribution of the regions considered in Tropical South America.

7.1.1 Parameters

The GP parameters (including the fitness function) are the same to the ones described in Section 5.2.2. For the classification phase, instead of the NCC algorithm, the Logistic Regression classification algorithm was used (LR). LR measures the relationship between the categorical dependent variable (class) and one or more independent variables (features) by estimating probabilities using a logistic function. It associates a probability distribution of the classes with each example, from which a confidence value can be deduced. This confidence score may help to identify regions that may not fit completely in none of the classes.

7.1.2 Dataset

The selected subset of regions in Tropical South America correspond to 250m $\times$ 250m quadrants whose centroids were visualized using Google Earth (GE) images from 2016 superimposed with MODIS grid cells of the same size using Series Views [20]. For each of these points, it was visually assessed whether the MODIS pixel at which the points fell into was completely filled with the corresponding labeled vegetation type (i.e., savanna or forest). Pixels that were not entirely filled with the corresponding vegetation type were either replaced by a neighboring pixel fulfilling these criteria or discarded (when a nearby suitable pixel could not be found). The centroid coordinate of each remaining pixel was then obtained and used to extract the satellite data used for classification purposes. The data was acquired using the Google Earth Engine (GEE) API to download the scenes. Two different data products were considered and compared:

1. **Landsat Surface Reflectance sensors (Landsat)[1]:** More specifically, Landsat 5

[1] `https://landsat.usgs.gov/landsat-surface-reflectance-high-level-data-products`. Last

Table 7.1: Specification of spectral bands for both sensors Landsat and MODIS.

	Landsat		MODIS	
Description	**Band**	**Wavelength** (μm)	**Band**	**Wavelength** (μm)
Blue	B1	0.45 - 0.52	B3	0.459 - 0.479
Green	B2	0.52 - 0.60	B4	0.545 - 0.565
Red	B3	0.63 - 0.69	B1	0.620 - 0.670
NIR	B4	0.76 - 0.90	B2	0.841 - 0.876
NIR 2	-	-	B5	1.230 - 1.250
SWIR	B5	1.55 - 1.75	B6	1.628 - 1.652
SWIR 2	B7	2.08 - 2.35	B7	2.105 - 2.155

Thematic Mapper (TM) and Landsat 7 Enhanced Thematic Mapper Plus (ETM+). The data from both sensors was merged, due to lack of availability to build time series. Statistical tests and qualitative inspection of the data were performed in order to assure that there is no significant variability between the data of both sensors, and therefore to confirm it would be safe to merge it.

2. **Moderate Resolution Imaging Spectroradiometer (MODIS)[2]:** The data from the satellites Terra and Aqua was merged too. According to the specifications,[3] the differences between both satellites do not involve the nature of data collected in the bands considered for this analysis.

The data was downloaded at the same spatial resolution for both sensors: 250 meters, so each region was covered by exactly one pixel. Although the sensors take daily snapshots, on May 31, 2003 the Scan Line Corrector of Landsat 7 failed, so the presence of gaps was considerable, preventing the construction of time series so, in order to reduce the impact of those gaps, the data was aggregated per month using the median, and months without data (due to huge gaps) were linearly interpolated. The time interval in which data was collected correspond to 17 years, from January, 2000 to August, 2016.

Regarding spectral resolution, Table 7.1 shows the information of the bands considered for both sensors, Landsat[4] and MODIS[5], associating the channel name with each band code and specifying the wavelength range covered by the band. For this analysis information of the of the visible (Blue, Green, and Red regions), near infrared (NIR), and shortwave infrared (SWIR) regions were used. Landsat TM products normally provide information of the thermal infrared region (LWIR) too (corresponding to band number 6) however, these SR products do not include it. As it can be seen in Table 7.1, the MODIS product provides an additional NIR channel.

accessed on February, 2017.

[2] `https://modis.gsfc.nasa.gov/about/`. Last accessed on February, 2017.

[3] `https://nsidc.org/data/modis/terra_aqua_differences`. Last accessed on February, 2017.

[4] `https://landsat.usgs.gov/landsat-5-history`. Last accessed on February, 2017.

[5] `https://modis.gsfc.nasa.gov/about/specifications.php`. Last accessed on February, 2017.

7.1.3 Evaluation protocol

As well as the experiments from Chapters 5 and 6, the comparison measure for classification is the normalized accuracy (see Section 5.3.1). The 5-fold cross-validation protocol is used too but, different from the approaches used in Chapters 5 and 6, the results reported will correspond to unseen data in this cross validation protocol.

Each of the first five years of the series (i.e., 2000 - 2004) is used as a a fold. As described in Section 5.2.3, a validation set is used too. Each fold run returns two spectral indices: one considering validation and another without considering it. For the sake of simplicity, from the ten indices learned by this protocol, only the one with the best normalized accuracy is reported and further analysed. The reported accuracies correspond to the classification of the data from the years that were not involved in the cross validation protocol (i.e., 2005 - 2016).

Baselines

The indices learned through the proposed method (GPVI) are compared to three of the most popular vegetation indices used to characterise forests and savannas [66].

1. **Normalized difference vegetation index (NDVI):** Normalized difference between the red and near infrared channels to measure vegetation greenness (see Equation 2.1) [63].

2. **Enhanced vegetation index (EVI):** Enhancement of sensitivity considering the blue channel [29].

$$EVI = \frac{2.5 \times (NIR - Red)}{NIR + 6.0 \times Red - 7.5 \times Blue + 1} \tag{7.1}$$

3. **Two-band enhanced vegetation index (EVI2):** Calibration of EVI so it must not depend on the blue band, which is not included in products of old sensors and normally presents noise problems [37].

$$EVI2 = \frac{2.5 \times (NIR - Red)}{NIR + 2.4 \times Red + 1} \tag{7.2}$$

7.2 Results

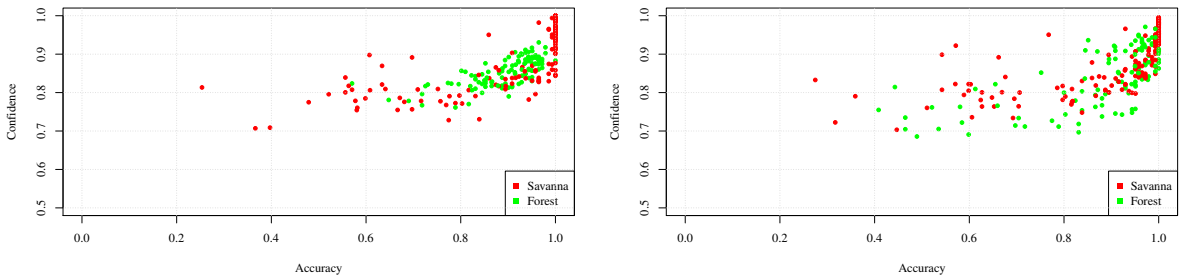
A total of six experiments are executed, corresponding to the three classification scenarios described above with the two sensors. For each experiment, the classification accuracy, the time series associated with the indices, and the relevance of the spectral bands are reported.

7.2.1 Classification accuracy

Table 7.2 shows the accuracies obtained in the experiment corresponding to the Forest/Savanna discrimination for both sensors. The columns **TFR** and **TSR** correspond to

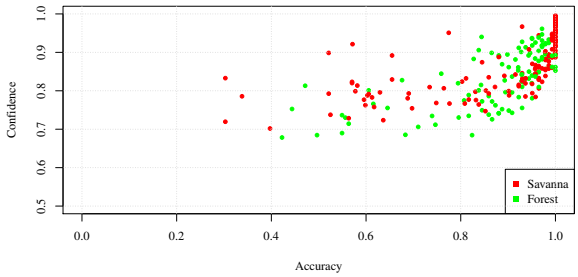
Table 7.2: Normalized accuracies and assertion rates per class for Forest/Savanna discrimination. The columns **TFR** and **TSR** correspond to the rate of correctly classified forests and savannas, respectively.

	Landsat			MODIS		
Sensor	**TFR**	**TSR**	**Acc.**	**TFR**	**TSR**	**Acc.**
NDVI	89.28	90.49	89.89	78.71	86.44	82.58
EVI	87.75	89.88	88.81	80.90	89.10	85.00
EVI2	87.34	89.48	88.41	78.24	87.93	83.09
GPVI	96.38	95.46	95.92	88.30	91.09	89.69

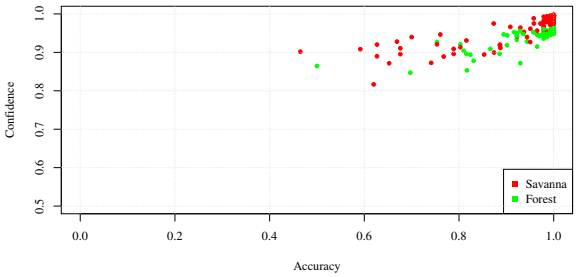


(a) NDVI

(b) EVI



(c) EVI2

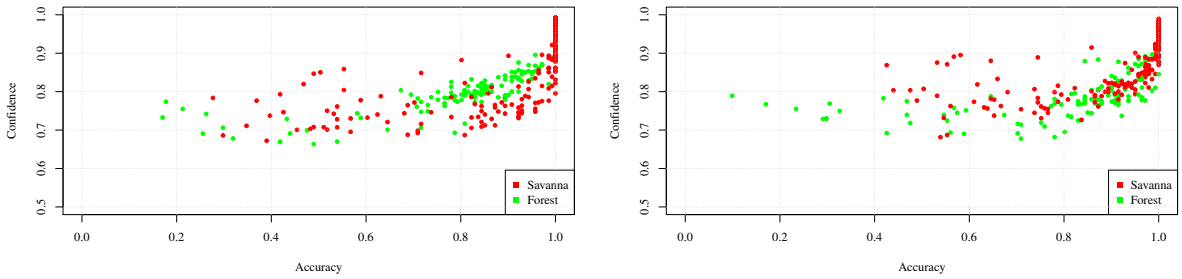


(d) GPVI

Figure 7.2: Confidence vs. accuracy in each region for Landsat sensor.

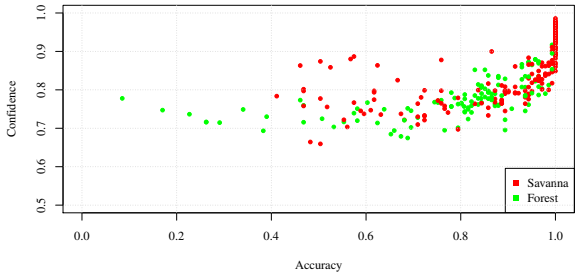
the rate of correctly classified forests and savannas, respectively. The baselines presented similar performances, which are clearly outperformed by GPVI by about 6% in the worst case with Landsat and 4% in the worst case with MODIS. In general, the performance for Landsat was better than the one for MODIS, and the rate of correctly classified savannas was higher than the rate of correctly classified forests.

For the Forest/Savanna discrimination scenario, an analysis regarding the confidence values provided by the logistic regression classification algorithm was performed. Figures 7.2 and 7.3 show scatter plots of mean accuracy vs. mean confidence score obtained in each region, for Landsat and MODIS respectively. It can be seen that GPVI presents, in average, higher confidence values than the baselines, which means less confusion in the classification task. The results for the MODIS sensor present a particular tendency of the algorithm to misclassify forest quadrants. Those regions presenting low accuracy and high confidence in various experiments can be suspicious of having been mislabeled by

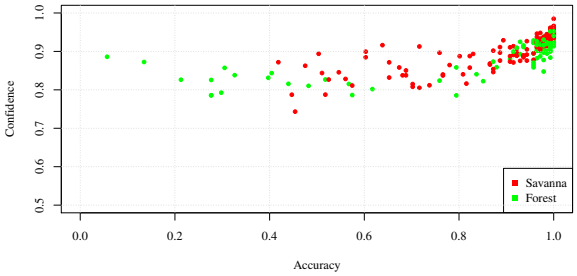


(a) NDVI

(b) EVI



(c) EVI2



(d) GPVI

Figure 7.3: Confidence vs. accuracy in each region for MODIS sensor.

Table 7.3: Normalized accuracies and assertion rates per class for Evergreen forest/Semi-deciduous forest discrimination. The columns **TEFR** and **TSDFR** correspond to the rate of correctly classified evergreen forests and semi-deciduous forests, respectively.

	Landsat			**MODIS**		
Sensor	**TEFR**	**TSDFR**	**Acc.**	**TEFR**	**TSDFR**	**Acc.**
NDVI	67.04	39.95	53.50	75.17	61.61	68.39
EVI	64.50	61.38	62.94	72.19	63.39	67.79
EVI2	63.93	56.61	60.27	69.95	61.08	65.52
GPVI	78.40	61.90	70.15	84.44	78.72	81.58

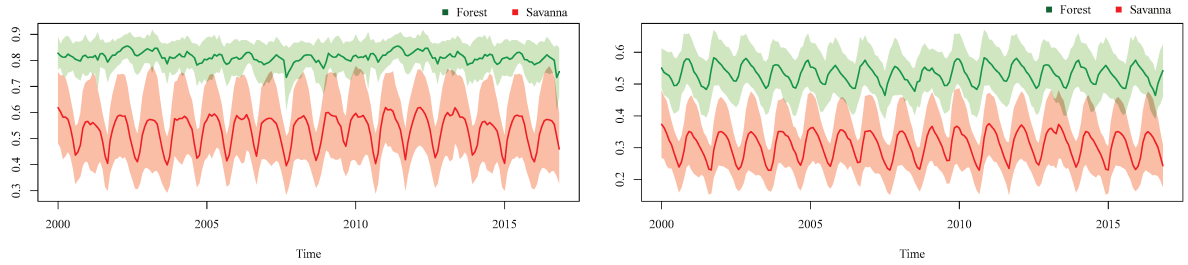
the specialists.

Table 7.3 shows the accuracies obtained in the experiment corresponding to the Evergreen forest/Semi-deciduous forest discrimination for both sensors. The columns **TEFR** and **TSDFR** correspond to the rate of correctly classified evergreen forests and semi-deciduous forests, respectively. The baselines presented similar performances, except for NDVI in Landsat, with a much inferior performance than the other baselines. In general the baselines behaved similarly for both sensors, and were outperformed by GPVI by about 7% in the worst case for Landsat, and by about 13% in the worst case for MODIS. For this scenario, the performance of GPVI in MODIS was much superior than Landsat, and Evergreen forest obtained a higher rate of correctly classified regions.

Table 7.4 shows the accuracies obtained in the experiment corresponding to the Savanna/Woodland savanna forest discrimination for both sensors. The columns **TSR** and **TWSR** correspond to the rate of correctly classified savannas and woodland savannas,

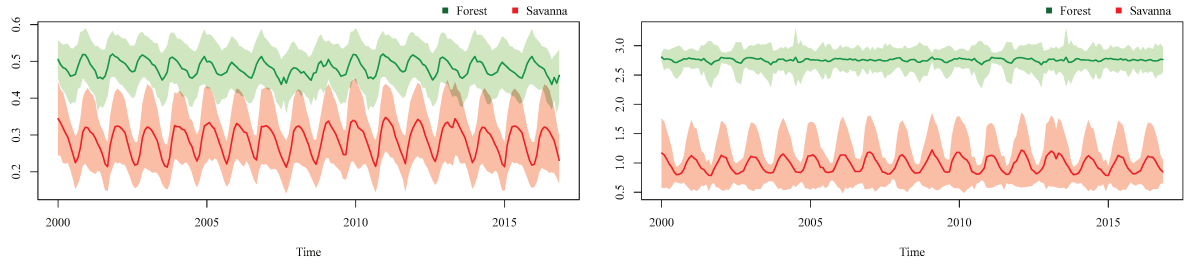
Table 7.4: Normalized accuracies and assertion rates per class for Savanna/Woodland savanna discrimination. The columns **TSR** and **TWSR** correspond to the rate of correctly classified savannas and woodland savannas, respectively.

	Landsat			MODIS		
Sensor	**TSR**	**TWSR**	**Acc.**	**TSR**	**TWSR**	**Acc.**
NDVI	57.92	73.70	65.81	59.89	70.99	65.44
EVI	58.37	65.61	61.99	62.91	62.28	62.60
EVI2	58.36	66.17	62.27	62.68	62.48	62.58
GPVI	71.06	74.89	72.98	58.62	76.34	67.48



(a) NDVI

(b) EVI



(c) EVI2

(d) GPVI

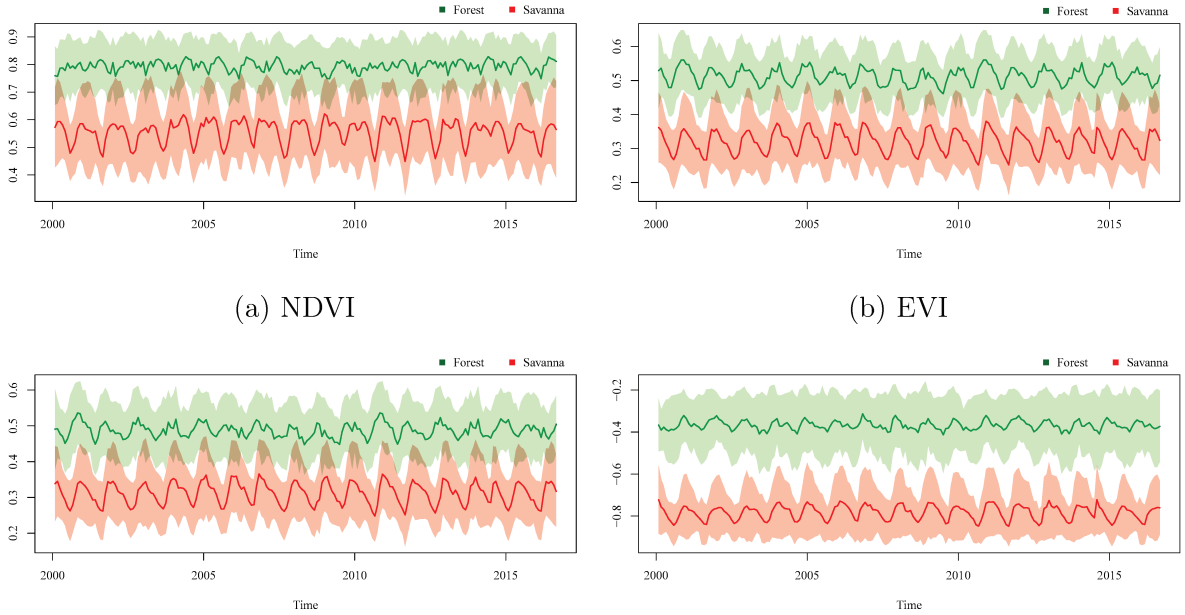
Figure 7.4: Time series for Forest/Savanna discrimination of the Landsat sensor.

respectively. The baselines presented similar performances for both sensors, and were outperformed by GPVI by about 7% in the worst case for Landsat, and by about 2% in the worst case for MODIS. For this scenario, the performance of GPVI in Landsat was superior than MODIS, and Woodland savanna obtained a higher rate of correctly classified regions only for Landsat; for MODIS, both classes obtained similar rates.

7.2.2 Time series comparison

The times series for GPVI and the baselines were plotted, taking the mean of the pixel values of all the regions belonging to the same class in each time stamp. The instant standard deviations of each class were plotted around the corresponding curve.

Figures 7.4 and 7.5 show the time series for Landsat and MODIS respectively in the Forest/Savanna scenario. Landsat presented a greater separation between the series than MODIS, for GPVI and the baselines. It is clear how GPVI makes easier to discriminate



(a) NDVI

(b) EVI

(c) EVI2

(d) GPVI

Figure 7.5: Time series for Forest/Savanna discrimination of the MODIS sensor.

the values of pixels for the two classes, regardless the instant of time.

The next scenario corresponds to Evergreen forest/Semi-deciduous forest. Figures 7.6 and 7.7 show the time series for Landsat and MODIS, respectively. Although the separation of GPVI is not as good as it was in the first scenario, it can be seen that it got much less saturated than the baselines. Here, the difference of behavior between the sensors is not so evident.

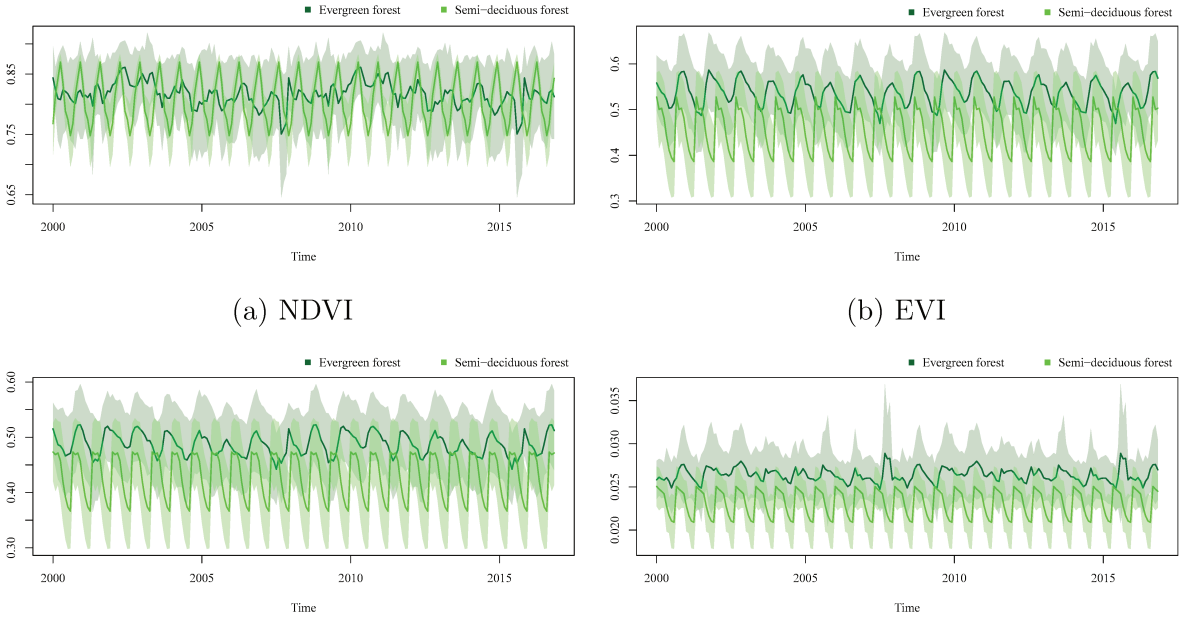
Figures 7.8 and 7.9 show the time series for Landsat and MODIS respectively for the Savanna/Woodland savanna scenario. Here, the saturation of all indices is more serious than the the others, showing that discriminating between these two classes is a hard problem, at least with the spectral information available.

7.2.3 Band frequency within the indices

A similar exercise to the one presented in Section 5.3.2 is done for the three scenarios. The number of times a band was present in the ten vegetation indices learned in the 5-fold cross validation protocol are counted and visualized in histograms. Figures 7.10, 7.11 and 7.12 show the number of occurrences of the bands with both sensors for the three scenarios.

To discriminate forests from savannas, there was an agreement between the two sensors, that the most frequent bands were NIR and SWIR, unlike the baselines, which consider the Red channel instead of the SWIR one. The sensors also agreed that the Blue channel (used in EVI) is more frequent than the Red one. The NIR2 channel was as frequent as the Blue one for MODIS, however, apparently the lack of this band did not give any disadvantage to the Landsat sensor to discriminate correctly (see Table 7.2).

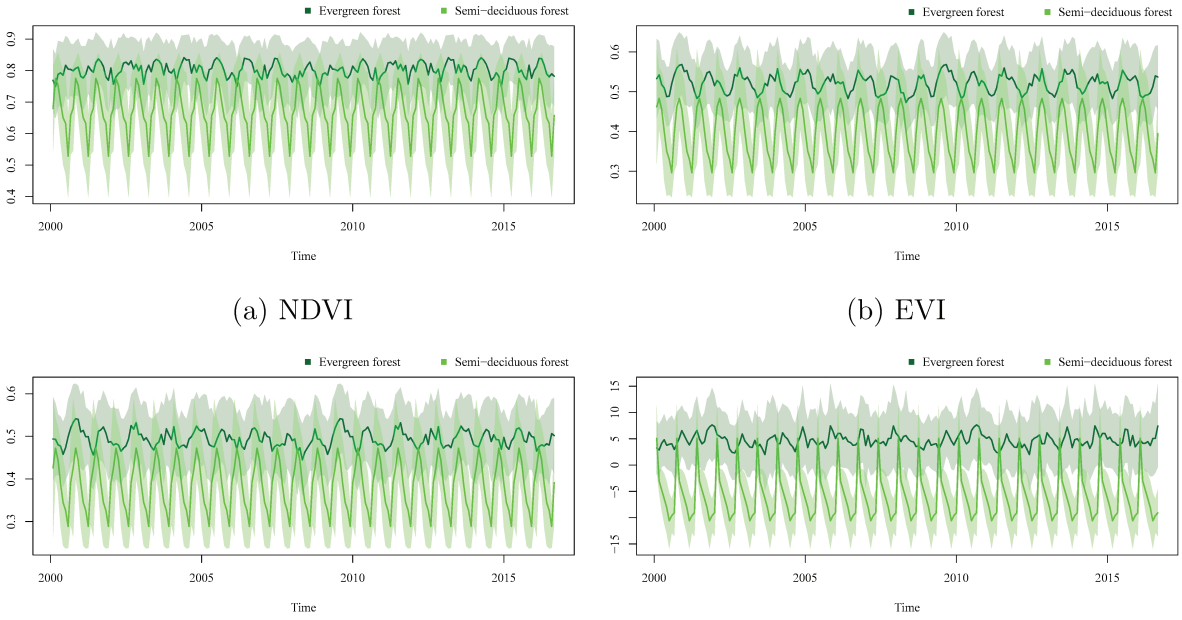
The other two scenarios did not obtain a satisfactory agreement between the sensors to determine the frequency of the bands in the indices. The separation by physiognomy



(a) NDVI (b) EVI

(c) EVI2 (d) GPVI

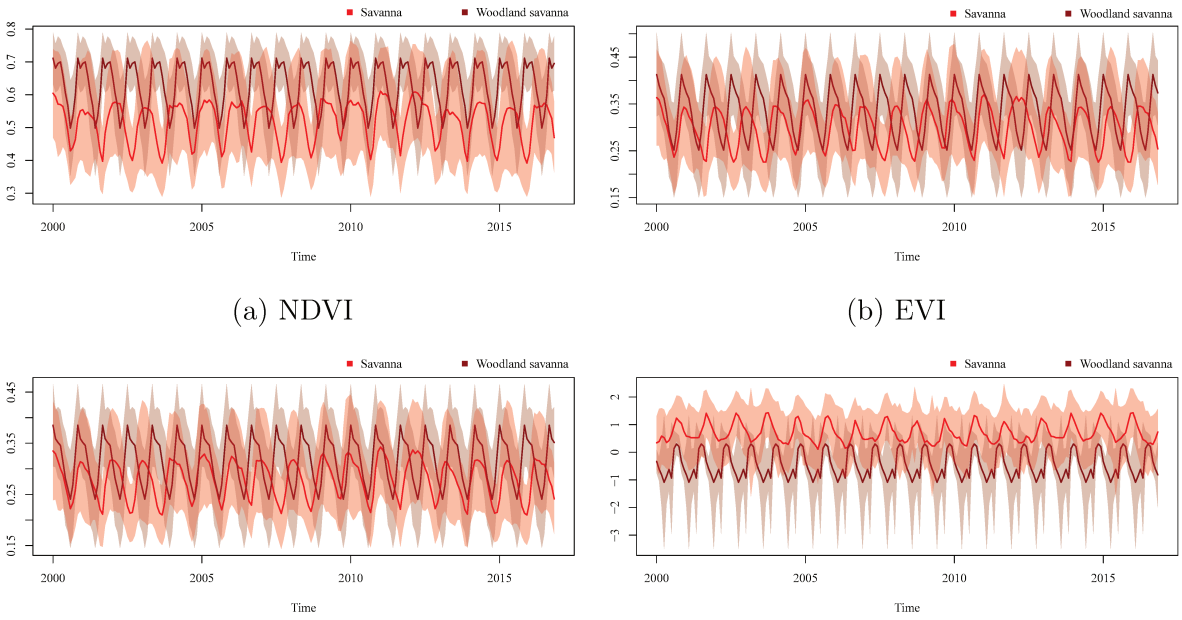
Figure 7.6: Time series for Evergreen forest/Semi-deciduous forest discrimination of the Landsat sensor.



(a) NDVI (b) EVI

(c) EVI2 (d) GPVI

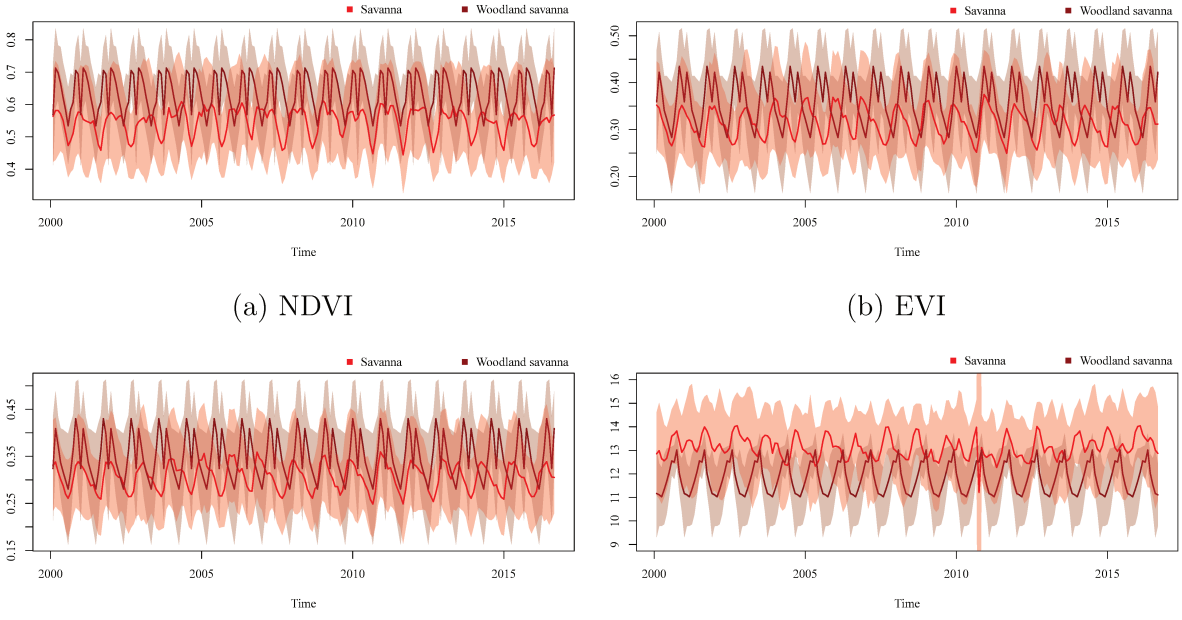
Figure 7.7: Time series for Evergreen forest/Semi-deciduous forest discrimination of the MODIS sensor.



(a) NDVI (b) EVI

(c) EVI2 (d) GPVI

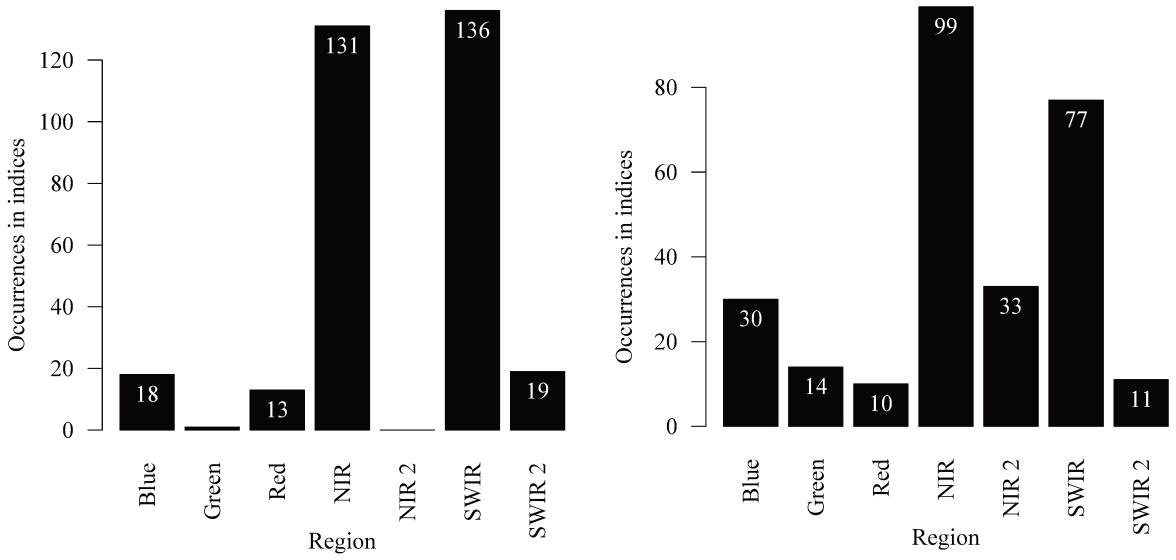
Figure 7.8: Time series for Savanna/Woodland savanna forest discrimination of the Landsat sensor.



(a) NDVI (b) EVI

(c) EVI2 (d) GPVI

Figure 7.9: Time series for Savanna/Woodland savanna forest discrimination of the MODIS sensor.



(a) Landsat (b) MODIS

Figure 7.10: Distribution of the frequencies of the spectral bands of the learned indices for Forest/Savanna discrimination.

between the same biome is clearly a harder problem.

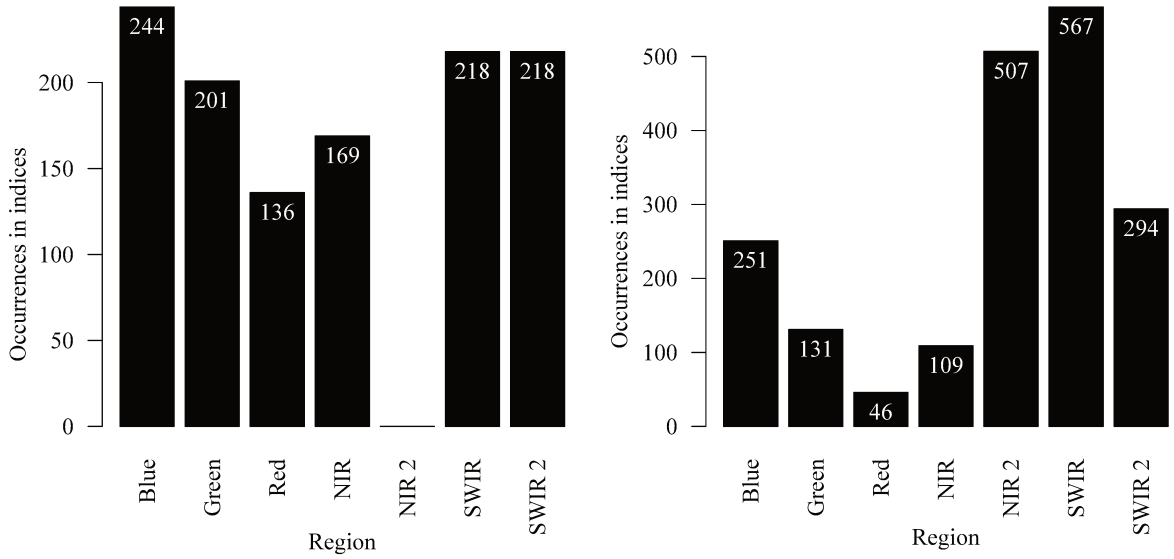
In particular, the Evergreen forest/Semi-deciduous forest scenario presents a clear relevance of the NIR2 and SWIR bands for the MODIS sensor, while for Landsat that difference is not so prominent (see Figure 7.11). The only agreement between the sensors is that the Red and NIR bands, the ones used in the baselines, are the less frequent. Looking at Table 7.3, the performance of GPVI for MODIS was about 11% superior to Landsat, which may suggest that the NIR2 band could have a high relevance at characterizing these classes.

The Savanna/Woodland savanna scenario which, according to the classification results (see Table 7.4), was the most difficult pair of classes to discriminate, presented no agreement between the sensors (see Figure 7.12). For example, the SWIR band was the most frequent for Landsat and the less frequent for MODIS, and the Red band, the most frequent for MODIS, was not so frequent for Landsat.

7.2.4 GPVIs

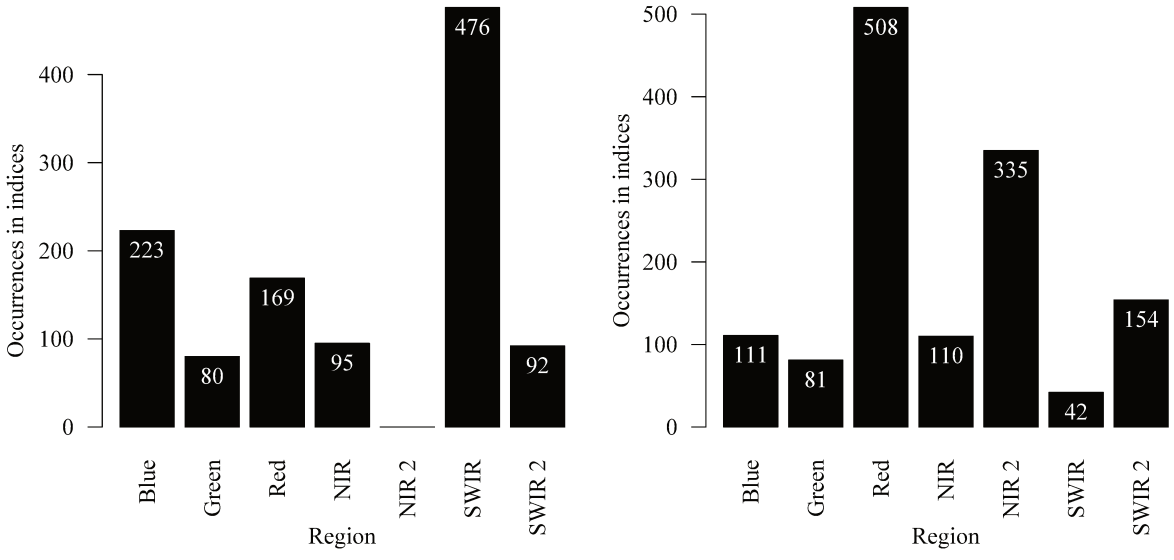
In this section, the six vegetation indices learned with the GP framework are presented. The first one is the index for Forest/Savanna discrimination with Landsat. It can be seen that the most recurrent bands are NIR and SWIR, and the operation **SWIR - NIR** and **SWIR2 % NIR** appear various times.

```
srt((Blue * rlog((SWIR - (NIR * SWIR2 - (SWIR - NIR)))))) - ((NIR - (SWIR - (srt((srt((SWIR - ((Blue *
   SWIR2) - (SWIR - NIR))) * Red) * rlog(rlog(rlog(srt(rlog(NIR) * ((Red * ((NIR - (rlog(rlog(Blue *
SWIR2 - (SWIR2 % NIR - ((SWIR - (srt(SWIR) - Blue * SWIR2)) * rlog(SWIR2 % NIR))))) - Blue)) * Red)) -
     (SWIR - (Blue * SWIR2 - (SWIR - NIR)))))))))) * rlog(SWIR2 % NIR)) - ((NIR - (SWIR - NIR)) %
                                    srt(SWIR))))) % srt(SWIR))
```



(a) Landsat (b) MODIS

Figure 7.11: Distribution of the frequencies of the spectral bands of the learned indices for Evergreen forest/Semi-deciduous forest discrimination.



(a) Landsat (b) MODIS

Figure 7.12: Distribution of the frequencies of the spectral bands of the learned indices for Savanna/Woodland savanna discrimination.

Forest/Savanna discrimination with MODIS, the NIR and SWIR bands are the most recurrent as well, and the operation **NIR2 % NIR** and **NIR % SWIR** appear various times.

```
srt(srt(NIR2 % NIR)) % ((NIR2 - srt(Green - (Green - (NIR2 + SWIR + (((Green % SWIR2) * (NIR % (NIR2 %
NIR))) * ((((((NIR % SWIR + Blue) + Blue) * rlog(Red % NIR)) - ((srt(srt(NIR % SWIR)) + Blue) + Blue))
  - (((NIR2 % NIR) * (Red % NIR)) * SWIR)) * ((((NIR % (NIR2 % NIR)) % SWIR) + Blue) + Blue))))))) -
                                    (((NIR % SWIR) + Blue) + Blue))
```

For the classification of different physiognomies within the same biome, the complexity of the learned indices increases significantly. Besides a greater quantity of operations, constants appear in the formulas. The next is the index for Evergreen forest/Semi-deciduous forest with Landsat. The operations **Blue % Red** and **SWIR2 % (Blue + SWIR)** appear various times.

```
 rlog(NIR * 529.24) % ((((SWIR2 * 699.14) % (Blue + SWIR)) - ((srt(Blue % Red) * (srt(SWIR) + ((srt(Blue % Red) * (srt(SWIR) + SWIR2 *
724.51)) * rlog(srt(rlog(((((SWIR2 * 724.51) % (Blue + SWIR)) - NIR * 599.64) - ((srt((SWIR2 - ((srt(Blue + Green) - (385.56 % (725.7 -
 NIR))) * rlog(Green))) * 313.30) - (385.56 % (755.42 - NIR))) - srt(SWIR2))) + (SWIR2 * ((SWIR2 - (SWIR2 - NIR)) * 239.55))) - SWIR2 -
   NIR))))) * rlog(srt(rlog(((NIR * 599.64 - Green) - ((srt(rlog(srt(rlog(SWIR2 - ((SWIR2 * 724.51) % (Blue + SWIR)))))) - (385.56 %
 ((rlog(srt(rlog(NIR * 599.64))) % (SWIR2 - (SWIR2 - ((((SWIR2 * 724.51) % (Blue + SWIR)) - NIR * 599.64) - (SWIR2 * (srt(Blue % Red) *
 313.3)))))) - Green))) - SWIR)) + (SWIR2 * (((((SWIR2 * 724.51) % (Blue + SWIR)) - (SWIR2 * 724.51)) - (srt(SWIR + Green) - (456.53 %
  ((SWIR2 * (SWIR + Green)) - (SWIR2 - NIR - ((srt(Blue % Red) * srt(SWIR) + ((rlog(SWIR2 - ((((srt(SWIR + Green) - (385.56 % (725.7 -
     NIR))) % (Blue + SWIR)) - NIR * 599.64) - (srt(Blue % Red) * (srt(SWIR) + ((srt(Blue % Red) * (srt(SWIR) + SWIR2 * 724.51)) *
rlog((rlog(srt(rlog(NIR * 599.64))) % (SWIR2 - (SWIR2 - ((((SWIR2 * 724.51) % (Blue + SWIR)) - NIR * 599.64) - (SWIR2 * srt(Blue % Red)
    * 313.3))))) - Green)))))) * (385.56 + (SWIR2 * 724.51))) * rlog(srt(rlog(srt(SWIR + Green) + SWIR2 - NIR) - SWIR2 - NIR))))) *
rlog(SWIR2 - NIR * 599.64 - NIR * 599.64 - NIR * 599.64)) - ((SWIR2 - SWIR2 - SWIR) * 313.3))))) + (SWIR2 * SWIR2 * 711.52)) * 313.3)))
                                        - (385.56 % (725.7 - NIR))))) - Green) - SWIR
```

The index for Evergreen forest/Semi-deciduous forest with MODIS is shown below. **SWIR** and **NIR2** are the more recurrent.

```
       ((srt((Green % Red) - 146.45 - Green) * (SWIR % (Blue + srt(NIR2) + SWIR2))) * (((rlog(Blue) - srt(SWIR2) - rlog(SWIR)) *
srt(srt(rlog(Blue) + srt(SWIR2) - rlog(SWIR) + srt(SWIR2) - rlog(SWIR)) - (((270.41 % SWIR2) % (377.91 % Red)) + ((((SWIR % NIR) * (SWIR
   % ((Blue + (523.85 % srt(Red % NIR2))) + SWIR2))) + srt(Green + srt(SWIR))) - rlog(SWIR))))) + rlog((NIR2 - Blue) * (((rlog(Blue) -
  (((NIR % (srt(((rlog(NIR) + (523.85 % Red)) + 523.85 % Red) + (NIR2 % SWIR2 - rlog(NIR))) - (((rlog(Green) - (NIR - (srt(rlog(SWIR) -
     146.45 - Green) * (SWIR % (Blue + srt(NIR2) + SWIR2))))) + srt(SWIR2 * 601.47)) + srt(srt((SWIR % (146.45 - Green)) + ((SWIR %
(((srt(NIR2 % SWIR2) * (187.95 + SWIR - (Green % Red))) + ((rlog(Blue) - srt((NIR2 % SWIR2) - 146.45 - Green)) * SWIR2)) - rlog(Blue) -
     Green)) * SWIR2)))))) * (SWIR - ((146.45 + (rlog(270.41 % SWIR2) + NIR2 - NIR2 % SWIR2)) + 509.05))) - (rlog(NIR) + srt(NIR2) +
     rlog(srt(SWIR2))))) * SWIR2) * srt(NIR2 - NIR2 % SWIR2))))) + (srt((rlog(Blue) + rlog(SWIR % srt(Red % NIR2))) + (srt(SWIR2) -
 rlog(SWIR) + (srt((srt(SWIR2) - rlog(SWIR)) % Red - 146.45 - rlog(NIR) + srt(SWIR2)) - rlog(SWIR)))) + (((srt((rlog(Blue) + rlog(SWIR %
srt(Red % NIR2))) + (srt(SWIR2) - rlog(SWIR) + (srt(((srt(SWIR2) - rlog(SWIR)) % Red) - 146.45 - rlog(NIR) + srt(SWIR2)) - rlog(SWIR))))
  - (NIR - (srt((Green % (srt(SWIR2) - rlog(SWIR))) - 146.45 + Green) * (SWIR % (Blue + srt(NIR2) + SWIR2))))) + srt(SWIR2 * 319.95)) -
                                    (rlog(146.45 + rlog(NIR) + Green) * (NIR2 - (NIR2 % SWIR2))))
```

The index for Savanna/Woodland savanna with Landsat is shown below. The operation **1 % SWIR** is recurrent.

```
((rlog(rlog(rlog(rlog(rlog(srt(SWIR)))))) * rlog(rlog((SWIR2 % (rlog(SWIR) * (rlog(SWIR2 * (rlog(rlog(srt((srt(SWIR) % (rlog(SWIR2 *
             rlog(737.76 % SWIR) * (486.8 % SWIR)) + (486.8 % SWIR))) + rlog(rlog(Blue) % rlog(rlog((rlog(srt(SWIR))) +
 rlog(rlog(rlog(rlog(srt(SWIR)))))))))) + rlog(rlog(rlog(SWIR2 * (rlog(486.8 % SWIR) * (634.11 % SWIR)))))) * (691.09 % SWIR)))))) -
                    SWIR))) + ((srt((rlog((557.12 % SWIR) % Blue) % (srt(rlog(737.76 % SWIR)) + (515.01 % ((SWIR +
(rlog(rlog(rlog(srt((rlog(srt(rlog(rlog(srt((SWIR2 % srt(SWIR)) - SWIR)))))) + (SWIR2 % (rlog(rlog(rlog(srt((SWIR2 % ((rlog(SWIR)) *
   (rlog((SWIR2 * ((rlog((rlog((srt(SWIR)))))) * (486.8 % SWIR))))))) - SWIR)))) * (SWIR + 462.26 % SWIR)))))))) * rlog(449.83 %
     srt(SWIR)))))) + rlog(SWIR2 % rlog(SWIR2 % (SWIR2 % rlog(Blue))))) + rlog(rlog(rlog(srt((SWIR2 % (rlog(SWIR) * rlog(SWIR2 *
  (rlog(rlog(srt(SWIR))) * (486.8 % SWIR))))) - SWIR))))) * rlog(((rlog(rlog(SWIR)) + (SWIR + ((SWIR + rlog(rlog(rlog(srt((SWIR2 %
(rlog(SWIR) * rlog(SWIR2 * rlog(rlog(srt(SWIR))) * (486.8 % SWIR)))) - SWIR))))) * rlog(557.12 % (SWIR + (462.26 % srt(SWIR)))))) *
  (rlog(rlog(srt(SWIR))) * (SWIR + (462.26 % srt(SWIR))))) + (rlog(srt((((557.12 % SWIR) % Blue) % (rlog(Blue) + 515.01 % SWIR)) +
rlog(SWIR2 % rlog(SWIR2 % (SWIR + SWIR))))) + rlog(rlog(rlog(SWIR2 * ((SWIR2 % (rlog(SWIR2 % rlog(Blue)))) * (737.76 % SWIR)))))) *
                                        rlog(515.01 % srt(SWIR)))))) * srt(SWIR)
```

The index for Savanna/Woodland savanna with MODIS is shown below.

```
(60.07 * Red + SWIR2 * NIR) % (NIR2 % NIR) + rlog(srt((((Red * 460.17 * rlog(rlog(Blue - SWIR))) * (NIR2 + Red * Red - 459.70 * 535.71 -
srt(((NIR2 + (rlog((472.68 * Red + SWIR2 * rlog(Blue)) * (230.3 * Red - ((srt(rlog(Blue)) + SWIR2) * NIR))) * Red)) - (459.70 * 535.71))
 - srt(rlog(Blue))))) + (((((((rlog(696.71 % (((472.68 * Red) + (SWIR2 * rlog(Blue))) * (230.3 * Red - ((472.68 * Red + SWIR2 + SWIR2) *
     NIR)))) + SWIR2) + Blue - SWIR) + ((NIR2 + NIR + SWIR2 * NIR) + rlog(Blue))) + rlog(Blue)) * (472.68 * Red - 459.70 * 535.71 -
 srt(472.68 * rlog(Blue)))) + SWIR2 + SWIR2) * (SWIR2 * (rlog((SWIR2 * ((rlog(Blue) + SWIR2) + rlog(Blue))) * NIR) * (Red - Blue))))) *
  ((((srt(rlog(Blue) + SWIR2) + NIR) + rlog(Blue)) * (((((NIR2 + NIR + SWIR2 * NIR + rlog(Blue)) * ((472.68 * Red - 459.70 * 535.71) -
 srt(SWIR2 * (472.68 * Red - 459.70 * 535.71 - srt((rlog(Blue) * (Red - Blue)) * rlog(Blue)))))) + (rlog(rlog((NIR2 + Red * Red - 459.70
 * 535.71) - srt(((472.68 * ((Red * 460.17 * (NIR2 + Red * Red)) * (NIR2 + Red * Red - 697.86 * 535.71 - SWIR2 * rlog(Blue)))) + SWIR2 *
rlog(Blue)) * (230.3 * Red - (srt(((NIR2 + (rlog(Blue) * Red)) - 459.70 * 535.71) - srt(Red * Red)) * NIR))))) + 130.18 * Red)) * (SWIR2
   * (rlog((SWIR2 * (rlog(Blue) + SWIR2 + rlog(Blue))) * NIR) * (Red - Blue)))) - srt(rlog(Blue) * (rlog(Blue) - NIR2 + NIR - (SWIR2 *
((rlog(Blue) + SWIR2 + 130.18 * Red * SWIR2 * NIR) * ((NIR2 + (rlog(Blue) * (Red - Blue) * WIR2 * 130.18 * Red * SWIR2 * NIR) - 459.70 *
 535.71) - srt(rlog(Blue))) + 472.68 * NIR))))) + SWIR2 * SWIR2 * 130.18 * Red * SWIR2 * NIR) * (SWIR2 * (rlog(Blue) * (Red - Blue))))))
```

7.3 Discussion

The use of spectral indices to analyse phenomena related to living things normally takes the time into account. In the case of biomes discrimination, traditional vegetation indices need to be propagated through the time in order to analyse its behavior and then classify the target objects. Figures 7.4 and 7.5 show that, for the traditional indices, sometimes savannas take high values that could be interpreted as forests, and vice versa, so it is not possible to point out one region in a specific time stamp, measure its index and conclude its class. With GPVI, on the other side, it is possible, showing that there exist complex interactions between the bands that characterize a class regardless the behavior through time of the objects. It is important to point that, since they are learned with many degrees of freedom, the GPVIs are not necessarily bounded like NDVI, for example, whose values are always between -1 and 1.

NDVI, EVI, and EVI2 are classified as *Greenness indices*, which aim to measure vegetation vigor, relying principally on the Red and NIR wavelengths [47]. In the case of forest-savanna discrimination, GPVI, at using the SWIR wavelength, could be classified more as a *Moisture index*. This agrees with several recent studies that found the potential of this kind of indices to detect woody vegetation in specifically these two biomes [6, 40, 41, 65].

As it was expected, the separation by physiognomy was harder, and obtained an inferior performance with respect to Forest/Savanna classification. Something that, for simplicity, was not documented, was the use of the GPVIs learned for Forest/Savanna in the discrimination to classify the different physiognomies, because they got saturated just as the traditional indices. One reason may be attributed to the fact that the fitness function aimed to increase inter-class distance, while decreasing intra-class distance, so the spatial distribution of forests and savannas might have become reduced, decreasing its variety. The formulation of other fitness functions can be done, in order to see if the physiognomy can be deduced naturally from general discrimination.

However, even training the GPVIs with the specific physiognomies, the separation obtained was not as good as it was expected. This could be because the regions provided in the sensors or its spectral resolution are not enough for such a specialized discrimination. Except for the case of Evergreen forest/Semi-deciduous forest discrimination, which presented a huge difference in the classification performance with MODIS due to the presence of the NIR2 channel, there were many inconsistencies between the relevance of bands in both sensors.

Chapter 8

Conclusions

A Genetic-Programming-based framework for automatic spectral index learning from ground-truth data has been presented, along with one classification scheme for binary classification tasks by applying the indices, and two fusion approaches to combine the indices for multi-class classification purposes. A Genetic Programming pipeline is employed to determine both the most suitable bands and their combination in binary problems. Later, these indices/classifiers are used in two different fusion methods. Performed experiments using well-known datasets demonstrate that the proposed methods are effective, yielding better or comparable results with respect to traditional methods in the literature.

The method was also used for a real-world problem of tropical biomes characterization in which a set of vegetation indices developed by specialists is traditionally used. The results show that the indices learned by means of the framework are superior for discrimination purposes.

8.1 Closing remarks

The results obtained as product of this research prove that Genetic Programming is able to find complex interactions between the reflectance of a target object in different regions of the spectrum, providing spectral indices which, despite their complexity, specialists can interpret and explain real-world phenomena coherently.

Regarding the research questions, the pattern in Figure 5.4a, showing that water absorption bands in the AVIRIS sensor naturally obtained a low relevance, suggests that GP effectively selects bands. Additionally, Tables 5.4 and 6.1 show that the LDA algorithm (the optimal linear projection of the bands) for dimensionality reduction was the only one that the proposed method could not outperform significantly.

8.2 Contribution

The main contribution of this work is a framework that can be used either as a white-box classification system or an analysis tool, by means of the spectral indices learned, that can go through other frameworks (e.g., time series classifiers). The indices can be further analysed in order to recognize interesting patterns made evident in the formulas. The GP

framework can be conditioned to support regression, only by changing the fitness function, and the concept of spectral indices in images applies to classification of regions of interest as well. The framework was implemented in Java, and the source code is available on Github[1] with the name of GPSI (Genetic-Programming-based Spectral Index).

The work documented in Chapters 5, 6 and 7 are suitable to be published. The framework for the binary scenario (Chapter 5) has already been published in the **2016 Conference on Graphics, Patterns and Images (SIBGRAPI)**[2] as a full paper named *Learning to combine spectral indices with genetic programming* [27]. The framework for the multi-class scenario (Chapter 6) has been accepted to the **2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)**[3] as a paper named *Fusion of Genetic-Programming-based Indices in Hyperspectral Image Classification Tasks*. Finally, the contribution for tropical biomes characterization and analysis (Chapter 7) is still under construction and is intended to be submitted to the multidisciplinary journal **Remote Sensing of Environment**[4].

8.3 Future work

Future work includes a more sophisticated (and context-dependent) analysis of the learned formulas to determine band relevance, since in this work only the number of occurrences of the bands in the indices were considered. Tree mining is an interesting option [53] to achieve this.

Since the GP framework can be extended to learn indices that solve regression problems, an interesting application would be to learn indices to describe diverse functional traits.

Considering that the fitness functions that aim at minimizing intra-class distance could yield indices that saturate easily for more specialized tasks, another future work would be the applications of fitness functions that ensure separation of classes without forming compact clusters. For multi-class classification recently proposed fusion approaches can be used [2, 18, 69] to combine GP indices.

A natural extension of this work consists in using the gray-scale images formed by the learned indices to perform classification of regions of interest (more suitable for the current paradigm in multi- and hyper-spectral imagery classification [1]), by including the use of image descriptors.

Finally, the serious problem of high variability of the search in hard spaces could be addressed with non-conventional heuristics in evolutionary algorithms [59, 68].

[1] `https://github.com/jfhernandeza/gpsi`

[2] `http://gibis.unifesp.br/sibgrapi16/index.php`. Last accessed on February, 2017.

[3] `http://igarss2017.org/`. Last accessed on February, 2017.

[4] `https://www.journals.elsevier.com/remote-sensing-of-environment`. Last accessed on February, 2017.

Bibliography

[1] E. A. Addink, F. M.B. Van Coillie, and S. M. De Jong. Introduction to the GEOBIA 2010 special issue: From pixels to geographic objects in remote sensing image analysis. *International Journal of Applied Earth Observation and Geoinformation*, 15:1 – 6, 2012. Special Issue on Geographic Object-based Image Analysis: GEOBIA. 68

[2] N. Alajlan, Y. Bazi, F. Melgani, and R. R. Yager. Fusion of supervised and unsupervised learning for improved classification of hyperspectral images. *Information Sciences*, 217:39 – 55, 2012. 68

[3] M. Aly. Survey on multiclass classification methods. Technical report, California Institute of Technology, 2005. 46

[4] P. Bajcsy and P. Groves. Methodology for hyperspectral band selection. *Photogrammetric Engineering and Remote Sensing Journal*, 70(7):793–802, 2004. 23

[5] A. Bhatt, S. K. Ghosh, and A. Kumar. Spectral indices based object oriented classification for change detection using satellite data. *International Journal of System Assurance Engineering and Management*, pages 1–10, apr 2016. 25

[6] G. Caccamo, L.A. Chisholm, R.A. Bradstock, and M.L. Puotinen. Assessing the sensitivity of MODIS to monitor drought in high biomass ecosystems. *Remote Sensing of Environment*, 115(10):2626 – 2639, 2011. 66

[7] J. B. Campbell and R. H. Wynne. *Introduction to Remote Sensing, Fifth Edition.* Guilford Publications, New York, NY, USA, 2011. 13, 16, 17

[8] C. Chion, J-a Landry, and L. Da Costa. A genetic-programming-based method for hyperspectral data information extraction: Agricultural applications. *IEEE Transactions on Geoscience and Remote Sensing*, 46(8):2446–2457, 2008. 25

[9] T. Costăchioiu and M. Datcu. Land cover dynamics classification using multitemporal spectral indices from satellite image time series. In *2010 8th International Conference on Communications*, pages 157–160, June 2010. 26

[10] M. Črepinšek, S. Liu, and M. Mernik. Exploration and exploitation in evolutionary algorithms: A survey. *ACM Computing Surveys (CSUR)*, 45(3):35, 2013. 21

[11] P. Cunningham. *Dimension Reduction*, pages 91–112. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. 20

[12] R. da S. Torres, A. X. Falcão, M. A. Gonçalves, J. P. Papa, B. Zhang, W. Fan, and E. A. Fox. A genetic programming framework for content-based image retrieval. *Pattern Recognition*, 42(2):283–292, February 2009. 29

[13] A. Damien. *Land and water use classification by means of spectral index based time series analysis*. PhD thesis, Delft University of Technology, 2015. 25

[14] V. de L. Dantas, M. Hirota, R. S. Oliveira, and J. G. Pausas. Disturbance maintains alternative biome states. *Ecology Letters*, 19(1):12–19, 2016. 53

[15] W. A. Dorigo, R. Zurita-Milla, A. J W de Wit, J. Brazile, R. Singh, and M. E. Schaepman. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. *International Journal of Applied Earth Observation and Geoinformation*, 9(2):165–193, 2007. 25

[16] L. C. B. dos Santos, S. J. F. Guimaraes, and J. A. dos Santos. Efficient unsupervised band selection through spectral rhythms. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):1016–1025, Sept 2015. 23

[17] Q. Du and H. Yang. Similarity-based unsupervised band selection for hyperspectral image analysis. *IEEE Geoscience and Remote Sensing Letters*, 5(4):564–568, October 2008. 23

[18] F. A. Faria, J. A. dos Santos, A. Rocha, and R. Torres. A framework for selection and fusion of pattern classifiers in multimedia recognition. *Pattern Recognition Letters*, 39(0):52 – 64, 2014. 68

[19] C. W. B. Fonlupt and D. Robilliard. Genetic Programming with Dynamic Fitness for a Remote Sensing Application. *Parallel Problem Solving from Nature - PPSN VI 6th International Conference*, 1917:191–200, 2000. 24

[20] R. M. Freitas, E. Arai, M. Adami, A. S. Ferreira, O. Y. Sato, Y. E. Shimabukuro, R. R. Rosa, L. O. Anderson, B. Friedrich, and T. Rudorff. Virtual laboratory of remote sensing time series: Visualization of modis evi2 data set over south america. *Journal of Computational Interdisciplinary Sciences*, pages 57–68, 2011. 54

[21] X. Geng, K. Sun, L. Ji, and Y. Zhao. A fast volume-gradient-based band selection method for hyperspectral image. *IEEE Transactions on Geoscience and Remote Sensing*, 52(11):7111–7119, Nov 2014. 23

[22] H. Gerstmann, M. Möller, and C. Gläßer. Optimization of spectral indices and long-term separability analysis for classification of cereal crops using multi-spectral rapideye imagery. *International Journal of Applied Earth Observation and Geoinformation*, 52:115 – 125, 2016. 25

[23] 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, Jan 1988. 24

[24] P. Groves and P. Bajcsy. Methodology for hyperspectral band and classification model selection. In *Workshop on Advances in Techniques for Analysis of Remotely Sensed Data*, pages 120–128, October 2003. 23

[25] B. Guo, S. R. Gunn, R. I. Damper, and J. D. B. Nelson. Band selection for hyperspectral image classification using mutual information. *IEEE Geoscience and Remote Sensing Letters*, 3(4):522–526, October 2006. 23

[26] R. Hedjam and M. Cheriet. Hyperspectral band selection based on graph clustering. In *Information Science, Signal Processing and their Applications*, pages 813 –817, July 2012. 23

[27] J. Hernández, J. A. dos Santos, and R. Torres. Learning to combine spectral indices with genetic programming. In L. A. F. Fernandes D. G. Aliaga, L. S. Davis and W. R. Schwartz, editors, *29th Conference on Graphics, Patterns and Images (SIBGRAPI)*, São José dos Campos, SP, Brazil, 2016. 68

[28] R. Huang and M. He. Band selection based on feature weighting for classification of hyperspectral data. *IEEE Geoscience and Remote Sensing Letters*, 2(2):156–159, April 2005. 23

[29] A. Huete, K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83(1–2):195 – 213, 2002. The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring. 56

[30] G. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, January 1968. 13, 19

[31] A. Ifarraguerri and M.W. Prairie. Visual method for spectral band selection. *IEEE Geoscience and Remote Sensing Letters*, 1(2):101–106, April 2004. 23

[32] M. Imani and H. Ghassemian. Band clustering-based feature extraction for classification of hyperspectral images using limited training samples. *IEEE Geoscience and Remote Sensing Letters*, 11(8):1325–1329, 2014. 24

[33] G. C. Tsuei J. C. Wu. Unsupervised cluster-based band selection for hyperspectral image classification. In *International Conference on Advanced Computer Science and Electronics Information*, page 562–565, July 2013. 23

[34] S. Jia, Y. Qian, and L. Shen. Unsupervised band selection for hyperspectral imagery classification without manual band removal. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):531–543, April 2012. 23

[35] S. Jia, G. Tang, J. Zhu, and Q. Li. A novel ranking-based clustering approach for hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing*, 54(1):88–102, 2016. 23

[36] X. Jia, B. Kuo, and M. M. Crawford. Feature Mining for Hyperspectral Image Classification. *Proceedings of the IEEE*, 101(3):676–697, March 2013. 13, 20

[37] Z. Jiang, A. R. Huete, K. Didan, and T. Miura. Development of a two-band enhanced vegetation index without a blue band. *Remote Sensing of Environment*, 112(10):3833 – 3845, 2008. 56

[38] I. T. Jolliffe. *Principal component analysis*. Springer series in statistics. Springer-Verlang, New York, NY, USA, 2002. 14, 24

[39] X. Kang, S. Li, and J. A. Benediktsson. Feature extraction of hyperspectral images with image fusion and recursive filtering. *IEEE Transactions on Geoscience and Remote Sensing*, 52(6):3742–3752, 2014. 24

[40] M. Karlson, M. Ostwald, H. Reese, J. Sanou, B. Tankoano, and E. Mattsson. Mapping tree canopy cover and aboveground biomass in sudano-sahelian woodlands using landsat 8 and random forest. *Remote Sensing*, 7(8):10017–10041, 2015. 66

[41] L. Kergoat, P. Hiernaux, C. Dardel, C. Pierre, F. Guichard, and A. Kalilou. Dry-season vegetation mass and cover fraction from swir1.6 and swir2.1 band ratio: Ground-radiometer and MODIS data in the sahel. *International Journal of Applied Earth Observation and Geoinformation*, 39:56 – 64, 2015. 66

[42] Z. Khan, F. Shafait, and A. Mian. Joint group sparse pca for compressed hyperspectral imaging. *IEEE Transactions on Image Processing*, 24(12):4934–4942, 2015. 24

[43] J. Kittler. Feature set search algorithm. In *Pattern Recognition and Signal Processing*, pages 41–60, 1978. 23

[44] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992. 14, 20, 22, 36

[45] D. A. Landgrebe. Hyperspectral data classification using nonparametric weighted feature extraction. In *IEEE International Geoscience and Remote Sensing Symposium*, volume 3, pages 1428–1430. IEEE, 2002. 24

[46] W. Li and Q. Du. A survey on representation-based classification and detection in hyperspectral remote sensing imagery. *Pattern Recognition Letters*, 83, Part 2:115 – 123, 2016. Advances in Pattern Recognition in Remote Sensing. 23

[47] Z. Liu, M. C. Wimberly, and F. K. Dwomoh. Vegetation dynamics in the upper guinean forest region of west africa from 2001 to 2015. *Remote Sensing*, 9(1), 2017. 26, 66

[48] F. Löw, P. Knöfel, and C. Conrad. Analysis of uncertainty in multi-temporal object-based classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105:91–106, 2015. cited By 12. 26

[49] V. Malyuto and Th. Schmidt-Kaler. Spectral indices in quantitative spectral classification from stellar libraries. *Astronomische Nachrichten*, 320(2):71–76, 1999. 25

[50] G. McLachlan. *Discriminant analysis and statistical pattern recognition*, volume 544. John Wiley & Sons, Hoboken, NJ, USA, 2004. 24

[51] M. Mitchell. *An Introduction to Genetic Algorithms*. A Bradford book. Bradford Books, 1998. 22, 37

[52] R. Y. M. Nakamura, L. M. Garcia Fonseca, J. A. dos Santos, R. da S. Torres, X. Yang, and J. P. Papa. Nature-inspired framework for hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing*, 52(4):2126–2137, April 2014. 23

[53] S. Nijssen. *Tree Mining*, pages 991–999. Springer US, Boston, MA, 2010. 68

[54] S. Patra, P. Modi, and L. Bruzzone. Hyperspectral band selection based on rough set. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10):5495–5503, 2015. 23

[55] S. J. Perkins, J. P. Theiler, S. P. Brumby, N. R. Harvey, R. B. Porter, J. J. Szymanski, and J. J. Bloch. Genie: a hybrid genetic algorithm for feature classification in multispectral images. *International Symposium on Optical Science and Technology*, 4120:52–62, 2000. 25

[56] J. C. Price. Calibration of satellite radiometers and the comparison of vegetation indices. *Remote Sensing of Environment*, 21(1):15 – 27, 1987. 18

[57] A.V. Prishchepov, V.C. Radeloff, M. Dubinin, and C. Alcantara. The effect of landsat etm/etm+ image acquisition dates on the detection of agricultural land abandonment in eastern europe. *Remote Sensing of Environment*, 126(1):195–209, 2012. 26

[58] C. Puente, G. Olague, S. V. Smith, S. H. Bullock, A. Hinojosa-corona, and M. A. González-botello. A Genetic Programming Approach to Estimate Vegetation Cover in the Context of Soil Erosion Assessment. *Photogrammetric Engineering and Remote Sensing*, 77(4):363–376, 2011. 18, 25

[59] R. Raghu and G. Jeyakumar. Empirical analysis on the population diversity of the sub-population in distributed differential evolution algorithm. *International Journal of Circuit Theory and Applications*, 2015. 68

[60] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971. 31

[61] P. J. Rauss, J. M. Daida, and S. Chaudhary. Classification of spectral imagery using genetic programming. In *Proceedings of the 2nd Annual Conference on Genetic and Evolutionary Computation*, pages 726–733. Morgan Kaufmann Publishers Inc., 2000. 25

[62] B. J. Ross, A. G. Gualtieri, F. Fueten, and P. Budkewitsch. Hyperspectral image analysis using genetic programming. *Applied Soft Computing*, 5(2):147–156, 2005. 25

[63] J. W. Rouse, Jr., R. H. Haas, J. A. Schell, and D. W. Deering. Monitoring Vegetation Systems in the Great Plains with Erts. *NASA Goddard Space Flight Center 3d ERTS-1 Symposium*, 351:309, 1974. 13, 18, 56

[64] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53 – 65, 1987. 30

[65] P. Rufin, H. Müller, D. Pflugmacher, and P. Hostert. Land use intensity trajectories on amazonian pastures derived from landsat time series. *International Journal of Applied Earth Observation and Geoinformation*, 41(1):1 – 10, 2015. 66

[66] A. Viña, A. A. Gitelson, A. L. Nguy-Robertson, and Y. Peng. Comparison of different vegetation indices for the remote assessment of green leaf area index of crops. *Remote Sensing of Environment*, 115(12):3468–3478, 2011. 56

[67] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. 31

[68] D. Whitley, S. Rana, and R. B. Heckendorn. The island model genetic algorithm: On separability, population size and convergence. *Journal of computing and information technology*, 7(1):33–47, 1999. 68

[69] J. Xia, J. Chanussot, P. Du, and X. He. Rotation-based support vector machine ensemble in classification of hyperspectral data with limited training samples. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1519–1531, March 2016. 68

[70] H. Yang, Q. Du, H. Su, and Y. Sheng. An efficient method for supervised hyperspectral band selection. *IEEE Geoscience and Remote Sensing Letters*, 8(1):138–142, January 2011. 23

[71] Y. Yuan, G. Zhu, and Q. Wang. Hyperspectral band selection by multitask sparsity pursuit. *IEEE Transactions on Geoscience and Remote Sensing*, 53(2):631–644, Feb 2015. 23

[72] G. Zhu, Y. Huang, J. Lei, Z. Bi, and F. Xu. Unsupervised hyperspectral band selection by dominant set extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 54(1):227–239, 2016. 23

Appendix A

Other fitness functions considered

In this appendix, two fitness functions besides the ones introduced in Section 4.2 are presented. They were not further considered because of its poor performance.

A.1 Bhattacharyya Distance

The Bhattacharyya Distance is a widely used measure of similarity between two probability distributions. Assuming normality, the distance between two distributions p and q is defined as:

$$D_B(p,q) = \frac{1}{4}\exp\left(\frac{1}{4}\left(\frac{\sigma_p^2}{\sigma_q^2} + \frac{\sigma_q^2}{\sigma_p^2} + 2\right)\right) + \frac{1}{4}\left(\frac{(\mu_p - \mu_q)^2}{\sigma_p^2 + \sigma_q^2}\right) \tag{A.1}$$

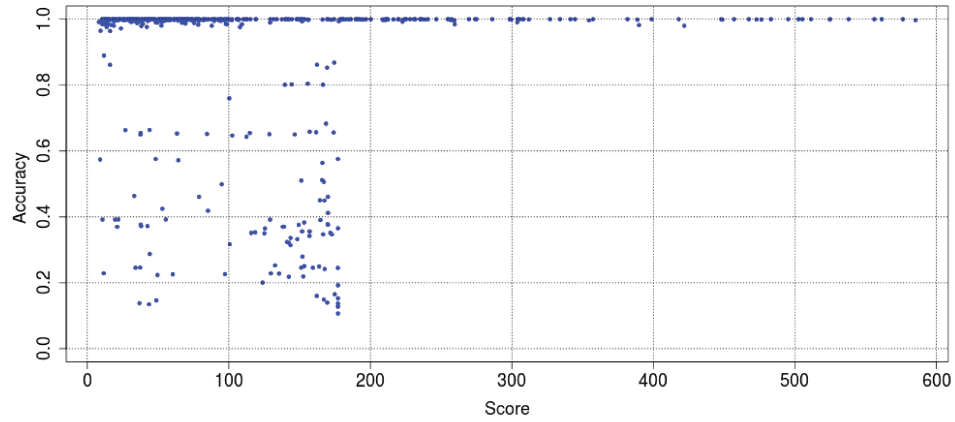
Figure A.1a shows that there is no a clear correlation of the distance with the obtained accuracy. This is because two distributions may be considered as distant, even if they overlap, as long as their standard deviations are significantly different.

A.2 Hellinger Distance

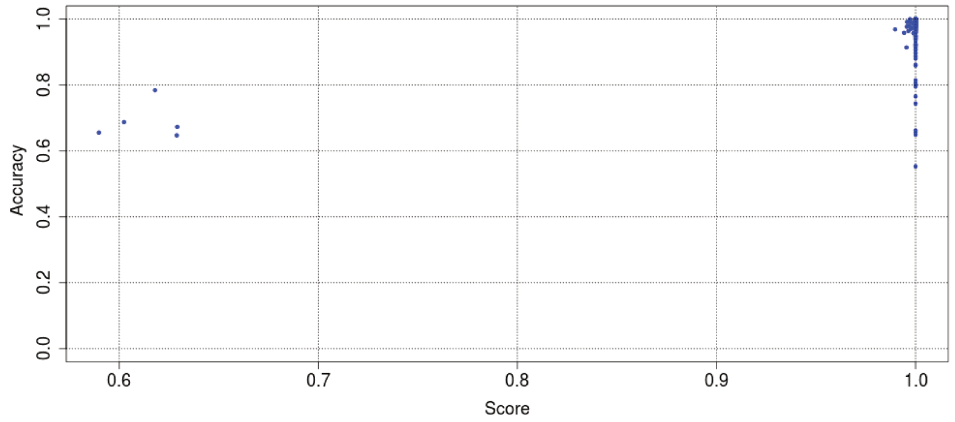
The Bhattacharyya Distance does not obey the triangle inequality. The Hellinger distance address this problem. It was calculated on the basis of the discrete probability distributions $p(x)$ and $q(x)$ of the pixel values, formed by their histograms. The distance score is defined as:

$$H(p,q) = \sqrt{1 - \sum_{x \in X} \sqrt{p(x)q(x)}} \tag{A.2}$$

Figure A.1b shows a better correlation than Bhattacharyya Distance.



(a) Bhattacharyya distance



(b) Hellinger distance

Figure A.1: Accuracy obtained vs. fitness score for the fitness functions that were no longer considered.